В этой книге много интересных примеров того, как сложнейшие технологии Big Data — методы анализа огромных объемов данных — применяются для решения важных задач из нашей повседневной жизни.

Сергей Мацоцкий, *председатель правления компании IBS*

ВИКТОР МАЙЕР-ШЕНБЕРГЕР | КЕННЕТ КУКЬЕР

BIG DATA

БОЛЬШИЕ ДАННЫЕ

РЕВОЛЮЦИЯ, КОТОРАЯ ИЗМЕНИТ ТО, КАК МЫ ЖИВЕМ, РАБОТАЕМ И МЫСЛИМ

Viktor Mayer-Schönberger
Kenneth Cukier

BIG DATA

A Revolution That Will Transform How We Live, Work, and Think

Виктор Майер-Шенбергер
Кеннет Кукьер

БОЛЬШИЕ ДАННЫЕ

Революция, которая изменит то,
как мы живем, работаем и мыслим

Перевод с английского Инны Гайдюк

Издательство «Манн, Иванов и Фербер»
Москва, 2014

Информация от издательства

Опубликовано с разрешения John Wiley & Sons

Майер-Шенбергер, В.

Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер ; пер. с англ. Инны Гайдюк. — М. : Манн, Иванов и Фербер, 2014.

ISBN 978-5-91657-936-9

С появлением новой науки открылась удивительная возможность с точностью предсказывать, что произойдет в будущем в самых разных областях жизни. Большие данные — это наша растущая способность обрабатывать огромные массивы информации, мгновенно их анализировать и получать порой совершенно неожиданные выводы.

По какому цвету покраски можно судить, что подержанный автомобиль находится в отличном состоянии? Как чиновники Нью-Йорка определяют наиболее опасные люки, прежде чем они взорвутся? И как с помощью поисковой системы Google удалось предсказать распространение вспышки гриппа H1N1?

Ключ к ответу на эти и многие другие вопросы лежит в больших данных, которые в ближайшие годы в корне изменят наше представление о бизнесе, здоровье, политике, образовании и инновациях.

От партнера издания

Любимая тема фантастической литературы прошлого века — «каким будет тот момент в будущем, когда машины станут умнее человека?». Кажется, мы сами не заметили, что уже живем в этом будущем. Сегодня человек может с помощью машины справляться с задачами, которые раньше считались практически неразрешимыми. В этой книге приводятся десятки примеров таких задач — от опережающего обнаружения зарождающихся эпидемий до профилактики тяжких преступлений. Многие из приведенных примеров поражают воображение и кажутся настоящей фантастикой!

Но самое интересное в этой книге — рассказ о том, почему ранее неразрешимые задачи сегодня становятся объектом внимания математиков и компьютерщиков. Авторы рисуют картину, как множество больших и маленьких вычислительных устройств, которыми наполнен современный мир, ежесекундно генерируют гигантские массивы цифровой информации. И как эта информация, собранная вместе и проанализированная с помощью современных высокопроизводительных компьютеров, позволяет получить качественно новое понимание того, что содержит эта информация. И как в конечном счете это позволяет отвечать на вопросы, которые раньше не имели ответов.

Этот переход количества накопленной человечеством информации в качество решения задач, стоящих перед нами, называют сейчас феноменом «больших данных», и сегодня это одно из самых обсуждаемых явлений в индустрии информационных технологий. О нем много говорят специалисты, но, пожалуй, еще очень мало знают обычные пользователи цифровых технологий.

Между тем мы уже живем в новой эпохе — эпохе больших данных. Изменения, которые несут новые информационные технологии, затрагивают жизнь каждого человека.

«Большие данные» — это масса новых задач, касающихся общественной безопасности, глобальных экономических моделей, неприкосновенности частной жизни, устоявшихся моральных правил,

правовых отношений человека, бизнеса и государства. Похоже, что в ближайшем будущем нам всем придется столкнуться с фантастическим уровнем прозрачности всей нашей жизни, действий и поступков. Этические вопросы, возникающие в связи с этим, в книге отчасти сформулированы, как и возможные ответы на них, однако только жизнь покажет, насколько правильно мы видим все риски и проблемы.

Очень хотелось бы, чтобы в будущих изданиях на тему «больших данных» среди рассматриваемых примеров нашлось достойное место и для ярких решений, созданных талантливыми российскими математиками и программистами, которые уже сейчас добились успехов в этой области. Наши разработки используются в больших энергетических сетях, крупнейших банках, в анализе информации в интернете и для работы со СМИ. У России огромный потенциал в этой области благодаря сильной математической школе и сложившейся за десятилетия качественной системе подготовки инженерных кадров. Наша страна может стать одним из флагманов нового глобального технологического тренда.

Надеемся, для многих читателей эта книга станет поводом задуматься над тем, что такое «большие данные» и каким образом эти технологии — такие неосязаемые и невесомые — стали силой, изменяющей мир. Развитие и внедрение технологий «больших данных» может дать уникальные конкурентные преимущества бизнесу, помочь построить более эффективное государство, предоставить новые возможности людям и в конечном итоге сделать нашу жизнь более удобной и безопасной. Кто знает, может быть, возникшие благодаря прочтению этой книги идеи дадут впоследствии импульс для развития такой перспективной индустрии «больших данных».

Сергей Мацоцкий,
председатель правления компании IBS

Глава 1

Наше время

В 2009 году был обнаружен новый штамм вируса гриппа — H1N1. Он включал в себя элементы вирусов, которые вызывают птичий и свиной грипп. Новый вирус быстро распространился и в считаные недели вызвал в государственных учреждениях здравоохранения по всему миру опасения, что надвигается страшная пандемия. Некоторые источники предупреждали о возможности масштабной вспышки эпидемии, подобной «испанке» 1918 года. Тогда от нее пострадало полмиллиарда человек, десятки миллионов погибли. Что хуже всего, против нового вируса не было вакцины. Единственная надежда органов здравоохранения состояла в том, чтобы замедлить распространение вируса. Но для этого требовалось знать его очаги.

В США, как и в других странах, центры по контролю и профилактике заболеваний (CDC) обязали врачей сообщать о новых случаях гриппа. И все-таки информация о возникшей пандемии каждый раз запаздывала на одну-две недели. Люди по-прежнему обращались к врачу лишь спустя несколько дней после первых признаков недомогания. Вдобавок время уходило на то, чтобы передать эту информацию в CDC. Организация лишь констатировала количество случаев каждую неделю. При быстром распространении заболевания отстать на две недели означало безнадежно опоздать. Из-за этой задержки государственные учреждения здравоохранения вынуждены были действовать вслепую в самые ответственные моменты.

За несколько недель до того, как сведения об H1N1 попали на первые полосы газет, инженеры интернет-гиганта Google опубликовали потрясающую статью в научном журнале Nature[1]. Она произвела настоящий фурор среди медицинских чиновников и программистов, но не привлекла интереса широкой аудитории. Речь шла о том, как компания Google может «предсказать» распространение зимнего гриппа в США не только в масштабах страны, но и в

отдельных регионах и даже штатах. Чтобы добиться такого результата, специалисты Google проанализировали поисковые запросы интернет-пользователей. Более трех миллиардов поисковых запросов, отправляемых в поисковую систему Google ежедневно со всего мира, составили огромный массив данных для обработки. Пригодилось и то, что Google хранит все поисковые запросы в течение многих лет.

Специалисты Google взяли 50 миллионов наиболее распространенных условий поиска, которые используют американцы, и сравнили их с данными CDC о распространении сезонного гриппа в период между 2003 и 2008 годами. Идея заключалась в том, что людей, подхвативших вирус гриппа, можно определить по тому, что они ищут в интернете. Предпринимались и другие попытки связать эти показатели с данными интернет-поиска, но никто не располагал таким объемом данных, вычислительными мощностями и статистическими ноу-хау, как Google.

В Google предположили, что в интернете существуют поисковые запросы на получение информации о гриппе (например, «средство от кашля и температуры»), но не знали, какие именно. Поэтому была разработана универсальная система, все действие которой сводилось к тому, чтобы находить корреляции между частотой определенных поисковых запросов и распространением гриппа во времени и пространстве. В общей сложности поисковая система Google обработала ошеломляющее количество различных математических моделей (450 миллионов) с целью проверки условий поиска. Для этого прогнозируемые значения сравнивались с фактическими данными CDC о случаях гриппа за 2007–2008 годы. Специалисты Google нашли золотую жилу: их программное обеспечение выявило сочетание 45 условий поиска, использование которых с математической моделью давало коэффициент корреляции между прогнозируемыми и официальными данными, равный 97%. Как и CDC, специалисты компании могли назвать территорию распространения гриппа. Но, в отличие от CDC, они делали это практически в режиме реального времени, а не спустя одну-две недели.

Таким образом, когда в 2009 году распространение вируса H1N1 достигло критических показателей, система оказалась гораздо более

полезным и своевременным индикатором[2], чем официальная статистика правительства с ее естественным отставанием из-за бюрократической волокиты. Сотрудники здравоохранения получили ценную информацию. Самое примечательное, метод компании Google позволяет обходиться без марлевых повязок и визитов к врачу. По сути, он создан на основе «больших данных» — способности общества по-новому использовать информацию для принятия взвешенных решений или производства товаров и услуг, имеющих большое значение. Благодаря этому методу к моменту приближения следующей пандемии мир будет владеть эффективным инструментом для ее прогнозирования, а значит, сможет предупредить ее распространение.

Здравоохранение — только одна из областей, в которых большие данные приносят ощутимую пользу. Они приводят к коренному преобразованию целых отраслей. Наглядный тому пример — покупка авиабилетов[3].

В 2003 году Орен Эциони[1] собрался лететь из Сиэтла в Лос-Анджелес на свадьбу своего младшего брата. За несколько месяцев до этого знаменательного события он купил авиабилет через интернет, зная, что чем раньше возьмешь билет, тем дешевле он обойдется. Во время перелета Эциони не удержался от любопытства и спросил попутчика, сколько тот заплатил за билет. Оказалось, что значительно меньше, хотя билет был куплен намного позже. От возмущения Эциони стал опрашивать других пассажиров — и все они заплатили меньше.

У большинства людей ощущение экономического предательства растаяло бы прежде, чем они сложили откидной столик и перевели спинку кресла в вертикальное положение. Но Эциони — один из передовых американских ученых в сфере компьютерных технологий. Будучи руководителем программы искусственного интеллекта в Вашингтонском университете, он основал множество компаний, занимающихся обработкой больших данных, еще до того, как термин «большие данные» приобрел известность.

В 1995 году Эциони помог создать одну из первых поисковых систем — MetaCrawler, которая, став главным онлайн-ресурсом, была

выкуплена компанией InfoSpace. Он стал одним из основателей Netbot — первой крупной программы для сравнения цен в магазинах, позже проданной компании Excite. Его стартап ClearForest для анализа текстовых документов приобрела компания Reuters. Эциони рассматривает мир как одну большую компьютерную проблему, которую он способен решить. И ему довелось решить немало таких проблем, после того как он окончил Гарвард в 1986 году одним из первых выпускников по специальности в области программирования.

Приземлившись, Эциони был полон решимости найти способ, который помог бы определить выгодность той или иной цены в интернете. Место в самолете — это товар. Все места на один рейс в целом одинаковы. А цены на них разительно отличаются в зависимости от множества факторов, полный список которых известен лишь самим авиакомпаниям.

Эциони пришел к выводу, что не нужно учитывать все нюансы и причины разницы в цене. Нужно спрогнозировать вероятность того, что отображаемая цена возрастет или упадет. А это вполне осуществимо, причем без особого труда. Достаточно проанализировать все продажи билетов по заданному маршруту, а также соотношение цен и количества дней до вылета.

Если средняя цена билета имела тенденцию к снижению, стоило подождать и купить билет позже. Если же к увеличению — система рекомендовала сразу же приобрести билет по предложенной цене. Другими словами, получилась новоиспеченная версия неформального опроса, который Эциони провел на высоте боле 9000 метров. Безусловно, это была сложнейшая задача по программированию. Но Эциони приступил к работе.

Используя 12-тысячную выборку цен за 41 день, с трудом собранную на сайте путешествий, Эциони создал модель прогнозирования, которая обеспечивала его условным пассажирам неплохую экономию. Система понимала только *что*, но не имела представления *почему*. То есть не брала в расчет переменные, влияющие на ценовую политику авиакомпании, например количество непроданных мест, сезонность или непредвиденную задержку рейса, которые могли снизить стоимость перелета. Ее задача заключалась только в составлении прогноза исходя из вероятностей, рассчитанных

на основе данных о других рейсах. «Покупать или не покупать, вот в чем вопрос», — размышлял Эциони. И назвал исследовательский проект соответственно — «Гамлет»[4].

Небольшой проект превратился в стартап Farecast с венчурным финансированием. Прогнозируя вероятность и значение роста или снижения цены на авиабилет, он дал возможность потребителям выбирать, когда именно совершать покупку. Он вооружил их ранее недоступной информацией. В ущерб себе служба Farecast была настолько прозрачной, что оценивала даже степень доверия к собственным прогнозам и предоставляла эту информацию пользователям.

Для работы системы требовалось большое количество данных. Для того чтобы повысить эффективность системы, Эциони раздобыл одну из отраслевых баз данных бронирования авиабилетов. Благодаря этой информации система создавала прогнозы по каждому месту каждого рейса американской коммерческой авиации по всем направлениям в течение года. Теперь для прогнозирования в Farecast обрабатывалось около 200 миллиардов записей с данными о рейсах, при этом потребителям обеспечивалась значительная экономия.

Брюнет с широкой улыбкой и ангельской внешностью, Эциони вряд ли походил на человека, который отказался бы от миллионов долларов потенциального дохода авиационной отрасли. На самом деле он нацелился выше. К 2008 году Эциони планировал применить этот метод в других областях, например к гостиничному бизнесу, билетам на концерты и подержанным автомобилям, — к чему угодно, где прослеживаются низкая дифференциация продукта, высокая степень колебания цен и огромное количество данных. Но прежде чем он успел реализовать свои планы, в его дверь постучалась корпорация Microsoft и выкупила службу Farecast за 110 миллионов долларов США[5], после чего интегрировала ее в поисковую систему Bing. К 2012 году система прогнозировала цены на авиабилеты для всех внутренних рейсов США, анализируя около триллиона записей. В 75% случаев система оказывалась права и позволяла путешественникам экономить на билете в среднем 50 долларов.

Farecast — это воплощение компании, которая оперирует большими данными; наглядный пример того, к чему идет мир. Эциони не смог бы создать такую компанию пять или десять лет назад. По его словам, «это было бы невозможно». Необходимое количество вычислительных мощностей и хранилище обошлись бы слишком дорого. И хотя важнейшим фактором, сыгравшим на руку, стали изменения технологий, изменилось еще кое-что — едва уловимое, но более важное: само представление о том, как использовать данные.

Данные больше не рассматривались как некая статичная или устаревшая величина, которая становится бесполезной по достижении определенной цели, например после приземления самолета (или в случае Google — после обработки поискового запроса). Скорее, они стали сырьевым материалом бизнеса, жизненно важным экономическим вкладом, используемым для создания новой экономической выгоды. Оказалось, что при правильном подходе их можно ловко использовать повторно, в качестве источника инноваций и новых услуг. Данные могут раскрыть секреты тем, кто обладает смирением и готовностью «слушать», а также необходимыми инструментами.

Данные говорят сами за себя

Приметы информационного общества нетрудно заметить повсюду: в каждом кармане найдется мобильный телефон, на каждом столе — компьютер, а в рабочих кабинетах по всему миру — большие ИТ-системы. Но сама информация при этом менее заметна. Полвека спустя с того времени, как компьютеры прочно вошли в жизнь общества, накопление данных достигло того уровня, на котором происходит нечто новое и необычное. Мир не просто завален небывалым количеством информации — это количество стало расти быстрее. Изменение масштаба привело к изменению состояния. Количественное изменение привело к качественному. В науках, таких как астрономия и геномика, впервые столкнувшихся со всплеском данных в середине 2000-х годов, появился термин «большие данные». Теперь эта концепция проникает во все сферы человеческой деятельности.

Для «больших данных» нет строгого определения. Изначально идея состояла в том, что объем информации настолько вырос, что рассматриваемое количество уже фактически не помещалось в памяти компьютера, используемой для обработки, поэтому инженерам потребовалось модернизировать инструменты для анализа всех данных. Так появились новые технологии обработки, например модель MapReduce компании Google и ее аналог с открытым исходным кодом — Hadoop от компании Yahoo. Они дали возможность управлять намного б*о*льшим количеством данных, чем прежде. При этом важно, что их не нужно было выстраивать в аккуратные ряды или классические таблицы баз данных. На горизонте также появились другие технологии обработки данных, которые обходились без прежней жесткой иерархии и однородности. В то же время интернет-компании, имеющие возможность собирать огромные массивы данных и острый финансовый стимул для их анализа, стали ведущими пользователями новейших технологий обработки, вытесняя компании, которые порой имели на десятки лет больше опыта, но работали автономно.

Согласно одному из подходов к этому вопросу (который мы рассматриваем в этой книге), понятие «большие данные» относится к операциям, которые можно выполнять исключительно в большом масштабе. Это порождает новые идеи и позволяет создавать новые формы стоимости, тем самым изменяя рынки, организации, отношения между гражданами и правительствами, а также многое другое.

И это только начало. Эпоха больших данных ставит под вопрос наш образ жизни и способ взаимодействия с миром. Поразительнее всего то, что обществу придется отказаться от понимания причинности в пользу простых корреляций: променять знание *почему* на *что именно.* Это переворачивает веками установленный порядок вещей и ставит под сомнение наши фундаментальные знания о том, как принимать решения и постигать действительность.

Большие данные знаменуют начало глубоких изменений. Подобно тому как телескоп дал нам возможность постичь Вселенную, а микроскоп — получить представление о микробах, новые методы сбора и анализа огромного массива данных помогут разобраться в окружающем мире с использованием способов, ценность которых мы

только начинаем осознавать. Но настоящая революция заключается не в компьютерах, которые вычисляют данные, а в самих данных и в том, как мы их используем.

Чтобы понять, на каком этапе находится информационная революция, рассмотрим существующие тенденции. Наша цифровая Вселенная постоянно расширяется. Возьмем астрономию.

Когда в 2000 году стартовал проект «Слоуновский цифровой обзор неба», его телескоп в Нью-Мексико за первые несколько недель собрал больше данных, чем накопилось за всю историю астрономии. К 2010 году его архив был забит грандиозным количеством информации: 140 терабайт. А его преемник, телескоп Large Synoptic Survey Telescope, который введут в эксплуатацию в Чили в 2016 году, будет получать такое количество данных каждые пять дней[6].

За подобными астрономическими цифрами не обязательно далеко ходить. В 2003 году впервые в мире расшифровали геном человека, после чего еще десять лет интенсивной работы ушло на построение последовательности из трех миллиардов основных пар. Прошел почти десяток лет — и то же количество ДНК анализируется каждые 15 минут с помощью геномных машин по всему миру[7]. В 2012 году стоимость определения последовательности генома человека упала ниже одной тысячи долларов. Эта процедура стала доступной широким массам. Что касается области финансов, через фондовые рынки США каждый день проходит около семи миллиардов обменных операций, из них около двух третей торгов решаются с помощью компьютерных алгоритмов на основе математических моделей, которые обрабатывают горы данных, чтобы спрогнозировать прибыль, снижая при этом по возможности риски.

Перегруженность в особенности коснулась интернет-компаний. Google обрабатывает более петабайта данных в день — это примерно в 100 раз больше всех печатных материалов Библиотеки Конгресса США. Facebook — компания, которой не было в помине десятилетие назад, — может похвастать более чем 10 миллионами загрузок новых фотографий ежечасно. Люди нажимают кнопку «Нравится» или пишут комментарии почти три миллиарда раз в день, оставляя за собой цифровой след, с помощью которого компания изучает предпочтения

пользователей[8]. А 800 миллионов ежемесячных пользователей службы YouTube компании Google каждую секунду загружают видео длительностью более часа[9]. Количество сообщений в Twitter увеличивается приблизительно на 200% в год и к 2012 году превысило 400 миллионов твитов в день[10].

От науки до здравоохранения, от банковского дела до интернета... Сферы могут быть разными, но итог один: объем данных в мире быстро растет, опережая не только наши вычислительные машины, но и воображение.

Немало людей пыталось оценить реальный объем окружающей нас информации и рассчитать темп ее роста. Они достигли разного успеха, поскольку измеряли разные вещи. Одно из наиболее полных исследований провел Мартин Гилберт из школы коммуникаций им. Анненберга при Университете Южной Калифорнии[11]. Он стремился сосчитать все, что производилось, хранилось и передавалось. Это не только книги, картины, электронные письма, фотографии, музыка и видео (аналоговые и цифровые), но и видеоигры, телефонные звонки и даже автомобильные навигационные системы, а также письма, отправленные по почте. Он также брал в расчет вещательные СМИ, телевидение и радио, учитывая охват аудитории.

По его расчетам, в 2007 году хранилось или отправлялось примерно 2,25 зеттабайта данных. Это примерно в пять раз больше, чем 20 лет назад (около 435 экзабайт). Чтобы представить это наглядно, возьмем полнометражный художественный фильм. В цифровом виде его можно сжать до файла размером в один гигабайт. Экзабайт состоит из миллиарда гигабайт. Зеттабайт — примерно в тысячу раз больше. Проще говоря, немыслимо много.

Если рассматривать только хранящуюся информацию, не включая вещательные СМИ, проявляются интересные тенденции. В 2007 году насчитывалось примерно 300 экзабайт сохраненных данных, из которых около 7% были представлены в аналоговом формате (бумажные документы, книги, фотоснимки и т. д.), а остальные — в цифровом. Однако совсем недавно наблюдалась иная картина. Хотя идея «информационного века» и «цифровой деревни» родилась еще в 1960-х годах, это действительно довольно новое явление, учитывая

некоторые показатели. Еще в 2000 году количество информации, хранящейся в цифровом формате, составляло всего одну четверть общего количества информации в мире. А остальные три четверти содержались в бумажных документах, на пленке, виниловых грампластинках, магнитных кассетах и подобных носителях.

В то время цифровой информации насчитывалось не так много — шокирующий факт для тех, кто уже продолжительное время пользуется интернетом и покупает книги онлайн. (В 1986 году около 40% вычислительной мощности общего назначения в мире приходилось на карманные калькуляторы, вычислительная мощность которых была больше, чем у всех персональных компьютеров того времени.) Из-за быстрого роста цифровых данных (которые, согласно Гилберту, удваивались каждые три с лишним года) ситуация стремительно менялась. Количество аналоговой информации, напротив, практически не увеличивалось.

Таким образом, к 2013 году количество хранящейся информации в мире составило 1,2 зеттабайта, из которых на нецифровую информацию приходится менее 2%[12].

Трудно представить себе такой объем данных. Если записать данные в книгах, ими можно было бы покрыть всю поверхность Соединенных Штатов в 52 слоя. Если записать данные на компакт-диски и сложить их в пять стопок, то каждая из них будет высотой до Луны. В III веке до н. э. считалось, что весь интеллектуальный багаж человечества хранится в великой Александрийской библиотеке, поскольку египетский царь Птолемей II стремился сохранить копии всех письменных трудов. Сейчас же в мире накопилось столько цифровой информации, что на каждого живущего ее приходится в 320 раз больше, чем хранилось в Александрийской библиотеке.

Процессы действительно ускоряются. Объем хранящейся информации растет в четыре раза быстрее, чем мировая экономика, в то время как вычислительная мощность компьютеров увеличивается в девять раз быстрее. Неудивительно, что люди жалуются на информационную перегрузку. Всех буквально захлестнула волна изменений.

Рассмотрим перспективы, сравнив текущий поток данных с более ранней информационной революцией. Она была связана с изобретением ручного типографского станка Гутенберга около 1450 года. По данным историка Элизабет Эйзенштейн, за 50 лет — с 1453 по 1503 год — напечатано около восьми миллионов книг. Это больше, чем все книжники Европы произвели с момента основания Константинополя примерно 1650 годами ранее[13]. Другими словами, потребовалось 50 лет, чтобы приблизительно вдвое увеличить информационный фонд всей Европы (в то время, вероятно, она представляла львиную долю всего мирового запаса слов). Для сравнения: сегодня это происходит каждые три дня.

Что означает это увеличение? Питер Норвиг, эксперт по искусственному интеллекту в компании Google, прежде работавший в Лаборатории реактивного движения НАСА, любит в этом случае проводить аналогию с изображениями[14]. Для начала он предлагает взглянуть на наскальные изображения лошади в пещере Ласко во Франции, которые относятся к эпохе палеолита (17 тысяч лет назад). Затем — на фотографию лошади или, еще лучше, работы кисти Пабло Пикассо, которые по виду не слишком отличаются от наскальных рисунков. Между прочим, когда Пикассо показали изображения Ласко, он саркастически заметил: «[С тех пор] мы ничего не изобрели»[15].

Он был прав, но лишь отчасти. Вернемся к фотографии лошади. Если раньше, чтобы нарисовать лошадь, приходилось потратить много времени, теперь ее можно запечатлеть гораздо быстрее. В этом и состоит изменение. Хотя оно может показаться не столь важным, поскольку результат по большому счету одинаков: изображение лошади. А теперь представьте, как делается снимок лошади, и ускорьте его до 24 кадров в секунду. Теперь количественное изменение переросло в качественное. Фильм коренным образом отличается от стоп-кадра. То же самое и с большими данными: изменяя количество, мы меняем суть.

Из курса физики и биологии нам известно, что изменение масштаба иногда приводит к изменению состояния. Обратимся к другой аналогии, на сей раз из области нанотехнологий, где речь идет об уменьшении объектов, а не их увеличении. Принцип, лежащий в

основе нанотехнологий, заключается в том, что на молекулярном уровне физические свойства меняются. Появляется возможность придать материалам характеристики, недоступные ранее. Например, медь, которая в обычном состоянии проводит электричество, на наноуровне обнаруживает сопротивление в присутствии магнитного поля, а серебро имеет более выраженные антибактериальные свойства. Гибкие металлы и эластичная керамика тоже возможны на наноуровне. Подобным образом при увеличении масштаба обрабатываемых данных появляются новые возможности, недоступные при обработке меньших объемов.

Иногда ограничения, которые мы воспринимаем как должное и считаем всеобщими, на самом деле имеют место только в масштабе нашей деятельности. Рассмотрим третью аналогию, и на сей раз из области науки. Для людей важнейшим физическим законом является гравитация: она распространяется на все сферы нашей деятельности. Но для мелких насекомых гравитация несущественна. Ограничение, действующее в их физической вселенной, — поверхностное натяжение, позволяющее им, например, ходить по воде. Но людям, как правило, до этого нет дела.

То же самое с информацией: размер имеет значение. Так, поисковая система Google определяет распространение гриппа не хуже, чем официальная статистика, основанная на реальных визитах пациентов к врачу. Для этого системе нужно произвести тщательный анализ сотен миллиардов условий поиска, в результате чего она дает ответ в режиме реального времени, то есть намного быстрее, чем официальные источники. Таким же образом система Farecast прогнозирует колебания цен на авиабилеты, вручая потребителям эффективный экономический инструмент. Однако обе системы достигают этого лишь путем анализа сотен миллиардов точек данных.

Эти два примера, с одной стороны, демонстрируют научное и общественное значение больших данных, а с другой — показывают, что с их помощью можно извлечь экономическую выгоду. Они знаменуют два способа, которыми мир больших данных готов радикально изменить все: от бизнеса и естественных наук до здравоохранения, государственного управления, образования, экономики, гуманитарных наук и других аспектов жизни общества.

Мы стоим на пороге эпохи больших данных, однако полагаемся на них ежедневно. Спам-фильтры разрабатываются с учетом автоматической адаптации к изменению типов нежелательных электронных писем, ведь программное обеспечение нельзя запрограммировать таким образом, чтобы блокировать слово «виагра» или бесконечное количество его вариантов. Сайты знакомств подбирают пары на основе корреляции многочисленных атрибутов с теми, кто ранее составил удачные пары. Функция автозамены в смартфонах отслеживает действия пользователя и добавляет новые вводимые слова в свой орфографический словарь. И это только начало. От автомобилей, способных определять момент для поворота или торможения, до компьютеров IBM Watson, которые обыгрывают людей на игровом шоу Jeopardy[2], — этот подход во многом изменит наше представление о мире, в котором мы живем.

По сути, большие данные предназначены для прогнозирования. Обычно их описывают как часть компьютерной науки под названием «искусственный интеллект» (точнее, ее раздел «машинное обучение»). Такая характеристика вводит в заблуждение, поскольку речь идет не о попытке «научить» компьютер «думать», как люди. Вместо этого рассматривается применение математических приемов к большому количеству данных для прогноза вероятностей, например таких: что электронное письмо является спамом; что вместо слова «коипя» предполагалось набрать «копия»; что траектория и скорость движения человека, переходящего дорогу в неположенном месте, говорят о том, что он успеет перейти улицу вовремя и автомобилю нужно лишь немного снизить скорость. Но главное — эти системы работают эффективно благодаря поступлению большого количества данных, на основе которых они могут строить свои прогнозы. Более того, системы спроектированы таким образом, чтобы со временем улучшаться за счет отслеживания самых полезных сигналов и моделей по мере поступления новых данных.

В будущем — и даже раньше, чем мы можем себе это представить, — многие аспекты нашей жизни, которые сегодня являются единственной сферой человеческих суждений, будут дополнены или заменены компьютерными системами. И это касается не только

вождения или подбора пары, но и более сложных задач. В конце концов, Amazon может порекомендовать идеально подходящую книгу, Google — оценить релевантность сайта, Facebook знает, что нам нравится, а LinkedIn предвидит, с кем мы знакомы. Аналогичные технологии будут применяться для диагностики заболеваний, рекомендации курса лечения, возможно, даже для определения «преступников», прежде чем они успеют совершить преступление.

Подобно тому как интернет радикально изменил мир, добавив связь между компьютерами, большие данные изменят фундаментальные аспекты жизни, предоставив миру небывалые возможности количественного измерения. Данные порождают новые услуги и инновации. И очень многое ставят под угрозу.

Количество, точность, причинность

По сути, большие данные представляют собой три шага к новому способу анализа информации, которые трансформируют наше представление об обществе и его организации.

Первый шаг описан во второй главе. В мире больших данных мы можем проанализировать огромное количество данных, а в некоторых случаях — обработать *все* данные, касающиеся того или иного явления, а не полагаться на случайные выборки. Начиная с XIX века, сталкиваясь с большими числами, общество полагалось на метод выборки. Сейчас он воспринимается как пережиток времен дефицита информации, продукт естественных ограничений для взаимодействия с информацией в «аналоговую эпоху». Понять искусственность этих ограничений, которые по большей части принимались как должное, удалось только после того, как высокопроизводительные цифровые технологии получили широкое распространение. Используя все данные, мы получаем более точный результат и можем увидеть нюансы, недоступные при ограничении небольшим объемом данных. Большие данные дают особенно четкое представление о деталях подкатегорий и сегментов, которые невозможно оценить с помощью выборки.

Принимая во внимание гораздо больший объем данных, мы можем снизить свои претензии к точности — и это второй шаг, который будет

рассмотрен в третьей главе. Когда возможность измерения ограничена, подсчитываются только самые важные показатели, и стремление получить точное число вполне целесообразно. Вряд ли вы сумеете продать скот покупателю, если он не уверен, сколько голов в стаде — 100 или только 80. До недавнего времени все наши цифровые инструменты были основаны на точности: мы считали, что системы баз данных должны извлекать записи, идеально соответствующие нашим запросам, равно как числа вносятся в столбцы электронных таблиц.

Этот способ мышления свойствен среде «малых данных». Измерялось так мало показателей, что следовало как можно точнее подсчитывать все записанное. В некотором смысле мы уже ощутили разницу: небольшой магазин в состоянии подбить кассу к концу дня вплоть до копейки, но мы не стали бы (да и не смогли бы) проделать то же самое с валовым внутренним продуктом страны. Чем больше масштаб, тем меньше мы гонимся за точностью.

Точность требует тщательной проверки данных. Она подходит для небольших объемов данных и в некоторых случаях, безусловно, необходима (например, чтобы проверить, достаточно ли средств на банковском счету, и выписать чек). Но в мире больших данных строгая точность невозможна, а порой и нежелательна. Если мы оперируем данными, большинство которых постоянно меняется, абсолютная точность уходит на второй план.

Большие данные неупорядочены, далеко не все одинакового качества и разбросаны по бесчисленным серверам по всему миру. Имея дело с большими данными, как правило, приходится довольствоваться общим представлением, а не пониманием явления вплоть до дюйма, копейки или молекулы. Мы не отказываемся от точности как таковой, а лишь снижаем свою приверженность к ней. То, что мы теряем из-за неточности на микроуровне, позволяет нам делать открытия на макроуровне.

Эти два шага приводят к третьему — отходу от вековых традиций поиска причинности, который мы рассмотрим в четвертой главе. Люди привыкли во всем искать причины, даже если установить их не так просто или малополезно. С другой стороны, в мире больших данных мы больше не обязаны цепляться за причинность. Вместо этого мы

можем находить корреляции между данными, которые открывают перед нами новые неоценимые знания. Корреляции не могут сказать нам точно, *почему* происходит то или иное событие, зато предупреждают о том, *какого оно рода*. И в большинстве случаев этого вполне достаточно.

Например, если электронные медицинские записи показывают, что в определенном сочетании апельсиновый сок и аспирин способны излечить от рака, то точная причина менее важна, чем сам факт: лечение эффективно. Если мы можем сэкономить деньги, зная, когда лучше купить авиабилет, но при этом не имеем представления о том, что стоит за их ценообразованием, этого вполне достаточно. Вопрос не в том *почему*, а в том *что*. В мире больших данных нам не всегда нужно знать причины, которые стоят за теми или иными явлениями. Лучше позволить данным говорить самим за себя.

Нам больше не нужно ограничиваться проверкой небольшого количества гипотез, тщательно сформулированных задолго до сбора данных. Позволив данным «говорить», мы можем уловить корреляции, о существовании которых даже не подозревали. В связи с этим хедж-фонды анализируют записи в Twitter, чтобы прогнозировать работу фондового рынка. Amazon и Netflix рекомендуют продукты исходя из множества взаимодействий пользователей со своими сайтами. А Twitter, LinkedIn и Facebook выстраивают «социальные графы» отношений пользователей для изучения их предпочтений.

Разумеется, люди анализировали данные в течение тысячелетий. И письменность в древней Месопотамии появилась благодаря тому, что счетоводам нужен был эффективный инструмент для записи и отслеживания информации. С библейских времен правительства проводили переписи для сбора огромных наборов данных о своем населении, и в течение двухсот лет актуарии собирали ценнейшие данные о рисках, которые они надеялись понять или хотя бы избежать.

В «аналоговую эпоху» сбор и анализ таких данных был чрезвычайно дорогостоящим и трудоемким. Появление новых вопросов, как правило, означало необходимость в повторном сборе и анализе данных.

Большим шагом на пути к более эффективному управлению данными стало появление оцифровки — перевода аналоговой

информации в доступную для чтения на компьютерах, что упрощало и удешевляло ее хранение и обработку. Это значительно повысило эффективность. То, на что раньше уходили годы сбора и вычисления, теперь выполнялось за несколько дней, а то и быстрее. Но, кроме этого, мало что изменилось. Люди, занимающиеся анализом данных, были слишком погружены в аналоговую парадигму, предполагая, что наборы данных имели единственное предназначение, в котором и заключалась их ценность. Сама технология закрепила этот предрассудок. И хотя оцифровка важнейшим образом способствовала переходу на большие данные, сам факт существования компьютеров не обеспечил этот переход.

Трудно описать нынешнюю ситуацию существующими понятиями. Для того чтобы в целом очертить изменения, воспользуемся датификацией *(data-ization)* — концепцией, с которой познакомим вас в пятой главе. Речь идет о преобразовании в формат данных всего, что есть на планете, включая то, что мы никогда не рассматривали как информацию (например, местоположение человека, вибрации двигателя или нагрузку на мост), путем количественного анализа. Это открывает перед нами новые возможности, такие как прогнозный анализ. Он позволяет обнаружить, например, что двигатель вот-вот придет в неисправность, исходя из его перегрева или производимых им вибраций. В результате мы можем открыть неявное, скрытое значение информации.

Полным ходом ведется «поиск сокровищ» — извлечение ценных идей из данных и раскрытие их потенциала путем перехода от причинности к корреляции. Это стало возможным благодаря новым техническим средствам. Но сокровища заключаются не только в этом. Вполне вероятно, что каждый набор данных имеет внутреннюю, пока еще не раскрытую ценность, и весь мир стремится обнаружить и заполучить ее.

Большие данные вносят коррективы в характер бизнеса, рынков и общества, о которых подробнее мы поговорим в шестой и седьмой главах. В XX веке особое значение придавалось не физической инфраструктуре, а нематериальным активам, не земле и заводам, а интеллектуальной собственности. Сейчас общество идет к тому, что новым источником ценности станет не мощность компьютерного

оборудования, а получаемые им данные и способ их анализа. Данные становятся важным корпоративным активом, жизненно важным экономическим вкладом и основой новых бизнес-моделей. И хотя данные еще не вносятся в корпоративные балансовые отчеты, вероятно, это вопрос времени.

Несмотря на то что технологии обработки данных появились некоторое время назад, они были доступны только агентствам по шпионажу, исследовательским лабораториям и крупнейшим мировым компаниям. Walmart[3] и CapitalOne[4] первыми использовали большие данные в розничной торговле и банковском деле, тем самым изменив их. Теперь многие из этих инструментов стали широкодоступными.

Эти изменения в большей мере коснутся отдельных лиц, ведь в мире, где вероятность и корреляции имеют первостепенное значение, специальные знания менее важны. Узкие специалисты останутся востребованными, но им придется считаться с большими данными. Помните, как в фильме «Человек, который изменил всё»[5]: на смену бейсбольным скаутам пришли специалисты по статистике, а интуиция уступила место сложной аналитике. Нам придется пересмотреть традиционные представления об управлении, принятии решений, человеческих ресурсах и образовании.

Большинство наших учреждений создавались исходя из предположения, что информация, используемая при принятии решений, характеризуется небольшим объемом, точностью и причинностью. Но все меняется: если данных чрезвычайно много, они быстро обрабатываются и не допускают неточности. Более того, из-за огромного объема информации решения принимают не люди, а машины. Темную сторону больших данных мы рассмотрим в восьмой главе.

Общество накопило тысячелетний опыт понимания и регулирования поведения человека. Но что делать с алгоритмом? Еще на ранних этапах обработки данных влиятельные лица увидели угрозу конфиденциальности. С тех пор общество создало массивный свод правил для защиты конфиденциальной информации. Однако в эпоху больших данных это практически бесполезная «линия Мажино»[6]. Люди охотно делятся информацией в интернете, и эта возможность —

одна из главных функций веб-служб, а не слабое место, которое нужно устранить.

Опасность для отдельных лиц теперь представляет не угроза конфиденциальности, а вероятность: алгоритмы будут прогнозировать вероятность того, что человек получит сердечный приступ (и ему придется больше платить за медицинское страхование), не выполнит долговые обязательства по ипотечному кредиту (и ему будет отказано в займе) или совершит преступление (и, возможно, будет арестован заранее). Это заставляет взглянуть на неприкосновенность волеизъявления и диктатуру данных с этической точки зрения. Должна ли воля человека превалировать над большими данными, даже если статистика утверждает иное? Подобно тому как печатный станок дал толчок для принятия законов, гарантирующих свободу слова (раньше они не существовали, так как практически нечего было защищать), в эпоху больших данных потребуются новые правила для защиты неприкосновенности личности.

Обществу и организациям во многом придется изменить способы обработки данных и управления ими. Мы вступаем в мир постоянного прогнозирования на основе данных, в котором, возможно, не всегда сможем объяснить причины своих решений. Что значит, если врач не может обосновать необходимость медицинского вмешательства, при этом не требуя согласия пациента полагаться на «черный ящик» (а именно так и должен поступать врач, опирающийся на диагноз, который получен на основе больших данных)? Придется ли в судебной системе менять стандартное понятие «вероятная причина» на «вероятностная причина» — и если да, то каковы будут последствия для свободы человека и его чувства собственного достоинства?

В девятой главе мы предлагаем ряд принципов эпохи больших данных, которые основаны на ценностях, возникших и закрепившихся в более знакомом нам мире «малых данных». Старые правила необходимо обновить в соответствии с новыми обстоятельствами.

Польза для общества будет огромной, поскольку большие данные помогут решению насущных глобальных проблем, таких как борьба с изменением климата, искоренение болезней, а также содействие эффективному управлению и экономическому развитию. При этом эпоха больших данных заставляет нас лучше подготовиться к

изменениям организаций и нас самих, которые произойдут в результате освоения технологий.

Большие данные — важный шаг человечества в постоянном стремлении количественно измерить и постичь окружающий мир. То, что прежде невозможно было измерять, хранить, анализировать и распространять, находит свое выражение в виде данных. Использование огромных массивов данных вместо их малой доли и выбор количества в ущерб точности открывают путь к новым способам понимания мира. Это подталкивает общество отказаться от освященного веками поиска причинности и в большинстве случаев пользоваться преимуществами корреляций.

Поиск причин стал своего рода религией современности. Большие данные в корне меняют это мировоззрение, и мы снова оказываемся в таком историческом тупике, где «Бог умер». То, в чем мы были непоколебимо уверены, в очередной раз меняется. На этот раз, по иронии судьбы, — за счет более надежных доказательств. Какая роль при этом отводится интуиции, вере, неопределенности, действиям вразрез доказательствам, а также обучению опытным путем? По мере того как мир переходит от поиска причинности к поиску корреляции, что нам нужно делать, чтобы продвигаться вперед, не подрывая глубинных основ общества, гуманности и прогресса, опирающихся на доводы? Эта книга намерена объяснить, в какой точке мы находимся и как сюда попали и какие выгоды и опасности нас ждут впереди.

Примечания к главе 1

Глава 2

Больше данных

Большие данные позволяют увидеть и понять связи между фрагментами информации, которые до недавнего времени мы только пытались уловить. По мнению Джеффа Йонаса, эксперта компании IBM по большим данным, нужно позволить данным «говорить». Это может показаться несколько тривиальным, ведь с древних времен люди воспринимали данные в виде обычных ежедневных наблюдений, а последние несколько столетий — в виде формальных количественных единиц, которые можно обрабатывать с помощью сложнейших алгоритмов[16].

В цифровую эпоху стало проще и быстрее обрабатывать данные и мгновенно рассчитывать миллионы чисел. Но если речь идет о данных, которые «говорят», имеется в виду нечто большее. Большие данные диктуют три основных шага к новому образу мышления. Они взаимосвязаны и тем самым подпитывают друг друга. Первый — это способность анализировать все данные, а не довольствоваться их частью или статистическими выборками. Второй — готовность иметь дело с неупорядоченными данными в ущерб точности. Третий — изменение образа мыслей: доверять корреляциям, а не гнаться за труднодостижимой причинностью. В этой главе мы рассмотрим первый из них — шаг к тому, чтобы использовать все данные, а не полагаться на их небольшую часть.

Задача точного анализа больших объемов данных для нас не новая. В прошлом мы не утруждали себя сбором большого количества данных, поскольку инструменты для их записи, хранения и анализа были недостаточно эффективными. Нужная информация просеивалась до минимально возможного уровня, чтобы ее было проще анализировать. Получалось что-то вроде бессознательной самоцензуры: мы воспринимали трудности взаимодействия с данными как нечто само собой разумеющееся, вместо того чтобы увидеть, чем они являлись на самом деле — искусственным ограничением из-за

уровня технологий того времени. Теперь же технические условия повернулись на 179 градусов: количество данных, которые мы способны обработать, по-прежнему ограничено (и останется таким), но условные границы стали гораздо шире и будут расширяться.

В некотором смысле мы пока недооцениваем возможность оперировать большими объемами данных. Основная часть нашей деятельности и структура организаций исходят из предположения, что информация — дефицитный ресурс. Мы решили, что нам под силу собирать лишь малую долю информации, и, собственно, этим и занимались. На что рассчитывали, то и получили. Мы даже разработали сложные методы использования как можно меньшего количества данных. В конце концов, одна из целей статистики — подтверждать крупнейшие открытия с помощью минимального количества данных. По сути, мы закрепили практику работы с неполной информацией в своих нормах, процессах и структурах стимулирования. Чтобы узнать, что представляет собой переход на большие данные, для начала заглянем в прошлое.

Не так давно привилегию собирать и сортировать огромные массивы информации получили частные компании, а теперь — и отдельные лица. В прошлом эта задача лежала на организациях с более широкими возможностями, таких как церковь или государство, которые во многих странах имели одинаковое влияние. Древнейшая запись о подсчетах относится к примерно 8000 году до н. э., когда шумерские купцы записывали реализуемые товары с помощью маленьких шариков глины. Однако масштабные подсчеты были в компетенции государства. Тысячелетиями правительства старались вести учет населения, собирая информацию.

Обратимся к переписям. Считается, что египтяне начали проводить их примерно в 3000 году до н. э. (как и китайцы). Сведения об этом можно найти в Ветхом и, конечно, Новом Завете. В нем упоминается о переписи, которую ввел кесарь Август, — «повелении сделать перепись по всей земле» (Евангелие от Луки 2:01). Это повеление и привело Иосифа с Марией в Вифлеем, где родился Иисус. В свое время Книга Судного дня (1086 год) — одно из самых почитаемых сокровищ Британии — была беспрецедентным, всеобъемлющим источником экономических и демографических сведений об

английском народе. В сельские поселения были направлены королевские представители, которые составили полный перечень всех и вся — книгу, позже получившую библейское название «Судный день», поскольку сам процесс напоминал Страшный суд, открывающий всю подноготную человека.

Проведение переписей — процесс дорогостоящий и трудоемкий. Король Вильгельм I не дожил до завершения книги Судного дня, составленной по его распоряжению. Между тем существовал лишь один способ избавиться от трудностей, сопряженных со сбором информации, —отказаться от него. В любом случае информация получалась не более чем приблизительной. Переписчики прекрасно понимали, что им не удастся все идеально подсчитать. Само название переписей — «ценз»[7] (англ. *census*) — происходит от латинского термина *censere*, что означает «оценивать».

Более трехсот лет назад у британского галантерейщика по имени Джон Граунт появилась инновационная идея. Чтобы вывести общую численность населения Лондона во время бубонной чумы, он не стал подсчитывать отдельных лиц, а воспользовался другим способом. Сегодня мы бы назвали его статистикой. Новый подход давал весьма приблизительные результаты, зато показывал, что на основании небольшой выборки можно экстраполировать полезные знания об общей картине. Особое значение имеет то, как именно это делалось. Граунт просто масштабировал результаты своей выборки.

Его система стала известной, хотя позже и выяснилось, что расчеты могли быть объективными только по счастливой случайности. Из поколения в поколение метод выборки оставался далеко не безупречным. Итак, для переписи и подобных целей, связанных с большими данными, основной подход заключался в грубой попытке подсчитать все и вся.

Поскольку переписи были сложными, дорогостоящими и трудоемкими, они проводились лишь в редких случаях. Древние римляне делали это каждые пять лет, притом что население исчислялось десятками тысяч. А в Конституции США закреплено правило проводить переписи каждые десять лет, поскольку население растущей страны насчитывает миллионы. Но к концу XIX века даже

это оказалось проблематичным. Возможности Бюро переписи населения не успевали за ростом данных.

Перепись 1880 года длилась целых восемь лет. Ее данные успели устареть еще до публикации результатов. По подсчетам, на подведение итогов переписи 1890 года требовалось 13 лет — смехотворный срок, не говоря уже о нарушении Конституции. В то же время распределение налогов и представительство в Конгрессе зависели от численности населения, поэтому крайне важно было своевременно получать точные данные.

Проблема, с которой столкнулось Бюро переписи населения США, напоминает трудности современных ученых и бизнесменов: поток данных стал непосильным. Объем собираемой информации превысил все возможности инструментов, используемых для ее обработки. Срочно требовались новые методы. В 1880-х годах ситуация оказалась настолько удручающей, что Бюро переписи населения США заключило контракт с Германом Холлеритом, американским изобретателем, на использование его идеи с перфокартами и счетными машинами для переписи 1890 года[17].

С большим трудом ему удалось сократить время на сведение результатов с восьми лет до менее одного года. Это было удивительное достижение, которое положило начало автоматизированной обработке данных (и заложило основу будущей компании IBM). Однако такой метод получения и анализа больших объемов данных обходился все еще слишком дорого. Каждый житель Соединенных Штатов заполнял форму, из которой создавалась перфокарта для подсчета итогов. Трудно представить, как в таких условиях удалось бы провести перепись быстрее чем за десять лет. Но отставание определенно играло против нации, растущей не по дням, а по часам.

Основная трудность состояла в выборе: использовать все данные или только их часть. Безусловно, разумнее всего получать полный набор данных всех проводимых измерений. Но это не всегда выполнимо при огромных масштабах. И как выбрать образец? По мнению некоторых, лучший выход из ситуации — создавать целенаправленные выборки, которые представляли бы полную картину. Однако в 1934 году польский статистик Ежи Нейман ярко

продемонстрировал, как такие выборки приводят к огромным ошибкам. Оказалось, разгадка в том, чтобы создавать выборку по принципу случайности[18].

Работа статистиков показала, что на повышение точности выборки больше всего влияет не увеличение ее размера, а элемент случайности. На самом деле, как ни странно, случайная выборка из 1100 ответов отдельных лиц на бинарный вопрос («да» или «нет») имеет более чем 97%-ную точность при проецировании на все население. Это работает в 19 из 20 случаев, независимо от общего размера выборки, будь то 100 000 или 100 000 000[19]. И трудно объяснить математически. Если вкратце, то с определенного момента роста данных предельное количество новой информации, получаемой из новых наблюдений, становится все меньше.

То, что случайность компенсирует размер выборки, стало настоящим открытием, проложившим путь новому подходу к сбору информации. Данные можно собирать с помощью случайных выборок по низкой себестоимости, а затем экстраполировать их с высокой точностью на явление в целом. В результате правительства могли бы вести небольшие переписи с помощью случайных выборок ежегодно, а не раз в десятилетие (что они и делали). Бюро переписи населения США, например, ежегодно проводит более двухсот экономических и демографических исследований на выборочной основе, не считая переписи раз в десять лет для подсчета всего населения. Выборки решали проблему информационной перегрузки в более раннюю эпоху, когда собирать и анализировать данные было очень трудно.

Новый метод быстро нашел применение за пределами государственного сектора и переписей. В бизнесе случайные выборки использовались для обеспечения качества производства, упрощая процессы контроля и модернизации и к тому же снижая расходы на них. Поначалу для всестороннего контроля качества требовалось осматривать каждый продукт, выходящий с конвейера. Сейчас достаточно случайной выборки тестовых экземпляров из партии продукции. По сути, случайные выборки уменьшают проблемы с большими данными до более управляемых. Кроме того, они положили начало опросам потребителей в сфере розничной торговли, фокус-

группам в политике, а также преобразовали большинство гуманитарных наук в социальные.

Случайные выборки пользовались успехом. Они же сформировали основу для современных масштабных измерений. Но это лишь упрощенный вариант — еще одна альтернатива сбора и анализа полного набора данных, к тому же полная недостатков. Мало того что ее точность зависит от случайности при сборе данных выборки — достичь этой случайности не так-то просто. Если сбор данных осуществляется с погрешностью, результаты экстраполяции будут неправильными.

Так, например, одна из ранних ошибок, связанных с выборкой, произошла в 1936 году, когда еженедельный журнал Literary Digest провел опрос двух миллионов избирателей и ошибочно спрогнозировал блестящую победу Республиканской партии на президентских выборах США. (Как оказалось, действующий президент Франклин Рузвельт, представитель Демократической партии, победил Альфреда Лэндона с перевесом в 523 голоса к 8 в коллегии выборщиков.) И дело было не в том, что выборка оказалась слишком маленькой, — не хватало элемента случайности. Выбирая участников опроса, специалисты Literary Digest использовали список подписчиков и телефонные каталоги, не понимая, что обе группы — и подписчики, и телефонные абоненты — относятся к более состоятельной категории населения и гораздо вероятнее проголосуют за республиканцев[20]. С этой задачей можно было бы справиться гораздо лучше и дешевле, используя часть выборки, но сформированную именно случайным образом.

Не так давно нечто подобное произошло в процессе опросов, связанных с выборами. Опросы проводились с помощью стационарных телефонов. Выборка оказалась недостаточно случайной из-за погрешности, вызванной тем, что люди, которые пользуются исключительно мобильными телефонами (более молодая и либеральная категория населения), не брались в расчет. Это привело к неправильным прогнозам результатов выборов. В 2008 году в период президентских выборов между Бараком Обамой и Джоном Маккейном главные организации по проведению анкетного опроса населения —

Gallup, Pew и ABC/Washington Post — обнаружили разницу в один-три пункта между опросами с учетом пользователей мобильных телефонов и без них. С учетом напряженности гонки это была огромная разница[21].

* * *

Большинство неудобств связаны с тем, что случайную выборку трудно масштабировать, поскольку разбивка результатов на подкатегории существенно увеличивает частоту ошибок. И это понятно. Предположим, у вас есть случайная выборка из тысячи людей и их намерений проголосовать на следующих выборах. Если выборка достаточно случайна, вполне вероятно, что настроения людей в рамках выборки будут разниться в пределах 3%. Но что если плюс-минус 3% — недостаточно точный результат? Или нужно разбить группу на более мелкие подгруппы по половому признаку, географическому расположению или доходу? Или если нужно объединить эти подгруппы в целевую группу населения?

Допустим, в общей выборке из тысячи избирателей подгруппа «обеспеченных женщин из северо-восточного региона» составила гораздо меньше сотни. Используя лишь несколько десятков наблюдений, невозможно точно прогнозировать, какого кандидата предпочтут *все* обеспеченные женщины в северо-восточном регионе, даже если случайность близка к идеальной. А небольшие погрешности в случайности выборки сделают ошибки еще более выраженными на уровне подгруппы.

Таким образом, при более внимательном рассмотрении интересующих нас подкатегорий данных выборка быстро становится бесполезной. То, что работает на макроуровне, не подходит для микроуровня. Выборка подобна аналоговой фотопечати: хорошо смотрится на расстоянии, но при ближайшем рассмотрении теряется четкость деталей.

Далее, выборка требует тщательного планирования и реализации. Данные выборки не смогут дать ответы на новые вопросы, если они не продуманы заранее. Поэтому выборка хороша в качестве упрощенного варианта, не более. В отличие от целого набора данных, выборка обладает недостаточной расширяемостью и эластичностью, благодаря

которым одни и те же данные можно повторно анализировать совершенно по-новому — не так, как планировалось изначально при сборе данных.

Рассмотрим анализ ДНК. Формируется новая отрасль индивидуального генетического секвенирования, что обусловлено грандиозным падением стоимости технологии и многообещающими медицинскими возможностями. В 2012 году цена декодирования генома упала ниже 1000 долларов США — неофициальной отраслевой отметки, при которой технология приобретает массовый характер. Так, начиная с 2007 года стартап Кремниевой долины 23andme[8] стал предлагать анализ ДНК всего за пару сотен долларов. Этот анализ позволяет выявить особенности генетического кода человека, которые повышают его предрасположенность к развитию определенных заболеваний, например рака молочной железы или проблем с сердцем. А объединяя информацию о ДНК и здоровье своих клиентов, 23andme рассчитывает выявить новые закономерности, которые невозможно обнаружить другим способом.

Компания секвенирует крошечную часть ДНК человека из нескольких десятков участков, которые являются «маркерами». Они указывают на определенную генетическую слабость и представляют собой лишь выборку всего генетического кода человека. При этом миллиарды пар оснований ДНК остаются несеквенированными. В результате 23andme может ответить только на те вопросы, которые связаны с заданными маркерами. При обнаружении нового маркера потребуется еще раз секвенировать ДНК человека (точнее, его соответствующую часть). Работа с выборкой, а не целым набором данных имеет свои недостатки: позволяя проще и быстрее находить нужные данные, она не в состоянии ответить на вопросы, которые не были поставлены заранее.

Легендарный руководитель компании Apple Стив Джобс выбрал другой подход к борьбе против рака, став одним из первых людей в мире, просеквенировавших всю свою ДНК, а также ДНК своей опухоли. Это обошлось ему в шестизначную сумму, которая в сотни раз превышала обычный тариф 23andme. Зато Стив Джобс получил не

просто выборку или набор маркеров, а целый набор данных, содержащий весь генетический код.

При лечении среднестатистического онкобольного врачам приходится рассчитывать, что ДНК пациента достаточно похожа на пробу, взятую для исследования. А у команды врачей Стива Джобса была возможность подбирать препараты, ориентируясь на их эффективность для конкретного генетического материала. Всякий раз, когда один препарат становился неэффективным из-за того, что рак мутировал и стал устойчивым к его воздействию, врачи могли перейти на другой препарат, «перескакивая с одной кувшинки на другую», как говорил Стив Джобс. В то время он язвительно заметил: «Я стану одним из первых, кто сумеет обойти рак, или одним из последних, кто умрет от него». И хотя его предсказание, к сожалению, не сбылось, сам метод получения всего набора данных (а не просто выборки) продлил жизнь Стива Джобса на несколько лет[22].

От малого к большему

Выборка — продукт эпохи ограниченной обработки информации. Тогда мир познавался через измерения, но инструментов для анализа собранных показателей не хватало. Теперь выборка стала пережитком того времени. Недостатки в подсчетах и сведении данных стали гораздо менее выраженными. Датчики, GPS-системы мобильных телефонов, действия на веб-страницах и Twitter пассивно собирают данные, а компьютеры могут с легкостью обрабатывать их.

Понятие выборки подразумевает возможность извлечь максимум пользы из минимума материалов, подтвердить крупнейшие открытия с помощью наименьшего количества данных. Теперь же, когда мы можем поставить себе на службу большие объемы данных, выборки утратили прежнюю значимость. Технические условия обработки данных резко изменились, но адаптация наших методов и мышления не поспевает за ней.

Давно известно, что цена выборки — утрата подробностей. И как бы мы ни старались не обращать внимания на этот факт, он становится все более очевидным. Есть случаи, когда выборки являются единственным решением. Однако во многих областях происходит

переход от сбора небольшого количества данных до как можно большего, а если возможно, то и всего: *«N = всё»*.

Используя подход «*N* = всё», мы можем глубоко изучить данные. Не то что с помощью выборки! Кроме того, уже упоминалось, что мы могли бы достичь 97%-ной точности, экстраполируя результаты на все население. В некоторых случаях погрешность в 3% вполне допустима, однако при этом теряются нюансы, точность и возможность ближе рассмотреть некоторые подгруппы. Нормальное распределение, пожалуй, нормально. Но нередко действительно интересные явления обнаруживаются в нюансах, которые невозможно в полной мере уловить с помощью выборки.

Вот почему служба Google Flu Trends полагается не на случайную выборку, а на исчерпывающий набор из миллиардов поисковых интернет-запросов в США. Используя все данные, а не выборку, можно повысить точность анализа настолько, чтобы прогнозировать распространенность какого-либо явления не то что в государстве или всей нации, а в конкретном городе[23]. Исходная система Farecast использовала выборку из 12 000 точек данных и хорошо справлялась со своими задачами. Но, добавив дополнительные данные, Орен Эциони улучшил качество прогнозирования. В итоге система Farecast стала учитывать все ценовые предложения на авиабилеты по каждому маршруту в течение всего года. «Это временные данные. Просто продолжайте собирать их — и со временем вы станете все лучше и лучше понимать их закономерности», — делится Эциони[24].

Таким образом, в большинстве случаев мы с удовольствием откажемся от упрощенного варианта (выборки) в пользу полного набора данных. При этом понадобятся достаточные мощности для обработки и хранения данных, передовые инструменты для их анализа, а также простой и доступный способ сбора данных. В прошлом каждый из этих элементов был головоломкой. Мы по-прежнему живем в мире ограниченных ресурсов, в котором все части головоломки имеют свою цену, но теперь их стоимость и сложность резко сократились. То, что раньше являлось компетенцией только крупнейших компаний, теперь доступно большинству.

Используя все данные, можно обнаружить закономерности, которые в противном случае затерялись бы на просторах информации. Так, мошенничество с кредитными картами можно обнаружить путем поиска нетипичного поведения. Единственный способ его определить — обработать все данные, а не выборку. В таком контексте наибольший интерес представляют резко отклоняющиеся значения, а их можно определить, только сравнив с массой обычных транзакций. В этом заключается проблема больших данных. А поскольку транзакции происходят мгновенно, анализировать нужно тоже в режиме реального времени.

Компания Xoom специализируется на международных денежных переводах и опирается на хорошо известные большие данные. Она анализирует все данные, связанные с транзакциями, которые находятся в обработке. Система подняла тревогу, заметив незначительное превышение среднего количества транзакций с использованием кредитных карт Discover Card в Нью-Джерси. «Система обнаружила закономерность там, где ее не должно быть», — пояснил Джон Кунце, президент компании Xoom[25]. Сами по себе транзакции выглядели вполне законно. Но оказалось, что они инициированы преступной группировкой, которая пыталась обмануть компанию. Обнаружить отклонения в поведении можно было, только изучив все данные, чего не сделаешь с помощью выборки.

Использование всех данных не должно восприниматься как сверхзадача. Большие данные не обязательно таковы в абсолютном выражении (хотя нередко так и есть). Служба Flu Trends базируется на сотнях миллионов математических модельных экспериментов, использующих миллиарды точек данных. Полная последовательность человеческого генома содержит около трех миллиардов пар оснований. Однако само по себе абсолютное число точек данных (размер набора данных) не делает их примером больших данных как таковых. Отличительной чертой больших данных является то, что вместо упрощенного варианта случайной выборки используется весь имеющийся набор данных, как в случае службы Flu Trends и врачей Стива Джобса.

Насколько значимо применение подхода «*N* = всё», отлично иллюстрирует следующая ситуация. В японском национальном спорте — борьбе сумо — выявилась практика договорных боев. Обвинения в проведении «боев в поддавки» всегда сопровождали соревнования в этом императорском виде спорта и строго запрещались. Стивен Левитт, предприимчивый экономист из Университета Чикаго, загорелся идеей научиться определять такие бои. Как? Просмотрев все прошлые бои без исключения. В своей замечательной исследовательской статье, опубликованной в American Economic Review[26], он описывает пользу изучения всех данных. Позже эта идея найдет свое отражение в его бестселлере «Фрикономика»[9].

В поиске отклонений Левитт и его коллега Марк Дагген просмотрели все бои за последние 11 лет — более 64 000 поединков. И попали в десятку. Договорные бои действительно имели место, но не там, где их искало большинство людей. Речь шла не о чемпионских поединках, которые могли фальсифицироваться. Данные показали, что самое занятное происходило во время заключительных боев турнира, которые оставались незамеченными. Казалось, что на карту поставлено немного, ведь у борцов фактически нет шансов на завоевание титула.

Одна из особенностей сумо в том, что борцам нужно победить в большинстве из 15 боев турнира, чтобы сохранить свое положение и доходы. Иногда это приводит к асимметрии интересов, например, если борец со счетом 7:7 сталкивается с противником со счетом 8:6. Результат имеет огромное значение для первого борца и практически безразличен второму. Левитт и Дагган обнаружили, что в таких случаях, скорее всего, победит борец, который нуждается в победе. На первый взгляд, это «подарок» одного борца другому. Но в тесном мире сумо все взаимосвязано.

Может, парень просто боролся решительнее, поскольку цена победы была столь высока? Возможно. Но данные говорят об обратном: борцы, которые нуждаются в победе, побеждают примерно на 25% чаще, чем следовало ожидать. Вряд ли дело лишь в одном адреналине. Дальнейший разбор данных также показал, что при следующей встрече тех же двух борцов тот, кто проиграл в предыдущем бою, в

три-четыре раза вероятнее выиграет, чем при третьем или четвертом спарринге.

Эта информация всегда была очевидной, была на виду. Но анализ случайной выборки может не выявить такие закономерности. Анализ больших данных, напротив, показывает ее с помощью гораздо большего набора данных, стремясь исследовать всю совокупность боев. Это похоже на рыбалку, в которой нельзя сказать заранее, удастся ли что-то поймать и *что именно*.

Набор данных не всегда измеряется терабайтами. В случае сумо весь набор данных содержал меньше бит, чем обычная цифровая фотография. Но так как анализировались большие данные, в расчет бралось больше данных, чем при случайной выборке. В этом и общем смысле «большой» — скорее относительное понятие, чем абсолютное (в сравнении с полным набором данных).

В течение долгого времени случайная выборка считалась хорошим решением. Она позволяла анализировать проблемы больших данных в предцифровую эпоху. Однако при выборке часть данных теряется, как и в случае преобразования цифрового изображения или песни в файл меньшего размера. Наличие полного (или почти полного) набора данных дает гораздо больше свободы для исследования и разностороннего рассмотрения данных, а также более подробного изучения их отдельных особенностей.

Подходящий пример — камера Lytro. Она стала революционным открытием, так как применяет большие данные к основам технологии фотографии. Эта камера захватывает не только одну световую плоскость, как обычные камеры, но и около 11 миллионов лучей всего светового поля. Точное изображение, получаемое из цифрового файла, можно в дальнейшем изменять в зависимости от того, на какой объект кадра нужно настроить фокус. Благодаря сбору всех данных не обязательно настраивать фокус изображения изначально, ведь он настраивается на любой объект изображения после того, как снимок уже сделан. Снимок содержит лучи всего светового поля, а значит, и все данные, то есть «*N* = всё». В результате информация лучше подходит для «повторного использования», чем обычные изображения, когда фотографу нужно выбрать объект фокусировки, прежде чем нажать на кнопку затвора.

Поскольку большие данные опираются на всю или максимально возможную информацию, точно так же мы можем рассматривать подробности и проводить новый анализ, не рискуя четкостью. Мы проверим новые гипотезы на любом уровне детализации. Это позволяет обнаруживать случаи договорных боев в борьбе сумо, распространение вируса гриппа по регионам, а также лечить раковые заболевания, воздействуя целенаправленно на поврежденную часть ДНК. Таким образом, мы можем работать на небывало глубоком уровне понимания.

Следует отметить, что не всегда необходимы все данные вместо выборки. Мы все еще живем в мире ограниченных ресурсов. Однако все чаще целесообразно использовать все имеющиеся данные. И если ранее это было невозможно, то теперь — наоборот.

Подход «*N* = всё» оказал значительное влияние на общественные науки. Они утратили свою монополию на осмысление эмпирических данных, а анализ больших данных заменил ранее востребованных высококвалифицированных специалистов по выборкам. Общественные дисциплины во многом полагаются на выборки, исследования и анкеты. Но если данные собираются пассивно, в то время как люди заняты обычными делами, погрешности, связанные с исследованиями и анкетами, сходят на нет. Теперь мы можем собирать информацию, недоступную ранее, будь то чувства, высказанные по мобильному телефону, или настроения, переданные в твитах. Более того, исчезает сама необходимость в выборках[27].

Альберт-Лазло Барабаши, один из ведущих мировых авторитетов в области сетей, и его коллеги исследовали взаимодействия между людьми в масштабе всего населения. Для этого они проанализировали все журналы анонимного мобильного трафика за четыре месяца, полученные от оператора беспроводной связи, который обслуживал около пятой части всего населения страны. Это был первый анализ сетей на общественном уровне, в котором использовался набор данных в рамках подхода «*N* = всё». Благодаря масштабу, который позволил учесть звонки миллионов людей в течение длительного времени, появились новые идеи, которые, скорее всего, не удалось бы выявить другим способом[28].

Команда обнаружила интересную закономерность, не свойственную небольшим исследованиям: если удалить из сети людей, имеющих множество связей в сообществе, оставшаяся социальная сеть станет менее активной, но останется на плаву. С другой стороны, если из сети удалить людей, имеющих связи за пределами их непосредственного окружения, оставшаяся социальная сеть внезапно распадется, словно повредили саму ее структуру. Это стало важным, но совершенно неожиданным открытием. Кто бы мог подумать, что люди с большим количеством близких друзей настолько менее важны в структуре сети, чем те, у кого есть более отдаленные связи? Выходит, что разнообразие высоко ценится как в группе, так и в обществе в целом. Открытие заставило по-новому взглянуть на то, как следует оценивать важность людей в социальных сетях.

Мы склонны думать, что статистическая выборка — это своего рода непреложный принцип (такой, как геометрические правила или законы гравитации), на котором основана цивилизация. Однако эта концепция появилась менее ста лет назад и служила для решения конкретной задачи в определенный момент времени при определенных технологических ограничениях. С тех пор эти ограничения весьма изменились. Стремиться к случайной выборке в эпоху больших данных — все равно что хвататься за хлыст в эпоху автомобилей. Мы можем использовать выборки в определенных обстоятельствах, но они не должны быть (и не будут) доминирующим способом анализа больших наборов данных. Все чаще мы можем позволить себе замахнуться на данные в полном объеме.

Примечания к главе 2

Глава 3

Беспорядочность

Число областей, в которых можно использовать все имеющиеся данные, неуклонно растет, однако увеличение количества приводит к неточности. В наборы данных всегда закрадывались ошибочные цифры и поврежденные биты. Эту проблему следует попытаться решить хотя бы потому, что это возможно. Чего нам никогда не хотелось, так это мириться с такими ошибками, считая их неизбежными. В этом и состоит один из основных переходов от малых данных к большим.

В мире «малых данных» сокращение количества ошибок и обеспечение высокого качества данных становились естественным и необходимым толчком к поиску новых решений. Поскольку собиралась лишь малая часть информации, мы заботились о том, чтобы она была как можно более точной. Поколения ученых оптимизировали свои инструменты, добиваясь все большей точности данных, будь то положение небесных тел или размер объектов под микроскопом. В мире, где правили выборки, стремление к точности принимало характер одержимости, сбор лишь ограниченного числа точек данных неминуемо вел к распространению ошибок, тем самым снижая точность общих результатов.

На протяжении большей части истории наивысшие достижения человека были связаны с завоеванием мира путем его измерения. Одержимость точностью началась в середине XIII века в Европе, когда астрономы и ученые взяли на вооружение как никогда точную количественную оценку времени и пространства — «меру реальности», выражаясь словами историка Альфреда Кросби.

Негласно считалось, что, если измерить явление, его удастся понять. Позже измерения оказались привязанными к научному методу наблюдения и объяснения — способности количественно измерять воспроизводимые результаты, а затем записывать и представлять их. «Измерить — значит узнать», — говорил лорд Кельвин. И это стало

основным постулатом. «Знание — сила», — поучал Фрэнсис Бэкон. В то же время математики и те, кто позже стал актуарием или бухгалтером, разработали методы, которые сделали возможным точный сбор и регистрацию данных, а также управление ими[29].

К XIX веку во Франции (в то время ведущей стране в мире по уровню развития науки) была разработана система строго определенных единиц измерения для сбора данных о пространстве, времени и не только. Другие страны перенимали эти стандарты. Дошло до того, что признанный во всем мире эталон единиц измерения стал закрепляться в международных договорах. Это явилось вершиной эпохи измерений. Лишь полвека спустя, в 1920-х годах, открытия в области квантовой механики навсегда разрушили веру в возможность достичь совершенства в измерениях. Тем не менее, не считая относительно небольшого круга физиков, инженеры и ученые не спешили расставаться с мыслью о совершенстве измерений. В деловой сфере эта идея даже получила более широкое распространение, по мере того как рациональные науки — математика и статистика — начали оказывать влияние на все области коммерческой деятельности.

Между тем множатся ситуации, в которых неточность воспринимается скорее как особенность, а не как недостаток. Взамен снижения стандартов допустимых погрешностей вы получаете намного больше данных, с помощью которых можно совершать новые открытия. При этом действует принцип не просто «больше данных — какой-то результат», а, по сути, «больше данных — лучше результат».

Нам предстоит иметь дело с несколькими видами беспорядочности. Это может быть связано с тем, что при добавлении новых точек данных вероятность ошибок возрастает. Следовательно, если, например, увеличить показатели нагрузки на мост в тысячу раз, возрастет вероятность того, что некоторые показатели будут ошибочными. Вы увеличите беспорядочность, сочетая различные типы информации из разных источников, которые не всегда идеально выравниваются. Или, определив причину жалоб, направленных в центр обработки заказов с помощью программного обеспечения для распознавания речи, и сравнив эти данные со временем, затраченным

со стороны оператора на их обработку, можно получить несовершенную, но полезную общую картину ситуации. Кроме того, беспорядочность иногда связана с неоднородностью форматирования. В таком случае, прежде чем обрабатывать данные, их следует «очистить». «Существуют буквально тысячи способов упомянуть компанию IBM, — отмечает знаток больших данных Дж. Патил, — от IBM до International Business Machines и Исследовательского центра Т. Дж. Уотсона»[30]. Беспорядочность может возникнуть при извлечении или обработке данных, поскольку путем преобразования мы превращаем их в нечто другое. Так, например, происходит, когда мы анализируем настроения в сообщениях Twitter, чтобы прогнозировать кассовые сборы голливудских фильмов. А беспорядочность сама по себе... беспорядочна.

Представьте себе, что вам нужно измерить температуру в винограднике. Если у вас только один датчик температуры на весь участок земли, необходимо убедиться, что он работает точно и непрерывно. Если же для каждой из сотен лоз установлен отдельный датчик, вероятно, рано или поздно какой-то из них станет предоставлять неправильные данные. Полученные данные могут быть менее точными (или более «беспорядочными»), чем от одного точного датчика. Любой из отдельно взятых показателей может быть ошибочным, но в совокупности множество показателей дадут более точную картину. Поскольку набор данных состоит из большего числа точек данных, его ценность гораздо выше, и это с лихвой компенсирует его беспорядочность.

Теперь рассмотрим случай повышения частоты показателей. Если мы возьмем одно измерение в минуту, то можем быть уверены, что данные будут поступать в идеально хронологическом порядке. Измените частоту до десяти или ста показателей в секунду — и точность последовательности станет менее определенной. Так как информация передается по сети, запись может задержаться и прибыть не по порядку либо попросту затеряться. Информация получится немного менее точной, но ввиду большого объема данных отказаться от строгой точности вполне целесообразно.

В первом примере мы пожертвовали точностью отдельных точек данных в пользу широты, получив взамен детали, которые не удалось бы обнаружить другим путем. Во втором случае отказались от точности в пользу частоты, зато увидели изменения, которые иначе упустили бы из виду. Такие ошибки можно устранить, если направить на них достаточно ресурсов. В конце концов, на Нью-Йоркской фондовой бирже производится 30 000 сделок в секунду, и правильная последовательность здесь чрезвычайно важна. Но во многих случаях выгоднее допустить ошибку, чем работать над ее предотвращением.

Мы можем согласиться с беспорядочностью в обмен на масштабирование. Один из представителей консалтинговой компании Forrester однажды выразился так: «Иногда два плюс два может равняться 3,9. И это достаточно хорошо»[31]. Конечно, эти данные не могут быть абсолютно неправильными, и мы готовы в некоторой степени пожертвовать точностью в обмен на понимание общих тенденций. Большие данные преобразуют цифры в нечто более вероятностное, чем точность. В этом процессе обществу придется ко многому привыкнуть, столкнувшись с рядом проблем, которые мы рассмотрим в этой книге. Но на сегодняшний день стоит просто отметить, что при увеличении масштаба беспорядочность неизбежна, и с этим нужно смириться.

Подобный переход можно заметить в том, в какой степени увеличение объема данных важнее других усовершенствований в вычислительных технологиях. Всем известно, насколько вычислительная мощность выросла за эти годы в соответствии с законом Мура, который гласит, что число транзисторов на кристалле удваивается примерно каждые два года. В результате компьютеры стали быстрее, а память — объемнее. Производительность алгоритмов, которые управляют многими нашими системами, также увеличилась, но осталась несколько в тени. По некоторым данным, вычислительные алгоритмы улучшились примерно в 43 000 раз в период между 1988 и 2003 годами — значительно больше, чем процессоры в соответствии с законом Мура[32]. Однако многие достижения, наблюдаемые в обществе благодаря большим данным, состоялись не столько за счет более

быстрых чипов или улучшенных алгоритмов, сколько за счет увеличения количества данных.

Так, шахматные алгоритмы изменились лишь немного за последние несколько десятилетий, так как правила игры в шахматы полностью известны и жестко ограничены. Современные компьютерные программы по игре в шахматы играют гораздо лучше, чем их предшественники, потому что лучше просчитывают свой эндшпиль[10]. И это им удается просто потому, что в систему поступает больше данных. Варианты эндшпиля при оставшихся шести (и менее) фигурах на шахматной доске полностью проанализированы, а все возможные ходы («*N* = всё») представлены в виде массивной таблицы, которая в несжатом виде заполнила бы более терабайта данных. Благодаря этому компьютеры могут безупречно вести все важные эндшпили. Ни один человек не сможет переиграть систему[33].

То, насколько можно усовершенствовать алгоритмы, увеличив количество данных, убедительно продемонстрировано в области обработки естественного языка — способа, с помощью которого компьютеры распознают слова, используемые нами в повседневной речи. Примерно в 2000 году Мишель Банко и Эрик Брилл из исследовательского центра Microsoft Research поставили задачу улучшить средство проверки грамматики — элемент программы Microsoft Word. Перед ними было несколько путей: улучшение существующих алгоритмов, поиск новых методов или добавление более сложных функций. Прежде чем выбрать один из них, они решили посмотреть, что будет, если существующие методы применить к гораздо большему количеству данных. Большинство исследований по машинному обучению алгоритмов полагались на корпусы[11], состоящие из миллиона слов, а то и меньше. Поэтому Банко и Брилл выбрали четыре алгоритма общего назначения и ввели в них на три порядка больше данных: 10 миллионов слов, затем 100 миллионов и, наконец, миллиард.

Результаты поразили. Чем больше данных подавалось на входе, тем лучше были результаты работы всех четырех типов алгоритмов. Простой алгоритм, который хуже всех справлялся с половиной миллиона слов, показал наилучший результат, обработав миллиард

слов. Степень точности возросла с 75 до более чем 95%. И наоборот, алгоритм, который лучше всех справлялся с небольшим объемом данных, показал наихудший результат при больших объемах. Следует отметить, что при этом его результат, как и результат остальных алгоритмов, значительно улучшился: с 86 до 94% точности. «Эти результаты показывают, что нам, возможно, понадобится пересмотреть свое представление о том, на что стоит тратить время и средства: на разработку алгоритмов или на развитие корпусов», — отметили Банко и Брилл в одной из своих научных статей на эту тему[34].

Итак, чем больше данных, тем меньше затрат. А как насчет беспорядочности? Спустя несколько лет после того, как Банко и Брилл начали активно собирать данные, исследователи компании Google, их конкурента, стали рассуждать в том же направлении, но еще более масштабно. Они взялись тестировать алгоритмы, используя не миллиард слов, а корпус из целого триллиона слов. Целью Google была не разработка средства проверки грамматики, а еще более сложная задача — перевод.

Концепция так называемого «машинного» перевода появилась на заре вычислительной техники, в 1940 году, когда устройства состояли из вакуумных ламп и занимали целую комнату. Идея стала особенно актуальной во времена холодной войны, когда в руки США попало огромное количество письменных и устных материалов на русском языке, но не хватало человеческих ресурсов для их быстрого перевода.

Специалисты в области компьютерных наук начали с того, что выбрали сочетание грамматических правил и двуязычный словарь. В 1954 году компания IBM перевела 60 русских фраз на английский язык на основе словарного запаса компьютера, состоящего из 250 пар слов, и шести правил грамматики. Результаты оказались многообещающими. В компьютер IBM 701 с помощью перфокарт ввели текст «Мы передаем мысли посредством речи» и получили на выходе We transmit thoughts by means of speech. В пресс-релизе по случаю такого события отмечалось, что было «благополучно переведено» 60 предложений. Директор программы профессор Леон Достерт из Джорджтауна заявил, что машинный перевод станет

«свершившимся фактом» предположительно через «лет пять, а то и три [года]»[35].

Первоначальный успех был обманчив. К 1966 году комитет по вопросам машинного перевода признал, что потерпел неудачу. Проблема оказалась сложнее, чем они предполагали. Суть перевода заключалась в обучении компьютеров не только правилам, но и исключениям. Этому трудно обучить компьютер в прямой форме. В конце концов, перевод состоит не только в запоминании и воспроизведении, как могло показаться раньше. Речь идет о поиске подходящих слов среди множества альтернативных вариантов. Что значит bonjour? «Доброе утро», «добрый день», «здравствуйте» или, может быть, «привет»? Все зависит от обстоятельств.

В конце 1980-х годов у исследователей из компании IBM родилась новая идея. Вместо того чтобы загружать словари и явные лингвистические правила в компьютер, они позволили ему автоматически вычислять статистическую вероятность того, что то или иное слово либо словосочетание на одном языке лучше всего соответствует аналогу на другом. В 1990-х годах в проекте компании IBM Candide был задействован десятилетний опыт переводов стенограмм заседаний канадского парламента, опубликованных на французском и английском языках, — около трех миллионов предложений[36]. Поскольку это официальные документы, их переводы были выполнены с соблюдением чрезвычайно высоких требований. По меркам того времени количество данных было огромным. Эта технология, получившая известность как «статистический машинный перевод», ловко превратила задачу перевода в одну большую математическую задачу. И это сработало. Компьютерный перевод неожиданно стал намного лучше. Однако вслед за начальным прорывом компании IBM не удалось внести каких-либо значительных улучшений, несмотря на большие вложения. В конечном счете проект был закрыт.

Менее чем через десять лет, в 2006-м, компания Google подалась в область перевода в рамках своей миссии «упорядочить мировую информацию и сделать ее полезной и всесторонне доступной». Вместо того чтобы использовать аккуратно переведенные на два языка

страницы текста, Google задействовала более массивный, но при этом гораздо более беспорядочный набор данных — глобальную сеть интернет. Разработанная система поглощала все переводы, которые ей только удавалось найти, с целью обучить компьютер. Она обрабатывала корпоративные сайты на нескольких языках, а также идентичные переводы официальных документов и отчетов межправительственных организаций, таких как Организация Объединенных Наций и Европейская комиссия. Даже переводы книг в рамках проекта по сканированию книг были пущены в дело. Вместо трех миллионов тщательно переведенных предложений, используемых в проекте Candide, по словам Франца Оча, главы службы «Google Переводчик» и одного из ведущих специалистов в этой области, система Google охватывала миллиарды страниц документов с широким спектром качества перевода. Корпус этой системы содержал триллион слов и насчитывал 95 *миллиардов* англоязычных предложений, пусть и сомнительного качества[37].

Несмотря на беспорядочность входящих данных, служба Google лучше других систем. Ее переводы точнее, хотя и весьма далеки от совершенства. К тому же эта служба во много раз полнее других: к середине 2012 года она охватила более 60 языков, а теперь даже способна принимать голосовой ввод на 14 языках для моментального перевода. Поскольку она рассматривает язык лишь как беспорядочный набор данных, по которому можно судить скорее о вероятностях явлений, чем о них самих, служба может выполнять переводы между языками, в переводах на которые представлено недостаточно прямых соответствий, чтобы создать систему. В таких случаях (например, для хинди и каталонского языка) английский язык служит своеобразным мостом. Кроме того, эта система более гибкая, чем другие подходы, поскольку может добавлять и удалять слова по мере того, как они входят в обиход или устаревают.

Google Переводчик работает хорошо не потому, что в его основе заложен более разумный алгоритм. Как это было у Банко и Брилла из корпорации Microsoft, причина тому — большее количество входящих данных (но не всех подряд). Так, например, компании Google удалось использовать в *десятки тысяч раз* больше данных, чем системе

Candide компании IBM. И все потому, что в Google принимались беспорядочные данные. Корпус из триллиона слов, выпущенный Google в 2006 году, состоял из разбросанных фрагментов интернет-контента. Он стал «обучающим набором», по которому вычислялась вероятность того, что именно последует за тем или иным английским словом. Это был огромный шаг вперед, в корне отличающийся от предшественника — знаменитого Брауновского корпуса с миллионом английских слов, созданного в 1960-х годах. Благодаря более объемным наборам данных развитие обработки естественного языка шло семимильными шагами. На нем были основаны как системы распознавания голоса, так и системы компьютерного перевода. «Простые модели с множеством данных по результатам превосходят более сложные модели, основанные на меньшем количестве данных», — отметил Питер Норвиг, гуру искусственного интеллекта в компании Google, в статье «Необоснованная эффективность данных», написанной в соавторстве с коллегами[38].

Однако, как поясняют Норвиг и его коллеги, ключевым элементом была беспорядочность: «В некотором смысле этот корпус — шаг назад по сравнению с Брауновским корпусом, ведь его данные взяты с неотфильтрованных веб-страниц, а значит, содержат неполные предложения, а также орфографические, грамматические и прочие ошибки. Такой корпус не имеет примечаний с добавленными вручную пометками частей речи. Но то, что он в миллион раз больше Брауновского корпуса, перевешивает эти недостатки».

Больше данных — лучше результат

Аналитикам, которые работают с обычными выборками, трудно свыкнуться с беспорядочностью, которую они всю жизнь стремились предотвратить или искоренить. Статистики используют целый комплекс стратегий в целях снижения частоты появления ошибок при сборе выборок, а также для проверки выборок на наличие потенциальных систематических ошибок перед объявлением результатов. Этот комплекс стратегий включает в себя сбор выборок, который осуществляется специально обученными специалистами в соответствии с точным протоколом. Реализация стратегий,

направленных на сокращение числа ошибок, — дорогостоящее удовольствие, даже при ограниченном количестве точек данных. Что немаловажно, эти стратегии становятся невозможными в случае сбора данных в полном объеме — не только из-за чрезмерной стоимости, но и потому, что при таком масштабе вряд ли удастся равномерно соблюсти строгие стандарты сбора. И даже исключение человеческого фактора не решило бы проблему.

Двигаясь в сторону больших данных, мы будем вынуждены изменить свое представление о преимуществах точности. Пытаясь мыслить привычными категориями измерений в цифровом взаимосвязанном мире XXI века, мы упускаем важный момент. Одержимость точностью — не более чем артефакт аналогового мира, находящегося в информационной изоляции, где данные поистине были редкостью. На тот момент измерение каждой точки данных было крайне важно для результата, поэтому большое внимание уделялось тому, чтобы не допускать в анализе систематические погрешности.

В наше время нет такого дефицита информации. При переходе на всеобъемлющие наборы данных, которые охватывают всё или почти всё рассматриваемое явление, а не только его мизерную часть, нам уже не приходится беспокоиться об отдельных точках данных, привносящих в анализ систематические погрешности. Вместо того чтобы искоренять каждый неточный бит (что со временем обходится все дороже), мы выполняем вычисления, принимая во внимание беспорядочность.

Возьмем для примера беспроводные датчики, внедряемые на производстве. По всей территории нефтеперерабатывающего завода BP Cherry Point в Блейне (Вашингтон) расставлены беспроводные датчики, образующие невидимую сеть, которая производит огромные объемы данных в режиме реального времени. Неблагоприятные окружающие условия — сильная жара и электрические механизмы — могут время от времени искажать показания, приводя к беспорядочности данных. Но огромное количество поступающей информации компенсирует эти трудности. Измеряя нагрузку на трубы непрерывно, а не через определенные промежутки времени, компания BP выяснила, что некоторые виды сырой нефти более едкие, чем

другие. Прежде это не удавалось определить, а значит, и предотвратить[39].

Получая огромные массивы данных нового типа, в некоторых случаях можно пренебречь точностью, если удается спрогнозировать общие тенденции. Мы живем как раз в условиях такого парадокса. Небольшой магазин может подсчитать прибыль к концу дня вплоть до копейки, но мы не стали бы (да и не смогли бы) проделывать то же самое с ВВП страны. В условиях перехода к большим масштабам меняется не только ожидаемая степень точности, но и практическая возможность ее достижения. Отношение к данным как к чему-то несовершенному и неточному (пусть поначалу и вопреки логике) дает возможность делать всеобъемлющие прогнозы, а значит, лучше понимать окружающий мир.

Получается, что беспорядочность не является неотъемлемой частью больших данных как таковых. Она скорее результат несовершенства инструментов, которые мы используем для измерения, записи и передачи информации. Если бы технологии вдруг стали совершенными, проблема неточности исчезла бы сама собой. Беспорядочность — не внутренняя характеристика больших данных, а объективная реальность, с которой нам предстоит иметь дело. И, похоже, она с нами надолго. Как правило, кропотливое повышение точности нецелесообразно с экономической точки зрения, поскольку польза от гораздо большего количества данных выглядит более убедительно. Происходит смещение центра внимания, как и в предыдущую эпоху, когда специалисты по сбору статистики отказались от наращивания размеров выборки в пользу увеличения случайности. Теперь же мы готовы мириться с незначительными неточностями в обмен на дополнительные данные.

В рамках проекта Billion Prices Project[12] можно найти занимательный пример. Каждый месяц американское Бюро статистики труда публикует индекс потребительских цен (ИПЦ), который используется для расчета уровня инфляции. Эти цифры крайне важны для инвесторов и компаний. Федеральная резервная система учитывает ИПЦ при решении вопроса о повышении или понижении процентных ставок. Основной оклад компаний увеличивается с поправкой на

инфляцию. Федеральное правительство учитывает величину оклада при расчете пособий (таких как пособие по социальному обеспечению), а также процента, выплачиваемого по некоторым облигациям.

Чтобы получить эти цифры, сотни сотрудников бюро по телефону, факсу или лично связываются с магазинами и офисами в 90 городах по всей территории США. В итоге они формируют отчет из 23 000 цен на все товары и услуги — от помидоров до такси. На это уходит около 250 миллионов долларов США в год. В такую сумму обходятся однородные, понятные и упорядоченные данные. А к моменту публикации они успевают устареть на несколько недель.

Как показал финансовый кризис 2008 года, такое отставание может быть непростительным. Ответственным лицам нужно быстрее получать показатели инфляции, чтобы действовать эффективнее. Но с традиционными методами, которые сосредоточены на сборе выборок и придают большое значение точности, это невозможно.

В ответ на это два экономиста из Массачусетского технологического института (МТИ), Альберто Кавелло и Роберто Ригобон, предложили альтернативу — взять курс на большие данные, отличающиеся гораздо большей беспорядочностью. Используя программное обеспечение для сканирования веб-страниц, они ежедневно собирают полмиллиона цен на товары. Эти данные беспорядочны, и не все собранные точки данных легко сопоставимы. Но, объединив собранные большие данные с глубоко продуманными системами анализа, в рамках проекта удалось обнаружить дефляционные колебания цен, последовавшие сразу за банкротством инвестиционного банка Lehman Brothers в сентябре 2008 года. Те же, кто привык ориентироваться на официальные данные ИПЦ, смогли увидеть это только в ноябре.

Проект МТИ вырос до пяти миллионов продуктов от 300 розничных торговцев в 70 странах и дал начало коммерческой компании PriceStats, которая используется банками и другими заинтересованными лицами для принятия взвешенных экономических решений. Безусловно, полученные цифры требуют осторожного истолкования и лучше демонстрируют тенденции в области ценообразования, чем точные цены. Но поскольку в данном случае

сведений о ценах гораздо больше и они поступают в режиме реального времени, это дает ответственным лицам значительное преимущество.

Беспорядочность в действии

Во многих общественных и технологических областях мы склоняемся в пользу беспорядочности, а не точности. Рассмотрим классификацию контента. На протяжении веков люди разрабатывали таксономии и индексы для хранения и извлечения материалов. Такие иерархические системы всегда были несовершенными, и это подтвердит каждый, кто не понаслышке знаком с библиотечной картотекой. В мире малых данных эти системы были достаточно эффективны. Однако стоило увеличить масштаб на много порядков — и эти системы, в которых все якобы идеально размещено, разваливаются. На сайте для обмена фотографиями Flickr в 2011 году хранилось более шести миллиардов фотографий почти от ста миллионов пользователей. Было бы бесполезно пытаться пометить каждую из фотографий в соответствии со стандартными категориями. Разве среди них найдется категория «Кошки, похожие на Гитлера»?

На смену понятным таксономиям и, как предполагается, совершенным классификациям приходят новые механизмы — более беспорядочные, зато гораздо более гибкие. Они легче адаптируются к миру, который непрерывно развивается и изменяется. Загружая фотографии на сайт Flickr, мы добавляем к ним теги, то есть назначаем любое количество текстовых меток, и используем их для упорядочения и поиска материала. Пользователи создают и добавляют теги по своему усмотрению. Нет единой стандартизированной, предопределенной иерархии, классификации или таксономии, которых следует придерживаться. Чтобы добавить новый тег, достаточно ввести его. Добавление тегов фактически стало стандартом классификации веб-контента, который используется на сайтах социальных сетей, таких как Facebook, а также в блогах и на прочих ресурсах. Благодаря этому стандарту стало гораздо удобнее бороздить просторы веб-контента, особенно нетекстового (изображений, видео, музыки), для которого поиск по словам не подходит.

Конечно, в тегах возможны опечатки. Такие ошибки привносят неточность (не в сами данные, а только в их порядок), а это наносит удар по традиционному способу мышления, основанному на точности. Но взамен беспорядочности того, как устроены наши коллекции фотографий, мы получаем гораздо больший спектр меток и, соответственно, более широкий доступ к своим фотографиям. Мы можем объединять поисковые теги для фильтрации своих фотографий такими способами, которые были недоступны прежде. Принять неточность, присущую методу меток, — значит принять естественную беспорядочность окружающего мира. Это лекарство от более точных систем, которые пытаются навязать суматошному миру ложную стерильность, делая вид, что все на свете можно четко систематизировать. Вокруг еще столько всего, что не укладывается в рамки такой философии!

Многие популярнейшие сайты не скрывают свою симпатию к неточности. Взглянув на значок Twitter или на кнопку «Нравится» на веб-странице Facebook, можно увидеть количество других людей, которые их нажали. Пока числа небольшие, например 63, каждое нажатие идет в расчет. Но при больших количествах нажатий указывается лишь приблизительное количество, например 4 тысячи. Нельзя сказать, что система не знает точных цифр. Просто с увеличением масштаба точность уже не играет большой роли. Кроме того, числа могут меняться так быстро, что на момент отображения будут уже неактуальны. Такого же принципа придерживается почтовая служба Gmail компании Google, в которой время последних сообщений указывается с точностью до минуты, например «11 минут назад», но более длительные интервалы округляются, например «2 часа назад».

Область бизнес-аналитики и аналитического программного обеспечения долгое время строилась вокруг обещания клиентам «единой версии правды» — популярного выражения среди поставщиков технологий в этих областях в 2000-х годах. Руководители произносили эту фразу без иронии. Некоторые так поступают и до сих пор. Под этой фразой подразумевается, что все, кто получает доступ к информационно-технологическим системам компании, могут использовать одни и те же данные. А значит, отделам маркетинга и продаж не придется спорить, чьи данные о количестве клиентов и

продаж правильнее, еще до начала встречи. Исходя из сказанного, их интересы могут во многом совпадать, если факты излагаются единообразно.

Идея «единой версии правды» кардинально меняется. И суть не в том, чтобы согласиться с тем, что единой правды не существует. Важно понять, что гнаться за ней — неблагодарное дело. Для того чтобы пожинать плоды освоения масштабных данных, нужно признать, что беспорядочность здесь — в порядке вещей, и не нужно тратить лишнюю энергию на то, чтобы от нее избавиться.

Мы даже можем наблюдать, как характерные черты неточности проникают в одну из наименее терпимых к ней областей — проектирование баз данных. Для обычных механизмов системы управления базами данных (СУБД) требуются точные и хорошо структурированные данные, которые не просто хранятся, а разбиваются на «записи» с полями. Каждое поле содержит информацию конкретного типа и длины. Например, в числовое поле длиной в семь цифр невозможно записать сумму, равную десяти миллионам и более. А в поле для телефонных номеров не получится ввести «недоступен». Приспособиться к таким изменениям можно, только изменив структуру базы данных. Мы все еще воюем с этими ограничениями на компьютерах и смартфонах, когда программное обеспечение отказывается принимать данные, которые мы хотим ввести.

Индексы тоже предопределены, и это ограничивает возможности поиска. А чтобы добавить новый индекс, его создают с нуля, затрачивая время. Обычные реляционные базы данных предназначены для работы в области разреженных данных, которые можно и следует тщательно проверять. В такой области вопросы, на которые нужно ответить с помощью данных, известны изначально, поэтому база данных служит именно для эффективного ответа на них.

Однако эта точка зрения на хранение и анализ данных все более расходится с реальностью. Теперь в нашем распоряжении имеются большие объемы данных разного типа и качества. Данные редко вписываются в определенные категории, известные изначально. И вопросы, на которые мы хотели бы получить ответ, тоже часто возникают только в процессе сбора данных или работы с ними.

Эти реалии привели к созданию новых структур баз данных. Старые принципы создания записей и предопределенных полей, отражающих четко заданную иерархию информации, остались в прошлом. Долгое время самым распространенным языком доступа к базе данных был SQL («структурированный язык запросов»). Само название говорит о его жесткости. Но в последние годы произошел переход в сторону так называемой технологии NoSQL, при которой в базах данных не требуется предопределенная структура записей. Допускаются данные различного типа и размера. При этом они все так же доступны для поиска. Беспорядок, который допускается в структуре таких баз данных, компенсируется тем, что для их хранения и обработки требуется больше ресурсов. И все же, учитывая резкое падение затрат на хранение и обработку, этот компромисс мы можем себе позволить.

Пэт Хеллэнд, один из ведущих мировых авторитетов по вопросам проектирования баз данных в корпорации Microsoft, в статье «Если у вас слишком много данных, то и “достаточно хорошо” — уже хорошо» (If You Have Too Much Data, then ‘Good Enough’ Is Good Enough) описывает это явление как фундаментальный переход. Определив несколько основных принципов традиционного проектирования баз данных, которые были подорваны беспорядочными данными различной точности и происхождения, он изложил такие выводы: «Мы больше не можем претендовать на то, чтобы жить в чистом мире [информации]. Обработка больших данных влечет за собой неизбежные потери информации. Зато вы получаете быстрый результат». «Не страшно, если мы получаем ответы с потерями, — зачастую это вполне соответствует бизнес-потребностям», — подытожил Хеллэнд.

Традиционное проектирование баз данных обещает стабильное получение единообразного результата. Спрашивая у бухгалтера о состоянии баланса, вы ожидаете получить точную сумму. А повторив свой запрос через несколько секунд, вы хотели бы, чтобы система выдала такой же результат при условии, что ничего не изменилось. Однако по мере роста объема данных и увеличения количества пользователей, имеющих доступ к системе, поддерживать такое единообразие становится все труднее.

Большие наборы данных не хранятся централизованно. Как правило, они распределяются между несколькими жесткими дисками и компьютерами. Для обеспечения надежности и скорости запись может храниться в двух или трех разных расположениях. Если обновить запись в одном расположении, то данные в других расположениях будут считаться неправильными, пока их тоже не обновят. Традиционным системам было свойственно ожидать завершения всех обновлений. Но это менее практично, когда данные широко распространены, а сервер атакуется десятками тысяч запросов в секунду. В таком случае беспорядочность — неплохое решение.

Типичным примером перехода, о котором идет речь, стала популярность системы Hadoop — конкурирующего аналога системы Google MapReduce с открытым исходным кодом. Hadoop отлично справляется с обработкой больших объемов данных, разбивая их на мелкие фрагменты и выделяя участки для других компьютеров. Она исходит из того, что оборудование может отказать, поэтому создает резервную копию. Система также предполагает, что поступающие данные не упорядочены и не выверены (а по факту и не могут быть выверены до обработки из-за поистине огромного объема). При типичном анализе данных в первую очередь требуется выполнить ETL (от англ. Extract, Transform, Load — «извлечение, преобразование, загрузка»), чтобы переместить данные в расположение для их анализа. Hadoop обходится без таких тонкостей. Напротив, исходя из того, что количество данных настолько велико, что их невозможно переместить, Hadoop анализирует данные на месте.

Результат, получаемый на выходе, не настолько точен, как в случае реляционных баз данных: на него нельзя рассчитывать при запуске космического корабля или при подтверждении реквизитов банковского счета. Но со многими менее важными задачами, где суперточный ответ не требуется (скажем, с задачами по сегментированию клиентов для проведения специальных маркетинговых кампаний), Hadoop справляется намного быстрее, чем другие. С помощью Hadoop компания по выпуску кредитных карт Visa сумела сократить время обработки тестовых записей, накопленных за два года (73 миллиарда транзакций) с одного месяца до каких-то 13 минут. Подобное сокращение времени обработки ведет к преобразованиям в деловой

сфере. Возможно, оно не годится для формального учета, зато исключительно полезно, когда некоторая погрешность вполне допустима[40].

Принимая беспорядочность, взамен мы получаем чрезвычайно ценные услуги, недоступные при использовании традиционных методов и инструментов, учитывая всю масштабность данных. По некоторым оценкам, только 5% всех цифровых данных «структурированы», то есть представлены в форме, подходящей для традиционных баз данных. Отказываясь от беспорядочности, мы теряем оставшиеся 95% неструктурированных данных, таких как веб-страницы и видео. Допуская неточность, мы открываем окно в непознанный мир открытий.

Общество пошло на два неявных компромисса, которые уже настолько укоренились в нашем быту, что воспринимаются как естественный порядок вещей. Во-первых, мы не замахиваемся на огромные массивы данных, поскольку исходим из того, что это невозможно. Но этот сдерживающий фактор становится все менее актуальным, и мы можем многого добиться, ориентируясь на подход «*N* = всё».

Второй компромисс — качество информации. В эпоху малых данных точность ставилась превыше всего, ведь тогда собирали только малую часть информации, поэтому она должна была быть как можно более точной. Во многом это актуально и сейчас. Но в большинстве случаев важнее не строго соблюсти точность, а быстро получить общее представление о данных или тенденциях их развития.

Представление о том, как использовать всю совокупность информации, а не ее часть, и постепенное осознание преимуществ менее точных данных коренным образом меняют взаимодействие людей с окружающим миром. По мере того как методы работы с большими данными становятся неотъемлемой частью повседневной жизни, общество в целом устремляется к всеобъемлющему, более широкому, чем раньше, пониманию явлений — своего рода мышлению «*N* = всё». Возможно, мы станем менее требовательными к точности и однозначности в областях, где полагались на четкость и определенность (пусть даже сомнительные). Мы согласимся с таким

подходом при условии, что взамен получим более полную картину явлений. Так на картинах импрессионистов мазки кажутся беспорядочными при ближайшем рассмотрении, но отступите на шаг — и вы увидите величественную картину.

Большие данные со свойственной им полнотой и беспорядочностью помогают нам ближе подойти к осознанию реального положения вещей, чем это удавалось в условиях зависимости от малых данных и точности. Призыв к частичным, но точным данным вполне понятен. Наше постижение мира, возможно, было неполным, а порой и вовсе неверным в условиях ограниченности данных, поддающихся анализу, зато они давали ощущение уверенности и обнадеживающей стабильности. Кроме того, поскольку мы могли собрать и изучить лишь ограниченный объем данных, не возникало непреодолимого желания получить их абсолютно все и рассмотреть со всех возможных сторон. В узких рамках малых данных мы могли гордиться точностью, но, даже измеряя все до мельчайших подробностей, упускали из виду более масштабную картину.

Большие данные могут потребовать, чтобы мы научились спокойнее относиться к беспорядочности и неопределенности. Представления о точности, которые, казалось бы, служат нам ориентирами (например, что круглые фигуры подходят круглым отверстиям, существует только один ответ на вопрос и т. п.), лучше поддаются изменениям, чем мы можем предположить. Вместе с тем такое предположение, принятое на веру, приближает нас к пониманию реального положения вещей.

Описанные изменения образа мышления знаменуют радикальные преобразования. Они ведут к третьему шагу, который может во многом подорвать устои общества, основанного на понимании причин всех событий. Вместе с тем поиск логических взаимосвязей между данными и выполнение действий с ними (что и является темой следующей главы) зачастую дают вполне достойный результат.

Примечания к главе 3

Глава 4

Корреляция

В 1997 году 24-летний Грег Линден на время отложил свою докторскую диссертацию в области искусственного интеллекта в Вашингтонском университете, чтобы поработать над местным стартапом по продаже книг в интернете. Этот онлайн-магазин появился всего два года назад, но уже вел оживленную торговлю. «Мне очень понравилась идея продавать книги, продавать знания, а еще помогать людям находить следующий источник знаний, с которым они с удовольствием бы ознакомились», — вспоминает Грег. Этим магазином был Amazon.com, и Линден был нанят в качестве инженера-программиста для обеспечения бесперебойной работы сайта.

Среди сотрудников компании Amazon были не только технари. В то время там работала дюжина литературных критиков и редакторов, которые писали отзывы и предлагали новые наименования. Хотя история сайта Amazon хорошо знакома большинству людей, мало кто помнит о том, что его контент первоначально создавался вручную. Редакторы выбирали наименования, которые рекомендовались на веб-страницах Amazon. Редакторский отдел отвечал за так называемый «голос Amazon», который по праву считался гордостью компании и источником ее конкурентного преимущества. Примерно в то же время вышла статья в Wall Street Journal, в которой сотрудников отдела чествовали как самых влиятельных литературных критиков страны, поскольку им удавалось стимулировать высокий уровень продаж.

Затем Джефф Безос, основатель и CEO[13] Amazon, начал экспериментировать с многообещающей идеей: что если рекомендовать конкретные книги отдельным клиентам в зависимости от их предыдущих покупок? С момента начала деятельности Amazon компания накопила массу данных о каждом клиенте: о покупках, о просмотренных, но не приобретенных книгах и времени, затраченном на их просмотр, а также о книгах, приобретенных одновременно.

Объем данных был настолько внушительным, что поначалу Amazon приходилось обрабатывать их обычным способом — путем отбора выборки и ее анализа с целью выявить сходство между клиентами. Рекомендации выходили приблизительными. Купив книгу о Польше, вы получили бы массу предложений по Восточной Европе, а купив книгу о детях — завалены подобной литературой. «Как правило, вам предлагались небольшие вариации на тему вашей предыдущей покупки. И так до бесконечности, — вспоминает Маркус Джеймс, литературный критик Amazon в 1996–2001 годах, в своих мемуарах Amazonia. — Создавалось ощущение, что вы отправились за покупками с бестолковым советчиком»[41].

Грег Линден нашел решение. Он понял, что рекомендательной системе, по сути, не нужно сравнивать одних людей с другими, что к тому же было технически обременительно. Нужно всего лишь найти ассоциации среди самих продуктов. В 1998 году Линден и его коллеги заявили патент на метод совместной фильтрации «предмет-предмет». Изменение подхода принесло большую пользу.

Поскольку расчеты проводились заранее, рекомендации выдавались молниеносно. К тому же они были универсальными и включали товары из разных категорий. Поэтому, когда компания Amazon расширила ассортимент, рекомендательная система могла предлагать не только книги, но и фильмы или, скажем, тостеры. Кроме того, рекомендации стали намного точнее, поскольку система использовала все данные. «В отделе шутили, что, если система отлично себя зарекомендует, на сайте Amazon достаточно будет показывать только одну книгу — ту, которую вы купите следующей», — вспоминает Линден[42].

Теперь перед компанией стоял выбор, что отображать: отзывы, написанные штатными литературными критиками Amazon, или контент, созданный компьютером (личные рекомендации, списки бестселлеров и пр.); то, что говорят критики, или то, на что указывают действия клиентов? Это в буквальном смысле была борьба человека против компьютера.

Линден сравнил продажи, которые последовали за отзывами литературных критиков, и контент, созданный компьютером. Разница

оказалась внушительной. По словам Линдена, материалы, полученные на основе данных, принесли практически в сто раз больше продаж. Возможно, компьютеру и было неизвестно, почему клиент, читающий Хемингуэя, пожелает приобрести Фрэнсиса Скотта Фицджеральда. Но, похоже, это не имело значения. Продажи текли рекой. Редакторам озвучили точный процент продаж, которые компания Amazon *недополучала* при каждой публикации их отзывов в интернете, и отдел распустили. «Мне было очень жаль, что результат редакторского отдела оказался ниже, — вспоминает Линден. — Но данные не лгут, а цена была очень высока».

Сегодня считается, что третью всех своих продаж компания Amazon обязана своим рекомендательным системам, а также системам персонализации. С их помощью компания не только вытеснила с рынка большие книжные и музыкальные магазины, но и сотни местных книготорговцев, которые думали, что их личный подход укроет их от ветра перемен. Работа Линдена поистине произвела революцию в сфере электронной коммерции, поскольку этот метод был подхвачен практически всеми. Компания Netflix, которая занимается сдачей фильмов напрокат в интернете, три четверти новых заказов получает благодаря рекомендациям[43]. Следуя примеру Amazon, тысячи сайтов могут рекомендовать продукты, контент, друзей и группы для подписки, не зная толком, чем это все может заинтересовать их пользователей.

Для рассматриваемой задачи знание *почему* может быть полезно, но не столь важно. А вот знание *что* приводит к конкретным действиям. Эта истина способна изменить помимо электронной коммерции многие отрасли. Продавцам из разных сегментов рынка долгое время твердили, что им нужно понять, что заставляет клиентов совершить покупку, понять причины их решений. Высоко ценились профессиональные навыки и многолетний опыт работы. Но большие данные показывают, что есть и другой, в некотором смысле более эффективный подход. Рекомендательным системам Amazon удалось выявить любопытные корреляции, не зная их первопричины. Так что знания *что*, а не *почему* вполне достаточно.

Прогнозы и предрасположенности

Корреляции полезны в области малых данных. Но по-настоящему они раскрывают свой потенциал в контексте больших данных. С их помощью мы можем рассматривать явления проще, быстрее и отчетливее, чем раньше.

По сути, корреляция — количественное выражение статистической связи между двумя значениями. Сильная корреляция означает, что при увеличении одних значений данных другие значения, вероятнее всего, тоже увеличатся. Такие корреляции мы наблюдали, когда описывали Google Flu Trends: чем больше людей в конкретном географическом регионе ищут определенные ключевые слова в поисковой системе Google, тем выше заболеваемость гриппом в этом регионе. С другой стороны, слабая корреляция означает, что при увеличении одних значений данных другие значения практически не изменятся. Так, если провести корреляцию между размером обуви людей и тем, насколько они счастливы, мы обнаружим, что размер обуви мало что может рассказать о счастье человека.

Корреляции помогают анализировать объекты, выявляя не принципы их работы, а полезные закономерности. Безусловно, даже сильные корреляции не идеальны. Вполне возможно, что похожее поведение двух объектов — не более чем совпадение. Нет никаких гарантий, что даже сильные корреляции сумеют объяснить каждый случай. Не каждая рекомендация книг на сайте Amazon безошибочна. Корреляции дают не определенность, а лишь вероятность. Но в случае сильной корреляции между явлениями высока вероятность, что они взаимосвязаны. Многие могут подтвердить это, указав на полку, уставленную книгами по рекомендациям Amazon.

Корреляции дают возможность определять ценные закономерности явлений, чтобы подмечать их в настоящем и прогнозировать в будущем. Например, если событие A часто сопровождается событием B, нужно следить за B, чтобы спрогнозировать A. Такой подход позволяет уловить, чего вероятнее всего ожидать от события A, даже если мы не можем измерить или проследить его напрямую. Более того, это позволяет нам спрогнозировать дальнейшие события. Конечно, корреляции не могут предсказывать будущее — они лишь могут

спрогнозировать его с определенной вероятностью. Но и это чрезвычайно ценно.

Walmart — крупнейшая в мире сеть розничной торговли, которая насчитывает более двух миллионов сотрудников. Ее объем продаж составляет около 400 миллиардов долларов — больше, чем ВВП большинства стран. Перед наплывом огромных массивов данных, порожденных интернетом, компания Walmart располагала, пожалуй, самым большим хранилищем данных среди коммерческих компаний в США. В 1990-х годах она произвела переворот в розничной торговле, внедрив учет всей продукции в виде данных с помощью сети Retail Link. Компания Walmart предоставила поставщикам возможность самим контролировать темпы и объемы продаж и запасов. Благодаря такой прозрачности Walmart удалось вынудить поставщиков самостоятельно заботиться о своей логистике. В большинстве случаев Walmart не выступает «собственником» продукта до момента продажи, тем самым снимая с себя риск обесценения запасов и снижая затраты. По сути, с помощью данных Walmart удалось стать крупнейшим комиссионным магазином.

О чем могут рассказать все эти накопленные данные, если их проанализировать должным образом? В сотрудничестве с экспертом в области обработки чисел Teradata (ранее — почитаемая корпорация NCR) компания Walmart стремилась выявить интересные корреляции. В 2004 году она взялась за изучение своих гигантских баз данных прошлых операций, которые включали не только информацию о товарах, приобретенных каждым клиентом, и общей сумме покупки, но и об остальных товарах в корзине, о времени суток и даже о погоде. Это дало компании возможность заметить, что перед ураганом росли объемы продаж не только фонариков, но и печенья PopTarts, а также сладких сухих американских завтраков. Поэтому, как только надвигалась буря, в магазинах Walmart поближе к витрине выкладывались коробки Pop-Tarts и припасы на случай урагана для удобства клиентов, снующих снаружи и внутри магазина, и, разумеется, для увеличения продаж[44].

В прошлом специалистам из главного офиса пришлось бы заранее собрать данные и проверить идею. Теперь же, имея столько данных и

улучшенные инструменты работы с ними, выявлять корреляции стало куда быстрее и дешевле.

Корреляционный анализ показал свою высокую эффективность задолго до больших данных. Эту концепцию в 1888 году выдвинул сэр Фрэнсис Гальтон, двоюродный брат Чарльза Дарвина, заметив взаимосвязь между ростом мужчин и длиной их предплечий. Математические расчеты, лежащие в основе корреляционного анализа, относительно просты и надежны. Благодаря этим характерным особенностям анализ стал одним из наиболее широко используемых статистических показателей. Но до перехода на большие данные корреляции имели ограниченную эффективность. Поскольку данные были скудными, а их сбор — дорогостоящим, специалисты по сбору статистики нередко интуитивно определяли вероятную закономерность, а затем собирали соответствующие данные и проводили корреляционный анализ, чтобы выяснить, насколько эта закономерность соответствовала действительности. В контексте службы Google Flu Trends это означало бы, что нужно предположить условия поиска, которые коррелируют с распространением гриппа, а затем провести корреляционный анализ, чтобы убедиться в правильности этих предположений. Учитывая набор данных Google из 50 миллионов различных условий поиска и более трех миллиардов запросов в день, интуитивно выбрать наиболее подходящие из них для тестирования не представляется возможным.

Таким образом, в эпоху малых данных корреляционный анализ утратил свою первостепенность. Даже сегодня термин «интеллектуальный анализ данных» в научных кругах звучит неодобрительно. Его противники острят: «Поиздевайтесь над данными достаточно долго — и они будут готовы признать что угодно».

Вместо того чтобы полагаться на простые корреляции, эксперты пытались интуитивно нащупать подходящие закономерности, исходя из гипотез в рамках определенных теорий — абстрактных представлений о принципах работы чего-либо. Затем эксперты получали соответствующие данные и проводили корреляционный анализ для проверки этих закономерностей. Если они оказывались ошибочными, эксперты, как правило, упрямо пробовали еще раз (на случай, если данные были собраны неправильно), пока, наконец, не

признавали, что исходная гипотеза (или даже теория, на которой она основана) требует доработки. Знания совершенствовались путем проб и ошибок, связанных с гипотезами. Процесс был очень медленным, поскольку личные и общие предубеждения мешали объективно оценить разработанные гипотезы, их применение и выбранные в итоге закономерности. И все это для того, чтобы в большинстве случаев в итоге узнать, что мы ошибались. Это был трудоемкий процесс, зато он годился для работы с малыми данными.

В эпоху больших данных невозможно определить переменные, которые следует рассматривать, лишь на основе личных предположений. Наборы данных слишком велики, а рассматриваемые области, пожалуй, слишком сложны. К счастью, многие ограничения, которые вынуждали нас применять подход на основе гипотез, уже не столь существенны. Теперь у нас настолько много данных и вычислительной мощности, что не приходится вручную выбирать одну закономерность или небольшую горстку наиболее вероятных, а затем изучать их по отдельности. Теперь сложные вычислительные процессы сами выбирают лучшую закономерность, как это было в службе Flu Trends, которая легко и точно обнаруживала лучшие условия поиска из 50 миллионов самых популярных запросов, протестировав 450 миллионов математических моделей.

Для того чтобы понимать окружающий мир, теперь не обязательно изучать рабочие гипотезы о том или ином явлении. А значит, не нужно развивать гипотезу о возможных поисковых запросах людей, чтобы узнать время и территорию распространения гриппа. Не нужно вдаваться в подробности того, как авиакомпании назначают цены на билеты. Не нужно заботиться о кулинарных вкусах покупателей Walmart. Вместо этого достаточно провести корреляционный анализ на основе больших данных, чтобы узнать, какие поисковые запросы наиболее характерны для гриппа, грядет ли рост цен на авиабилеты или чем обеспокоенные домоседы запасаются на время бури. Вместо подверженного ошибкам подхода на основе гипотез благодаря корреляциям между большими данными у нас есть подход, построенный на данных. И он может быть менее предвзятым, более точным и наверняка менее трудоемким.

В основе больших данных лежат прогнозы на основе корреляций. Они используются все чаще, и мы порой недооцениваем их новизну. Практическое применение прогнозов со временем будет только расширяться.

Для прогнозирования поведения отдельных лиц существует кредитная оценка заемщика. Компания Fair Isaac Company, известная как FICO, ввела это понятие в 1950-х годах. В 2011-м FICO ввела еще одно понятие — «оценка приверженности лечению». Она анализирует множество переменных, в том числе тех, которые, казалось бы, не имеют отношения к делу (например, как долго люди не меняли место жительства или работы, состоят ли они в браке и имеют ли собственный автомобиль), для того чтобы определить вероятность того, примет ли пациент назначенное лекарство. Оценка помогла бы медицинским сотрудникам экономить средства: они знали бы, кому следует делать напоминания. Между владением автомобилем и приемом антибиотиков нет причинно-следственных связей. Это чистой воды корреляция. Но она вдохновила исполнительного директора компании FICO гордо заявить на встрече инвесторов в 2011 году: «Мы знаем, что вы собираетесь делать завтра»[45].

Крупное кредитное бюро Experian предлагает продукт Income Insight, который прогнозирует уровень доходов людей на основе их кредитной истории. Проанализировав огромную базу данных кредитных историй в сравнении с анонимными данными о налогах, полученными из налоговой службы Америки, эта программа подготовила соответствующую оценку. В то время как проверка доходов определенного лица стоит около 10 долларов, Experian продает свою оценку менее чем за 1 доллар. Таким образом, в некоторых случаях использование закономерностей экономически выгоднее, чем волокита с получением нужных данных. Тем временем другое кредитное бюро, Equifax, продает «индекс платежеспособности» и «индекс дискреционных расходов», которые сулят прогноз благосостояния отдельных лиц[46].

Поиск корреляций находит все более широкое применение. Изучив идею использования кредитных отчетов и данных потребительского маркетинга, крупная страховая компания Aviva внедрила ее вместо

анализа образцов крови и мочи для определенных заявителей. Полученная информация помогала выявлять лиц, наиболее подверженных риску развития высокого артериального давления, диабета или депрессии. Этот метод основывался на данных об образе жизни, включая сотни переменных (таких как хобби, посещаемые сайты и время, затрачиваемое на просмотр телевизора), а также смете поступлений.

Прогнозная модель компании Aviva, разработанная компанией «Делойт», по праву считалась полезной для выявления рисков для здоровья. Свое намерение внедрить аналогичные проекты подтвердили страховые компании Prudential и AIG. Преимущество подхода заключалось в том, что он позволял заявителям избежать неприятных анализов. Этот подход экономил страховым компаниям по 125 долларов с человека, в то время как стоимость самого подхода на основе данных составляла около пяти долларов[47]. Некоторые ужаснутся, словно компании станут использовать кибердоносчиков, которые шпионят за каждым щелчком мыши. Возможно, люди подумали бы дважды, прежде чем посетить сайт экстремальных видов спорта или посмотреть комедийное шоу, прославляющее домоседов, если бы знали, что это может привести к повышению их страховых взносов. Это было бы страшным нарушением свободы взаимодействия с информацией. С другой стороны, польза системы состояла в том, что она способствовала бы увеличению количества застрахованных лиц. А это хорошо как для общества, так и для страховых компаний.

Корреляции между большими данными применялись и в американском розничном магазине сниженных цен Target, пример которого достоин подражания. Уже не первый год Target опирается на прогнозы, основанные на корреляциях между большими данными. В своем непривычно кратком отчете Чарльз Дахигг, бизнес-корреспондент New York Times, рассказал, откуда Target узнает, что женщина беременна, если она явно об этом не сообщала. Если коротко, нужно принимать в расчет все возможные данные и позволить корреляциям выявить нужные закономерности.

Знать о том, что в семье клиента ожидается пополнение, очень важно для магазинов розничной торговли, поскольку в этот

переломный момент в жизни пары ее торговое поведение открыто для перемен — разведки новых магазинов и новых брендов. Розничные продавцы сети Target обратились в свой отдел аналитики, чтобы узнать, возможно ли по модели покупок определенного человека судить о том, что он ожидает пополнение.

В первую очередь отдел аналитики обратил внимание на историю покупок женщин, которые зарегистрировались в реестре Target на получение подарка к рождению ребенка. Специалисты Target заметили, что популярной покупкой среди зарегистрировавшихся женщин примерно на третьем месяце беременности был лосьон без запаха. Спустя несколько месяцев женщины, как правило, покупали пищевые добавки (магний, кальций, цинк и пр.). В итоге компания выявила около двух десятков характерных продуктов, по которым каждому клиенту можно было присвоить оценку «прогнозируемой беременности». С помощью корреляций розничным магазинам даже удавалось определять дату родов с небольшой погрешностью, и они стали отправлять соответствующие купоны на каждом этапе беременности. Такое нацеливание рекламных кампаний и впрямь соответствовало названию компании — Target (англ. *цель*).

Поиск закономерностей в социальном контексте — лишь один из способов применения методов работы с большими данными. Не менее эффективны корреляции при работе с новыми типами данных, которые используются для решения повседневных задач.

В бизнесе все шире применяется метод прогностической аналитики для определения предстоящих событий. Это может быть алгоритм для выявления музыкальных хитов, который популярен в музыкальной сфере и позволяет звукозаписывающим лейблам лучше ориентироваться, на кого стоит делать ставки. Или же алгоритм предотвращения больших механических неисправностей и разрушений конструкции: все чаще на машинах, двигателях и элементах инфраструктуры, таких как мосты, размещают датчики для отслеживания получаемых данных (показателей тепла, вибрации, нагрузки, звука и пр.).

Если речь идет о поломке, она, как правило, происходит не сразу, а развивается постепенно, с течением времени. Собрав все данные, можно заметить явные признаки, предшествующие поломке:

жужжание и перегрев двигателя. Система сравнивает эту модель поведения с обычной и выявляет несоответствия. Обнаружив отклонения на ранней стадии, система отправляет предупреждение. Таким образом, вы успеете заблаговременно заменить поврежденную часть на новую и предупредить проблему. Система определяет, а затем отслеживает закономерности, тем самым прогнозируя будущие события.

Транспортная компания UPS с середины 2000-х годов использует прогнозный анализ для контроля своего 60-тысячного автопарка в США и выполнения своевременного профилактического обслуживания. Поломка на дороге причиняет массу неудобств, включая отправку запасного грузового автомобиля, задержки поставок и погрузок, а также привлечение дополнительных сотрудников. Поэтому в компании UPS существовало правило заменять отдельные части раз в два-три года. Но это было неэффективно, поскольку некоторые части оставались в хорошем состоянии. Благодаря измерению и отслеживанию деталей транспортного средства компания UPS сэкономила миллионы долларов, заменив только те части, которые нуждались в замене. Однажды компании даже удалось определить, что группа новых транспортных средств содержала бракованную деталь, которая неминуемо привела бы к неприятностям, не будь вовремя замечена[48].

Подобным образом к мостам и зданиям крепят датчики, чтобы отслеживать признаки износа. Такие же датчики внедряются на крупных химических и нефтеперерабатывающих заводах, где поломанная деталь оборудования может остановить все производство до момента ее замены. Стоимость сбора и анализа данных для принятия своевременных мер экономит средства по сравнению с тем, во что обходятся простои. Отметим, что прогностическая аналитика не в состоянии объяснить причину проблемы (из-за чего перегрелся двигатель — из-за потертого ремня вентилятора или плохо закрученного винта) — она только выявляет саму проблему. Корреляции показывают *что*, а не *почему*. Но, как видно, в большинстве случаев этого достаточно.

С помощью подобных методов обеспечивается нормальное функционирование человеческого организма. Когда к пациенту в больнице прикрепляют массу трубок, проводов и инструментов, формируется большой поток данных. Одна только ЭКГ выдает 1000 показателей в секунду. В настоящее время используется или хранится только часть получаемых данных. Большинство данных попросту выбрасывается, хотя и несет в себе важную информацию о состоянии пациента и его реакции на лечение. А в совокупности с аналогичными данными других пациентов эти сведения могли бы составить уникальную аналитическую картину того, какое лечение эффективно, а какое — нет.

Возможно, отсеивание данных было рациональным в то время, когда их сбор, хранение и анализ были дорогостоящими и трудоемкими. Но ситуация изменилась. Теперь Кэролин Макгрегор вместе с командой исследователей из Технологического института университета провинции Онтарио и компании IBM сотрудничает с рядом больниц для разработки программного обеспечения, которое получает и обрабатывает данные о состоянии пациента в режиме реального времени. Затем они используются для принятия более взвешенных диагностических решений в отношении преждевременно рожденных («недоношенных») младенцев. Система отслеживает 16 различных потоков данных, таких как частота сердечных сокращений, частота дыхания, температура, артериальное давление и уровень кислорода в крови, что вместе составляет около 1260 точек данных в секунду[49].

Система способна обнаружить едва уловимые изменения в состоянии недоношенных детей, которые сигнализируют о начале развития инфекции за сутки до появления явных симптомов. «Вы не можете увидеть их невооруженным глазом, но компьютеру это под силу», — поясняет доктор Макгрегор. Система полагается не на причинно-следственные связи, а на корреляции. Она сообщает, *что* происходит, а не *почему*. И это вполне отвечает ее назначению. Заблаговременное предупреждение позволяет врачам раньше и к тому же с более щадящим медицинским вмешательством приступить к лечению инфекции или же раньше узнать, что лечение неэффективно.

И то и другое благотворно сказывается на результатах лечения пациентов. В будущем эта технология наверняка будет реализована для всех пациентов и условий. И пусть алгоритм не принимает решения, зато компьютеры делают все от них зависящее, чтобы помочь медикам как можно лучше выполнять свои обязанности.

Поразительно, как с помощью анализа больших данных доктору Макгрегор удалось выявить корреляции, которые в известном смысле бросают вызов традиционным представлениям врачей. Она обнаружила, что выраженное постоянство жизненно важных показателей, как правило, служит предвестником серьезной инфекции. Звучит странно, ведь мы полагаем, что именно ухудшение этих показателей должно предшествовать полномасштабной инфекции. Можете представить себе поколения врачей, которые по окончании рабочего дня проверяют состояние пациента и, убедившись, что оно стабилизировалось, решают, что все в порядке и можно идти домой. И только безумный звонок медсестры посреди ночи разбудит их и сообщит, что, вопреки их предположению, состояние пациента резко пошло на ухудшение.

Полученные данные свидетельствуют о том, что стабильность состояния недоношенных детей не служит признаком улучшения, а скорее больше похожа на затишье перед бурей: тело как будто велит крошечным органам мобилизовать все силы и приготовиться к предстоящим трудностям. Но мы не можем быть абсолютно уверены, ведь это лишь корреляция — здесь нет места причинно-следственным связям. Чтобы выявить эти скрытые взаимосвязи среди множества составляющих, понадобилось непостижимое количество данных. Вне всякого сомнения, большие данные спасают жизни.

Иллюзии и иллюминации

В мире малых данных корреляционный анализ не был намного лучше или дешевле исследований причинно-следственных связей. Ввиду небольшого количества данных, как правило, и то и другое исследования начинались с гипотезы, которая затем проверялась и находила свое подтверждение либо опровергалась. Поскольку в обоих случаях отправной точкой служила гипотеза, оба подхода были

одинаково чувствительны к предвзятости и ошибочным предположениям. Необходимые данные для корреляционного анализа часто были недоступны, а их сбор влек за собой большие расходы. Сегодня при наличии огромного количества данных это не такие уж весомые препятствия.

Существует еще одно отличие, которое только начинает приобретать все большее значение. В эпоху малых данных в большинстве случаев корреляционный анализ ограничивался поиском линейных отношений, в частности из-за недостаточной вычислительной мощности. При таких отношениях усиление закономерности привело бы к определенным известным изменениям рассматриваемого явления. Но, безусловно, в жизни многое куда сложнее. Полноценный комплексный анализ определяет так называемые нелинейные отношения между данными. Наглядно их можно увидеть, когда данные нанесены на график. Для того чтобы выявить эти данные, нужно воспользоваться техническими инструментами. Нелинейные отношения не только гораздо подробнее линейных, но и более информативны для руководителей.

В течение многих лет экономисты и политологи считали, что счастье напрямую связано с уровнем доходов: чем больше доход, тем человек счастливее. Однако график данных показывает, что там, где статистические инструменты проводят линейную корреляцию, в игру вступают более сложные динамические изменения. При уровне доходов ниже 10 000 долларов каждое их увеличение приводило к большему ощущению счастья, но рост доходов выше этого уровня мало что менял. Если нанести эти данные на график, получилась бы скорее кривая линия, чем прямая, которую сулил статистический анализ.

Это стало важным открытием для политиков. При линейной корреляции было понятно: для того чтобы сделать народ счастливее, нужно увеличить его доходы. Но как только удалось определить нелинейные отношения, эта рекомендация изменила свой ракурс: нужно сосредоточиться на увеличении доходов бедных слоев населения, поскольку, как показали данные, это даст бо́льшую отдачу от затраченных средств[50].

Более сложные корреляционные отношения только добавляют беспорядочности. Неравномерность прививок от кори среди населения и суммы, которые люди тратят на здравоохранение, казалось бы, взаимосвязаны. Тем не менее корреляция представлена не в виде аккуратной линии, а несимметричной кривой. По мере того как расходы людей на здоровье растут, неравномерность охвата населения прививками, как ни странно, снижается, но если затраты на здравоохранение одного человека продолжают расти, неравномерность охвата прививками неожиданно увеличивается. Для сотрудников здравоохранения это важнейшее открытие, которое невозможно было бы совершить с помощью простого линейного корреляционного анализа[51].

Эксперты только начали разрабатывать необходимые инструменты для определения и сравнения нелинейных корреляций. Развитию методов корреляционного анализа способствует быстро растущий набор новых подходов и программ, которые способны выделять связи, отличные от причинно-следственных, с разных точек зрения, подобно тому как художники-кубисты изображали лицо женщины одновременно с нескольких ракурсов. Один из самых ярких примеров — быстро растущая область сетевого анализа. С ее помощью можно определять, измерять и рассчитывать самые разные узлы и связи — от друзей на Facebook до событий, предшествовавших судебным решениям, и сведений о том, кто кому звонит по мобильному телефону. Вместе эти инструменты предоставляют новые мощные способы отвечать на непричинные, эмпирические вопросы.

В эпоху больших данных корреляционный анализ вызовет волну новых идей и полезных прогнозов. Мы обнаружим связи, которые не замечали прежде, и поймем сложные технические и социальные движущие силы, суть которых уже давно перестали улавливать, несмотря на все усилия. А самое главное, корреляции помогают нам познавать мир, спрашивая в первую очередь *что*, а не *почему*.

Поначалу может показаться, что это противоречит здравому смыслу. Людям свойственно постигать мир сквозь призму причинно-следственных связей, исходя из убеждения, что все имеет свою причину, стоит только хорошенько присмотреться. Узнать причину,

которая стоит за тем или иным явлением, — разве не это должно быть нашим высшим устремлением?

Из глубины веков тянется философская дискуссия о том, существует ли причинность на самом деле. Если каждое явление имеет свою причину, то логика подсказывает, что мы, по сути, ничего не решаем. Выходит, человеческой воли на самом деле не существует, поскольку наши мысли и принимаемые решения имеют причину, которая имеет свою причину, и т. д. Вся линия жизни определяется причинами, которые приводят к определенным последствиям. Таким образом, философы спорили о роли причинности в нашем мире, а порой и противопоставляли ее свободе выбора. Однако обсуждение этой полемики не входит в наши планы.

Говоря о том, что люди смотрят на мир сквозь призму причинно-следственных связей, мы, как правило, имеем в виду два основных способа постижения мира: с помощью быстрых, иллюзорных причинно-следственных связей и путем медленных, методичных казуальных экспериментов. Корреляции между большими данными изменят роль и того и другого, и в первую очередь — нашего интуитивного желания искать причинно-следственные связи.

Мы склонны предполагать причины даже там, где их нет. Это не связано ни с культурой или воспитанием, ни с уровнем образования человека. Такова особенность человеческого мышления. Когда мы рассматриваем два последовательных события, наш ум одолевает желание увидеть связь между ними. Вот три предложения: «Родители Фреда прибыли поздно. Вот-вот должны были подойти поставщики. Фред злился».

Читая их, мы сразу интуитивно определяем, почему Фред злился: не потому что поставщики были уже на подходе, а потому что его родители припозднились. Это не следует из предоставленной информации. Однако мы не можем удержаться от умозаключения, что наши предположения — причинно-следственные связи, основанные на полученных фактах.

Дэниел Канеман, профессор психологии в Принстоне, который получил Нобелевскую премию по экономике в 2002 году, на этом примере показывает, что нам свойственны две формы мышления. Одна — быстрая и не требует больших усилий. Она позволяет делать

выводы за считаные секунды. Другая форма — медленная, трудоемкая и требует «обдумывания» того или иного вопроса[52].

Быстрый способ мышления по большей части склонен находить причинно-следственные связи даже там, где их нет. Он предвзято воспринимает информацию для подтверждения имеющихся знаний и убеждений. В древние времена быстрый способ мышления был полезен и помогал выжить в опасном окружении, где, как правило, приходилось принимать решения мгновенно и в условиях ограниченной информации, но зачастую он далек от установления истинной причины тех или иных следствий.

Канеман утверждает, что, увы, очень часто в повседневной жизни мозг ленится думать медленно и методично. Тогда в дело вступает быстрый способ мышления. В результате мы часто «видим» мнимые причинно-следственные связи, а значит, совершенно неправильно воспринимаем окружающий мир.

Подхватив грипп, дети нередко слышат от родителей, что заболели из-за того, что не носят шапку и варежки в холодную погоду. Однако между заражением гриппом и тем, чтобы одеться теплее, нет прямой причинно-следственной связи. Почувствовав недомогание после ресторана, мы интуитивно будем пенять на еду, которую съели там (и, возможно, обходить стороной этот ресторан в будущем), хотя внезапное острое расстройство пищеварения может быть вызвано и другими причинами, например, если пожать руку зараженному человеку. Быстрое мышление запрограммировано быстро переходить к казуальным выводам, которые выдает мозг. И это часто приводит нас к неправильным решениям.

Вопреки общепринятому мнению, внутреннее ощущение причинности не углубляет нашего понимания мира. Во многих случаях это не более чем мыслительный «сокращенный путь», который дает нам иллюзию понимания, а на самом деле оставляет в неведении. Так же как выборки упрощали задачу, когда мы не могли обработать все данные, наш мозг использует познание причинности, чтобы избежать долгих и мучительных раздумий.

В мире малых данных могло пройти немало времени, прежде чем становилось ясно, насколько предполагаемые причинно-следственные

связи ошибочны. В дальнейшем это изменится. Корреляции больших данных станут регулярно использоваться для опровержения предполагаемых причинно-следственных связей, убедительно показывая, что часто между следствием и его предполагаемой причиной мало, а то и вовсе нет статистической связи. А пока «быстрое мышление» заменяет нам масштабную и длительную проверку действительности.

Будем надеяться, что стремление познать мир заставит нас думать глубже (и размереннее). Но даже медленное мышление — второй способ, которым люди распознают причинные связи, — изменится ввиду корреляций между большими данными.

Категории причинности настолько прочно вошли в нашу повседневную жизнь, что мы полагаем, что причинные связи легко показать. Это не так. В отличие от корреляций, математика которых относительно проста, причинность не имеет очевидных математических «доказательств». Мы не можем с легкостью выразить ее в виде обычных уравнений. Таким образом, даже если думать медленно и старательно, то отыскать убедительные причинно-следственные связи непросто. Наш мозг привык к тому, что информации всегда недостаточно, поэтому мы склонны делать выводы на основе ограниченного количества данных. Хотя, как правило, внешних факторов слишком много, чтобы сводить результат к определенной причине.

Возьмем, к примеру, вакцину против бешенства. 6 июля 1885 года к французскому химику Луи Пастеру привели девятилетнего Йозефа Майстера, которого укусила бешеная собака. Пастер как раз работал над экспериментальной вакциной против бешенства. Родители Майстера умоляли Пастера применить вакцину, чтобы вылечить их сына. Он согласился, и Йозеф Майстер выжил. В прессе пошла слава о том, что Пастер спас мальчика от верной мучительной смерти.

Но спас ли на самом деле? Как оказалось, в среднем лишь один из семи человек, укушенных бешеной собакой, заболевает. Даже если предположить, что экспериментальная вакцина Пастера была эффективной, она понадобилась бы только в одном из семи случаев. С вероятностью около 85% мальчик выжил бы и так.

В данном случае считалось, что Йозеф Майстер вылечился благодаря введению вакцины. Но под вопросом остаются две причинно-следственные связи: одна — между вакциной и вирусом бешенства, другая — между укусом бешеной собаки и развитием болезни. Даже если первая связь верна, то вторая — лишь в редких случаях.

Ученым удалось решить вопрос наглядности причинно-следственных связей с помощью экспериментов, в которых можно было применить или исключить отдельно взятую предполагаемую причину. Если применение причины влияло на результат, это означало наличие причинно-следственной связи. Чем тщательнее контролировались обстоятельства, тем выше была вероятность того, что эта связь правильная.

Таким образом, как и корреляции, причинность редко удается (если вообще возможно) доказать. Можно лишь показать ее с высокой степенью вероятности. Но, в отличие от корреляций, эксперименты для подтверждения причинно-следственных связей, как правило, неприменимы на практике или ставят непростые этические вопросы. Какие эксперименты помогут определить лучшие среди 50 миллионов условий поиска, прогнозирующих грипп? А в случае прививки от бешенства — неужели мы смогли бы допустить мучительную смерть десятков, а может, и сотен пациентов в качестве «контрольной группы», которой не сделали прививку, имея нужную вакцину? Даже применимые на практике эксперименты остаются дорогостоящими и трудоемкими.

Расчет корреляций, как правило, проводится быстрее и с меньшими затратами. В отличие от причинно-следственных связей, существуют математические и статистические методы для анализа корреляций, а также необходимые цифровые инструменты для уверенной демонстрации силы взаимосвязей.

Корреляции не только ценны сами по себе, но и указывают способ исследования причинно-следственных связей. Демонстрируя потенциальную взаимосвязь между явлениями, они могут стать предметом дальнейшего исследования с целью убедиться в наличии причинно-следственной связи и выяснения ее причин. Этот недорогой и быстрый механизм фильтрации снижает затраты на причинно-

следственный анализ за счет специально контролируемых экспериментов. Благодаря корреляциям мы имеем возможность уловить важные переменные и с их помощью провести эксперименты для исследования причинности.

Однако необходимо проявить осторожность. Корреляции — мощный инструмент не только потому, что они показывают полную аналитическую картину, но и потому, что делают ее понятной. Но, как правило, эта картина омрачается, как только мы снова начинаем искать причинность. Kaggle —компания, которая организует открытые конкурсы по интеллектуальному анализу данных среди компаний, — провела конкурс по анализу качества подержанных автомобилей. Агент по продаже подержанных автомобилей предоставил данные, на основе которых конкурсанты-статистики должны были создать алгоритм, прогнозирующий, какие из автомобилей, представленных на аукционе перекупщиков, вероятнее всего, имеют неисправности. Корреляционный анализ показал, что вероятность неисправностей автомобилей, окрашенных в оранжевый цвет, гораздо ниже (примерно наполовину), чем среди остальных автомобилей.

Даже сейчас, читая об этом, мы тут же задумываемся, в чем причина. Может быть, владельцы оранжевых автомобилей — настоящие автолюбители и лучше заботятся о своих автомобилях? Может, индивидуальная покраска означает, что автомобиль обслуживался более внимательно? Или оранжевые автомобили более заметны на дороге, а значит, ниже вероятность их участия в ДТП и потому они в лучшем состоянии на момент перепродажи?

Быстро же мы попали в сети альтернативных причинных гипотез! Наши попытки пролить свет на положение вещей делают эти гипотезы еще более размытыми. Корреляции есть, и мы можем показать их математически, чего не скажешь о причинно-следственных связях. Так что было бы неплохо удержаться от попыток объяснить причину корреляций в поиске ответа на вопрос *почему* вместо *что*. Иначе мы могли бы смело советовать владельцам автомобилей красить свои развалюхи в оранжевый цвет, чтобы сделать их запчасти менее дефектными (что само по себе полный вздор).

Становится понятно, что корреляции на основе достоверных данных превосходят большинство интуитивно понятных причинно-

следственных связей, то есть результат «быстрого мышления». Растет и количество случаев, когда быстрый и понятный корреляционный анализ оказывается более полезным и, очевидно, более эффективным, чем медленное причинное мышление, воплощенное в виде тщательно контролируемых (а значит, дорогостоящих и трудоемких) экспериментов.

В последние годы ученые пытались снизить затраты на такие эксперименты, например, искусно сочетая соответствующие опросы для создания «псевдоэкспериментов». Благодаря этому можно было повысить рентабельность некоторых исследований причинности. Однако эффективность корреляций трудно превзойти. Кроме того, как мы говорили, корреляционный анализ сам по себе служит помощником в таких исследованиях, подсказывая экспертам наиболее вероятные причины.

Таким образом, наличие данных и статистических инструментов преобразует роль не только быстрых, интуитивно улавливаемых причинно-следственных связей, но и взвешенного причинного мышления. Когда нам нужно исследовать не само явление, а именно его *причину*, как правило, лучше начать с корреляционного анализа больших данных и уже на его основе проводить углубленный поиск причинно-следственных связей.

На протяжении тысячелетий люди пытались понять принципы мироздания, стараясь найти причинно-следственные связи. Какую-то сотню лет назад, в эпоху малых данных, когда не было статистики, оперировали категориями причинности. Но все меняется с приходом больших данных.

Причинно-следственные связи не утратят своей актуальности, но перестанут быть главным источником знаний о том или ином предмете. В эпоху больших данных то, что мы считаем причинностью, на самом деле не более чем частный случай корреляционной связи. Хотя порой мы по-прежнему хотим выяснить, объясняют ли причинно-следственные связи обнаруженную корреляцию. Большие данные, напротив, ускоряют корреляционный анализ. И если корреляции не заменяют исследование причинности, то направляют его и предоставляют нужную информацию. Наглядным примером служат загадочные взрывы канализационных люков на Манхэттене.

Задача с канализационными люками

Ежегодно несколько сотен люков в Нью-Йорке начинают тлеть из-за возгорания частей канализационной инфраструктуры. От взрыва чугунные крышки люков весом до 300 фунтов взмывают на высоту в несколько этажей, а затем с грохотом падают, подвергая опасности окружающих.

Con Edison, коммунальная компания, которая занимается электроснабжением Нью-Йорка, из года в год проводит регулярные проверки и техобслуживание люков. Раньше специалисты в основном полагались на волю случая, надеясь, что взрывоопасными окажутся именно те люки, которые планируется проверить. Такой подход был едва ли полезнее, чем блуждание по Уолл-стрит. В 2007 году компания Con Edison обратилась к статистикам Колумбийского университета, расположенного на окраине города, в надежде, что статистические данные о сети (например, сведения о предыдущих неполадках и инфраструктурных соединениях) помогут спрогнозировать, какие люки вероятнее всего небезопасны, и это позволит компании целенаправленно использовать свои ресурсы.

Это сложная проблема, связанная с большими данными. Общая протяженность подземных кабелей в Нью-Йорке — 94 000 миль (достаточно, чтобы обхватить Землю 3,5 раза). В одном только Манхэттене около 51 000 люков и распределительных коробок. Часть этой инфраструктуры построена еще во времена Томаса Эдисона (тезки компании), а один из 20 кабелей заложен до 1930 года. Сохранились записи, которые велись с 1880 года, но не систематизированные, поскольку их не собирались анализировать. Данные предоставили бухгалтерия и диспетчеры аварийной службы, которые вручную писали «заявки на устранение неисправностей». Назвать их беспорядочными — ничего не сказать. К примеру, один лишь термин «распределительная коробка» (англ. *service box*), обозначающий обычную часть инфраструктуры, был записан в 38 вариантах, в том числе: SB, S, S/B, S.B, S?B, S.B., SBX, S/BX, SB/X, S/XB, /SBX, S.BX, S &BX, S?BX, S BX, S/B/X, S BOX, SVBX, SERV BX, SERV-BOX, SERV/BOX и SERVICE BOX. Распознать все это предстояло компьютерному алгоритму.

«Взглянув на это, мы подумали, что нам не удастся проанализировать данные, поскольку они были невероятно сырыми, — вспоминает Синтия Рудин, статистик и руководитель проекта. — У меня имелись распечатки таблиц для всех видов кабелей. Вытаскивая какие-то из них, мы не могли удержать их в руках — все тут же летело на пол. И в этом всем нужно было разобраться. Без какой-либо документации. Мне оставалось только думать, как из всего этого извлечь пользу».

Для работы Синтии Рудин и ее команде следовало использовать все данные, а не только выборку, поскольку любой из десятков тысяч люков грозил оказаться бомбой замедленного действия. Таким образом, только подход «*N* = всё» мог прийти на помощь. Совсем не мешало бы продумать причинно-следственные связи, но на это ушла бы сотня лет, притом что правильность и полнота результатов оставались бы сомнительными. Лучшим решением этой задачи было найти корреляции. Синтию интересовал не столько вопрос *почему*, сколько *что*, хоть она и осознавала, что, когда команде феноменальных специалистов по статистике придется отвечать перед руководством Con Edison, им придется обосновать свой рейтинг. Прогнозы выполнялись компьютерами, но их потребителем выступал человек. А людям, как правило, нужны причины, чтобы понять.

Интеллектуальный анализ данных обнаружил те самые «золотые самородки», которые Синтия Рудин надеялась найти. Очистив беспорядочные данные для обработки с помощью компьютера, команда определила 106 прогностических факторов основной аварии, связанной с канализационными люками. Затем из них отобрали несколько самых сильных сигналов. Проверяя электросеть Бронкса, специалисты проанализировали все имеющиеся данные вплоть до середины 2008 года. Затем на основе этих данных спрогнозировали проблемные участки с расчетом на 2009 год и получили блестящий результат: из 10% первых по списку люков 44% были связаны с серьезными происшествиями.

Основными факторами оказались возраст кабелей и наличие неполадок в люках в прошлом. Как ни странно, эти сведения были полезными, поскольку легко объясняли руководству Con Edison, на чем основан рейтинг. Но, помилуйте, возраст и неполадки в прошлом?

Разве это не достаточно очевидно? И да и нет. С одной стороны, как любил повторять математик Дункан Уоттс (в своей книге[14]), «все очевидно, когда вы уже знаете ответ». С другой стороны, важно помнить, что модель изначально содержала 106 прогностических факторов. И не так уж очевидно, как их взвесить, а затем ранжировать десятки тысяч люков, учитывая множество переменных, связанных с каждым фактором. В итоге получаются миллионы точек данных, притом что сами данные изначально непригодны для анализа.

Этот случай наглядно демонстрирует, как данные находят новое применение для решения сложных задач реального мира. Для этого понадобилось изменить подход к работе и использовать все данные, которые удалось собрать, а не только их небольшую часть. Нужно было принять естественную беспорядочность данных, а не рассматривать точность как высший приоритет. К тому же пришлось рассчитывать на корреляции, не зная полностью причин, которые легли в основу прогнозирования.

Конец теории?

Большие данные меняют наш подход к познанию мира. В эпоху малых данных мы руководствовались гипотезами о том, как устроен мир, а затем старались проверить их путем сбора и анализа данных. В дальнейшем наше понимание будет зависеть от изобилия данных, а не от гипотез. Получая и анализируя данные, мы увидим связи, о которых и не подозревали раньше.

Гипотезы часто являются продуктом теорий естественных и социальных наук, которые помогают объяснить, а иногда и спрогнозировать события окружающего мира. По мере того как мир переходит от гипотез к данным, велико искушение решить, что теории тоже больше не нужны.

В 2008 году главный редактор журнала Wired Крис Андерсон высказал мнение, что «ввиду огромного потока данных научные методы уже неактуальны». В статье «Век петабайтов» он заявил, что это означает не что иное, как «конец теории». Традиционный процесс научного открытия (проверка гипотезы на достоверность с помощью модели основополагающих причин), по утверждению Андерсона, уже

отжил свое и заменен статистическим анализом корреляций, в котором нет места теории[53].

В подтверждение Андерсон пояснил, что квантовая физика стала практически полностью теоретической областью, поскольку эксперименты слишком сложные, дорогостоящие и слишком масштабные для реализации. Эта теория, как считает Андерсон, уже не имеет ничего общего с действительностью. Чтобы объяснить новый метод, он приводит в пример поисковую систему Google и генетическое секвенирование. «Это мир, в котором большие объемы данных и прикладная математика заменяют любые другие нужные инструменты, — пишет Андерсон. — При достаточном количестве данных числа говорят сами за себя. И петабайты позволяют сказать, что корреляций вполне достаточно».

Статья вызвала оживленное обсуждение, хотя Андерсон быстро отказался от своих смелых заявлений[54]. Но его основная идея достойна внимания. По сути, он считает, что до недавнего времени в стремлении проанализировать и понять окружающий мир нам требовались теории, которые проверялись на достоверность. В эпоху больших данных, напротив, основная идея состоит в том, что нам больше не нужны теории — достаточно взглянуть на данные. Предполагается, что все обобщенные правила (о том, как устроен мир, как ведут себя люди, что покупают потребители, как часто ломаются детали и т. д.) могут утратить свою актуальность, когда в ход идет анализ больших данных.

«Конец теории» позволяет предположить: несмотря на то что предметные области, такие как физика и химия, полны теорий, анализ больших данных не нуждается в каких-либо концептуальных моделях. Но это абсурд!

Большие данные имеют теоретическую основу. При анализе больших данных используются статистические и математические теории, а иногда и теоретические знания из области компьютерных наук. Да, это не теории о причинной динамике того или иного явления (например, гравитации), но все же теории! И, как было показано ранее, модели на основе этих теорий, лежащих в основе анализа больших данных, открывают полезные возможности прогнозирования. На самом деле анализ больших данных может предложить свежий взгляд

и новые идеи именно потому, что не обременен рамками традиционного мышления и присущими ему предубеждениями, которые неявно представлены в теориях конкретной области.

Поскольку анализ больших данных основан на теориях, эту основу невозможно игнорировать — более того, нужно признать, что она тоже влияет на результат. Все начинается с того, как мы выбираем данные. Их сбор может быть обусловлен удобством (доступны ли данные) или экономией (можно ли получить данные по дешевке). Наш выбор в данном случае зависит от теорий. Как полагают Дана Бойд[15] и Кейт Кроуфорд[16], наши находки зависят от того, что мы выбираем. В конце концов, специалисты Google использовали в качестве закономерности условия поиска, связанные с гриппом, а не с размерами обуви. Точно так же, анализируя данные, мы выбираем инструменты, которые опираются на теории. Наконец, интерпретируя результаты, мы снова применяем теоретические знания. Эпоха больших данных отнюдь не лишена теорий — они повсюду, со всеми вытекающими последствиями.

Большие данные не предрекают «конец теории», но принципиально меняют наше представление об окружающем мире. Обществу предстоит еще ко многому привыкнуть ввиду этих изменений. Многие учреждения столкнутся с новыми трудностями. Но огромные преимущества, которые мы получим, делают такой компромисс не только целесообразным, но и неизбежным. При этом следует отметить, как это произойдет. Большинство специалистов в области высоких технологий, поскольку сами занимаются их созданием, сказали бы, что все дело в новых инструментах — от быстрых чипов до эффективного программного обеспечения. Однако эти инструменты не настолько важны, как можно подумать. Более глубокая причина сложившихся тенденций лежит в том, что у нас появилось намного больше данных, так как стало фиксироваться больше факторов действительности. Об этом — в следующей главе.

Примечания к главе 4

Глава 5

Датификация

Мори Мэтью Фонтейн был многообещающим офицером военно-морского флота США. Получив новое назначение, в 1839 году он направился на бриг Consort. Его дилижанс внезапно съехал с дороги, опрокинулся, и Мори вылетел наружу. Жестко приземлившись, он сломал бедренную кость и вывихнул колено. Местный врач вправил ему коленный сустав, но бедренная кость срослась неправильно, и через несколько дней ее потребовалось повторно ломать. Из-за травм 33-летний Мори начал прихрамывать и стал непригоден к морской службе. Спустя почти три года, потраченных на оздоровление, он был назначен на офисную службу в ВМФ в качестве руководителя отдела со скучным названием «Депо карт и приборов».

И эта должность подошла ему как нельзя лучше! Будучи молодым штурманом, Мори задавался вопросом, почему корабли движутся по водной глади зигзагообразно, а не по прямой. Задавая этот вопрос капитанам, он слышал в ответ, что намного лучше держаться знакомого курса, чем рисковать и идти по малознакомому, который таит в себе скрытые опасности. Океан считался непредсказуемым царством, полным неожиданностей, волн и порывов ветра.

Имея опыт путешествий, Мори знал, что это не совсем так. Он во всем искал систему. Находясь в просторном порту в Вальпараисо (Чили), он заметил, что ветры дуют с точностью часов. Вечерний сильный ветер резко затихал на закате и сменялся легким бризом, будто кто-то щелкнул выключателем. Во время другого рейса Мори пересек теплые ярко-синие воды Гольфстрима, которые текут между темными стенами морских вод Атлантики по одному и тому же пути, словно река Миссисипи. Португальцы и вправду веками плавали по Атлантике, опираясь на постоянные восточные и западные ветры — пассаты (от древнеангл. «путь» или «курс», который стал ассоциироваться с торговлей).

Всякий раз, оказываясь в новом порту, мичман Мори отправлялся на поиски старых морских капитанов, чтобы перенять знания, основанные на опыте, который передавался из поколения в поколение. Так он узнал о приливах, ветрах и морских течениях, действующих с определенной закономерностью, о которой не прочтешь ни в одной книге и которой не увидишь ни на одной карте, что выпускались для моряков военно-морским флотом. Вместо этого в ВМФ полагались на карты порой столетней давности, многие из которых содержали значительные упущения или откровенные неточности. Занимая новую должность начальника депо карт и приборов, Мори стремился исправить это положение.

Со вступлением на пост он пополнил депо барометрами, компасами, секстантами и хронометрами. Он обратил внимание на множество хранившихся здесь книг по морскому делу, карт и схем. Среди материалов были заплесневелые ящики, забитые старыми журналами со всех прошлых плаваний капитанов ВМС. Предшественники рассматривали их как мусор, но Мори отряхнул пыль с покрытых пятнами морской соли книг и заглянул внутрь. Увиденное не оставило его равнодушным.

Здесь была как раз нужная информация: записи о ветре, водах и погоде в определенных точках, расписанные по датам. Некоторые из них были не слишком ценными, зато множество других изобиловали полезной информацией. Сведя их, Мори понял, что можно создать совершенно новую форму навигационной карты. Журналы были бессистемными. С чудаковатыми стишками и набросками на полях, они порой казались попыткой спастись от скуки в пути. Но были и сведения, которые пригодились. При помощи десятков «расчетчиков» (так назывались те, кто занимался расчетом данных) Мори начал трудоемкий процесс сведения информации, которая хранилась в истрепанных журналах.

Мори объединил данные и разделил всю Атлантику на блоки по пять градусов долготы и широты. Он отметил температуру, скорость и направление ветра и волн, а также соответствующий месяц, поскольку тенденции разнились в зависимости от времени года. Объединенные данные показывали определенные тенденции и указали более удачные маршруты.

Из поколения в поколение моряки передавали советы отправлять суда то в спокойные воды, то навстречу встречным ветрам и течениям. На одном из распространенных маршрутов — из Нью-Йорка в Рио-де-Жанейро — моряки, как правило, боролись со стихией, а не союзничали с ней. Американских шкиперов учили избегать опасных плаваний вдоль пролива к югу от Рио, поэтому суда легко скользили по юго-восточному курсу, а по пересечении экватора меняли его на юго-западный. Пройденное расстояние равнялось двум маршрутам через всю Атлантику. Как оказалось, в этом не было необходимости: они могли спокойно придерживаться прямого курса на юг.

Для большей точности Мори нужна была дополнительная информация. Он создал стандартную форму для регистрации данных судов и обязал все суда военно-морского флота США заполнять ее и сдавать по возвращении. Поскольку капитаны торговых судов жаждали получить его карты, Мори настоял, чтобы взамен они пустили в оборот свои журналы (тем самым образовав раннюю версию вирусной социальной сети). Мори объявил, что «каждое судно, которое выходит в открытое море, отныне может рассматриваться как плавающая обсерватория, храм науки». Для уточнения карт он искал другие точки данных (так же на основе алгоритма вычисления рейтингов веб-страниц PageRank была создана система Google, учитывающая больше сигналов). Мори поручил капитанам периодически бросать в море бутылки с записками, в которых указывать день, должность, преобладающие ветра и течения, а также вылавливать все бутылки, которые встречаются им на пути. Многие корабли вывешивали специальный флаг, чтобы показать, что они участвуют в обмене информацией (предвестники значков-ссылок «поделиться», которые отображаются на некоторых веб-страницах).

На основе данных сами собой вырисовывались естественные морские пути, где ветры и течения были особенно благоприятными. Карты Мори, как правило, сокращали долгое путешествие на треть, обеспечивая купцам значительную экономию. «Пока я не взял на вооружение ваш труд, я пересекал океан с завязанными глазами», — с благодарностью писал один из капитанов. Даже бывалые моряки, которые отвергали новомодные карты и полагались на старые пути, выполняли полезную функцию: если на их путешествие уходило

больше времени или они попадали в беду, это служило лишним доказательством в пользу системы Мори. К моменту публикации своей магистерской работы «Физическая география моря» в 1855 году Мори успел определить координаты 1,2 миллиона точек данных. «Таким образом, молодой моряк, вместо того чтобы брести на ощупь вперед, пока не наберется опыта... здесь сразу нашел бы рекомендации, основанные на опыте тысяч штурманов»[55], — писал Мори.

Его работа имела огромное значение для закладки первого трансатлантического телеграфного кабеля. А после трагического столкновения в открытом море он быстро разработал системы судоходных путей, которые используются по сей день. Он даже применил свой метод к астрономии: с открытием планеты Нептун в 1846 году Мори выдвинул прекрасную идею пересмотреть все архивные записи, где планета ошибочно упоминается как звезда, что позволило установить ее орбиту.

Выходец из Вирджинии, Мори редко упоминается в источниках американской истории. Возможно, это потому, что он ушел из флота во время Гражданской войны в США и служил шпионом в Англии на благо Конфедерации. Но несколькими годами ранее, прибыв в Европу, чтобы заручиться международной поддержкой для своих карт, в четырех странах Мори был посвящен в рыцари, а еще в восьми — награжден золотыми медалями, включая награду Святого Престола. И теперь лоцманские карты, изданные военно-морским флотом США, носят его имя.

Коммодор[17] Мори одним из первых осознал основополагающий принцип больших данных: огромный корпус данных обладает особой ценностью, которой нет в меньших количествах. Более того, он понял, что заплесневелые журналы ВМФ на самом деле представляют собой «данные», если из них извлечь и свести в таблицы соответствующую информацию. При этом он впервые использовал данные, в частности те сведения, которые никому не представлялись ценными, повторно. Подобно Орену Эциони из Farecast, который с помощью старых сведений о ценах в авиационной отрасли создал прибыльный бизнес, или инженерам Google, применившим старые поисковые запросы, чтобы понять распространение вспышек гриппа, Мори взял

целенаправленно созданную информацию (сведения о местоположении для безопасного путешествия) и преобразовал ее.

Его метод, в целом аналогичный современным методам работы с большими данными, был поразительным, учитывая, что Мори реализовывал его с помощью карандаша и бумаги. Это значит, что использование данных появилось намного раньше оцифровки. Сегодня мы часто объединяем эти понятия. Однако важно их различать. Уяснить, как данные получают из самых неожиданных областей, нам поможет более современный пример.

Сигеоми Косимицу, профессор Института передовых промышленных технологий в Токио, сумел извлечь данные из параметров, соотнесенных с ягодицами. Мало кому придет в голову, что сидячие позы несут в себе информацию, но это так. Контуры тела, позу и распределение веса сидящего человека можно оценить количественно и свести полученные цифры в таблицу. С помощью датчиков, размещенных в 360 разных точках сиденья автомобиля, Косимицу и группа инженеров снимают показатели давления, которое оказывают ягодицы водителя, оценивая каждую точку по шкале от 0 до 256 баллов. Получается цифровой код, уникальный для каждого человека. В ходе судебного разбирательства эта система способна отличить одного человека от другого с точностью до 98%.

Это исследование проводится не ради забавы. Технологию планируется использовать в качестве противоугонной системы автомобилей. Оборудованный такой системой автомобиль способен распознать «чужака» за рулем и потребовать пароль для запуска двигателя. Преобразование поз в данные представляет собой практичную услугу населению и потенциально прибыльный бизнес. Объединение данных может выявить связь между позой водителя и безопасностью на дорогах, например зафиксировать изменение позы перед дорожно-транспортным происшествием. Система способна также «почувствовать» замедление реакции из-за утомления и послать сигнал тревоги или автоматически нажать на тормоза. Она может не только обнаружить, что автомобиль украден, но и определить вора, так сказать, «со спины».

Профессор Косимицу обратился к материалу, который никогда не рассматривался как данные (вряд ли кому вообще пришло бы в голову,

что он обладает информационными качествами), и преобразовал его в цифровой, количественный формат. Таким же образом коммодор Мори взял материал, который казался практически бесполезным, и получил из него информацию, превратив его в поистине полезные данные. Это позволило использовать информацию по-новому и придало ей уникальную ценность.

Слово data (англ. *данные*) в переводе с латинского означает «данность», то есть «факт». Это понятие стало краеугольным камнем классического труда Евклида, в котором геометрия объясняется с точки зрения известных данных и таких, которые можно показать, чтобы сделать известными. Сегодня данные относят к некоторому процессу, который позволяет их записывать, анализировать и переупорядочивать. Пока не придуман подходящий термин для обозначения такого рода преобразований, которые выполняли коммодор Мори и профессор Косимицу. Назовем их *датификацией*, под которой подразумевается процесс представления явлений в количественном формате для дальнейшего сведения в таблицу и анализа.

Датификация — далеко не то же самое, что оцифровка, при которой аналоговая информация преобразуется в двоичный код (или последовательность единиц и нулей), считываемый компьютером. Оцифровка не являлась первичной функцией компьютеров. Эпоха компьютерной революции изначально была связана с вычислениями, как и предполагает этимология слова compute (англ. «вычислять»). Мы выполняли вычисления, которые занимали много времени (такие, как вычисления в таблицах траекторий ракет, расчеты для переписей и сведений о погоде). И лишь затем появилась оцифровка аналогового контента. Поэтому, когда Николас Негропонте из MIT Media Lab опубликовал свою эпохальную книгу Being Digital в 1995 году, одной из поднятых им тем был переход от атомов к битам. К началу 1990-х годов этот переход в значительной степени коснулся текстовых данных. По мере увеличения емкости хранилищ, процессоров и пропускной способности за последнее десятилетие это удалось сделать и с другими формами контента (изображениями, видео, музыкой и пр.).

Сегодня среди технологов негласно принято считать, что большие данные ведут свое начало с момента «кремниевой» революции. Но это не так. Безусловно, большие данные стали возможны благодаря современным ИТ-системам, но основная идея лишь продолжила древнейшие поиски человечества в области измерения, записи и анализа мира[56]. ИТ-революция, произошедшая в мире, очевидна. Основной акцент в ней приходился на «Т» — технологии. Пришло время переключиться на «И» — информацию.

Для того чтобы записывать информацию в количественной форме (датифицировать ее), нам нужно знать, как проводить измерения и записывать полученный результат. А для этого необходим правильный набор инструментов, а также желание количественно измерять и записывать. И то и другое — предпосылки датификации, и человечество разработало ее «строительные элементы» задолго до начала цифровой эпохи.

Мир, выраженный в количественных категориях

Возможность записи информации — одно из главных различий между примитивными и передовыми обществами. Основы счета, а также измерение длины и веса были древнейшими инструментами ранних цивилизаций. К началу III тысячелетия до н. э. идея записи информации значительно продвинулась вперед. Это произошло в долине Инда, Египте и Месопотамии. Повысилась точность измерений, да и сами они прочно вошли в повседневную жизнь. Эволюция письменности в Месопотамии обеспечила точный метод отслеживания производства и деловых операций. Это позволило ранним цивилизациям измерять окружающие объекты и явления, делать записи о них и извлекать их позднее. Измерение и запись способствовали созданию данных. Они же являются древнейшими основами датификации.

Так стало возможным воспроизводить продукты человеческой деятельности, например здания, записывая их размеры и строительные материалы. При этом можно было экспериментировать, изменяя отдельные размеры, чтобы создать нечто новое, что затем тоже подлежало бы записи. Можно было записывать коммерческие сделки,

чтобы знать, сколько урожая удалось собрать с поля (и сколько из него уйдет государству в виде налога). Появилась возможность прогнозирования и планирования, даже если они заключались в простом предположении, что следующий год будет таким же урожайным, как и текущий. Благодаря этому деловые партнеры могли отслеживать, сколько они должны друг другу. Без измерения и записей не появились бы деньги, поскольку не было бы данных для их обоснования.

Спустя столетия область применения измерений расширилась от длины и веса до площади, объема и времени. К началу I тысячелетия основные функции измерений узнал Запад. Существенным недостатком способа измерения в ранних цивилизациях являлось то, что он не был оптимизирован для вычислений, даже относительно простых. Система счета римских цифр малопригодна для численного анализа. Без позиционной системы нумерации из десяти основных цифр и десятичных чисел даже лучшим специалистам трудно давались умножение и деление больших чисел, а большинству остальных не хватало прозрачности даже в простом сложении и вычитании[57].

В Индии альтернативная система счисления появилась примерно в I веке. Она перекочевала в Персию, где была усовершенствована, а затем принята арабами, которые тоже значительно ее улучшили. Эта система стала основой арабских цифр, которыми мы пользуемся до сих пор. Крестовые походы, может, и несли абсолютное разрушение землям, на которые вторгались европейцы, но при этом знания мигрировали с востока на запад, и, пожалуй, самым значительным иноземным нововведением стали арабские цифры. Папа Сильвестр II, который занимался их изучением, выступил за их использование в конце первого тысячелетия. К началу XII века арабские тексты, описывающие данную систему, были переведены на латынь и распространились по всей Европе, дав начало математике.

Еще до того, как в Европе появились арабские цифры, вычислительный процесс улучшило использование счетных досок. На этих досках делались гладкие желобки, в которых размещались счетные метки для обозначения сумм. Складывали и вычитали, перемещая метки в определенных областях. Такой способ имел

значительные ограничения: было трудно одновременно рассчитывать очень большие и очень маленькие количества. А самое главное — недолговечность цифр на этих досках. Неверный шаг, небрежный удар — и цифра могла измениться, что приводило к неправильным результатам. Счетные доски годились для расчетов, но не для записи. Поэтому всякий раз, когда числа с доски необходимо было записать, их переводили обратно в неудобные римские цифры[58]. (Европейцы так и не переняли восточный способ подсчета с помощью абака[18], но это оказалось к лучшему, так как не дало увековечить на Западе использование римских цифр[59].)

Математика придала данным новый смысл: теперь их можно было *анализировать*, а не только записывать и при необходимости извлекать. Прошли сотни лет с момента введения арабских цифр (XII век) до их широкого распространения (конец XVI века). К началу XVI века математики уже гордились тем, что с помощью арабских цифр проводили расчеты в шесть раз быстрее, чем с помощью счетных досок. Окончательный успех арабским цифрам принесла эволюция еще одного инструмента датификации — двойной бухгалтерии.

Счетоводы изобрели письменность в III тысячелетии до н. э. Несмотря на развитие счетоводства в последующих столетиях, оно, по сути, оставалось централизованной системой учета конкретных сделок. Но так и не удалось реализовать механизм, благодаря которому счетоводы и их торговцы-работодатели могли бы в любой момент времени увидеть то, что интересовало их больше всего: является конкретный счет или целая компания прибыльной или нет. Ситуация изменилась в XIV веке, когда счетоводы Италии начали записывать операции одновременно в двух книгах. Изящество этой системы заключалось в том, что прибыль и убытки можно было легко свести в таблицы по каждому счету, просто добавив кредиты и дебеты. И «скучные» данные вдруг «заговорили», пусть даже сбивчиво и только в пределах выявления прибыли и убытков.

Сегодня двойная бухгалтерия, как правило, рассматривается только с точки зрения ее последствий для бухгалтерского учета и финансов. Однако она стала вехой в эволюции использования данных, так как позволила записывать информацию в виде «категорий», связывающих

счета между собой. Она работала по принятым правилам записи данных, став одним из самых ранних примеров стандартизированной системы записи информации. Бухгалтеры могли с легкостью разобраться в записях друг друга. Бухгалтерия была организована таким образом, чтобы сделать определенный тип запроса данных (расчет прибыли или убытков по каждому счету) быстрым и простым. Наконец, она предусматривала аудиторский след операций для более удобного прослеживания данных. Двойная бухгалтерия разрабатывалась с учетом встроенной «системы исправления ошибок», которая и сегодня не оставила бы равнодушными любителей технологий. Если запись в одной части бухгалтерской книги вызывала сомнения, можно было проверить соответствующую ей запись в другой.

Как и арабские цифры, двойная бухгалтерия не сразу стала успешной. Лишь спустя двести лет с момента изобретения этого метода вмешательство математика и купеческой семьи, наконец, изменило историю датификации.

Математик — это францисканский монах Лука Пачоли. В 1494 году он опубликовал учебник по коммерческой математике, рассчитанный на непрофессионалов в этой области. Благодаря своей популярности книга, по сути, являлась в то время учебником по математике. Кроме того, она стала первой книгой, полностью построенной на арабских цифрах, тем самым способствуя их укоренению в Европе. Наиболее долгосрочным вкладом была часть книги, посвященная бухгалтерии, где Пачоли четко объяснял систему двойного бухгалтерского учета. В течение последующих десятилетий часть, посвященную бухгалтерскому учету, отдельно издали на шести языках, и веками она оставалась настольной книгой по этому предмету.

Что касается купеческой семьи, это были знаменитые венецианские торговцы и меценаты — Медичи. В XVI веке они стали самыми влиятельными банкирами в Европе, в значительной степени благодаря тому, что использовали улучшенный способ записи данных — систему двойной записи. Учебник Пачоли и успех Медичи в его применении утвердили победу двойной бухгалтерии в качестве стандартной записи данных и с того момента закрепили использование арабских цифр.

Параллельно с достижениями в области записи данных развивалась идея измерения окружающего мира, которая подразумевала обозначения времени, расстояния, площади, объема и веса. Стремление познать природу через количественные категории определило развитие науки в XIX веке: ученые изобрели новые инструменты и агрегаты для измерения и регистрации электрических токов, атмосферного давления, температуры, частоты звука и т. п. Это была эпоха всеобщего определения, разграничения и обозначения. Увлечение этими процессами дошло до измерения черепа человека и его умственных способностей для выявления закономерностей между ними. К счастью, эта лженаука («френология») уже практически исчезла. Но желание все количественно измерить только усилилось.

Измерение объектов и явлений реального мира, а также запись получаемых данных процветали благодаря сочетанию подходящих инструментов и восприимчивого мышления. На этой благодатной почве и выросла датификация в ее современном понимании. Все составляющие датификации были готовы к использованию, однако в аналоговом мире этот процесс все еще оставался трудоемким и дорогостоящим. В большинстве случаев требовалось обладать бесконечным терпением или же посвятить этому делу всю жизнь. Примером тому служат тщательные ночные наблюдения за небесными телами, которые проводил астроном Тихо Браге[19] в 1500-х годах. В аналоговую эпоху случаи удачной датификации были редкостью. Как правило, им способствовало счастливое стечение обстоятельств (как в истории коммодора Мори, который был вынужден заниматься офисной работой, но имел в своем распоряжении целый склад журналов). Всякий раз результатом датификации исходной информации оказывались огромная ценность и потрясающие открытия.

Появление компьютеров повлекло за собой внедрение цифровых устройств для измерения и хранения данных, которые значительно повысили эффективность датификации, а также сделали возможным математический анализ данных для раскрытия их скрытой ценности. Проще говоря, оцифровка стала катализатором датификации, но никак не ее заменой. Процесс оцифровки (преобразование аналоговой

информации в формат, считываемый компьютером) сам по себе не является датификацией.

Когда слова становятся данными

Разница между оцифровкой и датификацией данных станет очевидной, если посмотреть на домен, где происходит и то и другое, и сравнить последствия. Рассмотрим такой пример. В 2004 году компания Google объявила невероятно смелый план — полностью оцифровать все книги, которые находятся в ее распоряжении (насколько это возможно с учетом законов об авторском праве), и дать возможность людям по всему миру искать и бесплатно просматривать книги через интернет. Чтобы совершить этот подвиг, компания объединилась с несколькими крупнейшими и наиболее престижными научными библиотеками мира и разработала машины для сканирования, которые могли бы автоматически перелистывать страницы, делая сканирование миллионов книг не только реализуемым, но и финансово жизнеспособным.

Первый текст, *оцифрованный* компанией Google, выглядел так. Каждую страницу отсканировали и записали в виде файла цифрового изображения в высоком разрешении, сохраненного на серверах Google. Страницы были преобразованы в цифровые копии, которые любой мог легко получить через интернет из любой точки мира. Однако при этом требовалось точно знать, какая книга содержит нужную информацию, иначе приходилось много читать, чтобы найти правильный отрывок. Текст невозможно было найти по словам или анализировать, поскольку его не датифицировали. Все, чем располагала Google, — это изображения, которые только люди могли превратить в полезную информацию.

И хотя это все равно было отличным инструментом — современной цифровой Александрийской библиотекой, более полезной, чем любая другая библиотека за всю историю, — Google этого показалось мало. Компания понимала, что эта информация хранила в себе ценнейший ресурс, который можно получить только в результате датификации. Поэтому специалисты Google пустили в ход программу оптического распознавания символов, которая могла распознать буквы, слова,

предложения и абзацы в цифровом изображении. В итоге получался датифицированный текст, а не оцифрованная картинка страницы.

Теперь информация со страниц была доступна не только для чтения, но и для обработки на компьютерах и для анализа с помощью алгоритмов. Благодаря этому текст становился индексируемым, а значит, доступным для поиска. Стал возможным бесконечный поток текстового анализа. Так, например, можно узнать дату первого упоминания определенных слов и фраз или выяснить, когда они стали популярными. Это позволяет нам по-новому взглянуть на распространение идей и развитие человеческого мышления на протяжении столетий и на многих языках.

Попробуйте сами. Служба Google NgramViewer (http://books.google.com/ngrams) создает график использования слов или фраз с течением времени, применяя в качестве источника данных весь перечень книг Google. Всего за несколько секунд мы можем обнаружить, что до 1900 года термин «причинность» (англ. causality) использовался чаще, чем «корреляция» (англ. correlation), но затем соотношение изменилось. Мы можем сравнить стили письма и понять, кто прав в спорах об авторстве. Кроме того, благодаря датификации стало гораздо легче обнаруживать плагиат в научных трудах, вследствие чего некоторые европейские политики, в том числе министр обороны Германии, были вынуждены уйти в отставку.

По оценкам, с момента изобретения печатного станка (середина XV века) опубликовано 129 миллионов различных книг. К 2010 году, пять лет спустя после запуска своего книжного проекта, компании Google удалось отсканировать более 15 миллионов наименований — существенную часть письменного наследия мира (более 12%). Это дало начало новой учебной дисциплине — «культуромике». Она представляет собой вычислительную лексикологию, которая пытается понять поведение человека и культурные тенденции путем количественного анализа текстов.

В ходе одного из исследований гарвардские ученые, обработав миллионы книг и более 500 миллиардов слов, выявили, что менее половины английских слов, которые встречаются в книгах, включены в словари. Они писали, что английский лексикон «состоит из лексической “темной материи”, которая не зафиксирована в

стандартных справочных источниках». Проведя алгоритмический анализ упоминаний о еврейском художнике времен нацистской Германии Марке Шагале, они могли бы показать, что подавление или цензура идеи, как и человека, оставляет «отпечатки, которые можно измерить количественно». Слова на страницах — словно окаменелости в осадочных горных породах, до которых приверженцы культуромики могут докопаться, словно археологи. Конечно, это влечет за собой огромное количество неявных предубеждений: отражают ли библиотечные книги истинное положение вещей в мире или показывают только то, что дорого авторам и библиотекарям? И все же культуромика дает интересные результаты.

Преобразование слов в данные открывает множество способов их применения. Конечно, их можно читать традиционным способом или анализировать с помощью компьютера. Но для Google как для образцовой компании, которая занимается обработкой больших данных, не было секретом, что информация имеет несколько потенциальных назначений, вполне оправдывающих ее сбор и датификацию. Так, например, с помощью датифицированного текста Google удалось улучшить свою службу машинного перевода. Как говорилось в третьей главе, система определяла отсканированные переводные книги и анализировала, какие слова и фразы на одном языке соответствуют словам и фразам на другом. Зная это, система обрабатывала перевод как огромную математическую задачу, в которой компьютер выясняет вероятности, чтобы определить наилучшие соответствия слов в разных языках.

Переход от цифровых изображений страниц к датифицированному тексту чреват ошибками. Даже очень сложные программы распознавания символов сталкиваются с трудностями из-за чрезвычайного разнообразия шрифтов, опечаток в тексте и выцветших чернил. Для слов, которые до сих не поддаются расшифровке с помощью специальных программ, компания Google поставила себе на службу хитрый способ получать непреднамеренную помощь от интернет-пользователей (об этом подробнее рассказано в следующей главе).

Конечно, Google не единственная компания, которая мечтала перенести богатое письменное наследие мира в эпоху компьютеров.

Она далеко не первая решила попробовать это осуществить. Проект «Гутенберг» (общественная инициатива по размещению различных произведений в интернете для общего пользования) был призван сделать тексты доступными людям исключительно для чтения. При этом не предусматривались дополнительные способы использования слов (в качестве данных), то есть не шла речь о повторном использовании. Подобным образом издатели в течение многих лет экспериментировали с электронными версиями книг. Но они тоже видели основную ценность книг в их содержании, а не в данных. На этом строилась их бизнес-модель. Издатели никогда не обращали внимания на данные, присущие тексту книги, и не позволяли этого другим. Они не видели в этом необходимости и попросту недооценивали потенциал данных.

Многие компании сейчас соперничают за успех на рынке электронных книг. Похоже, в этой области с большим отрывом лидирует компания Amazon с ассортиментом своих электронных книг Kindle. Однако стратегии компаний Amazon и Google в этой области значительно разнятся.

Компания Amazon получила в свое распоряжение датифицированные книги, однако не сумела найти новые способы применения текста в качестве данных. Джефф Безос, основатель и главный исполнительный директор компании, убедил сотни издателей выпустить книги в формате Kindle. Книги Kindle представляют собой не изображения страниц (в противном случае никто бы не смог изменить размер шрифта или отобразить страницы как на цветных, так и на черно-белых экранах) — их текст датифицирован, а не просто оцифрован. Компании Amazon удалось совершить с миллионами новых книг то, что Google усердно старается повторить с множеством старых.

Тем не менее книжный бизнес Amazon завязан на содержимом, которое читают, а не на анализе датифицированного текста. Справедливости ради стоит заметить, что компания наверняка сталкивается с ограничениями, которые консервативные издатели накладывают на использование информации, содержащейся в их книгах. В свою очередь компания Google, как хулиганка в области больших данных, стремящаяся выйти за рамки, конечно, не

испытывает таких ограничений — хлеб насущный ей обеспечивают клики пользователей, а не доступ к собственности издателей. Однако, не считая замечательной службы «статистически значимых слов» Amazon, которая использует алгоритмы для выявления неочевидных связей между темами книг, этот интернет-магазин так и не распорядился своей сокровищницей слов для анализа больших данных. Пожалуй, будет справедливо отметить, что, по крайней мере сейчас, Amazon осознает ценность оцифровки контента, а Google — ценность его датификации.

Когда местоположение становится данными

Один из самых весомых источников информации в мире, по сути, сам мир. Б*о*льшую часть истории человечества он не измерялся количественно и не использовался в форме данных. Безусловно, информацию представляет собой географическое положение объектов и людей: гора находится там, человек — тут. Но эту информацию необходимо преобразовать в данные. Для датификации местоположения требуется несколько составляющих: метод измерения площади земного шара вплоть до сантиметра, стандартизированный способ обозначения и инструмент для сбора и записи данных. Территория, координаты, инструменты. Определение количества, стандартизация, сбор. Только тогда мы сможем хранить и анализировать местоположение не как место само по себе, а как данные.

На Западе количественное измерение местоположения придумали греки. Около 200 года до н. э. Эратосфен изобрел систему координат (сродни широте и долготе) для демаркации местоположений. Со временем она утратила практическое применение, как и множество других хороших идей эпохи Античности. Полтора с половиной тысячелетия спустя (около 1400 года) копия птолемеевского труда «Руководство по географии» прибыла во Флоренцию из Константинополя ввиду того, что эпоха Возрождения и морская торговля возбудили живой интерес к науке и древним знаниям. Это стало сенсацией, и старые уроки Птолемея пригодились для решения современных задач в области навигации. С тех пор на картах

появились долгота, широта и масштаб. Позже систему улучшил фламандский картограф Герард Меркатор (в 1570 году), что позволило морякам выстраивать прямые маршруты в круглом мире.

Хотя к этому времени уже сформировался способ записи информации о местоположении, не существовал общепринятый формат для обмена ею. Требовалась единая система идентификации, так же как в интернете требуются доменные имена для работы электронной почты и других служб. Стандартизация долготы и широты заняла много времени и была, наконец, закреплена в 1884 году на Международной меридианной конференции в Вашингтоне (Колумбия), где 25 стран выбрали Гринвич (Англия) в качестве нулевого меридиана и нулевой долготы, и только Франция, считая себя лидером в международных стандартах, воздержалась от голосования. В 1940 году создана система координат «Универсальная поперечная проекция Меркатора» (UTM), согласно которой земной шар разделили на 60 зон для повышения точности.

Геопространственное положение теперь определяли, записывали, подсчитывали, анализировали и распространяли в стандартизированном числовом формате. Появилась возможность датифицировать положение. Однако из-за высокой себестоимости измерение и запись информации в аналоговом виде применялись редко. Изменить ситуацию могли инструменты для менее затратного измерения местоположения. До 1970-х годов единственным способом определения физического местоположения было использование ориентиров, астрономических созвездий, счисления пути и ограниченной технологии определения координат источника радиоизлучения.

Все изменилось в 1978 году после запуска первого из 24 спутников в рамках глобальной системы определения местоположения (GPS). Приемники на Земле, будь то автомобильная навигационная система или смартфон, триангулируют свое положение, отмечая разницу во времени, которое требуется для приема сигнала от спутников, расположенных на высоте более 20 000 км. В 1980-х годах систему впервые открыли для использования в гражданских целях, а в 1990-х она заработала в полную силу. Десятилетием позже ее точность была повышена в коммерческих целях. Система GPS воплотила

древнейшую мечту мореплавателей, картографов и математиков, предоставив технические средства для быстрого, относительно дешевого и не требующего специальных знаний измерения местоположения с точностью до одного метра.

Информацию нужно создавать. Ничто не мешало Эратосфену или Меркатору определять свое местоположение ежеминутно, будь у них такое желание, хотя на практике это вряд ли удалось бы осуществить. Первые приемники GPS ввиду сложности и дороговизны не были общедоступными и годились, скорее, для специальных нужд (например, для подводной лодки). Ситуацию изменили недорогие чипы, встроенные в цифровые устройства. Стоимость модуля GPS упала с сотни долларов в 1990-х годах до примерно доллара при нынешнем крупномасштабном производстве. Системе GPS нужно всего несколько секунд, чтобы определить местоположение и выдать координаты в стандартизированном формате. Так, запись 37° 14' 06" N 115° 48' 40" W означает, что вы находитесь на суперсекретной американской военной базе в отдаленной части штата Невада — «Зоне-51», где (возможно) находятся космические пришельцы.

В наше время GPS — одна из множества систем, предоставляющих данные о местоположении. В Китае и Европе реализуются конкурирующие спутниковые системы. А поскольку GPS не работает в помещении или среди высотных зданий, для определения положения на основе силы сигнала можно использовать триангуляцию между базовыми станциями сотовой связи или маршрутизаторами Wi-Fi-сети. За счет этого можно достичь еще большей точности данных о местоположении. Становится понятным, почему такие компании, как Google, Apple и Microsoft, создали собственные геолокационные системы, использующие преимущества GPS. Автомобилям Street View компании Google, делающим панорамные фотографии улиц, даже удалось собрать информацию о маршрутизаторах Wi-Fi-сети, а iPhone оказался «шпионским» смартфоном, который собирал данные о местоположении и Wi-Fi-сетях и отправлял их в компанию Apple без ведома пользователей (кроме того, аналогичные данные собирали телефоны Google Android, а также мобильная операционная система Microsoft)[60].

Теперь можно отслеживать не только людей, но и любые другие объекты. Благодаря беспроводным модулям, помещаемым в транспортные средства, датификация местоположения произвела революцию в области страхования. Данные позволяют подробно изучить время, маршрут и пройденное автомобилем расстояние, чтобы лучше оценить риски. В Великобритании водители могут приобрести страховку на автомобиль, исходя из времени и маршрута фактических поездок, а не только из годового показателя, вычисляемого на основе возраста, пола и последней записи. Такой подход к ценообразованию страховых услуг стимулирует примерное поведение. При этом изменяется сама природа страхования: происходит переход от учета объединенных рисков к рискам, основанным на действиях отдельных лиц. Отслеживание физических лиц по транспортным средствам также преобразует характер постоянных затрат, например на дороги и другие объекты инфраструктуры, связывая использование того или иного ресурса с водителями и другими субъектами. Все это было невозможно до того, как появился способ постоянного получения данных о географическом положении людей и объектов. Но это то, к чему мы идем.

Компания UPS использует «геолокационные» данные несколькими способами. Ее автомобили оснащены датчиками, модулями беспроводной связи и GPS, так что в случае задержек специалисты в главном офисе могут определить местоположение фургонов или спрогнозировать неисправности двигателя. Далее, это позволяет компании отслеживать работу сотрудников и изучать карту их маршрутов для дальнейшей оптимизации. Наиболее эффективный путь определяется, в частности, по данным предыдущих поставок, подобно тому как Мори составлял карты на основе более ранних морских плаваний.

По словам Джека Ливиса, начальника отдела управления процессами в компании UPS, программа аналитики дала колоссальные результаты. В 2011 году компании удалось сократить протяженность маршрутов на 30 миллионов миль, тем самым сэкономив три миллиона галлонов топлива и сократив выбросы углекислого газа на 30 тысяч тонн. Кроме того, повысилась безопасность и эффективность, поскольку алгоритм составляет маршруты с меньшим количеством

поворотов влево. Такие повороты нередко приводят к ДТП из-за того, что автомобилю приходится пересекать движение на перекрестках, к тому же они отнимают время и потребляют больше топлива, так как перед поворотом двигатель фургона работает на холостом ходу. Телеметрическая система позволяет предвидеть поломку деталей двигателя — прямо как Кэролин Макгрегор в Университете провинции Онтарио заблаговременно определяет заболевания у недоношенных детей, о чем шла речь в четвертой главе.

«Прогнозирование дало нам знание, — говорит Дж. Ливис из UPS и с уверенностью добавляет: — Но кроме знания есть еще кое-что — мудрость и прозорливость. В какой-то момент система станет настолько умной, что будет предсказывать проблемы и исправлять их раньше, чем пользователь успеет сообразить, что что-то не так».

Со временем широкое применение получила датификация местоположения людей. В течение многих лет операторы беспроводной связи собирали и анализировали информацию, чтобы улучшить уровень обслуживания своих сетей. Однако эти данные все чаще используются в других целях и собираются третьими лицами для новых услуг. Например, некоторые приложения для смартфонов накапливают информацию о местоположении независимо от того, имеет ли она отношение к функциям самого приложения. Цель других приложений — построить бизнес вокруг знания о местоположении пользователя. Яркий тому пример — веб-служба Foursquare, которая дает людям возможность «отметиться» в местах, которые они любят посещать. Компания получает доход от программ лояльности, а также рекомендуя рестораны и другие объекты, так или иначе связанные с местоположением.

Возможность собирать геолокационные данные о пользователях становится чрезвычайно ценной. На уровне отдельных лиц она позволяет нацеливать рекламу, исходя из местоположения человека или его предполагаемого пункта назначения. Эту информацию можно объединять для выявления определенных тенденций. Данные о местоположении массовых скоплений дают компаниям возможность обнаруживать пробки, не видя самих автомобилей, на основании количества и скорости перемещения телефонов вдоль шоссе. Компания AirSage ежедневно обрабатывает три миллиарда записей

геолокационных данных о перемещении миллионов абонентов сотовой связи для создания отчетов о ситуации на дорогах более чем в 100 городах по всей Америке в режиме реального времени. Две другие компании, которые занимаются геолокацией, Sense Networks и Skyhook, имея данные о местоположении, сообщают, в каких районах города активнее кипит ночная жизнь или сколько протестующих собралось на демонстрации.

Возможно, наиболее важным окажется некоммерческое использование геолокационных данных. Сэнди Пентлэнд, руководитель динамической лаборатории имени Хьюмана при МТИ, и бывший студент Натан Игл вместе открыли, по их словам, «интеллектуальный анализ действительности». Под этим подразумевается обработка больших объемов данных, получаемых с мобильных телефонов, для прогнозирования поведения людей. Они проанализировали передвижение людей и примеры звонков, чтобы определить, что человек заболел гриппом, прежде чем он сам это поймет. При вспышке смертельного гриппа можно спасти миллионы жизней, автоматически определяя, кого следует изолировать и где его найти. Но, как мы рассмотрим позже, попав в безответственные руки, интеллектуальный анализ действительности может привести к ужасающим последствиям[61].

Натан Игл, основатель стартапа Jana, базирующегося на данных о беспроводной связи, исследовал вопросы распространения заболеваний и процветания городов. Он обработал объединенные данные с мобильных телефонов около 500 миллионов человек в Латинской Америке, Африке и Европе, полученные более чем от 200 операторов беспроводной связи в 80 странах. В одном из исследований Игл и его коллега объединили данные о местоположении абонентов предоплаченной связи в Африке с суммами, которые те тратили на пополнение счета, и выяснили, что эти суммы сильно коррелируют с доходом: хорошо обеспеченные люди покупают больше минут за один раз. Одним из парадоксальных открытий Игла стало то, что трущобы не только являются центром нищеты, но и выступают в качестве экономических трамплинов[62]. Все эти примеры показывают косвенное использование данных о местоположении, которое не имеет ничего

общего с их первоначальным назначением — маршрутизацией мобильной связи. Напротив, как только информация о местоположении датифицируется, появляются новые области ее применения, позволяя извлечь из нее новую ценность.

Когда взаимодействия становятся данными

Некоторые границы датификации имеют личный характер: это наши отношения, переживания и настроения. Идея датификации лежит в основе многих социальных сетевых веб-служб. Социальные сети не только предоставляют нам платформу для поиска друзей и коллег, а также поддержания связи с ними, но и преобразуют нематериальные элементы нашей повседневной жизни в данные, которые можно использовать новыми способами. Так, Facebook датифицирует отношения. Они всегда представляли собой информацию, но официально не считались данными, пока не появился «социальный граф» Facebook. Twitter датифицирует настроения, предлагая людям способ легко записывать свои бессвязные мимолетные мысли и делиться ими с другими. LinkedIn датифицирует длительный профессиональный опыт (так же как Мори преобразовывал старые журналы), превращая эту информацию в прогнозы о нашем настоящем и будущем: с кем мы, возможно, знакомы и какую работу хотели бы получить.

Использование данных по-прежнему находится в зачаточном состоянии. Со стороны Facebook было весьма проницательно проявить терпение и не афишировать новые способы применения данных пользователей, зная, что эта информация могла быть шокирующей. Кроме того, компания все еще приспосабливает свою бизнес-модель (и политику конфиденциальности) к необходимому количеству и типу сбора данных. Поэтому большинство критических замечаний в адрес Facebook направлены на то, какие данные она способна получить, и гораздо меньше — на то, что с ними происходит на самом деле. Facebook охватывает более 850 миллионов активных пользователей в месяц, между которыми установлено более ста миллиардов дружественных связей. Получается, что социальный граф

представляет около 10% населения мира, сведения о которых датифицированы и находятся в руках одной компании.

Потенциальные сферы применения таких данных необычны. Некоторые начинающие компании в области потребительского кредитования рассматривают вопрос о разработке кредитной оценки на основе социального графа Facebook. Система оценки потенциальных заемщиков FICO использует 15 переменных, чтобы спрогнозировать, выплатит ли заемщик кредит. На основании внутреннего исследования один солидно финансируемый (но, к сожалению, анонимный) стартап выдвинул следующее предположение. О том, выплатит ли человек задолженность, красноречивее всего говорит поведение его друзей в аналогичной ситуации. Таким образом, обширные данные Facebook могут составить основу огромных новых бизнес-областей, которые выходят далеко за рамки поверхностного обмена фотографиями, обновления статуса и пометок «Нравится».

В Twitter данные используются не менее интересно. Более 100 миллионов человек ежедневно отправляют 250 миллионов кратких твитов, которые чаще всего представляют собой не что иное, как случайные обрывки фраз[63]. Компания дает возможность датифицировать мысли, настроения людей и взаимодействия между ними — то, что невозможно было получить ранее. Twitter заключила с компаниями DataSift и Grip соглашение на продажу доступа к данным (несмотря на то что все твиты являются общедоступными, «закулисный» доступ к ним платный). Многие компании проводят анализ твитов (иногда с помощью так называемого метода «анализа настроений»), чтобы собрать совокупные отзывы клиентов или оценить эффективность маркетинговых кампаний.

Два хедж-фонда — Derwent Capital в Лондоне и MarketPsych в Калифорнии — начали анализировать датифицированный текст твитов в качестве сигналов для инвестиций на фондовом рынке (при этом сохранив свои торговые стратегии в секрете; к примеру, они могли отдать предпочтение компаниям, специализирующимся на коротких продажах, а не на импульсной торговле). Обе компании теперь продают информацию трейдерам. В частности, хедж-фонд

MarketPsych совместно с медиакомпанией Thomson Reuters предлагает не менее 18 864 отдельных индексов по 119 странам. Эти индексы основаны на эмоциональных состояниях (оптимизм, подавленность, радость, страх, гнев и пр.) и даже таких факторах, как инновации, судебные разбирательства и конфликты, и обновляются ежеминутно. Данные используются не столько людьми, сколько компьютерами: математические гении Уолл-стрит (так называемые «кванты»[20]) с их помощью выявляют скрытые корреляции, которые можно превратить в прибыль[64]. А по словам одного из отцов анализа социальных сетей Бернардо Губермана, по частоте твитов на определенную тему можно спрогнозировать кассовые сборы кинокомпаний Голливуда. Вместе с коллегой из компании HP Губерман разработал модель для отслеживания скорости публикации новых твитов. Благодаря ей можно спрогнозировать успех фильма точнее, чем это делали рыночные прогнозисты[65].

Этим широта возможностей не ограничивается. Сообщения Twitter содержат всего 140 символов, однако метаданные, связанные с ними, несут много полезной информации. Метаданные («информация об информации») состоят из 33 отдельных элементов. Некоторые кажутся не слишком полезными (например, фоновый рисунок на странице пользователя Twitter или программное обеспечение, которое он использует для доступа к веб-службе), другие чрезвычайно интересны (например, используемый язык интерфейса службы, географическое положение пользователя, количество и имена людей, чьи твиты он читает и которые читают его твиты). Исследование, проведенное журналом Science в 2011 году, показало то, что невозможно было выявить прежде: перемены настроения людей имеют ежедневные и еженедельные закономерности, общие для всех культур во всем мире. Предметом анализа стали 509 миллионов твитов, полученных за два года от 2,4 миллиона пользователей из 84 стран. Настроения удалось датифицировать[66].

Датификация подразумевает перевод в анализируемую форму не только отношений и настроений, но и поведения людей, которое трудно было бы отследить иным способом, особенно в более широких группах населения и их подгруппах. Биолог Марсель Салатэ из

Университета штата Пенсильвания и инженер-программист Шашанк Ханделвал проанализировали твиты с целью убедиться, что вероятность того, что человек сделает прививку от гриппа, напрямую зависит от его отношения к прививкам как таковым. Важно отметить, что у них были метаданные о связях между пользователями Twitter, читающими твиты друг друга. Это позволило пойти дальше и выявить существование подгрупп непривитых людей. Такое волнующее открытие ставит под сомнение понятие «коллективного иммунитета», согласно которому проведение вакцинации среди большей части населения предотвращает вспышки заболеваний даже среди непривитых людей. Примечательно, что в отличие от других исследований, таких как Google Flu Trends, где объединенные данные использовались для рассмотрения вопроса о состоянии здоровья, анализ настроений, проведенный Салатэ, позволил обнаружить само *поведение* в отношении здоровья[67].

Первые находки уже показывают направление, в котором уверенно движется датификация. Подобно Google, социальные сети, такие как Facebook, Twitter, LinkedIn, Foursquare, Zynga и другие, сидят на сокровищнице датифицированной информации, проанализировав которую можно было бы пролить свет на динамику человеческого и социального поведения на всех уровнях — от личности до общества в целом.

Повсеместная датификация

Проявив немного фантазии, можно перевести в форму данных немыслимое число объектов и сделать при этом неожиданные открытия. В духе экстравагантных работ токийского профессора Косимицу компания IBM в 2012 году получила патент США на «систему безопасности помещений с использованием наземной вычислительной технологии». Говоря простым языком, это сенсорное напольное покрытие, подобное гигантскому экрану смартфона. Сфера его потенциального применения весьма обширна. Такой пол мог бы обнаруживать расположенные на нем предметы и определять, когда нужно включить свет в комнате или открыть двери. Более того, он опознавал бы людей по их весу, стоячей позе и походке. Сообщал,

когда кто-то упал и не может подняться. С помощью этой технологии торговые компании могли бы отслеживать поток клиентов в магазине. Таким образом, датификация напольного покрытия открывает безграничные возможности ее применения.

И это будущее не за горами. Возьмем, к примеру, движение Quantified Self («Измерение себя»). Его участники — разношерстная группа фанатов фитнеса, медицины и техники, которые измеряют каждый элемент своего тела и деятельности, чтобы улучшить качество своей жизни или по крайней мере узнать что-то новое, что раньше не удавалось измерить количественно. Пока что движение по отслеживанию личных показателей немногочисленное, но его ряды постоянно пополняются.

Благодаря смартфонам и недорогой вычислительной технике датификация наиболее важных аспектов жизни стала проще, чем когда-либо. Множество стартапов предоставляют людям возможность отслеживать свой сон путем измерения мозговых волн в течение всей ночи. Компания Zeo уже создала крупнейшую в мире базу данных активности во время сна и обнаружила различия в количестве фаз быстрого сна у мужчин и женщин. Компания Asthmapolis провела другой эксперимент: прикрепила к ингаляторам от астмы датчики, которые отслеживают местоположение с помощью GPS. Собранная информация позволяет выяснить, какие факторы окружающей среды провоцируют приступы астмы (например, близость к определенным видам посевных культур).

Компании Fitbit и Jawbone предлагают людям инструмент для оценки своей физической активности и сна. Владельцы браслетов компании Basis могут контролировать жизненно важные функции, в том числе частоту сердечных сокращений и электропроводность кожи, которые являются показателями стресса[68]. Получение данных становится проще и непринужденнее, чем когда-либо. Так, в 2009 году Apple подала заявку на патент для сбора данных о насыщенности крови кислородом, частоте сердечных сокращений и температуре тела через наушники-вкладыши[69].

Датификация принципов работы человеческого тела открывает широкое поле для изучения. Исследователи из Университетского

колледжа Йёвик в Норвегии и компания Derawi Biometrics разработали приложение для смартфонов, которое анализирует походку человека, чтобы использовать ее в качестве системы безопасности для разблокировки телефона[70]. Роберт Делано и Брайан Пэрисит из Технологического научно-исследовательского института штата Джорджия создали приложение iTrem, которое с помощью встроенного в телефон акселерометра контролирует тремор частей тела при болезни Паркинсона и других неврологических расстройствах. Это приложение удобно как для врачей, так и для пациентов. Пациенты получают возможность обойтись без дорогостоящих визитов к врачу, а медработники — удаленно отслеживать нарушения функций у людей и их реакцию на лечение[71]. По мнению исследователей в Киото, смартфон измеряет степень дрожания не настолько точно, как акселерометр, используемый в специализированном медицинском оборудовании. Однако разница в эффективности несущественна и делает показания приложения достаточно надежными[72]. Выходит, что немного беспорядочности не помеха точности.

В большинстве таких случаев мы получаем информацию и переводим ее в форму данных для повторного использования. Для этого годится практически любая информация, полученная где угодно. Стартап GreenGoose продает крошечные датчики движения, которые можно разместить на объектах, чтобы отслеживать частоту их применения. Прикрепив такой датчик на пачку зубной нити, лейку или коробку кошачьего туалета, вы сможете датифицировать гигиену полости рта и уход за растениями или домашними животными.

С тех пор как мир начал датифицироваться, использование информации стало настолько широким, насколько хватит фантазии. Мори раскрыл скрытую ценность данных путем кропотливого ручного анализа. Сегодня у нас есть инструменты (статистические данные и алгоритмы) и необходимое оборудование (компьютерные процессоры и хранилища), которые позволяют делать то же самое гораздо быстрее, в большем масштабе и во множестве различных областей. В эпоху больших данных можно извлекать пользу из самых неожиданных объектов.

Мы находимся в середине большого инфраструктурного проекта, который в некотором роде конкурирует с атрибутами прошлого — от римских акведуков до «энциклопедистов» эпохи Просвещения. Мы не в состоянии оценить проект по достоинству, поскольку он едва появился и мы полностью поглощены им. К тому же, в отличие от воды, текущей по акведукам, продукт нашего труда нематериален. Этот проект — датификация. Подобно остальным инфраструктурам, она приведет к фундаментальным изменениям в обществе.

Акведуки способствовали росту городов, печатные станки — просвещению, а газеты — подъему национального государства. Эти инфраструктуры имели дело с потоками (воды и знаний), так же как телефон и интернет. В отличие от них датификация — фундаментальное изменение действительности в человеческом понимании. Благодаря большим данным мы перестанем рассматривать окружающий мир как бесконечное множество событий, которые объясняются как природные или социальные явления, а взглянем на него как на область, состоящую в основном из информации.

Более века назад физики предположили, что не атомы, а информация является настоящей основой всего сущего. И пусть это звучит эзотерически, но во многом именно благодаря датификации мы теперь можем полномасштабно фиксировать и рассчитывать материальные и нематериальные аспекты существования и действовать в соответствии с ними.

Взглянув на мир с точки зрения информации — бескрайних просторов данных, которые нам предстоит постичь, — мы получим небывалое представление об окружающей действительности. Это мировоззрение охватит все сферы нашей жизни. Со временем датификация, которая затмит акведуки и газеты, станет конкурировать с типографией и интернетом, вручив нам инструменты для преобразования мира с помощью данных. Сейчас делом заняты самые продвинутые пользователи. Большие данные используются для создания новых форм ценности, которые мы рассмотрим в следующей главе.

Примечания к главе 5

Глава 6

Ценность

В конце 1990-х годов началось массовое засорение интернета. Программы, именуемые «спам-ботами», программировались на то, чтобы узнать последовательность действий для подписки на бесплатную учетную запись электронной почты, а затем использовать ее для массовой рассылки рекламных сообщений десяткам миллионов людей, переполняя почтовые ящики. Эти же роботы могли регистрироваться на сайтах, а затем оставлять сотни рекламных объявлений в разделах комментариев. Интернет превращался в неуправляемое, недружелюбное и недоброжелательное место. В частности, казалось, он перестал быть примером открытости и простоты использования, предлагающим такие возможности, как бесплатная электронная почта. Когда компании вроде TicketMaster предлагали приобрести в интернете билеты на концерты по принципу «кто не успел, тот опоздал», подлые программы скупали их все, опережая реальных людей.

В 2000 году новоиспеченный выпускник колледжа 22-летний Луис фон Ан загорелся идеей решить эту проблему: нужно заставить регистрирующегося доказать, что он человек. Луис нашел то, что легко давалось людям, но представляло трудности для компьютеров: опознать в процессе регистрации искаженные, трудно читаемые буквы. Люди смогут расшифровать их и ввести правильный текст в считаные секунды, но компьютер будет поставлен в тупик. Компания Yahoo реализовала эту идею и стремительно сократила атаки спам-ботов. Фон Ан назвал свое творение Captcha (англ. Completely Automated Public Turing Test to Tell Computers and Humans Apart — «полностью автоматизированный публичный тест Тьюринга для различения компьютеров и людей»). Пять лет спустя около 200 миллионов Captcha стали вводиться ежедневно.

Это принесло Луису фон Анну, выходцу из гватемальской семьи, которая владела кондитерской фабрикой, широкую известность и

работу преподавателя компьютерных наук в Университете Карнеги—Меллон, после того как ему была присвоена степень доктора философии. Благодаря своему изобретению в возрасте 27 лет он получил одну из престижных премий Фонда Макартуров[21] за «гениальность» в размере 500 тысяч долларов. Когда Луис понял, что каждый день миллионы людей тратили впустую около десяти секунд своего времени на ввод раздражающих букв и при этом огромное количество получаемой информации попросту выбрасывалось, он усомнился в гениальности своего изобретения[73].

Луис фон Ан искал способы более продуктивного применения человеческой вычислительной мощности. В итоге был создан тест-преемник с подобающим названием ReCaptcha. Теперь, вместо того чтобы вводить случайные буквы, люди набирают два слова из проектов по сканированию текстов, которые не удалось распознать с помощью компьютерной программы оптического распознавания символов. Одно слово подтверждает, что его уже вводили другие пользователи (и, следовательно, является сигналом того, что пользователь — человек), а другое — новое слово, которое нужно уточнить. Чтобы гарантировать точность, система отображает одно и то же случайное слово до тех пор, пока примерно пять разных пользователей не введут его без ошибок, и только тогда слово считается правильным. Таким образом, данные имеют как основное назначение (доказать, что пользователь является человеком), так и второстепенное — расшифровать непонятные слова из оцифрованных текстов. Система ReCaptcha оказалась настолько полезной, что в 2009 году компания Google решила внедрить ее в свой проект сканирования книг.

Выгода от системы огромна, если учесть, сколько нужно людей для выполнения такой работы. Более 200 миллионов ReCaptcha вводятся ежедневно. Примерно 10 секунд, затрачиваемых на эту операцию, — это в общей сумме около полумиллиона часов в день. Минимальная заработная плата в США в 2012 году составляла 7,25 доллара в час. Если бы для уточнения слов, которые компьютер не мог понять, пришлось обратиться на рынок труда, это обошлось бы примерно в 35 миллионов долларов в день, или более чем 1 миллиард долларов в год.

Но Луис фон Ан разработал систему, которая делает это, по сути, бесплатно.

История ReCaptcha подчеркивает, насколько важны повторные данные, особенно если это большие данные. В эпоху цифровых технологий мы осознали роль данных в поддержке операций, и нередко они сами становились товаром. В мире больших данных все снова меняется. Акцент переносится на потенциальное применение данных в будущем. Этот процесс влечет за собой далеко идущие последствия. Он влияет на то, как компании оценивают данные, имеющиеся в их распоряжении, и кому предоставляют к ним доступ. Он позволяет компаниям (а может быть, и вынуждает их) менять свои бизнес-модели, а также меняет отношение организаций к данным и способы их использования.

Информация всегда была необходима для рыночных сделок. Данные дают возможность проводить ценовые исследования, а те — определить объемы производства. Кроме того, на рынках давно торгуют определенными видами информации. Примеры тому — книги, статьи, музыка, фильмы, а также финансовая информация (такая как цены на акции). В последние несколько десятилетий подобная информация была объединена понятием личных данных. Специализированные брокеры данных в США, такие как Acxiom, Experian и Equifax, запрашивают кругленькие суммы за всеобъемлющие досье личной информации на сотни миллионов пользователей. С появлением Facebook, Twitter, LinkedIn, Foursquare и других платформ социальных сетей наши личные связи, мнения, предпочтения и примерный распорядок дня пополнили и без того огромный пул личной информации, уже имеющейся о каждом из нас.

Хотя ценность данных уже давно не вызывает сомнений, прежде они воспринимались как дополнение к основной коммерческой деятельности или как довольно ограниченные категории интеллектуальной собственности и личной информации. Но в эпоху больших данных *все* данные без исключения будут рассматриваться как ценные сами по себе.

Говоря «все данные», мы имеем в виду даже самые сырые, самые, казалось бы, обыденные отрывки информации. Это могут быть показатели датчика температуры на заводском механизме. Или поток

координат GPS в режиме реального времени, показатели акселерометра и уровень топлива в автомобиле — или в целом автопарке из 60 000 единиц. Или миллиарды старых поисковых запросов, или цены на все авиабилеты по всем рейсам коммерческих авиакомпаний США за прошедшие годы.

До недавнего времени не существовало простого способа сбора, хранения и анализа таких данных, что значительно ограничивало возможность извлечь из них потенциальную ценность. В знаменитом примере Адама Смита[22] производителю булавок, с которым он обсуждал разделение труда в XVIII веке, потребовались бы наблюдатели, постоянно присматривающие за сотрудниками, а также проведение измерений и подсчет выпущенной продукции с помощью бумаги и пера. Даже измерение времени было бы затруднительным, учитывая, что надежные часы в то время были редкостью[74]. Ограничения технической среды сформировали взгляды классических экономистов на устройство экономики — то, о чем они едва ли имели представление, так же как рыба не знает, что она мокрая. Поэтому, рассматривая факторы производства (земля, труд и капитал), они, как правило, упускали из виду роль информации. Хотя за последние два столетия стоимость сбора, хранения и использования данных успела снизиться, до недавних пор это по-прежнему оставалось относительно дорогим удовольствием.

Характерное отличие нашего времени состоит в том, что большинство ограничений, присущих сбору данных, исчезли. Технологии достигли того уровня, когда получение и запись огромных объемов данных стали достаточно доступными. Данные можно собрать пассивно, без особых усилий со стороны тех, о ком ведется запись, и даже без их ведома. А поскольку стоимость хранения значительно упала, оправдать хранение данных проще, чем удалить их. В таких условиях к вашим услугам намного больше данных и по более низким ценам, чем когда-либо. За последние 50 лет стоимость цифрового хранения урез*а*лась вдвое каждые два года, в то время как плотность хранимых данных увеличивалась в 50 миллионов раз[75]. В свете информационных компаний, таких как Farecast или Google, где на одном конце цифровой линии сборки поступают сырые факты, а на

другом выходит обработанная информация, данные начинают восприниматься как новый фактор производства.

Непосредственная ценность больших данных очевидна тем, кто их собирает. По сути, сбор данных производится с конкретной целью. Магазины собирают данные о продажах для надлежащего финансового учета. Заводы контролируют выпуск продукции, чтобы обеспечить ее соответствие стандартам качества. Сайты регистрируют все действия пользователей, вплоть до области перемещения мыши, чтобы проанализировать и оптимизировать контент, предоставленный посетителям. Первичное использование данных оправдывает сбор и обработку информации. Записывая информацию не только о книгах, которые покупают клиенты, но и о веб-страницах, которые они посещают, компания Amazon знает, что данные послужат для формирования персонализированных рекомендаций клиентам. Таким же образом Facebook отслеживает обновления статуса и пометки «Нравится» пользователей, чтобы подобрать подходящие рекламные объявления для показа на своем сайте с целью получения дохода.

В отличие от материальных объектов (употребляемой пищи, горящей свечи и пр.), ценность данных не уменьшается по мере их потребления. Данные можно обрабатывать снова и снова. Они представляют собой то, что экономисты называют «неконкурирующим» товаром. Им могут пользоваться несколько человек одновременно без ущерба друг для друга. К тому же, в отличие от материальных благ, информация не изнашивается по мере употребления. Amazon с помощью данных о прошлых операциях формирует рекомендации для своих клиентов и делает это неоднократно не только для тех клиентов, от которых получены данные, но и для многих других.

Поскольку ценность данных не ограничивается одним конкретным случаем, их можно употребить в дело многократно как с одной и той же целью, так и с разными. Особенно важен для нас второй случай, поскольку мы пытаемся понять, насколько ценной будет для нас информация в эпоху больших данных. Мы уже рассмотрели примеры реализации потенциала данных, когда сеть магазинов Walmart проанализировала старые квитанции продаж и заметила выгодную корреляцию между ураганами и продажами Pop-Tarts.

Все это означает, что абсолютная ценность данных намного превышает ту, которую удается извлечь при первичном использовании. Компании могут эффективно работать с данными, даже если первое или каждое последующее использование приносит лишь небольшую толику ценности.

«Альтернативная ценность» данных

Для того чтобы получить представление о том, как повторное использование данных отражается на их конечной ценности, рассмотрим электрические автомобили. Станут ли они способом транспортировки, зависит от схемы логистики, которая так или иначе связана со временем работы аккумулятора. Водители должны иметь возможность быстро и удобно подзарядить аккумуляторы автомобиля, а энергетические компании — гарантировать, что энергия, полученная транспортным средством, не дестабилизирует сеть. На то, чтобы путем проб и ошибок прийти к теперешнему эффективному распределению АЗС, ушли десятки лет, но нам пока неизвестно, какой окажется потребность в подзарядке электрических автомобилей и где следует размещать для них зарядные станции.

Что удивительно, это не столько инфраструктурная задача, сколько информационная, и большие данные являются важной частью решения. В ходе проведенного в 2012 году исследования IBM в сотрудничестве с калифорнийской компанией Pacific Gas and Electric Company и автопроизводителем Honda собрала огромное количество информации, чтобы ответить на вопросы о том, когда и где электрические автомобили будут подзаряжаться и как решить проблему источников электропитания. IBM разработала сложную интеллектуальную модель, основанную на многочисленных входящих данных, таких как уровень заряда аккумулятора, местоположение автомобиля, время суток и доступные разъемы на ближайших станциях зарядки электромобилей. Компания связала эти данные с текущим потреблением электросети, а также статистическими данными о закономерностях энергопотребления. Анализ огромных потоков данных в режиме реального времени и статистических данных из нескольких источников дал IBM возможность определить

оптимальное время и место для подзарядки электромобилей. Он также показал, где лучше всего строить станции для их зарядки[76]. С течением времени системе понадобится учитывать различия в ценах на таких станциях. Даже прогноз погоды придется брать в расчет (в солнечный день на близлежащих станциях, работающих на солнечной энергии, электричество будет в изобилии, но по прогнозу также может предстоять неделя дождей, в течение которой солнечные панели будут простаивать).

Система получает информацию, созданную с одной целью, и работает с ней повторно с другой — иными словами, данные переходят от первичного использования к вторичному. Это делает их гораздо более ценными с течением времени. Индикатор уровня заряда аккумулятора автомобиля сообщает водителю, когда требуется подзарядка. Энергетическая компания собирает данные об эксплуатации электросети, чтобы управлять ее стабильностью. Это примеры первичного использования. Оба набора данных находят вторичное применение — и новую ценность, когда рассматриваются с совершенно другой целью: определить, когда и где выполнять подзарядку, а также где строить новые станции обслуживания электромобилей. Помимо этих данных включается новая, вспомогательная информация — местоположение автомобиля и статистические данные о работе в сети. К тому же IBM использует данные не один раз, а многократно, постоянно обновляя свои сведения о потреблении энергии электромобилями, а также о нагрузке на электросеть.

Истинная ценность данных — как айсберг в океане. На первый взгляд видна лишь незначительная часть, в то время как все остальное скрыто под водой. Инновационные компании, которые понимают это, могут извлечь скрытую ценность и получить потенциально огромные преимущества. Проще говоря, ценность данных необходимо рассматривать с точки зрения всех возможностей их дальнейшего использования, а не только нынешнего. Мы могли убедиться в этом на многих рассмотренных примерах. Компания Farecast анализировала данные о продаже авиабилетов, чтобы прогнозировать будущие цены на авиабилеты. Компания Google повторно применила условия поиска,

чтобы узнать показатели распространения гриппа. Доктор Макгрегор собирала показатели жизненно важных функций младенцев, чтобы прогнозировать развитие инфекций. Мори многократно изучал старые капитанские журналы, чтобы выявить океанские течения.

И все-таки важность повторного применения данных недооценивается как в бизнесе, так и в обществе. Мало кто из руководителей нью-йоркской компании Con Edison мог предположить, что информация о кабелях со времен 1800-х годов и записи о техническом обслуживании могут пригодиться для предотвращения будущих аварий. Потребовалось новое поколение статистиков, а также новое поколение методов и средств, чтобы высвободить эту скрытую ценность данных. До недавних пор даже многим технологическим и интернет-компаниям не было известно, насколько ценным бывает повторное использование данных.

Данные можно наглядно представить в виде энергии, как ее видят физики. Это хранящаяся, или потенциальная энергия, которая дремлет в каждом из объектов, будь то сжатая пружина или мяч на вершине пригорка. Энергия в этих объектах находится в скрытом (потенциальном) состоянии, пока не будет высвобождена (например, если отпустить пружину или подтолкнуть мяч, чтобы он покатился вниз). Тогда она становится кинетической, поскольку они движутся и прилагают силу к другим объектам физического мира. После первичного использования данных их ценность остается прежней, но только в «спящем» состоянии. Она сохраняет свой потенциал, как пружина или мяч, вплоть до вторичного применения, когда преимущества данных раскроются с новой силой. В эпоху больших данных у нас, наконец, есть все необходимое (мышление, изобретательность и инструменты), чтобы высвободить их скрытую ценность.

В конечном счете ценность данных заключается в том, что можно получить от их всестороннего использования. Эти, по-видимому бесконечные, возможности служат альтернативами, но не с точки зрения финансовых инструментов, а с точки зрения практических вариантов выбора. Стоимость данных определяется суммой таких вариантов — так сказать, «альтернативной ценностью» данных. Раньше, задействовав данные по основному назначению, мы, как

правило, считали, что они свою миссию уже выполнили и теперь их можно окончательно удалить. Ведь, казалось бы, основная ценность получена. В эпоху больших данных все иначе: данные, как волшебный алмазный рудник, обеспечивают отдачу еще долго после того, как их номинальная ценность уже извлечена. Есть четыре мощных способа раскрыть альтернативную ценность данных: основное повторное использование, слияние наборов данных, поиск данных «2 в 1» и учет «амортизации» ценности данных.

Повторное использование данных

Классический пример инновационного повторного использования данных — условия поиска. На первый взгляд, информация становится бесполезной, как только ее первоначальное назначение достигнуто. Мгновенное взаимодействие между пользователем и поисковой системой приводит к подготовке списка сайтов и объявлений, тем самым выполняя определенную функцию, уникальную на тот конкретный момент. Но и старые запросы могут быть чрезвычайно полезными. Такие компании, как Hitwise, которая принадлежит брокеру данных Experian и занимается измерением веб-трафика, дают клиентам возможность проводить интеллектуальный анализ поискового трафика, чтобы выявить предпочтения потребителей. Маркетологи могут использовать Hitwise, чтобы узнать, какой цвет будет в моде этой весной — розовый или снова черный. Компания Google предоставляет пользователям открытый доступ к своей версии аналитики условий поиска. В сотрудничестве с BBVA, вторым по величине банком Испании, Google запустила службу бизнес-прогнозирования, чтобы анализировать сектор туризма и продавать в режиме реального времени экономические показатели, основанные на данных поиска. Банк Англии работает с поисковыми запросами, связанными с объектами недвижимости, чтобы уточнить тенденции роста или падения цен на жилье.

Компании, которые недооценили важность повторного использования данных, усвоили урок на собственном горьком опыте. В начале своей деятельности Amazon заключила сделку с компанией AOL по запуску технологии, лежащей в основе интернет-магазина

AOL. Для большинства людей это выглядело как обычная сделка внешнего подряда. «Но что на самом деле интересовало Amazon, так это данные о том, что пользователи ищут и покупают, поскольку это позволило бы повысить эффективность рекомендательной системы компании», — поясняет Андреас Вайгенд, бывший руководитель исследовательских работ в Amazon[77]. Бедняжка AOL так этого и не поняла. Она видела преимущества только с точки зрения первичного использования — продаж, в то время как в Amazon смекнули, что можно извлечь выгоду из вторичного использования данных.

Или возьмем первые шаги Google в области распознавания речи. В 2007 году был запущен голосовой телефонный справочник GOOG-411, который функционировал вплоть до 2010 года. Поисковый гигант не имел своей технологии распознавания речи, поэтому пришлось ее лицензировать. Компания заключила договор с лидером в этой области — компанией Nuance, которая была рада обзавестись таким ценным клиентом. Но Nuance плохо разбиралась в том, что касалось больших данных: в договоре не уточнялось, кто является держателем записей голосового перевода, поэтому Google сохраняла их для себя. Эти данные были необходимы для совершенствования технологии, но также годились для создания новой службы распознавания речи с нуля. На тот момент Nuance воспринимала себя как организацию, которая занимается лицензированием программного обеспечения, а не обработкой данных. Осознав свою ошибку, компания начала заключать сделки с мобильными операторами и производителями мобильных телефонов для внедрения своей службы распознавания речи, что позволило и Nuance собирать данные[78].

Ценность повторного использования данных — хорошая новость для организаций, которые собирают или имеют в своем распоряжении большие наборы данных, но пока с ними почти не работают (например, обычные компании, которые в основном функционируют вне интернета). Может оказаться, что они сидят на неиспользуемых информационных гейзерах. Некоторые компании, собрав данные и единожды их задействовав (а может, и не сделав этого вовсе), хранили данные лишь из-за низкой стоимости хранения. Ученые прозвали компьютеры с такой старой информацией «гробницами данных».

Технологические и веб-компании стоят первыми в очереди по освоению наплыва данных, поскольку собирают огромное количество информации, просто находясь в интернете, и опережают конкурентов в отрасли по ее анализу. При этом все компании остаются в выигрыше. Консультанты McKinsey & Company приводят в пример логистическую компанию (ее название они оставили анонимным). Компания обратила внимание на то, что в процессе доставки товаров она накапливала огромные ряды информации о поставках в глобальном масштабе. Учуяв возможности, она создала специальный отдел по продаже объединенных данных в форме деловых и экономических прогнозов — иными словами, офлайновую версию прошлого бизнеса Google, построенного на поисковых запросах[79].

Некоторые компании благодаря своему положению в цепочке создания ценности информации накапливают огромное количество данных, даже если не имеют в этом существенной необходимости или не практикуют их повторное использование. Так, например, операторы мобильной связи собирают информацию о местоположении своих абонентов, чтобы маршрутизировать их вызовы. Эти компании видят лишь узкое техническое назначение таких данных. Но их ценность значительно повышается при повторном использовании компаниями, которые распространяют персонализированную рекламу на основе местоположения. Иногда ценность формируют не отдельные точки данных, а их совокупность. Это дает возможность компаниям, таким как AirSage и Sense Networks, продавать информацию о том, где люди собираются по пятничным вечерам или насколько медленно ползут машины в пробках. Такая информация может служить для определения стоимости недвижимости или расценок для рекламных щитов.

Даже самая банальная информация может иметь особое значение, если направить ее в правильное русло. Вернемся к операторам мобильной связи: у них есть записи о том, где и когда телефоны подключались к базовым станциям, включая данные об уровне сигнала. Операторы уже давно используют эти сведения для тонкой настройки производительности своих сетей, решая, где добавить или обновить инфраструктуру. Но данные имеют и много других

потенциальных применений. С их помощью производители телефонов могут узнать, например, что влияет на уровень сигнала, чтобы улучшить качество приема сигнала на своих устройствах. Мобильные операторы сталкиваются с большим количеством юридических ограничений, которые, как правило, запрещают повторное использование данных или обмен ими ввиду конфиденциальности — изобретения эпохи малых данных. Во времена больших данных такие ограничения уже неактуальны.

Искусственно созданные данные

Иногда скрытую ценность можно раскрыть, только объединив один набор данных с другим, возможно, совершенно непохожим. По-новому комбинируя данные, можно добиться инновационных открытий, что подтверждает научное исследование, опубликованное в 2011 году. В нем шла речь о том, что мобильные телефоны повышают вероятность развития раковых заболеваний. Учитывая, что в мире насчитывается шесть миллиардов мобильных телефонов — практически по одному на каждого человека, — это очень важный вопрос. Множество исследователей искали подобную связь, но успеху препятствовали слишком маленькая выборка, недостаточная длительность изыскания или анализ только собственных данных, что чревато ошибкой. Тем не менее команда ученых из Датского онкологического общества разработала интересный подход, основанный на ранее собранных данных[80].

Датская база данных всех абонентов мобильной связи ведет начало с момента появления мобильных телефонов в 1985 году. Исследование охватило тех, кто пользовался мобильным телефоном с 1990 по 2007 год, за исключением корпоративных и других абонентов, чьи социально-экономические данные были недоступны. Получалось 358 403 человека. В Дании также существовал реестр всех онкологических больных, в котором числилось 10 729 человек, страдающих опухолями центральной нервной системы в обозначенный период. Объединив два набора данных, исследователи рассчитывали найти корреляции. Продемонстрируют ли владельцы мобильных телефонов более высокую заболеваемость раком, чем те, у кого их нет? И правда ли, что

абоненты, которые дольше пользуются мобильным телефоном, более подвержены раковым заболеваниям?

Несмотря на масштабы исследования, информация не была ни беспорядочной, ни неточной: оба набора данных составлялись с учетом строгих стандартов качества для медицинских и коммерческих целей. Информация собиралась в условиях, исключающих отклонения, несколькими годами ранее и по причинам, которые не имели ничего общего с целью этого исследования. Самое главное, что оно проводилось не на основе выборки, а близко к условию «*N* = всё»: учитывались почти каждый случай рака и почти каждый пользователь мобильного телефона (что в целом составило 3,8 миллиона человеко-лет владения мобильными телефонами). Благодаря тому что исследование охватывало почти все случаи, ученые могли контролировать подгруппы, например курящих.

В результате не было обнаружено, что увеличение риска развития рака связано с использованием мобильного телефона. Поэтому эти выводы вряд ли произвели фурор в средствах массовой информации, когда данные были опубликованы в британском медицинском журнале BMJ в октябре 2011 года. А вот если бы такая связь всплыла, о ней бы писали в первых полосах газет по всему миру, тем самым ознаменовав триумф методологии «искусственно созданных данных».

При анализе больших данных совокупность важнее отдельных частей, а при перекомпоновке совокупностей нескольких наборов данных получается еще более удачная совокупность. Современные интернет-пользователи знакомы с основными «мэшапами» — службами, которые по-новому объединяют несколько источников данных. Сайт недвижимости Zillow.com накладывает информацию о недвижимости и ценах на карту окрестностей в США, а также обрабатывает наборы данных о последних деловых операциях в районе и характеристиках объектов недвижимости, чтобы спрогнозировать стоимость конкретных домов в определенном районе.

Полученный результат полезен, поскольку наглядное представление данных делает их более понятными. Но это довольно упрощенный пример. В конце концов, не так уж трудно додуматься взять информацию относительно местоположения и наложить ее на карту. С большими данными мы можем пойти гораздо дальше. И датское

исследование рака показывает, какие перспективы перед нами открываются.

Расширяемые данные

Повторное использование данных нетрудно обеспечить, если продумать их расширяемость с самого начала. Это получается не всегда (ведь мысль о том, *что* можно выжать из данных, иногда приходит намного позже, чем они были собраны), однако способствовать многократному потреблению одного и того же набора данных можно разными способами. Некоторые розничные торговцы устанавливают в магазинах камеры наблюдения таким образом, чтобы не только обнаруживать магазинных воров, но и отслеживать передвижение клиентов по магазину и места, где они останавливаются, чтобы присмотреться. Такая информация полезна для разработки лучшей выкладки товаров в магазине, а также для оценки эффективности маркетинговых кампаний. Ранее камеры видеонаблюдения служили только для обеспечения безопасности и рассматривались не более чем статья расходов. Теперь они рассматриваются как инвестиции, которые могут увеличить доход.

Как ни странно, одной из компаний, которые достигли наибольшего успеха в сборе данных с учетом расширяемости, является Google. Ее автомобили Street View, вызывающие неоднозначную реакцию общества, разъезжают по улицам, не только делая снимки домов и дорог, но и собирая данные GPS, проверяя картографическую информацию и даже попутно захватывая названия Wi-Fi-сетей (а также, вероятно, на незаконных основаниях, контент, доступный в открытых беспроводных сетях). За одну поездку автомобиль Google Street View накапливает множество потоков дискретных данных. Расширяемость обеспечивается тем, что Google применяет данные и для первичного использования, и для целого ряда вторичных. Например, данные GPS не только улучшили картографическую службу компании Google, но и были незаменимы для работы ее самоуправляемых автомобилей[81].

Дополнительные расходы на сбор нескольких потоков данных или намного большего числа точек данных в каждом потоке, как правило,

невелики, поэтому имеет смысл собирать как можно больше данных, а также делать их расширяемыми, изначально рассматривая потенциальные виды вторичного использования. Благодаря этому увеличивается альтернативная ценность информации. Суть в том, чтобы искать наборы «2 в 1», когда один и тот же набор данных, собранных определенным образом, можно применять в различных целях. Так эти сведения приобретают двойное назначение.

Обесценение данных

Поскольку стоимость хранения цифровых данных резко упала, компании получили сильный экономический стимул сохранять их для повторного использования в тех же или аналогичных целях. Однако полезность данных небезгранична.

Компании Netflix и Amazon умело используют информацию о покупках клиентов, чтобы рекомендовать новые продукты. При этом у компаний возникает соблазн многократно использовать эти записи в течение многих лет. В такой ситуации можно было бы утверждать, что в рамках соблюдения обязательных нормативов (например, закона о неприкосновенности частной жизни) компаниям следует хранить цифровые записи всегда или по крайней мере пока это экономически целесообразно. Однако все не так просто.

Информация с течением времени теряет часть своей первичной пользы. В таких условиях дальнейшее использование старых данных может не только не добавить ценности, но и фактически нивелировать пользу более новых данных. Положим, вы купили книгу на сайте Amazon лет десять назад. Вряд ли она все еще отражает ваши интересы. Если Amazon будет отталкиваться от нее, рекомендуя вам другие книги, вы вряд ли их купите, а может, вообще перестанете обращать внимание на последующие рекомендации сайта. Поскольку рекомендации основываются на всех собранных данных, наличие устаревших данных сводит на нет всю пользу новых (все еще ценных).

Таким образом, у Amazon есть огромный стимул использовать данные ровно до тех пор, пока это продуктивно. Компания должна постоянно сортировать свою базу данных, удаляя информацию, которая уже утратила свою ценность. А как узнать, что данные стали

бесполезными? Ориентироваться исключительно на время не всегда эффективно. Поэтому Amazon и другие компании разработали сложные модели, которые позволяют отделить полезные данные от бесполезных. Если клиент просматривает или покупает книгу, которая была рекомендована на основе его предыдущей покупки, интернет-магазин берет на заметку, что старые покупки по-прежнему отражают текущие предпочтения клиента. Это позволяет оценить полезность старых данных и, следовательно, смоделировать более конкретную «степень обесценения».

Не все данные обесцениваются. Некоторые компании имеют веские причины хранить данные как можно дольше, даже если регулирующие органы или общество предпочли бы их удалить или сделать анонимными в кратчайший срок. Вот почему Google давно сопротивляется призывам удалить полные IP-адреса старых поисковых запросов (вместо этого спустя 18 месяцев удаляются только четыре последние цифры, чтобы сделать поисковый запрос анонимным). Компания оставляет за собой возможность сравнивать данные (например, поисковые запросы для предпраздничного шопинга) в годовом исчислении. Кроме того, сведения о местоположении пользователей, выполняющих поиск, помогают повысить релевантность результатов. Если большинство жителей Нью-Йорка набирают Turkey (англ. «Турция», «индейка») и открывают сайты, связанные со страной, а не птицей, алгоритм будет ранжировать эти страницы выше и для остальных нью-йоркцев. Даже если ценность данных для первичного использования снижается, их альтернативная ценность может оставаться высокой.

Понятие альтернативной ценности наводит на мысль, что организациям следует собирать как можно больше данных в пределах своих возможностей для их хранения, а также передавать эти сведения третьим лицам при условии, что они сохраняют за собой так называемые «сквозные» права (термин, заимствованный из патентного лицензирования). Если повторное использование данных дает определенный коммерческий результат, первоначальный владелец этих данных может получить свою долю. Разумеется, что организации, собирающие данные и владеющие ими, не могут вообразить все возможные способы их повторного применения.

Ценность выбросов данных

Повторное использование данных иногда производится в скрытой форме. Интернет-компании записывают данные обо всех действиях пользователей на своем сайте, а затем обрабатывают каждое отдельно взятое взаимодействие как «сигнал» обратной связи для персонализации сайта, улучшения обслуживания или создания нового цифрового продукта. Интересной иллюстрацией служит рассказ о двух средствах проверки правописания.

В течение двадцати лет корпорация Microsoft разрабатывала надежное средство проверки правописания для своей программы Word. Его работа заключалась в том, чтобы сравнивать часто обновляемый словарь правильно написанных терминов с потоком символов, вводимых пользователем. Известные слова сверялись со словарем, а похожие варианты, не зафиксированные в нем, система расценивала как опечатки и предлагала исправить. Из-за усилий, затрачиваемых на формирование и обновление каждого словаря, средство проверки правописания в Microsoft Word было рассчитано только на наиболее распространенные языки. Создание и поддержка системы обошлись компании в миллионы долларов.

Посмотрим, что сделала Google. Эта компания имеет, пожалуй, наиболее полное из современных средств проверки правописания практически для всех языков мира. Система постоянно совершенствуется и непрерывно добавляет новые слова — это результат ненамеренной деятельности людей, ежедневно использующих поисковую систему. Сделали опечатку в слове iPad? Не страшно, система и так поймет. Ввели Obamacare? Запрос принят! Это важнее, чем может показаться. Золотое правило поисковиков звучит так: 10% запросов вводятся с ошибкой. (Поскольку средство проверки правописания Google постоянно совершенствуется, люди не обращают особого внимания на правильный ввод поисковых запросов, ведь Google в любом случае прекрасно справится с их обработкой.)

Компания Google получила свое средство проверки правописания практически «даром». Оно основано на опечатках, которые вводятся в окне поиска среди трех миллиардов запросов, обрабатываемых ежедневно. Продуманная обратная связь указывает системе, что

пользователь на самом деле имел в виду. Пользователи могут непосредственно «сообщить» поисковой системе Google ответ на вопрос, отображаемый в верхней части страницы результатов (например: «Вы имели в виду *эпидемиология*?»), выбрав новый поиск с правильным термином. Или же веб-страница, на которую переходит пользователь, неявно сигнализирует о правильном написании, так как она, вероятно, сильнее коррелирует с правильно написанным словом, чем неправильным.

Система проверки правописания Google демонстрирует, что «плохие», «неправильные» или «дефектные» данные могут быть очень полезными. Интересно, что компания Google не первая загорелась этой идеей проверки правописания. Примерно в 2000 году Yahoo увидела возможность создания средства проверки правописания по опечаткам в запросах пользователей. Но идея не была реализована. Данные старых поисковых запросов рассматривались по большей части как балласт. Популярные когда-то поисковые системы Infoseek и Alta Vista в свое время тоже располагали наиболее полной базой данных слов с ошибками, но недооценили ее значимость. Их системы в ходе процесса, невидимого пользователям, рассматривали опечатки как «связанные термины» и выполняли поиск. Но эти системы были основаны на словарях (которые явно указывали системе, что правильно), а не на живом, динамичном взаимодействии с пользователем.

Только Google удалось разглядеть в отрывочных данных о взаимодействии пользователей поистине золотой песок, который можно было собрать и превратить в драгоценный слиток. Как считает один из ведущих инженеров Google, их средство проверки правописания работает на порядок лучше, чем средство Microsoft (хотя при некотором давлении инженер признал, что не проводил надлежащего исследования). Он высмеял идею «бесплатной» разработки. «Сырье (опечатки), возможно, и дается даром, но у Google на разработку системы ушло наверняка намного больше средств, чем у Microsoft», — сказал он, широко улыбаясь.

Разные подходы двух компаний чрезвычайно показательны. Корпорация Microsoft видела ценность средства проверки правописания только в одном — обработке текстов. Google, напротив,

ясно понимала его значение. Используя опечатки, она не только разработала передовое в мире средство проверки правописания, чтобы улучшить поиск, но и применила его ко многим другим службам, таким как «автозаполнение» в поисковой системе, Gmail, Google Диск и даже собственная система машинного перевода.

Для описания цифрового следа, который пользователи оставляют на сайте, был придуман специальный термин — «выбросы данных». Под ним подразумевается побочный продукт взаимодействия пользователей в интернете: где и что они нажимают, как долго смотрят на страницу, где проводят курсором мыши, что печатают и т. д. Многие компании разрабатывают собственные системы, для того чтобы собирать выбросы данных и перерабатывать их для улучшения существующей службы или разработки новой. В этом отношении, как ни странно, лидирует Google. Она применяет принцип рекурсивного «обучения на основе данных» во многих своих службах. Каждое действие пользователя считается «сигналом», который Google анализирует и передает обратно в систему.

Google четко знает, сколько раз пользователи искали тот или иной термин, а также другие связанные с ним термины или же переходили по ссылке, после чего (не найдя ничего ценного) возвращались на страницу поиска, чтобы начать заново. Компания знает, по каким ссылкам переходил пользователь (будь то восьмая ссылка на первой странице или первая ссылка на восьмой странице) и отказался ли он от поиска в целом. Возможно, Google и не была первой, у кого возникла такая идея, зато она реализовала ее с необычайной эффективностью.

Такая информация очень ценна. Если множество пользователей выбирают результат поиска в нижней части страницы результатов, система предположит, что он более актуален, и алгоритм ранжирования Google автоматически поместит его выше на страницах последующих поисков (то же самое относится к рекламным объявлениям). «Нам нравится учиться у больших, “шумных” наборов данных», — делится один из сотрудников Google[82].

Выбросы данных — это механизм, лежащий в основе многих компьютеризированных служб, таких как распознавание голоса, спам-фильтры, переводчики и других. Когда пользователь указывает в

программе распознавания голоса, что она неправильно поняла произнесенное слово, он, по сути, «тренирует» систему, совершенствуя ее.

Многие компании начинают подобным образом проектировать собственные системы сбора и использования информации. В начале деятельности компании Facebook ее специалисты по обработке данных изучили широкую базу выбросов данных и обнаружили, что пользователь чаще всего предпринимает то или иное действие (публикует материал, нажимает значок и пр.) по примеру своих друзей. Компания сразу модернизировала свою систему так, чтобы почти все действия пользователя становились известными его друзьям, и это вызвало новую волну активности на сайте.

Идея распространилась далеко за пределы интернет-сектора — в каждую компанию, у которой есть возможность собирать данные обратной связи с пользователем. Устройства для чтения электронных книг записывают большие объемы данных о литературных предпочтениях и привычках людей, которые ими пользуются: как быстро они читают страницу или раздел, пролистывают ли некоторые страницы, едва прочитав, или, может, вовсе не дочитывают книгу. Книги фиксируют, если читатели подчеркивают отрывки или делают заметки на полях. Возможность собирать такого рода информацию превращает чтение, которое долгое время считалось сугубо индивидуальным, в коллективную деятельность. Объединенные выбросы данных расскажут издателям и авторам то, что им ни за что не удалось бы узнать с помощью количественных измерений: предпочтения людей и свойственные им модели чтения. Это коммерчески ценная информация: компании — производители электронных книг могут продавать ее издателям для улучшения содержания и структуры книг. Компания Barnes & Noble проанализировала данные со своих устройств для чтения электронных книг Nook, в результате чего выяснила, что люди, как правило, забрасывали чтение длинных книг научного содержания на полпути. Это открытие вдохновило компанию на создание Nook Snaps — коротких тематических выпусков, посвященных актуальным вопросам, таким как здоровье и текущие события[83].

Программы дистанционного обучения, такие как Udacity, Coursera и edX, отслеживают взаимодействия студентов в интернете, чтобы определить наиболее удачные педагогические подходы. «Вместимость» аудитории порой превышает десятки тысяч студентов, что обеспечивает чрезвычайно большой объем данных. Теперь профессора могут увидеть, что многие студенты повторно просмотрели тот или иной отрывок лекции, и предположить, что определенный момент в ней был непонятен. Профессор Стэнфордского университета Эндрю Нг, преподавая курс машинного обучения в рамках программы Coursera, отметил, что около 2000 студентов неправильно поняли вопрос в домашнем задании, но выдали совершенно одинаковые ответы. Очевидно, они все делали одну и ту же ошибку. Но какую?

Проведя небольшое исследование, Эндрю понял, что студенты изменили порядок алгебраических уравнений в алгоритме. Впредь, если другие студенты сделают ту же ошибку, система не просто сообщит им, что что-то не так, но и посоветует проверить вычисления. Система также работает с большими данными, анализируя каждое сообщение на форуме, прочитанное студентами, и правильность выполненного ими домашнего задания. Это позволяет спрогнозировать вероятность того, что студент, прочитавший то или иное сообщение, правильно решит задание, а значит, определить какие сообщения наиболее полезны. Все это невозможно было узнать прежде. И эти знания могут навсегда изменить подход к преподаванию.

Выбросы данных могут дать компаниям огромные конкурентные преимущества, а также стать мощным рыночным барьером для конкурентов. Возьмем новую компанию, которая разработала интернет-магазин, социальную сеть или поисковую систему, намного лучшую, чем современные лидеры в этих областях — Amazon, Google или Facebook. Новой компании будет трудно конкурировать не только из-за отсутствия эффекта масштаба, сетевой выгоды или бренда, а еще и потому, что эффективность лидирующих компаний во многом связана с выбросами данных, собранными при взаимодействии с клиентами и включенными обратно в службу. Сможет ли новый сайт дистанционного обучения предложить ноу-хау, способное

посоревноваться в эффективности с теми, кто уже собрал гигантское количество данных, чтобы определить наиболее успешные подходы?

Ценность открытых данных

Считается, что сайты вроде Google и Amazon были первопроходцами в области больших данных, но это не так. Первоначальными сборщиками информации в массовом масштабе были государственные органы, и они по-прежнему дадут фору любой частной компании в том, что касается огромного объема управляемых данных. В отличие от держателей данных в частном секторе, государственные органы, как правило, обязывают людей предоставить информацию, а не убеждают или предлагают что-то взамен. Поэтому они и дальше будут собирать и накапливать огромные объемы данных.

Уроки больших данных применимы как к общественным, так и к коммерческим структурам; ценность данных правительственных структур по большому счету скрыта и может быть извлечена только путем инновационного анализа. Несмотря на преимущественное положение в этом отношении, государственные органы, как правило, не умеют эффективно ими распоряжаться. В последнее время стала популярной мысль о том, что лучший способ извлечь ценность из правительственных данных — предоставить эту задачу частному сектору и обществу в целом. И эта идея небезосновательна. Когда государство собирает данные, оно делает это от имени своих граждан и, следовательно, должно предоставить доступ к ним обществу, за исключением ограниченного числа случаев, связанных, например, с возможностью нанести вред национальной безопасности или правам на частную жизнь других людей.

Эта идея привела к несчетному количеству проектов «открытых государственных данных» по всему миру. Утверждая, что государственные органы являются лишь хранителями собираемой информации, а частный сектор и общество найдут ей инновационное применение, сторонники открытых данных призывают официальные органы открыто публиковать данные в общественных и коммерческих целях — разумеется, в стандартизированной форме, пригодной для

машинного считывания и обработки, иначе эту информацию можно будет назвать общедоступной только номинально.

Идея открытых государственных данных получила развитие, когда Барак Обама в свой первый полный рабочий день 21 января 2008 года издал президентский указ, обязывающий руководителей федеральных агентств выпускать как можно больше данных. «Перед лицом сомнений открытость имеет приоритетное значение», — наставлял Обама[84]. Это блестящее заявление, особенно в сравнении с мнением его предшественника, который поручил агентствам делать прямо противоположное. По указу Обамы был создан сайт data.gov — хранилище общедоступной информации от федерального правительства. Сайт стремительно вырос с 47 наборов данных в 2009 году до почти 450 000, получаемых из 172 агентств, к своему трехлетию в июле 2012 года.

Значительный прогресс достигнут даже в сдержанной Великобритании, где бо́льшая часть государственной информации защищена авторским правом, принадлежащим короне, а получение лицензии на ее применение (например, почтовых индексов для интернет-компаний на карте) — трудоемкий и дорогостоящий процесс. Правительство Великобритании издало указы для поощрения открытости информации и поддержки в создании Института открытых данных (одним из руководителей которого стал Тим Бернерс-Ли, изобретатель всемирной паутины WWW), чтобы содействовать новейшим способам использования открытых данных и высвободить их из цепких рук государства.

Европейский союз объявил инициативы относительно открытых данных, которые вскоре могут приобрести континентальный масштаб. Некоторые страны других континентов, такие как Австралия, Бразилия и Чили, уже выпустили и реализовали стратегии открытых данных. Помимо национального уровня растет число городов и муниципалитетов по всему миру, которые также приняли открытые данные. Не отстают от них и международные организации, включая Всемирный банк, который открыл сотни наборов данных экономических и социальных показателей, доступ к которым ранее был ограничен.

Тем временем вокруг данных сформировались сообщества веб-разработчиков и передовых «умов», стремящихся выяснить способы получения максимальной отдачи от данных, например Sunlight Foundation в США и Open Knowledge Foundation в Великобритании.

Одним из первых примеров возможностей использования открытых данных является американский сайт FlyOnTime.us. Он позволяет в интерактивном режиме узнавать, среди прочего, вероятность того, что ненастная погода приведет к задержке рейсов в конкретном аэропорту. Сайт объединяет информацию о рейсах и о погоде из официальных источников данных, которые находятся в свободном доступе в интернете. Его разработали сторонники открытых данных, чтобы наглядно показать полезность информации, которую накопило федеральное правительство. Кроме того что данные общедоступны, исходный код сайта тоже открыт, так что другие могут учиться на его примере, а также использовать его повторно.

FlyOnTime.us дает возможность данным «говорить», и они нередко сообщают неожиданные факты. Например, на сайте можно увидеть, что на рейсах из Бостона в нью-йоркский аэропорт Ла Гуардиа задержки из-за тумана длятся вдвое дольше, чем из-за снега. Большинство людей, слоняющихся в зале вылета, вряд ли бы об этом догадались, ведь снег кажется более весомой причиной задержки. Это одно из тех открытий, которые становятся возможными благодаря большим данным. В данном случае понадобилось обработать статистические данные о задержках рейса из Транспортного бюро США, текущую информацию о ситуации в аэропорту из Федерального управления гражданской авиации США, предыдущие отчеты о погоде из Национального управления океанических и атмосферных исследований, а также информацию о погодных условиях в режиме реального времени из Национальной метеорологической службы. FlyOnTime.us показывает, что не обязательно собирать или контролировать информационные потоки, чтобы получать данные и применять их с пользой, как это делают поисковые системы и крупные розничные торговцы.

Оценить то, что бесценно

Измерить ценность данных — как общедоступных, так и закрытых в корпоративных хранилищах — непростая задача. Рассмотрим события пятницы 18 мая 2012 года. В тот день 28-летний основатель Facebook Марк Цукерберг из главного офиса компании в городе Менло-Парк, Калифорния, дал символический звонок к открытию биржи NASDAQ. Отныне крупнейшая в мире социальная сеть, которая могла похвастать тем, что в ней зарегистрирован каждый десятый человек на планете, стала публичной компанией. Пакет акций тут же вырос на 11%, как в большинстве технологических компаний в их первый торговый день. Ожидалось практически чистое удвоение стоимости. Но в тот день произошло нечто странное: акции Facebook начали падать. Оказалось, произошел технический сбой в компьютерах NASDAQ, который временно приостановил торговлю. Но надвигалась более масштабная проблема. Почувствовав неприятности, биржевые андеррайтеры во главе с Morgan Stanley вынуждены были искусственно поддерживать котировки не ниже цены выпуска.

Накануне вечером банки Facebook оценили компанию в 38 долларов за акцию, что в общей сумме составляло 104 миллиарда долларов (для сравнения: это примерно рыночная стоимость компаний Boeing, General Motors и Dell Computers вместе взятых). Сколько на самом деле стоит Facebook? По результатам аудита финансовой отчетности за 2011 год, по которой инвесторы оценивали компанию, активы Facebook составили 6,6 миллиарда долларов. В их стоимость вошли аппаратные средства, патенты и другое материальное имущество. Что касается балансовой стоимости огромных запасов размещаемой информации, которая хранилась в корпоративном хранилище Facebook, она равнялась нулю. Точнее, вообще не была включена. И это притом что, по сути, главным ресурсом компании являются *данные*[85].

Ситуация становилась все более странной. Дуг Лэйни, вице-президент по исследованиям в компании Gartner, которая занимается изучением рынка, еще до первичного размещения акций (IPO) подсчитал, что в период между 2009 и 2011 годами компания Facebook собрала 2,1 триллиона единиц «монетизируемого контента», включая пометки «Нравится», опубликованные материалы, комментарии и пр.

При сопоставлении этих данных с оценкой IPO компании получалось, что каждый элемент, рассматриваемый как отдельная точка данных, стоил около четырех центов. Взглянув на эти результаты под другим углом, можно сделать вывод, что каждый пользователь Facebook (как источник собираемой информации) оценивался в 100 долларов.

Как объяснить огромное расхождение между стоимостью Facebook по стандартам бухгалтерского учета (6,6 миллиарда долларов) и тем, во сколько компанию первоначально оценил рынок (104 миллиарда долларов)? Внятного объяснения, пожалуй, нет. Скорее, существует всеобъемлющее соглашение, что нынешний метод определения корпоративной стоимости — исходя из «балансовой стоимости» компании (по сути, стоимости ее материальных активов) — уже не отражает реальной стоимости компании. Разрыв между «балансовой» и «рыночной» стоимостью (которую компания получила бы на фондовом рынке, будь она скуплена целиком) неуклонно рос на протяжении десятилетий[86]. В 2000 году Сенат США даже провел слушания по вопросам модернизации текущей модели финансовой отчетности, созданной в 1930-х годах, когда информационного бизнеса и не было как такового. Эта проблема затрагивает не только балансовый отчет компании — неспособность правильно оценивать стоимость компании может привести к бизнес-рискам и нестабильности рынка[87].

Разница между балансовой и рыночной стоимостью компании учитывается как «нематериальные активы». Она выросла примерно с 40% стоимости публичных компаний в США в середине 1980-х годов до 75% их стоимости в 2002-м[88]. Это внушительное расхождение. К таким нематериальным активам относятся бренд, талант и стратегия — все, что нематериально и не вписывается в формальную финансово-бухгалтерскую систему. Но все чаще к нематериальным активам относят и данные, которые хранятся и используются в компании.

В целом это означает, что в настоящее время нет эффективного способа оценки данных. В день открытия продажи акций компании Facebook разрыв между ее формальными финансовыми активами и неучтенными нематериальными составил около 100 миллиардов долларов — почти в 20 раз. Немыслимо! Подобные разрывы должны

быть (и будут) устранены, как только компании найдут способы отражать стоимость своих активов данных в балансовых отчетах.

Микроскопические шаги в этом направлении делаются. Руководитель высшего звена одного из крупнейших американских операторов беспроводной связи признался, что его компания осознала огромную ценность своих данных и озадачилась вопросом, следует ли рассматривать их как часть активов компании с точки зрения учета и отчетности по документам установленной формы. Как только юристы компании услышали об этой инициативе, они тут же ее остановили. По их утверждению, учет данных в бухгалтерской книге мог повлечь за собой юридическую ответственность за них, что было не такой уж удачной идеей[89].

Тем временем инвесторы тоже начали обращать внимание на альтернативную ценность данных. Цены на акции компаний, которые располагали данными или с легкостью их собирали, стали расти, в то время как могло наблюдаться падение рыночной цены других, менее удачливых позиций. Для этого совсем не обязательно, чтобы данные официально отображались в балансовых отчетах. Эти нематериальные активы найдут свое отражение в оценках рынков и инвесторов, хоть и не без труда, о чем свидетельствуют колебания цены на акции Facebook в первые несколько месяцев. По мере решения вопросов ответственности и трудностей с бухгалтерским учетом ценность данных почти наверняка станет отображаться в корпоративных балансах в виде нового класса активов.

Как будут оцениваться данные? Рассчитать их стоимость нельзя, просто сложив то, что было получено от первичного использования. Если б*о*льшая часть ценности данных скрыта и будет получена при неизвестном дальнейшем вторичном использовании, придется поразмыслить, как подступиться к их оценке. Подобные трудности возникали с ценообразованием опционов, до того как в 1970-х годах было выведено уравнение Блэка—Шоулза[23], а также при оценке патентов в условиях, когда аукционы, биржи, частные продажи, лицензирование и множество судебных разбирательств медленно формируют рынок знаний. В любом случае установление цены на

альтернативную ценность данных, безусловно, открывает широкие возможности для финансового сектора.

Начать можно с изучения стратегий, которые держатели данных применяют для извлечения ценности. Наиболее очевидная из них — возможность лицензировать данные третьим лицам. В эпоху больших данных акционеры, возможно, предпочтут соглашение, по которому будет выплачиваться процент от стоимости извлекаемых данных, а не фиксированная плата. Так издатели выплачивают авторам и исполнителям роялти — процент от продаж книг, музыки или фильмов. Этот подход также напоминает сделки с объектами права интеллектуальной собственности в области биотехнологий, согласно которым лицензиары могут потребовать роялти с любых последующих изобретений, основанных на их технологии. Таким образом, каждая из сторон имеет основания для максимального повышения ценности, получаемой от повторного использования информации.

Однако поскольку лицензиату не всегда удается извлечь полную альтернативную ценность данных, держатели данных могут отказаться предоставить к ним исключительный доступ. И тогда, пожалуй, «распущенность данных» грозит стать нормой. Таким образом держатели данных смогут подстраховаться.

Появился ряд рынков, желающих поэкспериментировать с различными способами ценообразования данных. Исландская компания DataMarket, основанная в 2008 году, обеспечивает свободный доступ к наборам данных других организаций, таких как Организация Объединенных Наций, Всемирный банк и Евростат, и получает доход на перепродаже данных от коммерческих поставщиков (например, компаний, занимающихся маркетинговыми исследованиями). Стартап InfoChimps, расположенный в Остине (Техас), выступает в роли информационного посредника, то есть площадки, где третьи лица могут делиться своей информацией, платно или бесплатно. Компания дает возможность любому держателю данных продать свои накопленные базы данных, равно как платформа eBay дает возможность людям продавать ненужные им вещи.

Корпорация Microsoft вышла на этот рынок со своим продуктом Windows Azure DataMarket, который призван сосредоточить внимание на высококачественных данных и контролирует размещаемые

предложения, подобно тому как компания Apple контролирует предложения в App Store. Microsoft видит ситуацию следующим образом: специалист по маркетингу, работая над таблицами Excel, может совместить табличные внутрикорпоративные данные с прогнозируемыми данными о росте ВВП, полученными из службы экономического консультирования. Он просто выбирает данные для покупки — и они мгновенно загружаются в соответствующие столбцы на экране.

Никто до сих пор не может сказать, чем обернутся модели оценивания стоимости. Но точно известно, что экономика начинает формироваться вокруг данных. При этом множество новых игроков получат ряд преимуществ, а у старых, вероятно, вдруг откроется новое дыхание.

Суть стоимости данных заключается в их неограниченном повторном использовании — альтернативной ценности. Сбор информации имеет решающее, но не исчерпывающее значение, поскольку существенная часть ценности находится в применении, а не хранении как таковом. В следующей главе мы поговорим о способах потребления данных на практике и компаниях по обработке больших данных, которые только-только выходят на рынок.

Примечания к главе 6

Глава 7

Последствия

В 2011 году в Сиэтле был запущен онлайн-стартап Decide.com с умными алгоритмами и фантастически смелыми амбициями. В целом он задумывался как механизм прогнозирования цен на миллиарды потребительских товаров. Но начать планировалось с относительно малого — всевозможных устройств от мобильных телефонов и телевизоров с плоским экраном до цифровых камер. Компьютеры вытягивали потоки данных с сайтов интернет-магазинов и отправлялись дальше на веб-поиски всевозможной информации о данном продукте и соответствующих цен.

Цены в интернете в течение дня постоянно меняются, динамически обновляясь на основе множества факторов. Поэтому компании приходилось постоянно собирать данные о них. И не только большие данные, но еще и большой «большой текст», так как система должна была анализировать слова, чтобы распознать, снят ли товар с продажи или планировался запуск новой модели, о котором следовало сообщить потребителям, поскольку это влияет на цены[90].

Год спустя Decide.com анализировал четыре миллиона продуктов с помощью более 18 миллиардов наблюдений за ценами. Стартапу удалось определить особенности розничной торговли, которые раньше невозможно было «увидеть», например то, что цены на устаревшие модели могут временно подняться, как только в продажу поступят новые модели. Большинство людей обратили бы внимание на более старую модель, полагая, что она обойдется дешевле. Но в зависимости от момента, когда они нажмут кнопку «Купить», есть вероятность, что они заплатят даже больше, чем стоит новая модель. Поскольку интернет-магазины все чаще используют автоматизированные системы ценообразования, система Decide.com может определить неестественные, алгоритмические скачки цен и предупредить потребителей. По внутренним подсчетам, точность прогнозов

компании составляет 77%, что позволяет покупателям экономить в среднем 87 долларов на каждом продукте.

На первый взгляд, Decide.com — один из многих перспективных стартапов, который стремится по-новому использовать информацию и честно получать оплату своих усилий. Но не данные делают сайт Decide.com особенным, а то, что он полагается на информацию, полученную по лицензии от сайтов интернет-магазинов, а также «бесплатно» собранную в интернете. Для этого не требуются технические знания: компания не делает ничего такого, что было бы по силам исключительно инженерам, к тому же только тем, которые работают на Decide.com. Несмотря на несомненную важность сбора данных и технических навыков, главное в деятельности Decide.com — идея. Компания мыслит категориями больших данных. Она сумела разглядеть возможности раньше других и поняла, какие данные нужно исследовать, чтобы раскрыть ценные секреты. И если кажется, что у Decide.com есть точки пересечения с Farecast (сайтом по прогнозированию цен на авиабилеты), на то существуют веские основания: оба сайта являются детищами Орена Эциони из Вашингтонского университета.

В предыдущей главе мы рассмотрели, как данные становятся новым источником ценности (в основном за счет так называемой альтернативной ценности) по мере их применения в новых целях. Основное внимание уделялось компаниям, которые собирают данные. Теперь рассмотрим компании, которые используют эти данные, их место в цепочке создания ценности информации, а также значение для организаций и физических лиц как с профессиональной, так и с бытовой точки зрения.

Компании, которые имеют дело с большими данными, можно отнести к одной из групп в зависимости от того, чем они располагают: данными, навыками и идеями.

К первой группе компаний относятся те, что имеют данные или хотя бы доступ к ним, но не обязательно обладают необходимыми навыками, чтобы извлечь из них ценность или придумать, чем они могут быть полезны. Лучший пример — компания Twitter, которая, безусловно, ценит огромный поток данных, проходящий через ее серверы, но предпочла передать их двум независимым компаниям на

правах лицензирования. Вторая группа компаний имеет навыки. Как правило, это консалтинговые компании, поставщики технологий и аналитики, которые имеют специальные знания и выполняют свою работу, но, вероятно, не имеют данных и не настолько изобретательны, чтобы придумать новейшие способы их использования. Так, компания Walmart обратилась к специалистам Teradata (компании по анализу данных) для того, чтобы найти корреляцию между ураганами и продажами Pop-Tarts. К третьей группе компании позволяет отнести способность мыслить категориями больших данных. Яркий пример — Пит Уорден, эксцентричный сооснователь компании Jetpac, которая рекомендует путешествия на основе фотографий, загружаемых пользователями на сайт. Успех некоторых компаний зависит не от данных или ноу-хау. Их главное преимущество — основатели и сотрудники, которые фонтанируют уникальными идеями использования данных, чтобы извлечь из них максимальную пользу.

Прежде компании больше внимания уделяли первым двум элементам: навыкам (которых не хватает) и данным (они в избытке). В последние годы появилась новая профессия — «специалист по обработке данных», сочетающая в себе навыки программиста, дизайнера, специалиста по статистике и инфографике и к тому же рассказчика. Специалистам по обработке данных не нужен микроскоп, чтобы сделать открытие. Их инструмент — базы данных. Консалтинговая компания McKinsey & Company прогнозирует острую нехватку таких специалистов и в настоящее время, и в будущем (об этом очень любят упоминать современные специалисты, чтобы потребовать повышения зарплаты)[91].

Между тем Хэл Вэриэн, главный экономист Google, в шутку называет профессию статистиков «самой сексуальной» работой. «Если вы хотите быть успешным, найдите то, что повсеместно и дешево, и станьте для него незаменимым дефицитным ресурсом. Данные так широкодоступны и настолько стратегически важны, что дефицит представляют собой знания, которые могут извлечь из них пользу, — говорит он. — Вот почему статистики, администраторы баз данных и специалисты по машинному обучению скоро займут невероятно выгодное положение»[92].

Делая акцент на навыках и преуменьшая важность данных, можно добиться лишь кратковременного успеха. По мере развития отрасли нехватка персонала будет ликвидирована, поскольку навыки, которые нахваливал Вэриэн, станут обычным явлением. Существует ошибочное мнение, что, поскольку данные в избытке, они бесплатны или же почти ничего не стоят. Данные являются *важнейшей* составляющей. Чтобы понять почему, рассмотрим разные части «цепочки создания ценности» больших данных и их вероятные изменения со временем; изучим по порядку каждую из групп: держатель данных, специалист по данным и мышление категориями больших данных.

Цепочка создания ценности больших данных

Основная составляющая больших данных — информация, поэтому целесообразно начать с первой группы — держателей данных. Они не обязательно являются создателями исходной базы данных, но в их руках находится доступ к информации и возможность ее использовать либо передать на правах лицензирования другим пользователям, которые сумеют извлечь из нее выгоду. ITA Software, одна из четырех главных сетей бронирования авиабилетов (после Amadeus, Travelport и Sabre), предоставила свои данные компании Farecast для прогнозирования цен на билеты, но самостоятельный анализ не проводила. Почему? ITA работала с данными исключительно по их прямому назначению. В конце концов, продажа авиабилетов — непростая задача, так что анализ не входил в компетенцию компании. Кроме того, у нее не было инновационной идеи (а значит, пришлось бы искать обходные пути вокруг патента Эциони).

Далее, компания решила не менять положение дел ввиду своего места в цепочке создания ценности информации. «Компания ITA уклонялась от проектов, предусматривающих коммерческое использование данных, слишком тесно связанное с доходами авиакомпании, — вспоминает Карл де Маркен, сооснователь ITA Software и ее бывший технический директор. — ITA имела доступ к информации особой важности, которая требовалась для предоставления услуг, и не могла позволить себе поставить их под

угрозу». Вместо этого она осторожно держала данные на расстоянии вытянутой руки, лицензируя их, но не используя. В итоге ITA продала данные за бесценок. Их основная ценность досталась Farecast: клиентам — в виде более дешевых билетов, а сотрудникам и владельцам Farecast — в виде доходов от рекламы, комиссий и, в конце концов, продажи компании[93].

Некоторые компании проницательно устраивались в центре информационных потоков, тем самым получая возможность масштабирования, а также извлечения пользы из данных. Такая картина наблюдалась в сфере кредитных карт. Годами высокая стоимость борьбы с мошенничеством вынуждала многие малые и средние банки отказываться от выпуска собственных кредитных карт и передавать эту функцию большим финансовым учреждениям, размах которых позволял инвестировать в технологии. При этом все сливки доставались компаниям вроде Capital One и MBNA банка Bank of America. Теперь более мелкие банки сожалеют о том, что так расточительно отнеслись к операциям с картами, поскольку это лишило их данных о структуре расходов, которые позволили бы им узнать больше о своих клиентах и продавать им специализированные услуги.

Крупные банки и эмитенты карт, такие как Visa и MasterCard, напротив, заняли тепленькое местечко в цепочке создания ценности информации. Оказывая услуги многим банкам и торговым компаниям, они видели больше операций по своим сетям и делали выводы о поведении потребителей. Их бизнес-модель перешла от простой обработки платежей к сбору данных. Вопрос теперь в том, что они с ними делают.

Компания MasterCard могла бы лицензировать данные третьим лицам для их дальнейшего использования (как это делала ITA), но предпочла анализировать данные самостоятельно. Подразделение MasterCard Advisors объединяет и анализирует 65 миллиардов операций, осуществляемых 1,5 миллиарда держателей карт в 210 странах, чтобы прогнозировать потребительские и бизнес-тенденции. Затем эта информация продается другим компаниям. Среди прочего компания обнаружила, что, если люди заправили автомобиль около

четырех часов дня, в течение часа они, скорее всего, потратят 35–50 долларов в продуктовом магазине или ресторане[94]. Эта информация могла бы пригодиться маркетологу, чтобы начать печатать купоны для близлежащих заведений на обороте бензозаправочных квитанций, выпускаемых в этот период.

Как посредник в информационных потоках MasterCard занимает весьма выгодное положение для сбора данных и получения из них выгоды. Только представьте себе будущее, в котором компании по выпуску платежных карт откажутся от своих комиссий по операциям и будут обрабатывать их бесплатно в обмен на доступ к большему количеству данных, чтобы получать доход от продажи еще более сложной аналитики, выполненной на их основе.

Во вторую группу входят компании, имеющие знания или технологии. MasterCard решила делать все собственными силами. Некоторые не могут сделать окончательный выбор, но часть компаний все же обращаются к специалистам. Например, консалтинговая компания Accenture сотрудничает с компаниями во многих отраслях промышленности для развертывания передовых технологий в области беспроводных датчиков и анализа собираемых ими данных. В 2005 году в ходе пилотного проекта в Сент-Луисе (штат Миссури) в десятке общественных автобусов были размещены беспроводные датчики, контролирующие работу двигателя для прогнозирования поломок и определения оптимального времени для регулярного техобслуживания. Один только вывод, что город может отсрочить плановую замену деталей с пробега в 200–250 тысяч километров до 280 тысяч километров, сэкономил 600 000 долларов на всем автопарке[95]. При этом именно клиент, а не консалтинговая компания собрал плоды ценности данных.

В сфере медицинских данных мы видим поразительный пример того, как внешние технологические компании могут предоставлять полезные услуги. Вашингтонский госпитальный центр в сотрудничестве с Microsoft Research проанализировал свои анонимные медицинские записи (демографические данные пациентов, анализы, диагностика, лечение и многое другое) за последние несколько лет, чтобы узнать, как снизить частоту повторных госпитализаций и

инфекционных заболеваний. Они составляют львиную долю расходов на здравоохранение, поэтому любое снижение их стоимости означало бы огромную экономию.

Методика позволила выявить несколько удивительных корреляций. Одним из результатов был список всех условий, которые увеличивали вероятность того, что выписанный пациент поступит на повторную госпитализацию в течение месяца. Некоторые из этих условий хорошо известны и не имеют простого решения. Так, пациент с застойной сердечной недостаточностью наверняка вернется, поскольку это заболевание трудно поддается лечению. Система выявила еще один неожиданный, но надежный прогностический фактор — психическое состояние пациента. Вероятность того, что человек будет повторно госпитализирован в течение месяца, заметно увеличивалась, если среди исходных жалоб пациента были слова «депрессия» и пр., что указывало на психическое расстройство.

Хотя эта корреляция ничего не говорит о причинности, она предполагает, что надлежащая психологическая помощь пациенту после выписки благотворно скажется и на его физическом здоровье. Это открытие может улучшить качество ухода, уменьшить количество повторных госпитализаций и снизить расходы на медицинское обслуживание. Данная корреляция была выявлена компьютером путем просеивания огромной базы данных, но человеку вряд ли удалось бы ее выявить самостоятельно. Корпорация Microsoft не вмешивалась в управление данными больницы. У нее не было гениальной идеи по их использованию. Да этого и не требовалось. Microsoft просто предложила правильный инструмент — свое программное обеспечение Amalga, чтобы извлечь ценную информацию.

Компании, компетентные в области больших данных, играют важную роль в цепочке создания ценности информации. Twitter, LinkedIn, Foursquare и другие компании имеют горы данных, которые нуждаются в обработке. Компании старого типа (такие как Ford и BP) тоже буквально утопают в данных, по мере того как все больше аспектов их деятельности и продуктов датифицируется. Как держатели данных они полагаются на специалистов в том, чтобы извлечь из них выгоду. Но, несмотря на престиж и солидные названия должностей в духе «ниндзя данных», работа технических экспертов не всегда так

заманчива, как может показаться. Они трудятся в алмазных копях больших данных, получая при этом внушительную зарплату. Но драгоценные камни достаются тем, кто владеет данными.

Третья группа — это компании и частные лица, которые мыслят категориями больших данных. Их сила в том, чтобы видеть возможности раньше других, даже если у них нет навыков и данных на реализацию. Возможно, именно нехватка этих ресурсов позволяет им взглянуть на ситуацию со стороны. Их разум не обременен стандартными ограничениями, и они видят то, чего можно достичь, пусть это практически трудноосуществимо.

Брэдфорд Кросс — живое олицетворение того, что значит мыслить категориями больших данных. В августе 2009 года в свои двадцать с лишним лет он и его четверо друзей создали FlightCaster.com. Как и FlyOnTime.us, их служба прогнозировала вероятность задержки рейсов в США, анализируя данные обо всех рейсах за последнее десятилетие и сопоставляя их со статистическими данными о прошлых и текущих погодных условиях.

Примечательно, что этого не сделали держатели данных. Никто не обнаружил желания или нормативно-правовой инициативы использовать данные таким образом. Ведь если бы источники данных — Бюро транспортной статистики, Федеральное управление гражданской авиации и Национальная метеорологическая служба США — осмелились предсказать задержку коммерческих рейсов, Конгресс, наверное, провел бы слушания, и чиновники получили бы по заслугам. Поэтому за дело взялась группа ребят в толстовках и с математическим образованием. Авиакомпании тоже не могли — и не хотели — строить такие прогнозы. Они пользовались преимуществами как можно более неясного положения дел. А прогнозы службы FlightCaster оказались настолько точными, что даже сотрудники авиакомпании стали ими пользоваться: поскольку авиакомпании не объявляют о задержке вплоть до последней минуты, они хоть и являются основным источником информации, но не самым своевременным.

Ребята мыслили категориями больших данных, и это вдохновило их на реализацию идеи: общедоступные данные можно обработать так, чтобы дать миллионам людей ответы на животрепещущие вопросы.

Служба FlightCaster Брэдфорда Кросса стала первопроходцем, но с большим трудом. В том же месяце, когда был запущен сайт FlightCaster (август 2009 года), энтузиасты из команды FlyOnTime.us начали в больших объемах собирать открытые данные, чтобы создать собственный сайт. В конечном счете преимущества, которыми наслаждалась компания FlightCaster, пошли на спад. В январе 2011 года Кросс и его партнеры продали свой стартап компании Next Jump, управляющей программами корпоративных скидок, в которых используются методы обработки больших данных.

Тогда Кросс обратил внимание на другую стареющую отрасль — новостные СМИ, увидев в ней нишу, которую мог бы занять внешний новатор. Его стартап Prismatic объединял и ранжировал контент со всего интернета на основе анализа текста, пользовательских настроек, популярности, связанной с социальными сетями, и анализа больших данных. Важно отметить, что система не делала различий между блогом подростка, корпоративным сайтом или статьей в Washington Post: если контент считался востребованным и популярным (что определялось по частоте просмотров и рекомендаций), он располагался в верхней части экрана.

Служба Prismatic стала отражением нового способа взаимодействия со СМИ, который присущ молодому поколению. Его суть в том, что источник информации не столь важен. И это унизительное напоминание СМИ о том, что общество в целом лучше осведомлено о событиях, чем они сами. Претенциозным журналистам приходится конкурировать с блогерами, которые могут днями не вылезать из своих халатов. Ключевым моментом является то, что служба Prismatic вряд ли появилась бы внутри самой медиаиндустрии, хоть она и собирает множество информации. Завсегдатаям бара Национального клуба печати не пришло в голову повторно использовать данные о потреблении СМИ в интернете. И специалисты по аналитике из Армонка (Нью-Йорк) или Бангалора (Индия) до этого не додумались. Зато Кросс, пользующийся дурной славой аутсайдера с растрепанными волосами и неторопливой речью, сумел предположить, что с помощью данных можно сообщать миру, на что следует обратить внимание, и делать это лучше редакторов New York Times.

Творческие аутсайдеры с блестящими идеями и их способность мыслить категориями больших данных напоминают происходившее на заре интернет-коммерции в середине 1990-х годов. Тогда первопроходцами становились те, кто не был обременен закоренелым мышлением или институционными ограничениями более старых отраслей. Так, хедж-фондовый специалист по статистике Джефф Безос основал книжный интернет-магазин, а разработчик программного обеспечения Пьер Омидьяр создал интернет-аукцион. Заметьте — не Barnes & Noble и Sotheby's. Современные лидеры с таким масштабным мышлением зачастую не располагают данными. Зато при этом у них нет корыстных интересов или финансовых стимулов, которые мешали бы им раскрыть потенциал своих идей.

Как мы уже убедились, бывают случаи, когда компания сочетает в себе сразу несколько характеристик, позволяющих оперировать большими данными. Возможно, Эциони и Кросс оказались впереди благодаря своей сенсационной идее, но кроме нее у них были навыки. Сотрудники Teradata и Accenture тоже времени зря не теряют и время от времени выдают отличные идеи. Прототипы идей по-прежнему помогают оценить роль каждой компании. Операторы мобильной связи, о которых шла речь в предыдущей главе, собирают гигантский объем данных, но испытывают трудности в его использовании. Однако они могут передать эти данные тем, кто сумеет извлечь из них новую ценность. Подобным образом компания Twitter с самого начала передала права лицензирования на свои «пожарные шланги данных» двум другим компаниям.

Некоторые компании располагают всеми инструментами для реализации возможностей, которые дают большие данные. Google собирает информацию (например, об опечатках в поисковых запросах), имеет великолепную идею создать с их помощью лучшее в мире средство проверки правописания и блестяще реализует ее своими силами. Учитывая множество других видов деятельности, компания Google получает выгоду от вертикальной интеграции в цепочку создания ценности больших данных, где она занимает все три позиции. В то же время Google предоставляет открытый доступ к некоторым своим данным через интерфейсы прикладного программирования (API), чтобы из них можно было извлечь

дополнительную ценность. Одним из примеров являются бесплатные карты Google, которые используются в интернете повсеместно — от списков недвижимости до сайтов государственных учреждений (хотя часто посещаемым сайтам все же приходится за них платить).

У Amazon есть и мышление, и знания, и данные. По сути, компания выстраивала свою бизнес-модель именно в таком (обратном по сравнению с нормой) порядке. Вначале у нее была только идея знаменитой рекомендательной системы. В объявлении о новом выпуске акций на фондовой бирже в 1997 году описание «совместной фильтрации» появилось раньше, чем компания Amazon узнала, как эта система будет работать на практике, и получила достаточно данных, чтобы сделать ее полезной.

И Google, и Amazon обладают равными возможностями, но руководствуются разными стратегиями. Приступая к сбору данных, компания Google сразу учитывает возможность их вторичного применения. Например, ее автомобили Street View собирали информацию GPS не только для картографической службы Google, но и для обучения самоуправляемых автомобилей[96]. Amazon, напротив, больше ориентирована на первичное использование данных и обращается к вторичному только в качестве бонуса. Например, ее рекомендательная система опирается на «сигналы» в виде действий пользователя на сайте, но компания ни разу не прибегла к полученной информации для непредусмотренных прогнозов (например, состояния экономики или вспышек гриппа).

Устройства для чтения электронных книг Amazon Kindle могут показать, на какой странице читатели оставили множество примечаний и подчеркнутых отрывков, но Amazon не продает эту информацию авторам и издателям. Маркетологов заинтересовали бы наиболее популярные отрывки, чтобы повысить продажи книг. Авторы хотели бы узнать, на каком месте их выдающихся произведений большинство читателей забрасывают чтение, и улучшить их. Издатели желали бы выявить темы, сулящие очередной бестселлер. Но Amazon оставляет это поле данных невспаханным.

С умом используя большие данные, можно преобразовать бизнес-модель компании и коренным образом изменить способы

взаимодействия с давними партнерами. Один из потрясающих примеров — история о том, как крупному европейскому автопроизводителю удалось перестроить коммерческие отношения с поставщиком запчастей с помощью данных, полученных в рабочих условиях (поскольку пример взят из частной практики аналитика, который занимался обработкой этих данных, мы, к сожалению, не вправе разглашать названия компаний).

Современные автомобили оборудованы чипами, датчиками и программным обеспечением, которые передают технические данные на компьютеры автопроизводителей во время техобслуживания. Типичный автомобиль среднего класса содержит около 60 микропроцессоров, и треть его себестоимости приходится на электронику[97]. Так что автомобили стали подходящими преемниками кораблей, которые Мори называл «плавающими обсерваториями»[98]. Информация о том, как части автомобиля ведут себя в полевых условиях (и повторное объединение такой информации для корректировки), может стать большим конкурентным преимуществом для компаний, которые ею владеют.

В сотрудничестве с внешней компанией по анализу данных автопроизводителю удалось выявить, что датчик обнаружения утечки топливного бака, производимый немецким поставщиком, не справлялся со своей задачей: на каждый правильный сигнал тревоги приходилось 16 ошибочных. Автопроизводитель мог передать эту информацию поставщику и потребовать регулировки. В эпоху более этичных деловых отношений он так и поступил бы. Но автопроизводитель изрядно потратился на аналитическое программное обеспечение, чтобы выявить проблему, и хотел с помощью полученной информации компенсировать часть своих инвестиций.

Итак, он задумался над вариантами. Стоит ли продавать данные? Как их оценивать? Что делать, если поставщик откажется исправлять ситуацию и компания останется с партией бракованных датчиков? К тому же было ясно, что разглашение информации позволит усовершенствовать аналогичные датчики в автомобилях конкурентов. Компания искала хитрый способ улучшить только свои автомобили. Наконец, автопроизводитель придумал. Он нашел способ

усовершенствовать датчик с помощью модернизированного программного обеспечения и запатентовал его. А затем продал патент поставщику, что с лихвой покрыло его расходы на аналитическое программное обеспечение.

Новые посредники данных

Кто получает наибольшую выгоду в цепочке создания ценности больших данных? В наше время — обладатели особого типа мышления и инновационных идей. Как показала эпоха интернет-магазинов, истинного успеха добивается тот, кто имеет преимущество первопроходца. Но это преимущество недолговечно. По мере развития эпохи больших данных другие лица перестроятся на новый тип мышления, и преимущества первопроходцев, условно говоря, пойдут на спад.

Возможно, вся суть ценности — в навыках? В конце концов, золотая жила ничего не стоит, если вы не можете извлечь золото. Однако история вычислительной техники говорит об обратном. Сегодня опыт управления базами данных, наука о данных, аналитика, алгоритмы машинного обучения и пр. пользуются высоким спросом. Но с течением времени, по мере того как большие данные проникают в повседневную жизнь, инструменты становятся все лучше и удобнее, а люди набираются опыта, относительная ценность навыков начинает снижаться. Подобным образом в 1960–1980-х годах навыками компьютерного программирования обладали уже многие. Компании, которые переносят производственные процессы за границу, сумели еще больше снизить ценность базовых навыков программирования. То, что когда-то считалось образцом технической смекалки, теперь лишь двигатель развития беднейших стран. Это не значит, что опыт работы с большими данными не важен. Просто он не является основным источником ценности, поскольку его можно получить из внешних источников.

Сегодня, на ранних этапах развития больших данных, идеи и навыки ценятся выше всего. Но в конечном счете ценность будет заключаться в самих данных. И не только потому, что появится больше способов применения информации, но и потому, что держатели

данных станут выше оценивать потенциал своих активов. В итоге они наверняка вцепятся в них еще крепче и назначат высокую цену за доступ для посторонних. (В продолжение метафоры с золотой жилой: наиболее ценным будет само золото.)

В истории долгосрочного роста выгоды держателей данных есть небольшой, но важный аспект, который стоит упомянуть. От случая к случаю станут появляться «посредники данных», способные собирать данные из нескольких источников, объединять их, а затем применять инновационным образом. Держатели данных не будут этому противиться, поскольку некоторую часть ценности данных можно извлечь только с их помощью.

В качестве примера можно привести Inrix — компанию из Сиэтла, которая занимается анализом дорожного движения. Она объединяет в режиме реального времени геолокационные данные о 100 миллионах автомобилей в США и Европе. Данные поступают от автомобилей BMW, Ford, Toyota и пр., из коммерческих автопарков такси и фургонов для доставки, а также с мобильных телефонов отдельных водителей (здесь следует отметить важную роль бесплатных приложений Inrix для смартфонов: пользователи получают бесплатную информацию о дорожном движении, а Inrix — их координаты). Полученную информацию Inrix объединяет с хронологическими данными о моделях дорожного движения, а также информацией о погоде и других факторах (например, местных мероприятиях), чтобы спрогнозировать плотность дорожного движения. Готовый «продукт» передается на автомобильные системы спутниковой навигации и используется государственными учреждениями и коммерческими автопарками.

Компания Inrix — типичный независимый посредник данных. Она получает информацию от многочисленных конкурирующих марок автомобилей и тем самым создает более ценный продукт, чем они могли бы создать самостоятельно. Каждый автопроизводитель, вероятно, получает сотни тысяч точек данных от автомобилей на дорогах и мог бы использовать их для прогнозирования дорожного движения, но его прогнозы были бы не очень точными или неполными. Качество улучшается по мере увеличения количества данных. Кроме того, таким компаниям может не хватать навыков, ведь

в их компетенцию входит изгибание металла, а не решение задач на распределение Пуассона. Так что у них есть основания поручить эту работу третьей стороне. Кроме того, хотя прогноз дорожного движения имеет большое значение для водителей, вряд ли он как-то влияет на выбор марки автомобиля при покупке. Поэтому конкуренты не против объединения усилий в таком виде.

Конечно, и раньше своей информацией делились многие отрасли, в частности лаборатории страховых компаний и сетевые секторы (например, банковское дело, энергетика и телекоммуникации), где такой обмен имеет важнейшее значение для предупреждения неприятностей; время от времени информацию могут требовать регулирующие органы. Компании по исследованию рынка, а также компании, специализирующиеся на отдельных задачах, таких как аудит тиража газетных изданий, уже десятки лет объединяют отраслевые данные. А некоторые торговые ассоциации считают это главной своей задачей.

Отличие нынешней ситуации в том, что данные выходят на рынок. И кроме основного значения, из данных извлекаются новые формы ценности. Например, информация компании Inrix полезнее, чем может показаться на первый взгляд. Ее анализ дорожного движения используется для оценки состояния местных экономик, поскольку он может дать представление о безработице, розничных продажах и не только. В 2011 году программа восстановлении экономики США начала трещать по швам, несмотря на заявления политиков об обратном. Это быстро выявил анализ дорожного движения: в часы пик на дорогах стало свободнее, что предполагало увеличение безработицы. Inrix продала свои данные в инвестиционный фонд, который с помощью моделей дорожного движения вокруг магазинов крупнейших розничных сетей выявляет объемы их продаж. Фонд использует эти данные для торговли акциями компаний до объявления их квартальных доходов. Согласно корреляции, чем больше автомобилей в районе магазина, тем выше его продажи.

В цепочке создания ценности больших данных стали появляться посредники иного типа. Одним из первых игроков на рынке стала компания Hitwise, впоследствии выкупленная компанией Experian. Hitwise заключала с поставщиками веб-служб сделки на получение

данных об их потоке «кликов» в обмен на дополнительный доход. Данные лицензировались за символическую фиксированную плату, а не как процент от приобретенной от них выгоды. Таким образом, основную часть ценности данных получала Hitwise, выступая в роли посредника. Другой пример — компания Quantcast, которая измеряет интернет-трафик на сайтах, позволяя их создателям узнавать подробнее о демографических данных посетителей, а также их предпочтениях, чтобы лучше нацеливать рекламные объявления. Компания распространяет свой интернет-инструмент бесплатно, позволяя сайтам отслеживать посещения. А взамен Quantcast может просматривать данные, и это помогает ей улучшить нацеливание.

Новые посредники заняли выгодное положение, не ставя под угрозу бизнес-модели держателей данных, с которыми сотрудничают. Одной из таких ниш является реклама, поскольку в ней сосредоточена основная часть данных и существует острая необходимость в их обработке для нацеливания рекламных объявлений. С ростом массовой датификации и по мере того, как в отраслях будет расти понимание, что они взаимодействуют с данными, независимые информационные посредники появятся и в других областях.

Посредники не обязательно являются коммерческими компаниями — среди них встречаются и некоммерческие. В 2012 году несколько крупнейших американских медицинских страховщиков создали институт Health Care Cost Institute. Их совокупный объем данных составил 5 миллиардов претензий (анонимных) от 33 миллионов физических лиц. Совместное использование записей позволило компаниям выявить тенденции, которые невозможно было бы увидеть, имея только собственные, меньшие наборы данных. Оказалось, что в 2008–2009 годах расходы США на медицинское обслуживание росли в три раза быстрее, чем инфляция, но с ярко выраженными отличиями на конкретном уровне: расходы на лечение в отделении неотложной хирургии выросли на 11%, в то время как в учреждениях сестринского ухода они, по сути, снизились[99]. Разумеется, страховщики никогда бы не передали свои ценные данные никому, кроме некоммерческого посредника. Такие организации вызывают меньше подозрений в

корыстных мотивах и могут создаваться с учетом прозрачности и подотчетности.

Множество компаний, имеющих дело с большими данными, наглядно демонстрируют, как меняется ценность информации. Собирая данные о ценах и новостях от партнерских сайтов на условиях распределения доходов, Decide.com получает комиссионные с каждой покупки на сайте, а компании, поставляющие данные, — свою часть прибыли. Это говорит об отраслевом развитии способа работы с данными, ведь в свое время ITA не получала комиссионных с данных, предоставляемых компании Farecast, — только базовый лицензионный сбор. Теперь поставщики данных могут претендовать на более привлекательные условия. Что касается следующего стартапа Орена Эциони, вполне вероятно, что он сам попытается стать поставщиком данных, поскольку ценность опыта постепенно сдает позиции в пользу идей и данных.

По мере того как ценность переходит к тем, кто управляет данными, изменяются и бизнес-модели компаний. У европейского автопроизводителя, заключившего со своим поставщиком сделку по поводу интеллектуальной собственности, была собственная сильная команда, которая занималась анализом данных, но ему пришлось обратиться за помощью к внешнему поставщику технологий. Технологическая компания получила гонорар за свою работу, но основная часть прибыли досталась автопроизводителю. Ввиду открывающихся возможностей технологическая компания изменила свою бизнес-модель таким образом, чтобы делиться с клиентами частью рисков и выгод. Она экспериментировала, работая за более низкую плату в обмен на часть выгод, полученных в результате анализа. (С большой долей вероятности можно утверждать, что в будущем поставщики автомобильных запчастей захотят добавить измерительные датчики в свою продукцию или будут настаивать на внесении пункта о технических данных в договор купли-продажи, чтобы постоянно совершенствовать комплектующие.)

Что касается компаний-посредников, их жизнь усложняется необходимостью постоянно проверять ценность данных, которыми они делятся. Компания Inrix начала собирать не только геолокационную информацию. В 2012 году она провела пробный анализ места и

времени поломок автоматических тормозных систем (АТС) по запросу автопроизводителя, который разработал собственную телеметрическую систему для сбора информации в режиме реального времени. Идея состояла в том, что, если АТС многократно срабатывает на одном и том же конкретном участке дороги, возможно, это связано с опасными условиями и следует рассмотреть альтернативные маршруты. Таким образом, Inrix получила возможность рекомендовать не только кратчайший, но и самый безопасный путь.

Однако в планы автопроизводителя не входило делиться данными с другими, как в случае с информацией GPS. Он настаивал на том, чтобы развертывать системы Inrix исключительно в своих автомобилях. Как видим, не разглашать информацию об этой функции оказалось выгоднее, чем объединить ее с данными других компаний, чтобы улучшить общую точность системы. И все-таки Inrix считает, что со временем все автопроизводители увидят пользу в объединении данных. И у нее есть веские основания придерживаться такой оптимистичной позиции, поскольку бизнес Inrix (как посредника) полностью держится на доступе к нескольким источникам данных.

Компании в области больших данных экспериментируют с различными корпоративными структурами. Inrix не «наткнулась» на свою бизнес-модель, как это часто случается у стартапов, а изначально рассматривала себя как посредника. Microsoft владела важнейшими технологическими патентами, но посчитала, что небольшая независимая компания (в отличие от крупной корпорации, известной своей агрессивной тактикой) будет воспринята более спокойно, сможет примирить конкурентов и получить максимальную отдачу от своей интеллектуальной собственности. Точно так же Вашингтонский госпитальный центр, который пользовался программным обеспечением Microsoft Amalga для анализа повторных госпитализаций пациентов, знал, каким образом употребить данные: первоначально система Amalga была собственным программным обеспечением отделения неотложной хирургии госпиталя и называлась Azyxxi, но в 2006 году она была продана корпорации Microsoft для дальнейшего усовершенствования[100].

В 2010 году компания UPS была продана в качестве штатного подразделения по анализу данных (UPS Logistics Technologies) частной инвестиционной компании Thoma Bravo. Теперь, работая под знаменем Roadnet Technologies, она чувствует себя свободнее и может анализировать маршруты более чем одной компании. Roadnet собирает данные от многих клиентов для предоставления услуг отраслевого сопоставительного анализа как компании UPS, так и ее конкурентам. По словам Лена Кеннеди, исполнительного директора Roadnet, будучи отделом по логистике в UPS, компания ни за что не получила бы доступ к наборам данных конкурентов своей родительский компании. Но Roadnet добилась этого, став независимой: конкуренты UPS начали более охотно предоставлять свои данные. В конечном счете все выиграли от повышения точности, которое стало возможным благодаря объединению данных.

О том, что именно данные, а не навыки или образ мышления станут самыми ценными характеристиками, говорят многочисленные сделки в области больших данных. Наиболее показательный пример: в 2006 году корпорация Microsoft вознаградила Эциони за идею, выкупив Farecast примерно за 110 миллионов долларов. Однако через два года Google заплатила уже 700 миллионов за данные от поставщиков Farecast — ITA Software.

Обесценивание экспертов

В фильме «Человек, который изменил всё» (о том, как бейсбольная команда «Окленд Атлетикс» стала чемпионом, применив аналитику и новые типы измерений) есть замечательные сцены, в которых старые седовласые скауты, собравшись за столом, обсуждают игроков. Зритель невольно съеживается — не только потому, что сцены демонстрируют, как принимаются решения, когда под рукой нет данных, но и потому, что каждый из нас наверняка сталкивался с ситуациями, когда определенность зависела от настроения, а не от науки.

— У него фигура настоящего бейсболиста... хорошая внешность, — говорит один скаут.

— У него отличный замах. От его биты мячи взрываются. Он бьет самым концом биты, да так мощно, что звук сломанной биты разносится по всему стадиону, — вмешивается хрупкий седой старичок со слуховыми аппаратами.

— Ужасный треск. И без усилий, — подтверждает другой скаут.

Третий скаут встревает в разговор:

— У него страшная подружка.

— Ну и что? — спрашивает скаут, ведущий встречу.

— Это признак неуверенности, — констатирует скептик.

— Ясно, — довольно говорит ведущий, готовый продолжить.

После ряда шутливых перепалок в беседу вступает скаут, который до этого отмалчивался:

— У этого парня есть характер, и это очень хорошо. Он из тех парней, которых видно за версту.

Другой добавляет:

— Да, на него приятно посмотреть. Он сыграет на поле заметную роль. Ему только нужно игровое время.

— Я просто говорю, что его подружка на троечку в лучшем случае!

Эта сцена прекрасно показывает недостатки человеческих суждений. То, что считается аргументированной дискуссией, по сути, не имеет конкретных оснований. Решения о заключении договоров с игроками на миллионы долларов принимаются на основе голой интуиции, без учета объективных показателей. Да, это всего лишь кино, но в реальной жизни все бывает столь же глупо. Сцена иронична в силу своей универсальности: такие же пустые рассуждения слышны повсюду — от залов заседаний правления в Манхэттене и Овального кабинета в Белом доме до кафе и обычных кухонь.

Фильм «Человек, который изменил всё», снятый по книге Майкла Льюиса, рассказывает правдивую историю Билли Бина — генерального менеджера «Окленд Атлетикс», который отбросил вековую традицию назначения игроков в пользу математически ориентированного подхода с новой системой показателей. Статистические подходы, такие как «средний уровень», канули в прошлое. На смену им пришли на первый взгляд непривычные суждения об игре, например «процент попадания на базу». Подход, основанный на данных, показал скрытую сторону спорта, которая, как

правило, ускользала от внимания за привычными атрибутами вроде арахиса и попкорна. Главное, чтобы игрок попадал на базу, и неважно, как он это делал — благодаря своей скорости или хитрости. Когда данные показали, что кража баз является неэффективной, со сцены ушел один из самых интересных, но наименее «продуктивных» элементов игры.

На фоне острой полемики Бин закрепил в руководстве метод, известный как «саберметрика» (аббревиатура англ. Society for American Baseball Research — Общество изучения американского бейсбола), который до этого не пользовался особой популярностью. Он бросил вызов догме скамейки запасных, как в свое время гелиоцентрические взгляды Галилея пошатнули авторитет католической церкви. В конечном счете этот метод дал возможность многострадальной команде Бина финишировать первой в Американской лиге сезона 2002 года, выиграв 20 игр подряд. С тех пор статистика вытеснила скаутов как крупных специалистов в спорте, а множество других команд стали усиленно перенимать саберметрику.

Подобным образом большие данные окажут существенное влияние на то, как решения, принимаемые на их основе, будут дополнять или отклонять человеческие суждения. Эксперты в предметной области и основные специалисты утратят часть своего блеска на фоне специалистов по статистике и аналитиков данных, которые не держатся за устаревшие способы ведения дел и позволяют данным «говорить». Эти новые сотрудники будут полагаться на корреляции без предубеждений и предрассудков. Точно так же Мори не принимал за чистую монету все, что умудренные опытом капитаны рассказывали о морских путях за кружкой пива в пабе. Выявляя практические истины, он полагался на объединенные данные. Метод Мори не объяснял, откуда берутся ветры и течения, но для моряков, которые ищут безопасный путь, вопрос *почему* был менее важен, чем *что* и *где*.

Авторитет экспертов в предметных областях ослабевает. Например, в СМИ контент, который создается и публикуется на сайтах, таких как Huffington Post и Gawker, систематически определяется данными, а не исключительно «нюхом» редакторов. Данные лучше, чем чутье опытных журналистов, показывают, что людям хотелось бы прочитать. Coursera, компания по дистанционному обучению, исследует все

собираемые ею выбросы данных (например, какой раздел видеолекции студенты просматривали повторно), чтобы узнать возможные неясные или особенно интересные моменты, которые следует учесть в разработке курсов. Раньше у преподавателей не было такой возможности, но ситуация изменилась и педагогика уже не станет прежней. Как мы упоминали, Джефф Безо уволил штатных редакторов Amazon, когда данные показали, что рекомендации, выявленные алгоритмическим путем, стимулировали больше продаж.

Это означает, что навыки, необходимые для достижения успеха в работе, меняются, как и ожидания, возлагаемые на сотрудников организаций. Доктору Макгрегор, которая занимается проблемами недоношенных детей в Онтарио, не обязательно было становиться лучшим врачом в больнице или главным авторитетом в области наблюдения за беременными, чтобы добиться наилучших результатов в лечении своих пациентов. У нее даже нет медицинского образования, разве что степень доктора в области компьютерных наук. Но она поставила себе на службу данные о пациентах, собранные более чем за десятилетний период, которые обрабатываются компьютером, а затем с ее помощью преобразуются в рекомендации по лечению[101].

Первопроходцы, проявившие себя в сфере больших данных, нередко являются специалистами из других областей: анализа данных, искусственного интеллекта, математики или статистики, которые применяют свои навыки в определенных отраслях. По словам главного исполнительного директора Kaggle Энтони Голдблума, победители конкурсов Kaggle (интернет-платформы для проектов на основе больших данных) редко приходят из сектора, в котором достигли высоких результатов: призовое место занял британский физик, разработавший алгоритмы для прогнозирования претензий по страхованию и выявлению неисправных подержанных автомобилей. Сингапурский страховой статистик победил в конкурсе с проектом прогноза биологических реакций химических соединений[102]. Инженеры отдела по машинному переводу Google отмечают свой успех в переводах на языки, которых никто из них не знает, а специалисты по статистике из отдела машинного перевода Microsoft

шутят, что качество переводов улучшается всякий раз, когда команду покидает лингвист.

Разумеется, эксперты в предметных областях не вымрут, но они наверняка утратят свое превосходство. Теперь им придется делить свои лавры со специалистами в области больших данных, а простые корреляции потеснят величие причинно-следственных связей. Это изменит наше отношение к знаниям, ведь мы склонны считать, что люди с узкой специализацией более ценны, чем с широкой: успех сопутствует более глубокому знанию предмета. Экспертные знания, как и точность, подходят для области «малых данных», где вечно не хватает нужной информации, поэтому в поисках правильного пути приходится полагаться на интуицию и опыт. В таких условиях опыт играет важнейшую роль, поскольку только длительное накопление скрытых знаний, которые нельзя передать, вычитать в книгах или даже попросту осознать, может помочь в принятии более взвешенных решений.

Но если у вас нет ничего, кроме данных, из них тоже можно извлечь огромную пользу. Те, кто проанализирует большие данные, увидят всю иррациональность традиционного мышления в прошлом не потому, что умнее, а потому, что имеют данные. (Кроме того, будучи посторонними наблюдателями, они позволят себе оставаться беспристрастными, в то время как эксперты предвзято отстаивают позиции своей предметной области.) Это говорит о том, что ценность сотрудника для компании будет измеряться другими мерками. Изменятся знания, связи и навыки, необходимые для профессиональной деятельности.

Знания в области математики, статистики и, возможно, общее представление о программировании и сетевой науке станут столь же неотъемлемыми требованиями к современным сотрудникам, какими были математическая грамотность столетие назад и общая грамотность в более раннюю эпоху. Ценность сотрудника начнет определяться не только тесными связями с коллегами и единомышленниками, но и широким кругом отношений с людьми целого ряда других профессий, чтобы знания могли циркулировать далеко за пределами исходных областей. Когда-то, чтобы быть превосходным биологом, нужно было знать множество других

специалистов в этой сфере. В этом смысле не многое изменилось. Но теперь, когда большие данные приобрели большое влияние, важна не только глубина опыта в предметной области. Сложную биологическую задачу можно успешно решить и при помощи астрофизика или дизайнера в области визуализации данных.

Видеоигры — одна из отраслей, где «лейтенанты» больших данных уже пробили себе путь локтями, чтобы встать в ряд с «генералами» экспертных знаний, попутно преобразуя саму отрасль. Рыночный сектор видеоигр ежегодно получает 10 миллиардов долларов прибыли, что превышает кассовые сборы Голливуда. Раньше компания разрабатывала игру, выпускала ее на рынок и надеялась, что та станет хитом. На основе данных о продажах компания готовила продолжение или начинала новый проект. Решения относительно темпа и элементов игры (таких как персонажи, сюжет, объекты, события и пр.) зависели от творческой фантазии дизайнеров, которые относились к своей работе с такой же серьезностью, как Микеланджело расписывал Сикстинскую капеллу. Это было искусство, а не наука, мир догадок и интуиции, как у скаутов из фильма «Человек, который изменил всё».

Но эти времена прошли. FarmVille, FrontierVille, FishVille компании Zynga и другие онлайн-игры являются интерактивными. Очевидно, это позволяет Zynga просматривать данные об использовании игр и вносить изменения, руководствуясь реальным опытом игроков. Поэтому, если игроки с трудом переходят с одного уровня на другой или склонны забрасывать игру в определенный момент из-за скуки, специалисты Zynga заметят это по данным и предпримут соответствующие меры. Менее бросается в глаза то, что компания адаптирует игры под особенности отдельных игроков. Так что существует не одна версия FarmVille — их сотни.

Аналитики больших данных в компании изучают, как на увеличение продаж виртуальных товаров влияет их цвет или выбор друзей. Например, когда данные показали, что игроки FishVille покупают полупрозрачных рыб в шесть раз чаще, чем остальных существ, компания Zynga предложила дополнительные разновидности таких рыб и хорошо на этом заработала. В игре Mafia Wars обнаружилось, что игроки охотнее всего покупают оружие с золотой каймой и

белоснежных домашних тигров[103]. Вряд ли разработчики игр, находящиеся в студии, узнали бы об этом сами. Это им подсказали данные. «Мы аналитическая компания, которая работает под видом игровой. Здесь всем заправляют числа», — говорит Кен Рудин, главный аналитик Zynga[104].

Происходит переход на решения, принимаемые на основе данных. Большинство людей приходят к решению, исходя из фактов, рассуждений и, пожалуй, во многом — догадок. «Буйство субъективных точек зрения возникает из ощущений в области солнечного сплетения», — говорится в памятных строках поэта Уистена Одена. Томас Дэвенпорт, бизнес-профессор в Бэбсон-колледже, Массачусетс, и автор многочисленных книг по аналитике, называет это явление «золотым нутром». Руководителям придает уверенность их внутреннее чутье, на которое они и полагаются. Но и здесь не обошлось без изменений: управленческие решения принимаются (или по крайней мере подтверждаются) прогнозным моделированием и анализом больших данных.

The-Numbers.com на основе баз данных и внушительного математического аппарата сообщает независимым голливудским продюсерам вероятный доход от того или иного фильма задолго до того, как отснят первый дубль. База данных компании обрабатывает около 30 миллионов записей о каждом коммерческом кинофильме США за последние десятилетия. Записи содержат сведения о бюджете, жанре, актерском составе, съемочной группе, наградах, доходах (включая американские и международные кассовые сборы, зарубежные права, продажу и аренду видеозаписей) и не только. «Компания разработала карту сети из миллиона взаимосвязей, таких как “этот сценарист работал с этим режиссером; этот режиссер работал с этим актером”», — объясняет основатель и президент компании Брюс Нэш.

The-Numbers.com умеет находить сложные корреляции, которые предсказывают доход от кинопроектов. Продюсеры предоставляют эту информацию студиям и инвесторам, чтобы получить финансовую поддержку. Повозившись с переменными, компания даже может подсказать клиентам, как увеличить их доход (или свести к минимуму

финансовые риски). В одном случае анализ показал, что проект будет иметь больше шансов на успех, если в главной мужской роли снимется актер «А-списка», номинированный на премию «Оскар», с гонораром в 5 миллионов долларов. В другом случае Нэш сообщил студии IMAX, что их проект окупится, только если его бюджет урезать с 12 до 8 миллионов долларов. «Это буквально осчастливило продюсера, чего не скажешь о кинорежиссере», — поделился Нэш.

Таким образом, вырисовывается определенный переход в принятии корпоративных решений (например, стоит ли снимать тот или иной фильм или с каким бейсболистом подписать контракт). Эрик Бриньолфссон, бизнес-профессор Массачусетского технологического института, и его коллеги сравнили показатели тех компаний, которые преуспели в принятии решений на основе данных, и тех, кто не придал этому подходу особого значения. Обнаружилось, что уровень производительности в таких компаниях на 6% выше, чем у тех, кто, принимая решения, не опирается на данные[105]. Такой подход дает значительное преимущество, хотя и кратковременное, поскольку все больше компаний применяют в своей практике подходы на основе больших данных.

Вопрос полезности

Благодаря тому что большие данные для многих компаний превращаются в источник конкурентного преимущества, изменится структура целых отраслей. Однако награды распределятся неравномерно. В выигрыше останутся крупные и мелкие компании, потеснив остальных.

Крупнейшие игроки, такие как Amazon и Google, продолжат расти. Но, в отличие от индустриальной эпохи, их конкурентное преимущество будет опираться на физические масштабы. Огромная техническая инфраструктура их центров обработки данных, несомненно, важная, но не самая значительная характеристика: ресурсы для цифрового хранения и обработки данных можно недорого арендовать всего за несколько минут. Компании могут регулировать необходимое количество вычислительной мощности на основе фактического спроса, тем самым превращая в переменную стоимость

то, что раньше считалось фиксированной. Это подрывает преимущества масштаба на основе технической инфраструктуры, которым уже давно пользуются крупные компании.

Масштаб все еще имеет значение, но его фокус сместился. Теперь важен масштаб данных. Под ним подразумевается наличие больших пулов данных и возможность легко получать еще больше. Таким образом, крупные держатели данных будут процветать, собирая и храня больше «сырых» материалов о своей деятельности, из которых можно извлечь выгоду при повторном использовании.

Задача победителей в области малых данных, равно как и «чемпионов», ведущих свою деятельность вне интернета (например, Walmart, FedEx, Proctor & Gamble, Nestle, Boeing и пр.), состоит в том, чтобы высоко ценить силу больших данных, а также стратегически подходить к сбору и анализу информации. И начинающие, и проверенные временем компании стараются занять в новых бизнес-областях положение, которое позволило бы им записывать огромные потоки данных. Пример тому — «набеги» Apple на мобильные телефоны. До появления iPhone мобильные операторы успели накопить потенциально ценные сведения об абонентах, но не сумели извлечь из них выгоду. Компания Apple, напротив, потребовала указать в своих договорах с операторами, что ей достанется б*о*льшая часть наиболее полезной информации. Собирая данные от десятков операторов по всему миру, Apple получает гораздо более полную картину использования мобильных телефонов, чем любой из операторов сотовой связи. Масштабное преимущество Apple основано на данных, а не на материальных ресурсах.

Большие данные открывают захватывающие возможности для всех. Умные и проворные мелкие игроки извлекут преимущества «масштаба без нагромождений» (цитируя знаменитую фразу профессора Бриньолфссона)[106]. Они обеспечат себе большое виртуальное присутствие при незначительных материальных ресурсах, а также широко внедрят инновационные решения при небольших затратах. И, что немаловажно, лучшие службы по обработке больших данных основаны прежде всего на инновационных идеях, а потому не обязательно требуют больших начальных инвестиций. Данные можно

лицензировать, а не приобретать, проводить анализ на недорогих «облачных» платформах, а расходы на лицензирование покрывать за счет процента от получаемых доходов.

Вполне вероятно, что все это касается не только пользователей данных, но и держателей, которые могут добавить к своим запасам данных веские преимущества (ведь более существенную выгоду обеспечивает только добавочная себестоимость). Во-первых, у держателей данных уже есть инфраструктура для хранения и обработки информации. Во-вторых, объединение наборов данных придает им особое значение. И, наконец, наличие интернет-магазина для получения данных значительно упрощает жизнь пользователей[107]. Более того, может возникнуть радикально новый тип держателей данных — частные лица. Поскольку ценность данных становится все более очевидной, держатели информации, имеющей к ним отношение (включая данные об их покупательских вкусах, предпочитаемых СМИ, о состоянии здоровья и пр.), окажутся в выигрышном положении.

И тогда потребители получат возможности, о которых и не мечтали. Отдельные лица смогут выбирать, кому лицензировать данные и на каких условиях. Конечно, кто-то начнет заламывать цены. А многие наверняка согласятся на повторное использование их данных бесплатно в обмен на лучшее обслуживание (например, точные рекомендации книг на сайте Amazon). Но для массы подкованных в цифровом плане пользователей идея маркетинга и продажи личной информации может стать столь же естественной, как ведение блога, публикация твитов или редактирование статей Википедии.

Для такого развития событий мало изменения взглядов и предпочтений пользователей. В настоящее время лицензирование личных данных было бы слишком трудоемким и дорогостоящим процессом и для пользователей, и для компаний с точки зрения заключения отдельных сделок с каждым из них. Скорее всего, появятся новые посредники, которые будут объединять данные многих пользователей и обеспечивать простой способ лицензирования данных, автоматизируя все операции. При достаточно низких затратах и доверии пользователей к таким посредникам, возможно, сформируется рынок личных данных, а частные лица станут

успешными держателями данных. Такие группы, как ID3, одним из основателей которой является Сэнди Пентлэнд — гуру аналитики личных данных в MIT Media Lab, уже работают над тем, чтобы превратить эту фантазию в реальность.

Пока нет таких посредников и их первых клиентов, пользователи, желающие стать держателями собственных данных, имеют очень скромные возможности. А для того чтобы не утратить их, прежде чем появятся посредники и инфраструктура для преуспевания частных держателей данных, пользователям имеет смысл раскрывать как можно меньше информации.

Для средних компаний большие данные не имеют весомого значения. «Преимущество крупных компаний — в их масштабе, а малых и проворных — в их расходах и инновациях», — утверждает Филип Эванс из Boston Consulting Group, отличающийся прозорливостью в области технологий и бизнеса[108]. Средние компании в традиционных секторах выживают благодаря своему размеру, который обеспечивает преимущества масштаба, но при этом достаточно компактен, чтобы не утратить гибкости, которой нет у крупных игроков. В мире больших данных нет минимального масштаба, по достижении которого компании придется вкладывать средства в производственную инфраструктуру. Пользователи больших данных, которые хотят преуспевать, но при этом оставаться гибкими, обнаружат, что им больше не нужно достигать порогового размера — можно благополучно процветать и при небольшом (или стать частью гиганта в области больших данных).

Большие данные вытесняют средние компании отрасли, заставляя их изменить масштаб (стать крупнее или меньше, но проворнее) или свернуть работу. Многие традиционные секторы — от сферы финансовых услуг до производства фармацевтических препаратов — перейдут на использование больших данных. Это не приведет к исчезновению всех средних компаний во всех секторах, но, безусловно, окажет давление на компании в секторах, особенно склонных к внедрению анализа больших данных.

Большие данные коренным образом изменят конкурентные преимущества стран. В период изобилия инноваций, когда

производство по большей части переместилось в развивающиеся страны, преимущество промышленно развитых стран состоит в том, что они располагают данными и знают, как их применить. Плохая новость: это преимущество не вечно. Когда остальные страны мира сумеют перенять эти технологии, как уже внедрили компьютерные вычисления и интернет, Запад утратит лидерство в области больших данных. Хорошая новость для энтузиастов из развитых стран: большие данные, скорее всего, усилят как сильные, так и слабые стороны компаний. Поэтому те, кто освоил работу с большими данными, смогут не только превзойти конкурентов, но и расширить сферу влияния.

Гонка за лидерство началась. Каждая компания может извлечь пользу из данных, действуя с умом. Так, поисковые алгоритмы Google учитывают выбросы данных пользователей для повышения качества результатов, а немецкий поставщик автомобильных запчастей на основе данных совершенствует свои комплектующие. Информация дает компаниям возможность не только оптимизировать имеющиеся продукты и услуги, но и создавать новые.

Несмотря на радужные перспективы, есть причины для беспокойства. Большие данные обеспечивают все более точные прогнозы об окружающем мире и нашей роли в нем. Мы можем оказаться не готовы к влиянию этих прогнозов на нашу частную жизнь и принятие решений, ведь наши мировоззрение и структура учреждений формировались в условиях дефицита, а не избытка информации. В следующей главе мы прольем свет на темную сторону больших данных.

Примечания к главе 7

Глава 8

Риски

Почти сорок лет, вплоть до падения Берлинской стены в 1989 году, Министерство государственной безопасности ГДР (нем. Ministerium für Staatssicherheit — Stasi (Штази)) шпионило за сотнями тысяч людей. Около ста тысяч штатных сотрудников вели наблюдения с улиц и из окон автомобилей. Они вскрывали письма и заглядывали в банковские счета, прослушивали квартиры и телефонные линии. Они заставляли влюбленных и супругов, родителей и детей шпионить друг за другом, подрывая важнейшие основы доверия между людьми. Итоговые материалы (в том числе не менее 39 миллионов единиц картотеки и 100 километров документов) подробно описывали самые сокровенные аспекты жизни простых людей. В ГДР был достигнут небывало масштабный уровень надзора.

Спустя 20 лет после развала ГДР о каждом из нас собирается и хранится больше данных, чем когда-либо. Мы находимся под постоянным наблюдением: расплачиваясь кредитной картой, общаясь по сотовому телефону или предъявляя номер социального страхования для удостоверения личности. В 2007 году британские СМИ подшучивали, что в радиусе всего 200 метров от лондонской квартиры, где Джордж Оруэлл писал свой знаменитый роман-антиутопию «1984», установлено более 30 камер наблюдения[109]. Задолго до появления интернета специализированные компании, такие как Equifax и Experian, собирали, упорядочивали и делали доступными сотни записей о каждом из около полумиллиарда человек по всему миру[110]. Интернет сделал процесс отслеживания более простым, дешевым и практичным. За нами шпионят не только тайные государственные службы с названиями из трех букв. Amazon отслеживает наши предпочтения в покупках, Google — просматриваемые веб-страницы, а Twitter — мимолетные мысли. Facebook успевает уловить все это сразу, наряду с нашими социальными отношениями.

Поскольку большие данные обещают ценные открытия тем, кто их анализирует, естественно ожидать стремительного увеличения числа тех, кто будет собирать, хранить и повторно использовать наши личные данные. Поскольку стоимость хранения будет так же стремительно падать, а аналитические инструменты — становиться все мощнее, размер и масштаб сбора данных станет расти не по дням, а по часам. Если эпоха интернета поставила под угрозу конфиденциальность, возможно ли, что большие данные усугубят эту проблему? Это ли не темная их сторона?

И не только она. Существенное свойство больших данных заключается в том, что изменение масштаба приводит к изменению состояния. Далее мы покажем, что это значительно усложняет защиту неприкосновенности частной жизни, но при этом ставит и новую задачу: судить и наказывать людей на основе прогнозов больших данных еще до того, как они совершат преступление. Это сводит на нет идею честности, справедливости и свободы воли и отвергает глубокомысленное принятие решений.

Существует еще одна опасность: мы рискуем стать жертвами диктатуры данных, в результате которой станем боготворить информацию и выходные данные анализов, а в конечном счете и злоупотреблять ими. Большие данные являются хорошим инструментом рационального принятия решений, если с ними вдумчиво обращаться. Если же ими орудовать неблагоразумно, они способны превратиться из мощного инструмента в оправдание репрессий, создавая неудобства клиентам и сотрудникам или, что еще хуже, нанося ущерб гражданам.

На кону гораздо больше, чем принято считать. Неспособность управлять большими данными с точки зрения конфиденциальности и прогнозирования или неправильное их толкование чреваты намного более глубокими последствиями, чем нацеливание рекламных объявлений в интернете. XX век буквально пропитан кровавыми примерами того, как данные способствуют ужасным злодеяниям. В 1943 году Бюро переписи населения США передало адреса кварталов американцев японского происхождения (но без названий улиц и номеров, чтобы поддержать иллюзию защиты конфиденциальности) в целях содействия их интернированию. Знаменитыми всеобъемлющими

голландскими записями об актах гражданского состояния воспользовались вторгшиеся нацисты для облавы на евреев. Изначальные пятизначные номера, нанесенные в виде татуировок на предплечья узников в нацистских концлагерях, соответствовали номерам перфокарт IBM Hollerith — комплексной системы учета узников концлагерей; обработка данных дала возможность совершать убийства в промышленных масштабах[111].

Несмотря на информационное мастерство, Штази многое было не под силу. Сотрудникам министерства стоило огромных усилий узнать, кто, куда, когда перемещается и с кем разговаривает. Основную часть этой информации теперь собирают операторы мобильной связи. В ГДР не могли спрогнозировать, кто станет диссидентом. Мы тоже не можем. Но правоохранительные органы начинают использовать алгоритмические модели для того, чтобы вычислять время и место патрулирования, узнавая предполагаемый ход развития событий. При этом риски, связанные с большими данными, соразмерны самим наборам данных.

Парализующая конфиденциальность

Велик соблазн ассоциировать угрозу конфиденциальности с ростом объема цифровых данных, проводя аналогию с системой надзора в антиутопии Дж. Оруэлла «1984». На самом деле ситуация гораздо сложнее. Во-первых, не все большие данные содержат личную информацию. Ее нет в данных датчиков на нефтеперерабатывающих заводах, в данных о работе заводских механизмов, о погодных условиях в аэропортах или о взрывах в канализационных люках. Компаниям BP и Con Edison не нужна была личная информация, чтобы извлечь выгоду из выполняемого ими анализа. По сути, анализ больших данных на основе такой информации практически ничем не угрожает конфиденциальности.

И все-таки основная часть создаваемых сегодня данных и вправду содержит личную информацию. Есть ряд довольно веских оснований для того, чтобы записывать ее как можно больше и хранить как можно дольше, при этом часто используя. Данные могут быть не похожи явным образом на личную информацию, но благодаря обработке

больших данных по ним можно легко проследить обратную связь с их автором.

Некоторые «умные» электросчетчики, которые внедряются в США и Европе, могут собирать от 750 до 3000 точек данных в месяц в режиме реального времени. Это гораздо больше, чем скудный поток информации о совокупном потреблении электроэнергии, который собирает обычный счетчик. Каждый прибор имеет уникальную «подпись нагрузки» при получении электропитания, которая позволяет отличить холодильник от телевизора, а телевизор — от подсветки для выращивания марихуаны. Таким образом, использование электроэнергии раскрывает личную информацию, будь то ежедневные привычки, медицинские условия или противозаконное поведение[112].

Однако не столько важно, увеличат ли большие данные риск нарушения конфиденциальности (а они увеличат), сколько изменится ли сам характер риска. Если угроза просто возрастет, то некоторые законы и правила о неприкосновенности частной жизни подойдут и для эпохи больших данных — потребуется лишь удвоить нынешние усилия. С другой стороны, если ситуация изменится, потребуются новые решения.

К сожалению, проблема все же приобретает новые очертания. Ценность больших данных не ограничивается первичным использованием — существенная ее часть, как мы уже поясняли, состоит во вторичном применении.

Это подрывает главную роль частных лиц в действующем законодательстве о неприкосновенности частной жизни. Сборщики данных должны сообщать им, какую информацию собирают и с какой целью. Чтобы начать сбор данных, сборщикам необходимо получить от частных лиц согласие. Хотя это и не единственный способ обработки личных данных законным путем, понятие «уведомления и согласия» стало краеугольным камнем политики конфиденциальности по всему миру. (На практике это вылилось в огромные примечания о конфиденциальности, которые мало кто читает, не говоря уже о том, чтобы понять, но это уже другая история.)

В эпоху больших данных самые инновационные способы их вторичного использования невозможно было представить на момент

их сбора. Как же компаниям уведомлять о цели, которая еще не придумана? И разве станут частные лица давать информированное согласие на неизвестное? А при отсутствии согласия, для того чтобы анализировать большие данные, содержащие личную информацию, потребуется обращаться к каждому лично, спрашивая разрешение на каждое повторное применение. Вы можете себе представить, как Google пытается связаться с миллиардами пользователей, чтобы получить от них разрешение на анализ их старых поисковых запросов с целью спрогнозировать грипп? Ни одна компания не возьмет на себя такие расходы, даже если бы это было технически возможно.

Альтернативный вариант — перед сбором получать согласие на любое дальнейшее использование их данных — тоже бесполезен. Такое разрешение «оптом» сводит на нет само понятие информированного согласия. В контексте больших данных проверенная временем концепция «уведомления и согласия» налагает слишком много ограничений для извлечения скрытой ценности данных и слишком бесполезна для защиты конфиденциальности частных лиц.

Кроме того, в эпоху больших данных технические способы защиты неприкосновенности частной жизни тоже сдают свои позиции. Если вся информация находится в наборе данных, ее извлечение само по себе может оставить след. Возьмем, к примеру, функцию Google Street View. Для ее создания собрали фотографии дорог и домов во многих странах (как и многие другие данные — но это спорный вопрос). В Германии компания Google столкнулась с массовым протестом общественности и СМИ. Люди опасались, что фотографии их домов и садов помогут бандам грабителей выбрать выгодные цели. Под давлением регулирующих органов Google согласилась предоставить домовладельцам возможность отказа от участия, которая позволяла размыть изображения их домов. Но результаты этой возможности заметны в Street View — вы видите размытые дома, а грабители могут расценить их как сигнал, что это отличная цель.

Такой технический подход к защите конфиденциальности, как анонимизация, тоже, как правило, неэффективен. Анонимизация подразумевает удаление из наборов данных всех личных идентификаторов (имя, адрес, номер кредитной карты, дата рождения,

номер социального страхования и пр.). Полученные данные можно анализировать без ущерба для чьей-либо конфиденциальности. Этот подход работает в мире малых данных. Большие данные упрощают повторное установление личности в связи с увеличением количества и разнообразия информации. Рассмотрим примеры с веб-поисками и оценками кинофильмов, которые, казалось бы, не позволяют установить личность.

В августе 2006 года компания AOL сделала общедоступными горы старых поисковых запросов под благовидным намерением дать исследователям возможность анализировать их в поисках интересных открытий. Набор данных из 20 миллионов поисковых запросов от 650 000 пользователей за период с 1 марта по 31 мая 2006 года был тщательно анонимизирован. Личные данные, такие как имя пользователя и IP-адрес, были удалены и замещены уникальным числовым идентификатором. Таким образом, исследователи могли связать между собой поисковые запросы от одного и того же человека, но не имели информации для установления его личности.

Тем не менее в течение нескольких дней сотрудники New York Times, связав поисковые запросы, такие как «одинокие мужчины за 60», «целебный чай» и «ландшафтный дизайнер в Лилбурне, Джорджия», успешно установили, что пользователь № 4417749 — это Тельма Арнольд, 62-летняя вдова из Лилбурна, штат Джорджия. «О Господи, это же вся моя личная жизнь! — сказала она журналистам Times, когда они наведались к ней в гости. — Я понятия не имела, что за мной подсматривают». Последовавшие за этим протесты общественности привели к увольнению технического директора и еще двух сотрудников AOL.

А всего два месяца спустя, в октябре 2006 года, служба проката фильмов Netflix сделала нечто подобное, объявив конкурс Netflix Prize. Компания выпустила 100 миллионов записей о прокате от около полумиллиона пользователей и объявила приз в размере одного миллиона долларов, который достанется команде исследователей, сумевшей улучшить систему рекомендации фильмов Netflix не менее чем на 10%. Личные идентификаторы были тщательно удалены. И снова пользователей удалось разоблачить: мать и скрытая лесбиянка из

консервативного Среднего Запада подала в суд на Netflix от имени псевдонима Jane Doe[113].

Сравнив данные Netflix с другими общедоступными сведениями, исследователи из Техасского университета быстро обнаружили, что оценки анонимизированных пользователей соответствовали оценкам людей с конкретными именами на сайте Internet Movie Database (IMDb). В целом исследования показали, что всего по шести оценкам фильмов в 84% случаев можно было верно установить личность клиентов Netflix. А зная дату, когда человек оценил фильмы, можно было с 99%-ной точностью определить его среди набора данных из полумиллиона клиентов[114].

В исследовании AOL личности пользователей можно было раскрыть по содержанию их поисковых запросов, а в конкурсе Netflix — путем сравнения с данными из других источников. В обоих случаях компании недооценили, насколько большие данные могут способствовать деанонимизации. Тому есть две причины: мы записываем больше данных и объединяем больше данных.

Пол Ом, профессор права в Университете штата Колорадо и эксперт по ущербу от деанонимизации, объясняет, что этот вопрос не так просто решить. При наличии достаточно большого количества данных идеальная анонимизация невозможна вопреки каким бы то ни было усилиям[115]. Хуже того, исследователи недавно показали, что не только обычные данные, но и «социальный граф» — связи между людьми в социальных сетях — также подвержены деанонимизации[116].

В эпоху больших данных три основные стратегии обеспечения конфиденциальности (индивидуальное «уведомление и согласие», возможность отказа от участия и анонимизация) во многом утратили свою эффективность. Уже сегодня многие пользователи считают, что их частная жизнь находится под угрозой. То ли еще будет, когда практика использования больших данных станет обычным явлением!

По сравнению с ситуацией в ГДР четверть века назад теперь вести наблюдение стало проще, дешевле и эффективнее. Возможность записи личных данных зачастую встроена в инструменты, которые мы используем ежедневно — от сайтов до приложений на смартфоне. Так, «черные ящики», установленные в большинстве автомобилей для

отслеживания активаций подушки безопасности, известны тем, что могут «свидетельствовать» против автовладельцев в суде в случае спора по поводу ДТП[117].

Конечно, когда компании собирают данные для улучшения своих показателей, нам не нужно опасаться слежки и ее последствий, как гражданам ГДР после прослушивания сотрудниками Штази. Мы не попадем в тюрьму, если Amazon узнает, что мы почитываем «красную книжечку» Председателя Мао Цзэдуна, а Google не изгонит нас за то, что мы искали Bing. Компании обладают определенным влиянием, но у них нет государственных правомочий принуждения.

Да, они не применяют таких жестких методов, как Штази, однако компании всех мастей накапливают базы личной информации обо всех аспектах нашей повседневной жизни, делятся ею с другими без нашего ведома и используют ее в неизвестных нам целях.

Не только частный сектор пробует силы в области больших данных. Государственные органы тоже. По данным расследования Washington Post в 2010 году, Агентство национальной безопасности США (АНБ) ежедневно перехватывает и сохраняет 1,7 миллиарда писем электронной почты, телефонных звонков и других сообщений[118]. По оценкам Уильяма Бинни, бывшего сотрудника АНБ, правительство собрало «20 триллионов операций» между американскими и другими гражданами: кто кому позвонил, написал по электронной почте, отправил денежный перевод и т. д.[119]

Для обработки этих данных США строят гигантские центры, такие как здание АНБ в Форт-Уильямс, Юта, стоимостью в 1,2 миллиарда долларов[120]. Все государственные органы, а не только спецслужбы по борьбе с терроризмом требуют больше информации, чем раньше. Когда список данных расширяется, включая сведения о финансовых операциях, медицинских картах, обновлениях статуса в Facebook и пр., их собирается невообразимое количество. Государственные органы не в состоянии обработать столько всего. Так зачем собирать?

Ответ на этот вопрос показывает, как изменился способ наблюдения в эпоху больших данных. В прошлом исследователи крепили щипковые зажимы к телефонным проводам, чтобы получить максимум информации о подозреваемом. Важно было как можно полнее изучить,

что он собой представляет. Сегодня иной подход. Новое мышление (в духе Google и Facebook) состоит в том, что люди — совокупность их социальных отношений, взаимодействий в интернете и связей с контентом. Чтобы полностью изучить человека, аналитикам нужно просмотреть как можно более широкий круг периферических данных — узнать не только с кем он знаком, но и с кем знакомы его знакомые и т. д. Раньше это было технически трудновыполнимо, а теперь — проще, чем когда-либо.

Однако сколько бы опасений ни вызывала способность бизнеса и правительства извлекать нашу личную информацию, в связи с большими данными возникает более актуальная проблема: использование прогнозов в вынесении приговора.

Вероятность и наказание

Джон Андертон, начальник специального полицейского подразделения в Вашингтоне, округ Колумбия, одним прекрасным утром врывается в пригородный дом за считаные секунды до того, как разъяренный Говард Маркс вот-вот вонзит ножницы в тело своей жены, которую он застал в постели с любовником. Для Андертона это всего лишь очередной день профилактики тяжких преступлений. «Как представитель отдела по профилактике преступлений округа Колумбия, — произносит он, — заявляю: вы арестованы по обвинению в будущем убийстве Сары Маркс, которое должно было произойти сегодня...»

Полицейские связывают Маркса, который кричит: «Я *ничего* не сделал!»

Начальный эпизод фильма «Особое мнение» изображает общество, в котором предсказания выглядят настолько точными, что полиция арестовывает частных лиц за еще не совершенные преступления. Людей сажают в тюрьму не за фактические действия, а за предсказанные, даже если на самом деле преступлений не произошло. Причиной тому является не анализ данных, а видения трех ясновидящих. Мрачное будущее, изображенное в фильме, показывает именно то, к каким угрозам может привести неконтролируемый анализ

больших данных: признание вины на основе индивидуальных предсказаний будущего поведения.

Мы уже видим первые ростки. Комиссии по условно-досрочному освобождению в тридцати штатах используют прогнозы, основанные на анализе данных, как фактор при принятии решений, стоит ли освобождать того или иного заключенного. Все чаще правоохранительные органы в Америке — от избирательных участков в Лос-Анджелесе до целых городов, таких как Ричмонд и Вирджиния, — используют «прогностический полицейский контроль», то есть с помощью анализа больших данных выбирают улицы, группы и частных лиц для дополнительной проверки просто потому, что алгоритм указал на них как на более склонных к совершению преступлений.

В Мемфисе программа под названием Blue CRUSH (англ. Crime Reduction Utilizing Statistical History — «снижение преступности на основе статистических данных») предоставляет полицейским относительно точные данные о зонах потенциальной угрозы с точки зрения места (в пределах нескольких кварталов) и времени (в пределах нескольких часов конкретного дня недели). Система, по всей видимости, помогает правоохранительным органам лучше распределять свои ограниченные ресурсы. Согласно одному из подсчетов, с момента создания системы в 2006 году количество основных имущественных и насильственных преступлений снизилось на четверть (хотя, конечно, нет никакой причинно-следственной связи, указывающей на то, что это как-то связано с Blue CRUSH)[121].

В рамках инициативы в Ричмонде, Вирджиния, полиция устанавливает корреляции между данными о преступлениях и дополнительными наборами данных, например датами выплаты зарплат в крупных компаниях города, а также датами местных концертов или спортивных мероприятий. Как показывает практика, они подтверждают, а иногда и уточняют подозрения полицейских о тенденциях в области преступности. Например, полиция Ричмонда давно предполагала, что за оружейными шоу следует резкий рост тяжких преступлений. Анализ больших данных доказал их правоту, но

с одной оговоркой: скачок преступности происходил через две недели после события, а не сразу после него[122].

Такие системы направлены на профилактику преступлений путем их прогнозирования вплоть до выявления частных лиц, которые могут их совершить. Большие данные здесь служат новым целям: с их помощью можно было бы предупреждать преступления. Звучит многообещающе. Разве не лучше остановить человека до совершения преступления, чем наказывать его после? Нам удалось бы избежать трагических происшествий. В итоге выиграли бы не только потенциальные жертвы, но и общество в целом.

Однако это скользкий путь. Если на основе анализа больших данных мы сможем прогнозировать возможных преступников, то вряд ли станем довольствоваться профилактикой преступлений. Вероятно, мы захотим наказать потенциальных виновников. Это вполне логично. Если мы просто вмешаемся, чтобы не допустить незаконные действия, предполагаемый преступник, освобожденный от наказания, может попробовать еще раз. Но мы надеемся удержать его от такой попытки, возлагая на него ответственность за свои действия (в том числе будущие).

Прогноз на основе наказания кажется шагом вперед по сравнению с практикой. Профилактика нездорового, опасного или незаконного поведения является краеугольным камнем современного общества. Мы ограничили условия для курящих, чтобы предупредить рак легких, требуем пристегивать ремни безопасности, чтобы предотвратить жертвы ДТП, и не пускаем на борт самолетов людей с оружием, чтобы не допустить угонов. Все эти профилактические меры ограничивают нашу свободу, но мы готовы их принять как небольшую плату взамен на прогнозирование гораздо большего ущерба.

Во многих случаях анализ данных уже работает на профилактику. С его помощью людей объединяют в группы по общему признаку, а затем соответственно оценивают их. Страховые таблицы свидетельствуют, что мужчины старше пятидесяти склонны к раку простаты. Поэтому, если вы относитесь к этой группе, возможно, вам придется больше платить за медицинскую страховку, даже если вы не больны. Студенты, бросившие вуз, воспринимаются как группа людей,

склонных не погашать кредиты, так что человек без высшего образования может получить отказ в кредите или будет вынужден оплачивать более высокие страховые тарифы. Кроме того, лица с определенными отличительными признаками подвергаются дополнительной проверке при прохождении контроля безопасности в аэропорту.

В современном мире малых данных такая методика получила название «профайлинг» (профилирование). Это поиск характерных ассоциаций в данных с последующим анализом тех, кто подходит под их описание. Это обобщенное правило, которое относится ко всем участникам группы. «Профайлинг» — весомое слово. Оно подразумевает не только дискриминацию в отношении определенных групп, но и при неправильном использовании означает «вину по ассоциации». Профайлинг имеет серьезные недостатки[123].

Используя большие данные, мы можем определять не группы, а конкретных лиц, что избавляет нас от существенного недостатка профайлинга: каждый прогностически подозреваемый превращается в виновного по ассоциации. В мире больших данных человек с арабским именем, рассчитавшийся наличными за билет в одну сторону в первом классе, больше не должен подвергаться вторичной проверке в аэропорту, если остальные данные указывают, что он, скорее всего, не террорист. Благодаря большим данным мы можем избежать ограничений профайлинга — этой смирительной рубашки групповых особенностей — и заменить их более подробными прогнозами на каждого человека.

Роль больших данных в признании виновности частных лиц состоит в том, что, хотя мы делаем то же, что и раньше (профайлинг), но делаем это лучше, тщательнее, с индивидуальным подходом и меньшей дискриминацией. Такой подход приемлем, если целью является предотвращение нежелательных действий. Но он таит в себе огромную опасность, если прогнозы больших данных послужат принятию решений о виновности и наказании за еще не совершенные поступки.

Наказывать исходя из вероятности будущего поведения — значит отрицать саму основу традиционного правосудия, когда сначала

совершается поступок, а затем уже человека можно привлечь к ответственности. В конце концов, думать о противоправных поступках не воспрещается, а вот совершать их — незаконно. Один из основополагающих принципов нашего общества состоит в том, что каждый несет ответственность за свой выбор действия. Если кого-то под дулом пистолета заставили открыть сейф компании, у него не было выбора и, следовательно, он не несет ответственности.

Если бы прогнозы больших данных были совершенными и алгоритмы могли предвидеть наше будущее с абсолютной точностью, мы не имели бы выбора, как поступать в будущем. Мы вели бы себя именно так, как предсказано. Если бы совершенные прогнозы были возможны, они бы отрицали человеческую волю, нашу способность жить свободной жизнью и, по иронии судьбы, из-за отсутствия выбора освобождали бы нас от любой ответственности.

Идеальное прогнозирование невозможно. Анализ больших данных, скорее, дает возможность прогнозировать наиболее вероятное поведение конкретного человека в будущем. Рассмотрим модель больших данных профессора Пенсильванского университета Ричарда Берка. Он утверждает, что эта модель может спрогнозировать, совершит ли убийство заключенный, если его выпустить условно-досрочно на поруки. В качестве исходных данных Берк использует бесчисленные переменные конкретных случаев, включая причину лишения свободы, дату первого преступления, а также демографические данные, такие как возраст и пол. Берк считает, что может прогнозировать будущее поведение с 75%-ной точностью. Что ж, неплохо. Но это также означает, что, если комиссия по условно-досрочному освобождению станет полагаться на анализ Берка, одно из ее четырех решений окажется ошибочным, то есть комиссия напрасно лишит свободы раскаявшихся заключенных либо отпустит на волю будущих убийц.

Основная проблема не в том, что общество подвергается большему риску, чем необходимо. Главная беда в том, что при такой системе мы наказываем людей, лишая их личной свободы, *прежде* чем они сделают что-то плохое. А путем предварительного вмешательства мы никогда не узн*а*ем, что произошло бы на самом деле. Мы не позволяем судьбе вмешаться и при этом привлекаем частных лиц к

ответственности за их возможные поступки, которые мы спрогнозировали. Такие прогнозы невозможно опровергнуть.

Это сводит на нет саму идею презумпции невиновности, которая лежит в основе нашей правовой системы и, по сути, нашего чувства справедливости. Поскольку мы несем ответственность за действия, которых, возможно, никогда не совершим, ответственность за спрогнозированные действия также отрицает способность людей делать нравственный выбор.

Опасность выходит далеко за рамки уголовного правосудия. Она охватывает все случаи человеческих суждений, в которых прогнозы больших данных используются для признания нашей виновности в будущих действиях. Сюда входят дела гражданских судов о совершении проступка по неосторожности, а также корпоративные решения по увольнению сотрудников.

Возможно, с такой системой общество стало бы более безопасным и эффективным, но разрушилась бы существенная часть того, что делает человека человеком, — наша способность выбирать действия и нести за них ответственность. Большие данные стали бы инструментом коллективизации человеческого выбора и отказа от свободы воли в нашем обществе.

Как уже говорилось в предыдущих главах, у больших данных множество преимуществ. И если они превратятся в самое мощное орудие дегуманизации, то не из-за свойственных им недостатков, а из-за того, *что* мы сделаем с прогнозами. Принуждая людей отвечать за спрогнозированные, но еще не совершенные действия, мы полагаемся на прогнозы больших данных, полученные на основе корреляций, и принимаем решения о виновности, которые должны учитывать причинные связи.

Большие данные помогают лучше понять текущие и будущие риски, а также скорректировать свои действия соответствующим образом. Их прогнозы помогают пациентам и страховщикам, кредиторам и потребителям. Но большие данные ничего не говорят о причинности. В отличие от них для признания «вины» — виновности частных лиц — требуется, чтобы подсудимый выбрал то или иное действие. Его решение служит причиной для последующего проступка. Именно потому, что большие данные основаны на непричинных корреляциях,

они непригодны для того, чтобы судить о причинности, а значит, и признавать чью-либо виновность.

Беда в том, что люди настроены смотреть на мир сквозь призму причин и следствий. Таким образом, большие данные находятся под постоянной угрозой неправильного использования — в целях установления причинности или подкрепления наших наивных предположений о том, насколько эффективнее стал бы процесс принятия решений о признании виновности, если бы мы вооружились прогнозами больших данных.

Это скользкий путь в мир, изображенный в кинофильме «Особое мнение», в котором индивидуальный выбор и свобода воли ликвидированы, личный моральный компас заменен интеллектуальными алгоритмами, а частные лица беспрепятственно подвергаются коллективному суду. В таких условиях большие данные угрожают сделать нас заключенными (возможно, в буквальном смысле) в рамках вероятностей.

Диктатура данных

Большие данные бесцеремонно вторгаются в частную жизнь и угрожают свободе, создавая для нас невиданные риски. При этом они усугубляют старую проблему — привычку полагаться на цифры, в то время как они гораздо более подвержены ошибкам, чем мы думаем. Пожалуй, наиболее яркий пример того, как последствия анализа данных могут завести в тупик, — история Роберта Макнамары.

Макнамара был мастером по части чисел. Будучи назначенным министром обороны США в период напряженности во Вьетнаме в начале 1960-х годов, он настаивал на повсеместном внедрении данных. Макнамара считал, что только применение статистической строгости поможет ответственным лицам, принимающим решения, понять сложную ситуацию и сделать правильный выбор. Мир, по его мнению, представлял собой массу непокорной информации, а если ее определить, обозначить, разграничить и количественно измерить, ее можно приручить и подчинить своей воле. Макнамара искал Истину в данных. Среди цифровых данных, которые обернулись против него, был «подсчет убитых».

Макнамара развил свою любовь к числам, еще будучи студентом Гарвардской школы бизнеса, а затем стал самым молодым доцентом — в 24 года[124]. Он применил свои навыки во время Второй мировой войны в составе элитной группы военного министерства США «Статистическое управление», которая внедрила процесс принятия решений на основе данных в крупнейшую бюрократическую систему в мире. До этого военный сектор был слеп. Ему не были известны, например, тип, количество и расположение запасных частей самолета. Одно лишь проведение комплексной инвентаризации в 1943 году сэкономило 3,6 миллиарда долларов[125]. Условием современной войны стало эффективное распределение ресурсов. Работа группы имела ошеломительный успех.

По окончании войны группа решила держаться вместе и применить свои навыки в интересах американских корпораций. В то время компания Ford испытывала некоторые трудности. Отчаявшись, Генри Форд II передал участникам группы вожжи правления. Они ничего не смыслили в военном деле, когда помогли выиграть войну, и были столь же невежественны в производстве автомобилей. Тем не менее «вундеркиндам» удалось изменить деятельность компании к лучшему.

Макнамара быстро поднялся по служебной лестнице, показывая точки данных по каждой ситуации. Задерганные руководители завода предоставляли все числа, которые он требовал, будь они правильными или нет. Когда вышел указ, предписывающий до начала производства новой модели израсходовать все имеющиеся детали старой, руководители линейных подразделений с раздражением просто сбрасывали лишние части в ближайшую реку. Руководство в штаб-квартире Ford одобрительно кивнуло, получив от заводских мастеров цифры, подтверждающие, что распоряжение было выполнено. А на заводе стали шутить, что теперь можно ходить по воде — из нее торчали ржавые части автомобилей 1950 и 1951 годов[126].

Макнамара был воплощением типичного руководителя середины XX века — рационального управленца, который полагался на числа, а не настроения и мог применить свои навыки для количественного измерения любой заинтересовавшей его отрасли. В 1960 году он был назначен президентом Ford и занимал эту должность всего несколько

недель, прежде чем президент Кеннеди назначил его министром обороны.

Когда обострился вьетнамский конфликт и США направили дополнительные войска, стало ясно, что это война характеров, а не территорий. Стратегия Америки заключалась в том, чтобы усадить Вьетконг[24] за стол переговоров. По этой причине военные успехи измерялись количеством убитых врагов. Эти данные публиковались в газетах и использовались как аргумент сторонниками войны, а для критиков служили доказательством их безнравственности. Подсчет убитых стал точкой данных, определившей новую эпоху.

В 1977-м, спустя два года после того, как последний вертолет поднялся с крыши посольства США в Сайгоне, отставной генерал армии Дуглас Киннард опубликовал масштабный опрос генералов под названием The War Managers[127]. Он показал трясину, в которой погрязло количественное измерение. Всего 2% американских генералов считали, что подсчет убитых был верным способом измерения военных успехов. Две трети сказали, что цифры часто были завышены. «Поддельные и совершенно бесполезные», — писал один из генералов в своих комментариях. «Нередко откровенно лживые», — считал другой. «Они были многократно преувеличены в основном из-за невероятного интереса со стороны таких людей, как Макнамара», — делился третий.

Подобно тому как заводские мастера Ford сбрасывали детали двигателей в реку, младшие офицеры порой подавали своему начальству внушительные цифры, чтобы сохранить свое место или продвинуться по службе. Они сообщали то, что начальство хотело услышать. Макнамара и его окружение полагались на цифры, буквально боготворя их. С превосходно уложенными волосами и безукоризненно завязанным галстуком, Макнамара чувствовал, что может понять то, что происходит на земле, только уставившись в таблицу — на все эти стройные ряды и столбцы, расчеты и графики, овладев которыми он, казалось бы, станет на одно стандартное отклонение ближе к Богу.

Использование данных и злоупотребление ими американскими военными во время войны во Вьетнаме свидетельствуют о том,

насколько ограниченной является информация в эпоху «малых данных». Этот урок необходимо усвоить, поскольку мир вступает в эпоху больших данных. Исходные данные могут быть низкого качества или необъективными. Их можно неправильно использовать и анализировать. Но, что хуже всего, данные могут не отражать то, что призваны количественно измерить.

Мы более уязвимы перед лицом «диктатуры данных», чем можем себе это представить, позволяя данным управлять нами как во благо, так и во вред. Угроза состоит в том, что мы бездумно позволяем связывать себе руки результатами анализов данных, даже если есть разумные основания полагать, что в них что-то не так. Еще один пример — одержимость собирать факты и числа просто ради данных или безосновательно оказывать им чрезмерное доверие.

Ввиду массовой датификации первое, к чему стремятся политики и бизнесмены, — получить как можно больше данных. «Мы верим в Бога — остальное дело за данными» — вот мантра современного руководителя, которая эхом разносится по офисам Кремниевой долины, заводским цехам и коридорам мэрии. Большие данные могут стать кладом в заботливых руках. Но неразумное обращение с ними чревато жуткими последствиями.

Образование катится вниз? Введите стандартизированные тесты для измерения результативности и примените санкции к учителям и школам, которые не дотягивают до нужного уровня. И если тесты и вправду могут оценить способности школьников, то вопрос о качестве преподавания или потребности в творческой, гибкой, современной рабочей силе остается открытым. Но данные не берут это в расчет.

Хотите предотвратить терроризм? Создайте многослойные списки людей для обязательного досмотра или запрета на вылет, чтобы обеспечить охрану порядка в небе. Впрочем, защита, которую такие списки предлагают, весьма сомнительна. Известен случай, когда сенатор от штата Массачусетс Тед Кеннеди, случайно попавший в список, был задержан и подвержен обыску только потому, что его имя и фамилия совпали с именем и фамилией другого человека в базе данных.

У тех, кто имеет дело с данными, в ходу выражение, отражающее суть некоторых проблем: «Мусор на входе — мусор на выходе».

Иногда причина в низком качестве исходной информации, но чаще — в злоупотреблении самим анализом. Из-за больших данных эти проблемы могут возникать чаще или с более существенными последствиями.

Вся деятельность компании Google, как уже было показано в этой книге на многочисленных примерах, построена на данных. Несомненно, они обусловили значительную долю успеха компании. Однако время от времени они же приводят ее к промахам. Сооснователи Google Ларри Пейдж и Сергей Брин длительное время запрашивали от соискателей их балл по тесту SAT (англ. Scholastic Assessment Test — «академический оценочный тест») при поступлении в колледж, а также средний балл при выпуске. Пейдж и Брин рассуждали так: первый показатель отражает потенциал кандидата, а второй — его достижения. Таким образом, состоявшиеся руководители в возрасте 40 лет, которые рассматривались на ту или иную должность, к своему откровенному недоумению, могли быть отсеяны из-за недобора баллов. Компания еще долгое время продолжала требовать эти цифры даже после того, как ее внутренние исследования показали, что между баллами и эффективностью работы нет корреляций[128].

Google следовало бы лучше знать, как не попасться на удочку ложной прелести данных, ведь показатели практически не оставляют места для изменений в жизни человека. Они не берут в расчет знания помимо академических. Они не могут отразить достоинства людей гуманитарных, а не научных и технических специальностей, где инновационные идеи легче измерить. Одержимость данными в кадровых целях вызывает особое недоумение ввиду того, что сами основатели Google являются выпускниками школ Монтессори, в которых особое внимание уделяется именно обучению, а не оценкам. Кроме того, такой подход повторяет прошлые ошибки американских технологических электростанций, в которых резюме кандидатов ставили выше их способностей. Какими были бы шансы Ларри и Сергея занять руководящие должности в легендарной корпорации Bell Labs, учитывая их незаконченное высшее образование доктора философии? По стандартам Google ни Билл Гейтс, ни Марк Цукерберг не получили бы место, так как не имеют высшего образования.

Зависимость компании от данных порой зашкаливает. Марисса Майер, в то время один из руководителей высшего звена Google, однажды дала задание сотрудникам проверить, какой из 41 оттенка синего наиболее популярен у пользователей, чтобы определить цвет панели инструментов на сайте[129]. Диктатура данных в Google была доведена до крайности и вызвала мятеж.

В 2009 году ведущий дизайнер Google Дуг Боумен уволился в гневе, потому что не выдержал постоянного количественного измерения всего и вся. «Недавно я участвовал в дискуссии по поводу того, какой должна быть ширина границы: 3, 4 или 5 пикселей. Меня попросили обосновать свой выбор. Я не могу работать в таких условиях, — написал он в блоге о своей отставке. — Когда в компании одни инженеры, они все превращают в инженерное решение вопросов. Сводят все к простым логическим задачам. Эти данные в конечном счете становятся костылем, тормозящим движение каждого решения, парализуя компанию»[130].

Гениальность не зависит от данных. Стив Джобс мог бы долгие годы непрерывно совершенствовать ноутбук Mac на основе отчетов об эксплуатации, но он воспользовался своей интуицией, а не данными, чтобы выпустить на рынок iPod, iPhone и iPad. Он полагался на свое шестое чувство. «Знать, чего хотят покупатели, не их забота», — сказал он репортеру, рассказывая, что не проводил исследование рынка перед запуском iPad[131].

В книге «Благими намерениями государства» антрополог Джеймс Скотт из Йельского университета рассказывает о том, как правительства, возводя в культ количественные измерения и данные, в конечном счете скорее ухудшают качество жизни людей, чем улучшают его. Они прибегают к картам для определения преобразований в обществах, но ничего не знают о людях на местах. С помощью огромных таблиц данных об урожаях они принимают решение о коллективизации сельского хозяйства, ничего в нем не смысля. Они берут на вооружение все несовершенные, естественные способы взаимодействия, которыми люди пользовались в течение долгого времени, и подстраивают их под свои нужды, иногда просто ради того, чтобы удовлетворить свое желание привести все к

исчисляемому порядку. Информация, по мнению Скотта, часто служит для расширения возможностей власть имущих[132].

Это диктатура данных с большой буквы. Из-за подобного высокомерия США начали войну во Вьетнаме, руководствуясь, в частности, количеством убитых, а не более разумными показателями. «Вы правы, что не все сложные человеческие ситуации, которые только можно представить, могут быть полностью сведены к линиям на графике, выражены в процентных точках на диаграмме или отражены в цифрах в балансе компании, — произнес Макнамара в 1967 году, в период нарастающих национальных протестов. — Но в действительности все может быть обосновано. И не измерять количественно то, что можно измерить, — все равно что довольствоваться меньшим, чем полный спектр причин»[133]. Если бы только правильные данные использовались должным образом, а не просто почитались за то, что они есть.

В течение 1970-х годов Роберт Макнамара удерживал пост главы Всемирного банка, а в 1980 году стал «голубем мира» — ярым критиком ядерного оружия и сторонником охраны окружающей среды. Позже в результате переоценки ценностей он написал мемуары «Взгляд в прошлое», в которых критиковал образ мышления, стоящий за военными действиями, и собственные решения на посту министра обороны. «Мы были неправы, совершенно неправы», — писал Макнамара, в то время как речь шла о масштабной военной стратегии. Однако по вопросу данных и, в частности, подсчета убитых он остался далек от раскаяния. Макнамара признался, что статистика была «недостоверной или ошибочной». «Но все факторы, которые вы можете подсчитать, вы обязаны подсчитать. Потеря убитыми — один из них...» Он умер в 2009 году в возрасте 93 лет, считаясь человеком умным, но не мудрым.

Соблазнившись большими данными, мы рискуем совершить страшную ошибку, как Макнамара, или настолько сконцентрироваться на данных и власти, которую они сулят, что будем не в состоянии оценить их ограничения. Чтобы наглядно представить эквивалент подсчета убитых в виде больших данных, достаточно снова вернуться к Google Flu Trends. Рассмотрим ситуацию (не такую уж невероятную),

когда смертельный грипп бушует по всей стране. Медицинские работники были бы признательны за возможность в режиме реального времени прогнозировать крупнейшие очаги с помощью поисковых запросов. Они бы знали, где нужна помощь.

Однако во время такого кризиса политические лидеры могут возразить, что знать наибольшие очаги заболевания и пытаться остановить их распространение недостаточно. Они призывают ввести режим всеобщего карантина (а не только для населения в охваченных регионах), по сути, излишнего. Большие данные дают возможность быть адресными и применять карантин только к отдельным пользователям, чьи поисковые запросы в значительной степени коррелируют с гриппом. Таким образом, мы получаем данные о тех, кого нужно изолировать. Федеральные агенты, вооруженные списками IP-адресов и информацией GPS о мобильных устройствах, могут объединить отдельные запросы веб-поиска в карантинные центры.

Может показаться, что это оправданно, однако в корне неправильно. Корреляция не означает причинности. Эти люди могут болеть гриппом, но могут и быть здоровыми. Их необходимо обследовать. В такой ситуации люди стали бы заложниками прогноза. Что еще более важно, они стали бы жертвами апологии данных, которые по самой природе своей не могут отразить информацию такого рода. Суть фактического исследования Google Flu Trends состоит в том, что условия поиска связаны со вспышкой. Но причины тому могут быть совершенно разными: например, сотрудники могли услышать, как кто-то в офисе чихнул, и решили поискать в интернете информацию о том, как защититься, а сами при этом здоровы.

Темная сторона больших данных

Большие данные предоставляют больше возможностей наблюдать за нашей жизнью, во многом упраздняя некоторые правовые средства защиты неприкосновенности частной жизни. Они также сводят на нет эффективность основных технических методов сохранения анонимности. Как и фактическое нарушение правопорядка, прогнозы больших данных относительно отдельных лиц могут повлечь за собой

наказание — однако не за действия, а за склонности. Такое положение дел отрицает свободу воли и унижает человеческое достоинство.

В то же время существует реальный риск того, что, поддавшись магии больших данных, люди станут руководствоваться ими в неподходящих условиях или же слишком полагаться на результаты анализов. Точность прогнозов будет возрастать, а с нею и желание все чаще пользоваться ими, подпитывая, в свою очередь, одержимость данными, раз они имеют такие широкие возможности. Такими были проклятие Макнамары и урок, который можно извлечь из его истории.

Нужно умерить увлечение данными, чтобы не повторить ошибку Икара, который гордился своей технической возможностью летать, но неправильно воспользовался ею и упал в море. В следующей главе мы рассмотрим способы, благодаря которым мы будем управлять данными, а не они нами.

Примечания к главе 8

Глава 9

Контроль

Изменение способов производства информации и взаимодействия с ней поневоле меняет правила самоуправления. А эти изменения, хотим мы того или нет, преобразуют основные ценности, которые общество должно защищать. Вспомним предыдущий наплыв данных, который произошел благодаря печатному станку.

До того как Гутенберг изобрел наборный шрифт (примерно в 1450 году), распространять идеи было нелегко. Книги в основном находились в монастырских библиотеках, строго охраняемых монахами в соответствии с правилами, которые католическая церковь предусмотрительно установила для защиты своего господства. Вне церкви нескольким университетам удалось собрать десятки или, быть может, пару сотен книг. Библиотека Кембриджского университета была основана в XV веке с фондом в 122 тома[134]. Серьезным препятствием на пути распространения информации являлась безграмотность.

Благодаря печатному станку Гутенберга стало возможным массовое производство книг и брошюр. Переведя Библию с латинского языка на немецкий и тем самым открыв ее для многих читателей, которые получили возможность узнать слово Божье без помощи священников, Мартин Лютер мог напечатать и распространить ее среди сотен тысяч людей. Поток информации превратился из ничтожного в огромный. В конечном счете общество установило новые правила для управления информационным взрывом, вызванным изобретением Гутенберга.

Были созданы законы (например, об авторском праве), призванные расширить возможности авторов и дать им правовой и экономический стимул творить. Когда светское государство объединило свою власть, интеллигенция той эпохи стала добиваться установления правил для защиты слова от правительственной цензуры. В итоге свобода слова превратилась в конституционную гарантию. Но, как всегда, права влекут за собой обязанности. По мере того как недобросовестные газеты вторгались в частную жизнь людей или порочили их

репутацию, возникали новые правила, чтобы оградить частную жизнь людей и дать им возможность подать в суд за клевету.

Изменились не только правила. Изменился и уровень доступности информации, что отразилось и на наших ценностях. В эпоху до печатного станка все управление сводилось к тому, чтобы спрятать всю текстовую информацию. Благодаря изобретению Гутенберга мы смогли по достоинству оценить, что значит широкое распространение информации в обществе. Столетия спустя мы предпочитаем получать как можно больше (а не меньше) информации, защищаясь от ее избытка не цензурой, а в первую очередь с помощью правил, ограничивающих злоупотребление информацией.

По мере того как мир движется в сторону больших данных, общество подвергается подобному «тектоническому» сдвигу. Большие данные заставляют нас пересмотреть фундаментальные представления о том, как стимулировать их рост и умерять потенциальный вред, поскольку они во многом меняют наш образ жизни и мышления. Однако, в отличие от печатной революции, на раздумья нам отведены не столетия, а, возможно, всего каких-то пара лет.

Защита частной жизни потребует от лиц, имеющих дело с личными данными, большей ответственности за свою политику и действия. Нам предстоит пересмотреть свое представление о справедливости, чтобы гарантировать человеческое право на свободу действий (и, конечно, соблюдение ответственности за эти действия). Понадобятся новые учреждения и эксперты (так называемые «алгоритмисты»), чтобы интерпретировать сложные алгоритмы, на основе которых формируются выводы из больших данных, и защищать интересы тех, кто может от этих выводов пострадать, например получить отказ в приеме на работу или хирургическом вмешательстве или не получить кредит из-за того, что о них «говорят» большие данные. Дело не в адаптации существующих правил, а в создании новых.

От безопасности к отчетности

На протяжении десятилетий важнейший принцип конфиденциальности во всем мире заключался в том, чтобы предоставить людям возможность самим решать, кто и как имеет

право обрабатывать их личную информацию. В век интернета это достойное правило превращается в шаблонную систему «уведомления и согласия». В эпоху больших данных, когда больше пользы приносит вторичное применение данных, далеко не всегда предсказуемое на момент их сбора, этот принцип уже не так актуален.

Намного разумнее было бы отменить практику индивидуального управления конфиденциальностью и заменить ее расширенной подотчетностью, которая предъявлялась бы к пользователям данных, повышая их ответственность за свои действия. Компании, работающие с данными, больше не смогли бы приводить в свое оправдание то, что человек разрешил их использовать. Напротив, им пришлось бы оценивать потенциальные опасности, с которыми могут столкнуться люди при вторичном применении их данных. И только убедившись, что уровень угрозы низкий (то есть возможный ущерб ограничен или гарантированно может быть снижен), компании могли бы воплощать в жизнь свои планы. А в случае неправильной оценки угроз или небрежной реализации планов компании можно было бы привлечь к ответственности за нанесенный ущерб. В свою очередь, правила должны предусматривать вторичное использование данных в большинстве случаев без явного согласия.

Приведем наглядный пример. Представьте себе, что профессор Косимицу, токийский эксперт по «задней части», продал противоугонное устройство для автомобиля, которое использует сидячую позу водителя в качестве уникального идентификатора. Предположим, что позже он повторно проанализировал полученную информацию, чтобы спрогнозировать уровень внимательности водителя (сонный, подвыпивший, раздраженный и т. п.) и отправить уведомления другим водителям, находящимся поблизости, во избежание аварий. При нынешних правилах конфиденциальности Косимицу потребовалось бы пройти еще один этап «уведомления и согласия», поскольку он ранее не получал разрешения на подобное применение информации. А с системой подотчетности пользователей данных ему достаточно было бы оценить опасности предполагаемого использования и, если они минимальны, продолжить задуманное, тем самым повышая безопасность дорожного движения.

Логично было бы переложить бремя ответственности с общества на тех, кто обрабатывает данные. Тому есть целый ряд причин. Лица, которые обрабатывают данные, гораздо лучше других знают, что с ними будут делать. Их оценка (или оценка нанятых ими экспертов) позволяет избежать проблем с выявлением конфиденциальных бизнес-стратегий. Возможно, самое главное — то, что эти лица получают большую часть преимуществ вторичного использования данных. Так что вполне справедливо привлекать их к ответственности за свои действия.

Безусловно, правительство тоже играет важную роль. Если пользователи данных произведут неточную оценку или будут действовать вразрез с предполагаемой оценкой, регулирующие органы привлекут их к ответственности путем распоряжений, штрафов и, возможно, даже уголовного преследования. Подотчетность пользователей данных должна иметь рычаги влияния. Регулирующие органы могут ей содействовать, например, определив основные категории допустимых видов применения или таких, для которых достаточно ограниченных мер по обеспечению безопасности. Это позволит стимулировать поиск новых приемов повторного использования данных. Для более рискованных инициатив регулирующие органы составят основные правила, по которым пользователи данных должны оценивать опасности, влияние на отдельных лиц и пути сведения к минимуму возможного ущерба. Цель в том, чтобы получить объективное и точное представление об угрозах конфиденциальности и понять, какие меры нужно предпринять.

Далее, с пользователей данных будет снята юридическая обязанность удалять личную информацию сразу после ее основного целевого использования, как того требует большинство нынешних законов о конфиденциальности. Это важное изменение, поскольку, как мы видели, только выявив скрытую ценность данных, современные коммодоры Мори могут максимально эффективно работать с данными для собственной (и общественной) выгоды. Взамен пользователи данных получат право на более длительное, хоть и не вечное хранение информации. Обществу необходимо уравновесить преимущества повторного использования данных и риски, вызванные их слишком широким разглашением.

Для того чтобы достичь такого равновесия, регулирующие органы, например, назначат срок удаления различных видов личных данных. Сроки повторного использования могут зависеть от неизбежного риска, связанного с данными, а также от ценностей, присущих различным обществам. Одни страны будут более осторожными, чем другие, так же как некоторые виды рассматриваемых данных могут быть более конфиденциальными, чем другие: база данных домашних адресов слепых людей в конкретном городе понадобится специалистам по городскому планированию, специализированным розничным магазинам и самим людям, а домашние адреса лиц, больных ВИЧ/СПИДом, относятся к разряду данных, о которых не всем хотелось бы распространяться.

В рамках такого подхода конфиденциальность личных данных защищается ограничением времени, на протяжении которого они могут храниться и обрабатываться. Кроме того, этот подход устраняет угрозу «постоянной памяти» — риск того, что никто не сможет скрыться от своего прошлого, поскольку цифровые записи всегда можно извлечь[135]. В противном случае наши личные данные повисли бы над нами как дамоклов меч, угрожая рано или поздно пронзить нас личными подробностями или напоминанием о неудачных поступках. Сроки также служили бы для держателей данных стимулом реализовать свой ресурс, пока есть такая возможность. На наш взгляд, это позволило бы достичь лучшего равновесия для эпохи больших данных: компании получили бы право дольше использовать личные данные, взяв на себя ответственность за это, а также обязательство удалить с устройства личные данные спустя определенный период.

В дополнение к этому переходу в управлении — от конфиденциальности по согласию к конфиденциальности через подотчетность — нам нужно найти и ввести в действие новые технические способы обеспечения защиты личных данных. Один из инновационных подходов содержит понятие «дифференциальной конфиденциальности», которая подразумевает намеренное размытие данных, чтобы запрос большого набора данных выдавал не точные результаты, а лишь приблизительные. Такой подход делает процесс

связывания определенных точек данных с конкретными людьми трудным и дорогостоящим[136].

Может показаться, что подобное перемешивание информации способно уничтожить ценные открытия. Но это совсем не обязательно или по крайней мере может служить удачным компромиссом. Эксперты в области политики и технологий отмечают, что Facebook использует дифференциальную конфиденциальность, когда сообщает информацию о своих пользователях потенциальным рекламодателям: полученные значения являются приблизительными и поэтому не могут помочь установить личности отдельных людей. Поиск ряда женщин азиатского происхождения, проживающих в Атланте и интересующихся аштанга-йогой, выдаст результат, например, «около 400», а не постоянное количество. Таким образом, информацию невозможно будет статистически свести к конкретному человеку[137].

Переход в управлении конфиденциальностью от согласия отдельных лиц к подотчетности пользователей данных является одним из основных и наиболее существенных изменений. Подобный переход необходим и в прогнозировании на основе больших данных, чтобы сохранить свободу человека и его ответственность.

Люди и прогнозирование

Суды привлекают людей к ответственности за совершенные действия. Когда судья оглашает свое беспристрастное решение после справедливого судебного разбирательства, это считается торжеством справедливости. В эпоху больших данных нам придется пересмотреть понятие справедливости, чтобы сохранить понятие «человеческого фактора» — свободы воли, согласно которой люди сами выбирают, как им действовать. Это простое понятие подразумевает, что люди могут и должны нести ответственность за свое поведение, а не склонности.

До появления больших данных эта фундаментальная свобода была очевидной, причем настолько, что вряд ли нуждалась в формулировке. В конце концов, на ней основан принцип работы нашей правовой системы: мы привлекаем людей к ответственности за свои действия, оценивая то, что именно они натворили. С помощью больших данных мы можем спрогнозировать действия человека, и порой достаточно

хорошо. Это создает искушение судить о людях не по тому, что они сделали, а по тому, что они сделают, судя по нашим прогнозам.

В эпоху больших данных нам придется расширить свое представление о справедливости и включить меры по обеспечению безопасности человеческого фактора, аналогичные тем, которые существуют для защиты процессуальной справедливости. Без этого само понятие справедливости может быть подорвано.

Учитывая человеческий фактор как обязательное условие, мы гарантируем, что органы государственной власти будут судить о нашем поведении исходя из наших реальных действий, а не анализа больших данных. Таким образом, мы должны нести ответственность перед ними только за совершенные действия, а не статистически прогнозируемые в будущем. А судя о предыдущих действиях, органы государственной власти не должны полагаться исключительно на анализ больших данных. Рассмотрим случай, когда две компании подозреваются в ценовом сговоре. К анализу больших данных вполне приемлемо прибегнуть для выявления возможного сговора, поэтому регулирующие органы могут провести расследование и завести дело с использованием традиционных средств. Но эти компании нельзя признать виновными только потому, что, по прогнозам больших данных, они, вероятно, совершили преступление.

Аналогичный принцип должен применяться и вне органов государственной власти, когда компании принимают важные решения о нас: нанять или уволить, предложить ипотеку или отказать в кредитной карте. Если они руководствуются исключительно прогнозами больших данных, необходимо обеспечить определенные меры безопасности. Во-первых, открытость — предоставление данных и алгоритма, лежащих в основе прогноза, который касается конкретного человека. Во-вторых, сертификацию — прохождение сертификации, в ходе которой алгоритм должен быть признан экспертной третьей стороной как обоснованный и достоверный. В-третьих, недоказуемость — определение конкретных путей, с помощью которых человек может опровергнуть прогнозы относительно себя (аналогично традиции в науке раскрывать любые факторы, которые могут подорвать результаты исследования).

Самое главное, гарантия человеческого фактора защищает нас от угрозы «диктатуры данных», когда данным придается больше смысла и значения, чем они заслуживают.

Не менее важно то, что мы защищаем индивидуальную ответственность. Ведь всякий раз, когда общество принимает решение, затрагивающее других, возникает большой соблазн избавить их от ответственности. Общество переходит к управлению рисками, то есть к оценке возможностей и вероятностей потенциальных результатов. При всей видимой объективности данных очень заманчиво звучит идея оградить процесс принятия решений от эмоциональных и личностных факторов, поставив алгоритмы на смену субъективным оценкам судей и оценщиков и формулируя свои решения уже не на языке ответственности, а оперируя категориями более «объективных» рисков и их предотвращения.

Ввиду прогнозов больших данных возникает сильное искушение изолировать людей, которые, судя по прогнозам, склонны к совершению преступлений, и во имя снижения риска регулярно подвергать их тщательным проверкам, даже если они чувствуют (не без оснований), что наказаны без суда и следствия. Предположим, такой алгоритм «охраны правопорядка», основанный на прогнозах, определил, что конкретный подросток в высшей степени склонен к совершению тяжкого преступления в ближайшие пять лет. В итоге по решению властей социальный работник будет ежемесячно наведываться к подростку, чтобы контролировать его и попытаться ему помочь.

Если подросток и его родственники, друзья, учителя или работодатели воспринимают эти визиты как клеймо (что вполне вероятно), то это вмешательство можно оценить как наказание — по сути, штраф за действия, которые никто не совершал. Впрочем, немногим лучше ситуация, если визиты рассматриваются не как наказание, а как простая попытка уменьшить вероятность криминальных событий — так сказать, способ минимизации рисков (в данном случае сводится к минимуму риск совершения преступления, которое подрывает общественную безопасность). Чем чаще привлечение людей к ответственности за свои действия заменяется мероприятиями по снижению рисков, тем больше в обществе

снижается ценность идеала индивидуальной ответственности. Государство, основанное на прогнозах, — в первую очередь государство-нянька. Отрицание ответственности человека за свои действия разрушает фундаментальную свободу людей выбирать свое поведение.

Если большинство решений на государственном уровне полагаются на прогнозы и желание снизить риски, наш личный выбор, а значит, и наша личная свобода действий больше не имеют значения. Где нет вины, там нет невиновности. Уступая такому подходу, мы не улучшаем, а скорее обедняем мир.

Основным стержнем управления большими данными является гарантия того, что мы продолжим судить других, принимая во внимание их индивидуальную ответственность, а не «объективно» обрабатывая числа, чтобы определить, являются ли те или иные лица преступниками. Только в таком случае мы будем относиться к ним по-человечески — как к людям, которые имеют свободу выбора своих действий и право быть судимыми за них. Это не что иное, как последствие наступления эпохи больших данных для нынешней презумпции невиновности.

Вскрытие «черного ящика»

Современные компьютерные системы принимают решения на основе явно запрограммированных правил, которым они должны следовать. Таким образом, если что-то пошло не так, а это неизбежно случается, мы можем вернуться и выяснить, почему компьютер принял то или иное решение. («Почему система автопилота подняла самолет на пять градусов выше, когда внешний датчик определил внезапное повышение влажности?») Сегодня компьютерный код можно открыть и проверить, а основания для решений системы независимо от их сложности — сделать понятными хотя бы для тех, кто разбирается в коде.

При использовании анализа больших данных отследить это станет гораздо сложнее. Основа прогнозов алгоритма зачастую может быть непосильной для человеческого понимания.

Когда компьютеры были явно запрограммированы следовать набору инструкций, как это было с одной из первых программ компании IBM для перевода с русского на английский (1954 год), человеку было легко понять, почему одно слово заменялось другим. Когда компания Google объединяет миллиарды страниц переводов, чтобы судить о том, почему английское слово light выводится на французском как lumière, а не léger (имеется в виду яркость, а не отсутствие тяжести), невозможно точно объяснить причину выбора: основа прогнозирования влечет за собой огромные объемы данных и обширные статистические вычисления.

Масштабы работы с большими данными выйдут далеко за рамки привычного для нас понимания. Так, корреляция, определенная компанией Google между несколькими условиями поиска и гриппом, стала результатом проверки 450 миллионов математических моделей. С другой стороны, Синтия Рудин первоначально разработала 106 прогностических факторов того, что канализационный люк может загореться, и сумела объяснить менеджерам компании Con Edison, почему ее программа выстроила места проверки именно в таком приоритетном порядке. «Объясняемость», как говорят в кругах исследования искусственного интеллекта, имеет большое значение для нас, смертных, которые, как правило, хотят знать не только факты, но и их причину. А что если бы вместо 106 прогностических факторов система автоматически создала 601, подавляющее большинство из которых имеют очень низкий вес, но вместе взятые повышают точность модели? Основа для любого прогноза была бы невообразимо сложной. Что тогда Синтия сказала бы руководителям, чтобы убедить их перераспределить свой скудный бюджет?

В таких случаях мы видим риск того, что прогнозы больших данных, а также алгоритмы и наборы данных, стоящие за ними, станут «черными ящиками», которые не дают ни малейшей прозрачности, подотчетности, прослеживаемости или уверенности. Для того чтобы предотвратить это, необходимы отслеживание и прозрачность больших данных, а также новые виды специальных знаний и учреждения, которые бы ими занимались. Эти новые игроки окажут поддержку в многочисленных областях, где общество должно внимательно изучить прогнозы и дать возможность пострадавшим требовать возмещения.

В обществе такое происходило и раньше, когда при резком увеличении сложности и специализации определенной области возникала острая необходимость в специалистах для управления новыми техническими средствами. Профессии, связанные с юриспруденцией, медициной, бухгалтерским учетом и инженерией, подверглись таким преобразованиям более ста лет назад. Не так давно появились консультанты по компьютерной безопасности и конфиденциальности. Они следят за тем, чтобы деятельность компании соответствовала передовой практике, определяемой такими органами, как Международная организация по стандартизации (созданная ввиду возникшей необходимости в разработке правил в этой области).

В эпоху больших данных потребуются люди, которые взяли бы на себя эту роль. Назовем их алгоритмистами. Они могли бы выступать как представители независимых органов, которые работают вне организаций, и как специалисты самих организаций, аналогично тому как компании нанимают и штатных бухгалтеров, и внешних аудиторов, которые проверяют их работу.

Новая профессия — алгоритмист

Новые профессионалы должны быть специалистами в области компьютерных наук, математики и статистики. Выступали бы они в качестве инстанций, контролирующих анализ и прогнозы больших данных. Алгоритмисты давали бы клятву в беспристрастности и конфиденциальности, как это делают бухгалтеры и другие специалисты в наше время. Они могли бы оценивать выбор источников данных, аналитических средств и средств прогнозирования (в том числе алгоритмов и моделей), а также интерпретацию результатов. В случае возникновения спора алгоритмисты получали бы доступ к соответствующим алгоритмам, статистическим подходам и наборам данных, которые подготовили данное решение.

Если бы в Министерстве внутренней безопасности США в 2004 году был штатный алгоритмист, он смог бы заблаговременно выявить ошибку, закравшуюся в черный список преступников, в который попал сенатор от штата Массачусетс Тед Кеннеди. Вспомним недавние

инциденты, где пригодились бы алгоритмисты. В Японии, Франции, Германии и Италии появились претензии от людей в том, что их позорила функция «автозаполнения» поисковой системы Google, которая выдает список наиболее распространенных условий запроса, связанных с их именем. Эта функция в значительной степени зависит от частоты предыдущих поисков: условия ранжируются в соответствии с их математической вероятностью. А кого бы не возмутило, если бы рядом с его именем отобразилось слово «зэк» или «проститутка», когда кто-то из потенциальных деловых партнеров или пассий решил поискать о нем информацию в Сети?

Мы рассматриваем алгоритмистов как рыночный подход для решения аналогичных проблем, который может оставить позади более навязчивые формы регулирования. Алгоритмисты удовлетворили бы потребность в обработке нового наплыва финансовой информации — так в начале XX века появились бухгалтеры и аудиторы. Обычным людям было трудно разобраться в обрушившемся на них потоке цифр. Возникла необходимость в объединении специалистов в гибкие, саморегулируемые структуры для защиты интересов общества. В ответ рынок породил совершенно новый сектор конкурирующих компаний, которые предлагали услуги финансового надзора. Таким образом новому поколению профессионалов удалось укрепить уверенность общества в экономике как таковой. Большие данные могут и должны использовать преимущества аналогичного повышения уверенности. И с этой задачей успешно справились бы алгоритмисты.

Внешние алгоритмисты

Внешние алгоритмисты могли бы выступить в роли независимых аудиторов для проверки точности и достоверности прогнозов больших данных по запросу клиента или правительства в судебном порядке или по решению регулирующих органов. Алгоритмисты также могли бы проводить аудит пользователей больших данных, нуждающихся в экспертной поддержке, и подтверждать обоснованность применения больших данных, допустим в технических средствах по борьбе с мошенничеством или системах *обращения ценных бумаг.* Наконец, они

могли бы консультировать государственные органы, как лучше всего использовать большие данные в государственном секторе.

По примеру медицины, права и пр. эту новую сферу деятельности можно регулировать кодексом поведения. Беспристрастность, конфиденциальность, компетентность и профессионализм алгоритмистов обеспечивались бы жестким порядком ответственности. В случае нарушения этих стандартов алгоритмисты подвергались бы судебным искам. Их можно было бы привлекать к участию в судебных процессах в качестве свидетелей-экспертов или назначать в качестве «придворных мастеров» (по сути, экспертов в определенной предметной области для оказания помощи судье) при рассмотрении особо сложных вопросов, связанных с большими данными, в ходе судебного разбирательства.

Кроме того, люди, пострадавшие от прогноза больших данных (пациент, которому отказали в хирургическом вмешательстве, заключенный, которому отказали в досрочном освобождении, или заявитель, которому отказали в ипотеке), могли бы обратиться за помощью к алгоритмистам, равно как к адвокатам, чтобы разобраться в этом решении и опротестовать его.

Внутренние алгоритмисты

Внутренние алгоритмисты — штатные специалисты организаций, которые контролируют деятельность, связанную с большими данными. Их задача — отстаивать интересы не только компании, но и людей, пострадавших в результате анализа больших данных, проводимого данной компанией. Внутренние алгоритмисты отвечают за операции с большими данными и являются первыми контактными лицами для таких потерпевших, а также проверяют анализ больших данных на целостность и точность, прежде чем будет оглашен результат. Для выполнения этой задачи алгоритмистам нужен определенный уровень свободы и непредвзятости в рамках организации, в которой они работают.

Может показаться нелогичным, что человек, работающий в компании, должен оставаться беспристрастным по отношению к ней. Но такое встречается достаточно часто. Один из примеров — отделы

по надзору в крупных финансовых учреждениях; далее — советы директоров во многих компаниях, которые несут ответственность перед акционерами, а не руководством. А многие медиакомпании, в том числе New York Times и Washington Post, нанимают омбудсменов, основной обязанностью которых является защита доверия общественности. Эти сотрудники работают с жалобами читателей и нередко публично подвергают суровой критике своего работодателя, если считают его виновным.

Еще более удачный аналог внутреннего алгоритмиста — специалист, который несет ответственность за злоупотребление личной информацией в корпоративной среде. В Германии компании, превышающие определенный размер (наличие в штате десяти и более человек, занятых обработкой личной информации), обязаны назначить представителя для защиты данных. Начиная с 1970-х годов штатные представители для защиты данных разработали профессиональную этику и корпоративный дух. Они регулярно встречаются для обмена передовым опытом и обучения, а также имеют собственные специализированные СМИ и проводят конференции. Кроме того, им удалось развить двойную лояльность: к своим работодателям *и* к своим обязанностям в качестве непредвзятого контролирующего органа[138]. Существование немецких представителей защиты корпоративных данных можно расценивать как успех в выполнении функций омбудсмена по защите корпоративных данных и укреплении ценностей конфиденциальной информации во всех сферах деятельности компании. На наш взгляд, алгоритмисты могли бы выполнять аналогичную функцию.

Раскрытие информации

Основной инструмент, который государственные органы используют для надзора за деятельностью граждан и компаний, — запрос на предоставление информации. Иногда раскрытия информации самого по себе достаточно, чтобы стимулировать соблюдение требований или отстаивать цели регулирования. Такой принцип лежит в основе законов, согласно которым компании, имеющие утечку больших данных, обязаны уведомлять об этом потребителей и регулирующие

органы. Как видно, угроза общественного неодобрения может стимулировать надлежащую профилактику. На эту же идею опираются экологические законы, которые требуют от компаний не снижения выбросов токсичных веществ, а лишь раскрытия информации о выделяемом их количестве: контроль и отчетность стимулируют сторонников внутри компании и в обществе в целом оказывать давление на компанию, чтобы снизить загрязнение. Сама лишь прозрачность информации может достигать социальных целей, которые трудно даются политическим путем[139].

Открытость станет важным способом контроля действий с большими данными и обеспечения надлежащей прозрачности для наборов данных, алгоритмов, предположений, статистических подходов и вытекающих из них решений. Проверки анализа больших данных могут потребоваться по решению суда, в рамках конкретного расследования или в качестве периодической меры (например, годовой финансовой отчетности для открытых акционерных компаний).

Конечно, прозрачность не означает, что компании будут разглашать конфиденциальную информацию. Публичное уведомление может содержать информацию о том, что организация проверяет или уже проверила свои прогнозные модели, не разглашая их суть. Такие уведомления характерны для современных проверок систем безопасности и конфиденциальности. Обязательная проверка и ограниченная огласка входят в требования, предъявляемые к компаниям, официально зарегистрированным на бирже до 2000 года; тогда компании должны были сообщать о своей готовности к полному изменению в своих отчетах на фондовой бирже.

Регулирующие органы США уже утвердили такой порядок в соглашениях на расследования Федеральной торговой комиссии (Federal Trade Commission, FTC), обязав Google и Facebook проводить аудит конфиденциальности раз в два года в течение 20 лет и предоставлять отчет FTC. Для компании Twitter срок аналогичных обязательств был установлен равным десяти годам. А после массовой утечки конфиденциальных данных о более чем 45 миллионах кредитных карт торговая сеть TJX, управляющая многочисленными магазинами уцененных товаров в США (T.J. Maxx и пр.), наряду с

брокерами данных Reed Elsevier и Seisint согласилась ежегодно проводить независимый аудит безопасности в течение последующих 20 лет и сообщать о результатах в FTC.

Такой подход имеет ряд преимуществ. Лучше обеспечивается соблюдение требований, поскольку контроль осуществляется периодически в течение длительного периода. Главный вопрос поднимается на самые высокие уровни управления, а не остается в ведении ИТ-вундеркиндов, которые заняты решением повседневных задач для поддержания работы систем и могут поскупиться на надлежащие меры безопасности ввиду ограниченности времени и бюджета. Кроме того, этот подход изначально гибкий, а значит, передовая практика и надлежащие меры безопасности будут со временем меняться с учетом новых технологий и взглядов. Опорой служит более рыночно ориентированный механизм проведения проверок — участие независимых специалистов, а не регулирующих органов, которые не всегда достаточно компетентны, чтобы проводить такие мероприятия.

Бароны данных

Данные в информационном обществе — все равно что топливо в эпоху промышленной революции: крайне важный ресурс, подпитывающий нововведения, на которые полагаются люди. Без обширного, динамичного снабжения данными и надежного рынка услуг эти нововведения могут исчезнуть.

В этой главе рассмотрены три фундаментальных перехода в управлении, благодаря которым мы можем быть уверены, что темную сторону больших данных удастся укротить. По мере развития зарождающейся отрасли больших данных возникнет дополнительная задача первостепенной важности — защита конкурентных рынков больших данных. Мы должны предотвратить появление «баронов данных» — современный эквивалент баронов-разбойников XIX века, которые подмяли под себя железные дороги, производство стали и телеграфные сети США.

Для контроля этих ранних промышленников в США установлены чрезвычайно гибкие антимонопольные правила. Первоначально

разработанные для железных дорог в 1800-х годах, позднее они были применены к другим компаниям, препятствующим потоку информации, от которой зависели компании, — от компании NCR Corporation (в 1910-х) до IBM (в 1960-х и далее), Xerox (в 1970-х), AT&T (в 1980-х), Microsoft (в 1990-х) и Google (в наше время). Технологии, впервые представленные ими, стали одним из основных компонентов «информационной инфраструктуры» экономики, и понадобилась сила закона, чтобы предотвратить их господство.

Для шумного рынка больших данных придется обеспечить условия, сопоставимые с конкурентной борьбой и надзором, которые уже успели закрепиться в этих технологических областях. Регулирующим органам потребуется найти равновесие между осторожными и решительными действиями. Антимонопольный опыт указывает, каким путем этого равновесия можно достичь. Но развитие технологий невозможно предугадать. Даже большие данные не могут спрогнозировать собственное развитие.

Антимонопольное регулирование обуздало злоупотребление властью. Удивительно, как превосходно принципы перемещаются из одного сектора в другой, а также между различными типами сетевых отраслей. Это словно вид мышечной регуляции, где каждая из технологий получает равную поддержку, что само по себе полезно, так как устанавливает равные условия для конкуренции, не предполагая ничего большего. Чтобы стимулировать здоровую конкуренцию в сфере больших данных, государственные органы должны применять антимонопольные правила. Кроме того, выступая одним из крупнейших в мире держателей данных, они должны выпускать свои данные публично. Подобные процессы мы наблюдаем уже сегодня.

Опыт антимонопольного регулирования заключается в том, что, определив всеобъемлющие принципы, регулирующие органы могут реализовать их, чтобы обеспечить необходимые гарантию и поддержку. Кроме того, три стратегии, которые мы обозначили: смещение защиты конфиденциальности от индивидуального согласия в сторону подотчетности пользователей данных, закрепление приоритетности человеческого фактора над прогнозами, а также создание нового класса аудиторов больших данных (алгоритмистов) —

могут служить основой эффективного и справедливого управления информацией в эпоху больших данных.

Как это часто бывало в истории других нововведений (от ядерных технологий до биотехнологий), люди сначала создают инструменты, которые могут им навредить, а затем изобретают механизмы, чтобы от них защититься. В этом смысле большие данные занимают место в ряду таких сфер жизни общества, которые ставят перед нами задачи, не имеющие единственно верного решения. Они поднимают текущие вопросы о том, как мы распоряжаемся окружающим миром. Каждое поколение должно решать эти вопросы заново. Наша задача — оценить опасность этих новейших технологий, поддержать их развитие и собрать плоды.

Как и печатный станок, большие данные приводят к изменению порядка самоуправления в обществе. Это заставляет нас по-новому решать вечные проблемы и новые задачи, опираясь на основные принципы. Чего мы не должны допустить, так это неуправляемого развития больших данных, когда формирование технологии становится неподвластно человеку. Нужно способствовать развитию технологий, не забывая о безопасности людей.

Примечания к главе 9

Глава 10

Что дальше?

Майк Флауэрс работал юристом в офисе окружного прокурора Манхэттена, занимаясь судебным преследованием по различным делам, от убийств до преступлений на Уолл-стрит, прежде чем перешел в одну из шикарных корпоративных адвокатских контор в начале 2000-х годов. Проведя год за скучной офисной работой, Майк решил уйти. Ему хотелось делать что-то более значимое, например помогать вершить перестройку в Ираке. Коллега сделал пару звонков вышестоящим лицам — и Флауэрс не успел опомниться, как направился в «Зеленую зону» (безопасный район для американских войск в центре Багдада) в составе группы юристов для суда над Саддамом Хусейном.

Его задача оказалась скорее логистической, чем юридической. Флауэрсу предстояло определить места предполагаемых массовых захоронений, чтобы знать, куда направить следователей на раскопки. Кроме того, ему нужно было благополучно переправить свидетелей в «Зеленую зону», обезопасив их от взрывов многочисленных СВУ (самодельных взрывных устройств), которые были страшной повседневной реальностью. Он увидел, что военные рассматривали эти вопросы как задачи обработки данных. Аналитики разведывательной службы, например, сочетали полевые отчеты со сведениями о местоположении, времени и жертвах прошлых СВУ, чтобы спрогнозировать наиболее безопасный маршрут на конкретный день.

По возвращении в Нью-Йорк два года спустя Флауэрс понял, что те методы являлись более эффективным способом борьбы с преступностью, чем он когда-либо имел, будучи прокурором. К тому же Флауэрс нашел поистине родственную душу в лице мэра Майкла Блумберга, который сколотил состояние на поставке банкам финансовой информации и ее анализа. Флауэрса определили в специальную оперативную группу по обработке чисел, которая должна

была разоблачить преступников, замешанных в скандале с ипотеками высокого риска в 2009 году. Работа группы оказалась настолько успешной, что уже через год мэр Блумберг попросил ее расширить сферу деятельности. Флауэрс стал первым городским «директором по аналитике». Его миссия заключалась в том, чтобы создать команду лучших ученых в области данных, которых только можно было найти, и с их помощью обрабатывать нетронутые городские залежи информации на благо всех и вся.

Флауэрс раскинул сеть для поиска подходящих людей: «Меня не интересовали очень опытные статистики, поскольку они могли не принять новый подход к решению проблем». В ходе собеседований со статистиками для проекта, связанного с финансовым мошенничеством, Флауэрс заметил, что они склонны проявлять скрытое беспокойство по поводу математических методов. «Я даже не задумывался о том, какая модель будет использоваться. Мне нужны были результаты, дающие основания для конкретных действий. Это все, что меня заботило», — говорит он. Флауэрс собрал команду из пяти человек (как он их назвал, «напарников»). Все, кроме одного, были экономистами по специальности, окончившими вуз всего год или два назад, без особого жизненного опыта в большом городе, но с определенным творческим потенциалом.

Одна из первых задач, с которыми они столкнулись, была связана с серьезным вопросом «незаконного переоборудования» — практикой разделения жилищ на множество мелких помещений, чтобы вместить в десятки раз больше людей, чем предусмотрено по проекту. Незаконно переоборудованные жилища не только имеют высокую пожароопасность, но и являются рассадниками преступности, наркомании, болезней и вредителей. Клубки проводов, опоясывающие стены, электроплиты прямо на покрывалах, люди, утрамбованные вплотную. В таких адских условиях люди мрут как мухи. В 2005 году двое пожарных разбились насмерть, пытаясь спасти людей в одном из приютов. Нью-Йорк ежегодно получает около 25 000 жалоб на незаконное переоборудование, но их обработкой занимается всего 200 инспекторов. При этом у них нет надежного способа отличить простые неудобства от реальной угрозы воспламенения. Флауэрс и его

напарники увидели в этом задачу, которую можно решить с помощью большого количества данных.

Они начали с составления списка всех 900 000 зданий в городе. Затем изучили наборы данных, полученные от 19 различных учреждений, в которых указывались наличие задержек в уплате налогов на недвижимость со стороны владельца здания, разбирательств по поводу взысканий по закладной, отклонений в оплате коммунальных услуг или их отключение за неуплату. Учитывались информация о типе здания и времени его постройки, визиты скорой помощи, уровень преступности, жалобы на грызунов и многое другое. Полученные данные сравнивались с упорядоченными по степени сложности данными о пожарах за последние пять лет. Тем самым планировалось выявить корреляции для создания модели, которая сможет прогнозировать, какие жалобы требуют наиболее быстрого реагирования.

Основная часть исходных данных была представлена в неподходящей форме. Отсутствовало единообразие в описании местоположения домов: каждые агентство и департамент, похоже, имели свой подход. Департамент строительства давал каждому зданию уникальный номер. У департамента по сохранению жилищного фонда была иная система нумерации. Налоговый департамент присваивал каждому объекту недвижимости идентификатор на основе района, квартала и участка. Полиция использовала декартову систему координат. Пожарные учитывали близость к «пожарным извещателям», связанным с расположением пожарной части (хотя сами пожарные извещатели уже упразднены). Напарники Флауэрса задействовали эти беспорядочные данные, разработав систему, которая учитывает радиус вокруг передней части здания на основе декартовых координат и добавляет геолокационные данные, полученные из других учреждений. Изначальные сведения были неточными, но огромное количество данных, загружаемых в систему, с лихвой компенсировало этот недостаток.

Команда не довольствовалась одними лишь математическими вычислениями. Напарники Флауэрса изучили работу инспекторов в полевых условиях. Они делали многочисленные заметки и выспрашивали у профессионалов мельчайшие подробности. Если

умудренный опытом начальник сообщал, что здание, к которому они подошли, не представляет угрозы, напарники хотели знать причину его уверенности. Он не мог ее точно сформулировать, но со временем напарники поняли, что он имел в виду новую кирпичную кладку снаружи здания. Это означало, что владелец заботился о здании должным образом.

Напарники вернулись в свои кабины, задаваясь вопросом, как внести в свои модели такой сигнал, как «свежая кирпичная кладка». В конце концов, кирпичи пока еще не датифицированы. Зато на выполнение любых фасадных кирпичных работ требовалось разрешение городских властей. Эта информация значительно улучшила прогностическую эффективность системы, указывая, какие здания, скорее всего, не представляли особого риска.

Аналитика неоднократно демонстрировала, что некоторые из освященных веками способов ведения дел не были лучшими, равно как скаутам из фильма «Человек, который изменил всё» пришлось смириться с недостатками своей интуиции. Например, раньше количество звонков с жалобами по горячей линии города «311» рассматривалось как индикатор наиболее серьезных проблем: чем больше звонков, тем серьезнее проблема. Но это оказалось ложной предпосылкой. Крыса, замеченная в шикарном Верхнем Ист-Сайде, могла обеспечить 30 звонков в час, но в районе Бронкса понадобилось бы не меньше армии грызунов, чтобы соседи соизволили набрать номер. Точно так же большинство жалоб на незаконное переоборудование могло быть связано с шумом, не вызвавшим каких-либо серьезных последствий.

В июне 2011 года Флауэрс с напарниками «щелкнули выключателем». Все жалобы, подходящие под категорию незаконного переоборудования, были пропущены через их систему на еженедельной основе. Напарники отобрали данные о 5% статистически наиболее пожароопасных зданий и передали их инспекторам для незамедлительной проверки. Полученные результаты ошеломили всех.

До применения анализа больших данных инспекторы в первую очередь проверяли жалобы, которые считали самыми неотложными. Но только в 13% случаев условия оказывались достаточно тяжелыми,

чтобы требовать выселения. Теперь инспекторы выдавали ордеры на выселение более чем в 70% случаев проверок. Большие данные позволили пятикратно повысить эффективность рабочего времени инспекторов. И результаты работы улучшились, так как можно было сконцентрировать усилия на самых серьезных проблемах. Обретенная эффективность имела и побочные преимущества. Пожары на незаконно переоборудованных участках в 15 раз чаще приводили к ранениям или гибели пожарных, поэтому новый подход тут же нашел признание в рядах пожарной службы. Флауэрс и его напарники были похожи на волшебников с хрустальным шаром, который позволяет заглянуть в будущее и предсказать, какие места наиболее опасны. Они взяли огромное количество данных, хранившихся долгие годы и практически не используемых с момента сбора, и применили их по-новому, извлекая реальную пользу. С помощью огромного массива информации напарникам удалось сделать ценные открытия, которые были бы невозможны при ее меньших количествах. В этом и есть суть больших данных.

Опыт нью-йоркских «алхимиков» в области аналитики наглядно демонстрирует множество тем, раскрытых в этой книге. Они использовали гигантский объем данных, а не его небольшую часть. Их список зданий в городе представлял собой не что иное, как массив данных «*N* = всё». Их не смутила беспорядочность данных, например информации о местоположении или записей скорой помощи. Преимущества большого количества данных перевесили недостатки меньшего количества нетронутой информации. Напарникам удалось достичь своих целей, поскольку многие характеристики города были представлены (пусть и непоследовательно) в виде данных, что позволило обрабатывать и использовать информацию для улучшения прогнозов.

Догадки экспертов, будь то напыщенные статистики или государственные служащие, отвечающие за горячую линию для жалоб, были вынуждены уступить место подходу, основанному на данных. Вместе с тем Флауэрс и его напарники постоянно сверяли свои модели с мнением опытных инспекторов, чьи советы помогли усовершенствовать систему. Однако важнейшей причиной

ошеломительного успеха программы был отказ от причинности в пользу корреляции.

«Меня не интересуют причинно-следственные связи, если только они не касаются конкретных действий, — поясняет Флауэрс. — Это не для меня. И, честно говоря, все эти разговоры о причинности полны неясностей. Не думаю, что день разбирательства по поводу взысканий по закладной и статистическая вероятность пожара в определенном здании хоть как-то взаимосвязаны. Я полагаю, было бы глупо так считать. И никто бы не объявил об этом во всеуслышание. Считается, что есть основные факторы. Но я даже не хочу в это вникать. Мне нужна конкретная точка данных, которая имеет определенную значимость и к которой у меня есть доступ. Если она значима, мы будем ее учитывать, а если нет — то нет. В общем, нам нужно решать реальные проблемы. И, откровенно говоря, я не могу себе позволить отвлекаться на причинность и прочую ерунду».

Когда данные говорят

Большие данные имеют огромное практическое значение как технология, которая служит решению животрепещущих повседневных проблем, но при этом порождает еще больше новых. Большие данные способны изменить наш образ жизни, труда и мышления. В каком-то смысле мы упираемся в больший тупик, чем во времена других эпохальных инноваций, значительно расширивших объем и масштабы информации в обществе. Мы стоим на зыбкой почве. Старые факты подвергаются сомнению. Ввиду больших данных необходимо пересмотреть понятия природы принятия решений, судьбы и справедливости. Мировоззрение, сотканное из понимания причин, теперь оспаривается доминированием корреляций. Обладание знанием, которое когда-то означало понимание прошлого, постепенно преобразовывается в способность прогнозировать будущее.

Эти вопросы намного важнее тех, которые возникали по мере запуска интернет-магазинов, повседневного использования интернета, входа в эпоху компьютеров или введения в обиход абака. Мысль о том, что стремление понять причины может быть переоценено и в большинстве случаев выгоднее отказаться от вопроса *почему* в пользу

вопроса *что*, предполагает, что эти вопросы оказывают существенное влияние на наш образ жизни и мышления. Однако они могут оказаться риторическими. По сути, эти вопросы — часть вечных дискуссий на тему места человека в мире и его поисков смысла жизни в суматохе хаотичного и непостижимого мира.

Большие данные ознаменовали момент, когда «информационное общество», наконец, начало оправдывать свое название. Всю собранную цифровую информацию теперь можно по-новому использовать в инновационных целях, открывая новые формы ценности. Для этого нужен иной тип мышления, который бросает вызов нашим учреждениям и даже нашему чувству идентичности. Ясно одно: объем данных будет неуклонно расти, равно как и возможности их обработки. Но если большинство людей рассматривают большие данные как технологический вопрос, сосредоточив внимание на аппаратном или программном обеспечении, мы считаем, что акцент необходимо перенести на то, что происходит, когда данные «говорят».

Мы можем собирать и анализировать больше информации, чем когда-либо. Нехватка данных отныне не определяет наши усилия для познания мира. Мы можем использовать значительно больше данных, а в некоторых случаях даже все. Но для этого придется взять на вооружение нестандартные способы обработки и, в частности, изменить свое представление об идеале полезной информации.

Вместо того чтобы ставить во главу угла точность, чистоту и строгость данных, мы можем — и это даже необходимо — несколько ослабить свои требования. Данные не должны быть заведомо ошибочными или ложными, но их беспорядочность не представляет особых проблем при многократном увеличении масштаба. Она может быть даже выгодной, так как, используя лишь небольшую часть данных, мы упускали из виду широкое поле подробностей, где обнаруживается масса знаний.

Поскольку корреляции можно найти гораздо быстрее и с меньшими затратами, чем причинность, им нередко отдается предпочтение. В некоторых случаях (например, при тестировании побочных эффектов препарата или проектировании важнейших частей самолета) по-прежнему понадобятся исследования причинно-следственных связей и

эксперименты в контролируемых условиях с тщательным контролем данных. Но для многих бытовых нужд вполне достаточно знать ответ на вопрос *что*, а не *почему*. Кроме того, корреляции больших данных способны указать перспективные направления для поиска причинности.

Быстрые корреляции позволяют экономить на покупке авиабилетов, прогнозировать вспышки гриппа и определять люки и перенаселенные здания, которые следует осмотреть, в условиях ограниченных ресурсов. Они же позволяют медицинским страховым компаниям принимать решения по страховой защите без медицинского осмотра и снижают стоимость напоминаний больным о приеме лекарств. На основании прогнозов, сделанных с помощью корреляций среди больших данных, выполняются переводы и создаются системы автоматического управления автомобилем. Walmart может узнать, какой сорт печенья Pop-Tarts положить сразу у входа в магазин, когда надвигается ураган (ответ: со вкусом клубники). Конечно, причинно-следственные связи не лишние, когда их удается уловить. Проблема в том, что зачастую их выявить непросто, и мы нередко обманываем себя, считая, что нам это удалось.

Все эти новые возможности в какой-то мере обеспечиваются новыми инструментами — от более быстрых процессоров и увеличенного объема памяти до более эффективного программного обеспечения и алгоритмов. Они, безусловно, играют важную роль, но больше данных у нас появляется благодаря постепенной датификации всего и вся. Надо отметить, что стремление измерить мир количественно появилось задолго до компьютерной революции. Но цифровые инструменты подняли датификацию на новый уровень. Мало того что мобильные телефоны могут отслеживать, кому мы звоним и куда идем, — те же данные дают возможность определить, что мы заболели. Вскоре они смогут дать понять, что мы влюблены.

Способность создавать что-то новое, успевать больше и делать все лучше и быстрее раскрывает огромную ценность данных, разделяя мир на победителей и проигравших. Основную (альтернативную) ценность информации обеспечит ее вторичное использование, а не только первичное, как принято считать. Таким образом, целесообразно собирать как можно больше самых разных данных и удерживать до тех

пор, пока это содержит добавочную ценность, а также давать возможность анализировать данные тем, кто имеет больше возможностей раскрытия их ценности (при условии разделения полученной выгоды).

Успеха добьются компании, которые сумеют попасть в центр информационных потоков и научатся собирать данные. Для эффективного использования больших данных требуются технические навыки и хорошее воображение — мышление категориями больших данных. Основная ценность достанется тем, кто владеет данными. При этом важным активом может оказаться не только та информация, которая на виду, но и выбросы данных, полученные от взаимодействия людей с информацией. Используя такие выбросы с умом, компания улучшит существующую службу или запустит совершенно новую.

Большие данные таят в себе огромные риски. Они стирают правовые и технические ограничения, с помощью которых мы пытаемся сохранить конфиденциальность, тем самым выявляя неэффективность существующих основных технических и правовых механизмов. Раньше было хорошо известно, что относится к личной информации: имена, номера социального страхования, идентификационные коды и пр. Защитить такую информацию было относительно нетрудно, заблокировав ее. Сегодня даже с помощью самых безобидных данных, если их накоплено достаточно много, можно установить личность. Попытки придать данным анонимную форму или скрыть их уже неэффективны. Кроме того, установление слежки за отдельными лицами теперь влечет за собой более глубокое вторжение в частную жизнь, чем когда-либо, поскольку органы власти хотят увидеть не только как можно больше информации о человеке, но и как можно более широкий спектр его отношений, связей и взаимодействий.

Независимо от того, насколько большие данные угрожают конфиденциальности, существует другая уникальная и тревожная проблема. Ввиду того что прогнозы больших данных становятся все более точными, их можно использовать для наказания людей за прогнозируемое поведение, то есть действия, которые им предстоит совершить. Такие прогнозы невозможно опровергнуть в очевидной форме, поэтому никто не в силах себя оправдать. Наказание на этой

основе отрицает понятие свободы воли и вероятность, пусть и небольшую, что подозреваемый выберет другой путь. Поскольку мы назначаем индивидуальную ответственность (и применяем наказание), человеческая воля должна быть неприкосновенна. Если будущее не оставит нам свободного поля деятельности, большие данные извратят саму суть человеческой природы: рациональное мышление и свободу выбора.

У нас пока нет надежных способов подстроить нормы и законы под специфику грядущего мира больших данных. Однако по мере постижения обществом их особенностей и недостатков его процветанию будут способствовать некоторые реформы. Мы в состоянии обеспечить свободный обмен информацией, учредив права исключения для данных, контролируя расстановку сил на рынке и поощряя государственные инициативы в поддержке идеи открытых данных. Мы можем расширить доступ к личной информации, установив способы ее приемлемого вторичного использования (для чего не понадобятся дополнительные разрешения), но в то же время ограничив сроки хранения и применения такой информации. Мы можем найти новые технические решения, например способы «размывания» признаков для установления личности. Прогнозы больших данных не должны служить назначению индивидуальной ответственности. Человеческая воля неприкосновенна. Наконец, людям нужно дать возможность исследовать алгоритмы и исходные данные, применявшиеся в ходе принятия решений, влияющих на их интересы (особенно если это влияние негативное). Для преодоления этой задачи необходимо новое поколение специалистов (алгоритмистов), призванных помочь анализировать и интерпретировать эффективность и законность инструментов и процессов обработки больших данных.

Большие данные станут неотъемлемой частью понимания и решения многих насущных глобальных проблем. Борьба с изменением климата требует анализа данных о загрязнении, чтобы понять, куда лучше всего направить усилия, и найти пути смягчения последствий проблем. Немыслимое количество датчиков, размещенных по всему миру (в том числе встроенных в смартфоны), позволяет моделировать ситуацию на более высоком уровне детализации. Улучшение

структуры здравоохранения и снижение затрат на него, особенно в беднейших странах мира, станет значительной частью программы автоматизации процессов, которые в настоящее время нуждаются в человеческих суждениях, но могли бы выполняться компьютерами (например, изучение биопсии раковых клеток или обнаружение признаков инфекции до ее полного развития).

Большие данные уже использовались на благо экономического развития и предотвращения конфликтов. Так, данные о передвижении владельцев сотовых телефонов показали участки африканских трущоб, которые являются средоточием бурной экономической активности. Кроме того, большие данные дали возможность обнаружить общины с наиболее обострившейся межэтнической напряженностью и показали, чем может обернуться кризис беженцев[140]. Со временем большие данные станут использоваться все чаще, поскольку технология находит применение во всех сферах жизни.

Большие данные позволяют не только делать лучше то, что мы уже умеем, но и изобретать что-то новое. Однако это не волшебная палочка. Они не установят мир во всем мире, не приведут к искоренению нищеты или появлению нового Пикассо. С помощью больших данных невозможно произвести на свет младенцев, зато можно спасти преждевременно рожденных. Со временем большие данные наверняка войдут почти во все аспекты нашей жизни. Возможно, их отсутствие даже станет вызывать легкое беспокойство сродни тому, когда мы ожидаем от врача направление на рентген для выяснения того, что не удалось выявить путем обычного медицинского обследования.

Поскольку большие данные входят в нашу жизнь, они вполне могут влиять на наше представление о будущем. Около пятисот лет назад изменилось восприятие человечеством времени в рамках движения к более светской, научно обоснованной и просвещенной Европе[141]. На заре человечества время считалось циклическим понятием, как и сама жизнь. Каждый день (и год) был очень похож на предыдущий, и даже конец жизни напоминал ее начало, поскольку стареющие взрослые снова становились беспомощны, как дети. Когда стало преобладать линейное восприятие времени, мир предстал в виде

развертывающейся вереницы дней — линии жизни, подвластной нашему влиянию. Если раньше прошлое, настоящее и будущее были слиты воедино, то теперь у человечества появилось прошлое, на которое можно оглянуться, и будущее, которого можно с трепетом ожидать, пока длится настоящее.

В то время как настоящее мы в силах формировать, будущее превратилось из чего-то абсолютно предсказуемого в нечто открытое и нетронутое — огромный пустой холст, который каждый мог заполнить в соответствии со своими ценностями и усилиями. Одна из характерных черт современности — то, что мы воспринимаем себя хозяевами своей судьбы, и это отличает нас от наших предков, для которых предопределенность в той или иной форме была нормой. Прогнозы больших данных делают полотно нашей жизни менее открытым, чистым и нетронутым. Наше будущее кажется в какой-то мере предсказуемым для тех, кто владеет технологией, чтобы это сделать. Похоже, это уменьшает нашу способность определять самим свою судьбу, а потенциальные возможности возлагает на алтарь вероятности.

В то же время большие данные могут означать, что мы всегда остаемся узниками своих предыдущих действий, которые модели прогнозирования используют против нас, претендуя на знание наших последующих действий: нам никогда не уйти от того, что случилось. «Прошлое — это лишь пролог», — писал Уильям Шекспир. Большие данные закрепляют это утверждение алгоритмически со всеми его достоинствами и недостатками. Но омрачит ли это нашу радость каждому восходу солнца или желание оставить в этом мире свой след?

Скорее всего, наоборот. Зная, что может произойти в будущем, мы примем надлежащие меры, чтобы предотвратить проблемы или улучшить результаты. Мы сможем заметить, кто из студентов начал «скатываться», задолго до выпускного экзамена. Мы выявим мельчайшие раковые опухоли и вылечим их, прежде чем они успеют разрастись. Мы узнаем о вероятности нежелательной подростковой беременности или преступности и сможем вмешаться, сделав все возможное, чтобы предотвратить вероятный исход. Мы предупредим пожары с потенциальными жертвами в многоквартирных зданиях Нью-Йорка, зная, какие из них проверить в первую очередь.

Ничто не предопределено, потому что мы всегда можем отреагировать на полученную информацию. Прогнозы больших данных не высечены на камне — это всего лишь наиболее вероятные результаты, а значит, при желании их можно изменить. Мы сами выбираем, как встретить и приручить будущее — словно Мори, отыскавший естественные пути среди огромной глади моря и ветров. Для этого не нужно понимать природу космоса или доказывать существование богов — достаточно больших данных.

Больше чем большие данные

Преобразуя свою жизнь с помощью больших данных — оптимизируя, улучшая, повышая эффективность и используя преимущества, — какую роль мы отводим интуиции, вере, неопределенности и новизне?

Большие данные учат нас тому, что более эффективные поступки и постоянное совершенствование, пусть и лишенные глубокого понимания, достаточно надежны. Твердо придерживаясь такого подхода, вполне можно преуспеть. Даже если вы не знаете, почему ваши усилия сказываются тем или иным образом, с большими данными вы добьетесь большего успеха, чем без них. Флауэрс и его напарники в Нью-Йорке, может, и не являются воплощением просвещенных мудрецов, но они и вправду спасают жизни. Так что большие данные не только повышают нашу эффективность, но со временем, вероятно, смогут дать то, что мы могли бы назвать мудростью.

Большие данные — нечто большее, чем холодный мир алгоритмов и автоматики. Существенную роль играют люди со всеми своими слабостями, заблуждениями и ошибками, поскольку эти черты — неотъемлемая часть творчества, интуиции и гениальности человека. Одни и те же беспорядочные умственные процессы ведут как к унижениям или упорству в заблуждениях, так и к успехам и обретению величия. Это наводит на мысль, что следует приветствовать некоторую неточность как своего рода часть человеческой природы, так же как мы учимся охватывать беспорядочные данные, поскольку они служат большой цели. В конце концов, беспорядочность является

важным достоянием мира и нашего мышления. Принять ее и считаться с ней — значит получить преимущества.

Вы спросите, какой толк от людей в условиях, когда решения опираются на данные, а интуиция противоречит фактам? Если бы все обращались к данным и использовали соответствующие инструменты, возможно, критическим отличием стал бы элемент непредсказуемости — человеческий фактор интуиции, риска, случайностей и ошибок.

В таких условиях неизбежно придется выкроить место для человека — его интуиции, здравого смысла и прозорливости, чтобы их не заглушили данные и машинные ответы. Главное преимущество человека заключается в том, чего не могут уловить и показать алгоритмы и кремниевые микросхемы, поскольку это нельзя выразить в виде данных. Мы имеем в виду не то, *что есть*, а то, *чего нет*, будь то пустое пространство, трещина в тротуаре или невысказанная либо пока еще не сформировавшаяся мысль.

Человеческий фактор имеет огромное значение для достижения прогресса в обществе. Большие данные означают, что мы можем экспериментировать быстрее и исследовать больше инициатив, при этом создавая больше инноваций. Искра изобретения — то, о чем не узнаешь из данных, и то, что не удастся подтвердить при любом их количестве, поскольку речь идет о том, чего пока не существует. Если бы Генри Форд спросил большие данные, чего хотят его клиенты, они бы ответили — более быстрых лошадей (мы перефразировали его крылатую фразу). В мире больших данных будут поощряться такие человеческие качества, как творчество, интуиция, риск и интеллектуальные амбиции, ведь наша изобретательность — источник прогресса.

Большие данные являются как инструментом, так и ресурсом и предназначены в большей степени информировать, чем объяснять. Они ведут людей к пониманию, но все еще могут вызывать недоразумения в зависимости от того, как с ними обращаться. Какими бы ослепительными ни были возможности больших данных, мы не должны позволять, чтобы их соблазнительный блеск затмил свойственные им недостатки.

Мы никогда не сможем собрать, сохранить или обработать всю совокупность мировой информации — максимальное количество «*N* =

всё» — с помощью существующих технологий. Лаборатория физики элементарных частиц ЦЕРН в Женеве собирает менее 0,1% информации, которая создается в процессе экспериментов, а остальное рассеивается, как дым, вместе с сопутствующими знаниями[142]. Но это вряд ли новая истина. Общество всегда было ограничено в инструментах, используемых для измерения и познания действительности — от компаса и секстанта до телескопа, радара и, наконец, GPS. Наши инструменты завтра могут стать вдвое, десятикратно или даже в тысячу раз мощнее, чем сегодня, основательно снизив значимость наших нынешних знаний. В скором времени наш мир больших данных покажется чем-то столь же забавным, как память 4 Кб бортового управляющего компьютера «Аполлон-11»[143].

Мы всегда сможем собирать и обрабатывать лишь малую часть совокупной всемирной информации, и она может быть только подобием действительности, словно тени на стенах пещеры Плато[25]. Поскольку информация не бывает идеальной, наши прогнозы так или иначе подвержены ошибкам. Но это не означает неправильности данных — просто они не бывают полными. Такое положение вещей не отрицает открытий со стороны больших данных, но все расставляет по местам. Большие данные не дают окончательных ответов, но и те, что есть, дают нам возможность дождаться лучших методов и, следовательно, лучших ответов. А между тем нам следует использовать большие данные с большой долей беспристрастности... и человечности.

Примечания к главе 10

Примечания

Глава 1

1. Статья о тенденциях распространения гриппа, опубликованная в научном журнале Nature: Jeremy Ginsburg et al. Detecting influenza epidemics using search engine query data // Nature. — 2009. — Vol. 457. — P. 1012–1014. URL: http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html
2. Дополнительное исследование службы Google Flu Trends (в соответствии с независимым дополнительным клиническим исследованием в госпитале Джона Хопкинса): Dugas et al. Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics // CID Advanced Access. — January 8, 2012. — DOI 10.1093/cid/cir883.
3. Покупка авиабилетов: Farecast — информация от Кеннета Кукьера: Kenneth, Cukier. Data, data everywhere // The Economist. — February 27, 2010. — P. 1–14. А также интервью с Эциони (2010–2012 гг.).
4. Статья Эциони «Гамлет»: Etzioni, Oren. To buy or not to buy: mining airfare data to minimize ticket purchase price / Oren Etzioni, C. A. Knoblock, R. Tuchinda, and. A. Yates // SIGKDD ’03. — August 24–27, 2003. URL: http://knight.cis.temple.edu/~yates//papers/hamlet-kdd03.pdf.
5. Сколько компания Microsoft заплатила за Farecast. Из сообщений СМИ, в частности: Secret Farecast buyer is Microsoft // Seattlepi.com. — April 17, 2008. URL: http://blog.seattlepi.com/venture/2008/04/17/secret-farecast-buyer-is-microsoft/?source=mypi.
6. Астрономия и секвенирование ДНК. Специальный отчет в журнале The Economist (см. выше): Data, data everywhere // The Economist. — February 27, 2010. — P. 1–14.
7. Секвенирование ДНК: Pollack, Andrew. DNA Sequencing Caught in the Data Deluge // New York Times. — November 30, 2011. URL:

http://www.nytimes.com/2011/12/01/business/dna-sequencing-caught-in-deluge-of-data.html?pagewanted=all.

8. Статистика Facebook: Facebook IPO prospectus // Facebook. — Form S-1 Registration Statement, US Securities And Exchange Commission. — February 1, 2012. URL: http://sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm.
9. Статистика YouTube: Page, Larry. Update from the CEO // Google, April 2012. URL: http://investor.google.com/corporate/2012/ceo-letter.html.
10. Количество твитов: Geron, Tomio. Twitter's Dick Costolo: Twitter Mobile Ad Revenue Beats Desktop On Some Days // Forbes. — June 6, 2012. URL: http://www.forbes.com/sites/tomiogeron/2012/06/06/twitters-dick-costolo-mobile-ad-revenue-beats-desktop-on-some-days/.
11. Информация и количество данных: Hilbert, Martin. How to measure the world's technological capacity to communicate, store and compute information? / Martin and Hilbert Priscila Lopez // International Journal of Communication. — 2012. URL: http://www.ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742.
12. По оценкам за 2013 год, объем сохраненной информации равен 1,2 зеттабайта, из которых нецифровая информация составляет менее 2% (из интервью Гилберта Кукьеру).
13. Печатный станок и восемь миллионов книг (больше, чем было выпущено с момента основания Константинополя): Eisenstein, Elizabeth L. The Printing Revolution in Early Modern Europe. — Cambridge: Canto/Cambridge University Press, 1993. — P. 13–14.
14. Аналогия Питера Норвига. Из бесед с Норвигом о его труде The Unreasonable Effectiveness of Data (написанном в соавторстве), в частности: Norvig, Peter. The Unreasonable Effectiveness of Data // Лекция в Университете провинции Британская Колумбия. — Видео YouTube. — 23.09.2010. URL: http://www.youtube.com/watch?v=yvDCzhbjYWs.

15. Пикассо об изображениях в Ласко: Whitehouse, David. UK Science shows cave art developed early // BBC News Online. — October 3, 2001. URL: http://news.bbc.co.uk/1/hi/sci/tech/1577421.stm.

Назад к тексту

Глава 2

16. О Джеффе Йонасе и о том, что «говорят» данные: беседа с Джеффом Йонасом (декабрь 2010 года, Париж).
17. История переписей в США: US Census Bureau. The Hollerith Machine (онлайн-материал). URL: http://www.census.gov/history/www/innovations/technology/the_hollerith_tabulator.html (последнее посещение — 25.07.2012).
18. Вклад Неймана: Kruskal, William. Representative Sampling, IV: the History of the Concept in Statistics, 1895–1939 / William Kruskal and Frederick Mosteller // International Statistical Review. — 1980. — Vol. 48. — P. 169–195, 187–188. Знаменитая статья Неймана: Neyman, Jerzy. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection // Journal of the Royal Statistical Society. — 1934. — Vol. 97, No. 4 . — P. 558–625.
19. Выборки из 1100 результатов наблюдений достаточно. Пример см. в статье: Babbie, Earl. Practice of Social Research. — 12th ed., 2010. — P. 204–207.
20. Подводные камни опросов: Crossen, Cynthia. Fiasco in 1936 Survey Brought ‘Science’ To Election Polling // Wall Street Journal. — October 2, 2006. URL: http://online.wsj.com/public/article/SB115974322285279370-_rk13XDUHmIcnA8DYs5VUscZG94_20071001.html?mod=rss_free.
21. Влияние сотовых телефонов: Estimating the Cellphone Effect. — September 20, 2008. URL: http://www.fivethirtyeight.com/2008/09/estimating-cellphone-effect-22-points.html.
22. Генетическое секвенирование Стива Джобса: Isaacson, Walter. Steve Jobs. — 2011.

23. Google Flu Trends: прогнозирование на уровне городов с 75%-ной точностью: Dugas et al. Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics // CID Advanced Access. — January 8, 2012.
24. Эциони о временных данных: интервью Кукьеру (октябрь 2011 года).
25. Исполнительный директор компании Xoom: Rosenthal, Jonathan. Special report: International banking // The Economist. — May 19, 2012. — P. 7–8.
26. Корректировка боев сумо: Duggan, Mark. Winning Isn't Everything: Corruption in Sumo Wrestling / Mark Duggan & Steven D. Levitt // American Economic Review. — 2002. — Vol. 92. — P. 1594–1605. URL: http://pricetheory.uchicago.edu/levitt/Papers/DugganLevitt2002.pdf.
27. Замена выборок: Savage, Mike. The Coming Crisis of Empirical Sociology / Mike Savage & Roger Burrows // Sociology. — 2007. — Vol 41. — P. 885–899.
28. Об анализе исчерпывающих данных, полученных от оператора мобильной связи: Onnela, J.-P. et al. Structure and tie strengths in mobile communication networks // Proceedings of the National Academy of Sciences of the United States of America (PNAS). — May, 2007. — Vol. 104. — P. 7332–7336. URL: http://nd.edu/~dddas/Papers/PNAS0610245104v1.pdf

Назад к тексту

Глава 3

29. Кросби: Crosby, Alfred W. The Measure of Reality: Quantification and Western Society. — 1997.
30. Множество способов сослаться на IBM: Patil, D. J. Data Jujitsu: The Art of Turning Data into Product // O'Reilly Media. — July 2012. URL: http://oreillynet.com/oreilly/data/radarreports/data-jujitsu.csp?cmp=tw-strata-books-data-products.
31. Идея о том, что «2 + 2 = 3,9»: Hopkins, Brian. Expand Your Digital Horizon With Big Data / Brian Hopkins and Boris Evelson // Forrester.

— September 30, 2011.

32. Белый дом: Report To The President And Congress Designing A Digital Future: Federally Funded Research And Development In Networking And Information Technology // President's Council of Advisors on Science and Technology. — December, 2010. — P. 71. URL: http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf.
33. Таблица шахматных эндшпилей. Наиболее полная общедоступная таблица шахматных эндшпилей, названная в честь ее создателей (Nalimovtableset), охватывает все варианты игры при шести (и менее) фигурах. Ее размер превышает 7 терабайт, и главная задача — сжатие содержащейся в ней информации. См.: Nalimov, E. V. Space-efficient indexing of chess endgame tables / E. V. Nalimov, G. McC. Haworth, and E. A. Heinz // ICGA Journal. — 2000. — Vol. 23, no. 3. — P. 148–162.
34. Эффективность алгоритма: Banko, Michele. Scaling to Very Very Large Corpora for Natural Language Disambiguation / Michele Banko & Eric Brill // Microsoft Research. — 2001. — P. 3. URL: http://acl.ldc.upenn.edu/P/P01/P01–1005.pdf.
35. Демоверсия IBM: слова и цитаты: IBM. 701 Translator: Press release // IBM archives. — January 8, 1954. URL: http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html. См. также: Hutchins, John. The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954. — November, 2005.
36. Проект IBM Candide: Berger, Adam L. et al. The Candide System for Machine Translation // Proceedings of the 1994 ARPA Workshop on Human Language Technology. — 1994. URL: http://aclweb.org/anthology-new/H/H94/H94–1100.pdf.
37. Корпус Google из 95 миллиардов предложений: Franz, Alex. All Our N-gram are Belong to You / Alex Franz and Thorsten Brants // Google blog post. — August 3, 2006. URL: http://googleresearch.blogspot.co.uk/2006/08/all-our-n-gram-are-belong-to-you.html.

38. Цитата из статьи Норвига: Halevy, A. The Unreasonable Effectiveness of Data / A. Halevy, P. Norvig, and F. Pereira // IEEE Intelligent Systems. — Mar./Apr., 2009. — P. 8–12. Обратите внимание, что ее название — вариация на тему знаменитой статьи Юджина Вигнера The Unreasonable Effectiveness of Mathematics in the Natural Sciences, в которой он рассматривает, почему физику можно аккуратно выразить в математических формулах, но они плохо годятся для гуманитарных наук. См.: Wigner, E. The Unreasonable Effectiveness of Mathematics in the Natural Sciences // Comm. Pure and Applied Mathematics. — 1960. — Vol. 13, no. 1. — P. 1–14.
39. Коррозия труб и враждебная среда связи в компании BP: Clarabut, Jaclyn. Operations Making Sense of Corrosion // BP Magazine. — 2011. — Issue 2. URL: http://www.bp.com/liveassets/bp_internet/globalbp/globalbp_uk_english/reports_and_publications/bp_magazine/STAGING/local_assets/pdf/BP_Magazine_2011_issue2_text.pdf.
40. Кукьер: трудности считывания данных по беспроводной связи: Data, data, everywhere // The Economist. — February 27, 2010. Система, безусловно, не является непогрешимой: причиной пожара на нефтеперерабатывающем заводе BP Cherry Point в феврале 2012 года оказались ржавые трубы.

Назад к тексту

Глава 4

41. Цитата Маркуса: Marcus, James. Amazonia: Five Years at the Epicenter of the Dot.Com Juggernaut // The New Press. — June, 2004. — P. 199.
42. Линден: интервью Кукьеру (март 2012 года).
43. Информация о ценах Netflix: Amatriain, Xavier. Netflix Recommendations: Beyond the 5 stars (Part 1) / Xavier Amatriain and Justin Basilico // Блог Netflix. — 6.04.2012.
44. Walmart и Pop-Tarts: Hays, Constance L. What Wal-Mart Knows About Customers' Habits // NYT. — November 14, 2004.

45. Примеры прогнозных моделей FICO, Experian и Equifax: Thurm, Scott. Next Frontier in Credit Scores: Predicting Personal Behavior // Wall Street Journal. — October 27, 2011. URL: http://online.wsj.com/article/SB10001424052970203687504576655182086300912.html.
46. Прогнозные модели Aviva: Scism, Leslie. Insurers Test Data Profiles to Identify Risky Clients / Leslie Scism and Mark Maremont // Wall Street Journal. — November 19, 2010. URL: http://online.wsj.com/article/SB10001424052748704648604575620750998072986.html. См. также: Scism, Leslie. Inside Deloitte's Life-Insurance Assessment Technology / Leslie Scism and Mark Maremont // Wall Street Journal. — November 19, 2010.
47. Там же.
48. Аналитическая работа UPS: интервью Кукьера Джеку Левису (март, апрель и июль 2012 года).
49. Недоношенные младенцы (на основе интервью с Макгрегор в январе 2010-го и апреле 2012 гг.). См. также: McGregor, Carolyn. Next Generation Neonatal Health Informatics with Artemis / Carolyn McGregor, Christina Catley, Andrew James, James Padbury // User Centered Networked Health Care, European Federation for Medical Informatics. 115 / A. Moen et al. (eds.). — IOS Press, 2011. — P. 117. Некоторые материалы взяты из специального отчета The Economist (2010 год).
50. О корреляции между показателями счастья и дохода: Genes, Culture and Happiness / R. Inglehart and H.-D. Klingemann. — MIT Press, 2000.
51. О кори, расходах на здравоохранение и новых нелинейных инструментах корреляционного анализа: Reshef, David et al. Detecting Novel Associations in Large Data Sets // Science. — 2011. — Vol. 334. — P. 1518–1524.
52. Канеман: Kahneman, Daniel. Thinking, Fast and Slow. — 2011. — P. 74–75.
53. Мнение Андерсона: Anderson, Chris. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete // Wired. — June 2008.

URL: http://www.wired.com/science/discoveries/magazine/16–07/pb_theory/.

54. Андерсон отказывается от своих слов: NPR. Search and Destroy // National Public Radio. — July 18, 2008. URL: http://www.onthe-media.org/2008/jul/18/search-and-destroy/transcript/.

Назад к тексту

Глава 5

55. Цитата Мори из книги: Maury. Physical Geography of the Sea. — Introduction.
56. Континуум. Множество людей содействовали истории развития больших данных, начиная с машины Холлерита, первого электронного компьютера, транзистора или редких упоминаний самого термина примерно с 1970-х годов. К его чести, Стефан Вольфрам повел историю развития с количественных измерений и записи слов. Рэй Курцвейл примерно в 2000 году изобразил графически способность человека взаимодействовать с информацией, начиная с первобытных обществ.
57. Альфред В. Кросби: Crosby, Alfred V. The Measure of Reality: Quantification and Western Society, 1250–1600. — Cambridge University Press, 1997. — P. 113.
58. Там же. С. 111–113.
59. О счетных досках и абаке. Как указывает Кросби (с. 112), европейцы так и не переняли арабский абак, иначе, вероятно, они бы значительно дольше придерживались римских цифр. Подсчеты с помощью арабских цифр производятся в шесть раз быстрее, чем с помощью счетных досок: Murray, Alexander. Reason and Society in the Middle Ages. — Oxford University Press, 1978. — P. 166/454.
60. Корпорация Microsoft собирает данные о Wi-Fi-сетях: McCullagh, Declan. Microsoft curbs Wi-Fi location database // CNET. — August 1, 2011.
61. Исследование Пентлэнда в журнале WSJ: Hotz, Robert Lee. The Really Smart Phone // Wall Street Journal. — April 22, 2011. URL:

http://online.wsj.com/article/SB10001424052748704547604576263261679848814.html.

62. Ссылка Игла на исследование по трущобам: Eagle, Nathan. Big Data, Global Development, and Complex Systems // Santa Fe Institute. — May 5, 2010. URL: http://www.youtube.com/watch?v=yaivtqlu7iM.
63. Данные Twitter: Tsotsis, Alexia. Twitter Is At 250 Million Tweets Per Day, iOS 5 Integration Made Signups Increase 3x // TechCrunch. — October 17, 2011.
64. Хедж-фонды используют Twitter: Cukier, Kenneth. Tracking social media: The mood of the market // The Economist online. — June 28, 2012. URL: http://www.economist.com/blogs/graphicdetail/2012/06/tracking-social-media.
65. Twitter и прогнозирование кассовых сборов Голливуда: Asur, Sitaram and Huberman, Bernardo A. Predicting the Future With Social Media. — HP.
66. Twitter и глобальные настроения: Golder, Scott A. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures / Scott A. Golder and Michael W. Macy // Science. — Vol. 333, no. 6051. — September 30, 2011. — P. 1878–1881.
67. Twitter и прививки от гриппа: Salathé, Marcel and Khandelwal, Shashank // PLoS Computational Biology. — October 2011.
68. Данные в рамках подхода Quantified Self («измерение себя»): Counting every moment // The Economist. — March 3, 2012.
69. Заявка на выдачу патента США 20090287067, 19.11.2009. Jesse Lee Dorogusker, Anthony Fadell, Donald J. Novotney and Nicholas R. Kalayjian. Наушники-вкладыши Apple для биоизмерений. Integrated sensors for tracking performance metrics // Правополучатель: Apple. Дата заявки: 23.07.2009.
70. Derawi Biometrics. Your walk is your PIN-code // Press release. — Feb. 21, 2011. URL: http://biometrics.derawi.com/?p=175.
71. Информация об iTrem: см. страницу проекта iTrem исследовательского центра Landmarc Research Center на сайте

Georgia Tech: http://eosl.gtri.gatech.edu/Capabilities/LandmarcResearchCenter/LandmarcProjects/iTrem/tabid/798/Default.aspx.

72. Исследователи из Киото о трехосевых акселерометрах: iMedical-Apps Team. Gait analysis accuracy: Android app comparable to standard accelerometer methodology // mHealth. — March 23, 2012.

Назад к тексту

Глава 6

73. История Луиса фон Ана (на основе интервью фон Ана Кукьеру в 2010 и 2011 годах). См. также: Ahn, Luis von. Luis von Ahn: Expert Q&A // NOVA scienceNOW. — July 6, 2009. Адрес в интернете: http://www.pbs.org/wgbh/nova/tech/von-ahn-captcha.html.

Thompson, Clive. For Certain Tasks, the Cortex Still Beats the CPU // Wired. — June 25, 2007. URL: http://www.wired.com/techbiz/it/magazine/15-07/ff_humancomp?currentPage=all.

Scanlon, Jessie. Luis von Ahn: The Pioneer of ‘Human Computation’ // Businessweek. — November 3, 2008. URL: http://www.businessweek.com/stories/2008-11-03/luis-von-ahn-the-pioneer-of-human-compu-tation-businessweek-business-news-stock-market-and-financial-advice.

Техническое описание технологии reCaptcha см. в статье: Ahn, Luis von et al. reCaptcha: Human-Based Character Recognition via Web Security Measures // Science. — September 12, 2008. — Vol. 321, no. 5895. — P. 1465–1468. URL: http://www.sciencemag.org/content/321/5895/1465.abstract.

74. Смит и булавочная фабрика: Smith, Adam. An Inquiry into the Nature and Causes of the Wealth of Nations. — 1776. — Book I, chapter one.

75. Хранение. Mayer-Schönberger, Viktor. Delete (second edition). — P. 63.

76. О том, сколько электроэнергии потребляют электромобили: IBM. IBM, Honda, and PG&E Enable Smarter Charging for Electric Vehicles

// Press release. — April 12, 2012. URL: http://www-03.ibm.com/press/us/en/pressrelease/37398.wss. См. также: Luthy, Clay. Guest Perspective: IBM Working with PG&E to Maximize the EV Potential // PGE Currents magazine. — April 13, 2012. URL: http://www.pgecurrents.com/2012/04/13/ibm-working-with-pge-to-maximize-the-ev-potential.

77. Amazon и данные AOL: интервью Андреаса Вайгенда, 2010 год.
78. Программное обеспечение Nuance и Google: специальный отчет The Economist, 2010 год.
79. Логистическая компания: Brown, Brad. Are you ready for the era of “big data”? / Brad Brown, Michael Chui, and James Manyika // McKinsey Quaterly. — October 2011. — P. 10.
80. Исследование Датского онкологического общества: Frei, Patrizia et al. Use of mobile phones and risk of brain tumours: update of Danish cohort study // BMJ. — 2011. URL: http://www.bmj.com/content/343/bmj.d6387.
81. GPS-записи и самоуправляемые автомобили Google Street View: Kirwan, Peter. This car drives itself // Wired UK. — January 2012. URL: http://www.wired.co.uk/magazine/archive/2012/01/features/this-car-drives-itself?page=all.
82. Цитата Мунди: специальный отчет The Economist на основе интервью (декабрь 2009 года).
83. Компания Barnes & Noble проанализировала данные со своих устройств для чтения электронных книг Nook: Alter, Alexandra. Your E-Book Is Reading You // WSJ. — June 29, 2012. URL: http://online.wsj.com/article/SB10001424052702304870304577490950051438304.html.
84. Открытая политика правительства Обамы: Barack Obama. Presidential memorandum. — White House, January 21, 2009.
85. О стоимости данных Facebook. Чтобы ознакомиться с превосходным исследованием несоответствия между рыночной и балансовой стоимостью IPO Facebook, см. статью: Laney, Doug. To

Facebook You’re Worth $80.95 // WSJ. — May 3, 2012. URL: http://blogs.wsj.com/cio/2012/05/03/to-facebook-youre-worth-80-95/.

86. Разрыв в стоимости материальных и нематериальных активов: Samek, Steve M. Prepared Testimony: Hearing on Adapting a 1930’s Financial Reporting Model to the 21st Century // US Senate Committee on Banking, Housing and Urban Affairs, Subcommittee on Securities. — July 19, 2000.
87. Там же.
88. Стоимость нематериальных активов: Strategy Maps: Converting Intangible Assets into Tangible Outcomes / Robert S. Kaplan, David P. Norton. — Harvard Business Review Press, 2004. — P. 4–5.
89. Данные как корпоративный актив американских операторов беспроводной связи: со слов высокопоставленного чиновника в межправительственном учреждении, которому об этом сообщил руководитель незадолго до интервью Кукьеру (ноябрь 2011 года).

Назад к тексту

Глава 7

90. Информация о сайте Decide.com. Электронная переписка Кукьера с Эциони (май 2012 года).
91. Отчет Дж. Маккинси: Manyika, James et al. Big data: The next frontier for innovation, competition, and productivity // McKinsey Global Institute. — May 2011. URL: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation. — P. 10.
92. Цитата Хэла Вэрэйна из интервью Кукьеру (декабрь 2009 года).
93. Цитата Карла де Маркена, ITA, из электронной переписки с Кукьером (май 2012 года).
94. Программа SpendingPulse компании MasterCard: интервью Кернса Кукьеру, а также на конференции The Economist’s The Ideas Economy: Information, Санта-Клара, Калифорния. — 8.06.2011.
95. Консалтинговая компания Accenture и город Сент-Луис (штат Миссури): интервью Кукьеру (февраль 2007 года).

96. О функции Google Street View и самоуправляемых автомобилях: Kirwan, Peter. This car drives itself // Wired UK. — January 4, 2012.
97. Автомобильные микропроцессоры и электроника, которая составляет треть стоимости автомобиля: Kirwan, Peter. This car drives itself // Wired UK. — January 4, 2012. URL: http://www.wired.co.uk/magazine/archive/2012/01/features/this-car-drives-itself?page=all.
98. То, что Мори называл «плавающими обсерваториями»: труд Мори «Физическая география моря».
99. Информация об институте Health Care Cost Institute: Kliff, Sarah. A database that could revolutionize health care // Washington Post. — May 21, 2012.
100. Единая интеллектуальная система Microsoft Amalga: Microsoft Expands Presence in Healthcare IT Industry With Acquisition of Health Intelligence Software Azyxxi // Microsoft press release. — July 26, 2006. URL: http://www.microsoft.com/en-us/news/press/2006/jul06/07–26azyxxiacquisitionpr.aspx.
101. Объем данных о пациентах, собранный Макгрегор более чем за десяток лет: из интервью Кукьеру (май 2012 года).
102. Цитата Голдблума из интервью Кукьеру (март 2012 года).
103. Анализ данных в компании Zynga: Wingfield, Nick. Virtual Products, Real Profits: Players Spend on Zynga's Games, but Quality Turns Some Off // Wall Street Journal. — September 9, 2011. URL: http://online.wsj.com/article/SB10001424053111904823804576502442835413446.html.
104. Цитата Рудин из Zynga: из интервью Рудин с Нико Вэше, цит. в статье: Simply Seven: Seven Ways to Create a Sustainable Internet Business / Erik Schlie, Jörg Rheinboldt, NikoWaesche. — Palgrave Macmillan, 2011.
105. Исследование Э. Бриньолфссона: Brynjolfsson, Erik. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? / Erik Brynjolfsson, Lorin Hitt and Heekyung Kim // Unpublished working paper. — April 2011.

106. Brynjolfsson, Erik. Scale without Mass: Business Process Replication and Industry Dynamics / Erik Brynjolfsson, Andrew McAfee, Michael Sorell, and Feng Zhu // HBS working paper. — September 2006. URL: http://www.hbs.edu/research/pdf/07-016.pdf also http://hbswk.hbs.edu/item/5532.html.
107. О постепенном увеличении масштаба держателей данных, см. также: Bakos, Yannis. Bundling Information Goods: Pricing, Profits, and Efficiency / Yannis Bakos and Erik Brynjolfson // Management Science. — Dec. 1999. — Vol. 45. — P. 1613–1630.
108. Филип Эванс: интервью авторам (2011 и 2012 гг.).

Назад к тексту

Глава 8

109. Камеры видеонаблюдения рядом с домом Оруэлла: Orwell, George. Big Brother is watching your house // The Evening Standard. — March 31, 2007. URL: http://www.thisislondon.co.uk/news/george-orwell-big-brother-is-watching-your-house-7086271.html.
110. О компаниях Equifax и Experian: Solove, Daniel J. The Digital Person: Technology and Privacy in the Information Age // NYU Press. — 2004. — P. 20–21.
111. Информация о компании IBM и холокосте: Black. Edwin. IBM and the Holocaust. — Crown, 2003.
112. Информация о конфиденциальности и интеллектуальных индикаторах: McNeil, Sonia K. Privacy And The Modern Grid // Harvard Journal of Law & Technology. — 2011. — Vol. 25, no. 1. URL: http://jolt.law.harvard.edu/articles/pdf/v25/25HarvJLTech199.pdf.
113. Компания Netfix вычислила частных лиц: Singel, Ryan. Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims // Wired. — December 17, 2009. URL: http://www.wired.com/threatlevel/2009/12/netflix-privacy-lawsuit/.
114. О выпуске данных компании Netflix: Narayanan, Arvind. Robust De-Anonymization of Large Sparse Datasets / Arvind Narayanan and Vitaly Shmatikov // Proceedings of the IEEE Symposium on Security

and Privacy. — 2008. — P. 111. URL: http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf. Arvind Narayanan and Vitaly Shmatikov. How to Break the Anonymity of the Netflix Prize Dataset. — ARVIX. — October 16, 2006.

115. О слабых местах в структуре анонимизации: Ohm, Paul. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization // 57 UCLA Law Review 1701. — 2010.
116. Об анонимности социального графа: Backstrom, Lars. Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography / Lars Backstrom, Cynthia Dwork, and Jon Kleinberg // Communication of the Association of Computing Machinery December. — 2011. — P. 133.
117. Автомобильные «черные ящики»: “Vehicle data recorders” Watching your driving // The Economist. — June 23, 2012. URL: http://www.economist.com/node/21557309.
118. Сбор данных в АНБ: Priest, Dana. A hidden world, growing beyond control / Dana Priest and William Arkin // Washington Post. — July 19, 2010. URL: http://projects.washingtonpost.com/top-secret-america/articles/a-hidden-world-growing-beyond-control/print/.
119. Gonzalez, Juan. Whistleblower: The NSA Is Lying–U.S. Government Has Copies of Most of Your Emails // Democracy Now. — April 20, 2012. URL: http://www.democracynow.org/2012/4/20/whistleblower_the_nsa_is_lying_us.
120. Binney, William. Sworn declaration in the case of Jewel v. // NSA. — July 2, 2012. URL: http://publicintelligence.net/binney-nsa-declaration/.
121. Примеры прогностического полицейского контроля: Vlahos, James. The Department Of Pre-Crime // Scientific American. — January 2012.
122. Там же.
123. Книга Харкорта: Harcourt, Bernard E. Against Prediction: Profiling, Policing, And Punishing In An Actuarial Age. — University of Chicago Press, 2006.

124. О пристрастии Макнамары к данным. Rosenzweig, Phil. Robert S. McNamara and the Evolution of Modern Management // Harvard Business Review. — December 2010. URL: http://hbr.org/2010/12/robert-s-mcnamara-and-the-evolution-of-modern-management/ar/pr.
125. Успех вундеркиндов во Второй мировой войне: Byrne, John. The Whiz Kids. — Doubleday, 1993.
126. О работе Макнамары в компании Ford: Halberstam, David. The Reckoning // William Morrow. — September 1986. — P. 222–245.
127. Статистические данные и цитаты из книги Киннарда: Kinnard, Douglas. The War Managers. — University Press of New England, 1977. — P. 7–25.

    В данном разделе использованы материалы из интервью по электронной почте с д-ром Киннардом при помощи его ассистента, за что авторы выражают свою благодарность.
128. Практика найма в Google, см.: Edwards, Douglas. I'm Feeling Lucky: The Confessions of Google Employee Number 59. — Houghton Mifflin Harcourt, 2011. — P. 9. См. также: Levy, Steven. In the Plex. — Simon & Schuster, 2011. — P. 140–141.
129. Тест 41 оттенка синего: Holson, Laura M. Putting a Bolder Face on Google // NYT. — March 1, 2009. URL: http://www.nytimes.com/2009/03/01/business/01marissa.html.
130. Уход в отставку ведущего дизайнера Google, цитата выписана без многоточия (для удобства чтения): Bowman, Doug. Goodbye, Google // Публикация в блоге. — March 20, 2009. Адрес в интернете: http://stopdesign.com/archive/2009/03/20/goodbye-google.html.
131. Цитата С. Джобса: Lohr, Steve. Can Apple Find More Hits Without Its Tastemaker? // NYT. — January 18, 2011. URL: http://www.nytimes.com/2011/01/19/technology/companies/19innovate.html.
132. Ссылка на книгу «Благими намерениями государства». Scott, James. Seeing Like a State: How Certain Schemes to Improve the

Human Condition Have Failed. — Yale University Press, 1998.

133. Цитата из речи Р. Макнамары, полученной из колледжа Милсапс в Джексоне, Миссисипи. — HBR 2010. (В цит. труде.)

Назад к тексту

Глава 9

134. О собрании книг библиотеки Кембриджского университета: Drogin, Marc. Anathema!: Medieval Scribes and the History of Book Curses. — Allanheld & Schram, 1983. — P. 37.
135. О сроке истечения данных: Mayer-Schönberger, Viktor. Delete. The Virtue of Forgetting in the Digital Age. — Princeton University Press, 2009.
136. «Дифференциальная конфиденциальность»: Dwork, Cynthia. A Firm Foundation for Private Data Analysis // Communications of the Association of Computing Machinery. — January 2011, 86.
137. Facebook и дифференциальная конфиденциальность: Chin, A. Differential Privacy as a Response to the Reidentification Threat / A. Chin & A. Klinefelter // The Facebook Advertiser Case Study. — 90 North Carolina Law Review Page. — 2012. Haeberlen, A. et al. Differential Privacy Under Fire. URL: http://www.cis.upenn.edu/~ahae/papers/fuzz-sec2011.pdf.
138. Роль немецких представителей защиты корпоративных данных: Mayer-Schönberger, Viktor. Beyond Privacy, Beyond Rights — Towards a “Systems” Theory of Information Governance // California Law Review. — 2010. — Vol. 1853. — P. 98.
139. О регулировании путем прозрачности: Full Disclosure The Perils and Promise of Transparency / Archon Fung, Mary Graham and David Weil. — Cambridge University Press, 2007.

Назад к тексту

Глава 10

140. Использование больших данных применительно к трущобам и моделям передвижения беженцев: Eagle, Nathan // Santa Fe Institute, 2008. (В цит. труде.)

141. Восприятие времени: Anderson, Benedict. Imagined Communities. Revised Edition. — Verso, 1991.
142. Эксперимент CERN и хранение данных: специальный отчет The Economist (2010 год).
143. Компьютерная система Apollo 11: Mindell, David A. Digital Apollo: Human and Machine in Spaceflight. — MIT Press, 2008.

Назад к тексту

[1] Директор исследовательского центра имени Тьюринга при Вашингтонском университете.

[2] Jeopardy! («Рискуй!») — телеигра, популярная во многих странах мира. Российский аналог — «Своя игра». Здесь и далее *прим. ред*.

[3] Walmart — американская компания-ретейлер, управляющая крупнейшей в мире розничной сетью.

[4] CapitalOne — американская банковская холдинговая компания, специализирующаяся на кредитах.

[5] «Человек, который изменил всё» (Moneyball) — биографическая спортивная драма режиссера Беннетта Миллера. На русском языке издана книга: *Льюис М.* Moneyball. Как математика изменила самую популярную спортивную лигу в мире. М. : Манн, Иванов и Фербер, 2014.

[6] Линия Мажино — система французских укреплений на границе с Германией.

[7] В Древнем Риме: перепись граждан с указанием имущества для определения их социально-политического, военного и податного положения.

[8] 23andme — частная компания в Маунтин-Вью, Калифорния, где разрабатываются новые биотехнологические методы.

[9] *Левитт С., Дабнер С.* Фрикономика. М. : Манн, Иванов и Фербер, 2011.

[10] Эндшпиль — заключительная часть шахматной партии.

[11] Лингвистическим корпусом называют совокупность текстов, собранных в соответствии с определенными принципами, размеченных по определенному стандарту и обеспеченных специализированной поисковой системой. Термин введен в употребление в 1960-х годах в связи с развитием практики создания корпусов, которому начиная с 1980-х способствовало развитие вычислительной техники.

[12] Billion Prices Project — проект в рамках учебной инициативы, в котором используются цены, ежедневно собираемые на сотнях сайтов розничных торговцев по всему миру, для проведения экономических исследований.

[13] Chief Executive Officer — главный исполнительный директор.

[14] *Уоттс Д.* Здравый смысл врет. Почему не надо слушать свой внутренний голос. М. : Эксмо, 2012.

[15] Дана Бойд — исследователь проблем молодежи и адвокат (Microsoft Research; Нью-Йоркский университет СМИ культуры и коммуникации; Гарвардский Беркман-центр).

[16] Журналист, персонаж фильма «Конец игры».

[17] Коммодор — офицерское звание в составе военно-морских сил; выше звания капитана корабля и ниже контр-адмирала.

[18] Абак — счетная доска, применявшаяся для арифметических вычислений в Древней Греции и Древнем Риме.

[19] Тихо Браге — датский астроном эпохи Возрождения. Первым в Европе начал проводить систематические и высокоточные астрономические наблюдения, на основании которых Кеплер вывел законы движения планет.

[20] На русском языке издана книга: *Паттерсон С.* Кванты. Как волшебники от математики заработали миллиарды и чуть не обрушили фондовый рынок. М. : Манн, Иванов и Фербер, 2014.

[21] Фонд Макартуров — один из крупнейших благотворительных фондов США.

[22] Имеется в виду пример разделения труда: один рабочий тянет проволоку, другой выпрямляет ее, третий обрезает, четвертый заостряет конец, пятый обтачивает один конец для насаживания головки; изготовление самой головки требует двух или трех самостоятельных операций; насадка ее составляет особую операцию, полировка булавки — другую; самостоятельной

операцией является даже завертывание готовых булавок в пакетики.

[23] Модель ценообразования опционов Блэка–Шоулза определяет теоретическую цену на европейские опционы, подразумевая, что если базовый актив торгуется на рынке, то цена опциона на него устанавливается самим рынком.

[24] Вьетконг — сокращение от «вьетнам конг шан» — «вьетнамский коммунист». Так в западной печати называли политических и общественных деятелей Демократической Республики Вьетнам и южновьетнамских партизан.

[25] Пещера Платона, Платонова пещера — аллегория, использованная Платоном в трактате «Государство» для пояснения своего учения об идеях.

Над книгой работали

Ответственный редактор *Мария Красовская*
Арт-директор *Алексей Богомолов*
Редактор *Татьяна Собко*
Верстка *Вячеслав Лукьяненко*
Корректоры *Лев Зелексон, Юлия Молокова*
Дизайн переплета *Сергей Хозин*
ООО «Манн, Иванов и Фербер»
mann-ivanov-ferber.ru
Электронная версия книги
подготовлена компанией Webkniga, 2013
webkniga.ru

Максимально полезные книги от издательства «Манн, Иванов и Фербер»

Наши электронные книги: http://www.mann-ivanov-ferber.ru/ebooks/

Если у вас есть замечания и комментарии к содержанию, переводу, редактуре и корректуре, то просим написать на be_better@m-i-f.ru, так мы быстрее сможем исправить недочеты.

Заходите в гости: http://www.mann-ivanov-ferber.ru/
Наш блог: http://blog.mann-ivanov-ferber.ru/
Мы в Facebook: http://www.facebook.com/mifbooks
Мы ВКонтакте: http://vk.com/mifbooks
Наш Twitter: https://twitter.com/mifbooks

Дерево знаний:
http://www.mann-ivanov-ferber.ru/promo/derevo-znanij/
Предложите нам книгу:
http://www.mann-ivanov-ferber.ru/about/predlojite-nam-knigu/
Ищем правильных коллег:
http://www.mann-ivanov-ferber.ru/about/job/

Для корпоративных клиентов:
Полезные книги в подарок:
http://www.mann-ivanov-ferber.ru/promo/presents-b2b/
Книги ищут поддержку:
http://www.b2b.mann-ivanov-ferber.ru/sponsorship/promo/
Корпоративная библиотека:
http://www.b2b.mann-ivanov-ferber.ru/corp-library/

Оглавление