БЕСТСЕЛЛЕР THE NEW YORK TIMES И AMAZON

СЕТ СТИВЕНС-ДАВИДОВИЦ

ВСЕ ЛГУТ

ПОИСКОВИКИ, BIG DATA И ИНТЕРНЕТ ЗНАЮТ О ВАС ВСЕ

СЕНСАЦИОННЫЕ ДАННЫЕ О ТОМ, ЧТО ЛЮДИ ДУМАЮТ И ДЕЛАЮТ **НА САМОМ ДЕЛЕ**

18+

IT BEST SELLER


BEST
SELLER

SETH STEPHENS-DAVIDOWITZ

EVERYBODY LIES:

BIG DATA, NEW DATA, AND WHAT THE INTERNET CAN TELL US ABOUT WHO WE REALLY ARE

СЕТ СТИВЕНС-ДАВИДОВИЦ

ВСЕ ЛГУТ

ПОИСКОВИКИ, BIG DATA И ИНТЕРНЕТ ЗНАЮТ О ВАС ВСЕ

БОМБОРА™
Москва 2018

УДК 316.3+004.738.5
ББК 60.5+32.973.202
С80

Seth Stephens-Davidowitz
EVERYBODY LIES

Стивенс-Давидовиц, Сет.

С80 Все лгут. Поисковики, Big Data и Интернет знают о вас все / Сет Стивенс-Давидовиц ; [пер. с англ. Л.И. Степановой]. — Москва : Эксмо, 2018. — 384 с. — (IT бестселлер).

ISBN 978-5-04-090836-3

Автор книги, специалист Google по Data Science, провел исследование, опираясь на науку о больших данных (Big Data), а также данные, которые может предоставить исследователю Интернет. В результате он получил сенсационные данные, полностью переворачивающие современные представления об обществе, в котором мы живем.

УДК 316.3+004.738.5
ББК 60.5+32.973.202

ISBN 978-5-04-090836-3

Оглавление

ЧАСТЬ III. БОЛЬШИЕ ДАННЫЕ: ОБРАЩАТЬСЯ С ОСТОРОЖНОСТЬЮ

ВСТУПЛЕНИЕ

Некогда философы мечтали о «микроскопе для мозга» — мифическом устройстве, отображающем на экране мысли человека. Социологи же активно искали инструменты, позволяющие понять действия человека. За время моей работы в качестве экспериментального психолога в моду входили различные инструменты, которые быстро разочаровывали ученых. Я перепробовал их все — рейтинговые шкалы, время реакции, расширение зрачка, функциональную нейровизуализацию, даже изучение пациентов, страдающих эпилепсией (они были рады скоротать время за экспериментами в ожидании приступа).

Но ни один из этих методов не позволил беспрепятственно заглянуть в разум. Проблема заключалась в необходимости грубого компромисса. Человеческие мысли — сложносоставное явление. В отличие от Вуди Аллена, который сводит «Войну и мир» к паре предложений, мы не просто думаем: «Это история о нескольких русских». Ученому трудно проанализировать предложения во всей их многомерной запутанности. Конечно, когда люди изливают свои души, мы можем наконец

постичь все богатство их потока сознания. Но монологи все равно не являются идеальным набором данных для тестирования гипотез. С другой стороны, если мы сосредоточимся на измерениях, легко поддающихся количественной оценке — таких как время реакции человека на слова или фотографии, — то сможем сформировать статистику. Но тем самым мы сведем сложную текстуру сознания к одному числу. Даже самые изощренные методики нейровизуализации могут рассказать нам, как мысль распределяется в 3D-пространстве, но не расскажет, о чем эта мысль.

Помимо этого, ученые-социологи учитывали действие закона малых чисел — Амос Тверски и Даниэль Канеман дали это название заблуждению, заключающемуся в том, что общие черты будут отражены в любой выборке населения, какой бы малой она ни была. Даже самые большие специалисты в области математики порой весьма печально ошибаются относительно того, сколько объектов нужно взять для исследования, прежде чем можно будет абстрагироваться от случайных отклонений данных и обобщить результат для всех американцев, не говоря уже обо всех Homo sapiens. Это тем более трудно, когда образец собирается по принципу удобства, например предлагая деньги на пиво второкурсникам.

Эта книга — о совершенно новом способе изучения сознания. Конечно, большие данные, полученные в результате интернет-поиска и других онлайн-исследований, — не энцефалоскоп. Но Сет Стивенс-Давидович показывает, что они дают удивительную возможность по-новому взглянуть на психику человека. Уединившись со своей клавиатурой, люди делают довольно странные

признания. Иногда потому (как на сайтах знакомств или при поиске профессиональных советов), что это имеет реальные последствия. А в других случаях потому, что эти действия, наоборот, не приводят ни к каким последствиям и люди могут раскрыться, признаться в наличии того или иного желания или страха без опасения, что кто-то отреагирует на это с ужасом.

В любом случае, люди не просто нажимают на кнопку или поворачивают ручку, но и набирают триллионы последовательностей символов, чтобы изложить свои мысли во всех их взрывоопасных комбинациях. Эти данные поступают из всех слоев общества. При этом люди оставляют цифровые следы, которые легко агрегировать и анализировать, принимая участие в незаметных экспериментах, меняющих стимулы и суммирующих ответы в реальном времени. И они с радостью предоставляют эти данные в огромных количествах. «Все лгут» — это больше, чем доказательство подобной концепции. Раз за разом открытия Стивенса-Давидовица переворачивали с ног на голову мои представления о согражданах и собственной стране. Откуда у Дональда Трампа столь неожиданная поддержка? В 1976 году Энн Лэндерс спросила своих читателей, сожалеют ли они о том, что у них есть дети — и была шокирована: большинство ответов оказались положительными. Не была ли она введена в заблуждение нерепрезентативной выборкой? Действительно ли интернет виноват в кризисе конца 2010-х годов — «информационном пузыре»? Что приводит к преступлениям на почве ненависти? Правда ли, что люди ищут шутки, чтобы посмеяться? Хотя мне нравится думать, что ничто не может меня шокировать, я все же

был в шоке от того, как в интернете раскрывается человеческая сексуальность — в том числе меня поразило открытие, что каждый месяц определенное количество женщин ищет «трахание плюшевых игрушек». Никакой эксперимент с использованием времени реакции, расширения зрачка или функциональной нейромедицины не смог бы никогда вскрыть этот факт.

Книга «Все лгут» обязательно понравится всем. Стивенс-Давидовиц с его неутомимым любопытством и терпением указывает новый путь для общественных наук XXI века. При наличии такого бесконечно увлекательного окна в мир человеческих страстей кому будет нужен энцефалоскоп?

Стивен Пинкер
Доктор наук, преподаватель MIT, автор книги «Чистый лист. Природа человека. Кто и почему отказывается признавать ее сегодня», 2017 г.

ПРЕДИСЛОВИЕ

КОНТУРЫ РЕВОЛЮЦИИ

«Разумеется, он проиграет», — сказали они.

По результатам республиканских предварительных выборов 2016 года эксперты пришли к выводу, что у Дональда Трампа нет никаких шансов, поскольку он оскорбил все возможные меньшинства. Опросы показали, сколь малое число американцев одобряет такое посягательство на их права.

Большинство опрошенных экспертов в то время также считали, что Трамп проиграет на всеобщих выборах. Слишком многие потенциальные избиратели говорили, что его манеры и взгляды вызывают у них отвращение.

Однако были факты, указывавшие на то, что на самом деле Трамп может выиграть как предварительные партийные, так и всеобщие выборы. И эти подсказки можно было найти в интернете.

Я эксперт в области интернет-данных. Ежедневно я отслеживаю цифровые следы людей, перемещающихся по ссылкам во всемирной паутине. По тому, на какие ссылки или клавиши они нажимают, я пытаюсь понять, чего они действительно хотят, что делают и кто они (да и мы все) есть на самом деле. Хочу рассказать, как я встал на этот необычный путь.

История началась — теперь кажется, что давным-давно, — с президентских выборов 2008 года. Социологи тогда вели долгие дискуссии: насколько сильны расовые предрассудки в Америке?

Барак Обама был выдвинут как первый афроамериканский кандидат в президенты США от лидирующей партии. Он победил, и довольно легко. Опросы показали, что раса не была тем фактором, который влиял на выбор американцев. Институт Гэллапа, например, проводил многочисленные опросы до и после первого избрания Обамы. Их вывод: американских избирателей не особо волновало, что Барак Обама черный[1]. Вскоре после выборов двое известных профессоров из университета Беркли[2] в Калифорнии внимательно изучили собранные в ходе исследований материалы, применяя сложнейшие методики обработки данных. В результате они пришли к аналогичному выводу.

Таким образом, во время президентства Обамы это стало общепринятым мнением, которое распространилось во многих СМИ и академических кругах. Источники, на которые восемьдесят с лишним лет опирались СМИ и ученые-социологи для понимания устройства нашего мира, утверждают, что подавляющее большинство американцев не волновало, что Обама — чернокожий, когда они решали, может ли он стать их президентом.

Эта страна, издавна запятнанная рабством и законами Джима Кроу*, казалось, наконец перестала судить о людях по цвету их кожи. Это вроде бы должно было указывать на то, что расизм в Америке на последнем издыхании.

---

* Неофициальное название законов о расовой сегрегации в США в период с 1890 по 1964 год. — *Прим. ред.*

Некоторые эксперты даже заявили, что мы живем в пострасовом обществе[3].

В 2012 году я был аспирантом в области экономики и разочаровался в выбранном мной направлении, будучи уверенным в том, что я уже довольно хорошо понимаю, как устроен мир, о чем люди думают и что их заботит в двадцать первом веке. А когда дело дошло до вопроса о предрассудках, я позволил себе поверить, исходя из того, что я читал в трудах по психологии и политологии, что явный расизм присущ весьма ограниченному проценту американцев и большинство из них — консервативные республиканцы, в основном живущие в глубинке на Юге.

Затем я обнаружил Google Trends.

Появление этого приложения в 2009 году прошло практически незамеченным. Оно позволяет пользователям определить, насколько часто то или иное слово или фраза появлялись в разных местах и в разное время, и преподносилось оно как инструмент для развлечения, например для обсуждения с друзьями, какие знаменитости сейчас популярны или какая одежда вошла в моду. Ранние версии программы даже включали шутливое предостережение о том, что «не стоит писать докторскую диссертацию», опираясь на такие данные, что сразу же побудило меня написать диссертацию на их основе*.

---

* Приложение Google Trends — источник большей части данных, содержащихся в моей работе. Однако, поскольку оно позволяет лишь сравнивать относительную частоту разных запросов, но не сообщает точное их число по какому-либо конкретному виду поиска, я обычно дополнял его результаты данными,

В то время данные поисковика Google, похоже, не считались достойным источником информации для серьезных научных исследований, ведь они не создавались как инструмент для изучения человеческой психологии. Google придумали для того, чтобы люди могли познавать мир, а не для того, чтобы исследователи изучали людей. Но оказалось, что следы, которые мы оставляем, выискивая крупицы знаний в интернете, чрезвычайно показательны.

Другими словами, люди, ищущие информацию, сами являются источником информации. То, когда и где они ищут факты, цитаты, шутки, места, людей, вещи или помощь, оказывается, может рассказать нам гораздо больше об их реальных мыслях, желаниях, опасениях и делах, чем можно себе представить. И особенно наглядно это проявляется тогда, когда люди не столько задают поисковику вопросы, сколько доверяются ему: «я ненавижу своего босса», «я пьян», «мой папа ударил меня».

Печатание слова или фразы в аккуратном белом окошке оставляет маленький реальный след. Помноженный на миллионы, в итоге он выявляет глубинные реалии.

полученными из Google Adwords — сервиса, который показывает, как часто осуществлялся каждый поиск. В большинстве случаев мне также удалось улучшить четкость изображения с помощью моего собственного алгоритма, написанного на базе Google Trends, который я описал в своей диссертации «Опыт использования данных Google», и в моей статье для *Journal of Public Economics* — «Уровень расовой неприязни к чернокожему кандидату: на основе данных, полученных с помощью Google». Диссертация, статья, полное объяснение данных и код, использовавшийся во всех оригинальных исследованиях, представленных в этой книге, доступны на моем сайте: sethsd.com. — *Прим. авт.*

Первое слово, которое я набрал в Google Trends, было «Бог». Я узнал, что штатами, в которых чаще всего в поисковых запросах в Google упоминается Бог, были Алабама, Миссисипи и Арканзас — так называемый Библейский пояс. И эти поиски чаще всего происходят по воскресеньям. В этом нет ничего удивительного, но любопытно, что поиск данных позволяет выявить настолько ясную картину. Я набрал Knicks* и увидел, что большинство запросов относится к городу Нью-Йорк. Ежу понятно. Тогда я набрал свое имя. «Мы сожалеем, — ответил мне Google Trends. — Не хватает поискового объема, чтобы показать результаты». Так я узнал, что Google Trends предоставляет данные только тогда, когда достаточно много людей выполняет один и тот же поиск.

Но сила поисковой системы Google не в том, чтобы выяснить, что наибольшей популярностью Бог пользуется на Юге, Knicks — в Нью-Йорке или что я не популярен нигде. Любой опрос может выявить это. Могущество и власть Google заключается в том, что люди рассказывают гигантской поисковой системе то, что они не могли бы сказать никому другому.

Возьмем, к примеру, секс (к этой теме я вернусь позднее и рассмотрю ее более подробно). Результатам опросов нельзя доверять, поскольку люди редко говорят правду о своей сексуальной жизни. Я проанализировал данные Всеобщего социального исследования[4], которое считается наиболее достоверным и авторитетным источником информации о поведении американцев.

---

* Сокр. от Knickerbockers — нью-йоркская баскетбольная команда (НБА). — *Прим. ред.*

По данным этого опроса, когда речь идет о гетеросексуальном контакте, женщины говорят, что они занимаются сексом в среднем пятьдесят пять раз в год, в шестнадцати процентах случаев используя презерватив. Это дает около 1,1 миллиарда презервативов в год. Но, по утверждению гетеросексуальных мужчин, ежегодно используется 1,6 миллиарда презервативов. По определению эти цифры должны совпадать. Так кто же говорит правду — мужчины или женщины?

Как оказалось — ни те, ни другие. По данным компании Nielsen, которая отслеживает поведение потребителей, ежегодно продается менее 600 миллионов презервативов[5]. Так что лгут и те и другие; единственное различие в том, насколько сильно.

Ложь на самом деле очень широко распространена. Мужчины, которые никогда не были в браке, заявляют об использовании в среднем двадцати девяти презервативов в год. Это число следует добавить к числу презервативов, продаваемых в Соединенных Штатах людям, состоящим в браке и одиноким, вместе взятым. Люди, состоящие в браке, наверное, тоже преувеличивают свою сексуальную активность. В среднем женатые мужчины в возрасте под шестьдесят пять говорят, что они занимаются сексом раз в неделю. Только один процент признается, что у них не было секса целый год. Замужние женщины сообщают о немного меньшем количестве секса, но совсем немного.

По результатам поиска в Google мы обнаружим менее яркую, но, как мне кажется, гораздо более правдоподобную картину. Больше всего жалоб на отсутствие секса в браке. Поисковый запрос «брак без секса» делается

в три с половиной раза чаще, чем запрос «несчастливый брак», и в восемь раз чаще, чем «брак без любви». Даже неженатые пары довольно часто жалуются на то, что они не занимаются сексом. Поисковый запрос «отношения без секса» уступает только запросам тех, кто ищет «жесткий секс». (Хочу подчеркнуть, что все эти данные предоставлены анонимно. Google, разумеется, не сообщает данные поиска конкретной личности.)

Поисковик Google позволил нам увидеть картину Америки, которая разительно отличается от той пострасовой утопии, которую показали результаты опросов. Помню, как я впервые набрал слово «ниггер» в Google Trends. Можете считать меня наивным, но, учитывая, насколько «токсично» это слово, я ожидал, что поисковый объем будет очень небольшим. Ребята, я был неправ. В Соединенных Штатах слово «ниггер» — или во множественном числе «ниггеры» — входило в поисковые запросы примерно столько же раз, сколько слова «мигрень», «экономист(ы)» и «Лейкерс». Я подумал, что, если связать это слово со словом «рэп», возможно, результат будет другим. Но нет. Слово, используемое в рэпе, почти всегда — «нигга». Какая же мотивация была у американцев, осуществлявших поиск со словом «ниггер»? Зачастую они ищут анекдоты, высмеивающие афроамериканцев. Но на самом деле только двадцать процентов поисковых запросов со словом «ниггер» включают и слово «анекдот», тогда как большинство подобных поисков включают фразы «тупые ниггеры» и «я ненавижу ниггеров».

И ежегодно — миллионы таких поисков. Множество американцев в уединении, находясь дома, делают шокирующе расистские запросы. Чем больше я занимался

этим исследованием, тем больше получал тревожной информации.

В первую ночь после выборов Обамы, когда большинство комментариев были хвалебными и признающими историческое значение его избрания, примерно один из каждых ста поисковых запросов Google, содержащих слово «Обама», также включал слова «ККК»* или «ниггер(ы)». Возможно, это не так уж много, учитывая тысячи нерасистских запросов в Google об этом молодом незнакомце с очаровательным семейством, который собирался взять на себя выполнение самой значимой в мире работы. В ночь выборов поисковых запросов и регистраций на Stormfront[6] — сайте белых националистов с неожиданно высокой популярностью в США — было более чем в десять раз больше, чем обычно. В некоторых штатах поисков по запросам «ниггер-президент»[7] было намного больше, чем по запросам «первый черный президент».

Темная сторона и неприязнь, которые не были выявлены традиционными методами, стали вполне очевидны после анализа поисковых запросов, которые делали люди.

Все эти запросы плохо согласуются с обществом, в котором расизм — незначительный фактор. В 2012 году я знал Дональда Дж. Трампа в основном как бизнесмена и ведущего реалити-шоу. Я, как и большинство людей, представить не мог, что спустя четыре года он станет серьезным кандидатом в президенты. Тем не менее, все эти неприглядные поисковые запросы нетрудно связать с успехом кандидата, который, используя злобные нападки

---

* Ku Klux Klan (*англ.*) — Ку-клукс-клан. — *Прим. ред.*

на иммигрантов, разжигая неприязнь и нетерпимость, часто играл на худших человеческих проявлениях.

Анализ поиска в Google также показал, что мы во многом имели неверное представление о локализации расистских настроений в стране. По опросам и традиционным представлениям, современный расизм базируется преимущественно на Юге и в основном среди республиканцев. Однако места с наивысшим уровнем расистских запросов были обнаружены — помимо Западной Виргинии, Южной Луизианы и Миссисипи — также в штатах Нью-Йорк, Пенсильвания, Западный и Восточный Огайо, Мичиган, промышленный и сельский Иллинойс. По данным Google, правильнее было бы противопоставить не Юг и Север, а Восток и Запад. Вы не получите подобного уровня запросов сильно к западу от Миссисипи. И распространение расизма не ограничивается средой республиканцев. Фактически расистские запросы в местах с высоким процентом республиканцев были не выше, чем в местах с высоким процентом демократов. Иными словами, анализ поиска в Google помог составить новую карту локализации расизма в США, и эта карта выглядела совершенно иначе, чем мы себе представляли. Дело в том, что республиканцы на Юге с большей вероятностью признаются в своем расизме, хотя и множество демократов на Севере имеют аналогичные взгляды.

Четыре года спустя эта карта окажется довольно значимой при объяснении политического успеха Трампа.

В 2012 году я использовал эту карту локализации расизма, разработанную по частоте запросов в Google, чтобы полностью пересмотреть ту роль, которую сыграла

расовая принадлежность Обамы, и увидел четкую картину. В районах страны с наибольшим количеством расистских поисковых запросов рейтинг Обамы был существенно ниже рейтинга Джона Керри, белого кандидата в президенты от Демократической партии. Такой результат в этих районах невозможно было объяснить никаким иным фактором, в том числе уровнем образования, возрастом, религиозностью или владением оружием. Расистские запросы не позволяли прогнозировать низкий уровень популярности ни для какого другого демократического кандидата. Только для Обамы.

В результате Обама потерял примерно четыре процента голосов по стране вследствие откровенного расизма. Это было намного больше, чем ожидалось, исходя из данных опросов. Барак Обама, конечно, был избран и переизбран президентом, в чем не последнюю роль сыграли очень благоприятные условия для демократов, но ему пришлось преодолеть намного больше трудностей, чем кому-либо, кто полагался на традиционные источники данных, которые в большинстве случаев были ошибочны. В стране было достаточно расистов, способных одержать победу на предварительных или всеобщих выборах не в столь благоприятный для демократов период.

Поначалу мое исследование было отклонено пятью научными журналами[8]. Многие из рецензентов — извините за брюзжание — заявили, что не могут поверить, будто так много американцев скрывают свой расизм. Это противоречило тому, что люди говорили при опросах. Кроме того, исследование поисковых запросов в Google казалось им очень странным способом получения данных.

Теперь, когда мы стали свидетелями инаугурации президента Дональда Дж. Трампа, мои результаты кажутся вполне убедительными.

Чем больше я изучал этот вопрос, тем больше понимал, что в Google есть много информации, которую не принимали во внимание при опросах и которая, помимо всего прочего, может быть полезна для понимания результатов выборов.

Например, информация о том, кто на самом деле будет принимать участие в выборах. Больше половины граждан, которые не голосуют, говорят исследователям, проводящим опросы непосредственно перед выборами, что они намерены пойти голосовать, что искажает оценку явки, в то время как данные о поиске в Google по фразам «как голосовать» или «где голосовать» за неделю перед выборами помогут более точно предсказать, где предполагается большая активность на избирательных участках.

Можно даже найти информацию о том, за кого они пойдут голосовать. Мы со Стюартом Гэбриэлом, профессором университета штата Калифорния, Лос-Анджелес, нашли удивительную подсказку для определения того, как именно люди планируют голосовать. Большой процент поисков, связанных с выборами, содержит запросы с именами обоих кандидатов. Во время выборов 2016 года, когда соперничали Трамп и Хиллари Клинтон, некоторые люди делали запрос: «выборы: Трамп — Клинтон». Другие искали: «Клинтон — Трамп, дебаты». По сути, двенадцать процентов поисковых запросов со словом «Трамп» включали и слово «Клинтон». Более

четверти поисковых запросов с фамилией Клинтон также содержали и фамилию Трампа.

Мы обнаружили, что эти, казалось бы, нейтральные поиски на самом деле могут дать нам некоторые подсказки о том, какого кандидата человек поддерживает.

Как? Все зависит от порядка, в котором кандидаты появляются в запросе. Наши исследования показывают, что человек со значительно большей вероятностью поставит имя кандидата, которого он поддерживает, первым в поисковом запросе, содержащем имена обоих кандидатов.

В ходе предыдущих трех выборов кандидат, фамилию которого ставили первым в поисковых запросах, набирал наибольшее число голосов. Что еще интереснее, порядок, в котором искали в сети кандидатов, позволял предсказать, чью сторону примет тот или иной штат.

Порядок, в котором имена кандидатов появляются в поисковых запросах, также содержит информацию, которую упускают при опросах. В 2012 году во время выбора между Обамой и республиканцем Миттом Ромни Нейт Сильвер, виртуозный статистик и журналист, точно предсказал результат во всех пятидесяти штатах. Однако мы обнаружили, что в тех штатах, которые чаще ставили Ромни перед Обамой в поисковых запросах, дела Ромни на самом деле были лучше, чем предсказал Сильвер. В штатах, которые чаще ставили Обаму перед Ромни, дела Обамы все-таки были лучше, чем предсказал Сильвер.

Этот показатель может содержать информацию, которая не выявляется при опросах, потому что избиратели либо обманывают сами себя, либо им неудобно раскрывать перед социологами свои истинные предпочтения.

Вероятно, если бы они в 2012 году говорили, что еще не определились, но при этом постоянно делали запросы: «выборы: Ромни — Обама», «дебаты Ромни — Обама» и «Ромни — Обама, выборы», это значило бы, что они планируют все же голосовать за Ромни.

Так что же, Google предсказал победу Трампа? Ну, нам еще предстоит проделать большую работу — и мне придется объединить свои усилия с большим числом других исследователей, — прежде чем мы поймем, как лучше всего использовать данные Google, чтобы предсказать результаты выборов. Это новая наука, и пока мы располагаем данными лишь по нескольким прошедшим выборам. Разумеется, я не говорю, что наступил момент — если он вообще когда-нибудь наступит, — когда можно полностью отказаться от опросов общественного мнения как инструмента, который помогает прогнозировать выборы.

Но могу сказать определенно, что в интернете можно было найти много свидетельств того, что у Трампа было больше шансов стать президентом, чем получалось на основании данных, собранных во время опросов.

Во время всеобщих выборов можно было заметить подсказки, свидетельствующие в пользу того, что электорат на стороне Трампа. Черные американцы говорили интервьюерам, что они в массе своей будут голосовать против Трампа. Но поисковые запросы в Google с выяснением информации о голосовании на участках с преобладанием афроамериканцев, показали, что их активность снижается. В день выборов Клинтон будет неприятно удивлена низкой явкой чернокожего населения.

Были даже признаки того, что неопределившиеся избиратели перешли на сторону Трампа. Мы с Гэбриэлом

обнаружили, что в ключевых штатах на Среднем Западе, в которых Клинтон надеялась одержать победу, намного больше поисковых запросов выстраивались как «Трамп — Клинтон», чем как «Клинтон — Трамп». Действительно, Трамп во многом обязан своим избранием тому, что он значительно превзошел там результаты своих показателей по опросам.

Но ключевой подсказкой — и я в этом абсолютно убежден, — которая помогла обнаружить основные признаки того, что Трампа может ждать успех — для начала на предварительных выборах — был все тот же скрытый расизм, который выявило мое исследование во время избрания Обамы. Анализ поисковых запросов в Google выявил озлобление и нетерпимость у значительного числа американцев, которые эксперты не замечали в течение многих лет. Эти данные показали, что мы жили в обществе, совершенно отличном от того, которое представляли нам ученые и журналисты, опираясь на опросы. Они выявили отвратительную, пугающую и повальную злость по отношению к кандидату, ожидающему, что избиратели отдадут за него свои голоса.

Люди часто лгут — и себе, и другим. В 2008 году американцы сообщили в ходе опросов, что их больше не волнует расовая принадлежность человека. Восемь лет спустя они избрали в качестве президента Дональда Дж. Трампа — человека, который ретвитнул ложное утверждение, что черные несут ответственность за большую часть убийств белых американцев, защищал своих сторонников, обвиненных в избиении чернокожих протестующих из Black Lives Matter (BLM) — интернационального движения активистов, выступающих против насилия

в отношении чернокожего населения, — на одном из митингов, и колебался, следует ли отвергать поддержку бывшего лидера Ку-клукс-клана. Тот же скрытый расизм, который повредил Бараку Обаме, помог Дональду Трампу.

В начале предварительных выборов Нейт Сильвер уверенно заявил, что у Трампа практически нет никаких шансов на победу. Но в ходе выборов становилось все яснее, что Трамп пользуется широкой поддержкой. Сильвер решил взглянуть на данные, чтобы понять, что же происходит? Каким образом Трампу удалось так успешно продвинуться вперед?

Сильвер заметил, что районы, где Трамп выступал успешнее всего, представляют собой странную карту. Трамп хорошо зарекомендовал себя в районах Северо-Востока и промышленного Среднего Запада, а также на Юге. На Западе он был принят заметно хуже. Сильвер начал искать параметры, объясняющие эту картину. Причина в безработице? Это религия? Это владение оружием? Уровень иммиграции? Оппозиция Обаме?

В итоге Сильвер пришел к выводу, что единственным фактором, который лучше всего коррелирует с поддержкой Дональда Трампа[9] на республиканских первичных выборах, было то, что я обнаружил четыре года назад. Трампа поддержали те области, жители которых сделали большинство поисковых запросов в Google со словом «ниггер».

Почти каждый день в течение последних четырех лет я занимался анализом данных Google. Это включало работу в качестве аналитика данных компании Google, которая наняла меня, узнав о моих исследованиях расизма.

И я продолжал работать с этими данными как автор редакционных статей и журналист газеты «Нью-Йорк таймс». Новые откровения не заставили себя ждать. Психические растройства, сексология, насилие над детьми, аборты, реклама, религия, здоровье — довольно серьезные темы. И этот набор данных, которого не существовало еще пару десятилетий назад, позволяет взглянуть на них совершенно по-другому. Экономисты и социологи постоянно охотятся за новыми источниками данных, так что позвольте мне быть откровенным: сегодня, я убежден, поиск в Google предоставляет самый важный набор данных о человеческой психологии, который когда-либо был собран.

Однако этот набор данных — не единственный инструмент для понимания нашего мира, предоставляемый интернетом. Вскоре я понял, что есть и другие золотоносные цифровые жилы. Я скачал всю Википедию, покопался в профилях Facebook и прошерстил Stormfront. Кроме того, PornHub, один из крупнейших порнографических сайтов интернета, дал мне свои полные данные по анонимному поиску и просмотрам видео, которые совершали люди со всего мира. Другими словами, я глубоко погрузился в то, что сейчас называют большими данными*. Затем я опросил десятки других специалистов — ученых, журналистов и предпринимателей, которые также проводят изыскания в этой новой сфере. Многие из их исследований будут обсуждаться в этой книге.

Но сначала я должен признаться: я не собираюсь давать точное определение того, что такое «большие

* В ориг. — Big Data — *Прим. ред.*

данные». Почему? Потому что это, по сути, довольно расплывчатое понятие. Большие — это сколько? 18 462 наблюдений — это малые данные, а 18 463 — уже большие? Я предпочитаю инклюзивное понимание того, что относится к этому классу: большая часть данных, с которыми я работал, была получена из интернета, но при обсуждении я буду принимать во внимание и другие источники. Мы переживаем взрывной рост количества и качества различных видов доступной информации. Новые потоки информации влились через Google и социальные сети. Некоторые из них — продукт оцифровки информации, которая раньше была спрятана в шкафах и папках, другие получены в результате увеличения ресурсов, выделяемых на маркетинговые исследования. Часть исследований, рассмотренных в этой книге, вообще не нуждаются в огромных массивах данных, вместо этого в них просто применяется новый творческий подход к данным, что особенно ценно в наш век переизбытка информации.

Так почему же именно большие данные обладают такой огромной мощью? Представьте себе все данные, которые разлетаются по интернету всего за день — по правде говоря, мы подсчитали объем такой информации. В начале двадцать первого века за день люди генерируют в среднем 2,5 миллиона триллионов байт данных[10].

И эти байты и есть ключ к разгадке.

Женщина скучает вечером в четверг. Она немного погуглила «приличные смешные видео». Она проверила свою электронную почту. Она отметилась в Twitter. Затем она гуглит «анекдоты про ниггеров».

Мужчине грустно. Он погуглил «симптомы депрессии» и «рассказы о депрессии». Затем разложил пасьянс.

Женщина видит в Facebook объявление о том, что ее подруга выходит замуж. Женщина не замужем, одинока, и она блокирует информацию о подруге.

Мужчина в перерыве между поисками информации о НХЛ и рэпе задает в поисковике вопрос: «Мечтать о поцелуях мужчины — это нормально?»

Женщина кликает на сюжет BuzzFeed про «15 милых кошек».

Мужчина видит ту же историю о кошках. Но на его экране она называется «15 самых очаровательных кошек». Он не кликает на ссылку.

Женщина гуглит: «Мой сын гений?»

Мужчина гуглит: «Как заставить мою дочь похудеть?»

Женщина в отпуске с шестью лучшими подругами. Все ее подруги постоянно говорят, как им весело. Она набирает в Google: «Одиночество вдали от мужа».

Мужчина, муж предыдущей женщины, в отпуске с шестью своими лучшими друзьями. Он набирает в Google: «Признаки того, что ваша жена изменяет».

Некоторые из этих данных содержат информацию, о которой в иной ситуации никто никогда не узнал бы. Если мы объединим все это, сохраняя анонимность, строго следя за тем, чтобы никто никогда не узнал о страхах, желаниях и поведении конкретных лиц, и добавим некоторые научные данные, мы начнем по-новому смотреть на людей — их поведение, их желания, их характеры.

Рискуя показаться пафосным, скажу: фактически я пришел к выводу, что новые данные, ставшие более

доступными в нашу цифровую эпоху, способны радикально расширить наше понимание человеческой природы. Микроскоп позволил нам увидеть в капле воды из пруда гораздо больше, чем мы думали. Телескоп показал нам в ночном небе намного больше того, что мы видели невооруженным глазом. И теперь новые цифровые данные открывают нам в человеческом сообществе многое из того, что было скрыто. Они могут стать нашими современными микроскопом или телескопом, и полученная ими информация, возможно, приведет к важнейшим, даже революционным открытиям.

В подобных высказываниях есть еще один рискованный момент: они могут восприниматься не только как пафосные, но и трендовые. Многие делали серьезные заявления о могуществе больших данных, не приводя никаких доказательств. Это побудило людей, скептически относящихся к большим данным, которых тоже немало, отвергнуть идею исследования больших массивов данных. «Я не говорю, что нет никакой информации в больших данных, — пишет публицист и статистик Нассим Талеб, — там масса информации. Проблема — основная — заключается в том, что иголку приходится искать в непрерывно растущих стогах сена».

Одна из основных целей этой книги — представить недостающие доказательства и показать, что можно сделать с большими данными, то есть как можно при желании находить иголки в непрерывно растущих стогах сена. Я надеюсь предоставить достаточно примеров того, как большие данные дают возможность по-новому взглянуть на человеческую психологию и поведение, чтобы вы могли увидеть очертания чего-то действительно революционного.

«Постой, Сэт, — могли бы вы сказать сейчас. — Ты обещаешь революцию. Ты так красноречиво разглагольствуешь об этих больших новых наборах данных. Но до сих пор ты использовал весь этот поразительный, впечатляющий, умопомрачительный, новаторский набор данных только для того, чтобы показать мне в основном два момента: в Америке много расистов и люди, особенно мужчины, сильно преувеличивают, говоря о том, как часто они занимаются сексом».

Я допускаю, что иногда новые данные просто подтверждают очевидное. Если вы считаете, что эти выводы были очевидны, подождите, пока не доберетесь до четвертой главы, где я предоставлю вам отчетливые и неопровержимые доказательства, полученные на базе поиска в Google, подтверждающие, что у мужчин существует серьезная озабоченность и неуверенность по поводу — чего бы вы думали? — размера своего пениса.

Это, я бы сказал, имеет определенную ценность в качестве доказательства того, о чем вы, возможно, уже подозревали, но не имели достаточно данных для подтверждения своих подозрений. Подозревать — это одно, доказать — совсем другое. Но если все, на что способны большие данные — подтверждение ваших подозрений, это не будет чем-то революционным. К счастью, большие данные могут гораздо больше. Снова и снова они показывают мне, что все происходит совсем не так, как я предполагал. Вот некоторые примеры, которые вы могли бы счесть достаточно впечатляющими и неожиданными.

Можно предположить, что основной причиной расизма является экономическая незащищенность и уязвимость. Вы, естественно, подозреваете, что, когда люди

теряют работу, их расизм усиливается. Но на самом деле при увеличении безработицы не увеличивается ни количество расистских поисковых запросов, ни число членов Stormfront.

Принято думать, что состояние тревожности в основном присуще жителям больших городов, где много высокообразованных людей. Городской невротик — это известный стереотип. Но количество запросов в Google, отражающих тревожность, таких как «симптомы тревожности» и «помощь при состоянии тревожности», как правило, выше в местах с низким уровнем образования, там, где меньше средний доход и где большая часть населения живет в сельской местности. То есть более высокий уровень числа поисковых запросов, связанных с тревожностью, в сельской местности, на севере штата Нью-Йорк, а не в самом Нью-Йорке.

Вы считаете, что теракт, в результате которого погибли десятки или сотни людей, автоматически приведет к широкому распространению массовой тревожности. Терроризм по определению должен внушать чувство страха. Я просмотрел поисковые запросы в Google, отражающие беспокойство, и отследил рост числа этих поисков по стране в последующие дни, недели и месяцы после каждой крупной террористической атаки в Европе или Америке, начиная с 2004 года. Итак, на сколько же в среднем выросло число поисковых запросов, связанных с тревожностью? Ни на сколько. Совсем.

Вы думаете, что люди чаще ищут анекдоты, когда им грустно. Многие из величайших мыслителей утверждали, что мы обращаемся к юмору как к обезболивающему. Юмор уже давно воспринимается как способ справиться

с огорчениями, болью, неизбежными разочарованиями в жизни. Как выразился Чарли Чаплин: «Смех — это тоник, способ расслабиться, забыть о страданиях».

Тем не менее в понедельник — день с репутацией самого несчастливого — уровень поиска шуток самый низкий. То же можно сказать про пасмурные и дождливые дни. И этот уровень резко падает после крупной трагедии, например, когда в результате взрыва двух бомб погибло трое и были ранены сотни людей во время Бостонского марафона 2013 года. На самом деле люди предпочитают шутки, когда дела идут хорошо, а не наоборот.

Иногда новый массив данных выявляет такие поступки, стремления или отношения, которые я бы даже никогда и предположить не мог. В эту категорию попадают многочисленные сексуальные предпочтения. Например, известно ли вам, что в Индии большинство поисковых запросов начинается со слов «мой муж хочет…». Например: «Мой муж хочет, чтобы я кормила его грудью»[11]. Этот запрос распространен в Индии гораздо больше, чем в других странах. Кроме того, уровень поиска по порносайтам изображений, где женщина кормит мужчину грудью, в Индии и Бангладеш в четыре раза выше, чем в любой другой стране. Я, конечно, никогда и не подозревал ни о чем подобном до того, как увидел эти данные.

Тот факт, что мужчины одержимы размером своего пениса, может, и не слишком неожиданный, но вот то, что вызывает наибольшую обеспокоенность у женщин, касаемо их тела, по данным Google, действительно вызывает удивление. Опираясь на эти новые данные, женским эквивалентом комплекса по поводу размера полового члена можно считать — выразительная пауза! — переживание

о том, как пахнет их вагина. Женщины выполняют почти столько же поисков, выражая озабоченность по поводу своих гениталий, как и мужчины, беспокоящиеся о размере своих. Главной заботой женщины является ее запах и то, как она может его улучшить. Разумеется, я не знал этого, пока не обнаружил такие данные.

Иногда новые данные показывают культурные различия, о которых я даже не задумывался. Вот один пример: очень по-разному люди по всему миру реагируют на беременность своих жен. В Мексике топ-запросы «моя беременная жена» включают фразы «frases de amor para mi esposa embarazada» (признание в любви моей беременной жене) и «poemas para mi esposa embarazada»[12] (стихи для моей беременной жены). В Соединенных Штатах топ поисковых запросов состоит из следующих фраз: «моя жена беременна — и что теперь?» и «моя жена беременна — что мне делать?».

Но эта книга больше, чем подборка странных фактов или единичных исследований, хотя в ней будет приведено много подобной информации. Поскольку эта методика совершенно новая и только набирает обороты, я изложу здесь некоторые идеи о том, как это работает и что делает ее столь революционной. Я также допускаю, что есть пределы больших данных.

Эйфория в связи с потенциальной информационной революцией вряд ли уместна. Большинство тех, кто без ума от больших данных, просто фонтанирует идеями применения этого колоссального массива информации. Такая одержимость не нова. До Google, Amazon и Facebook, до появления самого понятия «большие данные» состоялась конференция в Далласе — «Большие

и сложные массивы данных». Джерри Фридман[13], профессор статистики Стэнфордского университета и мой коллега по работе в Google, вспоминает, что на конференции 1977 года один уважаемый статистик заявил о том, что накопил невероятные, ошеломляющие пять гигабайт данных. Затем встал следующий выдающийся статистик и начал со слов: «Последний оратор говорил о гигабайтах. Это ничто. У меня — терабайты». Другими словами, акцент выступлений сместился на то, как много информации можно накопить, вместо того чтобы сделать упор на то, что с этими накопленными данными можно сделать или на какие вопросы можно найти ответы. «Тогда мне показалось забавным, — сказал Фридман, — что все надеялись поразить слушателей тем, насколько большой набор данных им удалось собрать. И это продолжается до сих пор».

Сегодня слишком много специалистов по анализу и обработке данных накопили большие массивы информации, но они дают нам слишком несущественные сведения, например, что баскетбольный клуб Knicks пользуется популярностью в Нью-Йорке. Слишком многие компании просто утонули в больших объемах данных. У них много терабайт информации, но мало своих идей. На мой взгляд, значение количества данных часто переоценивается. И это легко заметить, учитывая один небольшой, но очень существенный момент: чем важнее явление, тем меньше число наблюдений необходимо, чтобы его обнаружить. Вам достаточно один раз прикоснуться к горячей плите, чтобы понять, насколько это опасно. Но, возможно, вам придется тысячи раз пить кофе, чтобы понять, вызывает ли он у вас головную боль. Какой фактор серьезнее?

Очевидно, что горячая плита, которая в силу интенсивности своего воздействия позволяет получить мгновенный результат при минимальном объеме данных.

Поэтому самые сообразительные крупные компании, занимающиеся обработкой больших данных, зачастую обрезают имеющиеся в их распоряжении массивы. В компании Google основные решения принимаются на основе лишь малой толики имеющихся в их распоряжении данных. Вам не всегда нужны тонны информации для того, чтобы прийти к важным выводам. Нужны правильно подобранные данные. Главный аргумент в пользу того, что поисковые запросы в Google представляют собой ценнейшую информацию, состоит не в том, что их очень много, а в том, что люди в них весьма откровенны. Мы лжем друзьям, любовникам и любовницам, врачам, опросам и самим себе. Но Google дает возможность обсудить личные проблемы, в том числе с весьма компрометирующей информацией, такие как брак без секса, психическое нездоровье, неуверенность, враждебность по отношению к чернокожим.

Самое главное при работе с большими данными — умение задавать правильные вопросы, чтобы получить важные выводы. Как нельзя, случайно наведя телескоп на ночное небо, обнаружить там Плутон, нельзя, просто загрузив кучу данных, открыть тайны человеческой природы. Вам необходимо будет выделить наиболее перспективные для поиска фразы, например для Индии — это запросы в Google, которые начинаются со слов «мой муж хочет…».

Эта книга показывает, как лучше использовать большие данные, в ней подробно объясняется, почему эти

массивы информации имеют такое большое значение. И попутно вы узнаете много интересного из того, что я и другие люди уже открыли для себя с помощью этого метода, в том числе:

- Как много геев среди мужчин?
- Неужели реклама действительно работает?
- Почему Американский Фараон — лучшая скаковая лошадь?
- Ангажированы ли СМИ?
- Существуют ли оговорки по Фрейду?
- Кто мошенничает с налогами?
- Важно ли, в какой колледж пойти учиться?
- Можно ли выиграть на фондовом рынке?
- Где лучшее место, чтобы растить детей?
- Как истории разносятся по сети?
- О чем следует говорить на первом свидании, если вы хотите, чтобы было второе?

…И многое, многое другое.

Но прежде чем мы доберемся до этого, нужно обсудить базовый вопрос: зачем нам вообще все эти данные? И для этого я хочу представить вам мою бабушку.

ЧАСТЬ I

ДАННЫЕ, БОЛЬШИЕ И МАЛЫЕ

Папе и маме

Глава 1

ИНТУИЦИЯ ВАС ОБМАНЫВАЕТ

Если вам 33 года от роду и у вас уже несколько Дней благодарения подряд прошли без свиданий, скорее всего, возникнет тема выбора брачного партнера. И у каждого на этот счет свое мнение.

«Сету нужна сумасшедшая девчонка под стать ему», — говорит моя сестра.

«Ты с ума сошла! Ему нужна нормальная девушка, чтобы уравновешивать его», — заявляет брат.

«Сет не сумасшедший», — реагирует мать.

«Ты спятила! Конечно, Сет — настоящий псих», — заявляет отец.

Внезапно в разговор тихо вступает моя застенчивая, говорящая тихим голосом бабушка. Громкие агрессивные нью-йоркские голоса затихают, и все взгляды сосредотачиваются на небольшой старушке с короткими золотистыми волосами, говорящей с легким восточноевропейским акцентом.

«Сет, тебе нужна хорошая девушка. Не слишком красивая. Очень умная. Умеющая ладить с людьми, социальная, чтобы вы могли работать вместе. С чувством юмора, потому что у тебя хорошее чувство юмора».

Почему совет этой пожилой женщины выслушивается в моей семье с таким вниманием и уважением? Моя 88-летняя бабушка видела на своем веку больше, чем все остальные, сидевшие за столом. Она повидала множество браков, одни из которых были счастливыми, другие нет. И на протяжении десятилетий она составляла список качеств, делающих взаимоотношения успешными. За столом в День благодарения бабушка была источником самого большого числа данных. Моя бабушка сама была большими данными.

В этой книге я хочу развеять мифы о науке о данных.

Нравится нам это или нет, но информация играет все более важную роль в жизни каждого из нас — и эта роль будет становиться все значительнее. Сейчас в газетах встречаются целые разделы, полностью посвященные данным. В компаниях есть группы, единственной задачей которых является анализ собранных данных. Инвесторы дают десятки миллионов долларов стартапам, если те могут собрать и сохранить большие объемы данных. Даже если вы никогда не узнаете, как работает регрессия, и не можете рассчитать доверительный интервал, вы наверняка столкнетесь с большим количеством данных — на страницах книг, которые читаете, во время деловых встреч, в которых принимаете участие, в сплетнях, которые доходят до ваших ушей, в курилке или возле кулера, когда пьете воду.

Многих людей беспокоит такое развитие событий. Они запуганы данными, легко теряются и могут совсем

запутаться в мире чисел. Они думают, что количественное понимание мира предназначено для избранных левополушарных вундеркиндов, а не для них. Поэтому, едва столкнувшись с цифрами, готовы перевернуть страницу, закончить встречу или сменить тему разговора.

Я потратил десять лет на анализ различных данных, и за это время мне посчастливилось работать со многими из наиболее значимых в этой области людей. Один из самых важных уроков, которые я усвоил, заключается в том, что правильная работа с информацией не настолько сложна, как кажется многим. Лучшие примеры научной работы с данными на самом деле показывают, насколько она интуитивна[1].

Что же делает науку о данных столь интуитивной? По своей сути эта дисциплина занимается выявлением и отбором правильных данных, а также прогнозированием того, как одна переменная повлияет на другую. Люди постоянно этим занимаются.

Просто подумайте, как бабушка давала мне совет по поводу моих отношений. Она использовала большую базу данных об отношениях, загружавшуюся в ее мозг в течение практически всей жизни, — истории, которые она слышала от членов своей семьи, от друзей и знакомых. Сначала она ограничила данные для анализа примерами отношений, в которых мужчина имел многие из тех качеств, которые есть и у меня — чувствительность, склонность к самоизоляции, чувство юмора. Затем сосредоточилась на ключевых качествах известных ей в этой выборке женщин: насколько они были добрыми, умными, красивыми. Потом сопоставила эти ключевые качества женщин с важнейшим элементом

отношений: были ли они хорошими или нет. И, наконец, сообщила результат. Другими словами, она заметила закономерности и предсказала, как одна переменная будет влиять на другую. В этой ситуации бабушка выступила как специалист по работе с данными.

Вы тоже являетесь специалистом по работе с данными. Будучи ребенком, вы замечали: стоило начать плакать, как мама сразу обращала на вас внимание. Это тоже часть науки по работе с данными. Достигнув совершеннолетия, вы заметили, что, если слишком много ныть и жаловаться, люди начнут избегать общения с вами. Это тоже наука о данных. Когда люди меньше общаются с вами, у вас портится настроение, вы недовольны. Когда вы менее счастливы, вы менее дружелюбны, а когда вы менее дружелюбны, люди предпочитают держаться от вас еще дальше. Это наука о данных. Везде наука о данных. Повсюду наука о данных.

Поскольку она, таким образом, является практически естественным делом, я обнаружил, что в лучших вариантах анализа больших данных может разобраться практически любой умный человек. Если вы не можете понять, в чем суть исследования, проблема скорее всего не в вас, а в самом исследовании.

Вам нужны доказательства того, что научная работа с большими данными, как правило, является интуитивно понятной? Недавно я наткнулся на исследование, которое может оказаться одним из самых важных среди всех, проводившихся в течение последних нескольких лет. Оно также является одним из наиболее интуитивных, которые я когда-либо видел. Мне хочется, чтобы вы подумали не только о его важности, но и о том,

насколько оно естественно и похоже на то, что делала моя бабушка.

Этот эксперимент проводила команда ученых из Колумбийского университета и из Microsoft. Целью был поиск симптомов, позволяющих предсказать зарождение у людей рака поджелудочной железы[2]. При этом заболевании только три процента больных проживают больше пяти лет, но раннее обнаружение болезни может удвоить шансы пациента.

Какой метод применили исследователи? Они использовали данные десятков тысяч анонимных пользователей Bing — поисковика Microsoft. При этом выбирали пользователей, у которых недавно был диагностирован рак поджелудочной железы — основываясь на безошибочном поисковом запросе, например: «Мне только что диагностировали рак поджелудочной железы» или «Мне сказали, что у меня рак поджелудочной железы, чего ожидать?»

Далее ученые искали запросы относительно возникающих симптомов. Они сравнили данные небольшого количества пользователей, сообщивших о своем диагнозе не сразу, с теми, кто этого вообще не сделал. Другими словами, попытались выявить, какие симптомы беспокоили тех, кто признался в своем диагнозе только через несколько недель или месяцев.

Результаты оказались просто поразительными. Признаками рака поджелудочной железы оказались боль в спине, а затем пожелтение кожи. Поисковый запрос только о боли в спине по большей части не относился к раку. Аналогично, поисковый запрос «Несварение желудка, а потом боль в животе» свидетельствует о раке

поджелудочной железы, тогда как просто несварение желудка без болей не означает этого страшного диагноза. Исследователи смогли выявить от 5 до 15% случаев практически без ложных срабатываний. Может быть, это не выглядит особо удачным результатом, но если у вас рак поджелудочной железы, даже 10%-ная возможность удвоить шансы на выживание будет восприниматься как неожиданный подарок судьбы.

Неспециалисту изложенные в статье детали исследования будет трудно осмыслить в полной мере. Они включают в себя много технических терминов, таких как тест Холмогорова — Смирнова*, смысл которого, признаться, я уже забыл.

Однако обратите внимание, насколько естественно и интуитивно понятно это замечательное исследование на самом фундаментальном уровне. Ученые рассмотрели широкий спектр медицинских случаев и попытались связать симптомы с конкретным заболеванием. А знаете, кто еще использует эту методику, пытаясь выяснить, болен человек или нет? Мужья и жены, отцы и матери, медсестры и врачи. Исходя из своего опыта и знаний, они пытаются соединить лихорадку, головную боль, насморк и боли в желудке с различными недугами. Другими словами, специалисты из Колумбийского университета и Microsoft провели новаторское исследование с использованием самой обычной и очевидной методики, издавна используемой для диагностики.

---

* Это способ определить, насколько точно созданная модель соответствует данным. — *Прим. ред.*

Но подождите. Давайте сбавим скорость. Если методика наилучшей научной обработки данных является естественной и интуитивно понятной так часто, как я утверждаю, это поднимает фундаментальный вопрос о ценности больших данных. Если люди являются прирожденными специалистами по научной обработке данных, если сама наука о данных является интуитивно понятной, зачем нужны компьютеры и программное обеспечение статистической обработки информации? Зачем нужны тесты Холмогорова — Смирнова? Разве мы не можем просто использовать свою интуицию и все? Разве мы не можем поступать так же, как это делает моя бабушка, как работают медсестры и врачи?

Подобное ощущение усилилось после выхода бестселлера Малкольма Гладуэлла «Blink» («Миг»), в котором воспевается магия человеческих инстинктов. Гладуэлл рассказывает истории о людях, которые, полагаясь исключительно на свою интуицию, могут сказать, является ли статуя поддельной, еще до удара — промажет ли теннисист по мячу или сколько клиент готов заплатить — до того, как тот откроет рот. Герои этой книги не высчитывают регрессии, они не определяют доверительные интервалы и не запускают тесты Холмогорова — Смирнова, но при этом, как правило, делают удивительные прогнозы. Многие люди подсознательно поддерживают мнение Гладуэлла об интуиции — они доверяют своему нутру и своим чувствам. Фанаты романа наверняка восторженно подчеркнут мудрость моей бабушки и ее способность давать советы по поводу человеческих отношений без помощи компьютеров. Поклонники «Blink», уверен, менее склонны восхищаться моими исследованиями или другими

наработками, описанными в этой книге, поскольку здесь используются компьютеры. Если большие данные — компьютерные, а не информация от моей бабушки — революционны, следует доказать, что они способны на большее, чем наша интуиция, работающая без посторонней помощи. Хотя она, как отмечает Гладуэлл, зачастую и может выдавать просто потрясающие результаты.

Исследование, проведенное Колумбийским университетом и Microsoft, на примере строгих научных данных и компьютерных расчетов позволяет наглядно показать то, до чего интуиция никак не может дойти. Это также тот случай, когда важную роль играет объем информации. Иногда нашей интуиции просто не хватает опыта, на который она могла бы опереться. Маловероятно, что вы, ваши друзья или члены вашей семьи видели достаточно много случаев рака поджелудочной железы, чтобы уловить разницу между несварением желудка, сопровождаемым болью в животе, и обычным несварением желудка без болей. В какой-то момент массив поисковых запросов будет становиться все больше и больше, и в результате исследователи неизбежно найдут множество менее заметных закономерностей между симптомами и проявлениями этой болезни или других заболеваний, которые могут пропустить даже опытные врачи.

Более того, хотя наша интуиция, как правило, и может дать нам хорошее общее представление об устройстве мира, она нередко не дает точного результата.

Нам нужно больше данных, чтобы увеличить четкость изображения. Рассмотрим, например, влияние погоды на настроение. Вы, вероятно, полагаете, что люди будут чувствовать себя не слишком радостно скорее

при –12 градусах, чем при +21. Да, это действительно так. Но вы можете и не догадываться, насколько велико влияние этого перепада температур. Я искал корреляции между поисковыми запросами в Google относительно депрессии и целого ряда факторов, включая экономические условия, уровень образования и посещение церкви. Зимний климат перевешивает все остальное[3]. В зимние месяцы в теплом климате (например, на Гавайях) поисковых запросов относительно депрессии на 40% меньше, чем в районах с холодным климатом (таких, как Иллинойс). Но насколько значимо это влияние? Если у вас достаточно оптимистичное представление об эффективности антидепрессантов, вы с удивлением обнаружите: даже самые лучшие препараты снижают уровень депрессии всего лишь на 20%. Насколько можно судить по цифрам, предоставляемым Google, переезд из Чикаго в Гонолулу будет как минимум вдвое эффективнее, чем любое лекарство от зимней тоски*.

Иногда наша интуиция — если не направлять ее с помощью тщательного компьютерного анализа — может повести нас в совершенно неверном направлении. Собственный жизненный опыт и устоявшиеся предрассудки могут ослепить нас. Действительно, даже бабушка, которая в состоянии использовать свой многолетний опыт, чтобы дать лучший совет в плане личных отношений, чем остальная часть семьи, все равно имеет некоторые сомнительные представления о причинах крепости

---

* Если начистоту: вскоре после завершения этого исследования я переехал из Калифорнии в Нью-Йорк. Использовать факты для понимания того, что следует сделать — легко. Сделать это на самом деле — довольно сложно. — *Прим. авт.*

отношений. Например, она часто подчеркивала важность наличия общих друзей, считая это ключевым фактором, предопределившим успешность ее брака. Она проводила самые приятные вечера со своим мужем, моим дедушкой, в их небольшом дворике в Квинсе, Нью-Йорк, сидя на раскладных стульях и сплетничая с соседями.

Тем не менее, хоть я и рискую сделать свою любимую бабушку козлом отпущения, научные данные свидетельствуют о том, что ее теория неверна. Команда ученых-компьютерщиков недавно проанализировала самый большой набор фактов о человеческих взаимоотношениях[4] из когда-либо существовавших — Facebook. Они рассмотрели большое количество пар, которые в какой-то момент состояли «в отношениях». Некоторые из этих пар остались в них, другие перешли в статус «одиночка». Как выяснили ученые, наличие общей группы друзей является довольно существенным показателем того, что отношения НЕ продлятся долго. Вероятно, тусоваться каждый вечер со своим партнером и одной и той же небольшой группой людей не так уж здорово, а вот разные круги общения, возможно, помогают укрепить отношения.

Как видно, действуя только интуитивно и отказываясь от использования компьютеров, мы, порой, приходим к удивительным результатам. Но это может привести и к серьезным ошибкам. Бабушка, надо полагать, попалась в одну из когнитивных ловушек: иногда мы склонны преувеличивать значение собственного опыта. Если говорить языком специалистов по обработке и анализу данных, мы придаем намного большее значение фактам, взятым из одного источника — нас самих.

Бабушка была настолько сосредоточена на воспоминаниях о ее вечерних встречах с дедушкой и их друзьями, что не уделила достаточного внимания другим парам. Например, она упустила возможность рассмотреть ситуацию со своим деверем и его красоткой-женой, которая весь вечер болтала с небольшой постоянной группой друзей, но часто ссорилась с мужем. В конце концов они развелись. Бабушка забыла полностью рассмотреть историю моих родителей — ее дочери и зятя. Они нередко проводили вечера каждый сам по себе: мой отец играл в джаз-клубе или в мяч со своими друзьями, а мама отправлялась в ресторан или в театр со своими приятельницами, но это не мешало им счастливо прожить много лет в браке.

Полагаясь лишь на свою интуицию, мы также можем быть обмануты базовой человеческой склонностью к драматизации происходящего. Мы любим переоценивать важность всего, что может стать основой для незабываемого сюжета. Например, в ходе одного опроса выяснилось, что торнадо считается более распространенной причиной смерти[5], чем астма. Хотя на самом деле от астмы умирает примерно в 70 раз больше людей[6]. В смерти от астмы нет ничего впечатляющего, эти случаи не попадают в новости. А вот смерти от торнадо попадают.

Другими словами, полагаясь только на услышанное или на личный опыт, мы часто неправильно судим об устройстве мира. Несмотря на то, что методология правильной работы с фактами так же интуитивна, ее результаты обычно являются парадоксальными. Наука о данных использует естественное и интуитивное человеческое свойство — способность увидеть комбинации и связи и вдохнуть в них смысл, — и наполняет его силой, демонстрируя

нам, что мир устроен совершенно не так, как мы думали. Именно это и произошло, когда я исследовал прогностические показатели успешных выступлений в баскетболе.

В детстве у меня была одна, только одна мечта. Я хотел вырасти и стать экономистом и специалистом по обработке и анализу данных. Нет, я, конечно, шучу. Я отчаянно хотел стать профессиональным баскетболистом, чтобы пойти по стопам своего кумира Патрика Юинга[7], лучшего центрового «Нью-Йорк Никс» всех времен.

Иногда мне кажется, что внутри каждого ученого, занимающегося сбором, изучением и анализом данных, сидит ребенок, пытающийся выяснить, почему его детские мечты не сбываются. Поэтому неудивительно, что в последнее время я внимательно изучал показатели, необходимые для попадания в НБА. Результаты исследования оказались неожиданными. На самом деле они лишний раз продемонстрировали, как серьезная наука о данных может изменить ваше представление о мире и насколько нелогичными могут оказаться цифры.

Я рассмотрел следующий вопрос: у кого больше шансов добиться успеха в НБА — у бедняков или у представителей среднего класса?

Большинство людей полагает, что у первых. Житейская мудрость гласит: те, кто рос в трудных условиях, возможно, родился у одинокой матери-подростка, обретают драйв, необходимый для достижения максимального успеха в этом конкурентном виде спорта.

Такую точку зрения в интервью «Спортс иллюстрейтед» высказал Уильям Эллерби, школьный тренер по баскетболу в Филадельфии. «Дети из пригородов, как правило, играют для своего удовольствия, — сказал он. — Для городских же детей игра в баскетбол — вопрос жизни и смерти»[8]. Я, увы, был воспитан родителями, счастливо жившими в пригороде Нью-Джерси и состоявшими в браке. Леброн Джеймс, лучший игрок своего поколения, родился в бедной семье у 16-летней матери-одиночки в Акроне, Огайо.

Естественно, по результатам проведенного мной интернет-опроса[9], я предположил, что большинство американцев думают так же, как тренер Эллерби и я, — что большинство игроков НБА растут в бедности.

Верно ли это расхожее мнение?

Давайте посмотрим на факты. Не существует всеобъемлющего источника данных о социоэкономике игроков НБА. Но, проведя тщательное исследование целой кучи источников (basketball-reference.com, ancestry.com, бюро переписи США и некоторые другие), мы можем понять, какие семьи больше всего способствуют успеху в НБА. Обратите внимание: в этом исследовании были использованы различные источники данных, некоторые побольше, другие поменьше, одни онлайновые, другие — вне Сети. Интересно, что, активно черпая из новых цифровых источников, хороший специалист по анализу данных не гнушается пользоваться и старомодными — если это может принести пользу. Самый лучший способ получить правильный ответ на вопрос — объединить все доступные данные.

Первая релевантная информация — родина каждого игрока. Сначала я записал, сколько черных и белых мужчин родилось в 1980-х годах в каждом округе США.

Затем — сколько из них попали в НБА. При этом сравнил эти данные со средним доходом семьи в соответствующем округе. Я также проконтролировал расовую демографию округа, поскольку (но это тема для другой книги) чернокожие мужчины попадают в НБА примерно в 40 раз чаще, чем белые.

Факты говорят нам о том, что человек имеет значительно больше шансов попасть в НБА, если он родился в более богатом округе. Например, у черного парня, появившегося на свет в одном из самых богатых округов США, вдвое больше шансов попасть в НБА, чем у черного ребенка из беднейшего округа. Вероятность попадания в НБА белого малыша, родившегося в одном из самых богатых округов, на 60% выше, чем у белого ребенка из самого бедного округа.

Это говорит о том, что, вопреки расхожему мнению, бедные люди на самом деле имеют меньше шансов попасть в НБА. Однако эти данные не идеальны, поскольку многие богатые округи США — такие, например, как графство Нью-Йорк (Манхэттен) — включают в себя и бедные кварталы вроде Гарлема. Поэтому тяжелое детство теоретически все-таки может помочь вам попасть в НБА. Нам все еще нужно больше зацепок, больше данных.

Тогда я начал исследовать семьи игроков НБА. Информацию о них находил в новостях и в социальных сетях. Эта методология оказалась довольно трудоемкой, поэтому я ограничил анализ сотней чернокожих игроков, родившихся в 1980-х годах и набравших на площадке наибольшее количество очков. По сравнению со среднестатистическим афроамериканцем, вероятность рождения суперзвезды НБА у матери-подростка или

у незамужней матери на 30% меньше. Другими словами, семейные обстоятельства лучших чернокожих баскетболистов также позволяют предположить, что хорошая семья для достижения успеха является преимуществом.

Таким образом, ни средний уровень доходов в округе, ни семейный фон ограниченной выборки игроков не дают точной информации о детстве всех баскетболистов. Поэтому я все еще не был уверен в том, что полные семьи со средним доходом производят больше звезд НБА, чем неполные и малообеспеченные. Чем больше фактов мы можем собрать для ответа на этот вопрос, тем лучше.

Потом я вспомнил еще один момент, который мог бы существенно помочь. В работе двух экономистов, Роланда Фрайера и Стивена Льюитта, было высказано предположение, что имя афроамериканца — это показатель его социально-экономического статуса[10]. Фрайер и Льюитт просмотрели свидетельства о рождении в Калифорнии за 1980-е годы и обнаружили, что бедные, необразованные и одинокие чернокожие мамы дают своим детям не такие имена, как родители из среднего класса, образованные и состоящие в браке.

Выше вероятность, что детям из более состоятельных слоев будут даны более привычные имена вроде Кевина, Криса и Джона. А вот детей из неблагополучных семей, скорее всего, назовут уникальным именем, таким как Ноушон, Уник или Брейоншей. У афро-американских детей, рожденных в нищете, вдвое выше вероятность получения имени, которым не будет назван ни один другой ребенок, родившийся в том же году.

Так что насчет имен чернокожих игроков НБА? Они звучат скорее как имена среднего класса или как имена

бедняков? Баскетболисты, рожденные в Калифорнии в один и тот же период времени, имели уникальные имена в два раза реже, чем средний чернокожий мужчина того же возраста. Это статистически значимое отличие.

Вы знаете кого-то, кто считает, что НБА — это лига для детей из гетто? Скажите ему, чтобы он просто прислушался к репортажу со следующей игры. Предложите ему обратить внимание, как часто Расселл обходит Дуайта, а затем пытается проскользнуть мимо протянутой руки Джоша и передать мяч в ожидающие руки Кевина. Если бы НБА действительно была наполнена чернокожими парнями, вышедшими из бедных семей, репортаж звучал бы совершенно по-другому. В нем было бы намного больше упоминаний людей с такими именами, как у Леброна.

Итак, мы собрали три разных ключевых показателя — место рождения, семейное положение матерей лучших игроков и их имена. Ни один источник не идеален, но все они поддерживают одну и ту же версию. Чем выше социально-экономический статус, тем выше шанс попасть в НБА. Иными словами, общепринятое представление дало осечку.

Среди всех афроамериканцев, родившихся в 1980-х годах, около 60% не имели состоявших в браке родителей[11]. Но, по моим оценкам, среди чернокожих, рожденных в том десятилетии и попавших в НБА, значительное большинство выросло в полной семье. Другими словами, у большинства баскетболистов детство было иным, чем у Леброна Джеймса. Среди них было больше таких, как Крис Бош, росший в Техасе с двумя родителями, которые привили ему интерес к электронным гаджетам. Или как

Крис Пол, второй сын родителей, относящихся к среднему классу, из Льюисвилла, Северная Каролина.

Цель специалиста по обработке и анализу данных — понять мир. Как только находится кажущийся алогичным результат, можно попробовать взять больше научных данных и объяснить, почему мир устроен не так, как нам кажется. Почему, например, мужчины из семей среднего класса имеют преимущество в баскетболе по сравнению с выходцами из бедных семей? Есть как минимум два объяснения.

Во-первых, потому, что мужчины из бедных семей, как правило, ниже ростом. Ученым давно известно, что уход за детьми и правильное питание играют большую роль и способствуют здоровью. Именно поэтому средний человек в развитых странах сейчас на 10 см выше[12], чем полтора века назад. Статистика показывает, что американцы из бедных семей из-за плохих здравоохранения и питания в детстве вырастают более низкими[13].

Статистика также может рассказать нам о влиянии роста на попадание в НБА. Вы, несомненно, догадываетесь, что высокий рост — это преимущество для начинающего баскетболиста. Просто сопоставьте этот параметр у типичного игрока на площадке и у типичного фаната на трибунах*.

Насколько большое значение имеет высокий рост? Говоря о нем, баскетболисты иногда немного привирают, да и полного списка распределения ростов американских мужчин не существует. Но работая с грубой математической оценкой, можно прикинуть это распределение

---

* Средний рост игрока НБА — 201 см, средний рост американского мужчины — 179 см[14]. — *Прим. авт.*

и сопоставить его с ростом игроков НБА. Нетрудно убедиться, что влияние роста огромно — пожалуй, даже больше, чем мы могли бы подозревать. На мой взгляд, каждый дополнительный дюйм удваивает ваши шансы попасть в НБА. И это верно для всей шкалы. Мужчина ростом 170 см имеет вдвое больше шансов попасть в НБА, чем мужчина ростом 167,5 см. Мужчина ростом 211 см имеет вдвое больше шансов попасть в НБА, чем мужчина ростом 208,5 см. Оказывается, в НБА попадает всего один из двух миллионов мужчин ростом меньше 183 см. А для тех, чей рост превышает 213 см, шанс попасть в НБА составляет где-то один к пяти.

Обратите внимание: эти данные показывают, почему моя мечта о баскетбольной славе не сбылась. Дело не в том, что я был воспитан в пригороде. Дело в том, что мой рост 175 см и я белый (не говоря уж о том, что у меня очень медленная реакция). Кроме того, я ленив. И у меня плохо с выносливостью, ужасная подача, а иногда, когда мяч попадает ко мне в руки, и панические атаки.

Вторая причина, по которой некоторые мальчики из не слишком хороших семей, могут не попасть в НБА — отсутствие определенных социальных навыков. Используя данные о тысячах школьников, экономисты обнаружили, что в семьях с двумя родителями, относящимися к среднему классу[15], воспитание детей поставлено в целом существенно лучше. И там уделяют большое внимание выработке таких навыков, как дисциплинированность, настойчивость, целеустремленность и организованность.

Каким же образом недостаточно наработанные социальные навыки пускают под откос потенциально успешную баскетбольную карьеру?

Давайте посмотрим на историю Дага Ренна, одного из самых талантливых и перспективных баскетболистов 1990-х годов. Его тренер в колледже Джим Кэлан из университета Коннектикута, подготовивший многих будущих звезд НБА, заявил, что Даг прыгал выше любого человека[16], с которым он когда-либо работал. Но характер у Ренна[17] был очень сложным. Он был воспитан матерью-одиночкой на Блад Элли — в одном из самых неблагополучных районов Сиэтла. В Коннектикуте он постоянно конфликтовал с окружающими. Ему нравилось дразнить игроков, он постоянно изводил тренеров вопросами и, в нарушение правил команды, носил свободную одежду. У него также были проблемы с законом — он украл обувь из магазина и набросился с кулаками на сотрудников полиции. Терпению Кэлана пришел конец, и Дага выгнали из команды.

Второй шанс Ренн получил в университете Вашингтона. Но и там сполна проявилась его неспособность ладить с людьми. Парень ссорился со своим тренером из-за игрового времени, а с партнерами — из-за передач мяча. В общем, его выгнали из команды и здесь. Ренн не пришелся ко двору в НБА, поиграл за разные команды низших лиг, переехал к своей матери и в конечном счете попал в тюрьму. «Моя карьера закончилась, — сказал Ренн в интервью «Сиэтл Таймс» в 2009 году. — Мои мечты, мои стремления закончились. Даг Ренн мертв[18]. Как баскетболист я мертв. Все кончено». Ренн был талантлив и мог стать не просто игроком НБА, а легендарным игроком. Но он никогда не пытался справиться со своим характером, чтобы хотя бы остаться в команде колледжа. Возможно, если бы у него было более

радужное детство, он мог бы стать следующим Майклом Джорданом.

Кстати, Майкл Джордан как раз совершил впечатляющий рывок к вершинам. У него были огромное самомнение и высокая конкурентоспособность — его характер мало чем отличался от характера Ренна. Джордан был трудным ребенком[19]. В 12 лет его выгнали из школы за драку. Но у него по крайней мере было то, чего не хватило Ренну — хорошее воспитание, характерное для среднего класса. Его отец был инженером-механиком и начальником смены в «Дженерал Электрик»[20], а мать работала в банке. И они помогали ему принять решение в карьере.

Действительно, жизнеописание Джордана наполнено историями о том, как семья помогала ему обойти ловушки[21], в которые могла попасть эта талантливая и стремящаяся к постоянному соперничеству личность. После того, как Майкла выгнали из школы, мама взяла его с собой на работу. Ему не разрешили выйти из машины, вместо этого мальчишке пришлось сидеть в ней на стоянке и читать книги. После того, как его взяли в команду «Чикаго Буллз», родители, братья и сестры по очереди навещали его, чтобы убедиться, что он избегает искушений, которые приходят вместе со славой и деньгами.

Карьера Джордана закончилась не так, как у Ренна с его интервью в «Сиэтл Таймс». Майкл завершил свой славный путь речью перед введением его в баскетбольный Зал славы[22], которую смотрели миллионы людей. В своем выступлении Джордан сказал, что он всегда старался «концентрироваться только на хорошем — вы же знаете, как люди воспринимают вас, если вы их уважаете… как вас воспринимают публично». «Остановитесь

на минуту и подумайте о том, что вы делаете. И все это — благодаря моим родителям».

Факты говорят нам, что Джордан был абсолютно прав, поблагодарив своих женатых родителей, относящихся к среднему классу. Факты говорят нам, что в неблагополучных семьях, в неблагополучных общинах есть талантливые люди, которые вполне годятся для игры в НБА, но которые никогда туда не попадут. Эти люди имеют подходящие гены, имеют амбиции, но они никогда не занимались формированием характера, необходимого для того, чтобы стать суперзвездами баскетбола.

И, как подсказывает нам интуиция, даже пребывание в обстоятельствах настолько ужасных, что баскетбол становится «вопросом жизни и смерти», не помогает. Это отлично иллюстрируют истории вроде судьбы Дага Ренна. А факты подтверждают интуитивное представление.

В июне 2013 года Леброн Джеймс дал интервью[23] на телевидении после того, как во второй раз победил в чемпионате НБА. (С тех пор он уже победил и в третий раз.) «Я Леброн Джеймс — объявил он — из Акрона, штат Огайо. Городской житель. Я даже не должен был быть здесь». Twitter и другие социальные сети немедленно разразились критикой. Как мог такой высокоодаренный человек, которому еще в невероятно раннем возрасте прочили блестящее баскетбольное будущее, говорить о своем аутсайдерском статусе? На самом же деле любой, кто находился в похожих тяжелых начальных условиях, независимо от своих спортивных способностей, не имел бы никаких шансов. Другими словами, достижения Джеймса еще прекраснее и значительнее, чем кажутся на первый взгляд. И факты также подтверждают это.

ЧАСТЬ II

МОГУЩЕСТВО БОЛЬШИХ ДАННЫХ

Глава 2

ВОЗМОЖНО, ФРЕЙД БЫЛ ПРАВ?

Недавно я слышал, как идущего по улице мужчину обозвали — «penistrian» (игра слов: pedestrian — пешеход, penis — пенис; получается «членоход»). Вы уловили? «Penistrian» («Членоход») вместо «pedestrian» («Пешеход»). Я видел подобное во многих поисковых запросах. Человек видит, как кто-то шагает, и пишет слово «Penis» («пенис»). Это ведь должно что-то означать, правда?

Недавно я узнал об одном мужчине, которому ужасно хотелось банан в тот момент, когда он шел к алтарю навстречу своей будущей жене. Я видел подобное в подборках больших данных о фантазиях, которыми люди делятся в сети. Мужчина думает о поедании фрукта фаллической формы в момент, когда собирается жениться на женщине. Это же что-то значит?

Возможно, Фрейд был прав? С того самого момента, когда его теории впервые были вынесены на суд широкой общественности, самым честным ответом на этот вопрос будет пожатие плечами. Ясность внес Карл Поппер — австрийско-британский философ. Он утверждал,

что теорию Фрейда нельзя подделать. Не было никакого способа проверить, истинны они или ложны.

Фрейд мог бы сказать: человек, написавший «penistrian», возможно, проявил таким образом свое подавляемое сексуальное желание. А человек мог бы ответить, что он ничего не проявляет, что это вполне могла быть невинная опечатка — такая же, как, например, «pedaltrian» («педалеход»). Это просто ситуация из серии «он сказал, она сказала». Фрейд мог бы утверждать, что господин, мечтающий в день своей свадьбы о том, чтобы съесть банан, думает о пенисе — и это раскрывает его тайное желание выйти замуж за мужчину, а не жениться на женщине. На что сей джентльмен мог бы ответить, что он просто хотел банан. Идя к алтарю, он с тем же успехом мог бы думать о яблоке.

Не было никакого способа по-настоящему испытать теорию Фрейда. До настоящего времени.

Наука о данных делает многие моменты теории Фрейда опровержимыми, и это позволяет проверить ее на прочность. Начнем с фаллических символов во сне. Используя огромный массив данных из записанных снов, мы можем легко заметить, как часто в них появляются предметы фаллической формы. Еда — хороший объект, на котором можно сосредоточить свое внимание. Она появляется во многих снах, и многие продукты имеют форму фаллоса[1] — бананы, огурцы, сосиски и т. д.

Мы можем измерить факторы, которые заставляют нас видеть во сне одни продукты чаще других: как часто их едят, насколько вкусными находит их большинство людей, и — да! — действительно ли у них фаллический вид.

Можно протестировать два одинаково популярных продукта, один из которых имеет форму фаллоса. Насколько чаще в снах появляется тот или другой продукт? Если еда, имеющая форму фаллоса, не появляется в наших снах чаще продуктов другой формы, значит, фаллические символы не являются значимым фактором наших сновидений. Благодаря большим данным эту часть теории Фрейда можно реально опровергнуть.

Я получил сведения от Shadow — приложения, предлагающего пользователям записывать свои сны, — и проанализировал продукты, включенные в десятки тысяч снов.

В целом, что заставляет нас видеть во сне еду? Основным прогностическим фактором является то, как часто мы едим именно эти продукты. Вещество, которое мы видим во сне чаще всего — вода. В первую двадцатку продуктов из сновидений входят курица, хлеб, бутерброды и рис. Заметьте, все не по Фрейду.

Второй прогностический фактор, показывающий, насколько часто тот или иной продукт будет появляться в наших снах, это то, насколько вкусным мы его считаем. Два продукта, которые мы наиболее часто видим во сне, также не согласуются с теорией Фрейда: это шоколад и пицца.

А что относительно фаллической формы продуктов? Возможно, еда такой формы проникает в наши сны неожиданно часто? Ни в коей мере.

Бананы являются вторым по частоте появления в снах плодом. Но они также являются вторым по частоте употребления фруктом. Поэтому для объяснения того, почему мы так часто видим во сне бананы, Фрейд нам не нужен. Огурцы — седьмой по частоте появления в снах овощ. Но они занимают седьмое место в списке наиболее потребляемых овощей. Так что не надо объяснять их присутствие в наших снах формой. Хот-доги снятся гораздо реже, чем гамбургеры. Это верно, учитывая тот факт, что люди едят больше бургеров, чем хот-догов.

В целом, используя регрессионный анализ (метод, позволяющий ученым при сборе и анализе данных разделить воздействия нескольких факторов), я обнаружил: еда в форме фаллоса не появляется в наших снах с большей вероятностью, чем можно было бы ожидать при популярности каждого продукта. И это верно для всех фруктов и овощей. Таким образом, эта теория Фрейда является опровергаемой и, по крайней мере согласно собранной мной информации, ложной.

Далее рассмотрим оговорки по Фрейду. Великий психолог предположил, что наши устные или письменные оговорки или описки раскрывают наши подсознательные желания, часто сексуальные. Можем ли мы использовать большие данные, чтобы проверить это? Вот один из способов: посмотреть, не сводятся ли наши оговорки к сексуальным мотивам. Если наши подавленные сексуальные желания способны проникнуть в письмо или речь, должно быть огромное количество ошибок с внедрением таких слов, как «член» и «секс».

Вот почему я изучил набор из более 40 000 опечаток, собранных исследователями корпорации Microsoft[2]. Эти

данные включали ошибки, которые люди делали, но потом сразу же исправляли. Среди этих десятков тысяч ошибок во многих имелся различного рода сексуальный подтекст. Был там и вышеупомянутый «penistrian». И еще нашелся запрос, в котором напечатали «sexurity» вместо «security» («безопасность») и «cocks» (просторечное обозначение пенисов) вместо «rocks» («камни, скалы»). Но также имелось и множество невинных опечаток. Люди печатали «pindows», «fegetables», «aftermoons» и «refriderators».

Может быть, количество сексуальных опечаток необычно велико?

Чтобы проверить это, я использовал вышеуказанный набор для того, чтобы смоделировать, как часто люди путают определенные буквы. Сперва подсчитал, как часто они заменяют t и c, g и h. Затем написал программу, которая делала ошибки так же, как это могли бы сделать люди. Мы могли бы назвать эту программу Error Bot. Этот бот заменял t на c с той же частотой, что и люди в исследовании Microsoft. И g на h. И так далее. Я запустил программу, набирая те же слова, которые хотели напечатать люди в исследовании Microsoft. Другими словами, бот пытался набрать слова «пешеход», «скалы», «окна» и «холодильник». Но он так же часто, как люди, путал r и t и писал, например, вместо «rocks» — «tocks» («ягодицы»). И так же часто, как люди, путал r и c и писал вместо «rocks» — «cocks».

Так что же мы узнаем из сравнения программы Error Bot с обычными небрежными людьми? Сделав несколько миллионов ошибок, просто путая буквы так же, как это делают люди, Error Bot сделал множество опечаток по Фрейду. Вместо «seashell» программа писала — «sexshell», вместо «lipstick» — «lipsdick», вместо «luckiest» — «fuckiest» и делала много других подобных опечаток. И вот ключевой момент. Error Bot, у которого, конечно же, нет подсознания, делал ошибки с той же вероятностью, что и реальные люди, опечатки которых воспринимаются как сексуальные. С оговоркой — как мы, социологи, любим говорить, — что необходимо провести дополнительные исследования. Это означает, что сексуально ориентированные ошибки встречаются не чаще, чем просто случайные.

Иными словами, когда люди делают опечатки и пишут «penistrian», «sexurity» и «cocks», совсем необязательно существование какой-то связи между ошибками и запретным. Не факт, что посредством этих описок разум людей раскрывает свои тайные желания. Эти опечатки могут быть объяснены обычными промахами пальцев. Люди делают много ошибок. И если у вас это случается достаточно регулярно, в конце концов обязательно получится что-то вроде «lipsdick», «fuckiest» и «penistrian». Если обезьяна достаточно долго будет бить по клавишам, она в конце концов напишет «быть или не быть». Если человек достаточно долго печатает, он в конце концов может написать «penistrian».

Теория Фрейда о том, что оговорки демонстрируют содержание нашего подсознания, является, согласно моему анализу данных, ложной.

Большие данные говорят нам, что банан — это всегда просто банан, а «penistrian» — просто «pedestrian», но напечатанный с ошибкой.

Но неужели Фрейд промахнулся со всеми своими теориями? Не совсем. Когда я впервые получил доступ к данным PornHub, меня посетило откровение: я наконец нашел то, что показалось мне хоть в чем-то фрейдистским. По сути, это один из самых удивительных моментов, обнаруженных в ходе моей работы с данными: шокирующее количество людей, посещающих наиболее крупные порносайты, ищут изображение инцеста.

16 из 100 наиболее частых поисковых запросов мужчин на одном из самых популярных порносайтов PornHub посвящены видео инцеста. Честно предупреждаю: это довольно живописная картина. Среди них «брат и сестра», «мачеха трахает пасынка», «мама и сын», «мать трахает сына» и «реальные брат и сестра». Больше всего поисковых запросов по кровосмесительным связям мужчины делают относительно сцены с участием матери и сына. А женщины? Девять из ста наиболее частых поисковых запросов женщин на PornHub по поводу видео инцеста включают похожие образы, хотя пол родителя и ребенка, как правило, прямо противоположные. То есть женщины ищут кровосмесительные сцены с участием отцов и дочерей.

Нетрудно предположить в этих фактах хотя бы слабое эхо эдипова комплекса, описанного Фрейдом. Он предположил, что в детстве почти у всех возникает желание половых отношений с родителем противоположного пола, которое позже подавляется. Если бы австрийский психолог прожил достаточно долго, он мог бы применить

свои аналитические навыки к данным PornHub, где столь ярко и четко выражена совсем не подавленная заинтересованность взрослых людей к родителю противоположного пола.

Конечно, данные PornHub не могут точно показать, о ком фантазируют люди, когда смотрят подобное видео. Они и в самом деле представляют секс с собственным родителем? Поисковые запросы в Google могут подтвердить, что в мире есть много людей с подобными желаниями.

Рассмотрим все запросы, начинающиеся со слов «я хочу секса с…»[3]. На первом месте среди завершающих слов стоит «мама». В целом 82,7% поисковых запросов в подобной форме являются кровосмесительными. И это не связано с конкретной формулировкой. Например, при поиске в форме «меня привлекает…» признаний кровосмесительных желаний еще больше. Теперь я, рискуя разочаровать господина Фрейда, не исключаю, что это не особо распространенные поисковые запросы: ежегодно в США во влечении к своей матери признаются несколько тысяч человек. Кому-то даже придется подготовить господина Фрейда к новости о том, что поисковые запросы в Google (о чем в этой книге будет говориться позже) лишь иногда перекашиваются в сторону запретного.

Но все же… У людей много неподходящих желаний, которые, как мне казалось, должны чаще проявляться в поисковых запросах. Босс? Служащий? Студент? Терапевт? Пациент? Лучшая подруга жены? Лучшая подруга дочери? Сестра жены? Жена лучшего друга? Ни одно из этих желаний, проявившихся в поисковых запросах,

не может конкурировать с желанием обладать матерью. Возможно, в сочетании с данными PornHub это действительно что-то да значит.

Кстати, главное утверждение Фрейда о том, что сексуальность может быть сформирована в детстве, поддерживается данными Google и PornHub. Они дают понять, что мужчины по крайней мере сохраняют невероятное количество фантазий, связанных с детством. По данным поисковых запросов жен о своих мужьях, некоторые из самых популярных фетишей мужчин — желание носить памперсы и чтобы их кормили грудью. Особенно, как уже говорилось раньше, это распространено в Индии. Нельзя не упомянуть и о большой популярности порномультфильмов[4] — анимированных откровенных сексуальных сцен с участием персонажей из шоу, любимых мальчиками-подростками. Или рассмотрим вопрос о профессии женщин, чаще всего востребованной мужчинами в порно. Мужчины в возрасте 18–24 года чаще всего вводят в поисковый запрос профессию няни[5]. То же самое можно сказать и о мужчинах в возрасте 25–64 года, и о мужчинах от 65 лет и старше. Кроме того, для мужчин каждой возрастной группы в первую четверку наиболее привлекательных профессий входят учительница и черлидерша. Очевидно, что в формировании взрослых мужских фантазий первые годы жизни играют важную роль.

Я пока не в состоянии использовать все эти беспрецедентные данные о сексуальности взрослых для определения, как именно формируются сексуальные предпочтения. В течение следующих нескольких десятилетий социологи — и я в том числе — смогут создать новые

опровергаемые теории о сексуальности взрослых людей и проверить их с помощью фактов.

Но уже сейчас могу предсказать некоторые основные темы, которые, несомненно, станут частью теории о взрослой сексуальности, возникшей на базе большого объема данных. Она явно не будет идентична теории Фрейда с его отдельными, четко определенными универсальными стадиями детства и подавления. Но, основываясь на моем первом обзоре данных PornHub, я абсолютно уверен: в окончательный вердикт о взрослой сексуальности некоторые ключевые обозначенные Фрейдом темы обязательно будут включены. Главную роль будет играть детство человека. И его мать.

Наверное, еще десять лет назад было бы невозможно анализировать теорию Фрейда подобным образом. И конечно, нечто подобное было неисполнимо 80 лет назад, когда Фрейд был еще жив. Итак, давайте подумаем, почему эти источники данных смогли нам помочь? Благодаря такому упражнению мы поймем, из-за чего большие данные настолько могущественны.

Помните, мы уже говорили, что даже наличие целой россыпи фактов само по себе не позволит нам автоматически генерировать полезные выводы. Ученые переоценили объем данных. Но почему же тогда большие данные настолько могущественны? Почему они оказались способными революционно преобразовать наше видение самих себя? Я утверждаю, что существуют четыре уникальные особенности больших данных, и анализ Фрейда способен отлично проиллюстрировать это.

Прежде всего, вы, наверное, заметили: обсуждая теории Фрейда, мы довольно серьезно отнеслись к порнографии. Более того, в этой книге мы намерены достаточно часто обращаться к анализу порносайтов. Это довольно странно, ведь данные, полученные из такого источника, редко используются большинством ученых. Последние обычно удобно опираются на результаты традиционных опросов — и именно на них выстраивают свои карьеры. Но если немного подумать, становится ясно, что широкое использование данных порносайтов (а также поиск по ним и обработка полученных таким образом сведений) позволяет лучше понять человеческую сексуальность. На самом деле это, наверное, самое важное на свете. Получив такие данные, Шопенгауэр, Ницше, Фрейд и Фуко визжали бы от восторга, однако в то время, когда они жили, подобных данных не существовало. Их не было еще пару десятилетий назад, но они есть сейчас. Существует множество уникальных источников информации по различным темам, открывающих нам глаза в областях, о которых ранее мы могли только догадываться. *Способность предложить нам новые типы фактов — первая могущественная особенность больших данных.*

Данные порносайтов и поисковых запросов Google не только новые, они самые правдивые. В доцифровое время люди прятали свои постыдные мысли от других. В эпоху цифровых технологий они продолжают их прятать — но не от интернета и, в частности, не от сайтов вроде Google и PornHub, где поддерживается анонимность. Подобные сайты играют роль своего рода цифровой сыворотки правды — именно это позволило нам

открыть популярность темы инцеста. Большие данные позволяют нам наконец увидеть, чего люди хотят на самом деле, а не то, что они говорят или делают. *Предоставление самых правдивых фактов является второй могущественной особенностью больших данных.*

Поскольку сейчас существует огромное количество разнообразных сведений, можно найти содержательную информацию даже о самом небольшом популяционном срезе. Мы в состоянии сравнить, скажем, количество людей, видящих во сне огурцы, с теми, кто видит во сне помидоры. *Возможность пристально вглядеться в самые мелкие подмножества людских сообществ — это третья могущественная особенность больших данных.*

Большие данные обладают еще одной внушительной возможностью — той, которую я не использовал в своем кратком исследовании теории Фрейда, но которую я наверняка применю в будущем: она дает возможность проводить быстрые контролируемые эксперименты. Это позволяет определить причинно-следственную связь, а не просто корреляцию. Такие тесты в основном используются коммерческими предприятиями, но они станут мощным инструментом в руках социологов. *Возможность проводить многочисленные причинно-следственные эксперименты — это четвертая могущественная особенность больших данных.*

Теперь пришло время более подробно поговорить о каждой из этих великолепных особенностей и разобраться, почему большие данные настолько важны.

Глава 3

ПЕРЕОСМЫСЛЕНИЕ ДАННЫХ

В 6 часов утра в определенную пятницу каждого месяца улицы большей части Манхэттена будут практически пустыми. Магазины будут закрыты, их фасады скрыты за стальными ставнями, а в квартирах над ними будет темно и тихо.

Напротив, все этажи здания Goldman Sachs, всемирно известного инвестиционного банковского учреждения, расположенного в Нижнем Манхэттене, будут ярко освещены, его лифты будут сновать туда-сюда, поднимая тысячи людей, едущих к своему рабочему месту. К 7 утра большинство столов будут заняты.

Можно без сомнения назвать этот час здесь в любой другой день сонным. Однако в эту пятницу тут будут кипеть энергия и азарт, потому что в этот день должна прибыть информация, которая окажет значительное влияние на фондовый рынок.

Через несколько минут после появления она будет растиражирована на новостных сайтах. Еще через несколько секунд она начнет обсуждаться и рассматриваться со всех сторон — в Goldman и сотнях других финансовых компаний. Но основная часть действий

в области финансов в эти дни происходит за миллисекунды. Goldman и другие финансовые компании платят десятки миллионов долларов, чтобы получить доступ к оптоволоконным кабелям, сокращающим время передачи информации из Чикаго в Нью-Джерси на четыре миллисекунды (с 17 до 13). У финансовых фирм имеются алгоритмы[1] для чтения информации и торговли на ее основе, и все это происходит за мгновения. После получения важнейших для финансового рынка данных они будут действовать быстрее, чем вы моргаете.

Так что это за важные данные, которые так ценны для Goldman и ряда других финансовых институтов?

Месячная ставка по безработице.

Эта ставка, однако, оказывает такое огромное влияние на фондовый рынок, что финансовые учреждения сделали все от них зависящее для увеличения скорости получения этих данных, их анализа и реагирования в соответствии с полученной информацией. Последняя является результатом телефонного опроса, который проводит Бюро статистики труда, и к моменту опубликования она уже устареет примерно на три недели — или 2 миллиарда миллисекунд.

При том что фирмы тратят миллионы долларов для ускорения поступления потока информации на миллисекунды, вам может показаться более чем странным тот факт, что правительству для вычисления уровня безработицы требуется так много времени.

Действительно, ускорение получения этих цифр было одним из самых важных пунктов в повестке дня Алана Крюгера[2], когда он в 2011 году занял пост председателя президентского совета по экономике США при Бараке Обаме. Это ему не удалось. «Либо BLS (Бюро трудовой

статистики Министерства труда США) не хватает ресурсов, — заключил он, — либо их мышление застряло в XX веке».

Поскольку правительство в ближайшее время явно не наберет нужный темп, возникает вопрос: есть ли способ быстрее получить хотя бы приблизительное представление о статистике безработицы? В нашу высокотехнологичную эпоху, когда почти каждый клик любого человека в интернете где-то записывается, неужели нам действительно придется ждать несколько недель, чтобы выяснить, сколько людей остались без работы?

Одно из возможных решений родилось под влиянием работы бывшего инженера компании Google Джереми Гинзберга. Он заметил, что данные о состоянии здоровья, как и сведения по безработице, правительство выпускает с задержкой. Центрам по контролю и профилактике заболеваний требуется неделя для подготовки данных об эпидемии гриппа[3], хотя врачам и больницам было бы полезно иметь такие сведения как можно раньше.

Гинзберг подозревал, что заболевание гриппом напрямую связано с поисковыми запросами относительно его лечения. В сущности, люди сообщают о своих симптомах Google. Джереми решил, что эти запросы могут дать достаточно точную оценку текущему состоянию заболеваемости гриппом. И действительно, такие поисковые фразы как «симптомы гриппа» и «боль в мышцах» оказались важными показателями скорости распространения этого заболевания*.

---

* Первоначальная версия Google Flu имела существенные недостатки, поэтому исследователи недавно создали намного более успешную модель. — *Прим. авт.*

Тем временем инженеры компании Google создали сервис Google Correlate, дающий внешним исследователям средства экспериментирования с тем же типом анализа в достаточно широком диапазоне полей, а не только в здоровье. Исследователи могут взять любой ряд данных, которые они отслеживают, и посмотреть, какие поисковые запросы в Google наиболее явно коррелируют с ним.

Например, с помощью Google Correlate мы с Хэлом Варианом, главным экономистом Google, сумели выяснить, какие поисковые запросы позволяют наиболее точно отслеживать динамику изменения цен на жилье[4]. Когда последние растут, американцы, как правило, используют для поиска такие фразы, как «80/20 ипотека», «новый дом от застройщика» и «увеличение стоимости капитала». Когда же они падают, люди чаще всего ищут «процесс продажи без покрытия», «отрицательная ипотечная стоимость» и «снижение ипотечной задолженности».

Так может быть, поиск в Google можно использовать в качестве лакмусовой бумажки для оценки безработицы таким же образом, как он используется для оценки стоимости жилья или распространения эпидемии гриппа? В состоянии ли мы, просто оценивая запросы людей в Google, сказать, сколько из них не имеют работы? И можно ли сделать это достаточно точно до того, как правительство соберет и обнародует свои результаты опросов?

В один прекрасный день я ввел в Google Correlate запрос «Уровень безработицы в США в период с 2004 по 2011 год».

Как вы думаете, какие из триллионов запросов в Google за это время оказались наиболее тесно связаны с безработицей? Вы можете подумать, что это «биржа труда»

или что-то подобное. Да, количество таких запросов увеличилось, но не они были на самом верху списка. «Новые рабочие места»? Тоже много, но не первые.

Наиболее высокий уровень запросов за рассматриваемый мной период был со словами «Slutload». Вы верите? Чаще всего люди искали порнографический сайт с таким названием. Это может показаться странным — на первый взгляд. Но у безработных людей внезапно появляется очень много свободного времени. Многие из них застряли дома одни, и им скучно. Еще очень часто встречается запросов «игра «паук». Опять же, это не удивительно для группы людей, у которых, предположительно, внезапно оказалось очень много свободного времени.

Сейчас я не хочу спорить, но, основываясь на этом анализе, могу сказать: отслеживание «Slutload» или игры «паук» является лучшим способом прогнозирования уровня безработицы. Со временем могут появляться некоторые отклонения: безработные могут искать, например, «rawtube» — другой порносайт. Ни одно из этих условий само по себе не связано с увеличением числа безработных. Но в целом я обнаружил, что смесь подобных поисковых запросов позволяет адекватно оценивать уровень безработицы и является частью самой лучшей модели прогнозирования этого явления.

Данный пример иллюстрирует могущество больших данных: возможность переосмыслить то, что следует квалифицировать как данные. Часто наиболее ценным в больших данных является не их размер, а тот факт, что они могут предложить вам новые виды информации для исследования, которые никогда раньше не собирались.

До появления Google существовали сведения об определенных видах деятельности (например, о продаже билетов в кино), которые могут дать подсказки о том, каким количеством свободного времени располагают люди. Но возможность узнать, сколько из них раскладывают пасьянс или смотрят порно — это нечто новое, и это очень мощный ресурс. В данном случае эта информация способна помочь нам быстрее оценить состояние экономики — по крайней мере, до тех пор, пока правительство не научится быстрее проводить опросы и обобщать полученные данные.

Жизнь в кампусе Google в Маунтин-Вью, Калифорния, существенно отличается от той, которая кипит в штаб-квартире Goldman Sachs на Манхеттене. В 9 часов утра офисы Google почти пусты. Если в поле зрения оказывается кто-либо из работников, скорее всего, он пришел, чтобы съесть бесплатный завтрак — бананово-черничные блинчики, омлет и огуречную воду. Некоторых сотрудников может просто не быть в городе — они присутствуют на выездном заседании в Боулдере, в Лас-Вегасе или, возможно, принимают участие в бесплатном лыжном походе к озеру Тахо. Примерно в обеденное время волейбольная площадка и футбольное поле наполнятся людьми. Лучший буррито, который я когда-либо ел, был в мексиканском ресторане Google.

Как одна из крупнейших и наиболее конкурентоспособных технологических компаний в мире может быть настолько расслабленной и щедрой? Google собирает урожай больших данных так, как даже не снилось ни одной другой компании в мире. Это позволяет ей создать

автоматизированный денежный поток. А также стать главным героем данной книги, ведь поисковые запросы в Google на сегодняшний день являются доминирующим источником больших данных. Но важно помнить: успех Google основан на сборе нового типа данных.

Если вы живете достаточно давно и пользовались интернетом еще в XX веке, то можете вспомнить различные существовавшие тогда поисковые системы — в частности, MetaCrawler, Lycos, AltaVista. И вы, наверное, помните, что эти поисковые системы были в лучшем случае не особо надежными. Иногда, если вам везло, им удавалось найти то, что вы хотели. Но нередко они не справлялись с этой задачей. Если в конце 1990-х годов вы вводили в самых популярных поисковиках запрос «Билл Клинтон», на вершине списка результатов мог оказаться случайный сайт с заголовком «Bill Clinton Sucks» («Билл Клинтон сосет») или сайт с неприличными анекдотами о Клинтоне. Вряд ли это можно считать самой актуальной информацией о тогдашнем президенте США.

В 1998 году появился Google, и результаты его поиска были несомненно лучше, чем у любого из его конкурентов. Если вы в 1998 году вводили запрос «Билл Клинтон» в Google, вам выдавался его веб-сайт, адрес электронной почты Белого дома и лучшие биографии этого человека, которые тогда существовали в интернете. Работа Google казалась волшебством.

Что же изменили основатели компании Google Сергей Брин и Ларри Пейдж?

Другие поисковые системы находили для своих пользователей веб-сайты, в которых чаще всего фигурируют фразы, введенные в поисковый запрос. Если вы искали информацию о Билле Клинтоне, эти поисковики нашли бы в сети сайты с наибольшим числом упоминаний Билла Клинтона. Существует множество причин, по которым эта рейтинговая система была несовершенной, и одной из них было то, что ее легко обмануть. Сайт с анекдотами, на странице которого будет написано «Билл Клинтон Билл Клинтон Билл Клинтон Билл Клинтон Билл Клинтон», в выдаче будет выше, чем официальный сайт Белого дома*.

Брин и Пейдж нашли способ фиксировать новый тип информации, который был гораздо ценнее, чем простой подсчет слов. Нередко в публикуемых на сайтах статьях даются ссылки на другие ресурсы, которые могут быть полезными для понимания обсуждаемого вопроса. Например, если в статье в электронной версии «Нью-Йорк Таймс» упоминается Билл Клинтон, то читатели, кликнув на его имя, перейдут на официальный сайт Белого дома.

Каждый ресурс, создающий одну из таких ссылок, в некотором смысле, демонстрирует свою точку зрения на информацию по Биллу Клинтону. Брин и Пейдж сумели объединить все эти точки зрения на каждую тему. Их поисковик мог собрать мнения «Нью-Йорк Таймс»[6],

---

* В 1998 году, если вы искали «машина» в популярной до-Google поисковой системе, вас завалили бы адресами порносайтов[5]. Там было написано слово «машина» — часто белыми буквами на белом фоне, — чтобы обмануть поисковик. В результате эти сайты получали несколько дополнительных кликов от людей, желавших купить автомобиль, но отвлекшихся на порно. — *Прим. авт.*

миллионы рассылок, сотни мнений блогеров и все остальное, что есть в интернете. Поскольку множество людей считают, что самая релевантная ссылка по запросу «Билл Клинтон» — его официальный сайт, его большинство людей и ищут, набирая слова «Билл Клинтон».

Подобные ссылки были теми данными, которые не учитывали другие поисковые системы. Эти данные были невероятно предиктивны и определяли наиболее полезную информацию на заданную тему. Дело в том, что доминирование Google среди поисковых систем определяется не просто сбором большего количества данных, чем остальные — оно зиждется на нахождении *более качественных* данных. Меньше чем через два года после своего запуска компания Google, анализируя ссылки, стала самой популярной поисковой системой в интернете. Сегодня Брин и Пейдж вместе стоят больше 60 миллиардов долларов.

И Google, и все остальные поисковые системы пытаются использовать данные, чтобы помочь нам понять окружающий мир. Революционная суть больших данных не в том, чтобы собирать все больше и больше сведений. Она в том, чтобы собирать только нужные.

Но интернет — не единственное место, где можно собрать новые факты и где получение правильных данных может иметь революционные результаты. Эта книга во многом о том, как сведения из интернета способны помочь нам лучше понимать людей. В следующем подразделе, однако, мы не будем заниматься интернет-данными. Это даже не будет иметь ничего общего с людьми. Но описанная там история поможет проиллюстрировать основную идею этой главы: огромную ценность новых, нетрадиционных данных. И принципы, которым мы

можем научиться на этом примере, помогут нам понять суть опирающейся на цифровую базу революции в области данных.

ТЕЛО КАК ИНФОРМАЦИЯ

Летом 2013 года гнедой конь выше среднего роста с черной гривой стоял в деннике в небольшом сарае в штате Нью-Йорк. Он был одним из 152 однолеток, предназначенных для августовской продажи в Саратога-Спрингс, и одним из 10 тысяч годовалых лошадей, выставленных на аукцион в этом году.

Состоятельные мужчины и женщины, готовые раскошелиться и выложить огромные деньги за лошадь, хотят самостоятельно выбрать ей имя. В результате гнедой конь тогда еще не имел клички и, как и большинство лошадей на аукционе, вместо этого назывался по номеру денника — 85.

Чтобы выделить № 85 на этом аукционе, почти ничего не делалось. У него была хорошая родословная, но не исключительная. Его отец Pioneer of the Nile был хорошей скаковой лошадью, но другие дети Pioneer of the Nile не добивались особых успехов на скачках. Имелись и сомнения, основанные на экстерьере № 85: у него была царапина на лодыжке, отпугивавшая озабоченных покупателей, поскольку могла быть свидетельством травмы.

Владельцем № 85 был египетский пивной магнат Ахмед Заят, приехавший в Нью-Йорк продать одну лошадь и прикупить несколько других.

Как почти все владельцы, Заят нанял команду специалистов, которые должны были помочь ему выбрать

лошадей для покупки. Но его эксперты отличались от обычных. Типичными «знатоками», которых вы могли бы увидеть на подобном мероприятии, были мужчины среднего возраста, многие из которых приехали из Кентукки или сельской части Флориды, с низким уровнем образования, но чья семья испокон веков вращалась в конном бизнесе. Однако специалисты Заята работали в небольшой фирме под названием EQB. Ее глава не был лошадником, принадлежавшим к старой школе. Напротив, им был эксцентричный Джефф Седер, родившийся в Филадельфии и имевший множество гарвардских степеней.

Заят и раньше работал с EQB, так что процесс выбора был ему знаком. Седер с командой несколько дней оценивали бы предлагаемых лошадей, после чего вернулись бы к Ахмеду со списком из пяти лотов, которые они рекомендовали бы к покупке на замену № 85.

На этот раз, правда, все было по-другому. Команда Седера пришла к Заяту и сказала, что не в состоянии выполнить его просьбу. Эксперты просто не могли посоветовать ему купить ни одну из 151 лошади, выставленной на продажу в тот день. Вместо этого они высказали неожиданную и почти отчаянную просьбу: Заят ни в коем случае не должен продавать № 85. «Эта лошадь, — заявил эксперт из EQB, — не просто лучшая на аукционе, она лучшая лошадь года и, вполне возможно, десятилетия». «Продай свой дом, — упрашивали Заята специалисты, — но не продавай эту лошадь»[7].

Но на следующий день после недолгих торгов № 85 был куплен за 300 тысяч долларов человеком, называвшим себя Инкардо Блудстоком. Как позже выяснилось,

это был псевдоним, используемый Ахмедом Заятом. В ответ на мольбы Седера Заят купил свою собственную лошадь, что было почти беспрецедентным явлением. (Правила аукциона не позволяли Заяту просто снять лошадь с торгов, в результате чего ему пришлось совершать эту сделку под псевдонимом.) 62 лошади были проданы на том аукционе за более высокую цену, а две — даже дороже 1 миллиона долларов каждая.

Спустя три месяца Заят наконец выбрал имя для № 85: Американский Фараон. 18 месяцев спустя в жаркий субботний вечер в пригороде Нью-Йорка Американский Фараон стал первой за более чем три десятилетия лошадью, выигравшей тройную корону.

Что же такое знал Джефф Седер о № 85, по-видимому, неизвестное никому другому? Как этому выходцу из Гарварда удавалось так хорошо оценивать лошадей?

Я познакомился с Седером[8], которому тогда было 64 года, в июне в Окале, штат Флорида — более чем через год после того, как Американский Фараон завоевал тройную корону. Там проходил недельный осмотр двухлеток, завершившийся аукционом — таким же, как тот, на котором в 2013 году Заят купил свою собственную лошадь.

У Седера раскатистый голос, как у Мэла Брукса, копна волос, при ходьбе он заметно подпрыгивает. Он был одет в брюки с подтяжками цвета хаки, черную рубашку с логотипом своей компании, в ухе виднелся слуховой аппарат.

В течение последующих трех дней он рассказывал мне свою историю — в том числе и о том, как ему удается так хорошо предсказывать будущее лошадей. Вряд ли это был прямой путь. После окончания с отличием Гарварда

и Фи Бета Каппа* Седер там же получил юридическое образование и степень по бизнесу. В 26 лет он уже работал аналитиком в компании Citigroup в Нью-Йорке, но чувствовал себя несчастным и выгоревшим дотла. Однажды, сидя в атриуме нового офисного здания компании на Лексингтон-авеню он обнаружил, что внимательно рассматривает большую фреску, изображающую бескрайнее поле. Картина напомнила о его любви к сельской местности и лошадям. Дома Джефф посмотрел на себя в зеркало и увидел унылую фигуру в костюме-тройке. В тот момент он понял, что не хочет больше быть банкиром и ему не суждено жить в Нью-Йорке. На следующее утро он уволился с работы.

Седер переехал в сельскую часть Пенсильвании и занимал самые разнообразные должности в текстильной промышленности и даже в спортивной медицине, прежде чем смог посвятить жизнь своей страсти — прогнозированию успеха скаковых лошадей. Цифры на скачках приблизительные. Из тысячи двухлеток, представленных на аукционе Окала — одном из самых престижных, — может быть, всего пять когда-нибудь смогут выиграть скачки со значительным призовым фондом. А что будет с остальными 995 лошадьми? Примерно треть окажется слишком медленной[9]. Еще треть получит травму — скорее всего, потому, что их ноги не смогут выдерживать огромное напряжение бешеной скачки (каждый год на американских ипподромах умирают сотни лошадей[10] — в основном из-за переломов ног[11]). Оставшаяся

* Старейшее и самое почетное студенческое братство в США. — *Прим. ред.*

треть будет страдать тем, что можно назвать синдромом Бартлби. Писарь из рассказа Германа Мелвилла, перестает работать и отвечает на каждое требование работодателя словами: «Я не хочу». Многие лошади в начале своей карьеры, видимо, приходят к выводу, что они не обязаны работать, если им не хочется. Поначалу они могут бежать быстро, но в какой-то момент просто замедляются или вообще останавливаются. Зачем изо всех сил бежать по краю овального поля, когда у вас ломит копыта и суставы? «Я предпочитаю не напрягаться», — решают они. (Я испытываю слабость к Бартлби — как к лошадям, так и к людям.)

Как владельцам выбрать выгодную лошадь при таком количестве шансов ошибиться? Люди всегда верили, что самый лучший способ предсказать, будет ли лошадь побеждать, — проанализировать ее родословную. Быть специалистом по выбору лошадей — значит уметь разобрать по косточкам все, что только возможно, об отце, матери, дедушках, бабушках, братьях и сестрах интересующей клиента лошади. Например, агенты сообщают, что «большой размер лошади естественен, потому что в ее роду по материнской линии было много рослых коней».

Но существует одна проблема. Конечно, родословная очень важна, однако она все же может объяснить лишь малую часть успеха спортивной лошади. Рассмотрим послужной список братьев и сестер всех обладателей наиболее престижной ежегодной награды — титула «Лошадь года». Все они имеют идентичные наилучшие родословные. Тем не менее более трех четвертей из них не выигрывали крупные скачки[12]. Традиционный способ прогнозирования успеха оставляет много возможностей для совершенствования.

На самом деле неудивительно, что родословная не дает достаточной информации для точного прогноза. Представьте, что так подбирали бы людей. Например, владелец клуба НБА решил купить игроков в свою команду, исходя из их родословных — когда они еще были десятилетними детьми. Он бы нанял агентов, приказав им изучить Ирвина Джонсона[13], сына «Мэджика» Джонсона. «У него сейчас хороший рост, — сказал бы эксперт. — Это естественный рост, унаследованный от Джонсона. Потому же мальчик должен иметь отличные зрение, самоотдачу и скорость. Он кажется общительным, у него хороший характер. Уверенная походка. Представительный. Это хороший вариант». К сожалению, 22 года спустя рост этого человека составил 185 см (слишком низкий для профессионального баскетболиста). И Ирвин Джонсон стал модным блогером! Он может оказать серьезную помощь в разработке дизайна формы, но вряд ли сможет сделать что-либо полезное на баскетбольной площадке.

Помимо фэшн-блогера, владелец клуба НБА, собравшийся набрать себе команду таким же образом, как многие выбирают лошадей, скорее всего купит Джеффри и Маркуса Джорданов — сыновей Майкла Джордана. В колледже оба они показали себя вполне заурядными игроками. А вот «Кливленд Кавальерс» удача улыбнулась. Эту команду ведет вперед Леброн Джеймс, рост мамы которого был всего 165 см[14]. Или представьте себе страну, которая избирала бы своих лидеров на основе их родословных. Нами бы руководили такие люди, как Джордж Буш-младший. (Извините, не удержался.)

Агенты, помогающие выбрать лошадей, ориентируются не только на родословную, но и на другую информацию.

Например, они анализируют аллюры двухлеток и внимательно рассматривают предлагаемых лошадей. В Окале я часами общался с различными экспертами и в результате понял, что у них нет единого, общего для всех критерия поиска.

Добавьте к этим противоречиям и неясностям то, что у некоторых покупателей, похоже, бездонные кошельки — и вы получите рынок с довольно малой эффективностью. 10 лет назад лошадь под № 153 была двухлеткой, бегавшей быстрее всех и, казалось, выглядевшей для большинства агентов просто потрясающе. К тому же она обладала замечательной родословной, будучи потомком Северной Танцовщицы и Секретариата — двух величайших скаковых лошадей всех времен. Ирландский миллиардер и шейх из Дубая захотели купить ее и вступили на торгах в битву, очень быстро превратившуюся в борьбу двух гордынь. Сотни любителей лошадей стали свидетелями того, как ставки поднимались все выше и выше, пока двухлетний конь наконец не был продан за 16 миллионов долларов — на сегодняшний день это самая высокая цена, когда-либо заплаченная за лошадь. Позже № 153, получившая имя Зеленая Мартышка[15], поучаствовала в трех скачках, заработала всего 10 000 долларов и была отправлена на покой.

Седер никогда не увлекался традиционными методами оценки лошадей. Его интересовали только данные. Он планировал измерять различные показатели скаковых лошадей, а затем смотреть, какие из них коррелируют с показанными в забегах результатами. Важно отметить, что Седер выработал свой план на полтора десятилетия раньше, чем была изобретена Всемирная

паутина, но его стратегия во многом базируется на научных данных, и уроки, извлеченные из его рассказа, может применить любой, кто работает с большими данными.

В течение многих лет попытки Седера не приносили ничего, кроме разочарования. Он измерял размер ноздрей лошадей, создав первый и самый большой в мире массив подобных данных и соответствующих им возможных доходов. Джефф обнаружил, что размер ноздрей не может указать на успех. Потом он делал лошадям ЭКГ, чтобы исследовать их сердце. Он отрезал ноги мертвым коням, чтобы измерить объем их быстро сокращающихся мышц. Однажды он даже взял лопату, чтобы определить количество экскрементов лошадей — исходя из теории, что слишком большой их объем перед соревнованиями может замедлить бег. Ничто не коррелировало с результатами на скачках.

А затем, 12 лет назад, произошел первый большой прорыв. Седер решил измерить размер внутренних органов лошадей. Поскольку при существовавшей тогда технологии это было невозможно, он построил свой собственный портативный аппарат УЗИ. Результаты оказались поразительными. Джефф обнаружил, что размер сердца, и в частности левого желудочка, был мощным прогностическим фактором успеха лошади, одной из самых важных переменных. Другой орган, имевший большое значение — селезенка: лошади с небольшой селезенкой практически не имели шансов завоевать приз.

Сделал Седер и еще пару важных наблюдений. Он оцифровал видео тысяч бегущих галопом лошадей и обнаружил, что определенные аллюры коррелируют с успехом на ипподроме. Он также заметил, что некоторые

двухлетки начинают хрипеть, пробежав всего одну восьмую мили. Таких лошадей иногда продают даже за миллион долларов, но данные Седера показали: подобные «хрипуны» практически никогда не добиваются успеха. Таким образом, Джефф приказал помощнику сидеть возле финиша и отсеивать «хрипунов».

Из примерно тысячи лошадей, выставленных на аукционе Окала, десяток справился со всеми тестами Седера. Он полностью игнорировал родословную — за исключением того, как это будет влиять на цену коня при продаже. «Родословная может сказать нам, что у лошади очень маленький шанс быть замечательной, — говорит он. — Но если я вижу, что конь великолепен, какая мне разница, у кого он родился?»

Однажды вечером Джефф пригласил меня в свой номер в отеле «Хилтон» в Окале. Там он рассказал мне о своих детстве, семье и карьере. Показал фотографии жены, дочери и сына. Сказал, что был одним из трех еврейских учеников, перешедших в старшие классы в школе в Филадельфии, и что по окончании школы его рост был 145 см (позже, в колледже, он вырос до 173 см). Рассказал о своей любимой лошади Pinky Pizwaanski. Седер купил и назвал ее в честь одного гея-жокея. Он чувствовал, что Pinky-конь всегда старался изо всех сил, даже если и не был самым успешным.

И наконец, Седер показал мне файл, в котором содержались все данные о коне № 85, — файл, ставший наиболее успешным прогнозом в его карьере. Он разглашал свой секрет? Возможно. Но Джефф сказал, что его это не волнует. Важнее сохранения секрета для него было доказать свою правоту, показать всему миру, что эти 20 лет

копания во внутренностях, выгребания навоза и таскания с собой аппарата УЗИ принесли наконец результат.

Вот некоторые сведения о лошади № 85.

№ 85 (позже Американский Фараон), однолетка

	Процентиль
Рост	56
Масса	61
Родословная	70
Левый желудочек	99,61

Здесь четко и ясно видно, почему Седер и его команда так одержимо рекомендовали № 85. Процентиль его левого желудочка составлял 99,61!

Не только левый желудочек, но и все остальные важные органы, включая сердце и селезенку, были исключительно крупными. Вообще говоря, Седер обнаружил: когда дело касается скачек, чем больше левый желудочек, тем лучше. Но его размер может быть и признаком болезни — если другие органы невелики. У Американского Фараона все наиболее важные органы были больше среднего размера, а левый желудочек был просто огромен. Данные кричали о том, что № 85 уникален, таких лошадей была одна на 100 тысяч или даже на миллион.

Какую информацию ученые могут извлечь из проекта Седера?

Первое и, пожалуй, самое главное. Если вы собираетесь попробовать использовать новые данные для

революционного улучшения ситуации, лучше сперва задаться вопросом: где не срабатывают старые методы? Одержимость агентов-лошадников родословными оставила Седеру достаточно места для маневра. То же самое можно сказать и о победе Google над поисковыми системами, одержимыми подсчетом слов.

Одним из недостатков в попытке Google предсказать приближение эпидемии гриппа[16], используя данные поисковых запросов, было то, что вы можете сделать это очень хорошо и сами — просто используя данные прошлой недели и добавив сезонные корректировки. До сих пор ведутся споры о том, насколько сведения, полученные на основании поисковых запросов, лучше простой, но мощной модели. На мой взгляд, поиск в Google практичнее для измерения состояний, для которых существующие данные не столь показательны. Поэтому Google STD в долгосрочной перспективе может оказаться более полезным, чем Google Flu.

Второй урок заключается в том, что при попытке сделать прогноз не нужно всерьез задаваться вопросом, почему ваша модель работает. Седер не может полностью объяснить, почему левый желудочек имеет столь важное значение при прогнозировании успеха лошади. Он также не в состоянии точно сказать, почему на успех влияет именно величина селезенки. Возможно, когда-нибудь лошадиные кардиологи и гематологи и дадут ответ на эти вопросы. Но сейчас это не важно. Седер занимается прогнозированием успеха, а не его объяснением. То есть вы просто должны знать, что это работает, и не пытаться понять почему.

Например, Walmart использует данные о продажах во всех своих магазинах, чтобы знать, какие продукты

следует пока отложить. До урагана Фрэнсис — разрушительного шторма, обрушившегося на юго-восток США в 2004 году, — компания Walmart подозревала (и совершенно справедливо), что, когда город переживет удар стихии, покупательские привычки людей могут измениться. Эксперты компании изучили данные по продажам после предыдущих ураганов, стараясь понять, что именно люди, возможно, захотят купить. Какой товар оказался самым популярным? Клубничное печенье. За несколько дней до урагана этот продукт продается в семь раз быстрее, чем обычно.

На основе проведенного анализа в супермаркеты вдоль 95-го шоссе (по пути урагана) поехали грузовики с клубничным печеньем «Поп-Тартс»[17]. И действительно: в эти дни оно продавалось особенно хорошо.

Почему печенье «Поп-Тартс»? Наверное, потому, что оно не требует охлаждения или приготовления.

Почему клубничное? Понятия не имею. Но когда проносятся ураганы, люди сметают клубничное печенье.

Поэтому теперь за несколько дней до очередного урагана Walmart обязательно увеличивает количество этого продукта на полках. Причина взаимосвязи урагана с клубничным вкусом не имеет значения. Важно само ее наличие. Возможно, однажды ученые-диетологи выяснят связь между ураганами и выпечкой с начинкой

из клубничного джема. Однако пока мы ждем объяснений, при приближении ураганов Walmart будет по-прежнему заполнять свои полки клубничным «Поп-Тартс» и приберегать рисовые хлебцы для солнечных дней.

Такой же вывод можно сделать и из истории экономиста из Принстона Орли Эшенфелтера. То, чем для Седера были лошади, для Эшенфелтера было вино.

Немногим более 10 лет назад Эшенфелтер испытывал сильное раздражение. Он покупал много красного вина из региона Бордо во Франции. Иногда оно было вкусным и достойным своей высокой цены, но неоднократно случалось так, что оно вызывало сильное разочарование.

Почему, спрашивал Эшенфелтер, он должен платить одну и ту же цену за вино, вкус которого так сильно разнится?

Однажды Орли получил совет от знакомого журналиста и знатока вин. Существует способ выяснить, будет ли вино хорошим. Ключевым моментом, сказал друг Эшенфелтера, является погода во время вегетации винограда.

Орли заинтересовался. И начал выяснять, правда это или нет и не может ли он всегда покупать самое лучшее вино. Он скачал данные о погоде в Бордо за 30 лет. Собрал аукционные цены на вина: аукционы, проходящие через много лет после первой продажи вина, показывают, каким оно на самом деле было.

Результат оказался просто удивительным. Действительно, по большей части, качество вина может быть объяснено погодой во время вегетации. Фактически же его можно определить с помощью простой формулы,

которую мы могли бы назвать первым законом виноградарства:

> Цена = 12,145 + 0,00117 зимних дождей + 0,0614 средний рост температуры за сезон — 0,00386 дожди во время сбора.

Так почему же качество вина в Бордо определяется таким образом? Чем объясняется первый закон виноградарства? Есть некое объяснение формулы хорошего вина Эшенфелтера: тепло и ранний полив необходимы для того, чтобы виноград правильно созревал. Однако точные сведения о его прогностической формуле выходят за рамки любой теории и, вероятно, никогда не будут поняты до конца даже специалистами в этой области.

Почему сантиметр зимних дождей добавляет в среднем 0,1 цента к цене бутылки полностью созревшего красного вина? Почему не 0,2 цента? Почему не 0,05? Никто не может ответить на эти вопросы. Но если зимой выпало 1000 сантиметров дополнительных осадков, вы должны быть готовы платить за бутылку вина 1 дополнительный доллар.

Как бы то ни было, несмотря на то, что Эшенфелтер не знал точно, почему его регрессия действует именно так, все же использовал ее для покупки вина. По его словам, «это отлично срабатывало»[18]. Качество вина, которое он пил с того времени, заметно улучшилось.

Если ваша цель предсказать будущее — какое вино будет иметь приятный вкус, какие продукты нужно будет продавать, какие лошади будут бежать быстрее других, — вам не нужно слишком сильно беспокоиться

о том, почему ваша модель работает так, как работает. Просто пользуйтесь. Это второй урок, который можно извлечь из истории Джеффа Седера.

Заключительный урок, который можно извлечь из удачной попытки Седера спрогнозировать потенциального победителя Тройной короны, — вы должны быть открытыми и гибкими в определении того, что именно следует считать данными. Именно этого не хватало экспертам, оценивавшим шансы лошадей до Седера. Они проверяли время бега и родословную. Гений Джеффа заключался в том, что он стал искать информацию там, куда другие до него не смотрели — то есть нетрадиционные источники данных. Если ученые сумеют взять на вооружение такой свежий и оригинальный взгляд, это обязательно окупится.

СЛОВА КАК ДАННЫЕ

Однажды в 2004 году два молодых экономиста с опытом работы в СМИ, Мэтт Генцкоу и Джесси Шапиро, бывшие тогда аспирантами в Гарварде, прочитали о недавнем решении суда в Массачусетсе легализовать однополые браки.

Парни обратили внимание на нечто интересное: две газеты использовали разительно отличающиеся выражения, описывая одно и то же событие. «Вашингтон Таймс», имеющая репутацию консервативной, озаглавила статью «Гомосексуальная “свадьба” в Массачусетсе». А «Вашингтон пост», считающаяся либеральной, сообщила о «победе однополых пар».

Неудивительно, что различные новостные источники могут склоняться к разным мнениям, что газеты могут

пересказать одну и ту же историю в разном ключе. В течение многих лет Генцкоу и Шапиро размышляли, могут ли они использовать свое экономическое образование для того, чтобы понять причины этой предвзятости СМИ. Почему некоторые из них кажутся более либеральными, а другие — более консервативными?

Но у парней не было никаких идей о том, как им решать эту задачу — они не могли понять, каким образом систематически и объективно измерять субъективность СМИ.

Интересным для Генцкоу и Шапиро в истории о гей-браках было не то, что газеты разошлись во взглядах — их заинтересовало, *как* именно разнилось освещение событий. Речь идет о заметном смещении акцентов при выборе слов. В 2004 году слово «гомосексуалисты», которое использовала «Вашингтон Таймс», было старомодным и унизительным способом описания геев. А вот термин «однополые пары», который употребила «Вашингтон пост», подчеркивает, что отношения геев — просто еще одна форма любви.

Ученые задались вопросом: не может ли язык быть ключом к пониманию необъективности. Возможно, либералы и консерваторы последовательно использовали разные выражения? И можно ли слова, употребляемые газетами при описании той или иной истории, превратить в данные? И что эти сведения могут рассказать об американской прессе? Могли бы мы определить по словам, является пресса либеральной или консервативной? И могли бы мы понять, почему? В 2004 году это были не праздные вопросы. Миллиарды слов в американских изданиях больше не попадали на газетную бумагу или микропленку. Некоторые сайты сейчас

записывают каждое слово из каждой статьи почти каждой газеты в США. Генцкоу и Шапиро могли бы прошерстить эти сайты и быстро протестировать, в какой степени язык может показать перекос газеты в ту или иную сторону. Эти тесты помогли бы им улучшить наше понимание принципов работы СМИ.

Но прежде чем описывать их находки, давайте оставим на минутку историю Генцкоу и Шапиро, а также их попытки количественно описать газетный язык, и обсудим, как ученые уже использовали этот новый тип данных — слова — для более глубокого понимания человеческой природы.

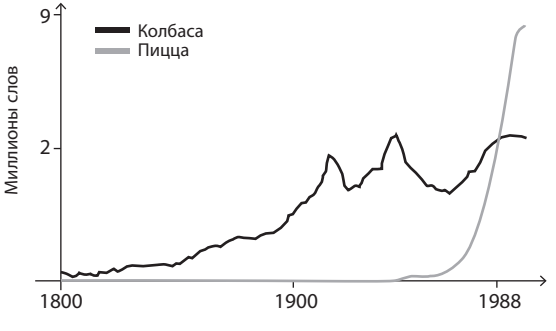
Конечно, язык всегда был предметом интереса социологов. Однако для его изучения, как правило, требуется внимательное чтение текстов. И превращение огромных кусков текста в данные раньше не представлялось возможным. Сейчас же, используя компьютеры и оцифровку, легко осуществить классификацию слов, взятых из огромного массива документов. Таким образом, язык стал предметом анализа больших данных. Ссылки, с которыми работает Google, также состоят из слов — равно как и поисковые запросы в Google, с которыми работаю я. Язык настолько важен в информационной революции, что заслуживает отдельного, посвященного только ему раздела книги. На самом деле сейчас он используется настолько широко, что появилось даже понятие «текст как данные».

Основной разработкой в этой области является Google Ngrams. Несколько лет назад два молодых биолога, Эрез Эйден и Жан-Батист Мишель, предложили своим помощникам одно за другим подсчитывать слова

в старых пыльных текстах — чтобы выяснить, как часто в них встречается та или иная лексика. Однажды Эйден и Мишель услышали о новом проекте компании Google по оцифровке книг со всего мира и почти сразу же сообразили: так в истории языка будет разобраться гораздо проще.

«Мы поняли, что наши методы безнадежно устарели, — рассказывал Эйден в интервью журналу «*Discover*». — Было понятно: конкурировать с этой всепобеждающей цифровой мощью невозможно». Поэтому они решили с ней сотрудничать. При помощи инженеров Google Эйден и Мишель создали сервис, осуществляющий поиск по определенному слову или фразе по миллионам оцифрованных книг. Потом приложение сообщает исследователям, как часто это слово или фраза появлялись ежегодно в период с 1800 по 2010 годы.

Так что же мы можем узнать по частоте, с которой слова или фразы появляются в книгах в разные годы? Прежде всего, о медленном росте популярности колбасы и относительно недавнем быстром росте популярности пиццы.



Но есть и гораздо более серьезные результаты. Например, Ngrams Google может показать, как формировалась наша национальная самобытность. Вот, скажем, увлекательный пример из книги Эйдена и Мишеля «Uncharted» («Неизведанное»).

Но сначала один вопрос. Как вы думаете, сегодня Соединенные Штаты — единая или разобщенная страна? Если вы принадлежите к большинству обычных людей, то скажете, что США сильно разобщены из-за высокого уровня политической поляризации. Можно даже сказать, что сегодня страна разобщена как никогда. Америка, в конце концов, теперь разделена по цвету: красные штаты — республиканские, синие — демократические. Но в книге «Uncharted» Эйдена и Мишеля есть один впечатляющий момент, демонстрирующий, насколько сильнее Соединенные Штаты были разобщены в прошлом. Об этом свидетельствуют слова, которые люди используют, говоря о своей стране.

Обратите внимание на слова, которые я использовал в предыдущем абзаце, говоря о разобщенности страны. Я писал: «США — разобщенная страна». Я говорил о США как о существительном в единственном числе. Это естественно, это правильная грамматика и стандартный вариант употребления слов. Уверен, вы этого даже не заметили.

Однако американцы далеко не всегда говорят подобным образом. На заре формирования Соединенных Штатов люди, упоминая свою страну, использовали множественное число. Например, Джон Адамс в докладе о положении дел в 1799 году говорил о «Соединенных Штатах и ИХ договорах с его британским Величеством». Если бы моя книга была написана в 1800 году, я бы сказал: «Соединенные

Штаты разобщены». Эта небольшая разница в использовании слов давно заинтересовала историков, поскольку предполагает существование момента, когда Америка перестала думать о себе как о совокупности штатов и начала думать о себе как о единой нации.

Так когда это произошло? Историки, как сообщает нам «Uncharted», никогда не знали этого точно, поскольку у них не было надежного способа прояснить ситуацию. Но многие уже давно подозревали, что это произошло во время Гражданской войны. Джеймс Макферсон, бывший президент американской исторической ассоциации и лауреат Пулитцеровской премии, отметил: «Война ознаменовала собой переход Соединенных Штатов из множественного числа к существительному единственного числа».

Но оказывается, что Макферсон был неправ. Google Ngrams обеспечил Эйдену и Мишелю надежный способ проверки. Они могли видеть, как часто в американских книгах употреблялись фразы «Соединенные Штаты являются…» и «США является…» — год за годом. Переход был достаточно постепенным и не ускорялся ни до Гражданской войны, ни после ее окончания.

Спустя 15 лет после Гражданской войны еще довольно часто писали «Соединенные Штаты являются…», а не «США является…», показывая, что страна лингвистически все еще была разделена. Военные победы опережали изменения в мышлении.

Это все об объединении страны. А как объединяются мужчина и женщина? Слова могут помочь и здесь.

Например, на основании того, о чем говорили конкретные мужчина и женщина во время первой встречи, мы можем предсказать, будет ли у них второе свидание.

Это продемонстрировала междисциплинарная команда Стэнфордского и Северо-Западного университетов — Дэниэл Макфарланд, Дэн Джуравски и Крейг Роулингс. Они общались с сотнями гетеросексуальных участников быстрых свиданий[19], пытаясь определить факторы, влияющие на возникновение контакта с партнером и желание пойти на вторую встречу с ним.

Сначала исследователи использовали традиционные данные. Они опросили участников быстрых свиданий, записав их рост, вес, увлечения, и проверили, насколько сильно эти факторы коррелируют с тем, с кем зафиксирована искра романтического интереса. В среднем женщины предпочитают мужчин выше себя ростом, разделяющих их увлечения; мужчины в среднем предпочитают более худощавых женщин, разделяющих их увлечения. Ничего нового.

Но ученые обнаружили и новую информацию. Они поручили участникам эксперимента взять с собой цифровые диктофоны. Таким образом удалось собрать все использовавшиеся в разговоре слова, выявить наличие смеха и вычленить тон голоса. Исследователи могли проверить, как

мужчины и женщины сигнализировали о своей заинтересованности и чем партнеры «зарабатывали» этот интерес.

Так о чем же говорят нам лингвистические данные? Во-первых, о том, как мужчина или женщина передает свою заинтересованность. Один из способов демонстрации того, что женщина привлекла мужчину, очевиден — он смеется над ее шутками. Еще один фактор, менее очевидный: в разговоре мужчина ограничивает диапазон оттенков голоса. Проводились исследования, показывающие, что монотонный голос часто воспринимается женщинами как мужской. Это означает, что мужчины, когда им нравится женщина, — возможно, подсознательно — преувеличивают свою мужественность.

А вот женщины сигнализируют о своей заинтересованности изменением диапазона оттенков голоса — они начинают говорить более мягко и более короткими фразами. Хорошей подсказкой о заинтересованности женщины являются используемые ею слова. Скажем, ей вряд ли нравится мужчина, если в ее речи встречаются слова и фразы уклонения от прямого ответа — такие, как «возможно» или «наверное».

Парни, если женщина начала подстраховываться высказываниями на любую тему — если ей «вроде бы» нравится ее напиток, или она «вроде как» зябнет, или «наверное» может поесть еще, — могу поручиться: она «вроде бы» «как бы» «наверняка» увлечена не вами.

Женщина *наверняка* заинтересована в вас, если она рассказывает о себе. Получается, если мужчине нравится женщина, самое прекрасное слово, которое он может услышать из ее уст — «я»: это знак того, что она чувствует себя комфортно. Помимо этого, женщина, скорее всего,

проявляет интерес, если использует самонаправленные фразы — такие как «Понимаете?», «Правда?» и «Я имею в виду». Почему? Ученые отметили, что эти фразы привлекают внимание слушателя. Они дружелюбные и теплые, они предполагают поддержание контакта с мужчиной — ну, вы понимаете, что я имею в виду?

Далее. Как мужчинам и женщинам следует общаться, чтобы заинтересовать партнера по свиданию? Статистика утверждает: у мужчин есть много способов говорить таким образом, чтобы увеличить свои шансы понравиться женщине. Дамам нравятся мужчины, которые соглашаются с ними. Поэтому неудивительно, что женщины скорее сочтут наметившийся контакт удачным, если мужчина смеется над их шутками и продолжает разговор на предложенные ими темы, а не постоянно меняет их, заводя разговор о том, о чем он хочет поговорить сам*. Женщинам также нравятся мужчины, выражающие им свои поддержку и сочувствие. Если мужчина говорит: «Это круто» или «Это потрясающе», женщины значительно чаще думают о возникшем контакте. Равно как и при использовании им таких фраз, как «Это тяжело» или «Тебе, должно быть, было грустно».

Для женщин есть плохие новости, поскольку статистика, кажется, подтверждает неприятную правду о мужчинах. Разговор играет лишь небольшую роль в их реакции

---

* Вот одна из теорий, над которыми я работаю: большие данные подтверждают все, что говорил покойный Леонард Коэн. Например, однажды он дал своему племяннику следующие советы по ухаживанию за женщинами[20]: «Слушай внимательно. Затем послушай еще немного. И когда ты решишь, что наслушался вдоволь, послушай еще». Похоже, именно это ученые и подтвердили своими исследованиями. — *Прим. авт.*

на женщин. При прогнозировании контакта со стороны мужчины внешность женщины перевесит все. Тем не менее есть одно слово, которое можно использовать, чтобы хоть немного повысить шансы на симпатию мужчины, и мы это уже обсуждали: «я». Мужчины более склонны заинтересовываться женщиной, которая рассказывает о себе. И, как отмечалось ранее, женщины также скорее готовы сообщить о возникшем интересе после свидания, где они рассказывали о себе. Таким образом, если на первом свидании пойдет предметный разговор о женщине, это очень серьезный знак. Дама свидетельствует о том, что ей комфортно в этих отношениях и она, похоже, ценит, что мужчина не перетянул на себя весь разговор. А джентльмену нравится, что женщина открылась ему навстречу. Так что второе свидание очень вероятно.

И наконец, в расшифровке записей свиданий был найден четкий индикатор проблем — знак вопроса. Если во время первого свидания было задано много вопросов, это практически исключает возможность второго — как со стороны мужчины, так и со стороны женщины. Это кажется нелогичным, ведь, кажется, вопросы — как раз признак интереса. Но не на первом свидании. На первом большое число вопросов — признак скуки. «Чем вы увлекаетесь?» «Сколько у вас братьев и сестер?» Так люди говорят, когда разговор глохнет. При этом многие удачные первые свидания могут включать в себя только один вопрос — в конце: «Ты встретишься со мной еще раз?» Если это единственный вопрос за всю встречу, скорее всего, ответ будет: «Да».

Мужчины и женщины говорят по-разному, не только когда пытаются завоевать друг друга. Они всегда говорят по-разному.

Команда психологов проанализировала слова, используемые в сотнях тысяч постов на Facebook[21]. Специалисты выясняли, как часто каждое слово употребляется как мужчинами, так и женщинами. В результате определились самые «мужские» и самые «женские» слова в английском языке.

Многие из этих слов, увы, были очевидны. Например, женщины говорят «покупки» и «мои волосы» гораздо чаще, чем мужчины. А последние говорят «футбол» и «Xbox» гораздо чаще, чем женщины. Чтобы утверждать то же самое, вам, наверное, не понадобилась бы команда психологов с их анализом больших данных.

Некоторые выводы, впрочем, оказались более интересными. Женщины используют слово «завтра» гораздо чаще мужчин — возможно, потому, что последние не настолько хорошо умеют загадывать наперед. Добавление буквы «о» к слову «so» (буквы «А» к слову «так») — одна из наиболее типичных женских лингвистических черт. Среди слов, которые непропорционально часто использовались женщинами, были «so», «sooo», «sooooo».

Может быть, тут дело в моем детском интересе к женщинам, которые не боятся случайно вырвавшихся ругательств, но я всегда думал, что представители обоих полов матерятся в равной степени. Но нет. В список слов, использующихся гораздо чаще мужчинами, чем женщинами, входят «черт», «трахает», «бред сивой кобылы», «лохи».

Здесь представлены облака слов, используемых в основном мужчинами, а затем тех, которые чаще всего употребляют женщины. Чем больше слово, тем чаще его используют представители соответствующего пола.

Мужчины

Женщины

Больше всего в этом исследовании мне нравится то, что новые данные предлагают нам выводы, которые существовали уже давно, но мы о них не знали. Мужчины и женщины всегда говорили по-разному. Но в течение сотен тысяч лет эта информация исчезала сразу же, как только звуки растворялись в пространстве. Теперь же она сохраняется в компьютерах и может быть проанализирована с помощью умных машин.

Возможно, учитывая мой пол, я должен был сказать: «Используемые слова, черт возьми, исчезают. Теперь мы можем отдохнуть от просмотра футбола и игры в Xbox и изучить это дерьмо. Ну, конечно, если будет не насрать на это».

Но не просто мужчины и женщины говорят по-разному. С возрастом люди тоже начинают использовать другие слова. Это может даже дать нам некоторые подсказки касательно процесса старения. Вот данные из того же исследования — слова, чаще всего используемые в Facebook людьми того или иного возраста. Я называю это распределение «Пить. Работать. Молиться». Подростки пьют. После 20 лет они работают. Когда им стукнет 30 и больше, они молятся.

Новый мощный инструмент для анализа текста иногда называют еще анализатором настроения. Теперь ученые в состоянии оценить, насколько счастливым или грустным является конкретный отрывок.

Как? Команда исследователей попросила большое число людей охарактеризовать десятки тысяч слов английского языка как положительные или отрицательные. Самыми положительными, согласно этой методике, стали «счастье», «любовь» и «круто». Наиболее негативными — «грустно», «смерть» и «депрессия». Таким образом на базе огромного набора слов был создан определитель настроения.

ПИТЬ. РАБОТАТЬ. МОЛИТЬСЯ

19–22 года

23–29 лет

30–65 лет

Используя его, можно измерить среднее настроение слов в текстовом отрывке. Если кто-то пишет: «Я счастлив, люблю и чувствую себя замечательно», анализатор отметит это как очень счастливый текст. Если кто-то пишет: «Мне грустно думать о смерти и депрессии», анализатор выдаст заключение, соответствующее очень грустному тексту. Остальные фрагменты будут располагаться где-то посередине.

Так что же можно узнать при помощи определителя настроения текста? Специалисты по анализу данных Facebook продемонстрировали одну замечательную возможность. Они могут оценивать валовое национальное счастье страны практически ежедневно. Если сообщения о своем состоянии люди склонны писать в позитивных тонах, страна в этот день считается счастливой. Если же тексты в основном будут негативными, день в стране явно не задался.

Одна из находок специалистов по анализу данных Facebook: Рождество — один из самых счастливых дней в году. Я был настроен скептически в отношении этого анализа, да и в целом в отношении всего проекта. Вообще, думаю, что многие люди тайно грустят в Рождество — потому что одиноки или поссорились со своей семьей. В целом я не склонен доверять обновленной информации Facebook в связи с нашей склонностью лгать онлайн о своей жизни (это мы обсудим в следующей главе).

Если вы одиноки и несчастны в Рождество, вам действительно захочется расстраивать всех друзей постом о том, как вы несчастны? Подозреваю, многие люди, проводящие безрадостное Рождество, публикуют в Facebook посты о том, как они благодарны за эту «замечательную, удивительную, поразительную, счастливую жизнь», тем самым повышая показатель валового счастья страны. Но если мы собираемся определить реальный уровень Валового Национального Счастья, следует использовать больше источников, чем просто обновления ленты в Facebook.

Заявление о том, что Рождество — это, в целом, радостное событие, будет похоже на правду. Обзор поисковых запросов в Google касательно депрессии и опросы Gallup также говорят о том, что Рождество является одним из самых счастливых дней в году. И, вопреки распространенному мифу, число самоубийств во время праздников снижается. Даже если в Рождество и встречаются грустные и одинокие люди, гораздо больше веселых и счастливых.

Сегодня, когда человек садится почитать, он большую часть времени проводит за внимательным изучением постов в Facebook. Но некогда, не так давно, люди читали

книги — и здесь анализ настроений может нам сообщить многое.

Команда ученых, возглавляемая Энди Рейганом из Калифорнийского университета и Школы информации в Беркли, скачала тексты тысяч книг и сценариев фильмов[22]. Затем исследователи определили, насколько счастливым или печальным был каждый фрагмент каждого текста.

Рассмотрим, например, книгу «Гарри Поттер и дары смерти». Ученые показали, как настроение повествования меняется вместе с описанием ключевых моментов сюжета.

Обратите внимание: многие взлеты и падения настроения, выявленные анализировавшей текст командой, соответствуют ключевым событиям.

Большинство историй имеют более простые структуры. Возьмем, например, трагедию Шекспира «Король Иоанн». В этой пьесе все идет гладко. Короля Иоанна Безземельного просят отказаться от престола. Он отлучен от церкви за неподчинение папе римскому. Вспыхивает

война. Его племянник умирает — возможно, в результате самоубийства. Другие люди умирают. И в конце умирает Иоанн, отравленный недовольным монахом.

А вот анализ настроений по ходу пьесы.

Другими словами, просто анализируя текст, компьютер смог показать, что события идут от плохого к худшему и к еще более худшему.

Или рассмотрим фильм «127 часов». Его основной сюжет выглядит следующим образом.

> Альпинист идет в поход по национальному парку Каньонлендс в штате Юта. Он знакомится с другими туристами, но затем расходится с ними. Внезапно он поскальзывается и сбивает непрочно стоявший камень, который зажимает его руку. Альпинист пытается различными способами освободиться, но каждый раз терпит неудачу. Он впадает

в отчаяние. Наконец он отрезает себе руку и убегает. Позже он женится, заводит семью, но продолжает ходить в горы — хотя теперь не забывает оставить записку, когда уходит.

А вот анализ настроений фильма, опять же, сделанный командой ученых Рейгана.

Так что же мы узнаем, изучив настроение тысяч подобных историй?

Специалисты по анализу данных обнаружили, что огромный процент историй вписывается в одну из шести относительно простых структур, обнаруженных командой Рейгана:

От нищеты к богатству (подъем)
От богатства к нищете (падение)
Человек в яме (падение, потом подъем)
Икар (подъем, потом падение)

Золушка (подъем, потом падение, потом подъем)
Эдип (падение, потом подъем, потом падение)

Возможны небольшие отклонения, не учитываемые простой схемой. Например, фильм «127 часов» относится к категории «Человек в яме», хотя есть моменты, когда эмоциональный фон временно улучшается. Но подавляющее большинство историй вписываются в одну из шести категорий. «Гарри Поттер и дары смерти» является исключением.

Нам еще нужно ответить на множество дополнительных вопросов. Например, как изменялась структура истории с течением времени? Становились ли с годами истории сложнее? Имеются ли культурные различия в типах историй? Какие типы историй люди любят больше всего[23]? Мужчин и женщин привлекают разные структуры историй или одинаковые? А как насчет людей из разных стран?

В конечном счете текст как данные может обеспечить нам беспрецедентное понимание того, что на самом деле хотят зрители. Это понимание может существенно отличаться от мнения писателей и создателей фильмов.

Рассмотрим исследование двух профессоров Уортонской школы — Ионы Бергера и Кэтрин Л. Милкмен. Они выясняли, какие типы историй привлекали людей больше всего, какие — позитивные или негативные — скорее попадут в список, которым делятся активнее всего по электронной почте (по данным «Нью-Йорк Таймс»). Исследователи скачивали каждую статью из «Нью-Йорк Таймс» в течение трех месяцев. Используя программу анализа настроений, профессоры расшифровывали

настроение статей. Скажем, «Премия «Тони» за меценатство» оказалась положительной историей. А вот «Слухи в интернете о самоубийстве корейской актрисы» и «Германия: умерла кормилица белого медвежонка» — что неудивительно — имели негативный характер.

Профессоры также фиксировали информацию о том, где именно каждая статья была размещена. На главной странице? Сверху справа? Сверху слева? Кроме того, они записывали информацию о времени выхода статьи. Поздно вечером во вторник? В понедельник утром?

Они могли сравнить две статьи (позитивную и негативную), оказавшиеся на сайте «Нью-Йорк Таймс» на одном и том же месте и вышедшие примерно в одно и то же время — чтобы посмотреть, какой из них люди будут активнее делиться по электронной почте.

Итак, какие статьи имеют больше откликов — позитивные или негативные?

Позитивные. Как заключают авторы исследования, «чем позитивнее контент, тем больше он имеет шансов быть распространенным в интернете».

Обратите внимание на этот неожиданный контраст с обычным журналистским представлением о том, что людей привлекают жестокие истории и рассказы о катастрофах. Действительно, СМИ вываливают на головы людей кучу мрачных статей. Пожалуй, нам есть что обсудить в редакционной поговорке: «Чем больше крови, тем сильнее притягивает». Однако исследование профессоров из Уортонской школы показывает, что на самом деле люди хотят видеть больше веселых историй. Они могут предложить новую поговорку: «Если что-то заставляет улыбаться, люди поделятся этим с другими».

Вот вам и вся правда о грустных и радостных текстах. Как бы вы могли определить, какие слова можно считать либеральными или консервативными? Что это говорит нам о современных СМИ? Это немного сложнее и возвращает нас к Генцкоу и Шапиро. Как вы помните, они были экономистами, заметившими, что браки геев по-разному описывались в двух разных газетах, и в этой связи задавшимися вопросом: не смогут ли они использовать язык для выявления политической предвзятости.

Первое, что сделали эти двое — проверили записи *стенограмм Конгресса*. Поскольку эти записи уже оцифрованы, ученые смогли скачать каждое слово, использованное в 2005 году каждым конгрессменом — как демократом, так и республиканцем. После чего они попробовали выяснить, какие фразы предпочитают использовать демократы, а какие — республиканцы.

И такие фразы действительно были. Вот несколько примеров в каждой категории.

Фразы, намного чаще используемые демократами	Фразы, наиболее часто используемые республиканцами
Налог на недвижимость	Налог на наследство
Приватизация системы социального обеспечения	Реформа системы социального обеспечения
Роза Паркс	Саддам Хусейн
Права работников	Права частной собственности
Бедняки	Государственные расходы

Что объясняют эти различия в лексике?

Иногда демократы и республиканцы используют разные формулировки для описания одного и того же понятия. В 2005 году республиканцы пытались сократить федеральный налог на наследство. Они, как правило, называют его «налогом на смерть» (это звучит как поборы с недавно усопших). Демократы же обозначили его как «налог на недвижимость» (что выглядит как налог на богатых). Аналогичным образом, республиканцы пытались превратить социальное страхование в индивидуальные пенсионные счета. Для них это была «реформа». Для демократов же это звучало более угрожающе — «приватизация».

Иногда различия в языке — это вопрос расстановки акцентов. Наверняка и республиканцы, и демократы с большим уважением относятся к Розе Паркс, герою борьбы за гражданские права. Но демократы чаще упоминают ее имя. Кроме того, обе партии считают Саддама Хусейна, бывшего президента Ирака, злым диктатором. Но республиканцы гораздо чаще упоминали его в своих попытках оправдать войну в Ираке. Аналогично, борьба за «права трудящихся» и забота о «бедняках» являются основополагающими принципами Демократической партии. «Право частной собственности» и урезание «госрасходов» — основные принципы республиканцев.

И эти различия в использовании лексики весьма существенны. Например, в 2005 году республиканцы в Конгрессе использовали фразу «налог на смерть» 365 раз, а «налог на недвижимость» — всего 46. У демократов картина оказалась прямо противоположной: 35 фраз «налог на смерть» и 195 — «налог на недвижимость».

Если эти слова могут сказать нам, является ли конгрессмен демократом или республиканцем, то ученые поняли, что их можно использовать и для определения газет правого или левого толка. Консервативные газеты делают на своих страницах примерно то же самое, что и республиканцы в конгрессе: последние предпочитают употреблять выражение «налог на смерть» — для убеждения людей противодействовать ему. Например, относительно либеральная «Вашингтон пост» использовала выражение «налог на недвижимость» в 13,7 раз чаще, чем словосочетание «налог на смерть». А более консервативная «Вашингтон Таймс» употребила фразу «налог на смерть» и «налог на имущество» примерно в одинаковых пропорциях.

Благодаря чудесам интернета Генцкоу и Шапиро смогли проанализировать лексику большинства национальных газет. Ученые использовали два вебсайта — newslibrary.com и proquest.com, — где имеется оцифрованный архив 433 газет. Затем они подсчитали, как часто там употреблялась тысяча подобных политически заряженных выражений — для определения политической ориентации самих СМИ. Самой либеральной по этому показателю оказалась «*Philadelphia Daily News*». А самой консервативной — «*Billings* (Montana) *Gazette*».

Когда у вас появляются первые обстоятельные мерила пристрастий такого широкого спектра СМИ, вы, пожалуй, можете ответить на самый важный вопрос о прессе: почему одни публикации демонстрируют сдвиг влево, а другие — вправо[24]?

Экономисты быстро сосредоточили свое внимание на одном ключевом факторе: политических настроениях

в том или ином регионе. Если он в целом либеральный — как Филадельфия и Детройт, — доминирующая газета, скорее всего, будет либеральной. Если же он более консервативен — как Биллингс и Амарилло, штат Техас, — основная часть газет там будет консервативной. Иными словами, факты убедительно свидетельствуют: газеты склонны давать своим читателям то, чего те хотят.

Вы можете сказать, что владелец газеты имеет влияние на направление взглядов своего издания. Но нет. Как правило, на политическую направленность газеты он влияет меньше, чем мы могли бы предположить. Обратите внимание на то, что происходит, когда один и тот же человек или компания владеет газетами на различных рынках. Рассмотрим компанию «Нью-Йорк Таймс». Генцкоу и Шапиро обнаружили, что она владеет как либеральной «Нью-Йорк Таймс» в Нью-Йорке, где около 70% населения являются демократами, так и (на момент исследования) консервативной «*Spartanburg Herald-Journal*» в Спартанбурге, Южная Каролина, где около 70% населения — республиканцы. Конечно, есть и исключения: новостная корпорация Руперта Мердока[25] владеет всеми признанной консервативной газетой «Нью-Йорк пост». Но в целом полученные данные свидетельствуют о том, что рынок определяет направленность газет в гораздо большей степени, чем воля хозяев.

Исследование имеет огромное влияние на наше представление о новостных СМИ. Многие люди, особенно марксисты, рассматривали американскую журналистику как нечто, управляемое кучкой богатых людей или корпораций с целью воздействия на массы, для того, чтобы подтолкнуть людей к определенным политическим

взглядам. Однако в статье Генцкоу и Шапиро показано: это не основная мотивация владельцев газет. Они, прежде всего, стремятся дать массам то, чего те хотят — чтобы владельцы газет могли стать еще богаче.

Да, есть же еще один вопрос — важный, спорный и, возможно, еще более провокационный. Куда, в среднем, больше склоняются американские СМИ — влево или вправо? Другими словами, СМИ в Америке скорее либеральные или консервативные?

Генцкоу и Шапиро обнаружили, что в основном газеты имеют левый уклон. Средняя газета, по используемым в ней словам, больше похожа на конгрессмена-демократа, чем на конгрессмена-республиканца.

«Ага! — могут завопить консервативно настроенные читатели. — Я же говорил!» Многие консерваторы давно подозревали, что газеты, пытаясь манипулировать массами, пишут предвзято — чтобы поддержать левые взгляды.

Нет, это неверно, отвечают авторы статьи. На самом деле либеральный уклон в газетах хорошо откалиброван и заточен на то, что читатели хотят увидеть. Последние, в среднем, имеют небольшой уклон влево. (У исследователей есть данные об этом). И газеты, в среднем, также имеют небольшой уклон влево — чтобы подать своим читателям ту точку зрения, которую они желают видеть.

В этом нет никакого великого заговора. Есть только капитализм.

Новостные СМИ, по данным Генцкоу и Шапиро, часто действуют как любая другая отрасль на планете. Точно так же, как супермаркеты выясняют, какое мороженое люди предпочитают, и заполняют им свои полки, газеты

выясняют, какие оценки люди хотят видеть, и заполняют ими свои страницы. «Это просто бизнес», — сказал мне Шапиро[26]. Вот что вы можете узнать, когда разберете на составные части и количественно оцените такие мудреные явления, как новости, анализ и мнения.

ИЗОБРАЖЕНИЯ КАК ДАННЫЕ

Традиционно когда ученые или бизнесмены хотели собрать информацию, они проводили исследования. Данные аккуратно формировались на основе чисел или флажков в окошках опросников. Сейчас все иначе. Дни структурированных, чистых и простых полученных в результате исследований данных закончились. Идя по жизни сегодня, мы повсюду оставляем свои грязные следы, которые и становятся основным источником информации.

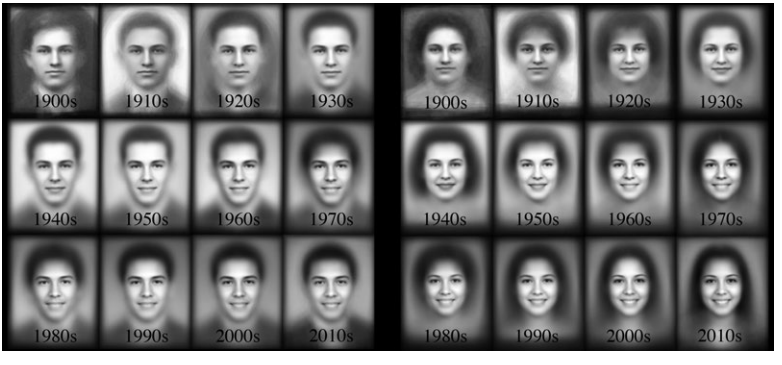
Как мы уже видели, слова — это данные. Клики — это данные. Ссылки — это данные. Опечатки — это данные. Бананы во сне — данные. Тон голоса — данные. Хрипы — данные. Сердечный ритм — данные. Размер селезенки — данные. Я утверждаю, что поисковые запросы — наиболее разоблачительные данные.

Оказывается, фотографии — тоже данные.

Так же, как слова, собранные в книгах или в периодике и хранившиеся на пыльных полках, фотографии были вытащены из фотоальбомов и картонных коробок и оцифрованы. Они тоже были превращены в биты и байты и запущены в облако. Как текст может преподать нам урок — продемонстрировав, например, как менялась манера людей излагать свои мысли, — так и фотографии

могут показать нам историю США — например, изменением способов позирования перед камерой.

Я считаю гениальным исследование группы четырех ученых-компьютерщиков в университетах Браун и Беркли. Они воспользовались достижением цифровой эпохи: многие вузы отсканировали ежегодные фотографии выпускников[27] и сделали их доступными онлайн. В интернете исследователи нашли 949 ежегодников с фотографиями учеников американских средних школ за период 1905–2013 годов. Это собрание включало десятки тысяч снимков. Используя компьютерные программы, ученые смогли создать по фотографиям «обычное» лицо каждого десятилетия. Другими словами, они смогли выяснить среднее расположение и конфигурацию носа, глаз, губ, волос. Здесь представлены типичные лица более чем за век — с разбивкой по полу:



Ничего не замечаете? Американцы — и особенно женщины — стали улыбаться. В начале XX века они фотографировались почти с каменным выражением, а в конце — просто сияли.

Так в чем причина? Разве американцы стали счастливее?

Нет. Ответить на этот вопрос помогли другие ученые. Причина — по крайней мере, на мой взгляд — просто удивительна. На заре фотографии люди относились к ней, как к живописи[28]. Они не могли сравнить этот процесс ни с чем другим. Таким образом, фотосюжеты были скопированы с сюжетов картин. А поскольку люди, позирующие художнику для картины, не могли сохранять улыбку в течение долгих часов, они принимали серьезный вид. Люди, которых снимал фотограф, также делали серьезные лица.

Что же в итоге заставило их поменять выражение лица? Бизнес, прибыль, маркетинг, конечно. В середине XX века Kodak — компания, продававшая пленку и камеры — была расстроена тем, что люди делают не слишком много снимков. Тогда была разработана стратегия приучения людей снимать все больше и больше. Рекламные кампании Kodak стали ассоциировать фотографии с понятием счастья. Их целью было заставить людей приобрести привычку делать фото всякий раз, когда они хотели показать другим, что в их жизни произошло нечто хорошее. Так что улыбки в ежегодниках являются результатом этой успешной кампании (как и большинство фотографий, которые вы видите в Facebook и в Instagram сегодня).

Но данные, полученные на основе фотографий, могут рассказать нам гораздо больше, чем обозначение периода времени, когда старшеклассники начали говорить: «Сы-ы-ыр». Удивительно, но снимки в состоянии поведать нам даже о положении дел в экономике.

Познакомьтесь с одной провокационно озаглавленной научной работой: «Измерение экономического роста из космоса». Когда у документа название вроде такого, можете держать пари: я обязательно прочитаю его. Авторы работы — Вернон Хендерсон, Адам Соригард и Дэвид Н. Вайль — начали с замечания, что во многих развивающихся странах существующие измерения валового внутреннего продукта (ВВП) являются неэффективными. Это происходит потому, что значительная часть экономической активности остается незафиксированной в бухгалтерских книгах, а ресурсы правительственных учреждений, в задачу которых входит измерение производительности экономики, довольно ограничены.

Авторы озвучили неординарную мысль. Они говорят, что могли бы измерять ВВП, исходя из того, насколько светло в этих странах ночью[29] — получив эту информацию из фотографий со спутника ВВС США, который делает полный оборот вокруг Земли 14 раз в сутки.

Почему свет в ночное время может быть хорошим показателем ВВП? В очень бедных частях мира людям трудно заплатить за электричество. В результате при плохих экономических условиях в домах резко уменьшалось количество света, которое жители позволяли себе включать ночью.

В Индонезии в результате Азиатского финансового кризиса 1998 года количество света ночью резко снизилось. А в Южной Корее в период с 1992 по 2008 год объем ночного освещения увеличился на 72%, что соответствовало исключительно высоким экономическим показателям в тот период. В Северной Корее в то же время количество освещения ночью уменьшилось, что

соответствовало удручающим экономическим показателям.

В 1998 году в южной части Мадагаскара было обнаружено большое скопление рубинов и сапфиров. Городишко Илакака из практически стоянки для грузовиков превратился в крупный торговый центр. До 1998 года там почти не было ночного освещения, а после обнаружения месторождения за пять лет количество света в ночное время взрывообразно увеличилось.

Авторы признают, что их оценка экономической активности по уровню ночного освещения далека от идеала. Вы не можете точно оценить состояние экономики только по тому, сколько света улавливает спутник. Авторы не рекомендуют использовать этот показатель для развитых стран, где существующие экономические данные дают более точную картину. И, по правде говоря, они обнаружили, что даже в развивающихся странах оценка количества света ночью столь же бесполезна, как и официальные показатели. Но сочетание ущербных правительственных данных и несовершенных показателей ночного освещения дает более точный результат, чем может обеспечить лишь один источник. Другими словами, с помощью снимков, сделанных из космоса, вы можете просто улучшить свое понимание уровня развития экономик той или иной страны.

Джозеф Райзингер, доктор наук в области информатики, разделяет разочарование авторов идеи о ночном освещении в отношении существующих баз данных с информацией об экономиках развивающихся стран. В апреле 2014 года Райзингер отметил, что Нигерия обновила информацию об объеме ВВП с учетом новых секторов,

которые тамошние чиновники, возможно, пропустили при опубликовании предыдущих оценок. По их нынешним оценкам, ВВП Нигерии сейчас на 90% выше[30].

«Это крупнейшая экономика в Африке, — сказал Райзингер[31], и его голос постепенно начал набирать силу. — Мы даже не знаем, какие основные параметры мы хотели бы знать об этой стране».

Он хотел найти способ более четко оценивать различные экономические показатели. Его решение — это отличный пример того, как можно переосмыслить данные и какова их реальная ценность.

Райзингер основал компанию «Premise», в которой работают группы сотрудников из развивающихся стран, вооруженные смартфонами. В чем заключается их работа? Фотографировать интересные происшествия, которые могут иметь какое-либо экономическое значение.

Сотрудники, вооружившись смартфонами, могут делать снимки АЗС или фруктовых корзин в супермаркетах. Они фотографируют одни и те же места снова и снова. Фотографии отправляются в головной офис компании, где вторая группа сотрудников — компьютерщики — превращают фотографии в информацию.

Специалисты компании могут проанализировать все — от длины очередей на заправках до того, сколько яблок лежит в корзине в супермаркете, и до цены этих яблок. На основе самых разных фотографий любых видов деятельности компания может начать оценивать уровень экономической активности и инфляции. В развивающихся странах длинные очереди на АЗС — основной индикатор экономических проблем. Равно как недозрелые яблоки и их отсутствие. Снимки, сделанные в Китае,

помогли обнаружить продовольственную инфляцию в 2011 году и продовольственную дефляцию в 2012 году — задолго до появления официальных данных.

«Premise» продает эту информацию банкам или хедж-фондам, а также сотрудничает со Всемирным банком.

Как и многие хорошие идеи, «Premise» продолжает приносить пользу. Недавно Всемирный банк заинтересовался размерами теневой экономики на Филиппинах, связанной с сигаретами. В частности, он хотел знать последствия недавних шагов правительства, включавших случайные рейды против производителей сигарет, не уплачивающих налоги. Что придумала компания «Premise»? Фотографировать табачные киоски на улице. Посмотрим, на скольких из них имеются акцизные марки, которые отличают законные сигареты. Было обнаружено, что эта часть теневой экономики, бывшая достаточно обширной в 2015 году, в 2016-м стала значительно меньше. Усилия правительства принесли результат, хотя для того, чтобы понять объем скрытого товара (нелегальных сигарет), требуются новые данные.

Как мы видели, эпоха цифровых технологий принесла совершенно новое понимание того, что считать данными, и из новой информации было сделано много интересных выводов. Знание причин, заставляющих СМИ смещать тональность своих выступлений влево или вправо, обеспечивающих успех первого свидания и возможность выявления хорошо развивающейся экономики — это только начало.

Неслучайно на основе этих новых данных было сделано немало денег — начиная с десятков миллиардов

господ Брина и Пейджа. Джозеф Райзингер также работает не в убыток себе. Обозреватели подсчитали, что годовой доход «Premise» сегодня составляет десятки миллионов долларов. Недавно инвесторы влили в компанию еще 50 миллионов[32]. Это означает, что некоторые из них считают «Premise» одним из самых выгодных предприятий в мире — в первую очередь, в области создания и использования фотографий. То есть в той же лиге, что и «Playboy».

Другими словами, новые типы данных имеют огромное значение как для ученых, так и для предпринимателей. При этом понятие данных в последнее время значительно расширилось. Сегодня специалисты не должны ограничивать себя узким или традиционным представлением о них. В наши дни фотографии очередей в супермаркетах — ценные данные. Наполнение полок там же — данные. Спелость яблок — данные. Фотографии из космоса — данные. Кривизна линии губ — тоже данные. Любая информация!

И все эти новые сведения мы наконец можем увидеть даже сквозь прикрывающую их ложь.

Глава 4

ЦИФРОВАЯ СЫВОРОТКА ПРАВДЫ

Все врут. О том, сколько выпили по дороге домой. О том, как часто ходят в тренажерный зал, сколько стоят эти новые туфли, будут ли читать эту книгу. Они говорят, что больны, когда вполне здоровы. Они говорят, что будут на связи, когда не собираются этого делать. Они утверждают, что говорят не о вас, хотя именно вас они и обсуждали. Они говорят, что любят вас, хотя на самом деле это не так. Они говорят, что счастливы, хотя в действительности хандрят. Они говорят, что им нравятся женщины, тогда как предпочитают мужчин. Люди врут друзьям. Боссам. Детям. Родителям. Они обманывают врачей и мужей. Лгут женам. Они врут сами себе.

И они — я совершенно в этом уверен — врут во время опросов.

Вот вам краткий обзор:

Вы когда-нибудь жульничали на экзамене? _____
Вы когда-нибудь мечтали кого-нибудь убить? _____

Вам когда-нибудь хотелось соврать? Многие люди при опросах занижают количество случаев неловкого

поведения и дурных мыслей. Они хотят хорошо выглядеть, хотя большинство опросов анонимны. Это называется «социально приемлемое смещение».

Одна серьезная статья 1950 года[1] представила веские доказательства того, как опросы могут пасть жертвой этого явления. Исследователи собрали из официальных источников данные о жителях Денвера: сколько процентов из них голосовали, давали деньги на благотворительность и имеют читательский билет в библиотеке. Затем они сами опросили денверцев — чтобы увидеть, насколько эти показатели совпадают с реальностью. Результаты оказались шокирующими. То, что люди сообщали в анкетах, сильно отличалось от сведений, собранных учеными. Хотя никто не подписывал анкету, все в основном преувеличивали свой регистрационный статус избирателя, стремление голосовать и участие в благотворительности.

	Сообщено в ходе опроса, %	Официальные данные, %
Регистрационный статус избирателя	83	69
Голосовал на последних президентских выборах	73	61
Голосовал на последних выборах мэра	63	36
Имеет читательский билет в библиотеке	20	13
Давал деньги на благотворительность	67	33

Изменилось ли что-либо за 65 лет? В век интернета отсутствие библиотечного читательского билета никого

больше не смущает. Но несмотря на изменение представлений о неудобном или нежелательном, стремление людей обманывать социологов остается весьма сильным.

Во время недавнего исследования выпускникам университета Мэриленда задавали различные вопросы об их жизни во время учебы[2]. Ответы сопоставлялись с официальными отчетами. Люди постоянно давали неверную информацию, что позволяло им выглядеть лучше, чем они были на самом деле. Меньше 2% опрошенных сообщили, что закончили обучение со средним баллом ниже 2,5 (в действительности таких было около 11%). А 44% заявили, что в прошлом году они сделали пожертвование университету (в действительности таких было около 28%).

Вполне возможно, именно ложь сыграла роль в провале опроса[3], не сумевшего предсказать победу Дональда Трампа в 2016 году. Исследования в среднем недооценивали его поддержку примерно на два процентных пункта. Некоторые люди могли постесняться сказать, что собираются голосовать за него. Другие, возможно, утверждали, что затрудняются ответить, в то время как все время были за Трампа.

Зачем люди дают ложную информацию в ходе анонимных опросов? Я спросил об этом Роджера Туранго, почетного профессора университета штата Мичиган и, возможно, лучшего специалиста в мире по отклонению от социально желательного поведения. «Наша слабость в отношении «белой лжи» является важной частью проблемы, — пояснил он. — В реальной жизни люди врут примерно в трети случаев. Эта привычка переносится и на опросы»[4].

То есть существует странная привычка — иногда мы лжем сами себе. «Имеет место нежелание признаться самому себе, что, мол, ты облажался как студент», — говорит Туранго.

Привычка лгать самому себе может объяснить, почему так много людей утверждают, будто их результат выше среднего[5]. Насколько велика эта проблема?

Более 40% инженеров одной компании заявили, что входят в 5% лучших работников.

Более 90% преподавателей колледжей говорят, что уровень их квалификации выше среднего. Четверть старшеклассников считают, что они входят в 1% лучших учеников по умению ладить с другими людьми. Если вы не честны с самим собой, то наверняка не будете честны и при опросе.

Еще один фактор, способствующий вранью при опросах — сильное желание произвести хорошее впечатление на незнакомца, проводящего собеседование (если опрос проводится в устной форме). Как добавляет Туранго, «приходит человек, который выглядит как ваша любимая тетя… Вы готовы рассказать вашей любимой тете, как в прошлом месяце курили марихуану?»*

* Еще одной причиной лжи[6] является стремление просто напакостить исследователям. Хотя в обычной жизни это не такая уж большая проблема, она огромна для любых опросов, связанных с подростками — и принципиально усложняет нам возможность разобраться в поведении этой возрастной группы.

Вы готовы признать, что не дали денег на свою старую добрую альма-матер?

Именно поэтому чем более обезличенными будут условия получения информации, тем больше честных ответов вы получите. Для этого лучше проводить интернет-опросы, а не телефонные, которые, в свою очередь лучше проводимых интервьюерами. Люди в большей степени готовы давать правдивые ответы, находясь в одиночестве, чем если рядом с ними в комнате есть кто-то еще.

Однако при проведении опроса на деликатные темы каждый метод обследования будет приводить к значительному искажению информации. В этом случае Турангo использует слово, часто применяемое экономистами: «стимул». У людей нет стимула говорить правду.

Так как же мы можем узнать, что в действительности они думают и делают?

В некоторых случаях имеются официальные источники данных, из которых мы можем узнать правду. Например, даже если люди лгут о своих благотворительных пожертвованиях, можно получить реальные цифры

Первоначально ученые обнаружили взаимосвязь: приемные дети (подростки) чаще ведут себя асоциально — в частности принимают наркотики, выпивают, прогуливают школу. Но последующие исследования показали, что эту взаимосвязь полностью сочинили 19% подростков, назвавших себя приемными, не будучи таковыми. Дальнейший анализ опросов выявил, что значительный процент подростков сообщил во время исследования, что их рост — более семи футов (213,5 см), а вес — более 400 фунтов (181,5 кг). В ходе другого опроса выяснилось, что 99% студентов, сообщивших о том, что у них установлены протезы, просто решили подшутить над учеными-исследователями. — *Прим. авт.*

по регионам от самих благотворительных организаций. Но когда мы пытаемся узнать о поведении, которое не заносится в официальные отчеты, или хотим выяснить, что люди думают — то есть их истинные убеждения, чувства и желания, — нет никакого другого источника информации, кроме того, что люди соизволят сказать в ходе опроса. До сих пор так и было.

И это второй важный аспект больших данных: некоторые интернет-ресурсы, служа своего рода цифровой сывороткой правды, заставляют людей признать то, что они не признают больше нигде. Вспомните о поисковых запросах в Google. Помните условия, делающие людей более честными. Онлайн? Есть. В одиночестве? Да. Нет человека, осуществляющего опрос? Все верно.

Есть еще одно огромное преимущество поиска в Google, заставляющее людей говорить правду — стимулы. Если вам нравятся расистские шутки, у вас нет причин поделиться этим фактом в ходе опроса. Тем не менее для поиска лучших новых расистских шуток онлайн они есть. Если вы считаете, что страдаете от депрессии, у вас нет мотива признать это во время опроса. Но он у вас есть, когда вы начинаете узнавать в Google о симптомах и возможных методах лечения.

Даже если вы врете самому себе, Google все же может узнать правду. За пару дней до выборов вы и некоторые из ваших соседей можете считать, что обязательно пойдете на избирательный участок и проголосуете. Но если ни вы, ни они не искали информацию о том, как и где голосовать, специалисты по поиску и обработке данных вроде меня могут сказать: явка в вашем районе будет низкой. Точно так же, возможно, вы еще не признались самому себе, что

страдаете от депрессии — даже когда делаете в Google запросы о неконтролируемом плаче и о том, что с трудом вылезаете из постели. Это будет ясно по поисковым запросам о депрессии, о которых я говорил ранее.

Проанализируйте свой опыт общения с Google. Полагаю, время от времени вы вводите в поле запроса нечто такое, что выдает поведение или мысли, в которых вы постеснялись бы признаться в приличном обществе. В самом деле, очевидно, что подавляющее большинство американцев сообщают Google некоторые очень личные сведения. Например, они ищут «порно» чаще, чем «погода»[7]. Это, кстати, мало согласуется с результатами одного опроса, в котором только около 25% мужчин и 8% женщин признаются в просмотре порнографии[8].

Вы могли бы отметить определенную честность в результатах поиска Google, глядя на то, как поисковик автоматически пытается выполнить ваши запросы. Его предложения основаны на наиболее распространенных поисковых запросах, сделанных другими людьми. Таким образом, автоматический подбор информации по запросам показывает нам, что люди ищут в Google. На самом деле автозаполнение может немного вводить в заблуждение. Google не вставляет определенные слова, которые он считает неуместными[9], вроде мата и «порно». Это означает, что, согласно автозаполнению, мысли человека, выраженные в Google, менее скабрезны, чем на самом деле. Но даже несмотря на это, все-таки порой происходят довольно пикантные недоразумения.

Если вы введете в строку поискового запроса «почему…», то первые два варианта в Google, в настоящее время автоматически завершающие вопросы — «Почему

небо голубое?» и «Почему существует високосный год?» Система предполагает, что это две наиболее распространенные цели поиска. Третий вариант: «Почему мои какашки зеленые?» И это автозаполнение в Google может вызвать у вас беспокойство. Сегодня, если вы введете в поисковую строку: «Нормально ли — хотеть…», первое предложение автозаполнения — «убить». Если вы введете: «Нормально ли — хотеть убить…», то первый вариант автозаполнения — «мою семью».

Хотите больше доказательств того, что поиск Google может показать другую картину мира, отличную от той, которую мы обычно видим? Рассмотрим поисковые запросы, связанные с сожалениями относительно решения иметь или не иметь детей. Прежде чем принять это решение, некоторые пары опасаются сделать неправильный выбор, и почти всегда вопрос ставится именно так: будут ли они жалеть, что у них нет детей. Люди в семь раз чаще спрашивают у Google, будут ли они жалеть об отсутствии детей, чем о наличии.

После принятии решения, родить (усыновить/удочерить) ребенка или нет, люди иногда признаются в том, что раскаиваются в сделанном выборе. Это может показаться шокирующим, но после принятия решения цифры меняются местами. Взрослые с детьми в 3,6 раза чаще сообщают Google, что жалеют о своем решении, чем взрослые без детей[10].

Есть один нюанс, который следует иметь в виду, читая эту главу: Google может показывать смещение в сторону неблаговидных мыслей, которые люди наверняка не могут обсуждать с кем-либо еще. При этом, если мы пытаемся раскрыть эти мысли, возможности Google могут

оказаться весьма полезны. И большое различие между сожалениями о решении иметь и не иметь детей может многое сообщить нам об этих неуместных мыслях.

Давайте на мгновение остановимся и задумаемся, что на самом деле означает поисковый запрос вроде: «Я сожалею, что завел(-а) детей». Компания Google позиционирует себя как источник, непосредственно из которого мы можем черпать информацию на такие темы, как погода, кто выиграл вчерашний матч или когда была возведена статуя Свободы. Но иногда мы печатаем в Google свои мысли без оглядки на цензуру, без особой надежды на то, что эта система даст нам прямой ответ. В этом случае поисковое окно служит своего рода исповедником.

Ежегодно фиксируются тысячи поисковых запросов вроде «Я ненавижу холодную погоду», «Меня раздражают люди» и «Мне грустно». Конечно, эти тысячи «мне грустно» представляют собой лишь малую часть из сотен миллионов людей, которым в этом году было тоскливо. Мои исследования показали: запросы, выражающие мысли вместо поиска определенной информации, принадлежат лишь небольшой части тех, кому эти идеи приходят на ум. Аналогично мои исследования показывают, что тысячи американцев, ежегодно вводящих в поисковик фразу «Я сожалею, что у меня есть дети», представляют собой лишь небольшую выборку из тех, кому в голову пришла эта мысль.

Дети, безусловно, огромная радость для многих — скорее всего, для большинства. И несмотря на страх моей мамы, что «ты и твой глупый анализ данных» может ограничить количество ее внуков, это исследование

не изменило моего желания иметь детей. Но это неприглядное сожаление интересно — это другой аспект человечества, о котором мы вряд ли могли бы узнать на основе традиционных данных. Наша культура постоянно заваливает нас изображениями прекрасных, счастливых семей. Большинство людей никогда даже не предполагали, что могут пожалеть о наличии детей. Но некоторые жалеют. И они не могут признаться в этом никому — кроме Google.

ПРАВДА О СЕКСЕ

Сколько геев среди американских мужчин? Это легендарный вопрос из исследований сексуальности. Тем не менее он остается одним из самых сложных для социологов, на него ищут ответ очень долго. Психологи уже давно не верят известной оценке Альфреда Кинси, опирающейся на опросы выборки заключенных и проституток, согласно которой 10% американских мужчин являются гомосексуалистами. Сейчас данные репрезентативных опросов сообщают о 2–3%. Однако сексуальные предпочтения прочно входят в число тех вопросов, отвечая на которые люди склонны лгать. Думаю, что могу использовать большие данные для самого лучшего ответа на этот вопрос изо всех когда-либо полученных.

Во-первых, надо более подробно остановиться на результатах опросов. Последние утверждают, что геев гораздо больше в толерантных штатах, чем в нетолерантных. Например, по данным исследования Gallup, доля

геев почти в два раза выше в Род-Айленде, штате с наибольшей поддержкой гей-браков, чем в Миссисипи, штате с самой низкой поддержкой гей-браков[11].

У этого есть две возможные причины. Например, геи рождаются в нетолерантных штатах и затем могут переехать в более толерантные. Кроме того, геи в нетолерантных штатах не готовы сообщать о своей ориентации — они скорее солгут при ответе на этот вопрос.

Некоторую информацию, связанную с переездами геев, можно почерпнуть из другого источника больших данных — Facebook, который позволяет пользователям указать, представители какого пола их интересуют. Около 2,5% мужчин-пользователей этой социальной сети интересуются мужчинами. Это примерно соответствует результатам исследований. И Facebook тоже показывает большие различия в проценте геев в штатах с высокой и низкой толерантностью: согласно Facebook, процент геев в Род-Айленде в два раза выше, чем в Миссисипи.

Facebook также может предоставить информацию о перемещениях людей. Я сумел распознать место постоянного проживания части пользователей Facebook, являющихся открытыми геями. Это позволило сразу оценить, сколько геев уехало из нетолерантных штатов в более терпимые районы страны. Ответ? Явно присутствует некоторая подвижность. Например, геи переезжали из Оклахома-Сити в Сан-Франциско. Но, по моему мнению, мужчины, упаковывающие свои CD с Джуди Гарленд и отправляющиеся в места с более свободными нравами, могут составлять не более половины разницы

числа гей-населения в толерантных и нетолерантных штатах*.

Кроме того, Facebook позволяет нам сосредоточить внимание на старшеклассниках. Это особая группа, поскольку школьникам редко удается выбирать место жительства. Если мобильность объясняет различие количества открытых геев в разных штатах, то эта разница не должна проявляться среди школьников — пользователей Facebook. Так что же нам говорят данные о последних? В нетолерантных штатах намного меньше открытых геев-школьников. Только две тысячи ребят — учащихся средней школы в Миссисипи являются открытыми геями. Так что мобильность — это еще не все.

Если одинаковое число мужчин-геев, родившихся в каждом штате, и мобильность не могут полностью объяснить, почему в некоторых регионах так много открытых геев, то очень большую роль должна играть скрытность. Что возвращает нас к Google, с которым люди оказались готовы поделиться очень многим.

Есть ли способ использовать поисковые запросы в области порно для того, чтобы проверить, сколько геев *на самом деле* имеется в каждом штате? Действительно есть. По моим оценкам, по всей стране — используя

---

* Некоторые могут считать оскорбительным то, что я ассоциирую увлечение Джуди Гарленд с предпочтением заниматься сексом с мужчинами — даже в шутку. И я, конечно, не имею в виду, что все геи — или даже их большинство — увлечены дивами. Но поисковые данные показывают: в этом стереотипе что-то есть. По моим оценкам, человек, ищущий информацию о Джуди Гарленд, в три раза чаще ищет и гей-порно, чем обычное. Большие данные говорят нам, что некоторые стереотипы все же справедливы. — *Прим. авт.*

данные поисковых запросов Google и Google AdWords — гей-порно составляет около 5% от поисковых запросов порно, сделанных мужчинами[12]. (Это относится к поиску таких терминов, как «Rocket Tube» — популярный порнографический гей-сайт, а также «порно с геями»).

А как это число варьируется в разных частях страны? В целом в толерантных штатах выполняется больше поисков гей-порно, чем в нетолерантных. Это имеет смысл, учитывая, что некоторые геи переехали из нетолерантных мест в более терпимые. Но отличия не столь велики, как у данных, полученных на базе опросов или Facebook. По моим оценкам, в Миссисипи 4,8% порнографических поисковых запросов связаны с гей-порно — это гораздо больше цифр, выявленных в исследованиях или в Facebook, и достаточно близко к 5,2% аналогичных запросов, сделанных в Род-Айленде.

Так сколько же американских мужчин являются гомосексуалистами? Цифра, полученная на базе поисковых запросов о мужской порнографии — примерно 5%, — кажется в достаточной степени истинной оценкой величины гей-сообщества в США. Есть еще один, менее очевидный, способ получить это число. Он требует некоторых научных данных. Мы могли бы использовать взаимосвязь между толерантностью и величиной открытого гей-населения. Потерпите минутку.

Мои предварительные исследования показывают: в рассматриваемом штате каждые 20 процентных пунктов поддержки гей-браков означают увеличение в этом штате количества мужчин, открыто идентифицирующих себя в качестве геев на Facebook, почти в полтора раза. Исходя из этого, мы можем оценить, сколько мужчин,

родившихся в гипотетически полностью толерантном штате (где 100% жителей поддерживает однополые браки), стали бы открытыми геями. По моим оценкам, это около 5% — что соответствует данным, полученным по базе поисковых порно-запросов. Ближайший пример мы могли бы найти в полностью толерантной среде школы для мальчиков в районе Калифорнийского залива. Около 4% из них — открытые геи на Facebook. Что, похоже, полностью соответствует моим подсчетам.

Должен отметить, что я еще не смог придумать метод оценки однополого влечения для женщин. Здесь данные поисковых порнографических запросов не менее полезны. Однако порнографию смотрит гораздо меньшее число женщин, что делает выборку менее представительной. А среди тех, кто все же смотрит ее — включая даже женщин, которых в реальной жизни преимущественно привлекают мужчины, — наибольшей популярностью пользуется лесбийская. В целом лишь 20% женщин, смотрящих видео на PornHub, лесбиянки.

5% геев среди американских мужчин — это, конечно, оценочная величина. Некоторые люди бисексуальны, другие — особенно молодежь — не определились в своей ориентации. Очевидно, вы не сможете определить это настолько же точно, насколько возможно подсчитать число голосующих или посещающих кинотеатры.

Но одно из следствий моей оценки совершенно очевидно: очень многие в США — особенно в нетолерантных штатах — до сих пор скрывают свои сексуальные предпочтения. Они не раскрывают их в Facebook. И не признаются в них во время опросов. И во многих случаях даже могут быть женаты на женщинах.

Получается, жены достаточно часто справедливо подозревают своих мужей в том, что они геи. Эти мысли проявляются в удивительно частом автозавершении: «Мой муж… гей». Слово «гей» — на 10% более вероятное завершение поискового запроса, начинающегося словами «Мой муж…», чем слово, занимающее второе место — «обманщик». Кроме того, оно встречается в восемь раз чаще, чем «алкоголик», и в 10 раз чаще, чем «страдает депрессией».

Возможно, очень показательно, что запросы, связанные с сексуальностью мужа, гораздо более распространены в наименее толерантных регионах. Штаты с самым высоким процентом женщин, задающих подобные вопросы — Южная Каролина и Луизиана. На самом деле в 21 из 25 штатов, где этот вопрос задается наиболее часто, поддержка однополых браков ниже, чем в среднем по стране.

Google и порносайты — не единственные полезные ресурсы информации, когда речь идет о мужской сексуальности. У больших данных существует еще больше доказательств того, что многие живут, скрываясь. Я проанализировал объявления на Craigslist в интернете — мужчины ищут «отношения на одну ночь» с мужчинами. В менее толерантных к геям штатах процент таких объявлений, как правило, выше. Наименее терпимыми являются Кентукки, Луизиана и Алабама.

Но вернемся к поиску данных в Google. Один из самых распространенных поисковых запросов, появляющихся непосредственно перед или после «гей-порно» — «тест на гомосексуальность». (Эти тесты осмеливаются определять, является ли мужчина гомосексуалистом или

нет.) В наименее толерантных штатах поиск «гей-теста» примерно в два раза выше, чем в среднем по стране.

Что же означает такое метание между «гей-порно» и «гей-тестом»? Предположительно это свидетельствует о довольно сильном — если даже не мучительном — смущении. Разумно предположить желание некоторых мужчин убедиться в том, что их интерес к гей-порно совсем не означает склонности к гомосексуализму.

Данные поиска в Google не позволяют нам увидеть статистику конкретного человека в течение некоторого времени. Однако в 2006 году AOL передала ученым выборку запросов своих пользователей. Вот, например, некоторые данные одного анонимного пользователя в течение шести дней.

Пятница 03:49:55	Бесплатные гей фото (!)
пятница, 03:59:37	Раздевалка гей фото
пятница 04:00:14	Гей фото
пятница 04:00:35	Гей секс фото
пятница 05:08:23	Проверка на гомосексуальность
пятница 05:10:00	Тест на полноценного гея
пятница 05:25:07	Гей-тесты для сомневающегося мужчины
пятница 05:26:38	Тесты на гомосексуальность
пятница 05:27:22	Тесты «я — гей?»
пятница 05:29:18	Гей фото

пятница 05:30:01	Фото обнаженных мужчин
пятница 05:32:27	Бесплатные фото обнаженных мужчин
пятница 05:38:19	Фото горячий гей-секс
пятница 05:41:34	Горячий гомосексуальный анальный секс
среда, 13:37:37	Тесты «я — гей?»
среда 13:41:20	Тесты на гомосексуальность
среда 13:47:49	Горячий гомосексуальный анальный секс
среда, 13:50:31	Бесплатное гей видео (!)

Из этого, безусловно, видно: указанный мужчина чувствует себя некомфортно со своей сексуальностью. Данные Google говорят нам, что подобных мужчин немало, и большинство из них, по правде говоря, живут в штатах, менее терпимых к однополым отношениям.

Чтобы поближе взглянуть на людей, стоящих за этими цифрами, я попросил врача-психиатра, практикующего в штате Миссисипи и специализирующегося на оказании помощи скрытым геям, узнать, не согласится ли кто-либо из его пациентов поговорить со мной. Один человек выразил готовность. Он сказал мне, что является профессором в отставке, ему около шестидесяти и он женат на одной женщине более 40 лет.

Около 10 лет назад, когда его захлестнул очень сильный стресс, он обратился к психологу и наконец узнал

о своей сексуальной ориентации. Он сказал, что всегда знал о своем влечении к мужчинам, но думал, что это общее свойство всех людей, о котором не принято говорить. Вскоре после начала терапии у него произошла первая и единственная гомосексуальная связь с одним из его студентов, которому было около 30 лет. Этот опыт он описывает как «потрясающий».

У них с женой нет сексуальных отношений. Он говорит, что чувствовал бы себя виноватым, если бы развелся или начал открыто встречаться с мужчиной. При этом сожалеет практически обо всех важнейших решениях в своей жизни.

Профессора в отставке и его жену ждет очередная ночь без любви и без секса. Несмотря на огромный прогресс, сохранение нетерпимости к геям приводит к тому, что миллионам других американцев приходится жить так же.

Вас, возможно, не шокирует тот факт, что 5% мужчин являются геями, а другие просто скрывают свою ориентацию. Но были времена, когда большинство людей были бы огорошены, узнав об этом. Да и сейчас еще встречаются места, где большинство людей подобная информация ошеломляет.

«В Иране нет гомосексуалов[13], в отличие от вашей страны, — настаивал в 2007 году президент Ирана Махмуд Ахмадинежад. — В Иране нет ничего подобного». Мэр Сочи Анатолий Пахомов, незадолго до того, как в его городе состоялись зимние Олимпийские игры 2014 года, сообщил: «У нас в городе нет геев»[14]. Но интернет-поведение показывает значительный интерес к гей-порно как в Сочи, так и в Иране[15].

В этой связи возникает очевидный вопрос: существуют ли в настоящее время в Соединенных Штатах какие-либо общие сексуальные интересы, которые до сих пор считаются шокирующими? Это зависит от того, что именно вы считаете общим и насколько легко вас привести в смятение.

Самые популярные поисковые запросы на PornHub включают такие слова, как «подросток», «секс втроем» и «минет» — для мужчин, а также фразы вроде «страстная любовь», «облизывание сосков» «мужчина лижет киску» — для женщин.

На основании информации, полученной от PornHub, мы узнали о некоторых фетишах — иначе мы бы даже не догадались об их существовании. Существуют женщины, ищущие «анальные яблоки» и «секс с мягкими игрушками». Существуют мужчины, ищущие «фетиш — сопли» и «распятие обнаженных». Но эти поисковые запросы редки — даже на этом огромном порносайте их всего около 10 в месяц.

При изучении данных PornHub становится совершенно очевидным еще один момент: там есть кто угодно для кого угодно. Женщины, что неудивительно, часто ищут «высоких» парней, «смуглых», «красавчиков». Но они также иногда посматривают на «низких» мужчин, «бледнокожих» и даже на «страшненьких». Встречаются женщины, интересующиеся «мужчинами-инвалидами», «пухлыми парнями с маленьким членом» и «жирными уродливыми старикашками». Мужчины часто ищут «худощавых» женщин, с «большими сиськами» и «блондинок». Но они также иногда посматривают на «толстых», женщин с «маленькой грудью» и с «зелеными волосами».

Встречаются мужчины, интересующиеся «лысыми женщинами», «карлицами», а также «женщинами без сосков». Эти данные могут порадовать тех, кто не высок, не темноволос, некрасив или не является худощавой блондинкой с большим бюстом*.

Как насчет других достаточно частых, но при этом довольно удивительных поисковых запросов? Среди 150 самых распространенных у мужчин, наиболее неожиданными для меня являются кровосмесительные — те, о которых я говорил в главе, посвященной Фрейду. Другие запросы, о которых мужчины обычно не говорят — «shemale» («транссексуал с женскими чертами лица и мужскими гениталиями»; 77-й по количеству запросов) и «бабушка» (110-е место в списке наиболее частых поисков). В целом около 1,4% поисковых запросов мужчин на PornHub относится к поиску женщин с пенисами. Примерно 0,6% (0,4% относится к мужчинам в возрасте до 34 лет) ищут изображения пожилых людей. Только 1 из 24 тысяч мужских запросов на PornHub упоминает детей — это может быть связано с тем, что PornHub, по понятным причинам, запрещает любые формы детской порнографии и владение подобной информацией является незаконным.

Одним из наиболее частых женских запросов является жанр порнографии, который — я предупреждаю вас — смутит и обеспокоит многих читателей: секс с насилием

---

* Думаю, эти данные также оказывают влияние на оптимальную стратегию знакомства. Ясно, что не стоит особенно беспокоиться и принимать отказ близко к сердцу. Постепенно вы найдете единомышленников, которых привлекают люди вроде вас. Как бы вы ни выглядели, существуют мужчины и женщины, которым нравится именно такой типаж. Верьте мне. — *Прим. авт.*

в отношении женщин. По меньшей мере 25% дам ищут порно, связанное с болью и/или унижением женщины — например «болезненный анальный секс с рыданиями», «публичное унижение» и «экстремально жестокая групповуха». Пять процентов ищут насильственный секс — «изнасилование» или «секс по принуждению», — хотя эти видео запрещены на PornHub. И количество подобных поисковых запросов у женщин как минимум в два раза больше, чем у мужчин. Если имеется порно, в котором насилие совершалось в отношении женщины, мой анализ данных показывает: почти всегда к нему обращается намного больше женщин, чем мужчин.

Конечно, когда пытаешься разобраться в этом, важно помнить, что между фантазией и реальной жизнью имеется большая разница. Да, среди посещающего PornHub женского меньшинства есть подмножество тех, кто ищет — пусть и безуспешно — сцены изнасилования. Естественно, это не значит, что такие женщины хотят быть изнасилованными в реальной жизни, и, безусловно, это не делает изнасилование менее ужасным преступлением. Данные, полученные на основании порно-запросов, говорят нам лишь о том, что иногда люди фантазируют о чем-то, чего не хотели бы пережить в реальной жизни и ни в коем случае не готовы совершить в отношении других.

Шкафы, в которых скрываются секреты, являются не только хранилищами фантазий. Когда дело доходит до секса[16], в них обнаруживается очень многое — например сведения о количестве секса в жизни людей.

Во введении я отмечал: в ходе опроса американцы сообщили о том, что они используют гораздо больше презервативов, чем продается ежегодно.

На этом основании вы могли бы решить: они чаще просто говорят, что пользуются презервативами во время секса, чем на самом деле делают это. Но исследования также свидетельствуют, что люди еще и преувеличивают частоту своего секса. Около 11% женщин[17] в возрасте от 15 до 44 лет говорят, что они сексуально активны, не беременны на данный момент и не используют средства контрацепции. Даже при относительно консервативных предположениях о том, сколько раз они занимаются сексом, ученые предположили бы, что каждый месяц 10% из них должны беременеть[18]. Но это было бы больше, чем общее число беременностей в США (1 из 113 женщин детородного возраста[19]). В нашей сексодержимой культуре, вероятно, трудно признать, что у вас просто нет столько секса, сколько вы говорите.

Но если вы ищете понимания или совета, у вас, опять же, есть стимул обратиться к Google. Там в 16 раз больше жалоб на то, что супруг (–а) не хочет секса, чем на то, что он (она) не желает разговаривать. Там в 5,5 раз больше сетований на то, что неженатый партнер (партнерша) не хочет секса, чем на то, что он (она) не отвечает на СМС.

Поисковые запросы в Google позволяют высказать предположение относительно неожиданной причины многих подобных отношений без секса. В Google в два раза больше жалоб на то, что парень не хочет заниматься сексом, чем на то, что этого не хочет девушка. Более того, самый частый вид подобных претензий девушек в поисковых запросах имеет вид: «Мой парень не хочет заниматься со мной сексом». (Поисковые запросы в Google не имеют разбивки по полу, но, как показывает предыдущий анализ, 95% мужчин являются гетеросексуалами,

поэтому мы можем предположить, что от мужчин исходит не слишком большое число запросов, включающих выражение «мой парень»).

Как это следует интерпретировать? Означает ли это, что парни воздерживаются от секса больше, чем девушки? Не обязательно. Как уже упоминалось ранее, поисковые запросы в Google могут искажаться, когда речь заходит о том, о чем люди стесняются говорить открыто. Мужчина более спокойно говорит друзьям об отсутствии сексуального интереса у его девушки, чем женщина подругам об отсутствии интереса у ее парня. Тем не менее, даже если данные Google и не означают, что парни действительно не хотят заниматься сексом вдвое чаще девушек, они все равно свидетельствуют о том, что парни избегают половых контактов чаще, чем готовы признаться.

Данные Google также указывают на причину того, почему люди могут так часто избегать секса: из-за сильной тревоги. Причем, по большей части, как раз причины этой тревоги люди и скрывают. Начнем с беспокойств, обуревающих мужчин. Давно известно, что они переживают по поводу уровня своего дохода, но степень этого беспокойства сильнее, чем принято считать.

Мужчины задают вопросы Google о своем половом органе чаще, чем о любой другой части тела — о легких, печени, ногах, ушах, носе, горле и мозге вместе взятых. Мужчины делают в Google больше запросов относительно того, как увеличить пенис, чем о том, как настроить гитару, сделать омлет или поменять колесо. Самая сильная мужская озабоченность по поводу стероидов, согласно их поискам в Google, связана не с потенциальным нанесением вреда здоровью. Нет, людей больше беспокоит,

не уменьшит ли прием стероидов размер их пениса. Когда мужчины задают вопросы в Google о возрастных изменениях, их больше волнует не изменения тела или мозга, а опасение, что их член уменьшится.

Примечание: один из наиболее распространенных вопросов о мужских гениталиях в Google — «Насколько большой у меня пенис?». То, что мужчины обращаются с подобным вопросом к поисковику, а не к обычной линейке, на мой взгляд, является типичным проявлением цифровой эпохи*.

Волнует ли женщин размер пениса? По данным запросов в Google, редко. На каждый женский поиск относительно размера пениса партнера приходится около 170 запросов мужчин о собственном половом органе. Правда, в тех редких случаях, когда женщины выражают обеспокоенность по поводу размера пениса партнера, они вовсе не обязательно имеют в виду, что он слишком маленький — более 40% жалоб связано с тем, что он слишком большой. «Боль» — самое частое первое слово в поисковых запросах Google, используемое вместе с фразой «…во время секса». («Кровотечение», «мочеиспускание», «плач» и «пукание» — вот пятерка наиболее часто употребляемых первых слов.) Однако только 1% мужчин, осуществляющих поисковые запросы в Google относительно размера своего пениса, ищут информацию о том, как сделать его меньше.

---

* Я хотел назвать эту книгу «Насколько большой у меня пенис? Что говорят поисковые запросы в Google о природе человека», но редактор предупредил, что с таким названием книга будет плохо продаваться — люди могут слишком смущаться, думая о покупке книги с таким названием в книжном магазине аэропорта. А вы что думаете? — *Прим. авт.*

Второй наиболее распространенный вопрос мужчин о сексе — как сделать так, чтобы он продолжался дольше. Опять же, эта неуверенность не совпадает с заботами женщин. Количество поисковых запросов о том, как добиться более быстрой кульминации у партнера и как замедлить ее — примерно одинаково. На самом деле, когда женщины озабочиваются мужским оргазмом, чаще всего их интересует не то, когда он происходит, а то, что он вообще не наступает.

Когда речь идет о мужчинах, мы нечасто говорим о проблемах, связанных с внешним видом их тела. И хотя вопросы внешности в основном, беспокоят женщин, перекос не столь велик, как можно предположить на основании сложившихся стереотипов. На основании моего анализа, проведенного с помощью Google AdWords (оценивалось количество посещаемых людьми сайтов), я определил: среди проявляющих интерес к красоте и здоровью — 42% мужчин, к похудению — 33% мужчин, к пластической хирургии — 39% мужчин. Среди всех запросов, начинающихся со слова «Как» и связанных с грудью, около 20% — мужские вопросы о том, как избавиться от груди.

Но даже если число мужчин, которым не хватает уверенности в отношении своего тела, больше, чем считается, женщины по-прежнему опережают их, когда дело доходит до тревоги о том, как они выглядят. Так что же можно узнать с помощью цифровой сыворотки правды о женском беспокойстве о своем внешнем виде? Каждый год в Соединенных Штатах более семи миллионов поисковых запросов связано с имплантами молочной железы. А официальная статистика сообщает: ежегодно осуществляет эту процедуру около 300 тысяч женщин.

Дамы также демонстрируют серьезную тревогу относительно вида своей попы, хотя многие из них недавно поменяли свое представление о том, что именно им в себе не нравится.

В 2004 году в некоторых областях США самым распространенным касательно изменения размеров попки был вопрос о том, как сделать ее меньше. А вот желающие увеличить ее в подавляющем большинстве были сосредоточены в районах с преобладанием чернокожего населения. Однако начиная с 2010 года желание иметь зад побольше распространилось и на остальную территорию страны. За последующие четыре года этот интерес утроился. В 2014 году уже во всех штатах было больше запросов о том, как увеличить ягодицы, чем как уменьшить их. Сегодня в США на каждые пять поисков об имплантах груди приходится один, связанный с увеличением размера ягодиц. (Спасибо, Ким Кардашьян!)

Соответствует ли женское предпочтение больших ягодиц мужскому? Что интересно, да. Поиск «Порно большие задницы», ранее в основном поступавший из черных регионов, в последнее время стал популярен на всей территории Соединенных Штатов.

Что еще хотят мужчины от женского тела? Как уже упоминалось ранее — и как большинство сочтет абсолютно очевидным, — мужчины предпочитают большую грудь. Примерно 12% поисковых запросов относительно порно включают слова «большая грудь» — это почти в 20 раз больше, чем число запросов порно с женщинами с маленькой грудью.

И все-таки это не свидетельствует о желании мужчин видеть женщин с грудными имплантами. Около 3% тех,

кто ищет порно женщин с большой грудью, прямо заявляют о том, что они хотят видеть натуральную.

Поисковые запросы в Google о своей жене и имплантах молочной железы равномерно распределяются между темами «Как уговорить ее вставить импланты» и недоумением «Почему она хочет их вставить».

Или рассмотрим самые распространенные поисковые запросы о груди девушки: «Обожаю сиськи моей подружки». Непонятно, что люди надеются найти в Google при подобной формулировке.

Женщины, как и мужчины, задают много вопросов по поводу своих гениталий. На самом деле они почти так же живо интересуются своими влагалищами, как и мужчины — своими пенисами. Женское внимание к своей вагине часто связано со здоровьем. Но по крайней мере 30% вопросов приходятся на совсем другие проблемы. Дамы хотят знать, как побрить промежность, сжимать ее и сделать ее вкус лучше. Поразительно, но, как отмечалось ранее, наиболее распространена озабоченность тем, как улучшить свой запах.

Женщины нередко обеспокоены тем, что их влагалище пахнет рыбой, затем уксусом, луком, аммиаком, чесноком, сыром, запахом тела, мочи, хлеба, отбеливателя, кала, пота, металла, ног, мусора и гнилого мяса.

В целом мужчины по поводу гениталий партнерши Google не особенно беспокоят. Они примерно столько же интересуются влагалищем, сколько женщины — пенисом.

Когда мужчины при поиске упоминают влагалище партнерши, они обычно жалуются на то, что женщин беспокоит больше всего — запах. В основном люди

пытаются выяснить, как сказать даме о плохом запахе, не задев ее чувств. Иногда, тем не менее, вопросы мужчин о запахе раскрывают их собственную неуверенность в себе. Обычно они задают вопрос о способах использовать запах для обнаружения измены — например, если ощущается запах презервативов или семени другого мужчины.

Что мы можем сделать со всеми этими скрытыми комплексами? Есть хорошие новости. Google дает нам законные основания поменьше беспокоиться. Многие из наших самых глубоких страхов о том, как нас воспринимают сексуальные партнеры, совершенно необоснованны.

Сидя в одиночестве у своего компьютера, люди не видят смысла врать и раскрывают самые сокровенные мысли, не довольствуясь поверхностными суждениями.

На самом деле мы все так заняты изучением собственных тел, что у нас остается слишком мало сил на то, чтобы судить о телах других людей.

Вполне вероятно, существует связь между двумя самыми актуальными проблемами, выявленными с помощью поисковых запросов в Google и связанными с сексом — отсутствием секса и неуверенностью в собственной сексуальной привлекательности и эффективности. Возможно, они связаны между собой. Может

быть, если бы мы меньше беспокоились о сексе, у нас его было бы больше.

Что еще можно выяснить о сексе на основании поисковых запросов в Google? Мы в состоянии сказать о соперничестве полов, увидеть, кто самый сильный, энергичный и щедрый. Можно выбрать все запросы, в которых выясняются способы лучшего орального секса с противоположным полом[20]. Кто чаще ищет информацию — мужчины или женщины? Кто более сексуально щедр — мужчины или женщины? Дамы, понятное дело. Суммируя все, я оцениваю соотношение как 2:1 в пользу женщин, ищущих советы о том, как лучше заниматься оральным сексом со своим партнером.

Когда мужчины ищут аналогичную информацию, их чаще всего не интересует способ угодить другому человеку. Они в основном интересуются тем, как получить оральный секс, а не тем, как обеспечить женщине оргазм (это один из моих излюбленных фактов, основанный на поисковых запросах в Google).

ПРАВДА О НЕНАВИСТИ И ПРЕДУБЕЖДЕНИИ

Секс и романтические отношения вряд ли являются единственными темами, окутанными туманом стыда — и, следовательно, люди предпочитают умалчивать не только о них. Многие — вполне разумно — предпочитают не распространяться о своих предрассудках. Полагаю, вы могли бы назвать прогрессом тот факт, что сегодня люди чувствуют: их будут осуждать за готовность судить о других на основе этнической принадлежности, сексуальной

ориентации или религии. Но многие американцы поступают так до сих пор. (Это еще один раздел, в котором я должен предупредить читателей о том, что он включает в себя материалы, способные обеспокоить и смутить их.)

В Google легко можно найти вопросы пользователей вроде «Почему чернокожие такие невоспитанные грубияны?» или «Почему евреи — это зло?». Ниже приведены в порядке убывания частоты пять слов с негативной коннотацией, используемых при поиске информации о различных этнических группах[21].

	1	2	3	4	5
Афроамериканцы	Хамы	Расисты	Дураки	Уроды	Ленивые
Евреи	Зло	Расисты	Уроды	Жадные	Корыстолюбивые
Мусульмане	Зло	Террористы	Злодеи	Жестокие	Опасные
Мексиканцы	Расисты	Дураки	Уроды	Ленивые	Тупые
Азиаты	Уроды	Расисты	Раздражают	Глупые	Жадные
Геи	Зло	Ошибка природы	Дураки	Раздражают	Самовлюбленные
Христиане	Глупые	Психи	Тупые	Помешанные	Ошибка природы

Среди этих стереотипов можно выделить некоторые признаки. Например, только с афроамериканцами

связан стереотип «хамы». Почти каждую группу считают «дураками», за исключением двух — евреев и мусульман. Стереотипно считают «злом» евреев, мусульман и геев — но не чернокожих, мексиканцев, азиатов и христиан.

Только мусульмане стереотипно ассоциируются с террористами. Когда американец-мусульманин слышит это определение в свой адрес, ответ может быть мгновенным и жестоким. Данные поисковых запросов в Google могут ежеминутно давать нам настоящие извержения ненависти, подпитываемые яростью.

Посмотрите, что произошло вскоре после массового расстрела в Сан-Бернардино, штат Калифорния, 2 декабря 2015 года. В то утро Ризван Фарук и Ташфин Малик встретились с коллегами Фарука, вооруженными полуавтоматическими пистолетами и полуавтоматическими винтовками, и убили 14 человек. В тот вечер — буквально через несколько минут после того, как СМИ впервые назвали имена стрелков, звучащие по-мусульмански, — значительное число калифорнийцев сообщили, что они хотят сделать с мусульманами: уничтожить их всех[22].

Самый популярный поисковый запрос в Google со словом «мусульмане» в то время в Калифорнии был: «Убивать мусульман». В целом американцы вводили запросы с этой фразой примерно с той же частотой, как и «рецепт мартини», «симптомы мигрени» и «состав команды «Ковбоев». Несколько дней после нападения в Сан-Бернардино практически каждый неравнодушный американец страдал «исламофобией». До нападения ненавистью были пропитаны около 20% всех поисковых запросов о мусульманах, но после него призывы к убийству мусульман присутствовали уже более чем

в половине всего объема поисков, связанных с мусульманами.

Эти расписанные по минутам поисковые данные в состоянии рассказать нам, как трудно может быть успокоить такой гнев. Через четыре дня после стрельбы тогдашний президент Обама выступил с обращением к гражданам страны в прайм-тайм. Он стремился успокоить американцев, говоря, что правительство способно остановить терроризм и — возможно, это даже более важно — усмирить эту опасную исламофобию.

Обама апеллировал к тому лучшему, что есть в нас, говоря о важности социальной интеграции и толерантности. Риторика была яркой и волнующей. «Лос-Анджелес Таймс» похвалила президента за «[предостережение] о том, чтобы не позволять страху затмить разум». «Нью-Йорк Таймс» назвала выступление как «жестким», так и «успокаивающим». Сайт Think Progress оценил его как «необходимый инструмент эффективного управления, направленный на спасение жизни американских мусульман». Иными словами, речь Обамы была признана большим успехом. Но так ли это?

Статистика поиска Google говорит об обратном. Вместе с Эваном Солтасом, на сей раз в Принстоне, я изучил данные. В своем выступлении президент сказал: «Это обязанность всех американцев любого вероисповедания — отказаться от дискриминации». Но в течение и вскоре после выступления количество поисковых запросов, в которых мусульман называли «террористами» и «злом», а также упоминались «насилие» и «жестокость», удвоилось. Еще Обама сказал: «Это наша обязанность — отказаться от религиозных предрассудков

в отношении тех, кого мы принимаем в этой стране». После чего негатив в поисковых запросах по поводу сирийских беженцев — в основном мусульман, — отчаянно искавших убежище, вырос на 60%. В то же время объем запросов о том, чем им можно помочь, сократился на 35%. Обама просил американцев: «Не стоит забывать, что свобода сильнее страха». И число запросов с текстом «убивать мусульман» во время его выступления утроилось. На самом деле во время и после речи Обамы количество почти всех поисковых запросов негативного толка в отношении мусульман, которые мы смогли придумать и проверить, увеличилось, а уровень позитивных по той же тематике значительно снизился.

Другими словами, казалось бы, Обама все говорил правильно. Все традиционные средства массовой информации поблагодарили его за целительные слова. Но новые данные из интернета, полученные с помощью цифровой сыворотки правды, показали: на самом деле речь дала результат, обратный желаемому, промахнувшись мимо своей главной цели. Согласно статистике из интернета, вместо того, чтобы успокоить разъяренную толпу, как все думали, Обама на самом деле распалил ее ярость. Порой то, что, как кажется, должно сработать, может иметь эффект прямо противоположный ожидаемому. Иногда для того, чтобы подправить инстинктивное стремление похвалить себя за хорошо проделанную работу, нам все же требуются независимые данные.

Так что должен был бы сказать Обама для подавления этой острейшей формы ненависти в Америке? Вернемся к этому позже. Сейчас же рассмотрим самый древний тип предрассудков в Соединенных Штатах — форму

ненависти, на самом деле занимающую первое место среди прочих. Ту, что была самой разрушительной и стала темой исследования, с которого началась эта книга. В моей работе со статистикой поисковых запросов в Google наиболее красноречивым фактом, связанным с проявлениями ненависти в интернете, является частота использования слова «ниггер».

Используемое в единственном или множественном числе, слово «ниггер» каждый год входит в семь миллионов поисковых запросов американцев. (Опять же, в песнях почти всегда это слово употребляется как «нигга», а не «ниггер», поэтому не следует считать, что на употребление этого слова оказала значительное влияние лирика хип-хопа.) Поиск «анекдотов о ниггерах» встречается в 17 раз чаще[23], чем «анекдоты о жидах», «тупые шутки», «анекдоты о латиносах», «анекдоты о китаезах» и «анекдоты о геях» вместе взятые.

Когда запросы со словом «ниггер(-ы)» или «анекдоты о ниггерах» встречаются чаще всего? Всякий раз, когда афроамериканцы упоминаются в новостях. Один из периодов, когда количество подобных запросов резко взметнулось вверх, случился сразу после урагана «Катрина». Тогда телевидение и газеты растиражировали изображения отчаявшихся чернокожих людей в Новом Орлеане, сражавшихся за свое выживание. Еще один всплеск подобных запросов наблюдался во время первых выборов Обамы. Примерно на 30% повышается объем поиска «анекдоты о ниггерах» в день памяти Мартина Лютера Кинга[24].

Пугающая повсеместность этого вида расовой дискриминации вызывает сомнения в сегодняшних представлениях об отсутствии в стране расизма.

Последний в Америке является настоящей головоломкой. С одной стороны, подавляющее большинство чернокожих американцев считают, что они страдают от предрассудков — и у них есть достаточно доказательств дискриминации со стороны полиции, во время собеседований при приеме на работу и при вынесении приговоров в суде. С другой стороны, очень немногие белые американцы готовы признаться в том, что они расисты.

До недавнего времени доминирующим объяснением политологов было следующее: это, по большей части, связано с широко распространенным *безотчетным* предубеждением. Белые американцы — теоретически — могут все понимать правильно, но у них есть подсознательное предубеждение, влияющее на их отношение к чернокожим. Ученые придумали гениальный способ проверить наличие такой предвзятости. Это называется имплицитный ассоциативный тест.

И он показал, что большинству людей требуется на несколько миллисекунд больше времени, чтобы связать черные лица с позитивными словами (как, например, «хороший»), чем с негативными (вроде «ужасный»). В отношении белокожего лица картина обратная. Дополнительное время является очевидным доказательством неявного безотчетного предубеждения — то есть предвзятости, которую человек может даже не осознавать.

Есть, однако, и альтернативное объяснение дискриминации, ощущаемой афроамериканцами и отрицаемой белыми: скрытый *явный* расизм. То есть речь идет о достаточно широко распространенном сознательном расизме, о котором люди неплохо осведомлены, но не желают в нем признаться — особенно во время опросов.

Вот о чем сообщают нам данные, полученные на основании анализа поисков в Google. Нет ничего неявного в запросе «анекдоты о ниггерах». И трудно представить, что американцы без явного расизма могут искать в Google слово «ниггер» с такой же частотой, как «мигрень» и «экономист» — и что это не оказывает существенного влияния на афроамериканцев. До появления статистики от Google у нас не было убедительного доказательства этой злобной неприязни. Сейчас оно есть. И теперь мы в состоянии показать, что это объясняет.

Это объясняет, как уже говорилось ранее, почему во многих регионах в 2008 и 2012 годах за Обаму было отдано так мало голосов. Кроме того, это также коррелирует с разницей в заработной плате у чернокожих и белых работников[25], о чем недавно сообщила одна команда экономистов. В областях, где, по моим наблюдениям, осуществляется больше всего расистских поисковых запросов, сильнее всего выражена и разница в оплате труда чернокожего и белого населения. А затем возникает феномен кандидатуры Дональда Трампа. Как уже отмечалось во введении, когда гуру опросов Нейт Сильвер искал географическую составляющую, наиболее сильно коррелирущую с поддержкой Трампа на первичных выборах республиканской партии в 2016 году, он нашел ее на карте расизма, которую разработал я. Этой переменной были поисковые запросы о «ниггерах».

Недавно ученые собрали воедино данные по штатам, связанные со скрытым предубеждением против чернокожих граждан. Они позволили мне сравнивать эффекты *явного* расизма, найденного с помощью поисковых запросов, и *скрытых* предубеждений. Например, я проверял,

насколько каждый вариант работал против Обамы на президентских выборах. Используя регрессионный анализ, я обнаружил, что области наибольшего распространения расистских поисковых запросов отлично показывают, где Обама недобрал голоса избирателей. Результаты имплицитно-ассоциативного теста мало что могут добавить.

Дискриминация, с которой регулярно сталкиваются в Соединенных Штатах чернокожие люди, скорее всего, подпитывается явной, а не скрытой враждебностью. Однако в случае других групп населения подсознательное предубеждение может иметь более существенное влияние. Например, я смог использовать анализ поисковых запросов в Google для того, чтобы найти доказательства скрытого предубеждения против девочек и молодых девушек.

И кто, могли бы вы спросить, станет иметь (и скрывать) предубеждение против девочек?

Их родители[26].

Вряд ли стоит удивляться тому, что родители маленьких детей часто чувствуют радостное возбуждение при мысли, что их чада могут быть одаренными. По сути, большинство поисковых запросов в Google, начинающиеся со слов «мой двухлетний ребенок…» заканчивается словами «одаренный», «талантливый». Но этот запрос касается мальчиков и девочек отнюдь не в равной степени. Родители в два с половиной раза чаще спрашивают «талантлив ли мой сын?», чем «талантлива ли моя дочь?». Они проявляют подобную предвзятость и при использовании других связанных с интеллектом фраз, которые могут стесняться произнести вслух — например, «мой сын гений?».

Родители наблюдают реальные различия между мальчиками и девочками? Возможно, первые чаще, чем

вторые, используют сложные слова или как-то иначе выказывают объективные признаки одаренности? Нет. Знаете, даже наоборот. В юном возрасте у девочек, как уже неоднократно было доказано, намного более обширный словарь и они используют более сложные предложения. В американских школах девочки на 9% чаще мальчиков входят в программу работы с одаренными детьми[27]. Несмотря на все это, родители почему-то видят больше одаренных мальчиков, чем девочек*. По сути, в каждом проверенном мной поисковом запросе, связанном с интеллектом (в том числе, с его отсутствием), родители чаще упоминают сыновей, а не дочерей. Намного больше запросов вида «неужели мой сын отстает в развитии» или «глупый», чем тех же сомнений в отношении дочерей. Однако в запросах с негативным звучанием (вроде «отстает» и «глупый») менее явно выражен перекос в сторону сыновей, чем в поиске с позитивными словами (такими как «талантливый» или «гениальный»).

Но тогда какие проблемы родители видят главными для своих дочерей? В первую очередь, все связанное с внешним видом. Рассмотрим вопросы о весе ребенка. Родители спрашивают в Google «У моей дочери избыточный вес?» примерно вдвое чаще, чем «У моего сына избыточный вес?». Они почти в два раза чаще спрашивают, как сбросить вес их дочери, чем задают аналогичный

---

* Для дальнейшей проверки гипотезы о том, что родители по-разному относятся к своим детям разного пола, я сейчас работаю над получением данных с сайтов для мам и пап. Эта статистика будет включать в себя информацию по гораздо большему количеству родителей, чем выборка, поисковыми запросами которой я оперирую сейчас. — *Прим. авт.*

вопрос, касающийся сына. И так же, как в случае с одаренностью, эти гендерные предрассудки не основаны на реальности. Избыточный вес имеют около 28% девочек и 35% мальчиков[28]. И несмотря на это родители видят — или волнуются — об избыточном весе девочек гораздо чаще, чем об ожирении мальчиков.

Кроме того, родители сообщают, что их дочь красива, в полтора раза чаще, чем что их сын красив. И они почти в три раза чаще сетуют на страшненькую дочку, чем говорят то же самое о сыне. (Каким образом, по их мнению, Google должен определить, является ребенок красивым или уродливым, сказать трудно.)

В общем, родители, похоже, в вопросах о сыновьях использовали позитивные слова чаще. Они более склонны спрашивать, «счастлив ли их сын» и менее готовы соединять воедино такие слова, как «сын» и «депрессия».

Либеральные читатели могут вообразить, что эти предубеждения более распространены в консервативных областях страны, но я не нашел этому никаких подтверждений. На самом деле мне не удалось найти значимой связи между любым из этих предрассудков и политическим или культурным обликом региона. Не существует доказательств того, что эти предубеждения уменьшились с 2004 года, когда для анализа впервые стали доступны поисковые данные Google. Кажется, эта предвзятость к девочкам распространена шире и укоренилась глубже, чем нам хотелось бы думать.

Сексизм — не единственная тема, где не срабатывают наши стереотипы о предубеждениях.

Vikingmaiden88 — пользователю Stormfront.org, самого популярного в Америке сайта, пропагандирующего

ненависть, 26 лет. Она любит читать романы и писать стихи. Ее подпись — это цитата из Шекспира. Я узнал все это из ее профиля и постов. Я также узнал, что Vikingmaiden88 нравится содержание сайта газеты, в которой я работаю — «Нью-Йорк Таймс». Она написала восторженный отзыв об одной из статей.

Недавно я проанализировал десятки тысяч таких профилей на Stormfront[29], в котором зарегистрированные участники могут указать место жительства, дату рождения, интересы и другую информацию.

Stormfront был основан в 1995 году Доном Блэком, бывшим лидером Ку-Клукс-клана. Наиболее популярные «социальные группы» на этом сайте — «Союз национал-социалистов» и «Поклонники и сторонники Адольфа Гитлера». За прошедший год, по данным Quantcast, ежемесячное посещение сайта составило примерно от 200 до 400 тысяч человек. В недавнем докладе Южного юридического центра по вопросам борьбы с бедностью было сказано, что за последние пять лет зарегистрированными пользователями Stormfront было совершено почти 100 убийств.

Члены Stormfront — не те, о ком бы я подумал.

Они, как правило, молоды — по крайней мере, согласно обозначенным ими датам рождения. Самый распространенный возраст, в котором люди регистрируются на этом сайте — 19 лет. Таковых здесь в четыре раза больше, чем 40-летних. Пользователи интернета и социальных сетей — в основном молодые люди. Но не настолько молодые.

Профили не имеют поля для указания пола. Но после просмотра всех постов и заполненных профилей случайной выборки американских пользователей выяснилось,

что пол большинства членов сайта можно вычислить: по моим оценкам, около 30% участников Stormfront — женщины.

Штаты с наибольшим количеством участников сайта на душу населения — Монтана, Аляска и Айдахо. Население этих регионов в подавляющем большинстве белое. Значит ли это, что люди, выросшие в среде с меньшим расовым разнообразием, больше пропитаны ненавистью?

Наверное, все же нет. Скорее, поскольку в этих штатах высок процент нееврейского белого населения, в них больше потенциальных членов групп, нападающих на евреев и темнокожих. Процент же целевой аудитории Stormfront на самом деле выше в районах с более значительным числом представителей этнических меньшинств. Это особенно верно в отношении членов данного сайта в возрасте 18 лет и моложе — следовательно, они не сами выбирают, где им жить.

Если говорить об этой возрастной группе, Калифорния — штат с наибольшим представительством различных национальных меньшинств. И количество членов Stormfront из этого штата на 25% выше, чем в среднем по стране.

Одна из самых популярных социальных групп на сайте — «В поддержку антисемитизма». Процент участников сайта, присоединившихся к ней, положительно соотносится с численностью еврейского населения каждой области. Так, количество входящих в эту группу жителей штата Нью-Йорк (региона с наибольшим числом еврейского населения) также выше, чем в среднем по стране.

В 2001 году Dna88 зарегистрировался на Stormfront, описывая себя как «хорошо выглядящий, расово подходящий» 30-летний разработчик интернет-сайтов,

живущий в «Евре-йорке». В следующие четыре месяца он написал более двухсот постов вроде «Преступления евреев против человечности» и «Еврейские кровавые деньги» и направлял людей на сайт jewwatch.com, позиционирующий себя как «научная библиотека» с материалами по «Сионистской уголовщине».

Члены Stormfront сетуют на то, что представители этнических меньшинств говорят на других языках и совершают различные преступления. Но наиболее интересными для меня были их жалобы по поводу конкуренции на рынке знакомств.

Человек, называющий себя Уильямом Лайоном Маккензи Кингом — в честь бывшего премьер-министра Канады, однажды заявившего: «Канада должна стать страной белых людей», — написал в 2003 году, что он изо всех сил старался «сдержать» свою «ярость», увидев белую женщину, «таскающуюся с уродливым черным ублюдком-полукровкой». В своем профиле Whitepride26, 41-летняя студентка из Лос-Анджелеса, говорит: «Я не люблю черных, латиносов и иногда азиатов — особенно когда мужчины находят их женщин более привлекательными [чем белых женщин]».

Некоторые политические события также играют определенную роль. День, когда был отмечен самый большой прирост членства в Stormfront за всю историю его существования, — однозначно, 5 ноября 2008 года, на следующий день после избрания Барака Обамы президентом. Однако в Stormfront не было замечено никакого повышенного интереса к кандидатуре Дональда Трампа[30], лишь небольшой подъем интереса сразу после его победы на выборах. Трамп поднялся на волне белого

национализма, однако нет никаких доказательств того, что именно он создал эту волну.

Избрание же Обамы вызвало всплеск белого националистического движения. Похоже, избрание Трампа стало результатом этого всплеска.

Экономика вроде бы не имеет никакой связи с регистрацией в Stormfront. Нет никакой связи между ежемесячной регистрацией новых членов сайта и уровнем безработицы в стране. Великая рецессия оказала различное влияние на разные штаты, но между ней и относительным увеличением количества поисковых запросов в Google о Stormfront нет видимой связи.

Но, возможно, наиболее интересным — и самым удивительным — может показаться выбор участниками сайта некоторых тем для разговора. Они похожи на то, о чем мы разговариваем с друзьями. Может, я слишком наивен, но мне представлялось, что белые националисты населяют совсем иной мир, чем мы с моими друзьями. Вместо этого они вовсю нахваливают «Игру престолов» и сравнивают достоинства сайтов знакомств вроде PlentyOfFish и OkCupid.

И ключевой факт, доказывающий, что пользователи Stormfront обитают в одном мире с остальными людьми: популярность среди них «Нью-Йорк Таймс». Не только VikingMaiden88 хвалит статьи этой газеты — она популярна среди многих членов Stormfront. На самом деле, если сравнить пользователей этого сайта с людьми, посещающими новостной сайт Yahoo, получается, что члены Stormfront интересуются nytimes.com* вдвое чаще.

---

* Сайт «Нью-Йорк Таймс». — *Прим. ред.*

Члены сайта, пропитанного ненавистью, просматривают «ой-такой-либеральный» nytimes.com? Как это может быть? Если значительное число членов Stormfront узнают новости из nytimes.com, это означает, что наше расхожее мнение о белых националистах неверно. Это также означает, что наше привычное представление о том, как работает интернет, также ошибочно.

ПРАВДА ОБ ИНТЕРНЕТЕ

Интернет — с этим согласны почти все — отталкивает американцев друг от друга, заставляя большинство людей прятаться на сайтах, ориентированных на таких, как они. Вот как описала ситуацию Кэсс Санштейн из юридической школы Гарварда: «Наш рынок общения стремительно движется к ситуации, когда люди ограничивают себя своей собственной точкой зрения. Либералы смотрят и читают в основном только либеральную прессу, умеренные — умеренную, консерваторы — консервативную, неонацисты — неонацистскую».

Это представление имеет смысл. Ведь интернет дает нам практически неограниченное количество источников, из которых мы можем потреблять новости. Я могу читать все что хочу. Вы можете читать все что хотите. VikingMaiden88 может читать все что захочет. И люди, если позволить им действовать самостоятельно, стремятся к поиску мнений, подтверждающих то, во что они верят. Таким образом, конечно, интернет способствует созданию крайне выраженной политической сегрегации[31].

С этим очевидным представлением связана одна проблема: данные говорят нам, что это неправда.

Улики против этой бытовой мудрости были получены в результате исследования 2011 года, проведенного Мэттом Генцкоу и Джесси Шапиро, двумя экономистами, чьи работы мы обсуждали ранее.

Генцкоу и Шапиро собирали данные о том, что и как просматривают представители большой выборки американцев. Сюда входили такие параметры, как идеология и самооценка: ученые выясняли, считают ли себя люди из выборки более либеральными или консервативными. Исследователи использовали эти данные для оценки политической сегрегации в интернете.

Как? Они провели интересный мысленный эксперимент.

Предположим, вы случайным образом выбрали двух американцев, которые — так получилось — посещают один и тот же новостной сайт. Какова вероятность, что один из них будет либералом, а другой — консерватором? Иными словами, как часто либералы и консерваторы «встречаются» на новостных сайтах?

Рассмотрим ситуацию дальше. Пусть и те и другие никогда не получают онлайн-новости в одном и том же месте. Иными словами, либералы посещают исключительно либеральные сайты, а консерваторы — исключительно консервативные. Если бы это было так, то шансы на то, что два американца, посетившие один новостной сайт, имеют противоположные политические взгляды, будут равны нулю. Интернет окажется идеально *сегрегирован*. Либералы и консерваторы никогда не смешаются.

Предположим, напротив, что и те и другие имеют одинаковые предпочтения мест ознакомления с новостями. Другими словами, и либерал, и консерватор равновероятно могут посетить один и тот же новостной сайт.

Если бы это было так, то шансы на то, что два американца на данном сайте имеют противоположные политические взгляды, будут равны примерно 50%. Интернет окажется абсолютно *десегрегированным*. Либералы и консерваторы перемешаются.

Итак, что же говорят нам полученные Генцкоу и Шапиро данные? В Соединенных Штатах шансы на то, что двое людей, посещающих один и тот же новостной сайт, имеют разные политические взгляды, составляют около 45%. Другими словами, интернет гораздо ближе к идеальной десегрегации, чем к идеальной сегрегации. Либералы и консерваторы постоянно «встречаются» там друг с другом.

Реально оценить отсутствие сегрегации в интернете позволяет сравнение с сегрегацией в других областях нашей жизни. Генцкоу и Шапиро смогли повторить свой анализ для различных видов взаимодействий вне интернета. Каковы шансы на то, что два члена семьи имеют разные политические взгляды? Два соседа? Двое коллег? Двое друзей?

Используя данные социологического опроса, ученые обнаружили: все эти цифры были ниже, чем у двух человек, посещающих один новостной сайт.

Вероятность того, что вы встретитесь с человеком, имеющим противоположные политические взгляды, %

На новостном сайте	45,2
Коллеги	41,6
Соседи, вне сети	40,3
Члены семьи	37
Друзья	34,7

Другими словами, у вас больше шансов встретить кого-то с противоположными политическими взглядами в интернете, чем в реальной жизни.

Почему же сегрегация в сети меньше, чем в обычной жизни? Есть два фактора, ограничивающие ее в интернете.

Во-первых, что довольно удивительно, источниками новостей в сети являются всего нескольких крупных новостных порталов. Мы обычно думаем об интернете как о средстве, обращенном к маргиналам. Действительно, там есть сайты для всех, независимо от их точки зрения. Есть места для высказывания сторонников свободного владения оружием и противников свободного владения оружием, борцов за свободу курения и активистов, призывающих к монетизации доллара, анархистов и белых националистов. Но все эти сайты вместе взятые составляют лишь небольшую долю новостного трафика интернета. На самом деле в 2009 году четыре сайта — Yahoo News, AOL News, msnbc.com и CNN.com — собрали более половины просмотров новостей. Первый из них остается наиболее популярным новостным сайтом у американцев — около 90 миллионов посетителей в месяц, или в 600 раз больше, чем у Stormfront. Сайты масс-медиа — такие как Yahoo News — обращаются к широкой, политически разнообразной аудитории.

Вторая причина практически нулевой сегрегированности интернета заключается в том, что многие люди с ярко выраженными политическими взглядами, едва начав злиться (и у них появляется желание спорить), сразу отправляются на сайты с противоположной точкой зрения. Политические наркоманы не ограничивают

себя сайтами, ориентированными на их мнение. Посетители thinkprogress.org и moveon.org — двух крайне либеральных сайтов — с большей, чем у среднего пользователя интернета, вероятностью зайдут на сайт правого толка foxnews.com. А посетители rushlimbaugh.com или glennbeck.com — двух крайне консервативных сайтов — с большей вероятностью, чем средний пользователь интернета, зайдут на более либеральный nytimes.com.

Исследование Генцкоу и Шапиро было основано на данных за 2004–2009 гг. — сравнительно раннего периода в истории интернета. Возможно, с тех пор сеть могла стать более разобщенной? Социальные медиа, и в частности Facebook, могли измениться? Понятно, что если наши друзья, как правило, разделяют наши политические взгляды, то рост соцсетей должен означать рост числа поддакиваний. Верно?

Опять же, все не так просто. Хотя мнение о том, что друзья на Facebook чаще всего придерживаются одинаковых политических взглядов, справедливо, данные, полученные группой ученых — Эйтаном Бакши, Соломоном Мессингом и Ладой Адамик, — показали: удивительное количество информации, которое люди получают на Facebook, исходит от пользователей с противоположными взглядами.

Как такое может быть? Разве наши друзья, как правило, не разделяют наши политические взгляды? Разделяют. Но Facebook может привести к более разнообразной политической дискуссии, чем общение вне сети — и тому есть одна важная причина. Люди в среднем имеют

гораздо больше друзей на Facebook[32], чем в реальной жизни. И эти неблизкие связи, скорее всего, будут соединять людей с противоположными политическими взглядами.

Другими словами, Facebook снабжает нас слабыми социальными связями[33]: знакомый по институту, психованный троюродный брат, друг друга друга, ну и все такое — вы наверняка знаете. Это люди, с которыми вы никогда не могли бы пойти в боулинг или отправиться на шашлыки. Вы не стали бы приглашать их на ужин. Но вам ничто не мешает добавить их в друзья в Facebook. И вы видите их ссылки на статьи, выражающие мнение, с которым вы бы иначе никогда не познакомились.

В целом интернет объединяет людей разных политических предпочтений. Средняя дама либеральных взглядов может проводить утро со своим либеральным мужем и либеральными детьми, день — с либеральными коллегами. Ее окружают либеральные наклейки на бамперах. А вечером ее ждут либеральные приятельницы по классу йоги.

Но когда она вернется домой и внимательно прочтет несколько консервативных комментариев на CNN.com или получит ссылку на Facebook от своего знакомого-республиканца, это может стать ее самым консервативным впечатлением дня.

Я, скорее всего, никогда не столкнусь с белыми националистами в моем любимом кафе в Бруклине. Но мы с VikingMaiden88 часто переписываемся на сайте «Нью-Йорк Таймс».

ПРАВДА О ЖЕСТОКОМ ОБРАЩЕНИИ С ДЕТЬМИ И АБОРТАХ

Интернет в состоянии дать нам представление не только о возмущающих нас взглядах, но и о возмутительном поведении. Действительно, данные Google могут эффективно предупреждать о проблемах, пропущенных всеми обычными источниками. Ведь при возникновении серьезных сложностей люди обычно обращаются к поиску онлайн.

Рассмотрим проблему жестокого обращения с детьми во время Великой рецессии.

Когда в конце 2007 года начался этот крупный экономический спад, естественно, многие эксперты обеспокоились тем, какое влияние это явление окажет на детей. Ведь многие родители могут испытывать сильные стресс и депрессию, а это основные факторы риска жестокого обращения: число подобных случаев может резко увеличиться.

Впрочем, после получения официальных данных стало казаться, что беспокойство необоснованно. Учреждения службы защиты детей сообщили об уменьшении количества сигналов о случаях жестокого обращения. Кроме того, эти выводы были наиболее оптимистичными в штатах, пострадавших от рецессии сильнее всего. «Мрачные прогнозы не сбылись»[34], — сказал в 2011 году в интервью «Ассошиэйтед Пресс» Ричард Геллес, специалист службы охраны здоровья и благополучия детей из университета Пенсильвании. Да, сколь бы нелогичным это ни казалось, похоже, число случаев жестокого обращения с детьми во время рецессии снизилось.

Но могло ли это действительно быть так — при том, что огромное число взрослых оказалось в крайне неприятной

ситуации? Я с трудом верил в это. Потому-то и обратился к статистике Google.

Оказалось, что некоторые дети вводили в поисковик совершенно трагические и душераздирающие запросы — такие как «Моя мама бьет меня» или «Мой папа ударил меня». Они представляют совсем иную — крайне мучительную — картину происходившего в этот нелегкий период. Количество подобных поисковых запросов во время Великой рецессии четко соответствовало изменению уровня безработицы.

Вот что, по моему мнению, произошло: снизилось число *сообщений* о случаях жестокого обращения с детьми, а не самих подобных прецедентов. По разным оценкам, до сведения властей доводился лишь небольшой процент случаев жестокого обращения с детьми. Во время рецессии многие люди, которые обычно сообщают о подобных прецедентах (например, учителя и полицейские) и занимаются ими (работники службы защиты детей), скорее всего, были перегружены работой или и вовсе остались без нее.

Действительно, есть еще одно свидетельство — на этот раз не от Google — того, что количество случаев жестокого обращения с детьми в период рецессии на самом деле выросло. Когда ребенок умирает из-за родительской жестокости или небрежности, об этом сообщается властям. Число таких смертей (хотя подобное все же происходило редко) возросло в штатах, пострадавших от рецессии сильнее всего.

Есть некоторая информация из Google, позволяющая подозревать людей из наиболее пострадавших регионов в жестоком обращении с детьми. На основании данных

за предшествующий рецессии период, а также учитывая общенациональные тенденции, можно утверждать: в штатах, сильнее пострадавших от экономического спада, возросло число поисковых запросов относительно жестокого обращения с детьми. На каждый 1% роста безработицы приходится 3% увеличения количества запросов со словами «жестокое обращение с детьми» или «беспризорники». Скорее всего, о большинстве событий, ставших причиной подобных запросов в Google, никто никогда не узнает, поскольку в этих регионах резко снизился объем отчетности.

Число поисковых запросов от страдающих детей увеличивается. Показатели детской смертности взлетают вверх. Растет количество запросов людей, подозревающих о злоупотреблениях. Но число официально зафиксированных случаев насилия снижается. Вследствие рецессии все больше детей сообщают Google о том, что родители ударили или даже избили их. Все больше людей подозревают, что они стали свидетелями насилия. Но перегруженные государственные службы способны реагировать на все меньшее количество подобных инцидентов.

Полагаю, можно с уверенностью заявить: в результате Великой рецессии случаев жестокого обращения с детьми стало больше, хотя традиционными способами это не выявляется.

Всегда, когда начинаю подозревать, что страдания людей не отражаются в официальной статистике, я обращаюсь к информации Google. Одно из потенциальных преимуществ получения этих новых данных и умения их интерпретировать — возможность оказания помощи наиболее уязвимым группам населения, мучения

которых в противном случае могут остаться незамеченными властями.

Таким образом, когда Верховный суд недавно изучал действие закона, ограничивавшего аборты, я обратился к Google. Я подозревал, что пострадавшие от этого закона женщины могли искать неофициальные способы прерывания беременности[35]. И они это делали. Число подобных запросов было наиболее велико в штатах, принявших указанный закон.

Данные, полученные на основании поисковых запросов, оказались не только полезными, но и весьма тревожными.

В 2015 году в Соединенных Штатах было произведено более 700 тысяч запросов в Google относительно самостоятельного прерывания беременности. Для сравнения, в том же году люди около 3,4 млн раз искали абортарии. То есть значительный процент женщин обдумывал возможность совершения аборта своими силами.

Женщины около 160 тысяч раз искали способы получения таблеток для прерывания беременности по неофициальным каналам — «купить таблетки для прерывания беременности онлайн» и «таблетки для прерывания беременности в свободной продаже». Они спрашивали у Google об аборте с помощью трав вроде петрушки или витамина С. Было обнаружено около 4000 запросов об аборте с помощью противозачаточных средств, в том числе около 1300 содержали четкую фразу: «Как сделать аборт с помощью противозачаточных средств?» Кроме того, нашлось несколько сотен запросов об аборте посредством введения в матку отбеливающих растворов и один — при помощи удара по животу.

Что пробуждает интерес к теме самостоятельного осуществления аборта? География распределения и даты запросов в Google указывают на вероятную причину: когда официально прервать беременность сложно, женщины начинают искать подпольные способы решения своей проблемы.

Уровень запросов об абортах домашними средствами был довольно стабильным в период с 2004 по 2007 год и начал расти в конце 2008-го, когда случился финансовый кризис, за которым последовала рецессия. В 2011 году число подобных запросов резко увеличилось, достигнув 40%. Институт Гуттмахера — организация по охране репродуктивных прав — заявил, что в 2011 году началось массовое закручивание гаек в отношении абортов: были приняты 92 государственных положения, ограничивающие прерывание беременности. А вот в Канаде, где еще не началось наступление на репродуктивные права женщин, увеличения запросов о самостоятельных абортах не произошло.

Больше остальных выросла частота поиска в Google способов самостоятельно прервать беременность в Миссисипи — штате, где на три миллиона человек аборты теперь делаются всего в одной клинике. По данным Института Гуттмахера, в восьми из десяти штатов с самым высоким числом запросов о самостоятельном прерывании беременности к медицинским абортам относятся негативно или очень враждебно. В 10 штатах с наименьшим числом запросов о домашних абортах ситуация ровно противоположная.

Конечно, на основании поисков в Google мы не можем сказать, сколько именно женщин успешно осуществили

аборты самостоятельно, но опыт показывает: их было много. Один из способов пролить свет на этот вопрос — сравнить данные абортов и родов.

В 2011 году, последнем, за который у нас есть полная картина абортов по регионам, женщины, живущие в штатах с меньшим числом абортариев, совершили намного меньше легальных прерываний беременности.

Сравните 10 штатов с наибольшим числом абортариев на душу населения (в список которых входят в том числе Нью-Йорк и Калифорния) с десяткой регионов с наименьшим их количеством на душу населения (список включает в том числе Миссисипи и Оклахому). Женщины из штатов с наименьшим количеством абортариев совершили на 54% меньше легальных абортов. Разница — в 11 абортов на каждую тысячу женщин в возрасте от 15 до 44 лет. При этом женщины, живущие в регионах с наименьшим количеством абортариев, чаще рожали. Однако этого недостаточно для компенсации снижения числа абортов. Разница — в шесть родов на каждую тысячу женщин детородного возраста.

Иными словами, похоже, в тех областях страны, где было труднее сделать аборт, существовали «потерянные» беременности. Официальные источники не говорят нам, что произошло с этими пятью не случившимися родами на каждую тысячу женщин.

Однако Google дает неплохие подсказки.

Мы не можем слепо доверять правительственным данным. Государство может сказать нам, что количество жестоких обращений с детьми или число абортов снизились, и политики отпразднуют это как свое достижение. Но, как мы уже видели, подобные результаты могут быть

следствием неверного метода сбора данных. Правда может быть иной, и иногда очень даже неприятной.

ПРАВДА О ВАШИХ ДРУЗЬЯХ НА FACEBOOK

В целом моя книга — о больших данных. Но эта глава в основном посвящена поиску в Google, где, как я понял, перед нами предстает ранее скрытый мир, сильно отличающийся от того, что мы видим вокруг себя. Но являются ли и другие источники больших данных цифровой сывороткой правды? На самом деле, многие из них — такие как Facebook — часто представляют собой полную ее противоположность.

В социальных сетях, так же, как и в опросах, у вас нет стимула проявлять честность. Наоборот, там — намного больше, чем в опросах — вам хочется лучше выглядеть. Ведь, прежде всего, ваше присутствие в интернете не анонимно — вы любезничаете с аудиторией и рассказываете о себе друзьям, членам семьи, коллегам, знакомым и незнакомым людям.

Чтобы понять, насколько неточными и необъективными могут быть сведения в соцсетях, можно рассмотреть относительную популярность респектабельного, высоколобого ежемесячного журнала «*Atlantic*» и газеты «*National Enquirer*», набитой сплетнями и сенсациями. Оба издания имеют схожие средние тиражи[36] — по нескольку сотен тысяч экземпляров. («*National Enquirer*» выходит еженедельно, так что на самом деле продает больше копий.) Сопоставимо и число поисковых запросов в Google о каждом из них.

Однако на Facebook примерно 1,5 миллиона[37] человек обсуждают в своих профилях статьи из «*Atlantic*» и только около 50 тысяч признаются, что читают «*National Enquirer* «или обсуждают ее содержание.

***«Atlantic»* и «*National Enquirer*»**
Сравнение на основании различных источников

Тираж	Ориентировочно 1 «*Atlantic*» на 1 «*National Enquirer*»
Поисковые запросы в Google	1 «*Atlantic*» на 1 «*National Enquirer*»
Лайки в Facebook	27 «*Atlantic*» на 1 «*National Enquirer*»

Данные о тираже являются эталоном для оценки популярности СМИ. С ними может сравниться статистика поисковых запросов в Google. А негативные отзывы о желтой газетенке в Facebook по большей части являются предвзятыми — соответственно, эта сеть является худшим источником данных для определения того, что нравится людям.

На Facebook такая картина во всем — как в отношении журналов, так и в плане любых житейских предпочтений. В соцсети мы выставляем свой улучшенный, окультуренный портрет, а не истинное лицо. В этой книге, в частности, в данной главе, я использую данные Facebook — но всегда помня об этой особенности.

Чтобы лучше понять, чего не хватает в информации из социальных сетей, вернемся на минутку к порнографии. Во-первых, нужно рассмотреть всеобщее убеждение о том, что в интернете преобладают чернуха и похабщина. Это

неправда. Основная часть контента в интернете отнюдь не порнографическая. Например, ни один из 10 наиболее посещаемых веб-сайтов[38] не связан с порнографией, поэтому популярность порно — надо признать, она весьма высока — не стоит преувеличивать.

Итак, внимательно оценив то, как нам нравится порнография и какую долю контента она занимает, можно утверждать: Facebook, Instagram и Twitter являются лишь очень ограниченной выборкой из того, что по-настоящему популярно в интернете. В Сети имеются большие подмассивы данных, которые невероятно популярны, но не особо бросаются в глаза.

Наиболее известным видео за все время (на момент написания этой книги) является «Gangnam Style» от Psy — тупое видео с поп-музыкой, в котором высмеиваются корейские модники. С момента дебюта в 2012 году только на YouTube его просмотрели около 2,3 миллиарда раз. И его популярность понятна — неважно, на каком сайте вы его нашли. Оно распространялось на различных социальных медиаплатформах десятки миллионов раз.

Наиболее известное порнографическое видео всех времен — «Отличное тело, отличный секс, отличный минет». Его просмотрели более 80 миллионов раз. Другими словами, на каждые 30 просмотров «Gangnam Style» приходится по крайней мере один просмотр «Отличного тела…». Если социальные медиа дали нам точное представление о том, какое видео люди смотрели, то «Отличное тело…» должны были перепостить миллионы человек. Тем не менее оно появилось в соцсетях всего

несколько десятков раз — и всегда на страничках порнозвезд, а не обычных пользователей. Люди явно не чувствуют необходимости афишировать друзьям свою заинтересованность в этом видео.

Facebook — это «средство для того, чтобы похвастаться друзьям о том, как в моей жизни все хорошо».

В мире Facebook среднестатистический взрослый человек вроде бы счастливо женат и отдыхал на Карибском море, просматривая «*Atlantic*». В реальном мире многие люди ругаются у касс супермаркетов, просматривают «*National Enquirer*» и игнорируют телефонные звонки супруга, с которым не спят вместе уже много лет. В мире Facebook семейная жизнь выглядит идеальной. В реальном же мире она — полный раздрай. Иногда это может быть такой сумбур, что некоторые даже жалеют, что завели детей. В мире Facebook кажется, что каждый молодой человек по субботам отправляется на крутую вечеринку. В реальном мире большинство сидят дома в одиночестве и не отрываясь смотрят сериалы на Netflix. В мире Facebook подруга постит 26 фотографий счастливого отпуска со своим бойфрендом. В реальности сразу же после этой публикации она пишет в Google: «Мой парень не хочет заниматься сексом со мной». Кстати, возможно, в это же время ее парень смотрит «Отличное тело, отличный секс, отличный минет».

Цифровая правда	Цифровая ложь
• Поисковые запросы • Просмотры • Клики • Выделения текста	• Посты в социальных сетях • Лайки в социальных сетях • Профили на сайтах знакомств

ПРАВДА О ВАШИХ КЛИЕНТАХ

Ранним утром 5 сентября 2006 года[39] Facebook представил значительное обновление своей главной страницы. В ранних версиях этой соцсети узнать, что делают их друзья, пользователи могли, лишь кликая по их профилям. Сайт считал большой удачей, если на нем одновременно находилось 9,4 миллиона человек.

Но после нескольких месяцев напряженной работы инженеры создали то, что было названо News Feed (новостной лентой), которая должна предоставлять пользователям информацию о том, что делают все их друзья.

Люди сразу же сообщили, что они возненавидели новостную ленту. Бен Парр, старшекурсник Северо-Западного университета, создал группу «Студенты против новостной ленты на Facebook». Он сказал, что «это слишком жутко, слишком навязчиво, и эту функцию следует убрать». В течение нескольких дней группа из 700 тысяч человек поддержала мнение Парра. Один из младших студентов Мичиганского университета сказал в интервью «*Michigan Daily*»: «Я серьезно испуган новым Facebook. Это заставляет меня чувствовать, будто за мной постоянно идет охота».

Дэвид Киркпатрик описал эту ситуацию в своем аккаунте, посвященном истории сайта и названном «Влияние Facebook: внутренняя история компании, которая объединяет весь мир». Он назвал введение новостной ленты «самым серьезным кризисом, с которым Facebook когда-либо сталкивался». Но когда Киркпатрик брал интервью у сооснователя и руководителя Facebook Марка Цукерберга, тот был невозмутим.

Причина? Цукерберг имел доступ к цифровой сыворотке правды: количеству кликов и визитов людей на Facebook. Вот что пишет Киркпатрик:

> «На самом деле Цукерберг знал, что людям понравилась новостная лента — неважно, что они говорили о ней в группах. У него на руках были данные, доказывающие это. В среднем люди стали тратить на Facebook больше времени, чем до запуска новостной ленты — значительно больше. В августе пользователи просмотрели 12 миллиардов страниц сайта, тогда как в октябре, с новостной лентой, уже 22 миллиарда.

И это были еще не все доказательства, имевшиеся в распоряжении Цукерберга. Даже вирусная популярность группы противников ленты новостей свидетельствовала о привлекательности нововведения. Группа смогла вырасти так быстро именно благодаря ленте. Очень многие пользователи смогли узнать о том, что их друзья присоединились к ней — и они узнали об этом из новостной ленты.

Другими словами, в то время как люди стали присоединяться к группам, во всеуслышание заявляющим

о неприятии того, что на Facebook теперь видны все подробности их жизни, они открывали новостную ленту, чтобы ознакомиться со всеми подробностями жизни своих друзей. И лента осталась. Сейчас на Facebook ежедневно приходит более миллиарда активных пользователей.

В своей книге «*От нуля к единице. Как создать стартап, который изменит будущее*» Питер Тиль, один из первых инвесторов в Facebook, утверждает, что большой бизнес строится на тайнах[40] — и неважно, тайны ли это природы или людские секреты. Джефф Седер, как уже рассказывалось в главе 3, сумел раскрыть загадку природы, выяснив, что по размерам левого желудочка можно спрогнозировать будущие успехи лошади. Google открыл секрет, насколько мощной силой могут быть информация и связи между людьми.

Тиль определяет «секреты людей» как «то, чего они не ведают о себе, или то, что они скрывают, поскольку не хотят, чтобы другие об этом знали». Иными словами, бизнес подобного рода строится на лжи.

Вы можете возразить, что сам Facebook основан на неприятном секрете о людях, который Цукерберг узнал во время учебы в Гарварде. В начале второго курса он создал сайт для своих сокурсников, который назвал Facemash*. Он был создан по образцу сайта под названием «Hot or not?» («Привлекательный или нет?»). На Facemash были выставлены фотографии двух гарвардских студентов, а их друзья могли оценивать, кто лучше выглядит.

---

* Face — лицо. Mash — флирт, любовная связь, болтушка. — *Прим. перев.*

Сайт второкурсника был встречен с возмущением. В редакционной статье «*Harvard Crimson*» Марка обвиняли в том, что он «потакает самым низменным страстям людей». Испаноязычные и афроамериканские группы обличили его в сексизме и расизме. Однако прежде чем администраторы Гарварда закрыли Цукербергу доступ в интернет (спустя несколько часов после запуска сайта), 450 человек успели просмотреть Facemash и 22 тысячи раз проголосовать за тот или иной снимок.

Цукерберг узнал важный секрет: люди могут утверждать, что они в бешенстве, могут что-то ругать, но они все равно заходят, смотрят и кликают.

Кроме того, он узнал еще одну важную вещь: при всем их профессионализме, серьезности, ответственности и уважении к частной жизни, люди (даже студенты Гарварда) проявляют большой интерес к оценке людей по внешности. Ему это подсказали просмотры и голосования. А позже — поскольку данные Facemash оказались слишком противоречивыми — он понял, насколько сильно могут быть заинтересованы люди в узнавании поверхностных фактов о других. Это позволило ему создать самое успешное предприятие своего поколения.

Компания Netflix выучила этот урок и использовала с самого начала своего существования: не доверяйте тому, что люди говорят — верьте только тому, что они делают.

Компания изначально позволила пользователям создавать список из фильмов, которые они хотели бы посмотреть в будущем, но на просмотр которых в данный момент у них нет времени. Таким образом, когда у людей появлялось больше времени, Netflix могла напомнить им об этих фильмах.

Тем не менее в этих данных было замечено нечто странное. Пользователи заполняли список большим количеством фильмов, но спустя несколько дней, когда им напоминали об этом, они редко на них кликали.

Так в чем была проблема? Спросите у людей, что они планируют посмотреть в течение ближайших нескольких дней, и они заполнят список желаний высоколобыми лентами вроде черно-белых документалок времен Второй мировой войны или серьезных иностранных фильмов. Однако несколько дней спустя они уже хотят смотреть только те фильмы, которые обычно смотрят: невзыскательные комедии или романтические ленты. Люди постоянно лгут самим себе.

Столкнувшись с таким парадоксом, Netflix перестал спрашивать людей о том, что они хотят посмотреть в будущем, а стала строить модель, основанную на миллионах кликов и просмотров. Компания начала приветствовать своих пользователей, предлагая им список кинолент, основанный не на теоретических предпочтениях посетителей, а на статистике наиболее частых просмотров. Результат: клиенты стали чаще посещать Netflix и смотреть больше фильмов.

«Алгоритмы знают вас лучше, чем вы знаете сами себя», — говорит Ксавье Аматриэн[41], бывший специалист по сбору данных в Netflix.

Насколько важно игнорировать то, что люди говорят вам

Что люди говорят	Реальность	В результате
Они не хотят отслеживать своих друзей.	Мало что в этом мире они хотят больше, чем подсматривать за своими друзьями и обсуждать их.	Марк Цукерберг, основатель Facebook, стоит 55,2 млрд долларов.
Они не хотят покупать продукты, которые производятся в подпольных цехах.	Они будут покупать хорошие, «недорогие» продукты.	Фил Найт, основатель компании Nike, стоит 25,4 млрд долларов.
Они хотят слушать новости по утрам.	Они хотят вечерами слушать о лилипутах, занимающихся сексом с порнозвездами.	Говард Штерн стоит 500 млн долларов.
Они не желают читать о связывании, доминировании и садомазохизме.	Они хотят читать о БДСМ-отношениях молодой выпускницы колледжа и бизнес-магната.	Продано 125 млн экземпляров книги «50 оттенков серого».
Они хотят, чтобы политики четко излагали свои позиции.	Они хотят, чтобы политики избавили их от подробностей, но выглядели решительными и уверенными в себе.	Дональд Трамп.

СПОСОБНЫ ЛИ МЫ ВЫДЕРЖАТЬ ПРАВДУ?

Части этой главы могут вас расстроить. Цифровая сыворотка правды выявила неизменное стремление судить о людях по их внешности и существование миллионов одиноких мужчин-геев, показала, что значимый процент женщин фантазирует об изнасиловании и что предубеждение против афроамериканцев достаточно широко распространено. Оказалось, что существуют скрытые пласты жестокого обращения с детьми и самостоятельных абортов. Выяснилось, что подспудно зреют яростные исламофобские настроения, лишь усилившиеся после обращения президента к народу с призывом к толерантности. Совсем невесело! После рассказов о моих исследованиях люди нередко приходят ко мне и говорят: «Сет, это все очень интересно, но так мрачно и утомительно!»

Я не могу притвориться, будто некоторые из этих данных не выявили мрачного подтекста. Если люди постоянно говорят нам лишь то, что, по их мнению, мы хотим услышать, значит, мы вообще слышим только нечто более комфортное, чем правда.

В среднем цифровая сыворотка правды показывает, что мир хуже, чем мы о нем думали.

Нужно ли нам знать это? Знать о поисковых запросах в Google, о данных с порносайтов, о том, кто и куда

кликает… Это может заставить вас думать: «Здорово. Теперь мы можем понять, кто мы есть на самом деле». Но вы можете подумать и иначе: «Ужасно. Теперь мы можем понять, кто мы есть на самом деле».

Но правда помогает — и не только Марку Цукербергу или другим желающим — привлечь клики или клиентов. Существует минимум три способа, с помощью которых эти знания в состоянии улучшить нашу жизнь.

Во-первых, вам может быть приятно узнать, что вы не одиноки в своей неуверенности в себе и в своем неловком поведении. Вам может быть нужно знать, что другие люди также не уверены в своем теле. Вероятно, многим — особенно тем, у кого не так много секса в жизни, — принесет облегчение информация о том, что не весь мир совокупляется, как кролики. И, возможно, это может оказаться ценным для старшеклассника из Миссисипи, влюбленного в квотербека, — знать, что, несмотря на низкую официальную численность открытых геев среди окружающих мужчин, вокруг все же много людей, испытывающих аналогичное притяжение.

Есть еще одна область — я ее еще не затрагивал, — где поиск в Google может показать, что вы не одиноки. В юности учитель, возможно, говорил вам: если у вас есть вопрос, вы должны поднять руку и задать его. Потому что если ты не понимаешь чего-то, то другие, вероятно, тоже этого не поняли. Если вы похожи на меня, значит, вы проигнорировали совет учителя и тихо сидели, боясь открыть рот. Вам казалось, что ваши вопросы слишком тупые (можно подумать, у других они очень глубокомысленные). Анонимные агрегированные данные Google могут раз и навсегда закрыть тему о том, насколько правы

были наши учителя. Ведь у других людей тоже накопилось много базовых вопросов.

Рассмотрим основные вопросы, возникшие в головах у американцев во время речи Обамы в Конгрессе в 2014 году[42].

Не только тебе было интересно: Основные вопросы, задаваемые в Google во время речи в Конгрессе

Сколько лет Обаме?

Кто сидит рядом с Байденом*?

Почему у Бейнера** зеленый галстук?

Почему Бейнер такой оранжевый?

Вы можете прочитать эти вопросы, и, думаю, они не в пользу нашей демократии. Видно, что людей больше волнует цвет чьего-то галстука или тона кожи, чем содержание речи президента и то, как она отразится на нас. Незнание, кто такой Джон Бейнер (в то время спикер Палаты представителей), тоже не очень хорошо говорит о нашем участии в политической жизни страны.

Я предпочитаю думать о таких вопросах как о демонстрации мудрости наших учителей. Это те темы, которые люди обычно не поднимают, опасаясь, что они слишком глупые. Но многие задаются ими и спрашивают у Google.

---

* Джо Байден — американский политик, член Демократической партии, был вице-президентом США во время президентства Обамы. — *Прим. ред.*

** Джон Бейнер — американский политик, член Республиканской партии, был спикером палаты представителей Конгресса США при Обаме. — *Прим. ред.*

На самом деле, полагаю, большие данные помогут обновить знаменитое выражение «Никогда не сравнивайте то, что внутри вас, с тем, что другие выставляют напоказ». В новой интерпретации это может звучать так: «Никогда не сравнивайте свои поисковые запросы в Google с постами других людей в социальных сетях».

Сравните описания женами своих мужей в соцсетях и в анонимных поисковых запросах.

Как чаще всего женщины описывают своих мужей

Посты в социальных сетях	В поисковых запросах
Самый лучший	Гей
Мой лучший друг	Придурок
Удивительный	Удивительный
Самый замечательный	Раздражающий
Такой милый	Посредственность

Читая посты других людей в соцсетях, но не их поисковые запросы, мы склонны преувеличивать число женщин, считающих своих мужей «лучшими», «замечательными» и «такими милыми»*. Мы склонны минимизировать количество женщин, уверяющих, что их мужья «ничтожества», «раздражают» и «придурки». Анализи-

* Я проанализировал данные в Twitter (благодарю Эмму Пирсон за помощь в их получении). В результаты не включено описание действий мужа прямо сейчас, что распространено в социальных сетях, но не соответствует действительности. Ведь даже эти описания сдвинуты в сторону желаемых действий — «работает» или «готовит». — *Прим. авт.*

руя же анонимные агрегированные данные, легко обнаружить, что мы — не единственные, кто считает брак, да и саму жизнь, довольно сложными. Вот теперь можно перестать сравнивать наши поисковые запросы с постами других людей в социальных сетях.

Второе преимущество цифровой сыворотки правды заключается в выявлении страдающих людей. Компания «Human Rights Campaign» попросила меня о сотрудничестве — помочь предоставить мужчинам в некоторых штатах возможность открыто заявить о том, что они геи. Для решения, куда лучше направить свои ресурсы, Human Rights хочет использовать анонимные агрегированные данные поиска Google. Аналогично, со мной связалась служба защиты детей — узнать, в какой части страны с детьми могут обращаться гораздо более жестоко, чем показывают официальные документы.

Есть удивительная тема, о которой я уже упоминал: вагинальные запахи. Когда я впервые писал о ней в «Нью-Йорк Таймс», то делал это в ироническом ключе. Тогда тема вызвала у меня желание похихикать.

Однако когда я начал изучать некоторые электронные доски объявлений и поисковые запросы, выяснилось: там было много постов молодых девушек, убежденных, что их жизнь оказалась разрушена из-за беспокойства по поводу вагинального запаха. Это не шутка. Со мной связались специалисты по половому воспитанию и поинтересовались, как наилучшим образом учесть некоторые данные из интернета для уменьшения паранойи среди молодых девушек.

Хотя я чувствовал себя немного не в теме при обсуждении всех этих вопросов, они были очень серьезны. И, считаю, мои научные данные способны им помочь.

И, что самое главное, эта цифровая сыворотка правды действительно способна привести нас от проблем к решениям. Когда нам многое становится более понятным, мы можем найти способы уменьшения проблем в общемировом масштабе.

Давайте вернемся к речи Обамы об исламофобии. Напомню, каждый раз, когда Обама утверждал, что мы должны больше уважать мусульман, те самые люди, которых он пытался убедить в толерантности, разъярялись все больше.

Тем не менее поисковые запросы в Google показывают: в его послании была одна строчка, сыгравшая роль триггера, запускающего отклик, на который и рассчитывал Обама. Он сказал: «Американские мусульмане — это наши друзья и наши соседи, наши сослуживцы, наши спортивные герои. И, да, они наши мужчины и женщины в военной форме, которые готовы умереть, защищая нашу страну».

После этой строки впервые более чем за год участились поисковые запросы в Google, в которых после слова «мусульмане» не набирали «террористы», «экстремисты» или «беженцы». Теперь там следовали «спортсмены», а затем — «солдаты». Причем «спортсмены» занимали первое место в течение суток после выступления президента.

Когда мы поучаем озлобленных людей, поисковые запросы показывают, что их ярость может только возрасти. Но тонко провоцируя любопытство людей, подбрасывая им новую информацию и предлагая новые образы разжигающей их ярость группы, можно повернуть мысли в другом, более позитивном направлении.

Через два месяца после той речи Обама выступил с еще одним телеобращением по вопросу исламофобии — на сей раз из мечети[43]. Возможно, кто-то в канцелярии президента прочитал в «Times» мою колонку, где обсуждалось, что сработало в предыдущем обращении, а что — нет. Как бы то ни было, содержание этой речи заметно отличалось от предыдущей.

Обама очень недолго говорил о важности толерантности. Вместо этого он в основном провоцировал любопытство людей и старался изменить их восприятие мусульман-американцев. Многие рабы из Африки были мусульманами, говорил Обама. У Томаса Джефферсона и Джона Адамса были собственные экземпляры Корана. Первая мечеть на территории США была построена в штате Северная Дакота. Американец-мусульманин проектировал небоскребы в Чикаго. Обама вновь говорил о мусульманских спортсменах и военнослужащих, а затем упомянул о мусульманах-полицейских и пожарных, учителях и врачах.

Мой анализ поисковых запросов в Google показал: эта речь была более успешной, чем предыдущая. В течение нескольких часов после послания президента значительно сократилось число поисковых запросов, полных ненависти.

Для понимания причин ненависти и способов снижения ее накала существуют и другие варианты использования статистики поисков. Например, мы можем посмотреть, как изменится число расистских запросов после того, как в местной команде появится черный квотербек. Или как изменятся сексистские запросы после того, как на высокую официальную должность будет избрана

женщина. Мы в состоянии отследить, как расизм реагирует на изменение общественного порядка, а сексизм — на новые законы о сексуальных домогательствах.

Также может быть полезно изучение предрассудков, скрывающихся в глубинах нашего подсознания. Например, мы могли бы прилагать дополнительные усилия, чтобы радоваться уму маленьких девочек и проявлять меньше беспокойства по поводу их внешности. Данные, полученные на основе поисковых запросов в Google и других достоверных источников в интернете, дают нам беспрецедентную возможность заглянуть в самые темные уголки человеческой психики. И порой, признаюсь, это бывает довольно тяжело. Но это подразумевает и расширение возможностей. Мы в состоянии использовать данные для борьбы с тьмой. Получение как можно большего массива информации о мировых проблемах — первый шаг к их устранению.

Глава 5

ПРИГЛЯДИМСЯ ПОВНИМАТЕЛЬНЕЕ

Мой брат Ной на четыре года младше меня. Большинство людей при первой встрече с нами говорят, что мы очень похожи. Мы оба слишком громко говорим, одинаково лысеем и с большим трудом сохраняем порядок в своих квартирах.

Но есть и различия. Я прижимист, а Ной покупает все самое лучшее. Я люблю Леонарда Коэна и Боба Дилана, а Ной — Cake и Beck.

Пожалуй, самым заметным отличием между нами является наше отношение к бейсболу. Я обожаю его и в частности «Нью-Йорк Метс». Ной же находит бейсбол невероятно скучным, и его ненависть к спорту уже давно стала неотъемлемой частью его личности*.

---

* Откроем секрет: когда я подбирал материалы для этой книги, Ной отрицал свою ненависть к такому типично американскому времяпрепровождению. Он признает, что ненавидит бейсбол, но считает, что доброта, любовь к детям и интеллект являются основными элементами его личности. И что негативное отношение к бейсболу не входит даже в первую десятку его основных характеристик. Однако я пришел к выводу: порой нам бывает трудно разобраться в себе. И объективно, как сторонний наблюдатель, я вижу, что ненависть к бейсболу — это действительно фундаментальная составляющая характера Ноя, в состоянии он признать это или нет. Так что я оставил его в покое. — *Прим. авт.*

Сет Стивенс-Давидовиц, обожает бейсбол

Ной Стивенс-Давидовиц, ненавидит бейсбол

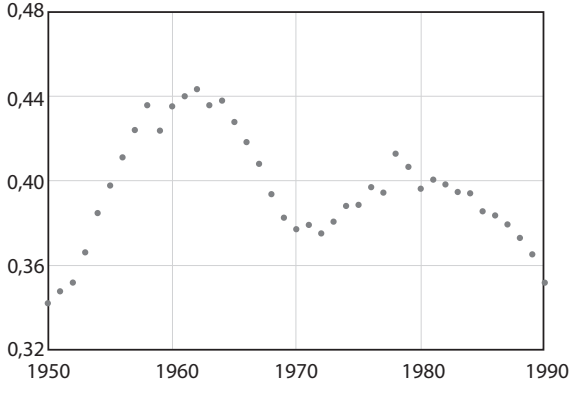
Как получилось, что двое парней со столь сходными генами, воспитанные одними и теми же родителями в одном и том же городе, имеют такие противоположные чувства к бейсболу? Что определяет, какими взрослыми мы вырастем? Или более принципиально: что не так с Ноем? Лежит ли ответ в области психологии развития, которая собирает, отфильтровывает и анализирует большие массивы баз данных взрослого человека, а также сопоставляет их с ключевыми событиями детства? Это может помочь нам решить данный, а также смежные вопросы. Такой процесс можно было бы назвать расширенным использованием больших данных.

Чтобы увидеть, как это работает, давайте рассмотрим одно из проведенных мной исследований. Оно касалось того, как детские впечатления влияют на выбор бейсбольной команды[1], за которую вы болеете — или готовы ли вы вообще переживать за какую-либо команду. Для этого исследования я использовал статистику Facebook о «лайках» бейсбольных команд. В предыдущей главе я отметил, что данные Facebook могут быть далеки от реальности во всем, что касается деликатных

вопросов. В данном же случае я полагаю, что никто, даже фанаты Филли*, не стесняются признаваться на Facebook в интересе к определенной команде.

Для начала я проанализировал возраст мужчин, лайкавших странички одной из двух нью-йоркских бейсбольных команд. Вот график, отражающий число болельщиков «Метс» в зависимости от года их рождения.

Процент болельщиков «Метс» в зависимости от года их рождения

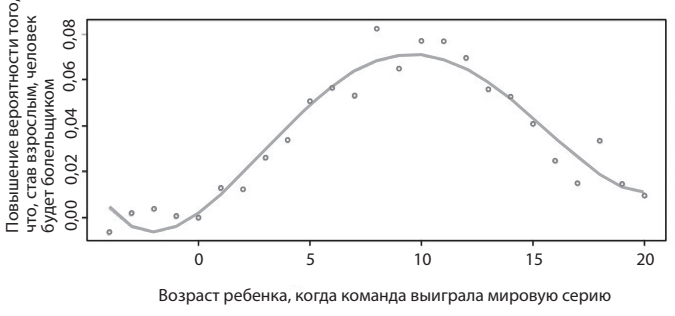


Чем выше точка, тем больше поклонников у команды. Ее популярность растет и падает, потом снова возрастает и опять падает. «Метс» были очень известны среди родившихся в 1962 и 1978 годах. Полагаю, поклонники бейсбола понимают, в чем тут дело. «Метс» выиграли всего две мировые серии — в 1969 и 1986 годах. В те годы мужчинам 1962 и 1978 г. р. было примерно по семь-восемь лет. Таким

* Шутливое название города Филадельфии и местной бейсбольной команды. — *Прим. ред.*

образом, важным прогностическим фактором симпатии к «Метс» — по крайней мере среди мальчиков — является тот факт, выиграла ли команда Мировую серию, когда им было около семи или восьми лет, или нет.

На самом деле мы можем расширить этот анализ. Я изучил информацию на Facebook, показывающую, сколько фанатов различных команд разного возраста поставили лайки любой из широкого выбора команд MLS* и обнаружил, что существует необычно большое количество мужчин 1962 года рождения, болеющих за «Балтимор Ориолс». И мужчин 1963 года рождения, поддерживающих «Питсбург Пайрэтс». Когда указанные команды были чемпионами, эти люди были восьмилетними мальчиками. Вычислив возрастной пик поклонников всех изученных мной команд, я составил следующий график:



В очередной раз мы видим: самый важный год в жизни мужчины[2], определяющий выбор его любимой бейсбольной команды во взрослом возрасте — плюс-минус

* Главная бейсбольная лига в США. — *Прим. ред.*

восемь лет. В целом, возраст от 5 до 15 является ключевым периодом для покорения сердца мальчика. В 19 или 20 лет этот показатель сокращается до 1/8 от пикового значения. И как раз в это время вопрос решится окончательно: либо парень полюбит какую-то команду на всю жизнь, либо вообще не будет интересоваться этим видом спорта.

Вы можете спросить: а что насчет женщин — любительниц бейсбола? Здесь зависимость выражена намного менее ярко, а возрастной пик, похоже, приходится на возраст около 22 лет.

Это мое любимое исследование. Оно касается двух моих самых обожаемых тем — бейсбола и источников моей взрослой неудовлетворенности. Я прочно подсел на бейсбол в 1986 году и страдал в одиночестве, болея за «Метс» — и болею до сих пор. Ной родился четыре года спустя и был избавлен от этого.

Сегодня бейсбол — не самая важная тема в мире, по крайней мере так неоднократно говорил мне мой консультант, доктор философии. Но моя методика может помочь нам решить подобные вопросы — в том числе показать, как люди формируют свои политические или сексуальные предпочтения, музыкальные вкусы и финансовые привычки. (Мне было бы особенно интересно узнать о происхождении сумасшедших идей моего брата в области последних двух тем.) Полагаю, мы увидим, что многие из наших взрослых привычек и интересов (даже те, которые мы считаем основополагающими) могут быть объяснены произвольными фактами — датой нашего рождения или тем, что происходило в те несколько ключевых лет, пока мы были молоды.

Конечно, подобные работы уже проводились, они касались происхождения политических предпочтений. Яир Гитца, главный научный сотрудник компании «Catalist», занимающейся анализом данных, и Эндрю Гельман, политолог, статистик Колумбийского университета, пытались проверить расхожую мысль о том, что большинство людей начинают с либеральных идей, но с возрастом скатываются в консерватизм. Это мнение выражено известной цитатой, часто приписываемой Уинстону Черчиллю: «Если человеку еще нет 30 и он не либерал, значит, у него нет сердца; если человеку уже за 30 и он не консерватор, значит у него нет мозгов».

Гитца и Гельман потратили 60 лет на обработку данных исследований, включивших в себя более 300 тысяч наблюдений за предпочтениями избирателей. И обнаружили, вопреки утверждению Черчилля, что подростки иногда придерживаются либеральных взглядов, а иногда — консервативных. То же самое касается людей среднего возраста и пожилых.

Их работа явно продемонстрировала, что политические взгляды на самом деле формируются точно так же, как и спортивные предпочтения. Есть критический период, оказывающий решающее влияние на всю остальную жизнь. Между 14 и 24 годами множество американцев формируют свое мнение, основываясь на славе президента. Популярный республиканец или непопулярный демократ — и многие молодые люди станут республиканцами. Наоборот — и очередное поколение пополнит колонны демократов. И эти взгляды, обретенные в ключевой период, у большинства американцев останутся на всю жизнь.

Чтобы увидеть, как это работает, сравните предпочтения американцев, родившихся в 1941 году, и тех, кто родился десятилетие спустя.

Представители первой группы достигли совершеннолетия во время президентства популярного республиканца Дуайта Эйзенхауэра. В начале 1960-х, несмотря на то, что этим людям было под 30, они в основном голосовали за представителя Республиканской партии. И даже старея, представители этого поколения постоянно склонялись к поддержке республиканцев.

Американцы, родившиеся на 10 лет позже — бебибумеры, — достигли совершеннолетия во время президентства Джона Ф. Кеннеди, чрезвычайно популярного демократа. Линдон Джонсон был изначально прославленным демократом. Ричард Никсон являлся республиканцем, который в конечном счете ушел в отставку с позором. Представители этого поколения всю свою жизнь склонялись к либеральному образу мыслей.

Имея все эти данные, исследователи смогли определить самый важный возраст для выработки политических взглядов — 18 лет.

Они обнаружили, что этот эффект импринтинга очень важен. Их модель предполагает, что в результате президентства Эйзенхауэра число республиканцев, родившихся в 1941 году, увеличилось на 10%. Кеннеди, Джонсон и Никсон увеличили количество демократов среди американцев, родившихся в 1952 году, на 7%.

Я дал понять, что скептически отношусь к данным исследования, но меня впечатляет количество рассмотренных откликов. В действительности подобная работа не могла быть сделана на основании одного небольшого

опроса. Чтобы увидеть, как именно меняются предпочтения с возрастом, ученым нужны были сотни тысяч наблюдений и обобщения многих исследований.

Для моего анализа любви или нелюбви к бейсболу объем данных также имел решающее значение. Мне нужно было узнать не только количество болельщиков каждой команды, но и разбить их по возрасту. Для этого требуются миллионы наблюдений — и Facebook вместе с другими цифровыми источниками способны предоставить нам подобную информацию.

Здесь вступает в игру объем исследуемой статистики. Нужно иметь много пикселей в фотографии, чтобы можно было увеличить четкость изображения одной ее малой части. Аналогично, необходимо много наблюдений в общем массиве данных для того, чтобы иметь возможность увеличить четкость одного небольшого подмножества — например, сказать, насколько популярна команда «Метс» среди мужчин 1978 года рождения. Небольшой опрос пары тысяч человек не будет достаточно большой выборкой.

Это третье достоинство больших данных: они позволяют рассмотреть вблизи мелкие сегменты большого массива — чтобы получить новую информацию о том, кто мы есть. Можем присмотреться и к другим параметрам помимо возраста. Если у нас есть достаточно информации, мы в состоянии понять, как ведут себя люди, живущие в определенных городах и поселках. Мы можем посмотреть даже, как они действуют ежечасно и ежеминутно.

В этой главе мы пристально посмотрим на поведение людей.

ЧТО НА САМОМ ДЕЛЕ ПРОИСХОДИТ В НАШИХ РЕГИОНАХ, ГОРОДАХ И ПОСЕЛКАХ?

Оглядываясь назад, все кажется удивительным. Но когда Радж Четти, ставший затем профессором в Гарварде, и его небольшая исследовательская группа впервые изучили довольно большой набор данных — налоговые записи всех американцев с 1996 года, — они не были уверены в какой-либо его пользе. Налоговая передала им эту информацию, поскольку ее руководство сочло, что исследователи могли бы использовать ее для прояснения последствий налоговой политики.

Первоначальные попытки Четти и его команды использовать эту статистику заводили их в многочисленные тупики. Их анализ последствий Федеральной налоговой политики и налоговой политики штатов приводил в основном к тем же выводам, которые получали все остальные исследователи, работавшие только с этой информацией. Возможно, результаты Четти, использовавшего сотни миллионов единиц данных налоговой службы, были немного более точными. Но получение практически такого же результата, как и у остальных, не является серьезным достижением социальной науки. Это не тот тип работы, о котором готовы писать в лучших научных журналах.

Более того, организация сбора и анализ всех данных налоговой службы занял много времени. Четти и его команда, потонув в информации, потратили на получение тех же результатов даже больше времени, чем все остальные ученые.

Стало казаться, что люди, скептически относившиеся к идее больших данных, были правы. Не нужно перелопачивать данные сотен миллионов американцев, чтобы разобраться в налоговой политике — опроса десяти тысяч человек оказалось бы вполне достаточно. Четти и его команда были, естественно, обескуражены.

И вот наконец ученые поняли свою ошибку. «Это не простое исследование, которое основано на большем массиве данных», — объясняет Четти[3]. Исследователи задавали слишком мало вопросов относительно данных, которые им были переданы. «Большие данные позволяют вам использовать совершенно другие конструкции, отличные от тех, которые применялись при опросах, — добавляет Четти. — Можно, например, более внимательно отнестись к географии распределения данных».

Другими словами, имея информацию о сотнях миллионов людей, Четти и его команда смогли определить закономерности, относящиеся к городам и различным регионам — большим и малым.

Будучи аспирантом Гарварда, я был в конференц-зале, когда Четти представил свои первые результаты, пользуясь данными налогового учета каждого американца. Социологи обращаются в своем творчестве к наблюдениям: сколько элементов у них имеется. Если социолог работает с опросом 800 человек, он говорит: «У нас восемь сотен наблюдений». Если он работает с лабораторным экспериментом, в котором принимали участие 70 человек, он скажет: «У нас есть семьдесят наблюдений».

«У нас есть 1,2 миллиарда наблюдений», — сказал Четти. Зрители нервно хихикнули.

И Четти с соавторами начали — сначала в конференц-зале, а затем в серии статей — демонстрировать нам важные новые выводы о жизни американского общества.

Рассмотрим такой вопрос: является ли Америка страной больших возможностей? Есть ли у вас шанс сколотить состояние, если ваши родители небогаты?

Традиционный способ ответа на этот вопрос — посмотреть на репрезентативную выборку американцев и сравнить ее с аналогичной статистикой других стран.

Вот данные по разным странам о равенстве возможностей. Был задан вопрос: какова вероятность того, что человек с родителями, входящими в 20% самых бедных жителей страны, попадет в 20% людей с наиболее высокими доходами?

Шансы человека с бедными родителями стать богатым (некоторые страны)

США	7,5
Великобритания	9,0
Дания	11,7
Канада	13,5

Как видите, у США *не самый высокий* результат.

Но в этом простом анализе не хватает конкретики. Команда Четти подобрала материалы по регионам и обнаружила, что шансы разбогатеть сильно различаются в зависимости от того, в какой части страны вы родились.

Шансы человека с бедными родителями стать богатым (по регионам США)	
Сан-Хосе, Калифорния	12,9
округ Вашингтон	10,5
Среднее значение по США	*7,5*
Чикаго, Иллинойс	6,5
Шарлотт, Северная Каролина	4,4

В некоторых частях Соединенных Штатов шанс бедного ребенка преуспеть равен шансу в любой развитой стране мира. В других частях США вероятность того, что бедный ребенок станет богатым, ниже, чем в любой развитой стране мира.

Эти результаты никогда не были бы получены при небольшом опросе, который содержал бы данные лишь о нескольких людях из Шарлотт и Сан-Хосе. Естественно, это не позволило бы создать такую разбивку по регионам, которую сделала команда Четти.

На самом деле ученые смогли еще более конкретизировать разбиение по географическому признаку. Поскольку они обладали столь большим массивом данных — информацией о каждом американце в стране, — то умудрились учесть даже небольшие группы людей, мигрировавших из города в город. И смогли понять, как это может повлиять на перспективы переехавших из Нью-Йорка в Лос-Анджелес, из Милуоки в Атланту, из Сан-Хосе в Шарлотт. Это позволило им проверить причины и следствия, а не только корреляцию (различия между этими понятиями мы обсудим в следующей главе). И, да — переезд

в «правильный» город в годы формирования личности значительно повлиял на конечный результат.

Так как, является ли Америка «страной больших возможностей»?

Ответ: ни да, ни нет. Некоторые регионы таковыми являются, а некоторые нет.

Как пишут авторы, «США лучше описывать как совокупность обществ, некоторые из которых являются «страной больших возможностей» с высоким уровнем мобильности в зависимости от поколения, а в других лишь небольшому числу детей удается выбраться из нищеты».

Так что можно сказать о тех частях Соединенных Штатов, где существует высокая мобильность доходов? Что делает некоторые места страны лучше, позволяя бедному ребенку добиться лучших условий жизни? Территории, где тратится больше средств на образование, предоставляют больше шансов. В местах с более религиозным населением и более низким уровнем преступности у детей также больше возможностей выбраться из нищеты. А вот регионы с большим количеством чернокожего населения уменьшают этот шанс. Что интересно, это относится не только к чернокожим детям, но и к живущим там белым. В местах с большим количеством матерей-одиночек ситуация хуже. Там этот эффект сказывается не только на детях одиноких матерей, но и на их ровесниках, растущих в полных семьях. Некоторые из полученных результатов свидетельствуют о несомненной важности окружения ребенка, его сверстников. Если у его друзей сложный семейный фон и мало возможностей, для избежания нищеты ему придется больше бороться.

Данные говорят нам о том, что некоторые регионы США обеспечивают детям больший шанс вырваться из нищеты. А в каких областях у людей больше шансов избежать встречи со «старухой с косой»[4]?

Мы предпочитаем думать, что смерть уравнивает всех. Никто не может ее избежать — ни нищий, ни король, ни бездомный, ни Марк Цукерберг. Все умрут.

Но если богатые не могут избежать смерти, они по крайней мере в состоянии отсрочить ее приход. Американские женщины, входящие в группу 1% людей с наиболее высоким доходом, в среднем живут на 10 лет дольше, чем американские женщины из 1% людей с наиболее низким доходом. У мужчин этот разрыв достигает 15 лет.

Как результаты различаются в разных регионах США? Зависит ли ваша продолжительность жизни от того, где вы живете? Разнятся ли эти данные для богатых и бедных? Команда Раджа Четти нашла ответы на эти вопросы — опять же за счет увеличения объема данных и разбивки их по географическому признаку.

Интересно, что средняя продолжительность жизни богатых американцев почти не зависит от того, где они живут. Если у вас есть излишек денег, вы можете ожидать, что проживете примерно 89 лет будучи женщиной или около 87 лет, если вы мужчина. Богатые люди везде стремятся развивать у себя здоровые привычки. В среднем они больше тренируются, лучше питаются, меньше курят и реже страдают от ожирения. Богатые могут позволить себе беговую дорожку, органические авокадо, занятия йогой. И они могут купить это все в любом уголке Америки.

У бедных история другая. Продолжительность жизни самых нищих американцев существенно варьируется в зависимости от того, где они живут. В самом деле, если обитать в подходящем месте, можно добавить пяток лет к продолжительности жизни бедного человека.

Так почему же в некоторых местах бедняки могут жить настолько дольше? Что такого есть в этих городах?

Вот четыре характеристики города. Три из них не коррелируют с продолжительностью жизни бедных, но одна связана с ней. Посмотрите, сможете ли вы догадаться, какая именно?

Что позволяет бедному человеку прожить в определенном городе значительно дольше?

Жители города значительно более религиозны.

В городе низкий уровень загрязнения.

В городе высокий процент жителей имеют медицинские страховки.

В городе живет много богатых людей.

Первые три — религия, окружающая среда и медицинское страхование — не коррелируют с продолжительностью жизни бедных. Переменная, имеющая решающее значение, по данным Четти и других исследователей, — число богатых людей, живущих в городе. Чем их больше, тем дольше живут и бедняки. Например, в Нью-Йорке они живут намного дольше, чем в Детройте.

Почему же наличие богачей является таким мощным фактором продолжительности жизни бедных людей?

Одну из гипотез — с которой можно поспорить — выдвинул Дэвид Катлер, один из авторов исследования и один из моих советчиков. Причиной может быть заразное поведение.

Существует большое количество исследований, показывающих, что привычки заразны[5]. Бедняки, живущие рядом с богачами, могут перенять у них ряд привычек. Некоторые из них — скажем, пафосная лексика — не могут оказать влияния на здоровье. Но другие — например физические тренировки — способны создать положительный эффект. Действительно, бедные люди, живущие рядом с богатыми, работают больше, меньше курят и реже страдают от ожирения.

Мне особенно нравится одно исследование команды Раджа Четти, получившей доступ к массиву данных налоговой инспекции. Ученые разобрались, почему одни люди уходят от налогов, а другие нет[6]. Объяснить причины этого явления немного сложнее.

Очень важно знать о существовании простого способа для самозанятых людей, имеющих одного ребенка, максимизировать сумму денег, получаемых от правительства. Если вы сообщите, что ваш налогооблагаемый доход ровно 9000 долларов в год, государство выпишет вам чек на 1377 долларов — эта сумма отражает скидку с подоходного налога, своего рода грант для работающих бедняков. Если вы сообщите о более высоком доходе, сумма налогов немедленно увеличится. Сообщите о меньшей сумме, и налоговый вычет уменьшится. Налогооблагаемый доход в размере 9000 долларов — это самый выгодный вариант.

И как ни странно, самозанятые люди с одним ребенком чаще всего сообщают о доходе именно в 9000 долларов.

Неужели эти американцы скорректировали свои графики работы специально для получения идеальной суммы дохода? Нет конечно. Когда таких работников проверяли в случайном порядке — очень редкое явление, — почти всегда выяснялось, что они заработали либо существенно меньше 9000 долларов, либо существенно больше.

Другими словами, они жульничали с налогами, делая вид, что заработали именно ту сумму, которая обеспечит им самый большой чек от государства.

Насколько типичным был этот вид налогового мошенничества и кто среди самозанятых людей с одним ребенком скорее всего его совершал? Согласно данным Четти и его коллег, распространенность этого вида жульничества очень сильно колебалась в зависимости от региона. Среди людей этой категории о заработке в 9000 долларов в Майами сообщили 30%, тогда как в Филадельфии — всего 2%.

Что нам укажет на обманщиков? Что известно о местах с наибольшим количеством подобных мошенников и о местах, где их меньше всего? Мы можем сопоставить уровни приписок с другими демографическими параметрами различных городов. В результате получается, что существует два мощных прогностических фактора: высокая концентрация в регионе людей, профессия которых предполагает налоговые льготы, и высокая концентрация налоговых специалистов.

На что указывают эти факторы? Четти и его команда дают объяснение. Ключевым катализатором подобного ухода от налогов была информация.

Большинство самозанятых налогоплательщиков с одним ребенком просто не знают, что магическое число для получения наибольшей суммы от государства — 9000 долларов. Но живя рядом с теми, кто знает — это могут быть соседи или налоговый консультант, — у них резко возрастают шансы пронюхать об этом.

На самом деле команда Четти нашла еще больше доказательств того, что знания ведут к обману. Когда американцы переезжали из региона с невысоким уровнем подобного мошенничества в район с довольно значительным, они перенимали этот трюк. Постепенно обман распространился от региона к региону по всей территории США. Уход от налогов оказался заразным, как вирус.

Теперь остановитесь на минуту и подумайте о том, насколько интересные результаты дало это исследование. Оно показало, что, когда речь пойдет об афере с налогами, самым важным будет не определить, кто честный, а кто нечестный. Самым важным будет понять, кто знает, как мухлевать, а кто — нет.

Поэтому когда кто-то говорит вам, что никогда не уклоняется от уплаты налогов, есть довольно высокая вероятность, что он врет. Исследование Четти показывает: многие мошенничали бы, если бы знали как.

Если вы хотите сжульничать со своими налогами (но я вам не рекомендую этого делать), вы должны жить рядом со специалистом в области налогообложения или быть соседом налоговых мошенников, которые могут подсказать вам, что нужно делать.

Если вы хотите, чтобы ваши дети стали знаменитостями, где вы должны жить?

Большие данные предоставляют возможность более пристально взглянуть на мир и получить действительно детализированный ответ на любой поставленный вопрос — и этот тоже.

Мне было любопытно, откуда приезжают самые успешные американцы, и вот однажды я решил ознакомиться с «Википедией»[7]. (Сегодня вы тоже можете это сделать.)

Немного программирования, и вот у меня есть набор данных о более чем 150 тысячах американцев, которых редакторы «Википедии» сочли достаточно заметными и достойными попасть в эту базу данных. Информация включала место и дату рождения, профессию и пол. Я соединил ее с региональными сведениями о рождении, собранными Национальным центром статистики департамента здравоохранения. После чего подсчитал шансы на попадание в Википедию людей, родившихся в каждом графстве США.

Можно ли сказать, что упоминание в «Википедии» является заметным достижением? Конечно, имеются некоторые ограничения. Редакторы «Википедии» больше внимания обращали на молодых мужчин, что может вызвать смещение выборки. А некоторых персонажей нельзя считать особо достойными. Например, Тед Банди попал в «Википедию», потому что убил десятки молодых женщин. В результате мне пришлось удалить преступников, что, впрочем, не оказало существенного влияния на результаты.

Я ограничил исследование беби-бумерами (людьми, рожденными в период между 1946 и 1964 годами), потому что у них было время на то, чтобы проявить себя — практически, целая жизнь. Примерно один из 2058 рожденных в Америке беби-бумеров был сочтен достаточно заметным, чтобы попасть в «Википедию». Около 30% — за достижения в области искусства или развлечения, 29% — спортсмены, 9% — политики и 3% — за научные результаты.

Первый поразительный факт, который я заметил — огромная географическая изменчивость вероятности достижения большого успеха, по крайней мере по меркам «Википедии». Ваши шансы стать заметной фигурой в значительной степени зависят от места вашего рождения.

В «Википедию» попал примерно один из 1209 беби-бумеров, рожденных в Калифорнии. Тогда как уроженцев Западной Виргинии там — один из 4496 беби-бумеров. При разбиении по округам результаты становятся еще более красноречивыми. До упоминания в «Википедии» добрался примерно один из 748 беби-бумеров, родившихся в графстве Саффолк, штат Массачусетс. А в некоторых других штатах процент успеха был в 20 раз меньшим.

Почему оказалось, что в некоторых частях страны гораздо легче штамповать влиятельных людей? Я внимательно осмотрел лучшие округа и выяснил, что почти все они вписываются в одну из двух категорий.

Во-первых — и это меня удивило, — во многих из этих регионов имеется большой студенческий городок[8]. Почти каждый раз, когда я видел название графства (например Уоштено, штат Мичиган), я узнавал, что там имеется классический университетский городок, в данном

случае — Энн-Арбор. В верхние 3% попадают такие округа, как: Мэдисон, штат Висконсин; Афины, штат Джорджия; Коламбия, штат Миссури; Беркли, штат Калифорния; Чапел-Хилл, штат Северная Каролина; Гейнсвилл, штат Флорида; Лексингтон, штат Кентукки; Итака, штат Нью-Йорк.

Почему так? Некоторые из попавших в «Википедию» людей вполне могут быть сыновьями и дочерьми преподавателей и аспирантов. Последние, как правило, достаточно умны (черта, которая в борьбе за большой успех может быть весьма полезной). И действительно, большое число выпускников колледжей в регионе является мощным прогностическим фактором успеха родившихся там людей.

Но, скорее всего, влияние имеет и еще кое-что — раннее приобщение ко всему новому. В городках, где располагаются колледжи, особенно хорошо развивается все, связанное с музыкой. У ребенка в университетском городке будет больше возможностей попасть на уникальные концерты, услышать передачи необычных радиостанций, там есть даже независимые музыкальные магазины. И дело не ограничивается искусством. Университетские городки поставляют довольно большой процент заметных бизнесменов. Возможно, раннее знакомство с передовым искусством и идеями способствует развитию умения организовывать бизнес.

Успех университетских городков касается не только регионов. Он связан и с расами тоже. Афроамериканцы недостаточно представлены в «Википедии» (за исключением спортсменов) — особенно, если говорить о бизнесе и науке. Это, несомненно, сильно связано с дискриминацией. Но в одном маленьком графстве, где 84% населения

1950 года рождения — черные, родилось почти столько же заметных беби-бумеров, сколько в графствах с наибольших процентом людей, упомянутых в «Википедии».

Из менее чем 13000 беби-бумеров, рожденных в округе Мэкон, штат Алабама, 15 попали в «Википедию» — или один из 852. И каждый из них — чернокожий. 14 из них были из города Таскджи, в котором расположился университет Таскджи, исторически «черный» колледж, основанный Букером Т. Вашингтоном. В списке выходцев из этого региона присутствуют судьи, писатели и ученые. На самом деле, черный ребенок, родившийся в Таскджи, имел такую же вероятность стать заметным не только в спорте, как и белый ребенок, родившийся в одном из городов с университетом, в котором учатся в основном белые.

Во-вторых, скорее всего, в «Википедию» попадут уроженцы округа, включающего в себя большой город. Наиболее высокая вероятность попадания в «Википедию» у тех, кто родился в Сан-Франциско, Лос-Анджелесе или Нью-Йорке. (Я объединил пять округов Нью-Йорка вместе, поскольку во многих статьях «Википедии» не указан район рождения.)

Урбанистические регионы, как правило, являются элементом модели успеха. Чтобы оценить значение возможности быть в молодости рядом с успешными профессионалами, сравните Нью-Йорк, Бостон и Лос-Анджелес. Среди них первый производит больше всего журналистов самого высокого уровня, второй — самых заметных ученых, а третий — самых знаменитых актеров. Помните, мы говорим о людях, которые родились там, а не переехали туда. И это справедливо даже после вычитания людей, чьи родители проявили себя в той же области.

Графства, в которых нет крупных городов с колледжами, демонстрируют гораздо худшие результаты, чем городские округа.

Мои родители, как и многие беби-бумеры, переехали от людных тротуаров к зеленым улицам — в моем случае из Манхэттена в округ Берген, Нью-Джерси. Потенциально это было ошибочным решением — по крайней мере с точки зрения воспитания детей-знаменитостей. Ребенок, рожденный в Нью-Йорке, на 80% вероятнее окажется в «Википедии», чем тот, кто появился на свет в графстве Берген. Это всего лишь корреляция, но можно сказать, что взросление рядом с великими идеями лучше, чем жизнь на большом заднем дворе.

Выявленный эффект мог бы быть даже сильнее, если бы у меня имелось больше сведений о том, где все эти люди жили в детстве — ведь многие из них выросли отнюдь не в том штате, где родились.

Успех университетских городков и больших городов поражает уже при беглом ознакомлении с данными. Но я копнул глубже и провел более сложный эмпирический анализ.

Это позволило мне вычислить существование еще одной переменной, ставшей сильным прогностическим фактором, способствующим занесению имени человека в «Википедию». Речь о доле иммигрантов в стране вашего рождения. Чем выше в регионе процент граждан, родившихся в другой стране, тем больше вероятность, что ребенок, появившийся там на свет, добьется заметного успеха. (Вот тебе, Дональд Трамп!) Если два места являются одинаковыми с точки зрения городского ландшафта и наличия колледжа, то из региона с большим числом иммигрантов выйдет больше выдающихся американцев. Почему?

Многие известные люди были детьми иммигрантов. Я сделал исчерпывающий обзор биографий 100 самых знаменитых белых беби-бумеров (по данным проекта Массачусетского технологического института «Пантеон», который также работает с данными «Википедии»). Большинство из них были работниками искусства. По крайней мере 13 родились у матерей-иммигранток — в том числе Оливер Стоун, Сандра Баллок и Джулианна Мур. Этот показатель более чем в три раза выше, чем в среднем по стране за этот период. (Многие имеют отцов-иммигрантов — в том числе Стив Джобс и Джон Белуши, — но эти данные трудно сравнивать со средними по стране, поскольку информация об отцах не всегда включается в свидетельства о рождении.)

А что насчет переменных, не влияющих на успех? Одна, которую я нашел, кажется довольно удивительной: неважно, сколько денег штат тратит на образование. В штатах со схожим процентом городских жителей расходы на образование никак не коррелируют с числом выросших там известных писателей, художников или руководителей предприятий.

Интересно сравнить мое изучение данных «Википедии» с одним из исследований, о котором уже говорилось ранее — команда Четти пыталась выяснить, какие регионы позволяют людям достичь верхней грани среднего класса. Я же попытался выяснить, какие области помогают им достичь славы. Результаты разительно отличаются.

Большие затраты на образование помогают детям достичь верхней грани среднего класса, но совсем не способствуют тому, чтобы они стали известными писателями, художниками или бизнес-лидерами. Многие из тех,

кто добился заметных успехов, ненавидели школу, а некоторые даже бросили учебу.

Как выяснила команда Четти, Нью-Йорк — не самое лучшее место для воспитания ребенка, если вы хотите, чтобы он достиг вершины среднего класса. А мое исследование показало, что это отличное место, если вы хотите дать ему шанс на славу.

Когда вы смотрите на факторы, обеспечивающие признание, существенные различия между регионами начинают обретать смысл. Многие штаты сочетают в себе все основные составляющие успеха. Вернемся к Бостону. Многочисленные университеты делают этот город котлом, в котором кипят инновационные идеи. В этом регионе живет множество чрезвычайно успешных людей, являющих отличный пример достижения успеха для молодежи. И это привлекает иммигрантов, чьи дети просто вынуждены использовать эти уроки.

Но что если область не имеет ни одного из этих качеств? Значит ли это, что ей суждено будет «выращивать» меньше суперзвезд? Не обязательно. Есть и другой путь: крайне узкая специализация. Отличным примером может быть округ Розо в штате Миннесота — небольшой сельский регион с малым количеством «понаехавших» и крупных вузов. Примерно один из 740 человек, родившихся здесь, оказался затем в «Википедии». Их секрет? Все девять были профессиональными хоккеистами, чему, несомненно, способствовало наличие хоккейных программ в местных школах и колледжах.

Так что, если вы не особенно жаждете стать звездой хоккея, но хотите обеспечить своим будущим детям все возможные преимущества, может, стоит переехать

в Бостон или Таскджи? Им это не повредит. Но есть и более серьезные вещи. Как правило, экономисты и социологи сосредотачиваются на том, как избежать негативных последствий — таких, как нищета и преступность. Но великая цель, стоящая перед обществом — не только подтянуть отстающих. Важно помочь как можно большему количеству людей выделиться. Возможно, как раз усилия по определению мест рождения сотен тысяч самых известных американцев и помогут создать какие-то первоначальные стратегии — в частности поощрение иммиграции, субсидирование университетов и поддержку искусства.

Обычно я изучаю данные по США. Поэтому, когда я пристально рассматриваю географическую информацию, то отбираю ее по нашим городам и поселкам — по таким регионам, как округ Мэйкон, Алабама или округ Розо, Миннесота. Но еще одно огромное (и все возрастающее) преимущество данных из интернета заключается в том, что подобным же образом можно легко собрать информацию со всего мира — и посмотреть, как и в чем различаются страны. А ученые, занимающиеся сбором и анализом данных, получают возможность прокрасться в антропологию.

Недавно я исследовал довольно необычный вопрос: как протекает беременность в разных странах мира? Я проверил число запросов в Google о беременных женщинах. И первым делом обнаружил поразительное сходство физических симптомов, на которые жалуются женщины.

Я проанализировал, как часто различные симптомы соединяются в поисковых запросах со словом

«беременна». Например, как часто вместе с «беременностью» искали «тошноту», «боли в спине» или «запор»? В Канаде и в Соединенных Штатах симптомы были очень схожи. В таких странах, как Великобритания, Австралия и Индия, они тоже были примерно одинаковы.

Похоже, беременные женщины во всем мире жаждут одного и того же. В США в Google чаще всего делают поиск по словам «хочется есть лед во время беременности». Следующие четыре варианта — желание съесть соленое, сладкое, фрукты и острую пищу. В Австралии список продуктов, о которых мечтают беременные, не очень отличается: соль, сладости, шоколад, мороженое и фрукты. А что насчет Индии? Похожая история: острая пища, сладости, шоколад, соль и мороженое. На самом деле пятерка желаемых продуктов практически одинакова во всех странах, информацию по которым я просмотрел.

Предварительные данные свидетельствуют: ни в одной части мира нет диеты или среды, где бы существенно менялось физическое ощущение беременности.

Но мысли, окружающие беременность, варьируются весьма значительно.

Начните с вопроса о том, что могут безопасно делать беременные женщины. Самые частые запросы в США: могут ли беременные женщины «есть креветки», «пить вино», «пить кофе» или «принимать “Тайленол”»?

Когда дело доходит до подобных обеспокоенностей, другие страны имеют мало общего как с Соединенными Штатами, так и друг с другом. Вариант с вином не входит в первую десятку вопросов в Канаде, Австралии и Великобритании. Проблемы на Зеленом континенте в основном связаны с употреблением во время беременности

молочных продуктов — особенно сливочного сыра. А в Нигерии, где интернетом пользуются 30% населения, самый частый вопрос — можно ли беременным пить холодную воду?

Реальны ли эти опасения? Когда как. Есть убедительные доказательства того, что беременные женщины подвергаются повышенному риску заражения листериями из непастеризованного сыра. Было установлено, что употребление слишком большого количества алкоголя негативно влияет на ребенка. В некоторых частях мира считается, что, когда мать пьет холодную воду, у плода может начаться пневмония — правда, я не знаю ни одного медицинского подтверждения этого факта.

Огромные различия в вопросах из разных стран мира, скорее всего, вызваны неиссякающим потоком информации, поступающим из разрозненных источников в каждой из них: официальные научные исследования, околонаучные изыскания, бабушкины сказки и обычный треп. Женщинам трудно определить, на что следует обращать повышенное внимание и, соответственно, о чем спрашивать у Google.

Глядя на популярные запросы типа «как… во время беременности», мы видим четкую разницу между странами. В Соединенных Штатах, Австралии и Канаде больше всего вопросов с текстом: «Как предотвратить растяжки во время беременности». А в Гане, Индии и Нигерии предотвращение растяжек даже не входит в пятерку основных проблем. Там женщины, как правило, больше озабочены занятием сексом или сном.

Пять самых частых поисковых запросов типа «как... во время беременности»

США	Индия	Австралия	Великобритания	Нигерия	ЮАР
Предотвратить растяжки	Спать	Предотвратить растяжки	Сбросить вес	Заниматься сексом	Заниматься сексом
Сбросить вес	Заниматься сексом	Сбросить вес	Предотвратить растяжки	Сбросить вес	Сбросить вес
Заниматься сексом	Иметь секс	Избежать растяжек	Избежать растяжек	Заниматься любовью	Предотвратить растяжки
Избежать растяжек	Секс	Спать	Спать	Сохранить здоровье	Спать
Остаться в форме	Беречь себя	Заниматься сексом	Заниматься сексом	Справиться с тошнотой	Справиться с тошнотой

Пятерка наиболее популярных поисковых запросов, начинающихся со слов «Может ли беременная женщина...»

США	Есть мелкие креветки	Пить вино	Пить кофе	Принимать «Тайленол»	Есть суши
Великобритания	Есть крупные креветки	Есть копченого лосося	Есть чизкейк	Есть моцареллу	Есть майонез
Австралия	Есть сливочный сыр	Есть крупные креветки	Есть бекон	Есть сметану	Есть сыр фета
Нигерия	Пить холодную воду	Пить вино	Пить кофе	Заниматься сексом	Есть моринга (съедобные растения)
Сингапур	Пить зеленый чай	Есть мороженое	Есть дуриан	Пить кофе	Есть ананас
Испания	Есть паштет	Есть хамон	Принимать парацетамол (обезболивающее)	Есть тунца	Загорать
Германия	Летать	Есть салями	Посещать сауну	Есть мед	Есть моцареллу
Бразилия	Красить волосы	Принимать Dipirona (болеутоляющее)	Принимать парацетамол	Ездить на велосипеде	Летать

Несомненно, рассматривая различные выборки данных, можно узнать намного больше о здоровье и культуре в разных уголках мира. Но мой предварительный анализ показывает: когда дело доходит до выхода за пределы нашей биологии, большие данные продемонстрируют нам, что люди даже менее сильны, чем мы думали.

КАК МЫ ЗАПОЛНЯЕМ ЧАСЫ И МИНУТЫ ЖИЗНИ

«Приключения молодого человека, основные интересы которого — изнасилования, особо яростное насилие и Бетховен».

Это было похоже на рекламу скандального фильма Стэнли Кубрика «Заводной апельсин». По сценарию, вымышленный молодой герой Алекс Делардж с пугающей отрешенностью совершал шокирующие акты насилия. В одной из самых известных сцен фильма он насиловал женщину, во все горло распевая «Поющие под дождем».

Почти сразу появились сообщения о подражателях. Действительно, группа мужчин изнасиловала 17-летнюю девушку, распевая именно эту песню. Фильм был запрещен к показу во многих европейских странах, и некоторые из наиболее шокирующих сцен были удалены из версии, показанной в Америке.

На самом деле есть много примеров того, как люди в реальной жизни подражают искусству[9] — когда мужчины, казалось, были просто загипнотизированы увиденным на экране только что. После показа фильма о бандитах «*Цвета*» произошла серьезная перестрелка. После показа фильма «*Нью-Джек-Сити*» последовали массовые беспорядки.

Возможно, наиболее тревожным оказался случай, когда через четыре дня после выхода фильма «Денежный поезд» мужчины использовали жидкость из зажигалок, чтобы поджечь кассу в метро — практически точно имитируя сцену, увиденную в кино. Единственное различие между вымышленным и реальным поджогами: в кино кассир сбежал, тогда как в реальной жизни он сгорел.

Существуют также некоторые свидетельства, полученные на основании психологических экспериментов: люди, посмотревшие фильм с жестокими сценами, выказывают больше гнева и враждебности, даже не имитируя точно ни одну из увиденных сцен[10].

Другими словами, рассказы и эксперименты показывают, что жестокие фильмы провоцируют агрессивное поведение. Но насколько велик эффект? Мы говорим об одном-двух убийствах в 10 лет или о сотнях каждый год? Рассказы и эксперименты не могут ответить на этот вопрос.

Чтобы понять, могут ли помочь в этом большие данные, два экономиста — Гордон Даль и Стефано делла Винья — слили воедино три больших набора данных за период с 1995 по 2004 год: ежечасные сведения ФБР о преступлениях, цифры кассовых сборов и степень насилия во всех фильмах с kids-in-mind.com.

Использованная ими информация была исчерпывающей — каждый фильм и каждое преступление, совершенное в каждый час по всей территории Соединенных Штатов. Это могло бы дать очень важные доказательства.

Ключом их исследования было то, что в одни выходные самый популярный фильм был очень жестоким[11], например «Ганнибал» или «Рассвет мертвецов», а в другие

выходные — позитивным, таким как «Сбежавшая невеста» или «История игрушек».

Экономисты могли точно сказать, сколько убийств, изнасилований и нападений было совершено в дни, когда показывали жестокий фильм, — и сравнить эти цифры с количеством убийств, изнасилований и нападений за выходные, когда показывали веселое, спокойное кино.

Так что же они выяснили? Увеличивалась ли преступность после жестоких фильмов, как предполагали некоторые экспериментаторы? Или оставалась прежней?

Экономисты обнаружили, что после показа популярного жестокого фильма преступность сокращалась.

Вы не ошиблись. По выходным, когда шел популярный жестокий фильм и миллионы американцев следили за людьми, убивающими других людей, число преступлений значительно сокращалось.

Когда вы получите этот странный и неожиданный результат, вашей первой мыслью будет: «Что я сделал неправильно?» Каждый экспериментатор тщательно все проверил. Никаких ошибок. Вторая мысль: «Есть какая-то переменная, объясняющая такие результаты?» Ученые проверили, не влияет ли на выводы время года. Нет. Они собрали данные о погоде, думая, что, возможно, она имеет значение. Нет, и она ни при чем.

«Мы проверили все предположения, все, что мы делали, — сказал мне Даль. — И не смогли найти никаких ошибок».

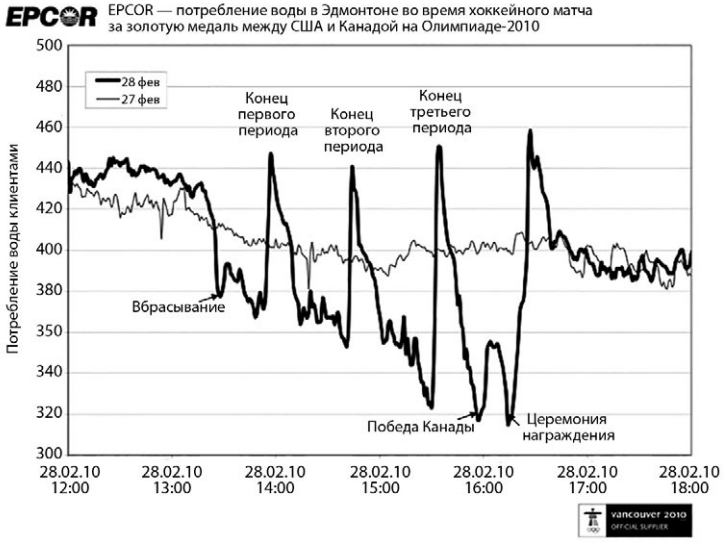
Несмотря на слухи, несмотря на лабораторные эксперименты, какими бы неожиданными ни казались результаты, демонстрация жестокости в фильмах вызывала резкое снижение уровня преступности. Как такое могло быть?

Чтобы найти ключ к разгадке, Даль и делла Винья решили использовать большие данные, проанализировав их повнимательнее. Традиционно результаты опросов дают информацию ежегодно или, в лучшем случае, ежемесячно. С толикой везения удалось бы получать данные по выходным дням. Теперь сравните: используя комплексные наборы данных, а не малые выборки (опросы), мы смогли составить почасовые и даже поминутные графики. Это позволило узнать о человеческом поведении намного больше.

Иногда, если это не жизненно важная информация, временные колебания даже забавны. EPCOR — коммунальная компания в Эдмонтоне, Канада, раскрыла поминутные данные о потреблении воды во время хоккейного матча за золотую медаль между США и Канадой на Олимпиаде-2010 (предположительно его смотрели 80% канадцев). Статистика говорит нам, что вскоре после завершения каждого периода потребление воды резко возрастало — туалеты в Эдмонтоне явно работали с максимальной нагрузкой.

Можно получить даже поминутные данные о поиске в Google[12], при этом откроются некоторые интересные закономерности. Например, число поисковых запросов «разблокировать игру» резко увеличивается в 8 утра по будням и достигает максимума в 3 часа дня — это, несомненно, ответ на попытки школ заблокировать доступ

к мобильным играм на своей территории без запрета на работу сотовых телефонов учащихся.



Количество поисковых запросов со словами «погода», «молитва» и «новости» достигает максимума около 5:30 утра — это доказывает, что большинство людей просыпаются гораздо раньше меня. Число поисковых запросов со словом «самоубийство» достигает пика в 12:36 дня, а минимума — около 9 утра. Это доказывает, что большинство людей утром гораздо менее несчастны, чем я.

Статистика показывает, что время между 2 и 4 часами утра — не лучшее для решения главных вопросов бытия. В чем смысл сознания? Существует ли свобода воли? Есть ли жизнь на других планетах? Популярность этих вопросов поздно ночью может быть результатом, в частности, использования каннабиса. Пик поисков с текстом «Как забить косяк?» приходится на период между 1 и 2 часами ночи.

Имея огромный набор данных, Даль и делла Винья смогли понять, как меняется уровень преступности по часам в те выходные, когда показывают фильмы. Они обнаружили, что снижение преступности в те выходные, когда были показаны фильмы с насилием — относительно других выходных, — начинается в самом начале вечера. Другими словами, преступность шла на убыль до начала показа жестоких сцен, когда люди еще только шли к кинотеатрам.

Можете догадаться, почему? Прежде всего, подумайте о тех, кто, скорее всего, предпочтет пойти смотреть жестокий фильм. Это молодые — особенно молодые — агрессивные мужчины.

Затем следует подумать о том, где обычно совершаются преступления. Это редко происходит в кинотеатре. Бывали исключения — в том числе в 2012 году, когда произошла стрельба в кинотеатре в Колорадо. Но, по большому счету, мужчины ходят в кинотеатры безоружными и сидят молча.

Предоставьте молодым агрессивным мужчинам шанс увидеть Ганнибала, и они пойдут в кино. Предоставьте молодым агрессивным мужчинам возможность посмотреть фильм «Сбежавшая невеста», и они откажутся. А вместо этого, возможно, пойдут в бар, клуб или в бильярдный зал, где уровень преступности потенциально выше.

Жестокие фильмы удерживают агрессивных людей от пребывания на улицах.

Головоломка решена? Не совсем. Статистика показала еще одну странность. Обозначенный эффект стартовал с началом показа фильмов, но не заканчивался

с окончанием лент, когда кинотеатр закрывался. В те вечера преступность была ниже и позже — с полуночи до 6 утра.

Даже если она снижалась в то время, когда молодые люди находились в кинотеатре, что ей мешало усилиться после того, как они выходили оттуда и больше не были ничем заняты? Ведь они только что посмотрели жестокий фильм, который, как показывают эксперименты, делает людей более злыми и агрессивными.

Вы в состоянии придумать какие-либо объяснения этому феномену? После долгих раздумий, исследователей — экспертов по преступности озарило. Они знали, что алкоголь является одной из основных причин преступности[13]. Кроме того, они не раз бывали в кинотеатрах США, поэтому знали, что там практически не продаются спиртные напитки. Действительно, ученые обнаружили: количество преступлений, связанных с употреблением алкоголя, в вечерние часы после жестоких фильмов снизилось.

Конечно, исследования Даля и делла Винья имели определенные ограничения. Ученые, например, не могли протестировать длительный эффект — чтобы понять, как долго может продолжаться снижение уровня преступности. Возможно, последовательное воздействие ряда жестоких фильмов в конечном счете приводит к еще большему насилию. Однако их исследование оценивает непосредственное влияние таких лент, что и было главной темой экспериментов. Вероятно, жестокий фильм влияет на некоторых людей и делает их необычайно злыми и агрессивными. Однако знаете ли вы, что именно совершенно точно негативно влияет на людей? Общение

с другими потенциально жестокими людьми. И пьянство*.

Сейчас это обрело смысл, которого, казалось, не было до того момента, пока Даль и делла Винья не начали анализ огромной горы данных[15].

Когда мы начинаем рассматривать информацию более детально, становится понятным еще один важный момент: мир сложен. Действия, предпринимаемые нами сегодня, могут иметь отдаленные последствия, большинство из которых непредсказуемы. Идеи распространяются — иногда медленно, а иногда экспоненциально,

---

* Эта история показывает, как нечто, кажущееся плохим, может оказаться хорошим, если мешает чему-то еще худшему. Эд Маккаффри, бывший принимающий игрок, получивший образование в Стэнфорде, использует этот аргумент для оправдания решения отправить всех четырех своих сыновей играть в футбол[14]: «У этих ребят море энергии, так что, если они не играют в футбол, то гоняют на скейтборде, лазают по деревьям, играют в салки во дворе или в пейнтбол. Я имею в виду, они не собираются сидеть и ничего не делать. Поэтому я смотрю на это таким вот образом. В футболе, по крайней мере, есть определенные правила… Мои дети уже бывали в отделении неотложной помощи — после падений с крыши, с велосипеда, со скейтборда и с деревьев. Думаю, вы понимаете… Да, спорт — это довольно суровое занятие. Но, кроме того, с моими ребятами там будет тренер и они, по крайней мере, не будут белками прыгать с горы и совершать разные безумства. Полагаю, это просто направление агрессии в организованное русло». Я никогда не слышал ничего похожего на аргументы Маккаффри, высказанные в интервью в передаче «*The Herd with Colin Cowherd*». А после прочтения статьи Даля и делла Винья я начал воспринимать эти мысли более серьезно. Преимуществом больших массивов реальных, а не лабораторных данных является то, что они могут выявить подобные эффекты. — *Прим. авт.*

как вирусы. Люди реагируют на стимулы самым непредсказуемым образом.

Эти связи и отношения, эти всплески и затухания не могут быть отслежены маленькими опросиками или другими традиционными способами получения и обработки данных. Мир слишком сложен и слишком многообразен для того, чтобы понять его с помощью небольших объемов информации.

НАШИ ДВОЙНИКИ

В июне 2009 года Дэвид «Биг Папи» Ортис с удовольствием смотрел на дело рук своих. За предыдущие полтора десятилетия Бостон влюбился в своего здоровяка родом из Доминиканской республики с дружелюбной улыбкой и щелью между зубами*.

Он принял участие в пяти победных играх Всех звезд, выиграл приз MVP** и помог Бостону впервые за 86 лет победить в чемпионате. Но в 2008 году, когда ему стукнуло 32, его успешная карьера явно подходила к концу. Его средний уровень упал на 68 пунктов, его процент пребывания на базе стал равен 76 очкам, а процент сильных ударов составил 114 очков. В начале сезона 2009 года результаты Ортиса упали еще ниже.

Вот как Билл Симмонс, спортивный журналист и страстный болельщик «Бостон Ред Сокс», описал происходившее в первые месяцы сезона 2009 года[16]: «Очевидно, что

---

* Дэвид Ортис — американский бейсболист доминиканского происхождения, играл за команду «Бостон Ред Сокс». — *Прим. ред.*

** Ежегодная награда, вручаемая самому выдающемуся игроку в матче Всех звезд Главной лиги бейсбола. — *Прим. ред.*

Дэвид Ортис уже не отличается в игре… Здоровенный бьющий выглядит как порнозвезда, тяжелоатлет, центровой НБА и мечта юных девиц: он сдал». Любители спорта доверяют своим глазам и глазам Симмонса: Ортис закончился. На самом деле Симмонс предсказал, что Ортис в скором времени окажется на скамейке запасных или даже уйдет из спорта.

Действительно ли Ортис закончился? Если бы в 2009 году вы были генеральным менеджером «Сокс», вы бы его убрали? И в более общем плане: как мы можем предсказать успешность бейсболиста в будущем[17]? И еще более обобщенно: как мы можем использовать большие данные для предсказания того, что люди будут делать в будущем?

Теория, которая заведет вас далеко в дебри науки о данных, такова: посмотреть на то, что делали сабермétрики (те, кто использовал данные для изучения бейсбола), и распространить это на другие области науки о сборе и анализе данных. Бейсбол стал одной из первых областей, породивших огромные массивы данных почти обо всем. И существовала целая армия умных людей, готовых посвятить жизнь тому, чтобы понять смысл этих данных. Сейчас почти каждый параметр изучен досконально. Бейсбол проложил дорогу, после него стало проще изучать все остальное.

Самый простой способ предсказать будущее игрока — предположить, будет ли он играть так же, как делает это сейчас. Если парень старался изо всех сил в течение последних полутора лет, можно предположить, что и в ближайшие полтора года он будет прикладывать такие же усилия.

Если следовать этой методологии, Бостон должен был попрощаться с Дэвидом Ортисом.

Однако есть и более актуальная информация. В 1980-х Билл Джеймс, которого многие считают основателем сaберметрики, подчеркнул важность возраста. Он обнаружил, что бейсболисты достигают расцвета достаточно рано — примерно к 27 годам. Но команды, как правило, игнорируют последующее снижение их активности и переплачивают за стареющих игроков.

Согласно этой более передовой методике оценки, «Сокс» нужно было обязательно убрать Дэвида Ортиса.

Но из-за привязки к возрасту можно что-то упустить. Не у всех игроков карьера протекает одинаково. Некоторые могут закончиться в 23, другие — в 32. Низкие бейсболисты стареют иначе, чем высокие, карьера толстых отличается от карьеры тощих. Бейсбольные статистики обнаружили: существуют различные типы игроков, каждый из которых стареет по-своему. Подобное распределение также не в пользу Ортиса: «здоровенные бьющие» действительно, в среднем, достигают пика раньше[18] и заканчивают карьеру вскоре после 30.

Если «Сокс» оценит его недавние матчи, возраст и физические параметры, администрация, без сомнения, должна разорвать контракт с Дэвидом Ортисом.

В 2003 году статистик Нейт Сильвер представил новую модель для прогнозирования результативности игрока, которую назвал PECOTA. Она оказалась лучшей — и самой крутой. Сильвер искал двойников бейсболистов. Вот как это работает. Нейт создал базу данных всех значительных игроков бейсбольной Лиги за все время — это более 18 тысяч человек. В нее была включена вся информация, которую удалось собрать: рост, возраст, телосложение, положение в команде, количество хоумранов,

средний уровень пробежек и число аутов за каждый год карьеры. Теперь нужно было найти 20 игроков, карьера которых была бы больше всего похожа на карьеру Ортиса — тех, кто играл примерно как он в свои 24, 25, 26, 27, 28, 29, 30, 31, 32 и 33 года. Другими словами, найти двойников. А потом посмотреть, как в дальнейшем развивались их карьеры.

Поиск двойников — это еще один пример использования детализации. Он фокусируется на небольшой группе людей, наиболее похожих на данного человека. И, как и любая детализация, результат получается тем точнее, чем больше данных у вас есть. Оказывается, двойники Ортиса[19] выдали совсем другой прогноз на будущее самого Ортиса. Среди них были Хорхе Посада и Джим Томе. Эти парни начинали свои карьеры немного медленно, а затем следовали удивительные всплески результативности. Около 30 лет они достигли уровня мирового класса, а затем, в первые годы после 30, потихоньку сдавали.

Тут-то Сильвер и предсказал, как сложится карьера Ортиса — на основании судеб его двойников. Он обнаружил, что те восстановили свои силы. В отношении поклонниц Симмонс, возможно, был бы прав. Но что касается двойников Ортиса, то здесь все иначе — они сдали, но затем вернулись.

Поиск двойника — лучшая методика, когда-либо использовавшаяся для прогноза результативности бейсболиста. Согласно ей, «Сокс» должны были потерпеть. Клуб действительно не стал рубить сплеча. И в 2010 году средняя результативность Ортиса выросла до 270. Он совершил 32 хоумрана и вошел в сборную Всех звезд. А затем входил в нее еще четыре года подряд.

В 2013 году, играя на своей традиционной позиции назначенного отбивающего, в возрасте 37 Ортис набрал 0,688 очка, а Бостон победил Сент-Луис в Мировой серии — 4:2.

Ортис был признан MVP (наиболее ценным игроком) Мировой серии*.

Едва дочитав статью о подходе Нейта Сильвера к оценке результативности игрока, я сразу же начал думать о том, может ли и у меня тоже быть двойник.

Поиск такового является перспективным во многих областях, а не только в спорте. Мог бы я найти человека, разделяющего мои интересы? Может быть, если бы я нашел кого-то, больше всего похожего на меня, мы могли бы проводить время вместе. Может быть, он бы знал рестораны, которые могли бы мне понравиться. Возможно, он мог бы познакомить меня с тем, чего я не знаю, и я бы заинтересовался этим.

Поиск двойников возможен даже по особенностям личности. И, как и любая детализация, сходство будет тем сильнее, чем больше у вас данных. Предположим, я буду искать двойника в наборе данных десяти человек. Я мог бы найти кого-то, кто разделяет мой интерес

---

* Читая эту часть книги, вы, наверное, можете сказать, что я склонен цинично относиться к историям с хорошим концом. На самом деле я хотел, чтобы эта история выглядела исключительно хорошо, поэтому убрал свой цинизм в сноску. Подозреваю, с помощью PECOTA просто обнаружилось, что Ортис употреблял стероиды, затем в какой-то момент перестал это делать, а затем начал снова. С точки зрения прогнозирования, это действительно круто, если PECOTA оказалась в состоянии докопаться до подобного. Правда, тогда история становится куда менее трогательной. — *Прим. авт.*

к книгам. Предположим, я буду искать двойника в наборе данных тысяч людей. Я мог бы найти кого-то, кому, как и мне, нравятся популярные книги о здоровье. Но предположим, что я буду искать двойника в наборе данных сотен миллионов людей. Тогда я мог бы найти кого-то, кто действительно похож на меня.

Однажды я провел поиск двойника в социальных сетях. Используя весь массив профилей Twitter, я искал людей, имеющих больше всего общих интересов со мной.

Вы, конечно, можете многое рассказать о моих интересах на основании информации в моем аккаунте в Twitter. В целом, я подписан примерно на 250 человек, разделяющих мою страсть к спорту, политике, комедиям, науке и мрачным еврейским певцам.

Так есть ли кто-нибудь во Вселенной, так же, как и я, подписанный на все эти 250 аккаунтов, мой твиттер-близнец? Конечно, нет. Двойники не идентичны нам, они лишь похожи на нас. И нет никого, кто подписан хотя бы на те же 200 аккаунтов. Или даже на 150.

Однако в конце концов я нашел аккаунт, подписанный на 100 пользователей из моих 250 — это Country Music Radio Today. Да неужели? Оказывается, Country Music Radio Today — это бот (его уже нет), который подписался на 750 тысяч профилей «Твиттера» в надежде, что они ответят ему тем же.

У меня есть бывшая подруга, которая, как я подозреваю, получила бы удовольствие от такого результата. Однажды она сказала, что я больше похож на робота, чем на человека.

Но шутки в сторону! Тот факт, что моим двойником стал бот, позволяет сделать важный вывод. Чтобы поиск

двойников оказался по-настоящему точным, следует стремиться не просто найти кого-то, любящего то же, что и вы. Нужно искать того, кто не любит то же, что не любите вы.

Мои интересы становятся очевидными не только на основании тех аккаунтов, на которые я подписываюсь, но и тех, которые я не выбираю. Я интересуюсь спортом, политикой, комедиями и наукой, а не едой, модой или театром. Мои подписки показывают, что мне нравится Берни Сандерс, но не Элизабет Уоррен*, Сара Сильверман, но не Эми Шумер**, «*New Yorker*» но не «*Atlantic*»***, мои друзья — Ной, Эмили Сэндс и Джош Готтлиб, но не Сэм Ашер. (Извини, Сэм, но твои посты в Twitter — это скукота.)

Из 200 миллионов аккаунтов в Twitter, у кого профиль похож на мой? Оказалось, мой двойник — пишущий для *Vox***** Дилан Мэтьюз. Это стало большим разочарованием с точки зрения улучшения использования социальных сетей, ведь я уже и так подписан на аккаунт Дилана в Twitter и Facebook и постоянно читаю его статьи в *Vox*. Поэтому знание о том, что именно он является моим двойником, ничего в моей жизни не изменило. Но это довольно круто — узнать о существовании человека, больше всех в мире похожего на вас. Особенно, если это кто-то, кем вы восхищаетесь. И когда я закончу

---

* Бернард Сандерс и Элизабет Уоррен — американские политики, принадлежат к демократической партии. — *Прим. ред.*

** Сара Сильверман и Эми Шумер — американские актрисы, выступающие в жанре стенд-ап. — *Прим. ред.*

*** Два старейших литературных журнала в США. — *Прим. ред.*

**** www.vox.com, американский новостной сайт. — *Прим. ред.*

эту книгу и перестану жить отшельником, может быть, мы с Мэтьюзом сможем общаться и обсуждать сочинения Джеймса Суровецки.

Поиск двойника Ортиса был важен для многих поклонников бейсбола, а поиск моего двойника был интересен только мне. Что еще могут показать такие исследования? Прежде всего, с помощью подобных данных многие крупнейшие интернет-компании стараются улучшить свои услуги и работу с пользователями. Amazon использует что-то вроде поиска двойников для вычисления книг, которые вы хотели бы купить. Там видят, что именно выбирают люди с вашими параметрами, и основывают на этом свои рекомендации.

Pandora делает то же самое, определяя, какие песни вы хотите слушать. Примерно так же Netflix узнает, какие фильмы вы хотели бы посмотреть. Результат получился просто ошеломляющим. Когда инженер Amazon Грег Линден в первый раз использовал поиск двойников для предсказания предпочтений читателей, и рекомендации оказались настолько точными, основатель Amazon Джефф Безос пал перед Линденом на колени с воплем: «Я тебя не достоин!»

Но самое интересное в поиске двойников не то, что он сейчас используется почти повсеместно, а то, что он часто не используется. Есть несколько крупных областей, работа которых может быть значительно улучшена путем персонализации. Возьмите, например, наше здоровье.

Исаак Коган, ученый и исследователь из Гарварда, пытается воплотить этот принцип в медицине. Он хочет собрать и организовать всю нашу медицинскую информацию так, чтобы вместо использования одинакового

подхода ко всем, врачи подыскивали бы похожих на вас пациентов. Затем они могли бы использовать более персонализированную диагностику и более целенаправленное лечение.

Коган считает это естественным развитием медицины, и даже не особо радикальным. «Что такое диагноз? — спрашивает он. — Диагноз, по сути, является утверждением, что вы оказались в той же ситуации, как и множество ранее изученных людей. Если я, не дай бог, диагностирую у вас инфаркт, то скажу, что у вас та же патофизиология, которую я уже видел у других людей с сердечным приступом».

Диагноз, по сути, является примитивным вариантом поиска двойника. Проблема в том, что наборы данных, которые используют врачи для его постановки, слишком маленькие. Сегодня диагноз основывается на опыте доктора, лечившего своих пациентов, и он может быть дополнен данными из научных статей о популяциях, с которыми работали другие исследователи. Как мы видели, поиск двойника может стать действительно полезной штукой — необходимо только, чтобы он включал в себя намного большую статистику.

Вот область, в которой большие данные на самом деле могут помочь. Так почему же на внедрение метода требуется столько времени? Почему он до сих пор широко не используется? Проблема заключается в сборе информации. Большинство медицинских заключений по-прежнему существуют только на бумаге и похоронены в папках. А те, которые оцифрованы, часто не могут быть использованы вследствие несовместимых форматов. «Мы нередко имеем больше информации о бейсболе,

чем о здоровье», — говорит Коган[20]. Но простые меры порой идут длинными путями. Ученый неоднократно говорил о «низко висящих плодах». Например, он считает, что даже просто создание базы данных, включающей информацию о росте и весе детей, а также обо всех возможных детских болезнях, стало бы революционным развитием педиатрии. После этого развитие каждого ребенка можно было бы сравнить с развитием любого другого ребенка. Компьютер помог бы найти детей, развитие которых идет по уже пройденному кем-то пути и автоматически предупредил бы обо всех тревожных моментах. Например, он был бы в состоянии обнаружить преждевременный рост ребенка, что в некоторых случаях может указывать на две возможные причины: гипотиреоз или опухоль мозга. Ранняя диагностика в обоих случаях принесет огромную пользу. «Подобные заболевания возникают достаточно редко — примерно одно на десять тысяч, — говорит Коган. — В остальном эти дети здоровы. Думаю, мы могли бы диагностировать болезнь раньше по крайней мере на год. Стопроцентно смогли бы».

Джеймс Хейвуд[21] — предприниматель, использующий другой подход к решению проблемы объединения медицинских данных. Он создал сайт PatientsLikeMe.com, где люди могут сообщать данные о своих заболеваниях, методах лечения и возникающих побочных эффектах. И Джеймс уже добился большого успеха в отношении ряда болезней.

Его цель заключается в сборе достаточного количества информации о людях со сходными состояниями — чтобы впоследствии каждый мог найти своего двойника по здоровью. Хейвуд надеется, что таким образом можно

будет найти людей нужных возраста и пола, с похожими историей и симптомами — и посмотреть, что им помогло. Это будет совсем другой тип медицины.

ИСТОРИИ, РАССКАЗАННЫЕ ДАННЫМИ

Во многих случаях детализация данных для меня ценнее локального поиска для конкретного исследования, поскольку она предлагает новый способ видения и описания жизненных процессов.

Когда люди узнают, что я — и ученый, занимающийся сбором и анализом данных, и писатель, они иногда делятся каким-либо фактом или результатами опроса. Я часто нахожу эти сведения скучными, обобщенными и лишенными жизни. Они не сообщают мне никаких интересных историй.

Помимо этого, друзья пытались уговорить меня начать читать различные романы и биографии. Но меня это тоже мало интересует. Я всегда спрашиваю себя: «Происходило ли подобное в других ситуациях? Каков более общий принцип?» Их истории кажутся мелкими и непоказательными.

Я попытался изложить в этой книге нечто, на мой взгляд, не имеющее аналогов. Оно основано на данных и цифрах; оно показательно и позволяет заглянуть далеко вперед. И при этом большие данные — настолько обширный материал, что позволяют представить себе описываемых ими конкретных людей. Когда мы составляем поминутный график расхода воды в Эдмонтоне, я вижу, как люди встают с дивана в конце хоккейного периода. Когда мы внимательно изучаем людей, переезжающих

из Филадельфии в Майами и начинающих мухлевать с налогами, я вижу, как они разговаривают со своими соседями и узнают о налоговых трюках. Когда мы детально анализируем статистику о бейсбольных болельщиках разного возраста, я вижу свое детство, детство брата, а также миллионы взрослых мужчин, все еще неистово болеющих за команды, завоевавшие их сердца, когда им было по восемь лет.

Рискуя в очередной раз впасть в пафос, я должен сказать: упомянутые в этой книге экономисты и ученые, занимающиеся сбором и анализом информации, создали не просто новый инструмент, но новый жанр. В этой главе и в большей части этой книги я попытался описать данные — настолько подробные и многочисленные, что позволяют нам добиться предельно точной детализации. Не ограничиваясь информацией о каком-либо конкретном обычном человеке, мы с их помощью все еще можем рассказывать разнообразные и запоминающиеся истории.

Глава 6

ВЕСЬ МИР — ЛАБОРАТОРИЯ

27 февраля 2000 года[1] в кампусе Google в Маунтин-Вью, начинался как обычный день. Светило солнце, велосипедисты крутили педали, массажистки занимались массажем, сотрудники увлажняли себе кожу огуречной водой. И вдруг в этот самый обычный день нескольким инженерам Google пришла в голову идея, оказавшая невероятное влияние на развитие интернета. Разработчики нашли наилучший способ заставить вас переходить на сайты, оставаться на них и возвращаться туда снова.

Прежде чем описывать то, что они сделали, мы должны поговорить о разнице между корреляцией и причинностью — это огромная проблема в области анализа данных, которой мы еще не уделили должного внимания.

СМИ каждый день бомбардируют нас результатами исследований на базе корреляций. Например, мы уже рассказывали, что физическое состояние у умеренно потребляющих алкоголь, как правило, лучше, чем у не умеющих остановиться. То есть наблюдается корреляция.

Значит ли это, что если пить немного, то здоровье улучшится — является ли это причинно-следственной

связью? Пожалуй, нет. Скорее, потреблять алкоголь в небольших дозах людям позволяет как раз хорошее здоровье. Социологи называют это обратной причинно-следственной связью. Или, возможно, существует независимый фактор, приводящий как к нежеланию много пить, так и к хорошему здоровью. Например, если вы проводите много времени с друзьями, это приводит к потреблению алкоголя и крепкому здоровью. Социологи называют это смещением с опущенной переменной.

Но как нам точнее установить причинно-следственную связь? Золотой стандарт — это рандомизированное контролируемое испытание. Вот как это работает. Людей наугад делят на две случайные группы. Одну, рабочую, просят сделать или взять что-то. Другую, контрольную, не просят. После чего наблюдают за реакцией каждой группы. Разница в результатах и является причинно-следственной связью.

Например, чтобы проверить, приводит ли умеренное употребление алкоголя к хорошему здоровью, можно случайным образом выбрать несколько человек. Некоторые из них будут пить один бокал вина в день в течение года, а другие не будут. А затем сравнить их состояние здоровья. Поскольку люди были разбиты на две группы случайным образом, нет никаких оснований ожидать, что в одной из них участники будут более здоровы или более социализированы. Вы можете поверить, что эффект вина совершенно обычен. Рандомизированные контролируемые испытания являются самым надежным доказательством в любой сфере деятельности. Если таблетка успешно прошла такой тест, ее можно начинать

продавать. Если она не может пройти его, ее не будет на аптечных полках.

Подобные эксперименты начинают все чаще использоваться в социальных науках. Эстер Дюфло, французский экономист из Массачусетского технологического института, возглавила кампанию за более широкое распространение таких исследований в экономике развития — области знаний, пытающейся найти наилучшие способы помочь беднейшим людям в мире. Рассмотрим эксперимент Дюфло и ее коллег, посвященный улучшению образования в сельских районах Индии, где более половины учащихся средних школ не могут прочитать простое предложение. Одной из потенциальных причин проблем является нехватка учителей. На данный момент в некоторых школах в сельских районах Индии не хватает более 40% преподавателей.

В чем суть теста Дюфло? Они с коллегами случайным образом разделили школы на две группы. В одной (рабочая группа) в дополнение к базовой заработной плате учителям каждый день платили небольшую сумму — 50 рупий, или около 1,15 долларов. В других преподаватели работали без дополнительной оплаты. Результаты были показательны. Когда учителям доплачивали, они в полтора раза реже пропускали работу[2]. Успеваемость школьников тоже существенно улучшилась — особенно это касалось молодых девушек. К концу эксперимента в школах, где учителям платили за приход на занятия, стало на 7% больше девочек, умеющих читать и писать.

Согласно статье в «*New Yorker*», когда Билл Гейтс узнал[3] о работе Дюфло, он был настолько впечатлен, что сказал ей: «Мы *должны* финансировать вас».

АЗБУКА A/B-ТЕСТИРОВАНИЯ

Итак, рандомизированные испытания являются золотым стандартом для доказательства причинно-следственных связей, и их использование распространилось на социальные науки. Теперь вернемся в офис Google в день 27 февраля 2000 года. Благодаря чему тогда произошла революция в интернете?

В тот день несколько инженеров решили провести эксперимент на сайте Google. Они случайным образом разделили пользователей на две группы. В рабочей была показана новая страница результатов поиска с 20 ссылками, а в контрольной — старая, с 10. Затем специалисты сравнили удовлетворенность представителей обеих групп, основываясь на том, как часто они возвращались в Google.

Революция? Поначалу это не казалось столь уж революционным. Я уже отметил, что подобные эксперименты использовались фармацевтическими компаниями и социологами. Так можно ли считать простой их перенос в другую область таким уж большим делом?

Ключевой момент — и это быстро поняли инженеры Google — заключался в том, что эксперименты в виртуальном мире имеют огромное преимущество перед исследованиями в реальном мире. Они так же убедительны, но менее ресурсоемки. По ходу дела Дюфло нужно было общаться со школами, организовать финансирование, платить части учителей и проверять уровень всех учащихся. Реальные эксперименты могут стоить тысячи или сотни тысяч долларов, и на их проведение могут уйти месяцы или годы.

В цифровом мире подобные исследования можно проводить дешево и быстро. Вам не нужно нанимать участников и платить им. Вместо этого можно просто написать строку кода и случайным образом составить группы. Для исследования вам не нужны пользователи — можно измерять перемещения мыши и клики. Нет необходимости вручную писать код и анализировать ответы — можно написать программу, которая будет автоматически делать это за вас. Вам не придется ни с кем связываться. Вам даже не придется объяснять людям, что они являются частью эксперимента.

Это четвертое преимущество больших данных: они позволяют проводить рандомизированные испытания, помогающие гораздо легче находить реальные причинно-следственные связи в любое время и практически в любом месте — важно только наличие доступа в интернет. В эпоху больших данных весь мир — большая лаборатория.

Понимание этого быстро распространилось в Google, а затем по всей Силиконовой долине, где рандомизированные испытания были переименованы в «A/B-тесты». В 2011 году инженеры Google провели семь тысяч A/B-тестов[4], и с тех пор их число только растет.

Если Google хочет знать, как заставить людей кликать на рекламу на его сайтах, компания может использовать в баннерах два оттенка синего: один для группы А, другой для группы Б, а затем сравнить количество кликов. Конечно, простота такого тестирования может привести к злоупотреблениям. Некоторые сотрудники считали, что, поскольку тестирование настолько легкое, Google утонет в экспериментах. В 2009 году

один несостоявшийся дизайнер уволился после того, как в ходе очередного A/B-тестирования был использован 41 незначительно отличающийся оттенок синего[5]. Но протест этого дизайнера против навязчивого исследования конъюнктуры рынка и в поддержку искусства практически не остановил распространение данной методологии.

Сегодня Facebook выполняет[6] тысячи A/B-тестов в день — это означает, что небольшое число инженеров за это время запускают больше рандомизированных контролируемых испытаний, чем вся фармацевтическая отрасль за год.

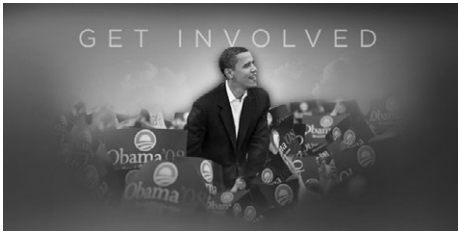
A/B-тестирование распространилось за пределы крупнейших технологических компаний. Бывший сотрудник Google Дэн Сирокер применил эту методику в первой президентской кампании Барака Обамы. Он выполнил A/B-тестирование дизайна главной страницы сайта кампании, полей электронной почты и формы пожертвований. Позже Сирокер основал компанию Optimizely[7], предоставляющую организациям услуги по экспресс-A/B-тестированию. В 2012 году за помощью к Optimizely обратились и Обама, и его соперник Митт Ромни — чтобы максимизировать количество регистраций, добровольцев и пожертвований. Ее услугами пользуются TaskRabbit и журнал «*New York*».

Чтобы понять, насколько ценно подобное тестирование, учтите: Обама использовал его для привлечения большего количества людей в свою предвыборную кампанию. Главная страница сайта президента изначально включала картинку с его изображением и кнопку под ней, приглашавшую людей: «Зарегистрируйтесь».

Было ли это наилучшим способом привлечь людей? С помощью Optimizely команда Обамы могла проверить, не помогут ли другие изображение и кнопка привлечь больше людей. Будут ли люди нажимать на кнопку чаще, если лицо Обамы на фото будет более торжественным? А если на кнопке будет написано: «Присоединяйтесь»? Пользователям были продемонстрированы различные комбинации картинок и кнопок, а затем подсчитано, сколько из них при каком варианте нажали на кнопку. Посмотрите варианты на следующей странице и попробуйте угадать выигрышную комбинацию.

Выиграли фотография семьи Обамы и кнопка «Узнайте больше». Это была победа. При использовании такой комбинации Обама получил на 40% больше зарегистрированных пользователей, что добавило кампании дополнительное финансирование в объеме примерно 60 млн долларов[8].

Тестируемые снимки



Присоединяйтесь к нам сейчас

Узнайте больше

Регистрируйтесь

И еще один большой плюс в том, что подобные тесты можно проводить дешево и легко: это освобождает нас от вечной зависимости от интуиции, которая, как отмечалось в главе 1, имеет свои ограничения. Основная причина важности A/B-тестирования заключается в том, что люди непредсказуемы. Интуиции часто не удается предсказать их реакцию.

Выигрышная комбинация

Было ли ваше шестое чувство право относительно оптимального вида сайта Обамы?

Вот еще несколько проверок для вашей интуиции. «*Boston Globe*» провел А/Б-тесты заголовков[9] — выяснить, какие из них привлекут наибольшее внимание людей, заставив их кликнуть на статью. Попробуйте угадать победителей:

Один из этих заголовков был признан лучшим с точки зрения числа кликов

	Заголовок А	Заголовок Б
1	Может ли дрон SnotBot спасти китов?	Может ли этот дрон помочь спасти китов?
2	Разумеется, «сдутые шарики» — самый распространенный поисковый запрос в Массачусетсе	Этот самый распространенный поисковый запрос в Массачусетсе может смутить
3	Конкурс на лучший перепихон стал причиной судебного процесса об изнасиловании	Никаких обвинений в отношении преподавателей школы после секс-скандала
4	Женщина сорвала куш на редкой бейсбольной карточке	Женщина заработала 179 000 долларов на редкой бейсбольной карточке
5	Ежегодный оперативный дефицит проектов MBTA удвоится к 2020 году	Приготовьтесь: дефицит MBTA почти удвоился
6	Как Массачусетс помог вам получить право на контроль за рождаемостью	Как Бостонский университет помог прекратить «преступления против целомудрия»
7	Когда в Бостоне открыли первое метро	Мультфильмы времени, когда в Бостоне открыли первое метро
8	Жертвы и их семьи на суде по делу об изнасиловании в начальной школе винят во всем современную культуру	Жертвы и их семьи на суде по делу об изнасиловании в начальной школе сделали заявление
	Парень в шапке «Free Brady» единственный сорвал пранк Майли Сайрус	Все восхищены фанатом, узнавшим переодетую Майли Сайрус

Догадались? Ответы приведены ниже

	Заголовок А	Заголовок Б	Победитель?
1	Может ли дрон SnotBot спасти китов?	Может ли этот дрон помочь спасти китов?	На 53% больше кликов на А
2	Разумеется, «сдутые шарики» — самый распространенный поисковый запрос в Массачусетсе	Этот самый распространенный поисковый запрос в Массачусетсе может смутить	На 986% больше кликов на Б
3	Конкурс на лучший перепихон стал причиной судебного процесса об изнасиловании	Никаких обвинений в отношении преподавателей школы после секс-скандала	На 108% больше кликов на Б
4	Женщина сорвала куш на редкой бейсбольной карточке	Женщина заработала 179 000 долларов на редкой бейсбольной карточке	На 38% больше кликов на А
5	Ежегодный оперативный дефицит проектов MBTA удвоится к 2020 году	Приготовьтесь: дефицит MBTA почти удвоился	На 62% больше кликов на Б
6	Как Массачусетс помог вам получить право на контроль за рождаемостью	Как Бостонский университет помог прекратить «преступления против целомудрия»	На 188% больше кликов на Б
7	Когда в Бостоне открыли первое метро	Мультфильмы времени, когда в Бостоне открыли первое метро	На 33% больше кликов на А

8	Жертвы и их семьи на суде по делу об изнасиловании в начальной школе винят во всем современную культуру	Жертвы и их семьи на суде по делу об изнасиловании в начальной школе сделали заявление	На 76% больше кликов на Б
9	Парень в шапке «Free Brady» единственный сорвал пранк Майли Сайрус	Все восхищены фанатом, узнавшим переодетую Майли Сайрус	На 67% больше кликов на Б

Полагаю, вы указали правильно больше половины ответов — возможно, вы рассуждали, на что вам самим захотелось бы кликнуть. Но, скорее всего, вы не все угадали правильно.

Почему? Что вы упустили? Каких выводов о поведении человека вам не хватает? Какие уроки вы извлекли из своих ошибок?

Обычно мы задаем такие вопросы после плохих прогнозов.

Но посмотрите, как трудно делать общие выводы из заголовков «*Globe*». В первом поменяли всего одно слово — «этот» на «SnotBot», что привело к победе. Можно предположить, что более подробное описание нравится людям больше. Но во втором заголовке «сдутые шарики» — конкретный термин — ведет к поражению. В четвертом заголовке «сорвала куш» опережает цифру «179 000 долларов». Это может означать, что выиграть помогает сленг. Но сленговое выражение «конкурс перепихона» проигрывает в третьем заголовке.

Урок, полученный в результате данного A/B-тестирования, в значительной степени отличается от привычных.

Кларк Бенсон[10], генеральный директор сайта новостей и развлечений ranker.com, при выборе заголовков и дизайна ресурса в значительной степени опирается на A/B-тестирование. «В конце дня вы ничего не сможете предложить, — говорит он. — Проверяйте буквально все».

Тестирование заполняет пробелы в нашем понимании человеческой природы. Они всегда будут существовать. Если бы мы, основываясь на жизненном опыте, могли знать точный отклик, тестирование не имело бы смысла. Но нам это недоступно, поэтому приходится использовать тесты.

Еще одна причина важности A/B-тестирования заключается в том, что оно четко показывает: казалось бы, небольшие изменения могут привести к огромным последствиям. Как говорит Бенсон, «меня все время поражает, как мелкие незначительные факторы способны оказать столь огромное влияние».

В декабре 2012 года компания Google изменила свою рекламу: на квадратиках появились направленные вправо стрелки[11].

Обратите внимание, как странно выглядят эти стрелки — они абсолютно ни на что не указывают. В самом деле, когда они появились впервые, многие посетители

Google отнеслись к ним критически[12]. Зачем добавлять бессмысленные стрелки на рекламе, задумались они.

Google защищает свои бизнес-секреты, поэтому ее сотрудники не стали сообщать, насколько ценными были эти стрелки. Единственная информация, которую удалось узнать: этот вариант победил в A/B-тестировании. Причина, по которой Google добавила их, заключатся в том, что очень много людей захотели туда нажать. В результате, это незначительное и, казалось бы, бессмысленное изменение принесло Google и ее рекламным партнерам кучу денег.

А вы можете найти эти небольшие хитрости, которые способствуют получению такой огромной прибыли? Вам следует проверить многие моменты — даже те, которые кажутся тривиальными. На самом деле пользователи Google неоднократно замечали, что реклама сначала чуть-чуть меняется, а затем возвращается к своему прежнему виду. В этих случаях они невольно становятся членами рабочей группы A/B-тестирования, но при этом получают лишь возможность видеть эти незначительные изменения.

Экспериментальное центрирование (не работает)

Лучшие продажи кейсов для iPad 2
The ZAGGmate™ противоударный алюминиевый кейс со встроенной блютус-клавиатурой
www.zagg.com

Эксперимент Green Star (не работает)

Foster's Hollywood Обзор ресторанов, Мадрид, Испания...
www.tripadvisor.co.uk > … > Madrid > Madrid Restaurants ▾ TripAdvisor ▾
★★★★★ Рейтинг 3 — 118 обзоров
Foster's Hollywood, Мадрид: см. 118 беспристрастных обзоров в Foster's Hollywood с рейтингами 3 от 3 до 5 по оценке TripAdvisor. Описаны 3647 из 6489 ресторанов

Эксперимент с новым шрифтом (не работает)

Новости динамичного фондового рынка
Бесплатные графики, новости и советы от экспертов UTVi. Приходите прямо сейчас!
UTVi.com/Stocks

Показанные выше изменения не пошли в дело. Они не сработали. Но они были частью процесса отбора победителей. Дорога к кликабельной стрелке вымощена уродливыми звездами, ошибочным расположением текста и бесполезными шрифтами.

Бывает весело угадывать, что же заставляет людей кликать на кнопки. Если вы демократ, то вам может быть приятно знать, что тестирование принесло Обаме больше денег. Но у А/Б-тестирования есть и темная сторона.

В своей замечательной книге «*Irresistible*» («Непреодолимость») Адам Алтер пишет о создании поведенческих зависимостей в современном обществе[13]. Многие люди сегодня признаются, что им бывает все труднее выключить интернет.

Мой любимый набор данных — поисковые запросы в Google — может дать нам некоторые подсказки относительно того, что люди считают наиболее привлекательным. По статистике Google, большинство зависимостей остаются прежними, такими, с которыми люди боролись в течение многих десятилетий — например, наркотики, секс и алкоголь. Но интернет начинает заявлять о своем присутствии в этом списке все громче — «порно» и Facebook входят в десятку самых распространенных зависимостей.

Самые сильные зависимости в 2016 году, по статистике Google[14]

Наркотики	Алкоголь	Азартные игры
Секс	Сахар	Facebook
Порно	Любовь	

A/B-тестирование может играть важную роль в создании так сильно затягивающей людей сети.

Тристан Харрис, «специалист по этике», слова которого цитируются в «*Irresistible*», объясняет, почему людям так трудно сопротивляться притяжению определенных сайтов: «С другой стороны экрана находятся тысячи людей, чья работа состоит в том, чтобы сломать вашу способность к саморегуляции».

И эти люди используют A/B-тестирование.

С помощью испытаний Facebook может определить, что конкретная кнопка конкретного цвета заставляет людей чаще возвращаться на сайт. Тогда исследователи меняют цвет кнопки. Затем они могут понять, что конкретный шрифт заставляет людей чаще возвращаться на сайт. Поэтому они меняют шрифт. Затем они могут выяснить, что получение людьми электронных писем в определенное время заставляет их приходить на сайт чаще. Поэтому они отправляют электронную почту именно в это время.

Довольно скоро Facebook стал весьма оптимизированным с точки зрения повышения количества времени, которое люди тратят на него.

Другими словами, используйте как можно больше выигравших результатов A/B-тестирования, и вы получите очень привлекательный сайт. Это тот тип обратной связи, которого не было у производителей сигарет.

A/B-тестирование чаще всего является инструментом игровой индустрии. Как говорит Алтер, проводится A/Б-тестирование различных версий игры World of Warcraft. В одной миссии можно убить кого-то. В другой — спасти кого-то. Дизайнеры могут дать игрокам различные образцы различных миссий, а затем посмотреть, какие из них будут пользоваться наибольшей популярностью. Они могли бы увидеть, например, что миссия по спасению кого-то привлекает на 30% больше пользователей. Если они испытывают много, очень много миссий, то начинают находить все больше и больше победителей A/B-тестирования. Эти 30% побед складываются до тех пор, пока не получится игра, в которую играют многие взрослые геймеры.

Если вы немного встревожены этим, знайте — я тоже. И в конце этой книги мы немного поговорим об этических последствиях, а также о других аспектах больших данных. Хорошо это или плохо, но эксперименты — главный инструмент в арсенале ученых. В этом наборе инструментов имеется еще одна форма исследований. Она была использована для того, чтобы задавать различные

вопросы — в том числе о том, насколько эффективно работает телевизионная реклама.

ЖЕСТОКИЕ, НО ПРОЛИВАЮЩИЕ СВЕТ НАТУРНЫЕ ЭКСПЕРИМЕНТЫ

22 января 2012 года «Нью-Инглэнд Пэтриотс» играли полуфинал чемпионата НФЛ* с «Балтимор Рэйвенс».

Оставалась минута игры. «Рэйвенс» проигрывали, но мяч у них. Следующие 60 секунд определят, какая команда сыграет в Суперкубке. Следующие 60 секунд смогут вознести футболистов на вершину славы. И последняя минута этой игры сделает то, что для экономиста гораздо важнее: эти последние 60 секунд помогут наконец понять раз и навсегда — работает ли реклама?

Информация о том, что реклама повышает продажи, очевидно, имеет решающее значение для нашей экономики. Но это безумно трудно доказать. По сути, это хрестоматийный пример того, как сложно различать корреляцию и причинно-следственные связи.

Нет никаких сомнений в том, что продукты, рекламируемые больше всего, и продаются лучше всего. Кинокомпания «20th Century Fox» потратила 150 миллионов долларов на промоушн фильма «Аватар», ставший в итоге самым кассовым фильмом всех времен. Но сколько из 2,7 миллиарда долларов, вырученных от продажи билетов, было получено благодаря мощному маркетингу? Отчасти причина этих безумных затрат на раскрутку проста: «20th Century Fox» точно знала, что продукт будет востребован.

---

* Лига американского футбола. — *Прим. ред.*

Фирмы считают, что знают, насколько эффективна их реклама. Экономисты в этом сомневаются. Профессор экономики Чикагского университета Стивен Левитт, сотрудничая с компанией по разработке и продаже электроники, был далеко не в восторге, когда фирма попыталась убедить его, что знает, насколько хорошо работает ее реклама. Левитта интересовало, почему руководство компании было настолько уверенно.

Ему пояснили, что каждый год накануне Дня отца расходы на телевизионную рекламу увеличиваются. И конечно, каждый год накануне Дня отца случаются самые высокие продажи. Но, возможно, это происходит просто потому, что множество детей покупают своим папам в подарок именно электронику, и особенно часто — именно в День отца, независимо от рекламы.

«Они вывернули наизнанку причинно-следственную связь», — говорил Левитт на лекции[15]. Конечно, реклама могла работать. Мы этого пока не знаем. «Это очень серьезная проблема», — добавляет он.

Поскольку очень важно решить эту проблему, фирмы хотят провести тщательные исследования. Левитт попытался убедить компанию провести контролируемые эксперименты, чтобы точно узнать, насколько эффективными были рекламные ролики. Поскольку A/B-тестирование на телевидении невозможно, нужно было понять, что происходит в некоторых регионах без использования рекламы.

Но представители фирмы ответили: «Ты с ума сошел? Мы не можем давать рекламу на двадцати рынках. Генеральный директор нас убьет». На этом сотрудничество Левитта с компанией закончилось.

Что возвращает нас к игре «Пэтриотс» с «Рэйвенс». Как результат футбольного матча поможет нам определить воздействие рекламы? Ну, он не в состоянии показать влияние той или иной конкретной рекламной кампании. Но он может продемонстрировать средний эффект рекламы многих крупных компаний.

Оказывается, в подобных играх проводится скрытый эксперимент с рекламой. Вот как это происходит. Ко времени проведения полуфинала компании уже купили и произвели рекламу для Супербоула. При этом когда фирмы решают, какую именно рекламу запустить, они еще не знают, какие команды сыграют в финале.

Но результаты полуфинала очень сильно повлияют на состав аудитории Супербоула. Каждая из двух команд привлечет огромное количество зрителей. Если победят «Пэтриотс», базирующиеся недалеко от Бостона, смотреть Супербоул будет гораздо больше людей из Массачусетса, чем из Мэриленда. И наоборот.

Для компаний это эквивалент подбрасывания монеты для определения, будут ли подвергнуты воздействию рекламы дополнительные десятки тысяч людей — в Балтиморе или в Бостоне. Это будет известно уже после того, как рекламные места куплены и реклама создана.

Теперь вернемся на поле, где Джим Нантц из CBS объявил окончательные результаты этого эксперимента.

> Билли Кандифф из «Рэйвенс» готовится сравнять счет ударом с поля и, по всей вероятности, перевести игру в овертайм. 32 ярда до ворот. Удар! Берегись! Берегись! Ой, как плохо… И «Пэтриотс» отправляются в Индианаполис на 46-й Супербоул.

Спустя две недели на Супербоуле доля аудитории в Бостоне составила 60,3, а в Балтиморе — 50,2. В Бостоне рекламу 2012 года будут смотреть 60 тысяч человек.

В следующем году в полуфинале встретятся те же две команды. На сей раз победит Балтимор. Дополнительная реклама на Супербоуле-2013 окажется в Балтиморе.

	Рейтинг Супербоул-2012 (играли в Бостоне)	Рейтинг Супербоул-2013 (играли в Балтиморе)
Бостон	56,7	48,0
Балтимор	47,9	59,6

Мы с главным экономистом Google Хэлом Вариан и экономистом университета Карнеги-Меллон Майклом Д. Смитом использовали эти два матча, а также все остальные Супербоулы в период с 2004 по 2013 год, чтобы проверить, работает ли — и если да, то насколько — реклама на Супербоуле. В частности, мы проверяли: когда компания рекламирует фильм во время этого события, происходит ли существенное увеличение продаж билетов в городах, представленных повышенным количеством зрителей?

Да, так и есть. Люди в городах команд-финалистов чемпионата посещают рекламируемые во время этого матча фильмы в значительно большем количестве, чем в других регионах. Там просто большее число людей увидело рекламу и решило пойти в кино.

Одно из альтернативных объяснений этого явления заключается в том, что, если ваша любимая команда участвует в Супербоуле, возрастает вероятность того, что

вы пойдете посмотреть кино. Тем не менее мы протестировали группу фильмов, имевших аналогичные бюджеты в аналогичные периоды, но не рекламировались во время Супербоула. Увеличения посещаемости в городах команд — участниц финала не наблюдалось.

Ладно, как вы уже догадались, реклама работает. Это не удивительно.

Но дело не только в том, что она работает. Реклама оказалась невероятно эффективна. В самом деле, когда мы впервые увидели результаты, то дважды, трижды и четырежды перепроверили их, чтобы убедиться в их подлинности — отклик получился невероятно высоким. Прокатчики в нашей выборке заплатили за рекламное место во время Супербоула в среднем около 3 миллионов долларов за фильм. А получилии 8,3 миллиона долларов от увеличения продаж билетов — то есть отдача от инвестиций составила 2,8:1.

Этот результат был подтвержден Уэсли Р. Хартманном и Дэниелом Клэппером, двумя другими экономистами, самостоятельно и раньше нас пришедшими к этой идее. Эти специалисты изучали рекламу пива и безалкогольных напитков во время Супербоула[16] — тоже учитывая тот факт, что ее могут чаще посмотреть в городах команд, претендовавших на победу. Эта реклама была очень дорогой, но и наши, и их результаты позволили предположить: она настолько эффективна с точки зрения повышения спроса, что на самом деле компании значительно недоплачивают за нее.

И что все это значит для наших друзей из компании по продаже электроники, с которой работал Левитт? Вполне возможно, реклама во время Супербоула более

эффективна, чем где-либо еще. По крайней мере, наше исследование позволяет предположить, что реклама на День отца — хорошая идея.

Один из позитивных моментов эксперимента с Супербоулом заключается в том, что не надо специально назначать кого-либо в рабочую и контрольную группы. Все произошло совершенно естественно, поэтому такие исследования и назвали натурными. Откуда возникло это преимущество? Дело в том, что произвольные контролируемые эксперименты, какими бы результативными и простыми они ни были, в эпоху цифровых технологий по-прежнему не всегда можно осуществить.

Иногда мы не можем совместить действия во времени. Иногда, как в случае с компанией по продаже электроники, у владельцев нет желания проводить эксперимент со своей рекламной кампанией, потому что они слишком сильно вложились в результат, чтобы проверять его.

Иногда исследования просто невозможны. Предположим, вас интересует, как страна реагирует на потерю своего лидера. Дойдет ли дело до войны? Перестанет ли функционировать экономика? Или ничего не изменится? Очевидно, мы не можем убить значительное количество президентов и премьер-министров и посмотреть, что произойдет. Это было бы не только невозможно, но и аморально. На протяжении многих десятилетий университеты накопили институциональные обзорные списки (IRBs), которые определяют, является ли предложенный эксперимент этическим или нет.

Поэтому если мы хотим узнать причинно-следственные связи в определенном сценарии, а проведение

исследования неэтично или иначе неосуществимо, что мы можем сделать? Мы можем использовать то, что экономисты, рассматривая понятие природы достаточно широко — вплоть до футбольных матчей, — назвали натурными экспериментами.

Хорошо это или плохо (ладно, явно плохо), в жизни часто есть место для значительной доли случайности. Никто не знает наверняка, что или кто руководит Вселенной. Но ясно одно: кто бы ни стоял во главе этого шоу — законы квантовой механики, Бог, прыщавый парень в трусах[17], моделирующий Вселенную на своем компьютере, — все они *выходят за рамки* того, что допускает IRB.

Природа постоянно экспериментирует с нами. Двух человек подстрелили. У одного пуля проходит в миллиметрах от жизненно важного органа. Другой погибает. Жизнь несправедлива. Но, если это вас утешит, плохие результаты позволяют экономистам учиться. Экономисты используют случайности для проверки причинно-следственных связей.

Из 43 американских президентов[18] 16 стали жертвами серьезных покушений, четверо были убиты. Причины того, что некоторые выжили, были, в основном, случайными.

Сравните Джона Кеннеди и Рональда Рейгана[19]. У обоих пули летели в наиболее уязвимые части тела. Пуля Кеннеди взорвалась в его мозгу, убив его. В случае Рейгана пуля остановилась в сантиметрах от сердца, что позволило врачам спасти ему жизнь. Рейган выжил, в то время как Кеннеди умер — просто повезло.

Эти покушения на жизнь руководителей и случайности, позволяющие одним выжить, а другим умереть, происходят во всем мире. Сравните Ахмата Кадырова

в Чечне и Адольфа Гитлера в Германии. Оба были в сантиметрах от взорвавшейся бомбы. Кадыров умер. Гитлер изменил свое расписание[20], вышел из заминированной комнаты на несколько минут раньше, чтобы успеть на поезд, и таким образом выжил. Кадыров умер мгновенно[21].

Мы можем использовать лишь случайность — убийство Кеннеди, но не Рейгана, — дабы увидеть, что, в среднем, происходит после гибели лидера страны. Это и сделали два экономиста — Бенджамин Джонс и Бенджамин Олкен. Контрольная группа здесь — любая страна в годы сразу после покушения на президента. Например, США в середине 1980-х годов. А рабочей группой является любая страна в первые годы после убийства. Например, США в середине 1960-х годов.

В чем же проявляется следствие гибели лидера? Джонс и Олкен обнаружили, что такие убийства кардинально меняют мировую историю, поскольку страны сворачивают на принципиально иной путь развития. Новый лидер принуждает ранее мирные страны вступать в войну, а ранее воюющие — мириться. Новый лидер вызывает разорение экономически развивающихся стран и резкий старт экономического развития разоренных стран.

На самом деле результаты натурного эксперимента с убийством опрокинули существовавшие несколько десятилетий общепринятые представления о том, как функционируют страны. Многие экономисты ранее склонялись ко мнению, что президенты в основном являлись бессильными марионеточными правителями, действующими по указке внешних сил. Но натурный эксперимент Джонса и Олкена показал, что это совсем не так.

Многие не считают это исследование покушений на мировых лидеров примером работы больших данных. Число убитых или раненых президентов было, конечно, невелико — равно как число войн, возникших и прекращенных в результате этого. Велик был объем данных, необходимых для определения параметров траектории развития экономики, но, по большей части, время, к которому они относились, предшествовало цифровой эпохе.

Тем не менее такие натурные эксперименты — хотя сейчас они используются почти исключительно экономистами — это мощное оружие, которое обретает все более важное значение в эпоху точных и больших наборов данных. Это инструмент, которым ученые, занимающиеся сбором и анализом информации, будут пользоваться еще очень долго.

И да, как должно быть уже ясно, экономисты играют важную роль в развитии науки о данных. По крайней мере мне бы хотелось так думать, потому что и я приложил к этому свою руку.

Где еще можно найти натурные эксперименты — или ситуации, когда случайный ход событий помещает людей в рабочую или контрольную группы?

Ярчайший пример — лотереи. Именно за это экономисты так любят их — не участвовать, что нерационально, а изучать. Если вылетит шарик с цифрой три, мистер Джонс разбогатеет. Если это будет шарик с цифрой шесть, мистер Джонсон проиграет.

Для проверки причинно-следственных связей возникновения денежной лавины экономисты сравнивают тех, кто выиграл в лотерею, с теми, кто купил билеты,

но проиграл. Подобные исследования, как правило, обнаруживают, что выигрыш в лотерею сделает вас счастливым в краткосрочной перспективе[22], но не в долгосрочной*.

Экономисты также могут использовать вероятностный характер лотерей для наблюдения за изменением жизни человека, чей сосед вдруг разбогател. Данные свидетельствуют: выигрыш вашего соседа в лотерею может повлиять на вашу собственную жизнь[23]. Например, если ему и вправду так повезло, вы с большей вероятностью купите дорогой автомобиль вроде BMW. Почему? Экономисты утверждают, что почти наверняка причиной является ваша ревность к ставшему богаче соседу, купившему дорогую машину. Спишем это на человеческую природу. Если г-н Джонсон видит г-на Джонса за рулем совершенно нового BMW, г-н Джонсон захочет такой же.

К сожалению, г-н Джонсон зачастую не может позволить себе BMW: экономисты обнаружили, что у соседей выигравших в лотерею людей значительно больше шансов обанкротиться[24]. Идти в ногу с Джонсом в данном случае невозможно.

Но натурные эксперименты не обязательно должны иметь такую же явно случайную природу, как лотерея. Едва вы принимаетесь выискивать случайности, как начинаете видеть их повсюду — и можете использовать их, чтобы понять, как действует наш мир.

Врачи являются частью натурного эксперимента. Время от времени правительство — по сути, по совершенно

---

* В известной статье 1978 года утверждалось, что выигрыш в лотерею сделает вас счастливым. Позже этот миф был во многом развенчан. — *Прим. авт.*

произвольным причинам — меняет формулы, использующиеся для вычисления оплаты работы врачей с пациентами, имеющими медицинскую страховку. Врачи в некоторых штатах обнаружили, что их оплата за определенные процедуры увеличилась. А их коллеги в других штатах констатировали уменьшение своих гонораров.

Два экономиста — Джеффри Клеменс и Джошуа Готлиб, бывшие одноклассники — проверили последствия этого произвольного изменения. Всегда ли врачи прописывают пациентам одинаковое лечение — такое, которое они считают наиболее необходимым? Или ими движут материальные стимулы?

Данные ясно показывают, что врачи могут быть мотивированы денежными стимулами[25]. В округах с более высоким возмещением за процедуры некоторые доктора постоянно назначают самые дорогие процедуры — например выполняют больше операций по удалению катаракты, колоноскопий и МРТ.

А затем встает главный вопрос: становится ли их пациентам лучше после выполнения всех этих лишних процедур? Клеменс и Готлиб сообщили только о «небольших улучшениях». Авторы не обнаружили их статистически значимого влияния на уровень смертности. Этот натурный эксперимент показывает: создание более сильных финансовых стимулов для врачей к назначению определенных процедур приводит к тому, что некоторые доктора начинают назначать лечение, не оказывающее существенного влияния на здоровье пациентов и не помогающее им продлить или улучшить свою жизнь.

Натурные эксперименты способны помочь ответить на вопросы жизни и смерти. Они также могут помочь

с ответами на вопросы, важные для некоторых молодых людей.

Stuyvesant High School (известная как «Стай») расположена в построенном за 150 миллионов долларов[26] 10-этажном бежевом кирпичном здании с видом на реку Гудзон — в нескольких кварталах от Всемирного торгового центра в Нижнем Манхэттене. Одним словом, «Стай» впечатляет. Она предлагает 55 углубленных программ[27] (AP, Advanced Placement) на семи языках, а также факультативы по еврейской истории, научной фантастике и азиатско-американской литературе. Примерно четверть выпускников[28] принимают в университеты Лиги плюща* или в какой-либо престижный колледж. В «Стайвесанте» преподает[29] профессор Гарварда физик Лиза Рэндалл, стратег Обамы Дэвид Аксельрод, лауреат премии Оскар актер Тим Роббинс, а также писатель Гари Штейнгарт. На церемонии вручения дипломов выступали[30] Билл Клинтон, Кофи Аннан и Конан О'Брайен.

Единственное, что впечатляет здесь еще сильнее, это стоимость обучения — 0 (ноль) долларов. Это государственная школа — наверное, лучшая в стране. Действительно, в недавнем исследовании по ранжированию

---

* Лига плюща — ассоциация восьми частных американских университетов на северо-востоке США. Это название происходит от побегов плюща, обвивающих старые здания в этих университетах. Считается, что члены лиги отличаются высоким качеством образования. В состав лиги входят Брауновский (Провиденс), Гарвардский (Кембридж), Йельский (Нью-Хейвен), Колумбийский (Нью-Йорк), Корнеллский (Итака), Пенсильванский (Филадельфия) и Принстонский (Принстон) университеты, а также Дартмутский колледж (Гановер). — *Прим. ред.*

всех государственных школ США было изучено 27 миллионов отзывов от 300 тысяч учеников и их родителей. В этом исследовании «Стай» заняла первое место[31]. Неудивительно, что амбициозные нью-йоркские родители среднего класса и их не менее амбициозные потомки одержимы стремлением попасть сюда.

Для Ахмета Йилмаза*, сына страхового агента и учителя в Квинсе, «Стай» был «ТОЙ САМОЙ средней школой».

«Рабочий класс и семьи иммигрантов видят в «Стае» возможность подняться по социальной лестнице, — объясняет Йилмаз. — Если ваш ребенок ходит в «Стай», он попадет в один из лучших двадцати университетов страны. Семье будет хорошо».

Итак, как можно попасть в школу «Стайвесант»? Вам придется поселиться в одном из пяти определенных районов Нью-Йорка и набрать больше определенного числа баллов на вступительных экзаменах. Вот и все. Никаких рекомендаций, эссе, унаследованного признания или прочих подкрепляющих действий. Один день, одно испытание, одна оценка. Если ваш результат выше определенного порога, вы приняты.

В ноябре каждого года около 27 тысяч нью-йоркских малышей садятся за парты, чтобы сдавать вступительные экзамены. Конкуренция серьезная. Изо всех детей, пришедших на экзамены, в «Стае» окажется меньше 5%[32].

Йилмаз объясняет, что его мама «работала как проклятая» и все свои деньги отдавала на подготовку сына к тесту. Проведя несколько месяцев подряд каждый будний день и все выходные за подготовкой к экзамену,

---

* Я изменил его имя и некоторые детали. — *Прим. авт.*

Йилмаз был уверен, что попадет в «Стай». Он до сих пор помнит тот день, когда получил конверт с результатами. Он ошибся в ответах на два вопроса.

Я спросил его, что он чувствовал. «Как вы думаете, каково это, когда ваш мир разваливается на осколки, а вы всего лишь школьник?» — ответил он.

Утешительный приз казался ему убогим — Bronx Science, другая эксклюзивная и достаточно престижная государственная школа. Но это был не «Стай». Кроме того, Йилмаз чувствовал, что Bronx Science была более специализированной школой, предназначенной для будущих технических специалистов. Четыре года спустя его не приняли в Принстонский университет. Он поступил в университет Тафтса, а после него попробовал несколько вариантов развития карьеры. Сегодня он является достаточно успешным сотрудником в одной из компаний, хотя говорит, что у него «отупляющая» работа, и платят ему не так, как хотелось бы.

Более чем десятилетие спустя Йилмаз признался, что иногда думает о том, как бы сложилась его жизнь, попади он в «Стай». «Все было бы по-другому, — говорит он. — Буквально у каждого, кого я знаю, жизнь сложилась иначе». Он считает, что попадание в среднюю школу «Стайвесант» обеспечило бы ему более высокие результаты вступительных экзаменов в университетах Принстона или Гарварда (он уверен, что оба этих вуза значительно лучше Тафтса) и, возможно, более высокооплачиваемую работу.

Люди любят перебирать гипотезы, и это может быть что угодно — от развлечения до самоистязания. Какой бы была моя жизнь, если бы я сделал шаг к той девочке или

к тому мальчику? Если бы я согласился на эту работу? Если бы я пошел в эту школу? Но на эти «что если» нет ответа. Жизнь — это не видеоигра. Вы не можете воспроизводить ее в различных сценариях до получения желаемого результата.

Известный писатель Милан Кундера замечательно сказал об этом в своем романе «Невыносимая легкость бытия»: «Жизнь человека происходит только один раз, и поэтому мы не можем определить, какие решения являются хорошими, а какие плохими. В каждой конкретной ситуации мы можем принять только одно решение; мы не получим второй, третьей или четвертой жизни, в которых смогли бы сравнивать различные варианты решений».

Йилмаз никогда не проживет жизнь, в которой ему удалось бы набрать на экзамене на два пункта больше.

Но, возможно, есть способ получить некоторое представление о том, насколько иной могла (или не могла) бы быть его жизнь — изучить большое число выпускников «Стайвесанта».

Тупой, наивный метод — сравнивать всех учеников, поступивших в «Стайвесант», и всех, кто не смог это сделать. Мы можем проанализировать результаты их тестов AP и экзаменов, а также в какие вузы они были приняты. Сделав это, мы бы обнаружили, что ученики, попавшие в «Стай», показали значительно более высокий результат в стандартизированных тестах и были приняты в существенно более престижные университеты. Но, как мы уже видели в этой главе, свидетельства такого рода сами по себе не убедительны. Может быть, причина гораздо более высоких результатов выпускников «Стайвесанта» в первую очередь заключается в том, что «Стай»

привлекает самых лучших учеников? Корреляция не доказывает причинность.

Для проверки *причинно-следственных связей* в средней школе «Стайвесант» мы должны сравнить две практически идентичные группы: тех, кто учился в «Стае», и тех, кто не учился там. Нам нужен натурный эксперимент. Но где мы можем найти данные?

Ответ: благодаря ученикам — таким, как Йилмаз, — показавшим результат, очень близкий к необходимому для поступления в «Стайвесант»*. Школьники, немного не дотянувшие до проходного балла, станут контрольной группой, а набравшие его — рабочей.

Нет никаких оснований подозревать, что ученики по обе стороны от проходного балла значительно отличаются по таланту или трудоспособности. От чего, в конце концов, зависит то обстоятельство, что один человек набирает на экзамене всего на один или два пункта больше, чем другой? Может быть, тот, кто недобрал пару баллов, просто поспал на десять минут меньше или съел недостаточно питательный завтрак? Может быть, тот, кто набрал более высокий балл, на тесте вовремя вспомнил особенно сложное слово, услышанное в разговоре с бабушкой, случившемся за три года до экзамена?

---

* При поиске коллег Йилмаза по несчастью, показавших результат чуть ниже проходного, я был поражен количеством людей от 20 до 50 лет, все еще помнящих то экзаменационное переживание из своего раннего подросткового возраста и с драматизмом говорящих о недополученных баллах. Среди них есть бывший конгрессмен и кандидат в мэры Нью-Йорка Энтони Вайнер, вспомнивший, что он не попал в «Стай», поскольку не добрал всего один балл. «Они не хотели меня», — сказал мне Вайнер в телефонном интервью. — *Прим. авт.*

По сути, эта категория физических экспериментов (использование четкого численного ограничения) настолько результативна, что получила у экономистов собственное название — «разрыв непрерывности». В любое время существует точное число, разделяющее людей на две разные группы — разрыв (экономисты могут сравнить) или регресс (результаты людей очень, очень близки к точке разрыва).

Два экономиста, Кит Чен и Джесси Шапиро, воспользовались этим методом в федеральных тюрьмах, чтобы испытать влияние тяжелых условий заключения на вероятность будущих преступлений[33]. Федеральным заключенным в США присваиваются баллы в зависимости от характера их правонарушения и криминальной истории. Сумма баллов определяет условия их содержания. Те, у кого достаточно высокий балл, отправятся в исправительные колонии, что означает меньше контактов с другими людьми, меньше свободы передвижения и, скорее всего, больше насилия со стороны охранников или других заключенных.

Опять же, было бы несправедливо сравнивать заключенных, отправленных в тюрьмы строгого режима, с теми, кого определили в тюрьмы с большим уровнем свободы. Среди первых будет больше убийц и насильников, среди вторых — наркоманов и мелких воришек.

Но те, кто находится чуть выше и чуть ниже разграничительного значения, оказались практически идентичны по причинам судимостям и истории. Однако эта незначительная разница приводит к получению совсем иного тюремного опыта.

Результат? Экономисты обнаружили, что заключенные, живущие в более суровых условиях, в большей

степени склонны к рецидивизму. Жесткие условия содержания в тюрьмах не столько сдерживали их от преступлений, сколько закаливали и делали более жестокими.

Итак, как «разрыв непрерывности» применяется к условиям средней школы «Стайвесант»? Команда экономистов из Массачусетского технологического института и университета Дьюка — Атила Абдулкадир-оглу, Джошуа Энгрист и Параг Патхак — провела исследование. Ученые сравнили жизнь школьников из Нью-Йорка — как попавших в «Стай», так и недобравших немного баллов на экзамене. Иными словами, они изучили судьбы сотен учеников, как и Йилмаз, не сумевших ответить на экзамене в «Стайвесанте» на один-два вопроса. А затем сравнили их с жизнями сотен школьников, показавших лучший результат и превысивших проходной балл на пункт-другой. Показателями успеха были AP-баллы*, SAT-баллы** и рейтинги колледжей, в которые они в итоге попали.

Потрясающие результаты этих ученых очень точно были обозначены в названии написанной по итогам исследования статьи: «Иллюзия элиты». Результаты средней школы «Стайвесант»[34]? Никаких. Nada. Ноль. No. Ученики, стоящие по обе стороны проходного балла, в итоге набрали практически одинаковые AP-баллы, неразличимые SAT-баллы и в равных пропорциях поступили в престижные вузы.

Ученые пришли к выводу: единственной причиной лучших жизненных достижений выпускников «Стая»

---

* От Advanced Placement exams — ежегодные годовые тесты для всех учеников в США. — *Прим. ред.*

** От Scholastic Assessment test — итоговый тест перед зачислением в колледжи в США (аналог ЕГЭ). — *Прим. ред.*

является то, что в эту школу попадают самые одаренные дети. Само по себе обучение в «Стае» *не является причиной того,* что его ученики лучше сдают экзамены или попадают в лучшие колледжи.

«Для большого числа учеников острая конкуренция на экзамене за место в школе, — писали экономисты, — не представляется оправданной для улучшения обучения».

Почему не важно, в какую школу ты ходишь? Ответ на этот вопрос могут дать несколько историй. Рассмотрим еще двух учеников — двух молодых жительниц Нью-Йорка Сару Кауфман и Джессику Энг. Обе с раннего возраста мечтали попасть в «Стай». Результат Кауфман оказался на единицу выше проходного балла. Добиться этого ей позволил ответ всего на один вопрос. «Я не думаю, что может быть еще что-то столь же захватывающее», — вспоминает Сара. Результат Энг оказался на один балл ниже проходного. Кауфман пошла в школу своей мечты, а Энг — нет.

Так как же сложилась их жизнь? Они обе сделали успешную и полезную карьеру — как и большинство людей, набравших на экзаменах наивысшие 5% баллов. По иронии судьбы, Энг с большим удовольствием вспоминает свои школьные годы. Bronx Science, где она училась, была единственной средней школы с музеем Холокоста. Джессика обнаружила, что ей нравится курирование, а затем она изучала антропологию в Корнелле.

Кауфман же чувствовала себя немного потерянной в «Стае», где все ученики уделяли чрезмерное внимание рейтингам. Ей казалось, что слишком большой акцент делается на тестировании, а не на обучении. Она назвала свой опыт «натуральной сборной солянкой». Но это был

опыт обучения. В школе она поняла, что будет учиться только в гуманитарном вузе, где бы делался больший акцент на обучение. И поступила в вуз своей мечты — Уэслианский университет. Там Сара обнаружила в себе страсть помогать другим людям, и сегодня она — адвокат по общественным делам.

Люди умеют приспосабливаться к обстоятельствам, и те, кто собираются стать успешными, найдут преимущества в любой ситуации. Факторы, приводящие вас на вершину — ваш талант и ваша энергия. Вам не нужны напутственные речи или другие преимущества, предоставляемые самыми лучшими модными школами.

Это только одно исследование. И его выводы, вероятно, ослабляет тот факт, что большинство учеников, не поступивших в «Стайвесант», оказались в другом отличном учебном заведении. Но появляется все больше доказательств того, что обучение в хорошей школе очень важно, хотя учеба в самой лучшей из них мало что добавляет к вероятности достижения успеха.

Возьмем колледж. Важно ли попасть в один из лучших университетов мира, такой, как Гарвард, или в серьезную школу, такую, как при университете штата Пенсильвания?

Опять же, есть четкая корреляция между экзаменационными данными в школе и тем, сколько денег люди будут зарабатывать. Спустя 10 лет после начала карьеры средний выпускник Гарварда зарабатывает 123 тысячи долларов[35]. А средний выпускник другого университета — 87 800 долларов.

Но эта корреляция не подразумевает причинно-следственной связи.

Два экономиста, Стейси Дейл и Алан Крюгер, нашли гениальный способ проверить причинно-следственную связь влияния элитных вузов на будущий заработок своих выпускников. У ученых был большой набор данных о старшеклассниках, в том числе такие, где они подавали документы в колледж, где они были приняты в колледж, учились в колледже, их семьи, их доходы во взрослом возрасте.

Чтобы собрать рабочую и контрольную группы, Дэйл и Крюгер сравнили студентов с похожими историями, которые были приняты одной и той же высшей школой, но выбрали другие. Некоторые студенты, которых готовы были принять в Гарвард, отправились в университет Пенсильвании. Возможно, чтобы быть ближе к девушке или к парню. Или потому, что там работал преподаватель, у которого они хотели учиться. Другими словами, эти студенты, по данным приемных комиссий, были столь же талантливы, как и те, кто учился в Гарварде. Но в итоге у них оказался разный опыт обучения.

Итак, нашлись два студента с похожими историями, но один поступил в Гарвард, а другой выбрал университет штата Пенсильвания. Что было дальше? Результаты исследователей оказались столь же впечатляющими, как и данные исследования о средней школе «Стайвесант». Эти студенты закончили карьеру с более или менее одинаковым достатком. Если будущая зарплата — это мерило, то похожие студенты, принятые в престижные школы, но желающие учиться в разных университетах, заканчивают на примерно одинаковых ступенях карьерной лестницы.

Наши газеты пестрят статьями о чрезвычайно успешных людях, учившихся в вузах Лиги плюща. Это, например, основатель Microsoft Билл Гейтс и основатели Facebook Марк Цукерберг и Дастин Московиц — они все учились в Гарварде. (Кстати, все они бросили учебу, что ставит дополнительные вопросы о ценности образования в Лиге плюща).

Есть также истории людей, которые были достаточно талантливы для поступления в университет Лиги плюща, но решили пойти в менее престижный вуз[36] и позже стали крайне успешны в жизни. Это, например, Уоррен Баффет[37], начавший карьеру в Уортонской школе бизнеса Пенсильванского университета, входящего в Лигу плюща, но переведшийся в университет Небраска-Линкольн, потому что тот был дешевле. Кроме того, он ненавидел Филадельфию и считал занятия в Уортоне скучными. Исследования показывают, что с точки зрения заработка выбор менее престижного учебного заведения — это прекрасное решение. По крайней мере для Баффета и ему подобных.

Эта книга называется «Все лгут». Это означает, что все люди врут — друзьям, в опросах и самим себе, — чтобы выглядеть как можно лучше.

Но и мир обманывает нас, предоставляя неверные, вводящие в заблуждение сведения. Мир показывает нам огромное количество успешных выпускников Гарварда,

но меньше успешных выпускников государственных вузов вроде университета штата Пенсильвания. В результате мы верим, что поступление в Гарвард дает огромное преимущество.

Придумав правильный натурный эксперимент, мы можем верно интерпретировать данные, предоставляемые миром, и разобраться, что действительно полезно, а что — нет.

Натурные эксперименты также применимы к задачам, о которых говорилось в предыдущей главе. Они часто требуют большей детализации, разделения на рабочую и контрольную группы: на города в случае с Суперкубком; на графства при исследовании цен на медицинские процедуры; на учеников, немного не дотянувших до проходного балла и набравших чуть больше, в эксперименте со школой «Стайвесант». И детализация, как обсуждалось в предыдущей главе, часто требует намного больших — всеобъемлющих — наборов данных, которые становятся все более доступными по мере того, как мир оцифровывается. Поскольку мы не знаем, когда природа намерена провести очередные натурные эксперименты, то не можем устроить небольшой опрос, чтобы измерить результаты. Для вычленения того, что нас интересует, нам нужно много уже существующих данных. Нам нужны большие данные.

Есть еще один важный момент, о котором следует упомянуть, говоря о натурных экспериментах (либо наших собственных, либо тех, которые устроила сама природа), подробно описанных в этой главе. Большая часть этой книги была сосредоточена на обучении пониманию

окружающего мира — с каким количеством расизма пришлось столкнуться Обаме, сколько человек действительно являются геями, насколько мужчины и женщины не уверены в своем теле. Но эти контролируемые или натурные эксперименты имеют и практическое применение. Они нацелены на улучшение процесса принятия решений, на помощь в понимании, какие действия полезны, а какие — нет.

Компании могут узнать, как заполучить больше клиентов. Правительства могут узнать, как использовать компенсацию для лучшей мотивации врачей. Студенты могут узнать, какие школы окажутся наиболее полезными. Эти эксперименты демонстрируют потенциал больших данных, способных заменить догадки, житейскую мудрость и ненадежные корреляции тем, что на самом деле работает правильно — *причинно-следственной связью*.

ЧАСТЬ III

БОЛЬШИЕ ДАННЫЕ: ОБРАЩАТЬСЯ С ОСТОРОЖНОСТЬЮ

Глава 7

БОЛЬШИЕ ДАННЫЕ-ШМАННЫЕ: ЧЕГО ОНИ НЕ МОГУТ?

«Сет, Лоуренс Саммерс хотел бы встретиться с вами», — получил я несколько загадочное письмо. Оно было от Лоуренса Каца, одного из моих ученых-советников. Кац не сказал, почему Саммерс заинтересовался моей работой, хотя позже я узнал, что ему это было известно.

Я сидел в приемной возле офиса Саммерса. После некоторой задержки бывший министр финансов США, бывший президент Гарвардского университета и лауреат крупнейших премий в области экономики предложил мне войти.

Саммерс начал встречу, зачитав мою статью о влиянии расизма на деятельность Обамы, которую распечатал для него его секретарь. Саммерс владеет методом скоростного чтения. В процессе он иногда высовывает кончик языка вправо, в то время как его глаза стремительно мечутся влево-вправо и вниз по странице. Саммерс, читающий текст научной работы, напоминает мне великого пианиста, исполняющего сонату. Он так

сосредоточен, что, кажется, забывает обо всем остальном. Меньше чем за пять минут он прочел статью в 13 страниц.

«Вы говорите, что поисковые запросы в Google со словом «ниггер» предполагают расизм, — сказал Саммерс. — Это похоже на правду. Они предсказывают, где Обама получит меньшую поддержку, чем Керри. Это интересно. Мы действительно можем сказать, что Обама и Керри похожи?»

«Они были классифицированы политологами как имеющие подобные идеологии, — ответил я. — Кроме того, нет никакой связи между расизмом и изменениями в Белом доме. Результат остается неизменным, даже если мы добавляем элементы демографии, посещение церкви и владение оружием». Так говорим мы, экономисты. Я был весьма воодушевлен.

Саммерс остановился и уставился на меня. Он ненадолго повернулся к настроенному на канал CNBC телевизору. Затем снова посмотрел на меня, потом на телевизор, потом опять на меня. «Ладно, мне нравится эта статья, — сказал Саммерс. — Над чем вы еще работаете?»

Следующие 60 минут были, возможно, самыми интеллектуально головокружительными в моей жизни. Мы с Саммерсом поговорили о процентных ставках и инфляции, о поддержании порядка и о преступности, о бизнесе и о благотворительности. Многие встречающиеся с Саммерсом люди подпадают под его обаяние. Мне посчастливилось разговаривать с этим, бесспорно, самым умным человеком, которого я когда-либо встречал. Саммерс показался мне невероятно мудрым. Новые идеи увлекают его больше, чем что-либо другое — и это,

кажется, нередко создает ему немалые проблемы. Он был вынужден оставить свой пост в Гарварде после того, как высказал предположение, согласно которому одна из причин нехватки женщин в науке может заключаться в намного большей вариативности IQ у мужчин. Если Саммерс находит какую-либо идею интересной, он, как правило, говорит об этом, даже если это режет чей-то слух.

После запланированного времени окончания нашей встречи прошло полтора часа. Разговор затягивался, но я до сих пор не имел понятия, зачем понадобился Саммерсу, когда мне нужно будет уходить и как я узнаю об этом. Такое впечатление, что на тот момент Саммерс и сам, вероятно, забыл, зачем устроил эту встречу.

И тогда он задал вопрос на миллион — или, возможно, миллиард — долларов. «Вы думаете, что на основе каких-либо данных сможете предсказать ситуацию на фондовом рынке?»

Ага! Вот наконец и выяснилась причина, по которой меня позвали сюда.

Саммерс не был первым, кто задал мне этот вопрос. Мой отец в основном поддерживал мои нетрадиционные научные интересы. Но однажды и он поднял эту тему. «Расизм, жестокое обращение с детьми, аборты, — сказал он. — А ты не можешь зарабатывать на этом деньги для себя?» Другие члены семьи и друзья тоже заговаривали об этом. Не говоря уже о коллегах и незнакомцах в интернете. Кажется, всем хотелось знать, могу ли я использовать поиск в Google и другие крупные базы данных для покупки акций. Теперь к ним присоединился бывший секретарь казначейства Соединенных Штатов. Это было уже серьезнее.

Так *могут* ли новые источники больших данных успешно предсказать, какие акции будут наиболее выгодны? Короткий ответ — нет.

В предыдущих главах мы обсудили четыре мощных достоинства больших данных. В этой поговорим об их ограничениях — о том, чего мы не можем сделать с их помощью и, порой, как мы не должны их применять. Я решил начать этот разговор с рассказа о нашей с Саммерсом неудачной попытке выиграть на фондовых рынках.

В главе 3 мы отмечали, что новые данные скорее будут полезны в случае неубедительности результатов уже осуществленных исследований в той или иной области. Это горькая правда: гораздо легче получить новые выводы по поводу расизма, жестокого обращения с детьми или абортов, чем о том, как функционирует бизнес. Это является следствием того, что на поиск даже малейшего преимущества в эффективности бизнеса брошены поистине огромные ресурсы. Конкуренция в области финансов крайне жесткая.

Саммерс, человек, не склонный воспевать похвалу чужому уму, был уверен, что хедж-фонды нас уже опередили. Во время нашей беседы я был очень впечатлен тем, насколько уважительно он говорил о них, а также его убежденностью в том, что они предвосхитили многие из моих предложений. В ответ я с гордостью поделился с ним придуманным мной алгоритмом, который позволил мне получать более полные данные с помощью Google Trends. Он сказал, что это очень здорово. Когда же я спросил, мог ли «Ренессанс», количественный хедж-фонд, придумать подобный алгоритм, он усмехнулся и сказал: «Да, конечно, они бы смогли догадаться».

Сложность конкурирования с хедж-фондами — не самая основная проблема, с которой мы с Саммерсом столкнулись, продумывая возможность использования новых больших наборов данных для победы на фондовых рынках.

ПРОКЛЯТИЕ ЧИСЛА РАЗМЕРНОСТЕЙ

Предположим, ваша стратегия прогнозирования на фондовом рынке — подбрасывание монетки. Но при этом она создана на основе тщательного тестирования. Вот ваша методика: вы наносите метки на тысячу монет — от 1 до 1000. Каждое утро в течение двух лет вы подбрасываете все монеты, записывая, падают они орлом или решкой, а затем смотрите, идет ли индекс Standard & Poor's в тот день вверх или вниз. Вы постоянно анализируете всю статистику. И вуаля! Вы что-то обнаружили. Получается, что при 70,3% подбрасываний монета №391 падает решкой вверх тогда, когда индекс S&P растет. Связь статистически значимая, ее уровень высокий. Вы нашли свою счастливую монету!

Теперь просто каждое утро подбрасывайте ее и покупайте акции, когда она выпадает решкой. Ваши дни в футболке и с ужином пустой лапшой закончились. Монета 391 — это ваш билет в хорошую жизнь!

Или нет.

Вы стали очередной жертвой одного из самых дьявольских аспектов «проклятия числа размерностей». Он может нанести удар, когда у вас имеется много переменных (или «размерностей») и не так много наблюдений: в данном случае, тысяча монет и 504 торговых дня за эти

два года соответственно. Одна из этих размерностей — монета 391 — скорее всего, счастливая. Уменьшите количество переменных — подбрасывайте всего сто монет. И вероятность того, что вам повезет, существенно уменьшится. Увеличьте число наблюдений, попытавшись предсказать поведение индекса S&P за 20 лет — и монеты постараются «не ударить в грязь лицом».

«Проклятия размерности» является серьезной проблемой при работе с большими данными, поскольку новые наборы данных никогда не дают нам экспоненциально больше переменных, чем традиционные источники — каждый поисковой запрос, каждая категория твитов и т. д. Многие люди, утверждающие, что способны прогнозировать динамику рынка, используя какой-то большой источник данных, просто оказались в плену этого проклятия. Все, что они действительно сделали — нашли эквивалент монеты 391.

Возьмем, к примеру, команду ученых-компьютерщиков из университета штата Индиана и университета Манчестера. Эти специалисты утверждали, что могут спрогнозировать динамику рынков, основываясь на сообщениях в Twitter[1]. Они построили алгоритм обработки каждодневного настроения твитов всего мира, используя методы, подобные анализу настроений, рассматриваемому в главе 3. Однако они учитывали не одно настроение, а множество — счастье, злость, доброту и многие другие. И обнаружили, что повышенное число твитов с выражением спокойствия — таких как «я спокоен» — позволяет предположить повышенную вероятность роста промышленного индекса Доу — Джонса через шесть дней. Для использования их результатов был основан хедж-фонд.

В чем здесь проблема?

Основная загвоздка заключается в том, что ученые протестировали слишком много элементов. Если вы в случайном порядке исследуете достаточно много переменных, одна из них окажется статистически значимой. Они изучили много эмоций, они тестировали каждую эмоцию за день, два, три, семь до дня, поведение фондового рынка в который пытались предсказать. И все эти переменные были использованы для того, чтобы попытаться объяснить взлеты и падения индекса Доу — Джонса всего за несколько месяцев.

За шесть дней до этого спокойствие не было легитимным прогностическим фактором фондового рынка. В тот момент оно было эквивалентом нашей гипотетической монеты 391 для больших данных. Хедж-фонд на базе твитов был закрыт через месяц после запуска вследствие малой отдачи[2].

Не только хедж-фонды, пытающиеся предсказать динамику рынков, страдали от «проклятия размерности». Те же проблемы возникли у ученых, пытавшихся найти генетические ключи, объясняющие, кто мы есть.

Благодаря проекту «геном человека» теперь можно собрать и проанализировать полную ДНК человека. Потенциал этого проекта казался огромным.

Возможно, нам удалось бы найти ген, ответственный за шизофрению. Может быть, мы могли бы обнаружить ген, вызывающий болезни Альцгеймера, Паркинсона и боковой амиотрофический склероз. Может быть, мы могли бы найти ген, отвечающий — ух ты! — за высокий уровень интеллекта. Есть ли ген, который в состоянии добавить кучу пунктов к IQ? Есть ли ген, создающий гения?

В 1998 году Роберт Пломин, видный поведенческий генетик, утверждал, что нашел ответ. Он получил набор данных, включавший ДНК и уровни интеллекта сотен студентов. Он сравнил ДНК «умников» (учащихся с IQ от 160 и выше) с ДНК студентов со средним уровнем IQ.

И обнаружил поразительную разницу в ДНК этих двух групп. Это различие было расположено в одном маленьком уголке 6-й хромосомы — неясный, но мощный ген, задействованный в метаболизме мозга. Одна версия этого гена, названного IGF2r, у более умных встречалась в два раза чаще.

«Сообщается о находке первого гена, связанного с высоким уровнем интеллекта», — запестрели заголовки «Нью-Йорк Таймс».

Можете задуматься о многочисленных этических вопросах, возникших после открытия Пломина. Следует ли разрешить родителям проводить тестирование детей на наличие гена IGF2r? Должны ли быть разрешены аборты, если у плода выявлен низкий уровень IQ? Можно ли генетически модифицировать людей, чтобы обеспечить им высокий уровень IQ? Коррелирует ли IGF2r с расой? Хотим ли мы знать ответ на этот вопрос? Следует ли продолжить исследования в области генетики, связанные с IQ?

Прежде чем специалисты по биоэтике, которым приходилось заниматься подобными острыми вопросами, занялись решением проблемы, перед генетиками — в том числе перед самим Пломиным — встал более простой вопрос: насколько точным был результат? Неужели правда, что IGF2r предопределяет уровень интеллекта? Неужели правда, что гении вдвое чаще являются носителями этого гена?

Нет. Через несколько лет после первого исследования, Пломин получил доступ к данным другой выборки людей, также включавшей ДНК и показатели IQ. На этот раз IGF2r с IQ не коррелировал. Пломин — и это показатель добросовестного ученого — отказался от своих заявлений.

Это, по сути, реализация общей схемы исследований в области генетики и IQ. Во-первых, ученые сообщили, что нашли генетический фактор, определяющий уровень IQ. Затем они получили новые данные и обнаружили, что исходное утверждение было неправильным.

Например, недавно группа ученых под руководством Кристофера Шабри исследовала 12 громких заявлений о вариантах генома, связанных с IQ. Специалисты изучили данные о 10 тысячах человек и не смогли воспроизвести корреляции ни для одной из 12 заявок[3].

В чем проблема во всех этих случаях? «Проклятие размерности». Геном человека, как теперь известно ученым, отличается миллионами элементов. То есть, попросту говоря, слишком много генов для тестирования.

Если вы анализируете достаточно много твитов, чтобы понять, коррелируют они с фондовым рынком или нет, то лишь случайно можете найти тот, который действительно коррелирует. Если вы испытываете достаточно много генетических вариантов, чтобы понять, коррелируют они с IQ или нет, то найдете нужный лишь случайно.

Как преодолеть «проклятие размерности»? Вы должны со смирением относиться к своей работе и не потерять голову из-за ее результатов. Вы должны проверять их с помощью дополнительных тестов. Например,

прежде чем ставить все свои сбережения на монету 391, стоит посмотреть, что будет происходить в течение ближайших нескольких лет. Социологи называют это «вневыборочным» тестом. И чем больше переменных вы включаете, тем скромнее надо быть. Чем больше переменных вы включаете, тем жестче должен быть «вневыборочный» тест. Важно также тщательно следить за проведением каждого исследования — тогда вы сможете точно сказать, с какой вероятностью вы стали жертвой «проклятия размерностей» и насколько скептически следует отнестись к результатам. Что возвращает нас к разговору с Ларри Саммерсом. Вот как мы пытались обогнать рынок.

Первая идея Саммерса заключалась в использовании поисковых запросов для прогноза продаж ключевых продуктов (например, iPhone), который мог бы пролить свет на дальнейшую динамику акций компании (например, Apple). Действительно, существует корреляция между поисковыми запросами относительно «айфонов» и величиной их продаж. Когда люди часто гуглят «айфон», вы можете биться об заклад, что этих телефонов продается много. Однако эта информация уже была заложена в цену акций Apple. Очевидно, когда у Google стали много спрашивать об «айфонах», хедж-фонды тоже выяснили, что они будут хорошо продаваться — независимо от того, были ли для этого использованы данные поисковых запросов или какой-то иной источник.

Следующая идея Саммерса касалась прогнозирования инвестиций в развивающиеся страны. Если большое число инвесторов в ближайшем будущем начнут вкладывать деньги в, скажем, Бразилию или Мексику,

то акции компаний в этих странах, несомненно, вырастут. Возможно, мы могли бы спрогнозировать рост инвестиций с помощью ключевых поисковых запросов в Google — например, «инвестировать в Мексику» или «инвестиционные возможности в Бразилии». Это оказалось тупиком. Проблема? Такие поисковые запросы были слишком редки. Вместо выявления значимых закономерностей эти данные постоянно перескакивали с одного на другое.

Мы пытались исследовать акции отдельных компаний. Возможно, если бы люди искали «GOOG», это означало бы, что они собираются купить акции Google. Подобные запросы, предположительно, дают понять, что эти акции будут прилично торговаться. Но они не прогнозируют, будет ли фондовый рынок расти или падать. Одним из основных ограничений является то, что эти поиски не скажут нам, заинтересован ли кто-то в покупке или в продаже акций.

Однажды я взахлеб делился с Саммерсом своей новой идеей: последние запросы «купить золото», по-видимому, коррелируют с будущим ростом цен на золото. Саммерс ответил, что я должен проверить это и убедиться в точности результата. Корреляция перестала работать — возможно, потому, что некоторые хедж-фонды также обнаружили данное соотношение.

В итоге за несколько месяцев мы не нашли ничего полезного. Несомненно, если бы мы искали корреляцию с рыночными показателями в каждом из миллиардов терминов поисковых запросов в Google, мы бы нашли тот, который сработает — пусть даже незначительно. Но это, скорее всего, стало бы нашей монетой 391.

ЧРЕЗМЕРНЫЙ АКЦЕНТ НА ТОМ, ЧТО МОЖНО ИЗМЕРИТЬ

В марте 2012 года Зои Чанс, профессор маркетинга[4] из Йельского университета, получила по почте маленький белый шагомер. Она решила изучить, как это устройство, измеряющее количество шагов, которое вы делаете в течение дня, и начисляющее за это баллы, может вдохновить вас больше заниматься спортом.

То, что произошло дальше, стало настоящим кошмаром больших данных. Чанс оказалась настолько одержима этим устройством и зависима от увеличения числа шагов, что стала ходить с ним везде — от кухни до гостиной, до столовой, до подвала, до своего кабинета. Она шагала рано утром, поздно ночью, почти целый день — 20 тысяч шагов за 24 часа. Она смотрела на шагомер сотни раз в день, и от ее человеческого общения остались только разговоры онлайн с другими пользователями шагомера — они обсуждали стратегии для улучшения результатов. Зои вспоминала, как положила шагомер на свою трехлетнюю дочь, когда та зашагала — потому что была одержима повышением результата.

Чанс стала настолько одержимой, что забыла, с чего все началось. Она забыла об основной цели достижения самого высокого результата — обретении хорошей физической формы, поэтому не позволяла дочери пройти даже несколько шагов без шагомера. При этом она не выполнила ни одного научного исследования. В конце концов она избавилась от этого устройства — после того, как однажды поздно вечером упала обессиленная при попытке сделать еще несколько шагов. Хотя Зои и является

специалистом по обработке и управлению данными, этот опыт очень сильно повлиял на нее. «Это заставило меня начать скептически относиться к возможности получить дополнительную информацию и понять, что лишние данные — это не всегда хорошо», — сказала Чанс.

Эта история, конечно, крайность, но она указывает на потенциальные проблемы, которые могут возникнуть у людей, использующих данные для принятия решений. Цифры могут оказаться соблазнительными[5]. Мы можем зациклиться на них и упустить из виду более важные вещи. Например, Зои практически перестала замечать все остальное в жизни.

Даже менее навязчивая влюбленность в цифры может иметь свои недостатки. Рассмотрим акцент на тестировании, которому в XXI веке в американских школах стали уделять особое внимание. На основе тестов учителя судят об успеваемости учеников. Конечно, стремление получить более объективные показатели успеваемости вполне понятно, но есть многое, что нелегко передать цифрами. Более того, все эти тесты заставляют многих учителей просто целенаправленно готовить учеников к ним. Некоторые даже, как было доказано в статье Брайана Джейкоба и Стивена Левитта, мошенничают при прохождении этих тестов[6].

Проблема заключается в следующем: то, что можно измерить — зачастую не совсем то, что нас интересует. Мы можем оценить, как студенты отвечают на вопросы, выбирая из нескольких ответов. Но мы не можем измерить критическое мышление, любопытство или развитие личности. Попытка увеличить один легко измеряемый показатель — результаты теста или количество

шагов в день — не всегда помогает достичь того, чего мы пытаемся добиться.

В попытках самоулучшения этой ошибки не избежал и Facebook. Компания обладает тоннами информации о том, как люди используют сайт. Легко увидеть, сколько лайков имеет конкретный пост, сколько раз по нему кликнули, сколько раз им поделились. Но, по данным Алекса Пейсаховича, специалиста по информации Facebook, которому я уже писал об этих важных моментах, ни один из этих параметров не дает ответ на более важные вопросы: на что похож опыт использования сайта? Соединяет ли тот или иной пост пользователей с их друзьями? Способен ли он чему-то научить? Заставил ли смеяться?

Или рассмотрим информационную революцию в бейсболе в 1990-х годах. Многие команды стали использовать все более сложные виды статистики вместо того, чтобы полагаться на старомодный человеческий метод — принимать решения. Легко было измерить количество атак и подач, но не работу на поле, поэтому некоторые команды стали недооценивать важность обороны. В своей книге «*The Signal and the Noise*» *(«Сигнал и шум»)* Нейт Сильвер указал, что, например, «Окленд Эйс», увлекшаяся данными, занесенными в «Moneyball», в середине 1990-х проигрывала от восьми до десяти игр в год именно из-за паршивой обороны.

Решение не всегда принимается благодаря увеличению объема информации. Чтобы большие данные работали лучше, нужна особая приправа: решение человека и небольшие исследования, которые мы могли бы назвать малыми данными. В интервью с Сильвером генеральный

менеджер и главный персонаж «Moneyball» Билли Бин заявил, что уже приступил к увеличению своего бюджета на сбор информации.

Чтобы заполнить пробелы в гигантском пуле данных, Facebook тоже должен был использовать старомодный подход: спрашивать людей о том, что они думают. Каждый день при загрузке новостей сотням пользователей Facebook задавались вопросы о том, что они там прочитали. Иными словами, Facebook теперь автоматически собирает данные (лайки, клики, комментарии) и дополняет их малыми данными («вы действительно хотите увидеть этот пост в своей Ленте новостей? Почему?»). Да, даже такой невероятно успешной и большой компании, как Facebook, иногда приходится использовать источник информации, всячески принижавшийся в этой книге ранее — небольшой опрос.

Действительно, из-за этого сбора малых данных в дополнение к основному массиву информации — огромному количеству кликов, лайков и постов — команда специалистов Facebook может взглянуть на статистику иначе, чем можно было предположить. В Facebook работают социальные психологи, антропологи и социологи — для поиска того, что не могут предоставить нам голые цифры.

Некоторые педагоги тоже становятся внимательнее к слепым пятнам в больших данных. Растет уровень национальных усилий по дополнению тестирования информацией, полученной из малых данных. Теперь стали широко распространены опросы студентов, возрос интерес к опросам родителей и наблюдениям за учителями (другими опытными преподавателями) во время урока.

«Руководство школьных округов понимает, что не следует сосредотачиваться исключительно на результатах тестов», — говорит Томас Кейн[7], профессор из Гарварда. Трехлетнее исследование Фонда Билла и Мелинды Гейтс подтверждает значение в образовании как больших, так и малых данных. Авторы проанализировали, что именно модель, основанная на оценках тестов, опросы учеников или наблюдения педагогов, является наилучшим вариантом оценки качества обучения школьников. Максимальный результат получается при объединении всех трех компонентов. «Каждый элемент вносит свой вклад в общую картину[8]», — заключают авторы доклада.

Как я выяснил в Окале, штат Флорида, на встрече с Джеффом Седером, на самом деле многие операции с большими данными используют малые данные — чтобы заполнить пробелы. Напомню, Седер, получивший образование в Гарварде — гуру в мире лошадей. Он использовал уроки, извлеченные из огромного числа экспериментов, что позволило ему правильно спрогнозировать успех Американского Фараона.

Поделившись со мной информацией, а также компьютерными файлами и расчетами, Седер признался, что у него было и секретное оружие — Пэтти Мюррей.

Мюррей, как и Седер, имеет высокий интеллект и элитарное образование — диплом Брин Маур. Она также переехала из Нью-Йорка в глубинку. «Я люблю лошадей больше, чем людей», — признается Пэтти. Но Мюррей немного более традиционна в плане подхода к выбору лошадей. Она, как и многие агенты-лошадники, лично осматривает их, наблюдает, как они двигаются,

проверяет их на наличие шрамов и синяков, а также беседует с их владельцами.

Затем Мюррей связывается с Седером, и они принимают окончательное решение относительно лошадей, которых будут рекомендовать. Мюррей вынюхивает проблемы коней — проблемы, которые Седер со всеми своими самыми инновационными и важными данными не отлавливает.

Я предсказываю революцию, основанную на открытиях больших данных. Но это не значит, что мы можем просто прошерстить информацию и получить ответ на любой вопрос или игнорировать этические соображения. И большие данные не исключают необходимости использования всего того, что люди развивали в течение тысячелетий, стремясь понять окружающий мир. Они просто дополняют друг друга.

Глава 8

БОЛЬШЕ ДАННЫХ — БОЛЬШЕ ПРОБЛЕМ? ЧЕГО НАМ НЕ СТОИТ ДЕЛАТЬ?

Иногда возможности больших данных настолько впечатляют, что становится страшно. Это ставит перед нами этические вопросы.

ОПАСНОСТЬ ВООРУЖЕННЫХ ДАННЫМИ КОРПОРАЦИЙ

Недавно три экономиста[1] — Одед Нецер и Ален Лемар из Колумбийского университета и Михал Херценштейн из университета Делавэр — искали способы предсказать вероятность погашения кредита заемщиком. Ученые использовали данные сайта взаимокредитования Prosper. Потенциальные заемщики указывают краткое обоснование необходимости кредита и какое обеспечение они могут предоставить, а потенциальные кредиторы решают, могут ли они предоставить деньги. В целом около 13% заемщиков[2] не выполняют своих обязательств по кредиту.

Оказывается, язык потенциальных заемщиков является сильным прогностическим фактором вероятности возврата ими кредита. И это важный показатель — даже если кредиторы имеют возможность проконтролировать другую значимую информацию о потенциальных заемщиках, в том числе их кредитные рейтинги и доходы.

Ниже перечислены 10 обнаруженных исследователями словосочетаний, которые обычно используются при подаче заявки на кредит. Пять из них коррелируют с оплатой кредита положительно, другие пять — негативно. Иными словами, первые пять, как правило, используются людьми, которым можно доверять, а вторые пять — теми, кому не стоит верить. Посмотрите, сможете ли вы догадаться, какие где.

Бог	более низкие процентные ставки	после уплаты налогов
обещаю	оплачу	больница
свободный от задолженности	выпускник	
минимальный платеж	спасибо	

Можно подумать — по крайней мере, надеюсь на это, — что вежливый, открыто религиозный человек, дающий честное слово, окажется среди тех, кто наиболее вероятно погасит кредит. На самом деле это не так. Как показывает статистика, честность таких людей — ниже среднего значения.

Вот несколько фраз, сгруппированных по степени вероятности погашения кредита.

Выражения, используемые в кредитных заявках людьми, которые наиболее вероятно вернут долг

свободный от задолженности	после уплаты налогов	выпускник
более низкие процентные ставки	минимальный платеж	

Выражения, используемые в кредитных заявках людьми, которые наиболее вероятно не вернут долг

Бог	оплачу	больница
обещаю	спасибо	

Прежде чем мы обсудим этические последствия этого исследования, давайте с помощью его авторов подумаем, что оно говорит о людях. Что мы должны понять на основании разделения слов на две категории?

Во-первых, рассмотрим выражения, на основании которых можно сделать предположение о большей вероятности выполнения платежей по кредиту. Такие словосочетания, как «низкая процентная ставка» или «после уплаты налогов» указывают на определенный уровень финансовой искушенности заемщика. Поэтому, пожалуй, не удивительно, что они коррелируют с его намерением вернуть кредит. Кроме того, если он или она говорит о своих позитивных достижениях — таких как «выпускник» и «свободный от задолженности», — больше вероятность того, что он или она оплатит и этот кредит.

Теперь рассмотрим выражения, предполагающие, что заемщик вряд ли собирается возвращать кредит. Вообще,

если кто-то говорит вам, что обязательно все оплатит, он не будет этого делать. Чем более уверенно дается обещание платежа, тем выше шанс его нарушения. Если кто-то пишет «я обещаю, что верну, да поможет мне бог», он относится к числу людей, возвращение кредита которыми наименее вероятно. Воззвание к вашему милосердию и апелляция к находящемуся в больнице родственнику также означает, что кредит вряд ли будет возвращен. На самом деле упоминание любого члена семьи — мужа, жены, сына, дочери, матери или отца — это знак того, что свои деньги назад вы не получите. Еще одно слово, указывающее на невозврат — «объяснить». Оно означает: если люди пытаются объяснять, почему они собираются погасить кредит, значит, они, скорее всего, не будут этого делать.

У авторов нет объяснения, почему благодарность в обращении свидетельствует о повышении вероятности невозврата кредита.

В целом, по данным этих исследователей, предоставление детального плана платежей с указанием выполненных в прошлом обязательств вовсе не свидетельствует о том, что данный заемщик погасит данный кредит. Давать обещания и взывать к милосердию — это явный признак того, что человек не будет возвращать долг. Независимо от причин (или от того, что говорит нам о человеческой природе это стремление давать обещания, которые в действительности никто не собирается выполнять) ученые обнаружили, что этот тест дает чрезвычайно ценную информацию для прогнозирования невозврата кредита. Люди, упоминающие бога, в 2,2 раза чаще не отдают долги — и это один из самых высоких показателей невозврата.

Но авторы также считают, что их работа поднимает этические вопросы. Хотя это было просто академическим исследованием, некоторые компании заявляют, что используют подобные интернет-данные при одобрении кредитов. Допустимо ли это?

Хотим ли мы жить в мире, где с помощью слов, которые мы пишем, можно предсказать, будем ли мы погашать кредит? Это как минимум жутковато — и порой просто страшно.

В ближайшем будущем потребителю, желающему получить кредит, возможно, придется обеспокоиться не только своей финансовой историей, но и своей активностью в интернете. И результат может зависеть от факторов, кажущихся абсурдными — например, используется ли в постах слово «спасибо» или упоминается ли бог. Далее. Что сказать о женщине, которая должна помочь своей сестре, оказавшейся в больнице, и которая, безусловно, оплатит кредит? Кажется ужасным отказать ей только на том основании, что в среднем люди, взывающие о помощи на медицинские расходы, часто врут. Мир, функционирующий таким образом, начинает выглядеть мерзко.

Это этический вопрос: есть ли у корпораций право судить о нашей пригодности получать их услуги, основываясь на абстрактных, но статистически прогностических критериях, не связанных непосредственно с этими услугами?

Оставим мир финансов. Давайте посмотрим на случаи, имеющие более серьезные последствия — например

при найме на работу. Работодатели при оценке кандидатов все чаще просматривают соцсети. Нет никаких проблем, если они ищут доказательства обливания грязью предыдущих работодателей или раскрытия конфиденциальных сведений с прошлого места работы. Можно даже найти некоторое оправдание отказу нанять кого-то, чьи записи в Facebook или в Instagram свидетельствуют о чрезмерном употреблении алкоголя. Но что если HR-специалисты находят совершенно безобидный показатель, соотносящийся с чем-то, что их беспокоит?

Исследователи из Кембриджского университета и Microsoft дали 58 тысячам американских пользователей Facebook различные тесты, касающиеся их личности и интеллекта. И обнаружили, что лайки на Facebook часто коррелируют с IQ[3], экстравертностью и добросовестностью. Например, люди, признающиеся в любви к Моцарту, грозам и картошке в виде спиралек, как правило, имеют более высокий IQ. А те, кто любит мотоциклы «Харлей-Дэвидсон», кантри-группу «Lady antebellum» или страницу «Мне нравится быть мамой», как правило, имеют более низкий показатель интеллекта. Некоторые из этих корреляций могут быть связаны с «проклятием размерности». Если вы протестируете достаточно много параметров, некоторые из них будут случайным образом коррелировать. Но часть интересов могут коррелировать с IQ вполне законно.

Тем не менее кажется несправедливым, что умный человек, которому — такое случается — нравятся «Харлеи», может не получить соответствующую его квалификации работу только потому, что он, сам того не понимая, сигнализировал о низком IQ.

Справедливости ради следует сказать, что эта проблема не нова. О людях уже давно судят по факторам, не связанным напрямую с производительностью — по твердости рукопожатия или по чистоте одежды. Но опасность информационной революции заключается в том, что по мере все большего оцифровывания нашей жизни эти приблизительные суждения могут становиться все более запутанными — и при этом все более навязчивыми. Улучшение прогнозирования может привести ко все более и более отвратительной дискриминации.

Более точные данные могут привести к другой форме сегрегации, которую экономисты называют «ценовой дискриминацией». Предприятия часто пытаются выяснить, какую плату они должны взимать за товары или услуги. В идеале они хотят брать с клиентов максимум того, что те готовы платить — таким образом будет извлекаться максимально возможная прибыль.

Большинство предприятий, как правило, в конечном итоге выбирают одну цену, которую готов заплатить каждый потребитель. Но иногда они знают, что члены определенной группы в среднем платят больше. Именно поэтому цены на билеты в кинотеатры для клиентов средних лет выше — у них доходы более высокие, чем у студентов или пенсионеров. Именно поэтому авиакомпании часто берут больше за билет с клиентов, купивших его в самую последнюю минуту. Это ценовая дискриминация.

Большие данные позволяют предприятиям существенно лучше изучить, за что клиенты готовы платить и как разделить людей на группы. Компания Optimal Decisions Group была пионером в использовании научных данных для определения цены, которую потребители

готовы платить за страховку. Как это было сделано? Специалисты компании использовали методологию, уже обсуждавшуюся в этой книге. Они нашли клиентов, наиболее похожих на тех, кто желал купить страховку в то время, и оценили, насколько высокую страховую премию те желают получить. Другими словами, был использован метод двойников. Поиск двойников — это здорово, если он помогает нам предсказать, вернется ли бейсболист к своему былому величию. Поиск двойников — это отлично, если он помогает нам вылечить кого-то. Но поиск двойников, помогающий корпорации выжать из вас все до последней копейки? Это уже не так круто. Мой брат-мот будет иметь право жаловаться, если с него возьмут больше, чем с меня-скряги.

Азартные игры — это та область, в которой возможность увеличения числа клиентов потенциально опасна. Большое казино использует нечто вроде поиска двойников для лучшего понимания своих клиентов. Цель? Извлечь максимально возможную прибыль и убедиться, что все больше ваших денег идет в его казну.

Вот как это работает. Казино полагает, что у каждого игрока есть «болевая точка». Это сумма убытков, которые достаточно сильно пугают его — настолько, что он или она не возвращается в казино в течение длительного периода времени. Предположим, например, что у Хелен «болевая точка» — 3000 долларов. Это означает, что если она потеряет их, то казино потеряет клиента — возможно, на несколько недель или месяцев. Если Хелен проиграет 2999 долларов, ей это не понравится. Кто, в конце концов, любит расставаться с деньгами? Но это не деморализует ее настолько сильно, чтобы завтра вечером она не вернулась.

Представьте на минуту, что вы — управляющий казино. И представьте, что Хелен пришла поиграть в игровые автоматы. Каков оптимальный результат? Понятно, вы хотите, чтобы она подошла как можно ближе к своей «болевой точке», но не ступила на нее. Вы хотите, чтобы она оставила в казино 2999 долларов — достаточно для того, чтобы принести вам большую прибыль, но не настолько много, чтобы больше к вам не вернуться.

Как это можно сделать? Ну, есть способы заставить Хелен перестать играть после того, как она потеряла определенную сумму. Например, вы можете предложить ей бесплатное питание. Принять такое предложение достаточно заманчиво, и она бросит автомат ради еды.

Но этот подход связан с серьезной проблемой. Откуда узнать «болевую точку» Хелен? Ведь у людей они разные. Для Хелен это 3000 долларов, а для Джона она может составлять 2000. А для Бена это может быть 26 000. Если вы убедите Хелен остановить игру после того, как она проиграла 2000 долларов, то потеряли прибыль. Если вы ждали слишком долго после того, как она проиграла 3000 долларов, то потеряли на некоторое время ее саму. Далее. Хелен может не захотеть сообщать вам о своей «болевой точке». Она даже может не знать о ней.

Так что же делать? Если вы дошли до этого места в книге, то, вероятно, сможете угадать ответ. Нужно использовать научные данные. Вы узнаете о клиентах все, что нужно — их возраст, пол, почтовый индекс и поведение в азартных играх. На основании информации о проигрышах, выигрышах, посещениях и пропусках можно оценить «болевые точки».

Вы собрали всю возможную информацию о Хелен и находите похожих на нее игроков — более или менее

двойников. Затем выясняете их «болевой порог». Вероятно, это будет та же сумма, что и у Хелен. Именно так и поступает казино «Harrah's», нанявшее компанию «Терабайт», которая имеет доступ к большим данным.

Скотт Гноу, генеральный менеджер «Терабайта», в своей замечательной книге «*Super Crunchers*» объясняет, что менеджеры казино, увидев, что клиент приближается к «болевой точке», подходят к нему и говорят: «Я вижу, у вас был тяжелый день. Я знаю, что вам понравится наш стейк-хаус. Предлагаю вам сейчас отвести жену на ужин за наш счет».

Это может показаться верхом щедрости — бесплатный обед. Но на самом деле это не так. Казино просто пытается заставить клиентов выйти из игры прежде, чем они потеряют так много, что больше никогда сюда не вернутся. Иными словами, менеджмент использует сложный анализ данных, чтобы постараться извлечь из клиентов как можно больше денег в долгосрочной перспективе.

Мы вправе опасаться, что все большее и большее использование онлайн-данных даст казино, страховым компаниям, кредиторам и другим юридическим лицам слишком большую власть над нами.

С другой стороны, большие данные позволяют и потребителям получить определенную компенсацию от предприятий, берущих с них слишком много или поставляющих некачественную продукцию.

Мощное оружие — сайты вроде Yelp, которые публикуют обзоры ресторанов и компаний, предоставляющих различные услуги. Недавнее исследование экономиста Майкла Лука из Гарварда показало, в какой степени те или иные бизнесы пострадали по милости Yelp[4]. Сравнивая отзывы с данными о продажах в штате Вашингтон,

он обнаружил: уменьшение числа звезд на Yelp* на одну снижает доходы ресторана на 5–9%.

Потребителям в их борьбе с бизнесом также помогают сайты, сравнивающие торговые площадки и отели — такие, как Kayak и Booking.com. Как обсуждалось во «*Фрикономике*»**, когда интернет-сайты начали публиковать отчеты о ценах разных страховых компаний, эти цены резко упали. Если страховщики берут слишком много, клиенты узнают об этом и найдут себе других. Какой оказалась общая экономия для потребителей? Один миллиард долларов в год.

Другими словами, данные в интернете могут подсказать компаниям, каких клиентов стоит избегать, а каких использовать. Они также могут подсказать клиентам, с какими фирмами не стоит связываться, а также какие из них пытаются их, клиентов, использовать. На сегодняшний день большие данные помогают обеим сторонам в борьбе друг с другом. Мы должны убедиться, что борьба по-прежнему честная.

ОПАСНОСТЬ ВООРУЖЕННЫХ ДАННЫМИ ПРАВИТЕЛЬСТВ

Когда ее бывший бойфренд пришел на вечеринку по поводу дня рождения, Адриана Донато поняла, что он расстроен. Ей показалось, что он сошел с ума. Она знала, что он боролся с депрессией. Когда он пригласил ее покататься на машине, Донато, 20-летняя студентка-зоолог,

---

* Американский сайт yelp.com, аналог российского otzovik.com — *Прим. ред.*

** Ориг. название Freaconomics. Книга Стивена Левитта и Стивена Дабнера, 2009 год. — *Прим. ред.*

не знала только одного. Она не знала, что ее бывший бойфренд, 22-летний Джеймс Стоунхэм, предыдущие три недели провел в поисках информации о том, как кого-то убить, и о наказании за убийство, вперемешку с редкими запросами о самой Адриане.

Если бы она знала это, она бы наверняка не села с ним в машину. И, скорее всего, он бы не зарезал ее в тот вечер.

В фильме «Особое мнение» экстрасенсы сотрудничают с полицией, чтобы предотвратить преступления еще до их возникновения. Следует ли для тех же целей предоставить большие данные и отделениям полиции? Нужно ли было по крайней мере предупредить Донато о поисковых запросах ее бывшего бойфренда, а полиции — допросить Стоунхэма?

Во-первых, следует признать: находится все больше доказательств того, что поисковые запросы в Google относительно преступной деятельности напрямую коррелируют с этой самой преступной деятельностью. Кристина Ма-Келламс, Флора Ор, Чжи Хен Баек и Ичиро Кавачи доказали, что количество запросов в Google, связанных с суицидом[5], сильно коррелирует с количеством самоубийств. Кроме того, мы с Эваном Солтасом обнаружили, что еженедельное число исламофобских поисковых запросов — например, с текстом «я ненавижу мусульман» или «надо убивать мусульман» — напрямую коррелирует с количеством преступлений против мусульман на этой неделе. Если большее число людей выполняет поисковые запросы с сообщением о своем желании что-то сделать, это значит, что большее число людей это сделают.

Так как же нам быть с этой информацией? Есть одна простая и достаточно бесспорная идея: мы можем

использовать данные по территориям. Если в каком-то городе сильно растет число поисковых запросов, связанных с самоубийством, мы можем быть уверены в том, что количество суицидов там также возрастет. Значит, местным властям или некоммерческим организациям пора запускать рекламу, объясняя, где люди могут получить психологическую помощь. Аналогично, если в городе сильно возросло число запросов «убивать мусульман», отделениям полиции стоит изменить принцип патрулирования улиц — например, можно направить больше сотрудников к местной мечети.

Но один шаг нам делать не слишком приятно: преследовать людей еще до того, как они совершат преступление. Нам кажется, что это вторжение в частную жизнь. С точки зрения этики, существует большая разница между возможностью правительства собирать информацию о поисковых запросах тысяч или сотен тысяч людей и возможностью полиции записывать аналогичные данные конкретного человека. С точки зрения этики, существует большая разница между защитой местной мечети и возможностью обшарить чужой дом. С точки зрения этики, существует большая разница между рекламой профилактики самоубийств и заключением кого бы то ни было в психиатрическую больницу против его воли.

Однако причина быть предельно осторожными с использованием информации личного характера выходит даже за рамки этики. Она заключается и в самих данных. С точки зрения науки о данных, есть большая разница между попытками предугадать вероятность определенных событий в городе и старанием предсказать действия отдельного человека.

Давайте вернемся к самоубийству. Каждый месяц в США делается около 3,5 миллиона связанных с суицидом поисковых запросов[6]. При этом в большинстве из них есть определенные намеки — слова «самоубийца», «самоубийство» и «способы самоубийства». Проще говоря, каждый месяц — больше одного запроса о самоубийстве на каждые 100 американцев. Как тут не вспомнить философа Фридриха Ницше: «Мысль о самоубийстве — это великое утешение: она помогает человеку преодолеть темноту ночи». Данные поисковых запросов в Google показывают, как это верно и насколько распространены мысли о суициде. Тем не менее каждый месяц в Соединенных Штатах совершается менее четырех тысяч самоубийств. То есть суицидальные мысли достаточно распространены, а самоубийства — нет. Поэтому полицейским нет смысла стучаться в двери каждый раз, когда кто-либо сообщает онлайн о желании вышибить себе мозги — если копы будут заниматься этим, у них ни на что больше не останется времени.

Или рассмотрим невероятно злобные исламофобские запросы. В 2015 году в Соединенных Штатах примерно 12 тысяч раз искали выражение «убивать мусульман». И, как сообщалось, на почве ненависти были убиты 12 мусульман[7]. Очевидно, что подавляющее большинство людей, делавших этот ужасающий запрос, не доходят до осуществления соответствующего действия.

Существуют математические выкладки, объясняющие разницу между предсказанием поведения конкретной личности и общей ситуацией в городе. Вот простой мысленный эксперимент. Предположим, в городе живет один миллион человек. И там есть одна мечеть. Предположим,

что если кто-то не обратится к Google с запросом «убить мусульман», то он нападет на мечеть с вероятностью всего 1 на 100 миллионов. Предположим, что кто-то начнет искать в Сети выражение «убить мусульман». Тогда эта вероятность резко возрастает до 1 к 10 тысячам. Предположим, что исламофобия взлетела до небес и число поисковых запросов «убить мусульман» выросло со 100 до 1000.

Как показывает математика, в этой ситуации шансы нападения на мечеть возросли примерно в пять раз — с 2 до 10 процентов. Но вероятность того, что человек, писавший «убить мусульман», действительно нападет на мечеть, по-прежнему остается только 1 на 10 тысяч.

Правильная реакция в этой ситуации — не сажать в тюрьму всех людей, искавших, как «убить мусульман». Не обыскивать их дома. Да, имеется крошечный шанс, что любой из них все же совершит преступление. Но правильным ответом, однако, было бы защитить мечеть, шансы которой подвергнуться нападению возросли до 10%.

Очевидно, что многие ужасные поисковые запросы не приводят к ужасным действиям.

Итак, хотя бы теоретически, существуют некоторые классы поисковых запросов, предполагающих достаточно высокую вероятность негативных последствий. И хотя бы теоретически возможно, например, что ученые, занимающиеся сбором и анализом данных, смогут в будущем построить модель, которая могла бы определить, что поисковые запросы Стоунхэма, связанные с Донато, должны вызвать серьезную обеспокоенность.

В 2014 году было зафиксировано около 6000 поисковых запросов с фразой «как убить подружку» и 400 убийств девушек. Если все эти убийцы перед совершением

преступления выполнили поиск в Google, это будет означать, что 1 из 15 человек перешел от слов к делу. Конечно, многие — вероятно, большинство людей, убивших своих подруг, — не искали информацию о преступлении именно таким образом. Это будет означать: истинная вероятность того, что конкретно этот поиск привел к убийству, намного ниже.

Но если бы ученые, занимающиеся анализом данных, смогли построить модель, вычисляющую угрозу в отношении конкретного человека — скажем, 1 к 100, — мы могли бы захотеть что-то сделать с этой информацией. По крайней мере человек, которому угрожает опасность, вероятно, имеет право быть проинформированным о ней — о существовании 1 шанса из 100 того, что он будет убит конкретным преступником.

Однако в целом мы должны быть очень осторожны при использовании поисковых данных для предсказания преступлений на индивидуальном уровне. Статистика ясно говорит нам: имеется много, очень много ужасных запросов, редко приводящих к ужасным последствиям. И пока все еще нет никаких доказательств, что правительство может с высокой долей вероятности спрогнозировать конкретное действие просто на основании изучения этих поисковых запросов. Поэтому нам следует быть очень осторожными и не позволять властям принимать меры на индивидуальном уровне, используя данные поисков. И дело не только в этических или юридических причинах. Дело — по крайней мере сейчас — в науке о данных.

Заключение

СКОЛЬКО ЛЮДЕЙ ДОЧИТЫВАЮТ КНИГУ ДО КОНЦА?

После подписания договора с издательством у меня было четкое видение того, как следует структурировать эту книгу. В начале — вы, возможно, помните — я описал сцену за столом у меня дома на День благодарения. Члены моей семьи обсуждали, в здравом ли я уме, и пытались выяснить, почему я к 33 годам никак не мог найти подходящую девушку.

Заключение к этой книге написалось практически само. Я хотел бы встречаться с хорошей девушкой и жениться на ней. Еще лучше — я хотел бы использовать большие данные, чтобы встретить подходящую девушку. Возможно, я смог бы вплести в книгу историю ухаживания, и тогда в заключении я бы описал, как все собрались вместе. А последние страницы включали бы описание дня моей свадьбы и любовное письмо к моей молодой жене.

К сожалению, жизнь не совпадает с фантазиями. Тот факт, что во время написания книги мне пришлось запереться в своей квартире и практически потерять связь

с реальным миром, наверняка не способствовал развитию романтической стороны моей жизни. И — увы — я еще не нашел себе жену. И, что еще важнее, мне все еще нужно новое заключение.

Я корпел над многими из моих любимых книг, пытаясь найти то, что позволит мне написать отличный эпилог. И пришел к выводу, что в лучших заключениях подчеркивается самая главная мысль, проходящая через всю книгу. Основная мысль этой книги следующая: социальная наука становится самой настоящей наукой. И эта новая настоящая наука призвана улучшить нашу жизнь.

В начале второй части я обсуждал критику Карлом Поппером трудов Зигмунда Фрейда. Я отмечал что Поппер не считал необычные мысли Фрейда научными. Но я кое-что не упомянул о критике Поппера. На самом деле это было нечто гораздо более важное, чем просто нападки на Фрейда. Поппер не считал, что какой-либо социальный ученый являлся ученым в строгом смысле этого слова. Карлу не нравилось отсутствие строгости в положениях, высказывавшихся этими так называемыми учеными.

Что двигало Поппером[1]? Общаясь с лучшими интеллектуалами своего времени — физиками, историками, психологами, — он заметил поразительную разницу. Когда говорили физики, Поппер верил им. Конечно, они порой совершали ошибки. Конечно, иногда — вследствие своих подсознательных предубеждений — они обманывались. Но они были вовлечены в процесс познания глубоких истин о мире, завершившийся появлением теории относительности Эйнштейна. Когда же говорили самые известные социологи, Карлу казалось, что они несут откровенную чепуху.

Поппер — не единственный, кто так по-разному относился к представителям разных наук. Все согласны, что физики, биологи и химики — это настоящие ученые. Для того чтобы найти объяснение явлениям физического мира, они проводят строгие эксперименты. И напротив, многие люди думают, что экономисты, социологи и психологи — неполноценные ученые, использующие бессмысленный жаргон, просто чтобы получить свою должность.

До последнего времени это было правдой. Но революционное появление больших данных все изменило. Если бы Карл Поппер был жив сегодня и посетил презентацию Раджа Четти, Джесси Шапиро, Эстер Дюфло или (ну, а все-таки!) мою, сильно подозреваю, что его реакция была бы совсем иной. По правде говоря, скорее всего, он задался бы вопросом, являются ли сторонники теории струн действительно учеными или они просто развлекаются умственной гимнастикой.

Если в городе показывают жестокое кино, число совершаемых преступлений растет или снижается? Если больше людей смотрит рекламу, произойдет ли увеличение продаж рекламируемого продукта? Если бейсбольная команда выигрывает, когда человеку исполняется 20, будет ли он продолжать болеть за нее в 40 лет? Это все — вопросы, на которые можно дать ответ «да» или «нет». И горы правдивой информации позволяют сделать это.

Это наука, а не псевдонаука.

Но это не означает, что революция в социальных науках придет в виде простых, вечных законов.

Марвин Мински, бывший ученый Массачусетского технологического института и один из первых, кто

взялся за изучение возможностей искусственного интеллекта, предположил, что в попытках копирования естественных наук, в которых удалось найти простые законы, верные везде и всегда, психология сбилась с пути.

Он считал, что человеческие мозги не могут быть объектом, подчиняющимся таким законам. Напротив, мозг, скорее всего — сложная система, в которой одна часть исправляет ошибки, возникшие в других. Экономика и политическая система могут быть не менее сложными.

Именно поэтому социальные науки вряд ли можно будет описать в виде аккуратной формулы вроде $e = mc^2$. В самом деле, если кто-то утверждает, что социальная наука может быть основана на сухих формулах, к этому следует отнестись весьма скептически.

Революционное преобразование социальных наук будет происходить постепенно, исследование за исследованием, поиск за поиском. Со временем мы начнем лучше понимать сложные системы строения человеческого сознания и общества.

Правильное заключение позволяет подвести итоги и наметить направление дальнейшей работы.

Что касается данной книги, это довольно легко. Наборы данных, о которых я говорил, революционны, но мало изучены. Еще многое нам только предстоит узнать. Честно говоря, подавляющее большинство ученых проигнорировали взрывное увеличение количества информации в цифровую эпоху. Самые известные в мире исследователи секса по-прежнему придерживаются испытанных и проверенных приемов. Они опрашивают несколько сотен человек об их желаниях, не собирая данные на сайтах вроде

PornHub. Большинство известных лингвистов мира анализируют отдельные тексты, по большей части игнорируя закономерности, выявленные при анализе миллиардов книг. Цифровая революция в основном не затронула методики, по которым учат аспирантов в областях психологии, политологии и социологии. Огромные практически неисследованные просторы информации, возникшие вследствие взрывного увеличения числа данных, заинтересовали лишь небольшое число дальновидных преподавателей, бунтующих студентов и любителей.

Но это изменится.

На каждую идею, о которой я говорил в этой книге, приходятся сотни не менее важных, лишь ждущих решения. Исследования, обсуждаемые здесь — это верхушка айсберга, царапины на поверхности.

Так что же еще мы прогнозируем?

Например, радикальное расширение методологии, использованной в одном из самых успешных исследований общественного здравоохранения. В середине XIX века английский врач Джон Сноу заинтересовался причиной вспышки холеры в Лондоне.

Он выдвинул гениальную идею[2]: сопоставить все случаи этой болезни в городе. Сделав это, он обнаружил, что заболевания в значительной степени группируются вокруг одного конкретного водяного насоса. После чего предположил, что болезнь распространяется через заражение воды — опровергнув тем самым расхожую мысль о плохом воздухе.

Большие данные — и детализация, которую они обеспечивают — делают этот тип исследования очень простым. При любом заболевании мы можем проанализировать

данные поисковых запросов в Google или других цифровых источниках о состоянии здоровья. Мы в состоянии найти на карте мира даже самые крошечные участки, где распространенность болезни является необычно высокой или необычно низкой. А затем оценить, что у них есть общего. Возможно, в воздухе? Или в воде? Или в социальных нормах?

Мы можем сделать это в отношении мигрени. Мы можем сделать это в отношении камней в почках. Мы можем сделать это в отношении беспокойства и депрессии, рака поджелудочной и болезни Альцгеймера, высокого кровяного давления и болей в пояснице, запоров и кровотечений из носа. Мы можем сделать это в отношении чего угодно. Анализ, некогда проведенный Сноу, мы могли бы провести 400 раз (некоторые исследования я начал уже во время написания этой книги).

Мы можем назвать это — применение простого метода и использование больших данных для проведения анализа несколько сот раз в течение короткого периода времени — наукой на высоком уровне. Да, социальные и поведенческие науки, безусловно, движутся к достижению таких позиций. Детализированные исследования в области медицины помогут этим наукам достичь требуемого масштаба. Этому также может поспособствовать использование A/B-тестирования. Мы обсуждали такой метод в контексте бизнеса — как добиться того, чтобы пользователи чаще кликали на рекламу. Сегодня эту эффективную методику используют повсеместно. Но A/B-тестирование можно применять для поиска ответов и на более фундаментальные — и социально значимые — вопросы, чем проблема кликов по рекламе.

Бенджамин Ф. Джонс[3] — экономист Северо-Западного университета, использующий A/B-тестирование для того, чтобы помочь детям лучше учиться. Он сумел создать платформу EDU STAR, которая позволяет школам случайным образом тестировать различные планы уроков.

Многие компании занимаются созданием образовательного программного обеспечения. Студенты входят в EDU STAR и случайным образом знакомятся с различными планами уроков. Затем они выполняют короткие тесты, призванные определить, насколько хорошо они разобрались с теми или иными заданиями. Иными словами, школы могут узнать, какое учебное программное обеспечение гарантирует лучшее усвоение материала.

EDU STAR, как и любая платформа на базе А/Б-тестирования, уже дает удивительные результаты. Один план урока, впечатливший представителей многих образовательных учреждений, позволял научить школьников работать с дробями. Считалось, что, если превратить математику в игру, ученики будут с бóльшим удовольствием узнавать новое и лучше выполнять тесты. Да? Неверно. Дети, изучавшие дроби посредством игры, проходили тесты хуже, чем те, кто знакомился с дробями стандартным способом.

Заинтересовать школьников в учебе — более захватывающее и социально полезное использование A/B-тестирования, чем его применение для того, чтобы заставить людей кликать на рекламу.

Средний американец спит каждую ночь 6,7 часа. Большинство из них хотят спать больше. Но вот наступает 11 вечера, и — спорт по телевизору или YouTube зовут.

Так что сон подождет. «Jawbone», компания, производящая гаджеты и имеющая сотни тысяч клиентов, проводит тысячи тестов в поисках решения, которое помогло бы пользователям сделать то, чего они так хотят — пойти спать пораньше.

«Jawbone» добилась отличного результата с помощью двойной цели. Сначала специалисты компании просят клиентов реализовать не самую амбициозную цель. Они отправляют им такое сообщение: «Похоже, вы мало спите в последние 3 дня. Попробуйте лечь спать в 23:30! Мы знаем, что обычно вы встаете в 8 утра». Затем у пользователя появляется возможность кликнуть на кнопку «Согласен».

Затем, в 22:30, «Jawbone» отправляет еще одно сообщение: «Вы хотели пойти спать в 23:30. Сейчас 22:30. Почему бы не начать сейчас?»

В «Jawbone» обнаружили, что такая стратегия привела к дополнительным 23 минутам сна. Компания не заставляет клиентов ложиться спать в 22:30, но заманивает их в постель пораньше.

Конечно, каждая часть этой стратегии должна быть оптимизирована путем долгих экспериментов. Если озвучить первоначальную цель — просить пользователей пойти спать в 11 вечера — слишком рано, мало кто согласится. Попросите пользователей лечь спать в полночь, и вы не многого добьетесь.

«Jawbone» использует A/B-тестирование для поиска эквивалента стрелки «вправо» в Google. Но вместо того, чтобы добиться еще нескольких кликов на рекламу партнеров Google, компания дает измученным людям еще несколько минут отдыха.

На самом деле для значительного увеличения успешности своих исследований целая армия психологов вполне может использовать инструменты Силиконовой долины. Я с нетерпением ожидаю первой статьи об этом — вместо описания пары быстрых A/B-тестов, проведенных с несколькими студентами.

Времена, когда ученые месяцами занимаются вербовкой небольшого числа старшекурсников для проведения одного теста, подходят к концу. Вместо этого аналитики будут использовать цифровые данные для тестирования нескольких сотен или тысяч идей за несколько секунд. Мы сможем узнать гораздо больше за гораздо меньшее время.

Данные в виде текста научат нас намного большему. Как распространяются идеи? Как создаются новые слова? Как исчезают слова? Как создаются шутки? Почему некоторые слова смешны, а другие — нет? Как развиваются диалекты? Держу пари, в течение 20 лет мы получим интересные ответы на эти вопросы.

Думаю, в качестве дополнения к традиционным тестам мы могли бы изучить поведение в Сети детей — естественно, анонимно, — чтобы понять, как они учатся и развиваются. Нет ли у них признаков дислексии? Развиваются ли у них зрелые интеллектуальные интересы? Есть ли у них друзья? Подсказки для ответов на все эти вопросы содержатся в тысячах кликов, которые каждый ребенок делает каждый день.

Есть и еще одна совершенно нетривиальная и намного более ценная область использования подобных методов.

В песне «Shattered» Мик Джаггер описывает все, делающее Нью-Йорк — это Большое Яблоко — таким волшебным. Смех. Радость. Одиночество. Крысы. Клопы.

Гордость. Жадность. Люди, одетые в бумажные мешки. Но большинство слов Джаггер посвящает описанию того, что делает этот город по-настоящему особенным: «секс, и секс, и секс, и секс».

То же и с большими данными. Благодаря цифровой революции нас ждут интересные открытия в здравоохранении и в науке о сне. В обучении. В психологии. В языке. И в сексе, в сексе, в сексе, в сексе.

В настоящее время я изучаю вопрос: сколько существует аспектов сексуальности? Мы обычно думаем, что люди бывают геями или натуралами — и все. Но сексуальность явно сложнее. Как среди геев, так и среди натуралов существуют различия — например, некоторым мужчинам нравятся блондинки, а другим — брюнетки. Могут ли эти предпочтения быть столь же сильными, как и предпочтения по полу? Другой вопрос: откуда они берутся? Так же, как мы в состоянии выяснить ключевой возраст, определяющий приверженность бейсболу или политические взгляды, возможно, нам удастся найти ключевой возраст формирования сексуальных предпочтений человека во взрослой жизни? Чтобы узнать ответы, вам придется купить мою следующую книгу под рабочим названием «Все (все еще) лгут».

Существование порно — и данных, полученных в этой области — дало возможность революционного развития исследований человеческой сексуальности.

Для того чтобы естественные науки начали менять нашу жизнь, требуется время. Постепенно были созданы пенициллин, спутники и компьютеры. Аналогично, может потребоваться время для того, чтобы большие данные смогли обеспечить серьезные успехи социальных

и поведенческих наук, помогающих нам любить, учиться и жить. Но я считаю, что некоторые подвижки уже есть, и надеюсь, что в этой книге вы смогли увидеть хотя бы контуры такого развития событий. Надеюсь, некоторые из вас, прочитав эту книгу, помогут сдвинуть дело с мертвой точки.

Чтобы правильно написать заключение, автор должен прежде всего думать о том, почему он написал эту книгу. Какую цель он пытался достичь?

Думаю, самой важной причиной, по которой я взялся за эту книгу, стал обретенный в результате жизненный опыт, очень поспособствовавший моему развитию. Знаете, немногим более 10 лет назад вышла удивительная книга «Фрикономика». В ней описывались исследования Стивена Левитта, превосходного экономиста из Чикагского университета. Левитт считался «пройдошистым экономистом». Казалось, он получил возможность использовать данные для получения ответа на любой вопрос, который смог придумать его изворотливый ум. Мошенничают ли борцы сумо? Проявляют ли дискриминацию геймеры? Предлагают ли нам риэлторы то, что купили бы сами?

Я тогда только-только закончил колледж со специализацией по философии и слабо представлял, что собираюсь делать в жизни. После прочтения «Фрикономики» я понял: хочу сделать то же, что и Стивен Левитт. Я хотел прорваться через горы данных, чтобы выяснить, как устроен наш мир на самом деле. Я хотел последовать за ним и решил получить степень кандидата экономических наук.

За прошедшие 12 лет многое изменилось. В некоторых исследованиях Левитта были обнаружены ошибки обработки данных. Он говорил некоторые политически некорректные вещи о глобальном потеплении. Из-за этого «Фрикономика» впала в немилость в интеллектуальных кругах.

Но, думаю, несмотря на некоторые допущенные им ошибки, прошедшие годы показали важность того, что Левитт пытался сделать. Он рассказывал нам, что сочетание любопытства, творчества и большого объема данных могут значительно улучшить наше понимание мира. На основании полученной информации можно рассказать множество интересных историй — и это было неоднократно доказано.

Я надеюсь, что эта книга сможет оказать такое же влияние на других людей, какое «Фрикономика» оказала на меня. Я надеюсь, что какой-то молодой человек, читающий это прямо сейчас, так же, как и я когда-то, находится на распутье и не знает, что хочет делать в жизни. Если у вас есть немного умения работать со статистическими данными, творческая жилка и любопытство — добро пожаловать в мир анализа информации.

По сути, эта книга, если я могу настолько смело выразиться, может рассматриваться как следующий уровень «Фрикономики». Основное различие между исследованиями, описанными там и здесь — это их притязания. В 1990-е годы, когда делал себе имя Левитт, было не так много данных. Он гордился ответами на причудливые вопросы, полученными на основе анализа данных. И он в значительной степени игнорировал важные темы в тех областях, где информации не существовало. Сегодня,

однако, везде имеется так много данных, что имеет смысл сразу перейти к наиболее значительным и самым глубоким вопросам, вникнуть в суть того, что значит быть человеком.

У анализа данных блестящее будущее. Я подозреваю, что следующий Кинси будет заниматься анализом данных. Следующий Фуко будет заниматься анализом данных. Следующий Фрейд будет заниматься анализом данных. Следующий Маркс будет заниматься анализом данных. Следующий Солк вполне может тоже заняться анализом данных.

Во всяком случае, я попытался сделать правильное заключение. Но пришел к выводу, что великие заключения способны на большее. Много большее. Великое заключение должно быть ироничным. Оно должно заставить действовать. Оно должно быть одновременно глубоким и веселым, насыщенным юмором и грустным. В великом заключении в одном-двух предложениях должно быть суммировано все то, о чем говорилось раньше, и все, что случится в будущем. Отличная книга должна заканчиваться умным, веселым, задорным, провокационным бабахом!

Теперь, возможно, пришло время поговорить немного о том, как я пишу. Я не очень плодовитый писатель. В этой книге всего около 75 тысяч слов, что совсем немного для такой емкой темы.

Но недостаток многословности я компенсирую одержимостью. За пять месяцев я написал 47 набросков моей первой колонки в «Нью-Йорк таймс», посвященной сексу — это две тысячи слов. Некоторые главы этой книги

я правил 60 раз. Я могу часами искать правильное слово для предложения в сноске.

Большую часть прошлого года я прожил отшельником. Только я и мой компьютер. Я жил в фешенебельной части Нью-Йорка и почти не выходил на улицу. На мой взгляд, эта книга — мой magnum opus, лучшее, что я сотворил в жизни. И я был готов пожертвовать всем, чтобы все сделать правильно. Я хотел отшлифовать каждое слово в этой книге. В моем телефоне скопилось множество писем, на которые я забывал ответить, и сообщений, которые я проигнорировал*.

После 13 месяцев напряженной работы я наконец смог отправить издателю завершенный проект книги. Одной части, правда, не хватало — заключения.

Я объяснил Дениз, моему редактору, что это может занять еще несколько месяцев. Я сказал ей, что, по-моему, мне потребуется на него еще шесть месяцев. Заключение, на мой взгляд — самая важная часть книги. И я только

---

* Поскольку все лгут, вы должны задаться вопросом: какая часть сказанного мной — правда. Может, я не такой уж трудоголик. Может быть, я не так уж и упорно работал над этой книгой. Может быть, я, как и многие люди, преувеличиваю, рассказывая о том, как много работаю. Может быть, среди 13 месяцев «напряженной работы» были и такие, когда я вообще не работал. Может быть, я не жил как отшельник. Возможно, если проверить мой профиль в Facebook, выяснится, что в этот период предполагаемого отшельничества я вовсю общался с друзьями. Или, может быть, если я и был отшельником, то отнюдь не добровольно. Может быть, я провел много ночей в одиночестве, будучи не в состоянии работать и напрасно надеясь, что кто-нибудь захочет связаться со мной. Возможно, никто мне и не писал сообщений. Все лгут. Каждый рассказчик ненадежен. — *Прим. авт.*

начал учиться писать отличные заключения. Излишне говорить, что Дениз была довольна.

Затем однажды один мой знакомый прислал мне исследование Джордана Элленберга. Элленберг — математик, работающий в Университете Висконсина — поинтересовался, сколько людей на самом деле дочитывают книги до конца. Он придумал гениальный способ проверить это с помощью больших данных. «Amazon» сообщает, сколько людей цитируют несколько строк из книг. Элленберг понял, что может сравнить, как часто цитирование берется из начала книги и из конца. Это дало бы ориентировочную информацию о склонности читателей дочитывать книгу до последней страницы. В результате выяснилось, что более 90% читателей закончили роман Донны Тартт «*Щегол*». В отличие от этой книги, только около 7% дошли до конца опуса лауреата Нобелевской премии экономиста Даниэля Канемана «*Думай медленно... Решай быстро*». По этой грубой оценке, менее 3% читателей добрались до финала много обсуждавшейся книги экономиста Томаса Пикетти «*Капитал в XXI веке*». Другими словами, люди не стремятся дочитывать трактаты экономистов[4].

Один из важных выводов этой книги заключается в том, что мы *всегда* должны идти туда, куда нас ведут большие данные, — и действовать соответственно. Надеюсь, большинство читателей внимательно проследят за всем, здесь написанным, и постараются выявить закономерности, связывающие содержание последних страниц с тем, о чем говорилось раньше. Но, как бы я ни старался оттачивать фразы, большинство людей наверняка прочитают первые 50 страниц, ознакомятся

с некоторыми интересными фактами и двинутся дальше по своим делам.

Таким образом, я завершаю эту книгу единственным адекватным способом — в соответствии с данными о том, что люди делают, а не с тем, что они говорят. Я собираюсь выпить пива с друзьями и перестать работать над этим чертовым заключением. Большие данные говорят мне: мало кто из вас все еще читает эту книгу.

БЛАГОДАРНОСТИ

Эта книга — плод совместных усилий многих людей. Ее идеи разрабатывались в то время, когда я учился в Гарварде, работал в Google специалистом по обработке и анализу данных и писал для *«Нью-Йорк Таймс»*.

Хэл Вариан, с которым мы работали в Google, оказал большое влияние на мое понимание идей этой книги. Насколько я могу судить, Хэл постоянно опережает время лет на 20. В его книге «*Information Rules*» *(«Информация решает все»)*, написанной совместно с Карлом Шапиро, будущее было предсказано, по большей части, удивительно точно. А его статья «Предсказание настоящего», написанная совместно с Юньянгом Чоем, во многом легла в основу революции больших данных в социальных науках, описанной в этой книге. Он также удивительно добрый наставник, что могут подтвердить многие трудившиеся с ним люди. Во время совместной работы над статьей Хэл нередко делает бóльшую часть, а затем настаивает, чтобы ваше имя стояло в списке авторов первым. Такое сочетание гения и щедрости, как у Хэла, встречается нечасто.

Моим процессом сочинительства во многом руководил Аарон Ретика, который был моим редактором в «*Нью-Йорк Таймс*» и обсуждал со мной каждую статью. Аарон — человек поистине энциклопедических знаний. Он разбирается в музыке, истории, спорте, политике, социологии, экономике и бог знает в чем еще — практически во всем. Благодаря ему в статьях, подписанных моим именем, появилось очень много интересных и важных сведений. Другими игроками в нашей команде являются Билл Марч, чьи графические материалы по-прежнему сводят меня с ума, Кевин Маккарти и Гита Данешьо. В эту книгу вошли отрывки из их статей, перепечатанные с их разрешения. Стивен Пинкер, любезно согласившийся написать предисловие, давно уже стал моим героем. Он установил высокую планку для современной книги по социологии — глубокое исследование основ человеческой природы, использование самых свежих данных из различных областей знаний. В своей дальнейшей работе я буду стараться никогда не опускаться ниже заданного им высокого уровня.

Моя диссертация, из которой выросла эта книга, была написана под руководством таких блестящих и терпеливых советников, как Альберто Алесина, Дэвид Катлер, Эд Глэзер и Лоуренс Кац. Дениз Освальд — замечательный редактор. Если вы хотите знать, насколько хорошо ее редактирование, сравните эту книгу с моим первым вариантом. На самом деле вы не сможете этого сделать, поскольку я не намерен никому его показывать, настолько он меня смущает. Я также благодарю всю команду HarperCollins, в том числе Майкла Баррса, Линн Грейди, Лорен Джанек, Шелби Мейзлик и Эмбер Оливер.

Эрик Лупфер, мой агент, видел потенциал этого проекта с самого начала. Он сыграл важную роль в формировании предложения и помог его реализовать.

Я благодарен Мелвису Акосте за превосходную проверку фактографического материала, а также многим другим людям, от которых я узнал много нового и важного для моей профессиональной и научной жизни — Сьюзен Атей, Шломо Бенарци, Джейсон Бордофф, Дэниэл Бауэрс, Дэвид Брукман, Бо Каугилл, Стивен Дельпом, Джон Донахью, Билл Гейл, Клаудия Голдин, Сюзанна Гринберг, Шейн Гринштейн, Стив Гроув, Майкл Хойт, Дэвид Лейбсон, А. И. Магнусон, Дана Мэлони, Джеффри Олдхэм, Питер Орзаг, Дэвид Рейли, Джонатан Розенберг, Майкл Шварц, Стив Скотт, Рич Шавельзон, Майкл Д. Смит, Лоуренс Саммерс, Джон Вавер, Майкл Уиггинс и Цин Ву.

Я благодарен Тиму Рекварту и NeuWrite за помощь в разработке текста. Я также благодарен Кристоферу Чабрису, Раджу Четти, Мэтту Генцкоу, Соломону Мессингу и Джесси Шапиро за помощь в интерпретации моих исследований.

Я спросил Эмму Пирсон и Катю Собольски, могут ли они дать мне какие-либо советы по различным главам этой книги. Они решили — по непонятным мне причинам — сперва прочитать всю книгу, а затем дать мудрые советы по каждому пункту.

Моя мать, Эстер Давидовиц, неоднократно читала всю книгу и помогала кардинально улучшить ее. Она также научила меня, например, следовать моему любопытству, независимо от того, куда это приведет. Когда я беседовал со своим руководителем относительно содержания

моей диссертации, он спрашивал меня: «Что ваша мама думает о том, чем вы занимаетесь?» Он полагал, что моей маме может быть стыдно из-за того, что я изучаю секс и другие запретные темы. Но я всегда знал, что она гордится разнообразием моих интересов.

Многие люди читали различные части будущей книги и давали полезные комментарии. Я благодарю Эдуардо Асеведо, Корен Апичелла, Сэма Ашера, Дэвида Катлера, Стивена Дабнера, Кристофера Глазека, Джессику Голдберг, Лорен Голдман, Аманду Гордон, Якоба Лешно, Алексея Пейсаховича, Ноя Поппа, Рамона Руйяра, Грега Собольски, Эвана Солтаса, Ноя Стивенс-Давидовица, Лорен Стивенс-Давидовиц и Джейн Янг. Вообще-то Джейн была моей лучшей подругой в то время, когда я писал эту книгу, за что я особенно ей благодарен.

Я благодарю за помощь в сборе данных Бретта Голденберга, Джеймса Роджерса и Майка Уильямса из MindGeek, Роба Маккууна и Сэма Миллера из Baseball Prospectus. Я благодарен Фонду Альфреда Слоуна за финансовую поддержку.

Во время написания этой книги в какой-то момент я серьезно застрял — настолько, что был готов бросить все и отказаться от проекта. Тогда мы с отцом, Митчеллом Стивенсом, поехали в деревню. В течение недели папа заставил меня собраться. Он брал меня с собой на прогулки, во время которых мы обсуждали любовь, смерть, успех, счастье и текст книги. А позже он сел рядом со мной и заставил заново просмотреть каждое предложение книги. Я бы не смог закончить этот труд без него. Все ошибки, безусловно, мои собственные.

ПРИМЕЧАНИЯ

Предисловие

[1] Katie Fretland, «Gallup: Race Not Important to Voters» («Гэллап: Раса не важна для избирателей»), The Swamp, *Chicago Tribune,* June 2008.

[2] Alexandre Mas and Enrico Moretti, «Racial Bias in the 2008 Presidential Election» («Расовые предрассудки на президентских выборах 2008 года»), *American Economic Review* 99, no. 2 (2009).

[3] 12 ноября 2009 года в эпизоде своего шоу Лу Доббс сказал, что мы живем в «пост-расовом обществе». 27 января 2010 года на его шоу Крис Мэттьюс сказал, что президент Обама был «по всем признакам, пост-расовым». Другие примеры см. Michael C. Dawson and Lawrence D. Bobo, «One Year Later and the Myth of a Post-Racial Society» («Год спустя или миф об обществе»), *Du Bois Review: Social Science Research on Race* 6, no. 2 (2009).

[4] Подробную информацию обо всех этих расчетах можно найти на моем сайте sethsd.com в формате CSV под заголовком «секс-данные». Данные общего социального обследования могут быть найдены по адресу http://gss.norc.org/.

[5] Данные, предоставленные автором.

[6] Авторский анализ с помощью Google Trends. Я тоже собрал данные на всех членов Stormfront, как описано в Seth Stephens-Davidowitz, «The Data of Hate» («Данные о ненависти»), *New York Times*, 13 июля 2014 года, sr4. Соответствующие данные могут быть обнаружены в sethsd.com в разделе под заголовком «Stormfront».

[7] Анализ автором трендов с помощью данных Google. Штаты, для которых это справедливо — Кентукки, Луизиана, Аризона и Северная Каролина.

[8] Этот документ был опубликован как Seth Stephens-Davidowitz, «The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data» («Уровень расовой враждебности для чернокожего кандидата: опыт использования данных поисковых запросов в Google»), *Journal of Public Economics* 118 (2014). Более подробную информацию об исследовании можно найти здесь. Кроме того, данные можно найти на моем сайте, sethsd.com в разделе под заголовком «расизм».

[9] «Самая сильная корелляция с поддержкой Трампа в поисковых запросах Google — слово «черномазый». Другие также сообщали об этом» (28 февраля 2016 года, твит). Смотрите также Nate Cohn, «Donald Trump's Strongest Supporters: A New Kind of Democrat» («Убежденные сторонники Дональда Трампа: новый тип демократа»), *New York Times,* December 31, 2015, A3.

[10] «Bringing Big Data to the Enterprise» («Привлечение Больших Данных к работе на предприятии»), ИБМ, https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

[11] Это обсуждается Seth Stephens-Davidowitz, «What Do Pregnant Women Want?» («Чего хочет беременная женщина?»), *New York Times*, 17 мая 2014года, SR6.

[12] Stephens-Davidowitz, «What Do Pregnant Women Want?» («Чего хочет беременная женщина?»)

[13] Я брал интервью у Джерри Фридмана по телефону 27 октября 2015 года.

Глава 1. Интуиция вас обманывает

[1] Я говорю о той части их анализа, которую хорошо знаю — о части, пытающейся объяснить и предсказать поведение человека. Я не говорю об искусственном интеллекте, который пытается, скажем, водить машину.

[2] John Paparrizos, Ryan W. White, and Eric Horvitz, «Screening for Pancreatic Adenocarcinoma Using Signals from Web Search Logs:

Feasibility Study and Results» («Скрининг поджелудочной железы аденокарцинома, используя сигналы из журналов веб-поиск: технико-экономическое обоснование и результаты»), *Journal of Oncology Practice* (2016).

[3] Это исследование обсуждается в Seth Stephens-Davidowitz, «Dr. Google Will See You Now» («Доктор Google теперь видит вас»), *New York Times*, 11 августа 2013, SR12.

[4] Lars Backstrom and Jon Kleinberg. «Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook» («Романтические отношения и дисперсия социальных связей: сетевой анализ статуса отношений на Facebook»), in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2014).

[5] Kahneman, *Thinking, Fast and Slow* («Думай медленно, решай быстро»).

[6] Между 1979 и 2010 годами, в среднем, 55,81 американцев погибли от ураганов и 4216,53 умерли от астмы. Посмотрите ежегодную статистику США погибших от ураганов в Национальной Метеорологической службе: http://www.spc.noaa.gov/climo/torn/fatalmap.php и тенденцию заболеваемости и смертности от астмы — в американской легочной ассоциации, эпидемиологии и статистики.

[7] Мое любимое видео Юинга «Patrick Ewing's Top 10 Career Plays» («10 лучших игр за карьеру Патрика Юинга»), на Ютуб, размещено 18 сентября 2015 года, https://www.youtube.com/watch?v=Y29gMuYymv8; и «Patrick Ewing Knicks Tribute» видео на Ютуб, опубликовано 12 мая 2006 года, https://www.youtube.com/watch?v=8T2l5Emzu-I.

[8] S. L. Price, «Whatever Happened to the White Athlete?» («Что случилось с белым спортсменом?»), *Sports Illustrated*, 8 Декабря 1997 года.

[9] Этот опрос потребителей Googlee я провел 22 октября 2013 года. Я спросил: «Где, по вашему мнению, родились большинство игроков НБА?» Были два варианта ответов: «бедные кварталы» и «кварталы среднего класса»; 59,7% опрошенных выбрали «бедный район».

[10] Roland G. Fryer Jr. and Steven D. Levitt, «The Causes and Consequences of Distinctively Black Name» («Причины и последствия явно чернокожих имен»), *Quarterly Journal of Economics* 119, no. 3 (2004).

[11] «Центр по контролю и профилактике заболеваний, США, 2009», Таблица 9, внебрачные дети с подробной разбивкой по расам, происхождению и возрастам матери: США, 1970–2006 годы.

[12] Крис Пол: «Не просто типичный качок: интересы форварда «Майами Хит» Криса Боша выходят далеко за рамки баскетбола», PalmBeachPost.com, 15 февраля 2011 года, http://www.palmbeachpost.com/news/sports/basketball/not-just-a-typical-jock-miami-heat-forward-chris-b/nLp7Z/; Dave Walker, «Chris Paul's Family to Compete on 'Family Feud'» nola.com, October 31, 2011, http://www.nola.com/tv/index.ssf/2011/10/chris_pauls_family_to_compete.html.

[13] «Почему наш вид становится выше?» Scientific American, http://www.scientificamerican.com/article/why-are-we-getting-taller/. Интересно, что американцы перестали расти. Amanda Onion, «Why Have Americans Stopped Growing Taller?» («Почему американцы перестали расти?»), ABC News, 3 июля 2016 года, http://abc news.go.com/Technology/story?id=98438&page=1. Я утверждаю: одной из причин наблюдаемого огромного притока игроков НБА, родившихся в других странах, является то, что другие страны догоняют США по росту. Количество родившихся в США в период с 1946 по 1980 годы баскетболистов НБА ростом более 180 см увеличилось в 16 раз. С тех пор этот показатель выровнялся, поскольку американцы перестали расти. Между тем, число игроков ростом более 210 см из других стран существенно возросло. Иностранные баскетболисты чрезвычайно высокого роста приезжают из таких стран, как Турция, Испания и Греция, где в последние годы отмечается заметное улучшение здоровья детей и увеличение роста взрослых.

[14] Carmen R. Isasi et al., «Association of Childhood Economic Hardship with Adult Height and Adult Adiposity among Hispanics/Latinos: The HCHS/SOL Socio-Cultural Ancillary Study», *PloS One* 11, no. 2 (2016); Jane E Miller and Sanders Korenman, «Poverty and Children's

Nutritional Status in the United States» («Бедность и детское питание в США»), *American Journal of Epidemiology* 140, no. 3 (1994); Harry J. Holzer, Diane Whitmore Schanzenbach, Greg J. Duncan, and Jens Ludwig, «The Economic Costs of Childhood Poverty in the United States» («Экономические последствия нищеты у детей в Соединенных Штатах»), *Journal of Children and Poverty* 14, no. 1 (2008).

15 Cheryl D. Fryar, Qiuping Gu, and Cynthia L. Ogden, «Anthropometric Reference Data for Children and Adults: United States, 2007–2010» («Антропометрические справочные данные для детей и взрослых: США, 2007–2010»), статистика департамента здравоохранения, серич 11, № . 252 (2012).

16 Tim Kautz, James J. Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans, «Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success» («Поощрение измерительных навыков: совершенствование когнитивных и некогнитивных навыков, содействующих успеху в жизни»), National Bureau of Economic Research Working Paper 20749, 2014.

17 Desmond Conner, «For Wrenn, Sky's the Limit» (Для Ренна небо — не предел»), *Hartford Courant*, *Hartford Courant*, 21 октября 1999 года.

18 История Дага Ренна была рассказана в Percy Allen, «Former Washington and O'Dea Star Doug Wrenn Finds Tough Times» («Для Дага Ренна, бывшей звезды Вашингтона, и О'Деа наступают трудные времена»), *Seattle Times*, 29 марта 2009 года.

19 Там же.

20 Melissa Isaacson, «Portrait of a Legend» («Портрет легенды»), ESPN.com, 9 сентября 2009 года, http://www.espn.com/chicago/columns/story?id=4457017&columnist=isaacson_melissa. Хорошую биографию Джордана написал Роланд Лейзенби, Roland Lazenby, *Michael Jordan: The Life* (Boston: Back Bay Books, 2015).

21 Barry Jacobs, «High-Flying Michael Jordan Has North Carolina Cruising Toward Another NCAA Title», *People*, 19 марта, 1984.

22 Isaacson, «Portrait of a Legend» («Портрет легенды»).

23 Речь Майкла Джордана в баскетбольном Зале славы, видео на Ютуб, опубликовано 21 февраля 2012 года, https://www.youtube.

com/watch?v=XLzBMGXfK4c. Наиболее интересный аспект речи Джордана не в том, что он был так несдержан, говоря о родителях, а в том, что он все еще чувствовал потребность указать на обиды начала своей карьеры. Возможно, обида на всю жизнь — необходимое условие для того, чтобы стать величайшим баскетболистом всех времен.

[24] «Я Леброн Джеймс из Акрона, штат Огайо», видео на YouTube, опубликовано 20 июня 2013 года, https://www.youtube.com/watch?v=XceMbPVAggk.

Глава 2. Возможно, Фрейд был прав?

[1] Я посчитал, что продукты имеют форму фаллоса, если их длина значительно больше их ширины и они, как правило, круглые. Я насчитал: огурцы, кукуруза, морковь, баклажаны, кабачки и бананы.

[2] Набор данных может быть загружен на https://www.microsoft.com/en-us/download/details.aspx?id=52418. Ученые попросили пользователей Amazon Mechanical Turk описать изображения. Они проанализировали логи кликов и отметили любой момент, когда кто-то исправлял слово. Более подробную информацию можно найти в Yukino Baba and Hisami Suzuki, «How Are Spelling Errors Generated and Corrected? A Study of Corrected and Uncorrected Spelling Errors Using Keystroke Logs» («Как исправлять орфографические ошибки? Исследование корректируемых и некорректируемых ошибок с помощью журналов нажатия клавиш»), Proceedings of the Fiftieth Annual Meeting of the Association for Computational Linguistics, 2012.

[3] Полные данные — предупреждение: в графическом виде — выглядят следующим образом:

«Я хочу заниматься сексом с...»

	Ежемесячное число поисковых запросов Google с указанной фразой
мамой	720
сыном	590
сестрой	590
кузиной	480
отцом	480
парнем	480
братом	320
дочерью	260
другом	170
подругой	140

[4] Например, порно — это одно из самых распространенных слов в поисковых запросах Google для различных чрезвычайно интересных анимационных программ, как показано ниже.

Мультфильмы с порно (чаще всего запрашиваемые в Google)

Гриффины порно	Смотреть Симпсонов	Футурама порно	Скуби Ду игры
эпизоды Гриффины	Симпсоны порно	Футурама Лила	Скуби Ду мультфильм
Гриффины бесплатно	Симпсоны фильм	Футурама онлайн	Скуби Ду велма

[5] По расчетам автора, это самые популярные женские профессии в поисковых порнозапросах мужчин, с разбивкой по возрасту последних:

Профессии в поисковых запросах порно у мужчин с разбивкой по возрасту

18–24	25–64	65+
Няня	Няня	Няня
Учительница	Инструктор по йоге	Черлидерша
Инструктор по йоге	Учительница	Врач
Черлидерша	Черлидерша	Учительница
Врач	Агент по недвижимости	Агент по недвижимости
Проститутка	Врач	Медсестра
Агент по недвижимости	Проститутка	Инструктор по йоге
Медсестра	Секретарша	Секретарша
Секретарша	Медсестра	Проститутка

Глава 3. Переосмысление данных

1 Matthew Leising, «HFT Treasury Trading Hurts Market When News is Released» («Как показывают данные, HFT Treasury Trading наносит удар по рынку»), Bloomberg Markets, 16 декабря, 2014 года; Nathaniel Popper, «The Robots Are Coming for Wall Street» («Роботы идут на Уолл-Стрит»), *New York Times Magazine*, 28 февраля 2016, MM56; Richard Finger, «High Frequency Trading: Is It a Dark Force Against Ordinary Human Traders and Investors?» («Высокочастотная торговля: это темные силы против простых трейдеров, и инвесторов?») *Forbes*, 30 сентября 2013 года, http://www.forbes.com/sites/richardfinger/2013/09/30/high-frequency-trading-is-it-a-dark-force-against-ordinary-human-traders-and-investors/#50875fc751a6.

2 Я брал интервью у Алана Крюгера по телефону 8 мая 2015 года.

3 Исходный документ — Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry

Brilliant, «Detecting Influenza Epidemics Using Search Engine Query Data» («Обнаружение эпидемий гриппа с помощью поискового запроса данных») *Nature* 457, no. 7232 (2009). Недостатки в исходной модели обсуждались в David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, «The Parable of Google Flu: Traps in Big Data Analysis» («Притча о гриппе в Google: ловушки в анализе Больших Данных»), *Science* 343, no. 6176 (2014). Исправленная модель представлена ShihaoYang, Mauricio Santillana, and S. C. Kou, «Accurate Estimation of Influenza Epidemics Using Google Search Data Via ARGO» («Точная оценка эпидемии гриппа с использованием данных поиска в Googlee и с помощью «АРГО»), *Proceedings of the National Academy of Sciences* 112, no. 47 (2015).

4 Seth Stephens-Davidowitz and Hal Varian, «A Hands-on Guide to Google Data» («Практическое руководство по данных Google»), мимеограф, 2015.

5 Sergey Brin and Larry Page, «The Anatomy of a Large-Scale Hypertextual Web Search Engine» («Анатомия крупномасштабной гипертекстовой поисковой системы»), 7-я Международная конференция, посвященная Всемирной сети, 14–18 апреля 1998 года, Брисбен, Австралия.

6 John Battelle, *The Search: «How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture» (*«Как Google и его конкуренты переписали правила бизнеса и изменили нашу культуру»), New York: Penguin, 2005.

7 Хорошее обсуждение этого вопроса можно найти в Steven Levy, «*In the Plex: How Google Thinks, Works, and Shapes Our Lives*» («Как Google думает, работает и определяет нашу жизнь»), Нью-Йорк: Саймон и Шустер, 2011.

8 Эта цитата была также включена в Joe Drape, «Ahmed Zayat's Journey: Bankruptcy and Big Bets» («Жизнь Ахмеда Заята: банкротство и большие ставки»), *New York Times*, 5 июня 2015 года, A1. Однако, в статье ошибочно приписывают цитату Седеру. На самом деле это сказал другой член его команды.

9 Я брал интервью у Джеффа Седера и Пэтти Мюррей в Окале, штат Флорида, в период с 12 по 14 июня 2015 года.

[10] Причины провала скачек были приблизительно оценены Джеффом Седером, исходя из его опыта в этом бизнесе.

[11] Дополнительные таблицы статистических данных и база данных травм лошадей, http://jockeyclub.com/pdfs/eid_7_year_tables.

[12] «Программа патологоанатомических исследований», California Animal Health and Food Laboratory System, 2013.

[13] Avalyn Hunter, «A Case for Full Siblings» («Дело братьев»), Bloodhorse, 18 апреля 2014 года, http://www.bloodhorse.com/horse-racing/articles/115014/a-case-for-full-siblings.

[14] Melody Chiu, «E. J. Johnson Loses 50 Lbs. Since Undergoing Gastric Sleeve Surgery» («Е. Дж. Джонсон теряет 50 кг после перенесенной операции рукавной резекции желудка»), *People*, 1 октября 2014 года.

[15] Eli Saslow, «Lost Stories of LeBron, Part 1» (СПотерянные рассказы о Леброне, Часть 1»), ESPN.com, October 17, 2013, http://www.espn.com /nba/story/_/id/9825052/how-lebron-james-life-changed-fourth-grade-espn-magazine.

[16] См. Sherry Ross, «Малышка на 16 миллионов», New York *Daily News*, 12 марта 2006 года, и Jay Privman, «The Green Monkey, Who Sold for $16M, Retired» («Зеленая мартышка, которая была продана за 16 млн долларов, отправилась на отдых»), ESPN.com, 12 февраля 2008 года, http://www.espn.com/sports/horse/news/story?id=3242341. Видео аукциона «Лошадь за 16 млн долларов», видео на Ютуб, опубликовано 1 ноября 2008 года, https://www.youtube.com/watch?v=EyggMC85Zsg.

[17] Sharad Goel, Jake M. Hofman, Sebastien Lahaie, David M. Pennock, and Duncan J. Watts, «Predicting Consumer Behavior with Web Search» («Прогнозирование поведения потребителя на базе вебпоиска»), *Proceedings of the National Academy of Sciences* 107, no. 41 (2010).

[18] Constance L. Hays, «What Wal-Mart Knows About Customers' Habits» («Что знает Wal-Mart о привычках клиентов»), *New York Times*, 14 ноября 2004 года.

[19] Я опросил Орли Ашенфельтера по телефону 27 октября 2016 года.

[20] Daniel A. McFarland, Dan Jurafsky, and Craig Rawlings, «Making the Connection: Social Bonding in Courtship Situations» («Создание

связей: социальные связи в ситуациях ухаживания»), *American Journal of Sociology* 118, no. 6 (2013).

[21] Jonathan Greenberg, «What I Learned From My Wise Uncle Leonard Cohen» («Что я узнал от моего дяди Леонарда Коэна»), *Huffington Post*, 11 ноября 2016 года.

[22] H. Andrew Schwartz et al., «Personality, Gender, *and Age in the Language of* Social Media: The Open-Vocabulary Approach» («Личность, пол и возраст в языке СМИ: применение открытого словаря»), *PloS One* 8, no. 9 (2013). В документе также разбиты на группы ответы людей на личностные тесты.

[23] Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds, «The Emotional Arcs of Stories Are Dominated by Six Basic Shapes» *EPJ Data Science* 5, no. 1 (2016).

[24] Jonah Berger and Katherine L. Milkman, «What Makes Online Content Viral?» («Что делает онлайн-контент вирусным?»), *Journal of Marketing Research* 49, no. 2 (2012).

[25] Эти исследования опираются на статью Matthew Gentzkow и Jesse M. Shapiro, «What Drives Media Slant? Evidence from U.S. Daily Newspapers» («Куда направлен уклон СМИ? Данные из ежедневных газет США»), *Econometrica* 78, no. 1 (2010). Хотя, когда этот проект стартовал, Генцкоу и Шапиро были всего лишь аспирантами, теперь они звезды в мире экономистов. Генцкоу, ныне профессор Стэнфордского университета, в 2014 году получил медаль Джона Бейтса Кларка и был объявлен ведущим экономистом моложе сорока лет. Шапиро, ныне профессор университета Браун — редактор престижного журнала политической экономики. Их совместная статья относительно уклона СМИ является одним из самых цитируемых статей в библиографии каждого.

[26] То, что консервативная «Нью-Йорк пост» принадлежит Мердоку, можно объяснить тем фактом, что Нью-Йорк очень большой, и в нем могут существовать газеты с различными точками зрения. Однако, очевидно, что «Пост» постоянно теряет деньги. См., например, Joe Pompeo, «How Much Does the 'New York Post' Actually Lose?» («Сколько теряет сегодня «Нью-Йорк Пост»?), Politico,

30 августа 2013 года, http://www.politico.com/media/story/2013/08/how-much-does-the-new-york-post-actually-lose-001176.

[27] Я взял интервью у Мэтта Генцкоу и Джесси Шапиро 16 августа 2015 года в Королевском отеле Бостон.

[28] Kate Rakelly, Sarah Sachs, Brian Yin, and Alexei A. Efros, «A Century of Portraits: A Visual Historical Record of American High School Yearbooks» («Век портрета: визуальный исторический портрет американского школьника»), доклад, представленный на Международной конференции по компьютерной визуализации, 2015. Фото печатается с разрешения авторов.

[29] См., например, Christina Kotchemidova, «Why We Say 'Cheese': Producing the Smile in Snapshot Photography» («Почему мы говорим «сыр»: Создание улыбки на фотографии»), *Critical Studies in Media Communication* 22, no. 1 (2005).

[30] J. Vernon Henderson, Adam Storeygard, and David N. Weil, «Measuring Economic Growth from Outer Space» («Измерение экономического роста из космоса»), *American Economic Review* 102, no. 2 (2012).

[31] Kathleen Caulderwood, «Nigerian GDP Jumps 89% As Economists Add in Telecoms, Nollywood» («ВВП Нигерии подскочил на 89%, когда экономисты добавили в «Телекомс», «Нолливуд»), *IBTimes*, 7 апреля 2014 года, http://www.ibtimes.com/nigerian-gdp-jumps-89-economists-add-telecoms-nollywood-1568219.

[32] Я взял интервью у Джо Райзингера по телефону 10 июня 2015 года.

[33] Leena Rao, «SpaceX and Tesla Backer Just Invested $50 Million in This Startup» («Spacex и Tesla Backer просто вложили 50 миллионов долларов в стартап»), *Fortune*, 24 сентября 2015 года.

Глава 4. Цифровая сыворотка правды

[1] Hugh J. Parry and Helen M. Crossley, «Validity of Responses to Survey Questions» («Достоверность ответов на вопросы анкеты»), *Public Opinion Quarterly* 14, 1 (1950).

[2] Frauke Kreuter, Stanley Presser, and Roger Tourangeau. «Social Desirability Bias in CATI, IVR, and Web Surveys» («Смещение социальной желательности в CATI, IVR и веб-опросах», *Public Opinion Quarterly* 72(5), 2008.

[3] Относительно статьи, утверждающей, что ложь может быть проблемой, мешающей предугадать поддержку Трампа, см. Thomas B. Edsall, «How Many People Support Trump but Don't Want to Admit It?» («Сколько людей поддерживают Трампа, но не хотят признать это?»), *New York Times*, 15 мая 2016 года, SR2. Но аргумент, что это не самый серьезный фактор, вы найдете в статье Andrew Gelman, «Explanations for That Shocking 2% Shift» («Объяснение того, что вызвало 2%-ный сдвиг»), с *Statistical Modeling, Causal Inference, and Social Science*, 9 ноября, 2016 года, http://andrewgelman.com/2016/11/09/explanations-shocking-2-shift/.

[4] Я брал интервью у Роджера Туранго по телефону 5 мая 2015 года.

[5] Это обсуждается в Adam Grant, «*Originals: How Non-Conformists Move the World*» («Оригиналы: как нонконформисты перемещаются по миру») (Нью-Йорк: Викинг 2016). Оригинальный источник — David Dunning, Chip Heath, and Jerry M. Suls, «Flawed Self-Assessment: Implications for Health, Education, and the Workplace» («Ущербная самооценка: последствия для здоровья, образования и работы»), *Psychological Science in the Public Interest* 5 (2004).

[6] Anya Kamenetz, «Mischievous Responders' Confound Research on Teens» («Хулиганство респондентов, ошибочные исследования подростков»), nprED, 22 мая 2014 года, http://www.npr.org/sections/ed/2014/05/22/313166161/mischievous-responders-confound-research-on-teens. Исходные исследования, обсуждаемые в данной статье — Joseph P. Robinson-Cimpian, «Inaccurate Estimation of Disparities Due to Mischievous Responders» («Неточная оценка неравенства из-за хулиганства респондентов»), исследователь в области образования *Education al Researcher* 43, № . 4 (2014).

[7] https://www.google.com/trends/explore?date=all&geo=US&q=porn, weather.

[8] Amanda Hess, «How Many Women Are Not Admitting to Pew That They Watch Porn?» («Сколько женщин не признаются в том, что они смотрят порно?»), 11 октября 2013, *Slate,* http://www.slate.com/blogs/xx_factor/2013/10/11/pew_online_viewing_study_percentage_of_women_who_watch_online_porn_is_growing.html.

[9] Nicholas Diakopoulus, «Sex, Violence, and Autocomplete Algorithms» («Секс, насилие и алгоритмы автозаполнения»), *Slate,* 2 августа 2013 года, http://www.slate.com/articles/technology/future_tense/2013/08/words_banned_from_bing_and_google_s_autocomplete_algorithms.html.

[10] По моим оценкам, каждый месяц насчитывается около 1730 поисковых запросов американцев в Google в различных формулировках, где явно говорится: они сожалеют, что завели детей. И только около 50 выражают сожаление о том, что у них нет детей. Насчитывается около 15,9 млн бездетных американцев в возрасте старше 45 лет. Есть около 152 миллионов американцев, у которых есть дети. Это означает, что люди, имеющие детей, примерно в 3,6 раза чаще выражают сожаление Google, чем люди без детей. Очевидно, как уже упоминалось в тексте книги и подчеркивается здесь еще раз, эти исповеди в Google очень небольшого числа людей. По-видимому, подобные чувства достаточно сильны — настолько, что люди на мгновение забыли: Google не может помочь им.

[11] Эти оценки получены из статьи Нейта Сильвера «How Opinion on Same-Sex Marriage Is Changing, and What It Means» («Как меняется отношение к однополым бракам и что это значит»), FiveThirtyEight, 26 марта 2013 года, http://fivethirtyeight.blogs.nytimes.com/2013/03/26/how-opinion-on-same-sex-marriage-is-changing-and-what-it-means/?_r=0.

[12] Как уже обсуждалось, Google Trends не может разбить информацию по полу. Google AdWords разбивает просмотр страниц для различных категорий по половому признаку, однако эти данные гораздо менее точны. Чтобы оценить запросы по полу, сначала я использовал данные поиска для получения региональной оценки процента поисковых запросов гей-порно по штатам. Потом проанализировал эти данные по полу с помощью Google Adwords. Еще один способ получить гендерную статистику — использовать данные PornHub. Однако выборка там может быть слишком селективной, поскольку многие геи могут вместо PornHub использовать сайты, ориентированные только на гей-порно. PornHub сообщает, что гей-порно мужчины спрашивают меньше, чем

можно предположить на основании поисковых запросов Google. Однако это подтверждает, что не существует тесной взаимосвязи между толерантностью к гомосексуализму и гей-порно. Все эти данные доступны на моем сайте sethsd.com в разделе «секс».

13 «У нас нет геев в Иране», — сказал иранский президент аудитории «Лиги плюща», Daily Mail.com, 25 сентября 2007 г., http://www.dailymail.co.uk/news/article-483746/We-dont-gays-Iran-Iranian-president-tells-Ivy-League-audience.html.

14 Brett Logiurato, «Sochi Mayor Claims There Are No Gay People in the City» («Мэр Сочи утверждает, что в его городе геев нет»), *Sports Illustrated*, 27 января 2014.

15 Согласно данным, полученным с помощью Google AdWords, ежегодно выполняются десятки тысяч поисковых запросов «гей-порно». В порнографических поисковых запросах процент гей-порно примерно одинаков и в Сочи, и в США. Google AdWords не включают данные по Ирану. PornHub также не сообщает сведения по Ирану. Однако PornMD изучал их поисковые данные и сообщил, что пять из десяти поисковых запросов в Иране были относительно гей-порно. Они включает «папочкина любовь» и «бизнесмен в отеле», как указано в опросе Джозефа Патрика Маккормика относительно гей порно «Survey Reveals Searches for Gay Porn Are Top in Countries Banning Homosexuality» («Наблюдение показывает, что больше всего поисковых запросов «гей-порно» исходит из штатов, где гомосексуализм запрещен», *PinkNews,* http://www.pinknews.co.uk/2013/03/13/survey-reveals-searches-for-gay-porn-are-top-in-countries-banning-homosexuality/. По данным Google Trends, около 2% порно-запросов в Иране относятся к гей-порно, что ниже, чем в Соединенных Штатах, но по-прежнему предполагает широкий интерес к этому виду секса.

16 Stephens-Davidowitz, «Searching for Sex.» («Поиск секса»). Данные для этого раздела можно найти на моем сайте sethsd.com, в разделе «секс».

17 Современные средства контрацепции среди женщин в возрасте 15–44 года: США, 2011–2013, центры по контролю и профилактике заболеваний, http://www.cdc.gov/nchs/data/databriefs/db173_table.pdf#1.

[18] David Spiegelhalter, «Sex: What Are the Chances?» («Секс: каковы шансы?»), BBC News, 15 марта 2012 года, http://www.bbc.com/future/story/20120313-sex-in-the-city-or-elsewhere.

[19] Примерно 6,6 млн беременностей каждый год и 62 млн женщин в возрасте от 15 до 44 лет.

[20] Как уже упоминалось, я не знаю пол выполняющих поисковые запросы в Google людей. Я предполагаю, что подавляющее большинство запросов относительно того, как выполнить куннилингус, принадлнежит мужчинам, а подавляющее большинство запросов относительно выполнения орального секса принадлежит женщинам. Дело в том, что большинство людей — гетеросексуалы. А гомосексуалам наверняка легче научиться доставлять удовольствие однополым партнерам.

[21] Анализ автора с помощью Google AdWords.

[22] Evan Soltas and Seth Stephens-Davidowitz, «The Rise of Hate Search» («Вспышка ненависти в поисковых запросах»), *New York Times*, 13 декабря 2015 года, SR1. Данные и более подробную информацию можно найти на моем сайте sethsd.com в разделе «Исламофобия».

[23] Авторский анализ тенденций данных Google.

[24] Авторский анализ данных Google Trends.

[25] Ashwin Rode and Anand J. Shukla, «Prejudicial Attitudes and Labor Market Outcomes» («Предрассудкив и ситуация на рынке труда»), mimeo, 2013, мимеографированный бюллетень, 2013.

[26] Seth Stephens-Davidowitz, «Google, Tell Me. Is My Son a Genius?» («Google, скажи мне, мой сын гений?»), *New York Times*, 19 января 2014 года, SR6.

[27] «Gender Equity in Education: A Data Snapshot» («Гендерное равенство в образовании: моментальный снимок данных»), отдел гражданских прав Департамента образования США, июнь 2012, http://www2.ed.gov/about/offices/list/ocr/docs/gender-equity-in-education.pdf.

[28] Центр данных здоровья детей и подростков, http://www.childhealthdata.org/browse/survey/results?q=2415&r=455&a=3879&p=1.

[29] Стивенс-Давидович «Данные о ненависти», соответствующие данные загружены в sethsd.com в разделе под названием «Stormfront».

[30] Количество запросов в Google о Stormfront в октябре 2015 и 2016 годов было примерно равно. Оно резко контрастирует с ситуацией во время первых выборов Обамы. В октябре 2008 года процент поисковых запросов о Stormfront возрос почти на 60% по сравнению с предыдущим октябрем. На следующий день после того, как Обама был избран, число поисковых запросов в Google со словом «Stormfront» возросло примерно в 10 раз. На следующий день после избрания Трампа количество запросов относительно Stormfront выросло примерно в 2,5 раза. Это было примерно как в день после избрания Джорджа Буша в 2004 году, и во многом может отражать интересы политических наркоманов.

[31] Matthew Gentzkow and Jesse M. Shapiro, «Ideological Segregation Online and Offline» («Идеологическая сегрегация онлайн и оффлайн»), *Quarterly Journal of Economics* 126, no. 4 (2011).

[32] См. Ben Quinn «Social Network Users Have Twice as Many Friends Online as in Real Life» («Пользователи социальных сетей имеют в два раза больше друзей в интернете, чем в реальной жизни»), *Guardian*, 8 мая, 2011 года. В данной статье рассматривается исследование 2011 года, выполненное Cystic Fibrosis Trust, которое показало: средний пользователь социальной сети имеет 121 онлайн-друга, тогда как в реальной жизни их всего 55. По данным исследования Pew Research 2014 года, средний пользователь Facebook имеет более 300 друзей. См. Aaron Smith, «6 New Facts About Facebook» («6 новых фактов о Facebook»), 3 февраля 2014 года, http://www.pewre search.org/fact-tank/2014/02/03/6-new-facts-about-facebook/.

[33] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic, «The Role of Social Networks in Information Diffusion» («Роль социальных сетей в распространении информации»), труды 21-й Международной конференции по Всемирной паутине, 2012.

[34] «Исследование: жестокое обращение с детьми и спад экономики в США», Associated Press, 12 декабря 2011.

[35] Seth Stephens-Davidowitz, «The Return of the D.I.Y. Abortion» («Возвращение криминальных абортов»), *New York Times*, 6 марта 2016 года, SR2. Данные и более подробную информацию можно найти на моем сайте sethsd.com, в разделе «самопроизвольный аборт».

[36] Alliance for Audited Media, Consumer Magazines, http://abcas3.auditedmedia.com/ecirc/magtitlesearch.asp.

[37] Расчеты автора 4 октября 2016 года, выполненные с помощью Ads Manager в Facebook.

[38] «Список самых популярных сайтов», Википедия. Согласно данным Alexa, который отслеживает поведение в интернете, 4 сентября 2016 года самым популярным порносайтом был XVideos, и это был 57-й по популярности сайт. По данным SimilarWeb по состоянию на 4 сентября 2016 года, самым популярным порносайтом был XVideos, и это был 17-й по популярности сайт. В первую десятку, согласно данным Alexa, входят Google, YouTube, Facebook, Baidu, Yahoo!, Amazon, Wikipedia, Tencent QQ, Google India и Twitter.

[39] Это история, рассказанная Дэвидом Киркпатриком, «*The Facebook Effect: The Inside Story of the Company That Is Connecting the World*» *(*«Эффект Facebook: внутренняя история компании, которая объединила мир»), New York: Simon & Schuster, 2010.

[40] Peter Thiel and Blake Masters, *Zero to One: Notes on Startups, or How to Build the Future*» *(*«От нуля до единицы: заметки о стартапах, или как построить будущее»), New York: The Crown Publishing Group, 2014).

[41] Я брал интервью у Ксавье Аматриэна по телефону 5 мая 2015 года.

[42] Авторский анализ тенденций данных Google.

[43] «Президент выступает перед исламским сообществом в Балтиморе», видео на Ютуб, опубликовано 3 февраля 2016 года, https://www.youtube.com/watch?v=LRRVdVqAjdw.

Глава 5. Приглядимся повнимательнее

[1] Seth Stephens-Davidowitz, «They Hook You When You're Young» («Они цепляют тебя в детстве»), *New York Times*, 20 апреля

2014 года, SR5. Данные для данного исследования можно найти на моем сайте sethsd.com в разделе «Бейсбол».

2 Yair Ghitza and Andrew Gelman, «The Great Society, Reagan's Revolution, and Generations of Presidential Voting» («Великое общество, революция Рейгана и поколения президентских выборов»), неопубликованная рукопись.

3 Я брал интервью у Раджа Четти по телефону 30 июля 2015 года.

4 Raj Chetty et al., «The Association Between Income and Life Expectancy in the United States» («Связь между доходом и продолжительностью жизни в США»), 2001–2014, *JAMA* 315, no. 16 (2016).

5 Julia Belluz, «Income Inequality Is Chipping Away at Americans' Life Expectancy» («Влияние неравенства доходов на продолжительность жизни американцев»), vox.com, 11 апреля 2016 года.

6 Raj Chetty, John Friedman, and Emmanuel Saez, «Using Differences in Knowledge Across Neighborhoods to Uncover the Impacts of the EITC on Earnings» («Использование различий в знаниях по местностям, раскрывающее влияние EITC на заработки»), *American Economic Review* 103, no. 7 (2013).

7 Это из Seth Stephens-Davidowitz, «The Geography of Fame» («География славы»), *New York Times*, 23 марта 2014 года, SR6. Данные можно найти на моем сайте seths.com в разделе «Википедия, рождаемость по округам». За помощь в скачивании и анализе округа рождения каждого участника «Википедии» я благодарю Ноя Стивенса-Давидовича.

8 Дополнительные доказательства ценности городов, см. Ed Glaeser, *Triumph of the City* (New York: Penguin, 2011). (Глэзер был моим научным руководителем в аспирантуре).

9 David Levinson, ed., *Encyclopedia of Crime and Punishment* (Thousand Oaks, CA: SAGE, 2002).

10 Craig Anderson et al., «The Influence of Media Violence on Youth» («Влияние насилия в СМИ на молодежь»), *Psychological Science in the Public Interest* 4 (2003).

11 Gordon Dahl and Stefano DellaVigna, «Does Movie Violence Increase Violent Crime?» («Насилие в кино и рост насильственных преступлений»), *Quarterly Journal of Economics* 124, no. 2 (2009).

[12] Seth Stephens-Davidowitz, «Days of Our Digital Lives» («Дни наших цифровых жизней»), *New York Times*, 5 июля 2015 года, sr4.

[13] Anna Richardson and Tracey Budd, «Young Adults, Alcohol, Crime and Disorder, *Criminal Behaviour and Mental Health*» («Молодежь, алкоголь, преступность и беспорядки, преступное поведение и психическое здоровье»), 13, no. 1 (2003); Richard A. Scribner, David P. MacKinnon, and James H. Dwyer, «TheRisk of Assaultive Violence and Alcohol Availability in Los Angeles County» («Риск агрессии, насилия и наличие алкоголя в округе Лос-Анджелес»), *American Journal of Public Health* 85, no. 3 (1995); Dennis M. Gorman, Paul W. Speer, Paul J. Gruenewald, and Erich W. Labouvie, «Spatial Dynamics of Alcohol Availability, Neighborhood Structure and Violent Crime» («Пространственная динамика доступности алкоголя, структура местности и насильственные преступления»), *Journal of Studies on Al cohol* 62, no. 5 (2001); Tony H. Grubesic, William Alex Pridemore, Dominique A. Williams, and Loni Philip-Tabb, «Alcohol Outlet Density and Violence: The Role of Risky Retailers and Alcohol-Related Expenditures» («Плотность продаж алкоголя и насилие: роль рискованной розничной торговли и расходы, связанные с алкоголем»), *Alcohol and Alcoholism* 48, no. 5 (2013).

[14] «Эд Маккэффри с самого начала знал, что Кристиан будет хорош», видео на YouTube, опубликовано 3 декабря, 2015 года, https://www.youtube.com/watch?v=boHMmp7DpX0.

[15] Исследователи получили очень многое от использования этих данных о преступлениях, разбитых на небольшие промежутки времени. Пример? Жалобы на бытовое насилие возрастают сразу после того, как футбольная команда города проигрывает в матче, в котором ожидалась победа. См. David Card and Gordon B. Dahl, «Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior» («Насилие в семье и футбол: эффект неожиданного агрессивного поведения»), *Quarterly Journal of Economics* 126, no. 1 (2011).

[16] Bill Simmons, «It's Hard to Say Goodbye to David Ortiz» («Трудно сказать «Прощай» Дэвиду Ортису»), ESPN.com, 2 июня 2009 года, http://www.espn.com/espnmag/story?id=4223584.

[17] Это обсуждается в Nate Silver, *«The Signal and the Noise: Why So Many Predictions Fail — But Some Don't»* («Сигнал и шум: почему многие предсказания не сбылись, а некоторые попали в точку»), New York: Penguin, 2012.

[18] Ryan Campbell, «How Will Prince Fielder Age?» («Сколько лет будет Принцу Филдеру?»), 28 октября 2011 года, http://www.fangraphs.com/blogs/how-will-prince-fielder-age/.

[19] Эти данные были любезно предоставлены мне Робом МакКоуном из «Baseball Prospectus».

[20] Я брал интервью у Исаака Когана по телефону 15 июня 2015 года.

[21] Я брал интервью у Джеймса Хейвуда по телефону 17 августа 2015 года.

Глава 6. Весь мир — лаборатория

[1] Эта история обсуждается, среди прочих, в книге Brian Christian «The A/B Test: Inside the Technology That's Changing the Rules of Business» («A/B-тест: внутри технологии, которая меняет правила бизнеса»), 25 апреля 2012, http:// www.wired.com/2012/04/ff_abtesting/.

[2] Esther Duflo, Rema Hanna, and Stephen P. Ryan, «Incentives Work: Getting Teachers to Come to School» («Стимулы работы: что заставляет педагогов приходить в школу»), *American Economic Review* 102, no. 4 (2012).

[3] Ian Parker, «The Poverty Lab» («Лаборатории бедности»), *New Yorker*, «Нью-Йоркер», 17 мая 2010 года.

[4] Christian, «The A/B Test».

[5] Douglas Bowman, «Goodbye, Google» («Прощай, Google»), 20 марта 2009 года, http://stopdesign.com/archive/2009/03/20/goodbye-google.html.

[6] Eytan Bakshy, «Big Experiments: Big Data's Friend for Making Decisions» («Большие эксперименты: друзья Больших Данных при принятии решений»), 3 апреля 2014 года, https://www.facebook.com/notes/facebook-data-science/big-experiments-big-datas-friend-for-making-decisions/10152160441298859/. Источники информации о фармацевтических исследованиях можно найти в «How many clinical

trials are started each year?» («Сколько клинических испытаний проводится каждый год?»), https://www.quora.com/How-many-clinical-trials-are-started-each-year.

[7] Я брал интервью у Дэна Сирокера по телефону 29 апреля 2015 года.

[8] Dan Siroker, «How Obama Raised $60 Million by Running a Simple Experiment» («Как Обама собрал 60 млн долларов, запустив простой эксперимент»), ноябрь 29, 2010, Optimizely blog, https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a -simple-experiment/.

[9] A/B-тесты и их результаты были представлены автором. Некоторые подробности о тестировании Boston Globe можно найти в «The Boston Globe: Discovering and Optimizing a Value Proposition for Content», Marketing Sherpa Video Archive, https://www.marketingsherpa.com/video/boston-globe-optimization-summit2. Сюда включены записанный разговор между Питером Дусеттом из «Глоб» и Памелой Марки из «MECLABS».

[10] Я брал интервью у Кларка Бенсона по телефону 23 июля 2015 года.

[11] «Увеличение текстовой рекламы в Google Display Network», внутри Adsense, 3 декабря 2012 года, https://adsense.googleblog.com/2012/12/enhancing-text-ads-on-google-display.html.

[12] См., например, «Large arrows appearing in google ads — please remove» («Большие стрелки появляющиеся в гугл-рекламе — пожалуйста, удалите»), DoubleClick Publisher Help Forum, https://productforums.google.com/forum/#!topic/dfp/p_TRMqWUF9s.

[13] Adam Alter «*Irresistible*: *The Rise of Addictive Technology and the Business of Keeping Us Hooked*» («Непреодолимо: рост технологий, вырабатывающих привыкание и умение держать нас на крючке»), New York: Penguin, 2017.

[14] Авторский анализ тенденций данных Google.

[15] Это обсуждается в видео, в настоящее время выставленном на странице Freakonomics в Harry Walker Speakers Bureau, http://www.harrywalker.com/speakers/authors-of-freakonomics/.

[16] Wesley R. Hartmann and Daniel Klapper. «Super Bowl Ads» («Реклама на Супер Боуле»), неопубликованная рукопись, 2014.

[17] Чтобы получить веские доводы в пользу того, что мы живем в компьютерной симуляции, см. Nick Bostrom, «Are We Living in a Computer Simulation?» («Возможно, мы живем в компьютерной симуляции?»), *Philosophical Quarterly* 53, no. 211 (2003).

[18] Сотрудники «Лос-Анджелес Таймс», «U.S. Presidential Assassinations and Attempts» («Убийства президентов США и покушения на них»), «Лос-Анджелес Таймс», 22 января 2012 г. http://timelines.latimes.com/us-presidential-assassinations-and-attempts/.

[19] Jones and Olken, «Do Assassins Really Change History?» («Действительно ли убийцы меняют историю?»), «Нью-Йорк Таймс», 12 апреля 2015 года, SR12.

[20] Эта история также описана в Jones and Olken, «Do Assassins Really Change History?» («Действительно ли убийцы меняют историю?»).

[21] Ужасное видео нападения можно увидеть на «Parade surprise (Чечня 2004) », видео на Ютуб, опубликованное 31 марта 2009 года, https://www.youtube.com/watch?v=fHWhs5QkfuY.

[22] с. 281 следствие гибели лидера: Benjamin F. Jones and Benjamin A. Olken. «Hit or Miss? The Effect of Assassinations on Institutions and War» («В яблочко или промах? Влияние убийств на государственные ингституты и войну»), *American Economic Journal: Macroeconomics* 1, no. 2 (2009).

[23] Этот момент рассмотрен Джоном Тирни в статье «How to Win the Lottery (Happily)» («Как выиграть в лотерею (счастливо)»), *New York Times*, 27 мая 2014 года, D5. Тирни обсуждает исследование Benedicte Apouey and Andrew E. Clark, «Winning Big but Feeling No Better? The Effect of Lottery Prizes on Physical and Mental Health» («Большой выигрыш не принес счастья? Влияние выигрыша в лотерею на физическое и душевное здоровье»), *Health Economics* 24, no. 5 (2015); Jonathan Gardner and Andrew J. Oswald. «Money and Mental Wellbeing: A Longitudinal Study of Medium-Sized Lottery Wins» («Деньги и психическое благополучие: долговременное исследование средних выигрышей в лотерею»), *Journal of Health Economics* 26, no. 1 (2007); а также Anna Hedenus, «At the End of the Rainbow: Post-Winning Life Among Swedish Lottery Winners»

(«Конец радуги: жизнь победителей шведской лотереи после выигрыша»), неопубликованная рукопись, 2011. Тирни также упоминает знаменитую статью 1978 года ученых Philip Brickman, Dan Coates, and Ronnie Janoff-Bulman: «Lottery Winners and Accident Victims: Is Happiness Relative?» («Победители лотереи и жертвы несчастных случаев: счастье относительно?», *Journal of Personality and Social Psychology* 36, no. 8 (1978), в которой показано, что выигрыш в лотерею не сделает вас счастливым. Эта статья основывалась на очень небольшой выборке.

[24] См. статью Peter Kuhn, Peter Kooreman, Adriaan Soetevent, and Arie Kapteyn, «The Effects of Lottery Prizes on Winners and Their Neighbors: Evidence from the Dutch Postcode Lottery» («Влияние выигрыша в лотерею на победителей и их соседей: опыт голландской Почтовой лотереи»), *American Economic Review* 101, no. 5 (2011), и Sumit Agarwal, Vyacheslav Mikhed, and Barry Scholnick, «Does Inequality Cause Financial Distress? Evidence from Lottery Winners and Neighboring Bankruptcies» («Вызывает ли неравенство финансовый кризис? Информация о победителях в лотереях и разрушении добрососедских отношений»), рабочие материалы, 2016.

[25] Agarwal, Mikhed, and Scholnick, Does Inequality Cause Financial Distress?» («Вызывает ли неравенство финансовый кризис?»)

[26] Jeffrey Clemens and Joshua D. Gottlieb. «Do Physicians' Financial Incentives Affect Medical Treatment and Patient Health?» («Влияют ли финансовые стимулы на лечение и здоровье пациентов?»), *American Economic Review* 104, no. 4 (2014). Отметим, что подобные результаты не означают, что врачи — злодеи. На самом деле, итоги могли бы быть и более удручающими, если бы дополнительные процедуры, назначенные докторами, когда им было заплачено за то, чтобы они их назначили, действительно спасали жизнь. Если бы это было так, это бы означало, что врачи должны получать оплату для того, чтобы назначить лечение, спасающее жизнь. Вместо этого результаты, полученные Клеменсом и Готлибом, показывают: врачи будут назначать процедуры, спасающие жизнь, независимо от того, сколько денег им дают на их осуществление. Что касается процедур, которые не так уж и нужны,

врачам нужно хорошо заплатить, чтобы они их назначили. Можно сформулировать это по-другому: врачи не обращают слишком много внимания на финансовые стимулы в случае ситуации, когда болезнь угрожает жизни пациента, но очень заинтересованы в оплате при обращении к ним по малозначимым причинам.

[27] Robert D. McFadden and Eben Shapiro, «Finally, a Face to Fit Stuyvesant; a High School of High Achievers Gets a High-Priced Home», *New York Times,* 8 сентября, 1992.

[28] Курсы доступны на веб-сайте Стайвесанта, http://stuy.enschool.org/index.jsp.

[29] Anna Bahr, «When the College Admissions Battle Starts at Age 3», *New York Times*, 29 июля 2014 года, http://www.nytimes.com/2014/07/30/upshot/when-the-college-admissions-battle-starts-at-age-3.html.

[30] Sewell Chan, «The Obama Team's New York Ties», *New York Times*, 25 ноября 2008 года; Evan T. R. Rosenman, «Class of 1984: Lisa Randall», *Harvard Crimson*, 2 июня 2009 года; «Gary Shteyngart on Stuyvesant High School: My New York», YouTube, видео размещено 4 августа 2010 года, https://www.youtube.com/watch?v=NQ_phGkC-Tk; Candace Amos, «30 Stars Who Attended NYC Public Schools», *New York Daily News*, 29 мая 2015 года.

[31] Carl Campanile, «Kids Stuy High Over Bubba; He'll Address Ground Zero School's Graduation», New York Post, 22 марта 2002 года; United Nations Press Release, «Stuyvesant High School's 'Multicultural Tapestry' Eloquent Response to Hatred, Says Secretary-General in Graduation Address», 23 июня 2004 года; «Conan O'Brien's Speech at Stuyvesant's Class of 2006 Graduation in Lincoln Center», YouTube, видео опубликовано 6 мая 2012 года, https://www.youtube.com/watch?v=zAMkUE9Oxnc.

[32] См. https://k12.niche.com/rankings/public-high-schools/best-overall/.

[33] Pamela Wheaton, «8th-Graders Get High School Admissions Results», Insideschools, 4 марта 2016 года, http://insideschools.org/blog/item/1001064–8th-graders-get-high-school-admissions-results.

[34] M. Keith Chen and Jesse M. Shapiro, «Do Harsher Prison Conditions Reduce Recidivism? A Discontinuity-Based Approach», *American Law and Economics Review* 9, no. 1 (2007).

[35] Atila Abdulkadiroglu, Joshua Angrist, and Parag Pathak. «The Elite Illusion: Achievement Effects at Boston and New York Exam Schools», *Econometrica* 82, no. 1 (2014). То же отсутствие результата независимо обнаружили Will Dobbie and Roland G. Fryer Jr., «The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools», *American Economic Journal: Applied Economics* 6, no. 3 (2014).

[36] См. http://www.payscale.com/college-salary-report/bachelors.

[37] Stacy Berg Dale and Alan B. Krueger. «Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables», *Quarterly Journal of Economics* 117, no. 4 (2002).

[38] Alice Schroeder, «*The Snowball: Warren Buffett and the Business of Life*», New York: Bantam, 2008.

Глава 7. Большие данные-шманные: Чего они не могут?

[1] Johan Bollen, Huina Mao, and Xiaojun Zeng, «Twitter Mood Predicts the Stock Market», *Journal of Computational Science* 2, no. 1 (2011).

[2] James Mackintosh, «Hedge Fund That Traded Based on Social Media Signals Didn't Work Out», *Financial Times*, 25 мая 2012.

[3] Christopher F. Chabris et al., «Most Reported Genetic Associations with General Intelligence Are Probably False Positives», *Psychological Science* (2012).

[4] Эта история обсуждается на TEDx Talks, «How to Make a Behavior Addictive: Zoe Chance at TEDx Mill River», Видео YouTube, пост от 14 мая 2013 года, https://www.youtube.com/watch?v=AHfiKav9fcQ. В интервью были конкретизированы некоторые детали истории — такие, как цвет шагомера. Я интервьюировал Чанс по телефону 20 апреля 2015 года, и по email 11 июля 2016 года и 8 сентября 2016 года.

[5] Это раздел взят из материала Алекса Пейсаховича и Сета Стивенс-Давидовича «Как не утонуть в цифрах», «Нью-Йорк Таймс», 3 мая 2015, ИР6.

[6] Brian A. Jacob and Steven D. Levitt, «Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating», *Quarterly Journal of Economics* 118, no. 3 (2003).

[7] Я брал интервью у Томаса Кейна по телефону 22 апреля 2015 года.

[8] Bill and Melinda Gates Foundation, «Ensuring Fair and Reliable Measures of Effective Teaching», http://k12education.gatesfoundation.org/wp-content/uploads/2015/05/MET_Ensuring_Fair_and_Reliable_MeasuresPractitionerBrief.pdf.

Глава 8. Больше Данных — больше проблем? Чего нам не стоит делать

[1] Oded Netzer, Alain Lemaire, and Michal Herzenstein, «When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications», 2016.

[2] Peter Renton, «Another Analysis of Default Rates at Lending Club and Prosper», 25 октября 2012, http://www.lendacademy.com/lending-club-prosper-default-rates/.

[3] Michal Kosinski, David Stillwell, and Thore Graepel, «Private Traits and Attributes Are Predictable from Digital Records of Human Behavior», *PNAS1*10, no. 15 (2013).

[4] Michael Luca, «Reviews, Reputation, and Revenue: The Case of Yelp», неопубликованная рукопись, 2011.

[5] Christine Ma-Kellams, Flora Or, Ji Hyun Baek, and Ichiro Kawachi, «Rethinking Suicide Surveillance: Google Search Data and Self-Reported Suicidality Differentially Estimate Completed Suicide Risk», *Clinical Psychological Science* 4, no. 3 (2016).

[6] Здесь используется методика, изложенная на моем сайте в примечаниях. Я сравниваю поиск в Google «самоубийства» с поиском «как завязать галстук». В 2015 году насчитывалось 6,6 миллионов запросов в Google «как завязать галстук». А поисков в категории «самоубийства» было в 6,5 раз больше. $6{,}5 \times 6{,}6 / 12 = 3{,}5$.

[7] Bridge Initiative Team, «When Islamophobia Turns Violent: The 2016 U. S. Presidential Election», May 2, 2016, можно найти на http://bridge.georgetown.edu/when-islamophobia-turns-violent-the-2016-u-s-presidential-elections/.

Заключение. Сколько людей дочитывают книгу до конца?

[1] Карл Поппер, «*Conjectures and Refutations*», London: Routledge & Keagan Paul, 1963.

[2] Сопоставить все случаи этой болезни в городе: Simon Rogers, «*John Snow's Data Journalism: The Cholera Map That Changed the World*», *Guardian*, 15 марта 2013 года.

[3] Я взял интервью у Бенджамина Джонса по телефону 1 июня 2015 года. Эта работа также обсуждается в статье Аарона Чаттерджи и Бенджамина Джонса, «*Harnessing Technology to Improve K-12 Education*» («Использование технологий для повышения уровня K-12 образования»), Hamilton Project Discussion Paper, 2012.

[4] Jordan Ellenberg, «The Summer's Most Unread Book Is…», *Wall Street Journal*, July 3, 2014.

Научно-популярное издание

IT БЕСТСЕЛЛЕР

Сет Стивенс-Давидовиц

ВСЕ ЛГУТ

ПОИСКОВИКИ, BIG DATA И ИНТЕРНЕТ ЗНАЮТ О ВАС ВСЕ

Директор редакции *Е. Капьёв*
Ответственный редактор *Е. Истомина*
Младший редактор *Е. Минина*
Художественный редактор *П. Петров*

В коллаже на переплете использована фотография:
Amanda Carden / Shutterstock.com
Используется по лицензии от Shutterstock.com

ООО «Издательство «Эксмо»
123308, Москва, ул. Зорге, д. 1. Тел.: 8 (495) 411-68-86.
Home page: **www.eksmo.ru** E-mail: **info@eksmo.ru**

Өндіруші: «ЭКСМО» АҚБ Баспасы, 123308, Мәскеу, Ресей, Зорге көшесі, 1 үй.
Тел.: 8 (495) 411-68-86.
Home page: www.eksmo.ru E-mail: info@eksmo.ru.
Тауар белгісі: «Эксмо»
Қазақстан Республикасында дистрибьютор және өнім бойынша
арыз-талаптарды қабылдаушының
өкілі «РДЦ-Алматы» ЖШС, Алматы қ., Домбровский көш., 3«а», литер Б, офис 1.
Тел.: 8(727) 2 51 59 89,90,91,92, факс: 8 (727) 251 58 12 вн. 107; E-mail: RDC-Almaty@eksmo.kz
Өнімнің жарамдылық мерзімі шектелмеген.
Сертификация туралы ақпарат сайтта: www.eksmo.ru/certification

Сведения о подтверждении соответствия издания согласно законодательству РФ
о техническом регулировании можно получить по адресу: http://eksmo.ru/certification/

Өндірген мемлекет: Ресей
Сертификация қарастырылмаған

Подписано в печать 18.01.2018.
Формат 60x90$^{1}/_{16}$. Печать офсетная. Усл. печ. л. 24,0.
Тираж экз. Заказ

ISBN 978-5-04-090836-3

ЛЮДИ СКЛОННЫ ПРЕУВЕЛИЧИВАТЬ И НЕ ДОГОВАРИВАТЬ, ОПРОСЫ НЕ ПОКАЗЫВАЮТ ВСЕЙ КАРТИНЫ, ИССЛЕДОВАНИЯ НЕДОСТАТОЧНО РЕПРЕЗЕНТАТИВНЫ – В ОБЩЕМ, ЛГУТ ВСЕ...

КРОМЕ BIG DATA!

ТАК ЛИ МЫ ТОЛЕРАНТНЫ, КАК ПОКАЗЫВАЮТ СОЦИОЛОГИЧЕСКИЕ ИССЛЕДОВАНИЯ?

ПОЧЕМУ В ОБЩЕНИИ С ДРУЗЬЯМИ ПОДДЕРЖИВАЕМ ОДНОГО КАНДИДАТА, А ГОЛОСУЕМ ЗА ДРУГОГО?

НАСТОЛЬКО ЛИ СЕКСУАЛЬНО АКТИВНЫ, КАК ЗАЯВЛЯЕМ ПРИ ОПРОСАХ?

Автор этой книги, специалист Google по Data Science, провел собственное исследование, опираясь на современную науку о данных, и пришел к сенсационным результатам.

Автор подводит нас к мысли, что данные – это жизненная необходимость для современного мира. И это правда. Современная наука о данных за последнее десятилетие сделала огромный скачок вперед. Сейчас проще найти отрасль деятельности, в которой применяются Большие данные, нежели обратное. И хотя эта книга не научит вас пользоваться технологиями обработки данных и механизмами их интерпретации, она расскажет о том, какими они бывают и в какую сторону все движется. В ней скорее примеры и размышления, а не инструкция по применению. А дальше все в ваших руках.

ЛЕОНИД ЧЁРНЫЙ, ДИРЕКТОР ПО РАЗВИТИЮ БИЗНЕСА «РАМБЛЕР ИНТЕРНЕТ ХОЛДИНГ»

БОМБОРА

Бомбора — это новое название Эксмо Non-fiction, лидера на рынке полезных и вдохновляющих книг. Мы любим книги и создаем их, чтобы вы могли творить, открывать мир, пробовать новое, расти. Быть счастливыми. Быть на волне.

bomborabooks
www.bombora.ru

ISBN 978-5-04-090836-3