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»* (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

---

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

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* 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,

---

* 23andme — частная компания в Маунтин-Вью, Калифорния, где разрабатываются новые биотехнологические методы.

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»*.

---

* *Левитт С., Дабнер С.* Фрикономика. М. : Манн, Иванов и Фербер, 2011.

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.

издательство
МАНН, ИВАНОВ И ФЕРБЕР
Максимально полезные книги