Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим

Автор:
Кеннет Кукьер

Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим

Виктор Майер-Шенбергер

Кеннет Кукьер

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.

Виктор Майер-Шенбергер, Кеннет Кукьер

Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим

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 - 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 (http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html)]. 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 - 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.], 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 - 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.).].

В 2003 году Орен Эциони[4 - Директор исследовательского центра имени Тьюринга при Вашингтонском университете.] 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»[5 - 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 (http://knight.cis.temple.edu/~yates//papers/hamlet-kdd03.pdf).].

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[6 - 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 (http://blog.seattlepi.com/venture/2008/04/17/secret-farecast-buyer-is-microsoft/?source=mypi).], 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. Эционi 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 бо?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[7 - Астрономия и секвенирование ДНК. Специальный отчет в журнале The Economist (см. выше): Data, data everywhere // The Economist. – February 27, 2010. – P. 1–14.].

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[8 - 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 (http://www.nytimes.com/2011/12/01/business/dna-sequencing-caught-in-deluge-of-data.html?pagewanted=all).]. 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[9 - 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 (http://sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm).]. А 800 миллионов ежемесячных пользователей службы YouTube компании Google каждую секунду загружают видео длительностью более часа[10 - Статистика YouTube: Page, Larry. Update from the CEO // Google, April 2012. URL: http://investor.google.com/corporate/2012/ceo-letter.html (http://investor.google.com/corporate/2012/ceo-letter.html).]. Количество сообщений в Twitter увеличивается приблизительно на 200 % в год и к 2012 году превысило 400 миллионов твитов в день[11 - 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/ (http://www.forbes.com/sites/tomiogeron/2012/06/06/twitters-dick-costolo-mobile-ad-revenue-beats-desktop-on-some-days/).].

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[12 - 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 (http://www.ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742).]. 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 %[13 - По оценкам за 2013 год, объем сохраненной информации равен 1,2 зеттабайта, из которых нецифровая информация составляет менее 2 % (из интервью Гилберта Кукьеру).].

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[14 - 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.]. 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[15 -

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 (http://www.youtube.com/watch?v=yvDCzhbjYWs).]. 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»[16 - 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 (http://news.bbc.co.uk/1/hi/sci/tech/1577421.stm).].

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[17 - Jeopardy! («Рискуй!») – телеигра, популярная во многих странах мира. Российский аналог – «Своя игра». Здесь и далее прим. ред.], – 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[18 - Walmart – американская компания-ретейлер, управляющая крупнейшей в мире розничной сетью.] и CapitalOne[19 - CapitalOne – американская банковская холдинговая компания, специализирующаяся на кредитах.] 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»[20 - «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.]: 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»[21 - Линия Мажино – система французских укреплений на границе с Германией.]. 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.

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[22 - О Джеффе Йонасе и о том, что «говорят» данные: беседа с Джеффом Йонасом (декабрь 2010 года, Париж).].

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»[23 - В Древнем Риме: перепись граждан

с указанием имущества для определения их социально-политического, военного и податного положения.] (англ. census) – происходит от латинского термина censere, что означает «оценивать».

Конец ознакомительного фрагмента.

notes

Примечания

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 (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

Директор исследовательского центра имени Тьюринга при Вашингтонском университете.

5

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 (http://knight.cis.temple.edu/~yates//papers/hamlet-kdd03.pdf).

6

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 (http://blog.seattlepi.com/venture/2008/04/17/secret-farecast-buyer-is-microsoft/?source=mypi).

7

Астрономия и секвенирование ДНК. Специальный отчет в журнале The Economist (см. выше): Data, data everywhere // The Economist. – February 27, 2010. – P. 1–14.

8

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 (http://www.nytimes.com/2011/12/01/business/dna-sequencing-caught-in-deluge-of-data.html?pagewanted=all).

9

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 (http://sec.gov/Archives/edgar/data/1326801/000119312512034517/d287954ds1.htm).

10

Статистика YouTube: Page, Larry. Update from the CEO // Google, April 2012. URL: http://investor.google.com/corporate/2012/ceo-letter.html (http://investor.google.com/corporate/2012/ceo-letter.html).

11

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/ (http://www.forbes.com/sites/tomiogeron/2012/06/06/twitters-dick-costolo-mobile-ad-revenue-beats-desktop-on-some-days/).

12

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 (http://www.ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742).

13

По оценкам за 2013 год, объем сохраненной информации равен 1,2 зеттабайта, из которых нецифровая информация составляет менее 2 % (из интервью Гилберта Кукьеру).

14

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.

15

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 (http://www.youtube.com/watch?v=yvDCzhbjYWs).

16

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 (http://news.bbc.co.uk/1/hi/sci/tech/1577421.stm).

17

Jeopardy! («Рискуй!») – телеигра, популярная во многих странах мира. Российский аналог – «Своя игра». Здесь и далее прим. ред.

18

Walmart – американская компания-ретейлер, управляющая крупнейшей в мире розничной сетью.

19

CapitalOne – американская банковская холдинговая компания, специализирующаяся на кредитах.

20

«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.

21

Линия Мажино – система французских укреплений на границе с Германией.

22

О Джеффе Йонасе и о том, что «говорят» данные: беседа с Джеффом Йонасом (декабрь 2010 года, Париж).

23

В Древнем Риме: перепись граждан с указанием имущества для определения их социально-политического, военного и податного положения.
