
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 2019

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, 2019. — 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

Вступление 7
Предисловие. Контуры революции 11

ЧАСТЬ I. ДАННЫЕ, БОЛЬШИЕ И МАЛЫЕ

Глава 1. Интуиция вас обманывает 39

ЧАСТЬ II. МОГУЩЕСТВО БОЛЬШИХ ДАННЫХ

Глава 2. Возможно, Фрейд был прав? 63
Глава 3. Переосмысление данных 75
Тело как информация 84
Слова как данные 98
Изображения как данные 124
Глава 4. Цифровая сыворотка правды 133
Правда о сексе 142
Правда о ненависти и предубеждении 161
Правда об интернете 176
Правда о жестоком обращении с детьми и абортах 182
Правда о ваших друзьях на Facebook 188
Правда о ваших клиентах 192
Способны ли мы выдержать правду? 198

Глава 5. Приглядимся повнимательнее 207
Что на самом деле происходит в наших регионах, городах и поселках? 215
Как мы заполняем часы и минуты жизни 237
Наши двойники 245
Истории, рассказанные данными 255
Глава 6. Весь мир — лаборатория 257
Азбука A/B-тестирования 260
Жестокие, но проливающие свет натурные эксперименты 274

ЧАСТЬ III. БОЛЬШИЕ ДАННЫЕ: ОБРАЩАТЬСЯ С ОСТОРОЖНОСТЬЮ

Глава 7. Большие данные-шманные: Чего они не могут? 301
Проклятие числа размерностей 305
Чрезмерный акцент на том, что можно измерить 312
Глава 8 Больше данных — больше проблем? Чего нам не стоит делать? 319
Опасность вооруженных данными корпораций 319
Опасность вооруженных данными правительств 329

Заключение. Сколько людей дочитывают книгу до конца? 335
Благодарности 351
Примечания 355

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-Давидовиц 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.*