ТЕМНЫЕ ДАННЫЕ

Практическое руководство по принятию правильных решений в мире недостающих данных

ДЭВИД ХЭНД

ПРЕЗИДЕНТ БРИТАНСКОГО КОРОЛЕВСКОГО СТАТИСТИЧЕСКОГО ОБЩЕСТВА

Дэвид Хэнд

Темные данные. Практическое руководство по принятию правильных решений в мире недостающих данных

«Альпина Диджитал»

2020

УДК 004.6
ББК 32.972

Хэнд Д.

Темные данные. Практическое руководство по принятию правильных решений в мире недостающих данных / Д. Хэнд — «Альпина Диджитал», 2020

ISBN 978-5-96-145893-0

Человечество научилось собирать, обрабатывать и использовать в науке, бизнесе и повседневной жизни огромные массивы данных. Но что делать с данными, которых у нас нет? Допустимо ли игнорировать то, чего мы не замечаем? Британский статистик Дэвид Хэнд считает, что это по меньшей мере недальновидно, а порой – крайне опасно. В своей книге он выделяет 15 влияющих на наши решения и действия видов данных, которые остаются в тени. Например, речь идет об учете сигналов бедствия, которые могли бы подать жители бедных районов, если бы у них были смартфоны, результатах медицинского исследования, которые намеренно утаили или случайно исказили, или данных, ставших «темными» из-за плохого набора критериев для включения в выборку. Хэнд также рассказывает о том, какие меры могут сгладить эффект «темных данных» и как их можно обратить себе на пользу. Книга будет интересна широкому кругу читателей, интересующихся дата-сайенс, программированием и статистикой.

УДК 004.6
ББК 32.972

ISBN 978-5-96-145893-0

Содержание

Дэвид Хэнд
Темные данные. Практическое руководство по принятию правильных решений в мире недостающих данных

Переводчик *М. Белоголовский*
Редактор *В. Ионов*
Главный редактор *С. Турко*
Руководитель проекта *А. Василенко*
Корректоры *Е. Аксёнова, А. Кондратова*
Компьютерная верстка *К. Свищёвв*
Художественное оформление и макет *Ю. Буга*

⁂

Посвящается Шелли

Предисловие

Перед вами необычная книга. Почти все, что издается на эту тему – будь то популярная литература о больших или открытых данных, обработке данных или пособия по статистическому анализу, – основывается на том, что у вас уже есть. Речь идет об информации, хранящейся в компьютере, ящиках рабочего стола или аудио-, видеозаписях вашего смартфона. Но эта книга совсем о другом. Она о данных, *которых у вас нет.* Возможно, вы пытаетесь получить их прямо сейчас или когда-то безуспешно пытались сделать это, а может быть, ошибочно полагаете, что они у вас имеются. Как бы то ни было, речь пойдет о данных, которых у вас нет.

Я утверждаю и далее продемонстрирую это на многих примерах, что отсутствующие данные важны не менее тех, которыми мы располагаем. Вы сможете сами убедиться, что неизвестные нам данные являются причиной многих заблуждений, порой имеющих катастрофические последствия. Я покажу, как и почему это происходит. Затем я расскажу, как этого можно избежать – на что именно стоит обращать внимание, чтобы обойти неприятности. А в завершение, когда вы поймете, как возникают темные данные и как они создают нам проблемы, я покажу, как с их помощью перевернуть с ног на голову традиционное представление об анализе данных и, если вы достаточно проницательны, глубже вникнуть в свою область, улучшить процесс принятия решений и выбора действий.

Мое собственное понимание темных данных развивалось постепенно, на протяжении всей карьеры. Я благодарю всех, кто подкидывал мне проблемы, которые, как я постепенно осознал, были не чем иным, как проблемами темных данных. Я выражаю признательность всем, кто вместе со мной искал способы их решения. Сферы, где возникали эти проблемы, варьировались от медицинских исследований и фармацевтической промышленности до государственной и социальной политики, финансового сектора и производства – ни одна сфера человеческой деятельности не свободна от рисков, которые несут с собой темные данные.

Отдельно хочу поблагодарить тех, кто любезно согласился пожертвовать своим временем, чтобы прочитать рукопись этой книги, а именно Кристофороса Анагностопулоса, Нила Ченнона, Найла Адамса и трех анонимных читателей от издательства. Они помогли мне избежать неловкости перед вами, сократив число допущенных ошибок. Питер Таллак, мой агент, помог найти идеального издателя для этой работы, любезно давал мне советы и направлял работу над книгой в целом. Мой редактор из издательства Princeton University Press Ингрид Гнерлих была мудрым и ценным гидом в вопросах оформления проекта. Наконец, я особенно признателен своей жене профессору Шелли Ченнон, за ее вдумчивую критику моих рукописей. Благодаря ее вкладу книга стала значительно лучше.

Имперский колледж, Лондон

Часть I
Темные данные
Происхождение и последствия

Глава 1
Темные данные
Незримая сила, которая формирует наш мир

Призрак данных

Как-то во время прогулки я встретил странного пожилого человека, который что-то высыпал на пешеходную дорожку примерно через каждые 15 м. Я не смог сдержать любопытства и поинтересовался, что это он такое делает.

– Рассыпаю слоновий порошок, – совершенно серьезно ответил он. – Слоны не выносят его запах, поэтому держатся подальше.

– Постойте, но в наших краях нет слонов, – улыбнулся я.

– Вот именно! – воскликнул он. – Это очень эффективное средство.

Этот забавный случай служит хорошим прологом для вещей куда более серьезных, о которых я собираюсь рассказать.

Каждый год корь убивает почти 100 000 человек. Один из 500 заболевших умирает от осложнений, многие страдают от необратимой потери слуха или от поражения головного мозга. К счастью, для Соединенных Штатов это редкое заболевание – например, в 1999 г. было зарегистрировано всего 99 случаев. Однако внезапная вспышка кори в январе 2019 г. привела к тому, что в штате Вашингтон была объявлена чрезвычайная ситуация. Некоторые штаты также сообщили о резком увеличении числа случаев заражения корью[1]. Подобное отмечалось и в других местах. На Украине в середине февраля 2019 г. число заразившихся превысило 21 000[2]. В Европе в 2017 г. было отмечено 25 863 случая, а в 2018 г. – уже более 82 000[3]. С 1 января 2016 г. по конец марта 2017 г. в Румынии зарегистрировано более 4000 случаев заражения и 18 летальных исходов.

Корь – коварное заболевание, распространяющееся незаметно, поскольку симптомы проявляются лишь через несколько недель после инфицирования. Болезнь поражает организм намного раньше, чем обнаруживаются ее признаки.

Это не означает, что корь нельзя предотвратить. Простая вакцинация способна иммунизировать организм, эффективно снижая риск заражения. И, действительно, национальные программы вакцинации, подобные тем, которые проводились в Соединенных Штатах, доказали свой успех. В результате большинство родителей в странах, где осуществляются такие программы, никогда не видели и тем более не испытывали на себе ужасных последствий этого заболевания.

---

[1] https://blog.uvahealth.com/2019/01/30/measles-outbreaks/, accessed 16 April 2019.

[2] http://outbreaknewstoday.com/measles-outbreak-ukraine-21000-cases-2019/, accessed 16 April 2019.

[3] https://www.theglobeandmail.com/canada/article-canada-could-see-large-amount-of-measles-outbreaks-health-experts/, accessed 16 April 2019.

Именно поэтому, когда родителям рекомендуют делать детям прививку от кори – заболевания, которого они и в глаза не видели, которым не болели ни их друзья, ни соседи и которое Центр по контролю и профилактике заболеваний признал неэндемичным для Соединенных Штатов, – они принимают такой совет с изрядной долей скепсиса.

Вакцинировать от того, чего вроде бы нет? Это то же самое, что использовать слоновий порошок.

Правда, в отличие от слонов, риск заражения все-таки существует, причем такой же реальный, как и раньше. Просто информация и данные, которые нужны родителям для принятия решений, отсутствуют, и риски становятся неочевидными.

Для многочисленных видов отсутствующих данных я использую обобщающий термин «*темные данные*». Темные данные скрыты от нас, и этот факт означает, что мы рискуем недооценить опасность, сделать неправильный вывод и принять неверное решение. Иначе говоря, наше неведение становится причиной ошибок.

Понятие «темные данные» возникло из аналогии с другим, физическим, термином – темной материей. Около 27 % Вселенной состоит из этого таинственного вещества, которое не взаимодействует со светом или каким-либо другим электромагнитным излучением и потому остается невидимым. Поскольку темная материя не видна, когда-то астрономы не подозревали о ее существовании. Но затем наблюдения за вращением галактик показали, что звезды более удаленные от центра движутся ничуть не медленнее звезд, расположенных ближе к центру галактики, что противоречит нашему пониманию гравитации. Эта аномалия вращения галактик на сегодняшний день объясняется предположением, что галактики имеют более значительную массу, чем та, о которой мы можем судить по звездам и другим видимым в телескопы объектам. Поскольку эта дополнительная масса не видна, ее назвали темной материей. И она может быть весьма значительной: согласно оценкам, наша галактика Млечный Путь содержит в 10 раз больше темной материи, чем обычной.

Темные данные ведут себя аналогично темной материи: мы не видим их, они не обнаруживаются, но все же способны оказывать существенное влияние на наши выводы, решения и действия. И, как я покажу на дальнейших примерах, если не осознать саму вероятность существования чего-то неизвестного, то последствия такой слепоты могут быть катастрофическими и даже фатальными.

Цель этой книги – исследовать, как и почему возникают темные данные. Мы рассмотрим различные виды темных данных, проследим, что приводит к их появлению, и выясним, как не допустить этого. Мы разберемся с тем, какие меры имеет смысл предпринимать, когда становится ясно, что темные данные все же имеются. А еще мы посмотрим, как этими данными, несмотря на их отсутствие, можно воспользоваться. Хотя это кажется странным, даже парадоксальным, но мы можем обернуть наше незнание себе во благо, учась принимать более правильные решения и повышая эффективность своих действий. На практике разумное использование неизвестности означает более крепкое здоровье, дополнительные деньги и меньшие риски. Я вовсе не имею в виду сокрытие информации от других (хотя, как мы увидим, намеренно скрытые сведения – это весьма распространенный вид темных данных). Речь идет о гораздо более тонких методах, которые могут стать выгодными для всех.

Темные данные принимают различные формы, возникают по разным причинам, и эта книга среди прочего содержит классификацию *типов* темных данных, обозначаемых как *DD-тип x*. Всего я насчитал 15 таких *DD-типов*, но не берусь утверждать, что эта классификация является исчерпывающей. Учитывая большое разнообразие причин, по которым возникают темные данные, не исключено, что полная классификация просто невозможна. Более того, многие образцы темных данных соединяют в себе несколько *DD-типов* – они могут действовать независимо друг от друга, а могут проявлять некое подобие синергии, усиливая негативный эффект. Но, несмотря на это, обладание информацией о *DD-типах* и изучение темных данных

на конкретных примерах помогает вовремя выявить проблему и защититься от возможных угроз. Список *DD-типов,* упорядоченных по сходству, вы найдете в конце этой главы, а в главе 10 я опишу их более подробно. В книге есть указания на то, где можно встретить примеры того или иного *типа*, однако я намеренно не пытался перечислить все возможные места существования темных данных – в этой книге такой подход был бы излишним.

Давайте перейдем к одному из таких примеров. В медицине понятие «травма» означает повреждение с возможными долговременными последствиями. Травмы являются одной из наиболее серьезных причин сокращения продолжительности жизни и инвалидности, а также самой распространенной причиной гибели людей в возрасте до 40 лет. Компьютерная база данных TARN является самой большой медицинской базой данных о травмах в Европе. В нее стекаются данные о полученных травмах из более чем 200 больниц, в числе которых 93 % всех больниц Англии и Уэльса, а также больницы в Ирландии, Нидерландах и Швейцарии. Безусловно, это очень большой объем данных для прогнозирования и изучения эффективности медицинского вмешательства при травмах.

Доктор Евгений Миркес и его коллеги из Лестерского университета в Великобритании провели исследование этой базы данных и выяснили: из 165 559 зарегистрированных травм исход 19 289 случаев оказался неизвестным[4]. «Исход» в данном случае определяется тем, выживает пациент или нет в течение 30 дней после травмы. Иначе говоря, 30-дневная выживаемость неизвестна для более чем 11 % пациентов. Этот пример иллюстрирует распространенную форму темных данных – *DD-тип 1: данные, о которых мы знаем, что они отсутствуют*. Иначе говоря, нам известно, что травмы у этих пациентов чем-то закончились, – мы просто не знаем, чем именно.

Можно, конечно, сказать: «Нет проблем, давайте просто проанализируем 146 270 пациентов, для которых исход известен, и будем делать выводы и прогнозы на основе этой информации». В конце концов, 146 270 тоже немало – в сфере медицины это уже большие данные. Поэтому мы можем смело утверждать, что понимание, основанное на этих данных, будет верным.

Но так ли это на самом деле? Возможно, 19 289 недостающих случаев сильно отличаются от других. В конце концов, их необычность уже в самой неизвестности исхода, так почему же они не могут отличаться и чем-то другим? Как следствие, анализ 146 270 пациентов с известными исходами может быть ошибочным по отношению к общей совокупности пациентов с травмами. Таким образом, действия, предпринимаемые на основе подобного анализа, могут быть в корне неверными и привести к ошибочным прогнозам, ложным предписаниям и несоответствующим режимам лечения с неблагоприятными и даже фатальными последствиями для пациентов.

Давайте возьмем нарочито неправдоподобную, крайнюю ситуацию: предположим, что все 146 270 человек с известными исходами выжили и выздоровели без лечения, а 19 289 с неизвестными исходами умерли в течение двух дней после обращения в больницу. Если бы мы игнорировали последних, то неизбежно пришли бы к выводу, что беспокоиться не о чем – ведь все пациенты с травмами выздоравливают сами собой. Исходя из этого, мы бы просто не стали их лечить, ожидая естественного выздоровления. И вскоре были бы шокированы и озадачены тем фактом, что более 11 % пациентов умерли.

Прежде чем продолжить, я должен вас успокоить – в реальности все обстоит не так уж плохо. Во-первых, приведенный выше сценарий действительно наихудший из возможных, а во-вторых, доктор Миркес и его коллеги являются экспертами по анализу недостающих данных. Они прекрасно осознают опасность и разрабатывают статистические методы решения

[4] E. M. Mirkes, T. J. Coats, J. Levesley, and A. N. Gorban, “Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes.” *Computers in Biology and Medicine* 75 (2016): 203-16.

проблемы, о которых мы поговорим позже. Я привел такой ужасающий пример лишь для того, чтобы показать: *вещи могут быть не такими, какими кажутся*. В самом деле, если бы мне нужно было сформулировать основную идею этой книги, она бы, пожалуй, звучала примерно так: хотя иметь много данных полезно, большие данные, то есть объем, – это еще далеко не все. И то, чего вы не знаете, те данные, которых у вас нет, могут быть важнее для понимания происходящего, чем те, которыми вы располагаете. Во всяком случае, как мы увидим дальше, проблемы темных данных – это не только проблемы больших данных: они характерны и для малых наборов данных. Они вездесущи.

Пример с базой данных TARN, конечно, преувеличен, но он служит предупреждением. Возможно, результаты 19 289 пациентов не были зарегистрированы именно *потому*, что все они умерли в течение 30 дней. Ведь если исход заносился в базу на основании опроса пациентов через 30 дней после обращения, чтобы оценить их состояние, то никто из умерших просто не ответил на вопросы. Если бы мы не допускали возможность этого, то никогда бы не фиксировали смерть таких пациентов.

На первый взгляд это кажется нелепым, но в реальности такие ситуации возникают довольно часто. Допустим, модель прогнозирования эффективности того или иного лечения основывается на результатах предыдущих пациентов, которые получали такое лечение. Но что, если время лечения предыдущих пациентов было недостаточным для достижения результата? Тогда для некоторых из них конечный исход окажется неизвестен, а модель, построенная только на известных результатах, будет вводить в заблуждение.

Похожая ситуация возникает и с опросами, когда *отсутствие ответов* становится источником затруднений. Исследователи обычно имеют некий идеальный список людей, от которых они хотели бы получить ответы, но, как правило, отвечают не все. Если все те, кто отвечает, каким-то образом отличаются от тех, кто этого не делает, то у исследователей появляется основание усомниться в достоверности статистической сводки для данной группы населения. В конце концов, если бы некий журнал затеял опрос своих подписчиков, задав им единственный вопрос: «Отвечаете ли вы на журнальные опросы?», тот факт, что 100 % ответивших скажут «да», еще не говорил бы о том, что все подписчики отвечают на подобные опросы.

Предыдущие примеры иллюстрируют первый тип темных данных. Мы знаем, что данные для пациентов TARN существуют, даже если не все значения учтены. Мы знаем, что у людей в списке опроса были ответы, даже если они их не давали. В общем, мы знаем, что существуют некоторые значения данных, но не знаем, какие именно.

Следующие примеры познакомят нас с другим типом темных данных – *DD-тип 2: данные, о которых мы не знаем, что они отсутствуют.*

Многие города сталкиваются с проблемой выбоин в дорожном покрытии. Вода попадает в мелкие трещины, замерзает зимой, расширяя их, а колеса автомобилей довершают разрушительную работу. В результате у машин портятся колеса и подвеска. Бостон решил бороться с этой проблемой с помощью современных технологий. Он выпустил приложение для смартфона, которое использовало внутренний акселерометр устройства, чтобы определять тряску автомобиля, проехавшего по выбоине, а затем с помощью GPS автоматически передавать ее координаты городским властям.

Фантастика! Теперь люди, обслуживающие шоссе, будут точно знать, куда ехать, чтобы залатать выбоины. Однако это элегантное и дешевое решение реальной проблемы, основанное на современных технологиях анализа данных, не учитывает того, что владельцы автомобилей и дорогих моделей смартфонов с акселерометрами концентрируются в более богатых районах. Это повышает вероятность того, что выбоины на дорогах в районах победнее не будут обнаружены, а значит, аварийная опасность таких дорог будет все возрастать. Вместо того чтобы решить проблему в целом, такой подход усугубляет социальное неравенство. Ситуация в этом

примере отличается от ситуации с базой данных TARN, когда мы точно знали, что отсутствуют некоторые данные. Здесь мы этого не знаем.

Вот еще одна иллюстрация темных данных такого рода. В конце октября 2012 г. сильнейший ураган, получивший название «Сэнди»[5], обрушился на восточное побережье Соединенных Штатов. На тот момент это был второй по разрушительности ураган в истории США и крупнейший в истории атлантический ураган, причинивший ущерб в $75 млрд и унесший жизни более 200 человек в восьми странах. «Сэнди» затронул 24 штата (от Флориды на юге до Висконсина и штата Мэн на севере страны) и спровоцировал закрытие финансовых рынков из-за отключения электроэнергии. Надо признать, что поэтому он стал еще и косвенной причиной всплеска рождаемости спустя девять месяцев после описываемых событий.

Ураган «Сэнди» также стал настоящим триумфом современных СМИ. Ураган сопровождался шквалом сообщений в твиттер, который позволяет обсуждать происходящее сразу же и с тем, кто непосредственно участвует в событии. Вообще, социальные платформы – это способ быть в курсе событий в реальном времени, и «Сэнди» стал именно таким событием. В период с 27 октября по 1 ноября 2012 г. было опубликовано более 20 млн твитов об урагане. Очевидно, что это идеальный материал, на основе которого можно получить непрерывную картину стихийного бедствия по мере его развития – вы видите, какие районы пострадали больше всего и куда направить экстренную помощь.

Однако спустя какое-то время анализ показал, что наибольшее количество твитов о «Сэнди» пришло с Манхэттена и лишь немногие поступали из таких районов, как Рокуэй и Кони-Айленд. Означало ли это, что Рокуэй и Кони-Айленд пострадали не так серьезно? Метро и улицы Манхэттена были затоплены, это правда, но едва ли его можно назвать самым пострадавшим районом даже в пределах Нью-Йорка. Причина того, что из каких-то районов было послано меньше твитов, заключалась не в том, что ураган пощадил их, а в том, что на их территории оказалось меньше пользователей твиттера и меньшее число смартфонов, чтобы отправить твит.

Давайте снова представим себе крайний вариант этой ситуации. Если бы ураган «Сэнди» полностью уничтожил какой-нибудь населенный пункт, то оттуда вообще бы не поступало никаких твитов и создалось бы впечатление, что там все просто замечательно. Но на самом деле мы опять имеем дело с темными данными.

Примеры второго типа темных данных, когда мы не знаем, что чего-то не достает, встречаются не менее часто, чем примеры первого типа. Они варьируются от необнаруженных мошенничеств до незафиксированных убийств, выпадающих из результатов опроса жертв преступлений.

Как-то на информационном брифинге бывший министр обороны США Дональд Рамсфелд охарактеризовал темные данные второго типа, да так удачно, что его высказывание стало знаменитым: «Есть известные неизвестные; то есть мы знаем, что есть какие-то вещи, которых мы не знаем. Но есть также неизвестные неизвестные – те, о которых мы не знаем, что мы их не знаем»[6]. Этот замысловатый пассаж стал объектом насмешек для разнообразных СМИ, но их критика была несправедливой. То, что сказал Рамсфелд, было сущей правдой и имело глубокий смысл.

Эти первые два типа темных данных только начало. Далее мы познакомимся со множеством других, которые вкупе и составляют основу этой книги. Как вы увидите, темные данные разнообразны и до тех пор, пока мы не осознаем, что наши данные могут быть неполными; наблюдение чего-либо не означает наблюдения всего; процедура измерения может быть неточной; а то, что мы измеряем, на самом деле может оказаться не тем, что мы хотим измерить, мы

[5] https://www.livescience.com/24380-hurricane-sandy-status-data.html.

[6] D. Rumsfeld, Department of Defense News Briefing, 12 February 2002.

рискуем получать результаты, далекие от истины, что зачастую и происходит. Тот факт, что никто не слышит, как в лесу падает дерево, не означает, что оно падает бесшумно.

Так вы думаете, у вас есть все данные?

Покупатель подходит к кассе супермаркета, выкладывает на ленту выбранные товары, лазер сканирует их штрихкоды, и каждый раз кассовый аппарат издает звуковой сигнал, сообщая, что суммирует цены. В результате этой процедуры покупатель получает чек и расплачивается. Однако история его покупки на этом не заканчивается. Данные о купленных товарах и их стоимости отправляются в базу данных. Позже статистики и аналитики будут изучать их, создавая картину поведения покупателей на основе того, что они купили, какие из товаров были куплены вместе и, конечно, какие клиенты покупали эти товары. Казалось бы, здесь просто нельзя ничего пропустить. Данные о транзакциях собираются во всех случаях, кроме отключения электроэнергии, сбоя кассового аппарата или мошенничества.

Вроде бы собираются все данные. Иначе говоря, в базу попадают данные не по *некоторым* транзакциям или *некоторым* купленным товарам, а по *всем* транзакциям, совершенным *всеми* покупателями, и по *всем* товарам в конкретном супермаркете. Такие данные еще называют исчерпывающими.

Однако так ли это? Ведь собранные данные описывают то, что произошло на *прошлой* неделе или в *прошлом* месяце. Конечно, польза от них несомненна, но если мы управляем супермаркетом, то, вероятно, нам будет интересно, что произойдет завтра, на следующей неделе или через месяц. Мы бы хотели знать, кто, что, когда и сколько купит в будущем. Какие товары могут закончиться на полках, если не заказать их впрок? Как могут измениться предпочтения людей в отношении брендов? Другими словами, нам нужны данные, которые не собираются. Это связано с самой природой времени, и здесь фигурируют темные данные *DD-тип 7: данные, меняющиеся со временем*.

Помимо этого, интересно узнать, *как вели бы себя* люди, если бы мы, скажем, более плотно заставили товарами полки, или разместили их как-то иначе, или изменили часы работы супермаркета. Такие данные называются *контрфактуальными*, поскольку они противоречат реальным фактам – они о том, что случилось бы, если бы произошло нечто, чего на самом деле не происходило. Контрфактуальные данные классифицируются как *DD-тип 6: данные, которые могли бы существовать*.

Излишне говорить, что контрфактуальные данные интересуют не только менеджеров супермаркетов. Все мы принимаем те или иные лекарства и при этом, разумеется, доверяем врачу, который их прописал, предполагая, что лекарства прошли тестирование и были признаны эффективными. Но как бы вы себя чувствовали, если бы вдруг обнаружили, что ваши лекарства не были проверены? И не было собрано данных о том, помогают ли они вообще? Вдруг они делают только хуже? А если они даже и были протестированы и рекомендованы, то ускоряют ли эти лекарства на самом деле процесс выздоровления? А может быть, их не сравнивали с другими препаратами, чтобы оценить эффективность? В истории со слоновьим порошком такое сравнение принятых мер с бездействием быстро показывает, что для отпугивания слонов *отсутствие действия так же эффективно,* как и применение порошка. (А это, в свою очередь, может привести к следующему, не менее полезному выводу, что никаких слонов, которых надо отпугивать, просто нет.)

Возвращаясь к понятию «исчерпывающие данные», стоит отметить, что часто контекст делает *явно* бессмысленной саму возможность иметь «все» данные. Возьмите, например, свой вес. Узнать его легко – достаточно встать на весы. Однако уже не так легко будет повторно получить те же данные. Даже если сразу же встать на весы снова, результат, скорее всего, будет немного другим, особенно если попытаться измерить его с точностью до грамма. Никакие

физические измерения нельзя считать абсолютно точными в результате погрешностей или случайных колебаний, возникающих вследствие очень незначительных изменений условий (*DD-тип 10: ошибки измерения и неопределенность*). Для решения этой проблемы ученые, измеряющие параметры какого-либо явления – скажем, скорость света или заряд электрона, проводят серию измерений, а затем усредняют значения. Можно сделать тысячи и миллионы измерений, но очевидно, что невозможно сделать «все» измерения. В этом контексте просто не существует понятия «все», а значит, не существует и исчерпывающих данных.

Следующий тип темных данных хорошо иллюстрируется примером знаменитых лондонских автобусов. Если вам доводилось на них ездить, то, скорее всего, вы помните, что они, как правило, набиты битком. И все же данные показывают, что средняя заполняемость одного автобуса составляет всего 17 человек. Но чем можно объяснить это кажущееся противоречие? Кто-то манипулирует цифрами?

Немного поразмыслив, вы поймете, что ответ довольно прост – в основном мы попадаем в автобусы в часы пик, именно поэтому они и набиты битком. Вот почему большинство людей видит автобусы переполненными. В то же время о пустом автобусе будет просто некому сообщить, что он пуст (разумеется, не считая водителя). Этот пример иллюстрирует темные данные *DD-тип 3: выборочные факты*. Иногда, впрочем, это может быть необходимым следствием сбора данных, и в таком случае мы получаем *DD-тип 4: самоотбор*. Я приведу два моих любимых примера, похожих и в то же время несопоставимых по своему масштабу.

Первый – известная карикатура, на которой изображен человек, стоящий перед большой картой, какие обычно висят на вокзалах. В центре карты находится красная точка с надписью «Вы здесь». «Как?! – думает потрясенный человек. – Как они узнали это?» Они узнали, потому что отталкивались от простого факта, что *каждый*, кто смотрит на эту красную точку, должен находиться непосредственно перед ней. Мы имеем дело с очень узкой выборкой, отсекающей всех, кто находится в другом месте.

Данные могут быть собраны, только если имеется кто-то или что-то для их сбора, например измерительный прибор. Второй пример самоотбора связан с *антропным принципом*, который, по сути, говорит, что Вселенная должна быть такой, какая она есть, а иначе нас бы просто не существовало и мы бы не смогли наблюдать ее. У нас нет данных из разных вселенных по одной простой причине – мы там не были. Это означает, что любые выводы, которые мы делаем, неизбежно ограничиваются нашей Вселенной (а точнее, вселенными такого же типа): как и в случае с бостонскими выбоинами, может происходить масса всего, о чем мы не знаем.

Из этого примера наука может извлечь для себя важный урок. Теория может идеально согласовываться с данными, но сами данные имеют ограничения. И это относится не только к сверхвысоким температурам, геологическим эпохам или космическим расстояниям. Если вы экстраполируете теорию за пределы, в которых были собраны данные, то всегда есть вероятность того, что она окажется недействительной. Экономические теории, основанные на данных, собранных в период процветания, часто оказываются несостоятельными во время рецессии, а законы Ньютона работают только тогда, когда речь не идет о крошечных объектах, высоких скоростях и прочих крайностях. В этом и заключается суть темных данных *DD-тип 15: экстраполяция за пределы ваших данных*.

У меня есть классная футболка от сайта веб-комиксов xkcd.com, на которой общаются два персонажа. Один говорит: «Раньше я думал, что корреляция подразумевает причинность». В следующем кадре он продолжает: «Потом я прошел курс статистики, и теперь я в этом не уверен». Другой персонаж говорит ему: «Похоже, курс помог», а первый отвечает: «Возможно, но не факт»[7].

[7] http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636, accessed 31 July 2018.

Корреляция просто показывает, что две вещи меняются синхронно, например положительная корреляция означает, что когда одно становится большим, то и другое увеличивается, а когда первое уменьшается, то и второе поступает точно так же. Это в корне отличается от причинно-следственной связи. Говорят, что одно становится *причиной* другого, если изменения первого приводят к изменениям второго. Но проблема в том, что две вещи могут изменяться вместе, но при этом изменения одной не являются причиной изменений другой. Например, наблюдения в начальной школе показывают, что дети с более значительным словарным запасом в среднем выше. Но вряд ли вам придет в голову, что причиной этого являются родители, которые, желая иметь более рослое потомство, нанимают репетиторов для расширения словарного запаса своих детей. Намного вероятнее, что существуют какие-то темные данные, третий фактор, который объясняет корреляцию, например разный возраст детей. Когда персонаж на моей майке говорит «Возможно, но не факт», он признает, что пройденный курс статистики мог изменить его понимание, но при этом допускает наличие и других причин. Далее в книге мы еще столкнемся с поразительными примерами темных данных этого типа, а именно с *DD-типом 5: неизвестный определяющий фактор*.

Существуют и другие типы темных данных, о которых мы будем говорить. Напомню, что цель этой книги – рассказать о существующей на сегодня классификации темных данных, объяснить способы их идентификации, наглядно продемонстрировать оказываемое ими влияние и показать пути решения проблем, которые они вызывают, а также то, как темные данные можно использовать. Список типов темных данных приводится в конце этой главы, а краткое описание каждого из них вы найдете в главе 10.

Не было ничего необычного, поэтому мы не придали этому значения

Следующий пример служит иллюстрацией того, что темные данные могут иметь катастрофические последствия и что они не являются специфической проблемой больших наборов данных.

28 января 1986 г. на 73-й секунде полета на высоте около 15 км космический челнок Challenger превратился в гигантский огненный шар в результате неисправности ракеты-носителя. Отсек с экипажем какое-то время еще продолжал двигаться по восходящей траектории, достиг отметки 19 км и рухнул в Атлантику. Все семь членов экипажа погибли.

Впоследствии президентская комиссия установила, что руководители среднего звена NASA нарушили правила безопасности, требующие передачи данных по цепочке управления. Все объяснялось экономическими причинами: необходимо было уложиться в график, ведь дата старта уже переносилась с 22-го на 23-е, потом на 25-е, а затем и на 26 января. Поскольку прогноз погоды на этот день обещал неприемлемо низкую температуру, запуск снова отложили на день. Обратный отсчет прошел нормально, индикаторы показали, что замок люка закрылся должным образом. Однако к тому моменту поднялся сильный ветер, и запуск шаттла вновь пришлось отложить.

В ночь на 27 января состоялась трехчасовая телеконференция между представителями компании Morton Thiokol, построившей разгонные ступени, сотрудниками NASA в Центре космических полетов Маршалла и людьми из Космического центра Кеннеди. Ларри Уир из Центра космических полетов Маршалла попросил представителей Morton Thiokol проверить возможное влияние низких температур на твердотопливные ракетные двигатели. В ответ команда Morton Thiokol указала на то, что при низких температурах уплотнительные кольца становятся более жесткими.

Уплотнительные кольца представляли собой манжеты из резиноподобного материала с диаметром поперечного сечения около 6 мм, которые устанавливались по окружности в

стыки между четырьмя сегментами ракетного двигателя. Твердотопливные ракетные ускорители имели 45 м в высоту и 11 м в диаметре. Во время запуска зазор величиной 0,1 мм, который в обычных условиях полностью герметизировался уплотнительными кольцами, открывался максимум до 1,5 мм и оставался открытым в течение каких-то 0,6 секунды.

Роберта Эбелинга из Morton Thiokol беспокоило то, что при низких температурах повышение жесткости уплотнительных колец может привести к потере способности герметизировать зазоры между сегментами, пока они будут в течение 0,6 секунды оставаться увеличенными на 1,4 мм. На телеконференции Роберт Лунд, вице-президент Morton Thiokol, заявил, что рабочая температура уплотнительного кольца не должна быть ниже границы подтвержденной температуры запуска 53 ºF (около 12 ºC). За этим последовала довольно горячая дискуссия, продолжавшаяся и после окончания конференции на уровне личных бесед. По ее итогам Morton Thiokol пересмотрела свою позицию и согласилась рекомендовать запуск.

Ровно через 58,79 секунды с момента старта из правого ракетного двигателя в районе последнего стыка вырвалось пламя. Оно быстро превратилось в мощную струю, которая выломала стойки, соединяющие ракетный двигатель с внешним топливным баком. Двигатель развернуло и ударило сначала о крыло орбитального аппарата, а затем о топливный бак, в результате чего этот резервуар, наполненный жидкими водородом и кислородом, попал в струю пламени. На 64-й секунде полета поверхность бака получила повреждения, а еще через 9 секунд огромный огненный шар поглотил Challenger, и он разлетелся на несколько больших частей[8].

Мы не должны ни на секунду забывать, что космические полеты всегда связаны с риском. Ни одна миссия, даже при самых хороших условиях, не является безопасным предприятием – риск просто не может быть сведен к нулю. И всегда существуют противоречивые требования.

Кроме того, как и в любом другом подобном инциденте, установить какую-то одну причину произошедшего бывает довольно сложно. Было ли это вызвано нарушением правил безопасности, неоправданным давлением на менеджеров по экономическим соображениям, следствием ужесточения бюджета или, возможно, влиянием СМИ, которые после семикратного откладывания запуска предыдущего челнока Columbia встречали каждую новую задержку саркастическими насмешками? Вот что сказал, например, известный журналист Дэн Ратер в выпуске вечерних новостей в понедельник, 27 января, после того, как старт Challenger был отложен в четвертый раз: «Еще одна дорогостоящая и позорная задержка запуска космического челнока. На этот раз виноватыми оказались плохой болт на крышке люка и гром среди ясного неба». А может быть, причина кроется в политическом давлении? В конце концов, интерес к этому запуску был значительно выше, чем к предшествующим, потому что в число экипажа впервые вошел рядовой гражданин США, учительница Криста Макалиф и на вечер 28 января было запланировано выступление президента.

В таких ситуациях обычно переплетаются несколько факторов. Их запутанные и неопределенные взаимодействия могут привести к неожиданным последствиям. Но в нашем случае был еще один фактор: темные данные.

После катастрофы комиссия, возглавляемая бывшим госсекретарем Уильямом Роджерсом, обратила внимание на то, что не все результаты полетов, которые показывали опасное состояние уплотнительных колец, были включены в диаграмму, обсуждаемую на телеконференции (темные данные *DD-тип 3: выборочные факты*, а также *DD-тип 2: данные, о которых мы не знаем, что они отсутствуют*). На с. 146 отчета сказано следующее: «Менеджеры сопоставляли с температурой окружающей среды лишь те полеты, во время которых были зафиксированы критические состояния уплотнительных колец, но не рассматривали частоту их воз-

[8] https://er.jsc.nasa.gov/seh/explode.html.

никновения на основе данных всех полетов»[9]. Именно в этом и заключается истинная причина трагедии: *данные некоторых полетов не были включены в анализ*. Ранее я уже показал, к каким проблемам может привести такое игнорирование данных.

Далее в докладе говорится: «При таком сопоставлении [то есть с использованием ограниченного набора данных] не было заметно отклонений от нормы в распределении критических состояний уплотнительного кольца по всему диапазону температур при запуске от 53 до 75 ºF [от 12 до 24 ºC}». Это означает, что нет очевидной зависимости между температурой воздуха и числом уплотнительных колец, показывающих критическое состояние. Тем не менее «если рассматривать всю историю полетов, включая “нормальные” полеты без каких-либо разрушений или прорывов газа, результаты сопоставления существенно отличаются». Иначе говоря, если вы включите все данные, то получите другую картину. Фактически не включенные в анализ полеты, которые осуществлялись при более высоких температурах, с гораздо большей вероятностью не имели проблем, и это были те самые темные данные, не учтенные на графике. Ведь если вывод о том, что, чем выше температура, тем меньше вероятность возникновения проблемы, верен, то верно и обратное: чем температура ниже, тем выше вероятность возникновения этой проблемы. А согласно прогнозу температура воздуха на момент запуска была 31 ºF или около 0 ºC.

В этом же разделе доклада сделан следующий вывод: «Анализ полной истории температур при запуске указывает на то, что критическое состояние уплотнительного кольца становится *почти неизбежным*, если температура стыка меньше 65 ºF [18 ºC]» (курсив мой).

Ситуация проиллюстрирована ниже на двух диаграммах. На рис. 1, a показана диаграмма, которая обсуждалась на телеконференции. Это график зависимости количества поврежденных уплотнительных колец при каждом запуске от температуры в градусах Фаренгейта. Так, при 53 ºF – самой низкой температуре воздуха при запусках в прошлом – три уплотнительных кольца достигали критического состояния, а при 75 ºF, что было самой высокой температурой, при которой осуществлялся запуск, критического состояния достигли два уплотнительных кольца. Мы видим, что нет устойчивой связи между температурой при запуске и числом поврежденных уплотнительных колец.

Однако если мы добавим отсутствующие данные по запускам, при которых не наблюдалось критических состояний уплотнительных колец, то получим совсем иную картину, изображенную на рис. 1, b. И закономерность становится очевидной. Фактически *все* запуски, которые произошли при температуре ниже 65 ºF, приводили к критическому состоянию уплотнительных колец, и лишь 4 из 21 запуска, осуществленных при более высоких температурах, дали подобный результат. На диаграмме четко видна закономерность – чем ниже температура, тем выше риск. И что еще хуже, прогнозируемая температура была намного ниже минимальной, при которой ранее проводились запуски (*DD-тип 15: экстраполяция за пределы ваших данных*).

Отсутствующие данные имеют решающее значение для понимания происходящего.

В истории Challenger, однако, остался один загадочный момент. Хотя официальному расследованию потребовался не один месяц, чтобы сделать выводы о причинах аварии, цена акций Morton Thiokol упала на 11,86 % прямо в день катастрофы. При этом изменения цены акций компании даже на 4 % были редкостью. Котировки акций других компаний, принимавших участие в создании ракеты-носителя, также упали, но существенно меньше. Такое ощущение, что рынок знал о настоящей причине аварии. Неужели снова темные данные?

[9] https://xkcd.com/552/; отчет комиссии Роджерса см. https://forum.nasaspaceflight.com/index.php?topic=8535.0.

Сила темных данных

Этот последний пример показывает, насколько катастрофическими могут стать ситуации, когда не обращают внимания на темные данные. А они, по всей видимости, представляют реальную опасность. Однако картина все же не настолько мрачная. Оказывается, само осознание факта существования темных данных уже может дать нам преимущество. Что-то вроде принципа дзюдо для науки о данных; и в этом дзюдо есть конкретные приемы, которые я опишу в части II книги, а пока просто назову несколько из них.

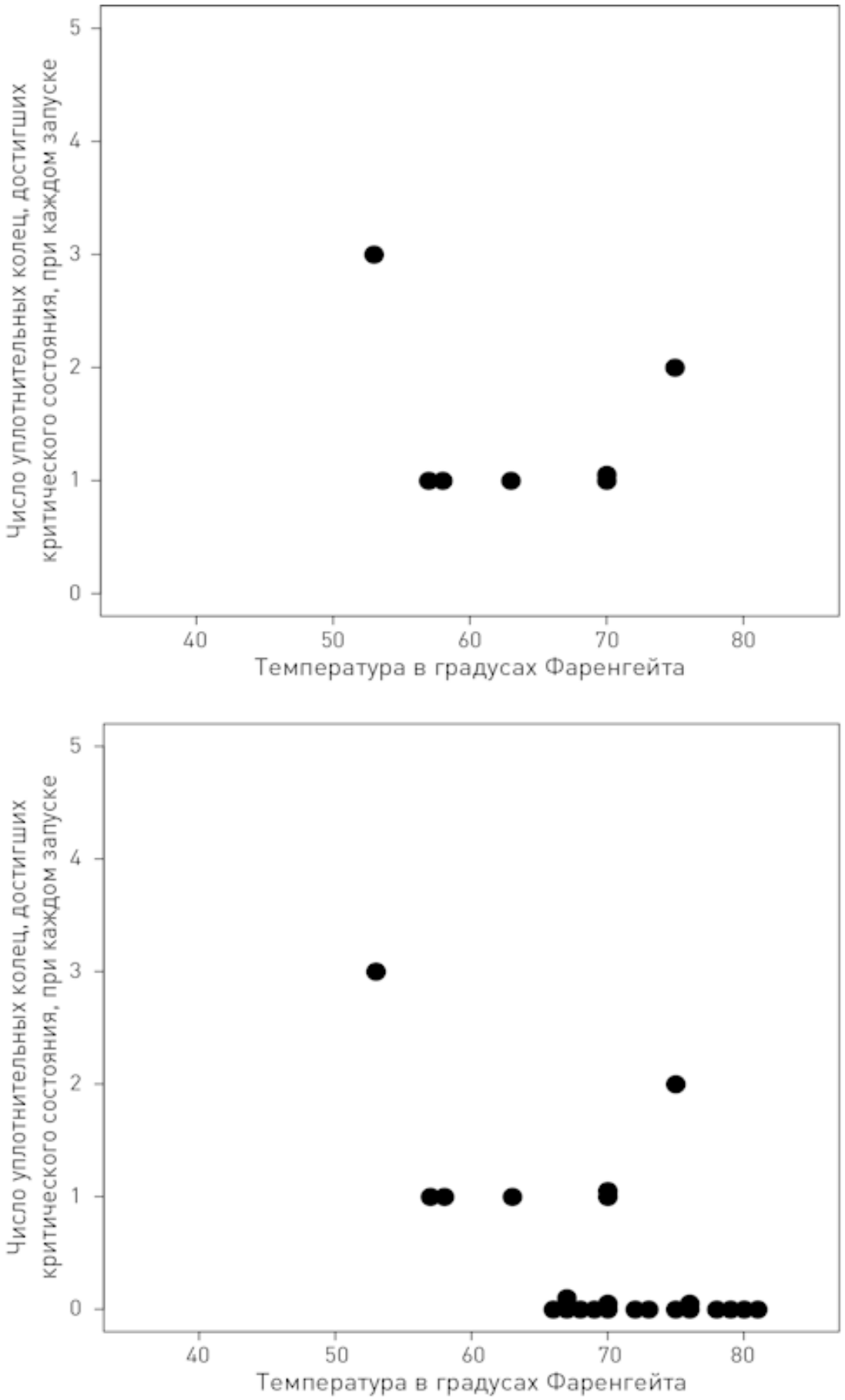

Рис. 1. a — данные, изученные в ходе предстартовой телеконференции челнока Challenger; b — полные данные

В главе 2 пойдет речь о так называемых рандомизированных контролируемых исследованиях. В главе 9 мы вновь вернемся к ним, но рассмотрим с иного ракурса. Для примера возьмем медицинские исследования, когда сравнивают два метода лечения и при этом назначают их двум группам пациентов. Однако просто разделить людей на группы недостаточно. Если известно, кому какое лечение назначено, это может повлиять на результаты – исследователи могут относиться к одной из групп более внимательно, чем к другой. Например, когда сравнивают новый непроверенный метод лечения со стандартным, исследователи, порой даже не осознавая этого, склонны тщательнее отслеживать побочные эффекты и проводить измерения в первой группе. Чтобы преодолеть эту потенциальную необъективность, в подобных исследованиях распределение методов лечения скрывают от исследователей (*DD-тип 13: намеренно затемненные данные*). В таких случаях говорят о *слепом* исследовании, чтобы указать на темные данные.

Другой хорошо известный метод, использующий темные данные, – выборочные опросы. Возможно, мы захотим узнать мнение горожан или покупателей конкретной продукции, но выяснять мнение всех без исключения слишком затратно. К тому же это занимает много времени, и мнения могут измениться. Альтернативой тотальному опросу является опрос отдельных представителей группы. Мнения тех, кто не попадает в наш опрос, и будут темными данными. Вроде бы такая стратегия выглядит рискованно – она явно напоминает историю с базой данных TARN. Но оказывается, что, используя продуманные методы отбора людей для опроса, мы можем получить точные и достоверные ответы, при этом быстрее и дешевле, чем если бы обращались к каждому.

Третий способ заставить темные данные работать на нас заключается в так называемом сглаживании данных. В главе 9 мы увидим, что этот метод сродни выявлению незамеченных и не поддающихся наблюдению видов темных данных (*DD-тип 14: фальшивые и синтетические данные*) и позволяет получить более точные оценки и прогнозы.

Другие способы использования темных данных, которые носят весьма экзотические названия, мы также рассмотрим в главе 9. Некоторые из них широко применяются в таких областях, как машинное обучение и искусственный интеллект.

Всюду вокруг нас

Как мы видим, темные данные вездесущи. Они могут появляться повсеместно и где угодно, а их наиболее опасное свойство заключается в том, что мы по определению не можем быть уверенными в их *отсутствии*. Это означает, что необходимо постоянно быть начеку и задавать себе вопрос: «*Что мы упускаем?*»

Не потому ли многие мошенничества остаются незамеченными, что полиция ловит лишь неумелых преступников, а настоящие «мастера» продолжают «творить»? Берни Мэдофф основал свою фирму Bernard L. Madoff Investment Securities LLC в 1960 г., а арестован был лишь в 2008 г. Когда его приговорили к 150 годам тюремного заключения, ему исполнился уже 71 год – можно сказать, что ему практически все сошло с рук.

А множество потенциально излечимых больных, которых мы вовремя не диагностируем? Разве это не происходит лишь потому, что болезни на ранней стадии имеют гораздо меньше симптомов, чем в своей тяжелой форме?

Опасны ли социальные сети? Ведь они отражают только то, что мы уже знаем и чему верим, не посягая на нашу точку зрения, поскольку отбирают факты и события в пределах нашей зоны комфорта. Или, что еще хуже, те рассказы, которые люди выбирают для публикаций в социальных сетях, могут создавать у нас ложное представление о том, что жизнь всех

остальных людей удивительно легка и прекрасна, а это прямой путь к депрессии – ведь в своей жизни мы встречаем так много препятствий.

Мы привыкли думать о данных как о числах. Но данные необязательно должны быть числами, включая и темные данные. Вот вам пример, в котором отсутствующей критической информацией является одна буква.

Арктическим экспедициям 1852, 1857 и 1875 гг. поставлялось Arctic Ale – пиво с особо низкой температурой замерзания, изготовленное Сэмюэлем Аллсоппом. Альфред Барнард, написавший историю британского пивоварения, попробовал этот эль в 1889 г., описав его как напиток «приятного коричневого оттенка, обладающий вкусом вина и орехов и таким шипением, словно был сварен только что… Из-за большого количества оставшегося неферментированного экстракта, его следует рассматривать как чрезвычайно ценный и питательный продукт»[10]. Как раз то, что нужно в арктических экспедициях.

В 2007 г. бутылка из партии 1852 г. была выставлена на аукционе eBay со стартовой ценой $299. Продавец, у которого она хранилась в течение 50 лет, неправильно написал название пива, пропустив одну «p» в слове «Allsopp». Как следствие, предмет не обнаруживался поисковыми запросами любителей винтажного пива, так что поступило только две заявки. Из них победила заявка 25-летнего Даниэля Вудула, который предложил целых $304. Стремясь определить ценность покупки, Вудул тут же вновь выставил бутылку на продажу, но на этот раз с правильным названием. В ответ было подано 157 заявок с максимально предложенной ценой $503 300.

В этом случае одна пропущенная буква стоила полмиллиона долларов[11]. Это наглядный пример того, что потеря информации может привести к значительным последствиям. Как мы увидим далее, полмиллиона долларов – ничто по сравнению с убытками в других ситуациях, связанных с отсутствием данных. Они способны разрушать судьбы, уничтожать компании и, как в случае с Challenger, приводить к гибели людей. Короче говоря, отсутствующие данные важны.

В случае с Arctic Ale чуть большее внимание помогло бы избежать проблемы. Небрежность, безусловно, одна из самых распространенных причин появления темных данных, но далеко не единственная. Неприятный факт заключается в том, что данные могут стать темными по очень широкому ряду причин, и далее в книге мы увидим это.

Заманчиво считать темные данные исключительно тем, что можно было бы получить, но по каким-то причинам не удалось. Безусловно, это самый очевидный вид темных данных. Отсутствующие данные по заработной плате в опросе, в котором часть респондентов отказалась разглашать эту информацию, конечно, являются темными данными, но также ими является и уровень заработной платы безработных, которые не получают ее и, следовательно, просто не могут назвать. Ошибки измерения и неточности скрывают истинные значения; обобщая данные (например, вычисляя средние значения), мы теряем детали; неверные формулировки запросов искажают смысл того, что мы хотим узнать. В более общем понимании любую неизвестную характеристику некоей генеральной совокупности (статистики часто используют термин «*параметр*») можно рассматривать как темные данные.

Поскольку число возможных причин возникновения темных данных, по сути, не ограничено, знание того, на *что* следует обращать внимание, является чрезвычайно важным для предотвращения ошибок и просчетов. Именно с этой целью в нашей книге и представлено описание *DD-типов*. Они не охватывают все возможные причины (например, небрежность,

---

[10] R. Pattinson, Arctic Ale: History by the Glass, issue 66 (July 2102), https://www.beeradvocate.com/articles/6920/arctic-ale/, accessed 31 July 2018.

[11] В действительности оказалось, что победившая заявка была шуткой и участник торгов не собирался платить. Но даже при этом Вудул мог рассчитывать на приличную прибыль: частный коллекционер из Шотландии недавно продал с аукциона бутылку из экспедиции 1875 г. за £3300, что равняется примерно $4300.

допускающую включение в окончательный результат исследования данных пациентов, которые наблюдались недостаточно длительное время), но обеспечивают более общую систематику (например, проводят различие между данными, о которых мы знаем, что они отсутствуют, и данными, о которых мы этого не знаем). Понимание этих *DD-типов* может помочь вам защититься от ошибок, оплошностей и угроз, вытекающих из самого факта незнания. В этой книге представлены, а в главе 10 обобщены следующие *DD-типы*:

- *DD-тип 1: данные, о которых мы знаем, что они отсутствуют;*
- *DD-тип 2: данные, о которых мы не знаем, что они отсутствуют;*
- *DD-тип 3: выборочные факты;*
- *DD-тип 4: самоотбор;*
- *DD-тип 5: неизвестный определяющий фактор;*
- *DD-тип 6: данные, которые могли бы существовать;*
- *DD-тип 7: данные, меняющиеся со временем;*
- *DD-тип 8: неверно определяемые данные;*
- *DD-тип 9: обобщение данных;*
- *DD-тип 10: ошибки измерения и неопределенность;*
- *DD-тип 11: искажения обратной связи и уловки;*
- *DD-тип 12: информационная асимметрия;*
- *DD-тип 13: намеренно затемненные данные;*
- *DD-тип 14: фальшивые и синтетические данные;*
- *DD-тип 15: экстраполяция за пределы ваших данных.*

Глава 2
Обнаружение темных данных
Что мы собираем, а что нет

Темные данные со всех сторон

Данные не возникают сами собой. Они не существуют с начала времен, ожидая, пока их проанализируют. Кто-то должен собрать их. И разные методы сбора данных, как вы догадываетесь, порождают разные типы темных данных.

В этой главе мы рассмотрим три основных метода создания наборов данных, а также пути возникновения темных данных, связанные с каждым из них. Следующая глава посвящена дополнительным осложнениям, которые темные данные могут вызывать в разных ситуациях.

Итак, вот три основные стратегии создания наборов данных.

● Сбор данных обо *всех* интересующих нас объектах.

Именно к этому стремятся, например, во время переписи населения. Точно так же инвентаризации преследуют цель максимально детализировать все позиции на складе или в любом другом месте. В 2018 г. ежегодная инвентаризация в лондонском зоопарке, которая занимает около недели, показала, что в данной организации насчитывается 19 289 животных – от филиппинских крокодилов до беличьих обезьян, пингвинов Гумбольдта и двугорбых верблюдов (в случае муравьев, пчел и других социальных насекомых подсчитывались колонии). В главе 1 мы уже отмечали, что супермаркеты собирают данные обо *всех* покупках. То же самое касается налогов, операций по кредитным картам и персонала. Не менее подробно регистрируются спортивная статистика, книги на полках библиотек, цены в магазинах и многое другое. Во всех этих примерах каждая единица – будь то объект или человек – детализируется для формирования набора данных.

● Сбор данных о *некоторых* элементах совокупности.

Альтернативой полной переписи населения является сбор данных в рамках ограниченной выборки. Репрезентативная выборка крайне важна в нашем контексте, и мы подробно рассмотрим ее взаимосвязь с проблемой темных данных. Проще говоря, порой приходится собирать только те данные, которые легче собрать. Чтобы понять, как ведут себя покупатели в принципе, вы можете понаблюдать за теми, кто пришел в магазин сегодня. Для того чтобы узнать, сколько времени у вас отнимает дорога до работы, вы можете просто ежедневно на протяжении месяца следить за продолжительностью поездки. Бывают ситуации, когда просто не нужно измерять все: чтобы увидеть динамику изменения цен на продукты питания, вам не нужна информация о каждой покупке, а для определения среднего веса песчинки ни к чему взвешивать каждую из них. В главе 1 мы уже видели, что само понятие «измерение всего» может быть лишено смысла. Полнота данных, например о вашем росте, будет ограничена только теми измерениями, которые вы проведете.

Несколько лет назад, еще до начала эры легкодоступных больших наборов данных, мы с коллегами опубликовали «Справочник по небольшим наборам данных»[12], включающий в себя 510 массивов реальных данных, на примере которых преподаватели могут иллюстрировать концепции и методы статистики. В справочнике приведены результаты 20 000 бросков игральной кости, данные о сроках беременности, толщине роговицы глаза, длительности нерв-

[12] D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski, A Handbook of Small Data Sets (London: Chapman and Hall, 1994).

ных импульсов и множество других наборов данных, очень немногие из которых описывают генеральные совокупности целиком.

● Изменение условий.

Первые две стратегии помогают собрать так называемые данные наблюдения. Вы просто измеряете значения, которые присущи объектам или людям, никак не меняя условия, в которых проводятся измерения. Вы не даете людям лекарств, чтобы отследить их реакцию, не просите выполнить какое-либо задание, чтобы подсчитать, сколько времени это займет, не меняете удобрения, чтобы посмотреть, какие из них дают самый обильный урожай, не пробуете разную температуру воды, чтобы понять, как она влияет на вкус чая. Если же вы *меняете* условия сбора данных, иначе говоря, *вмешиваетесь,* то такие данные называются экспериментальными. Экспериментальные данные особенно важны, потому что они могут дать информацию о контрфактуальности (*DD-тип 6: данные, которые могли бы существовать*), упомянутой в главе 1.

Хотя у всех трех методов сбора данных есть немало общих недостатков, связанных с темными данными, для каждого из них характерны и свои особые проблемы. Мы начнем с рассмотрения первой стратегии сбора данных, претендующей на полный охват.

Извлечение, отбор и самоотбор данных

Компьютеры оказали революционное влияние на все аспекты нашей жизни. Где-то это влияние проявляется очевидным образом, например в программном обеспечении, которое я использую для подготовки рукописи этой книги, или в системе бронирования авиабилетов, а где-то оно не так заметно, если речь идет, скажем, о встроенных компьютерах, управляющих тормозами и двигателем автомобиля, или о начинке какого-нибудь копировального аппарата.

Но независимо от того, очевидна или нет роль компьютеров, во всех случаях в машины поступают данные – измерения, сигналы, команды – и обрабатываются ими, чтобы принять решение или выполнить какую-либо операцию. Казалось бы, по завершении операции можно попрощаться с данными, однако зачастую этого не происходит. Данные все чаще сохраняют, отправляют в базы данных и там аккумулируют. То же самое происходит и с побочными или, как их еще называют, выхлопными данными (по аналогии с выхлопными газами), которые в дальнейшем помогают добиться лучшего понимания, усовершенствовать системы или восстановить картину событий, если что-то пошло не так. Черный ящик в самолете является классическим примером такого рода систем.

Выхлопные данные, описывающие людей, называются *административными*[13]. Особая сила административных данных заключается в том, что они сообщают не то, *что люди говорят о своих действиях* (как, например, в случае опросов), а то, *что они делают на самом деле*. Такие данные показывают, что люди купили, где они это купили, что они ели, какие поисковые запросы делали и т. д. Считается, что административные данные намного точнее демонстрируют реалии общества, чем ответы людей на вопросы об их действиях и поведении. Это привело к накоплению правительствами, корпорациями и рядом других организаций гигантских баз данных, описывающих наше поведение. Нет сомнения в том, что эти базы данных представляют собой очень ценный ресурс, настоящую золотую жилу в сфере знаний о человеческом поведении. Сделанные на их основе выводы помогут усовершенствовать процесс принятия решений, повысить корпоративную эффективность и лучше продумать государственную политику – конечно, при условии, что эти выводы будут точными и не подвергнутся влиянию темных данных. Кроме того, когда данные, которые мы хотели бы сохранить в темноте, ста-

[13] D. J. Hand, “Statistical challenges of administrative and transaction data (with discussion),” *Journal of the Royal Statistical Society*, Series A181 (2018): 555-605.

новятся известны другим, возникают риски нарушения конфиденциальности. Мы вернемся к этому вопросу чуть дальше, а пока давайте поищем темные данные, причем в самых неожиданных местах.

Один из очевидных и очень серьезных недостатков административных данных кроется в самом их преимуществе: они сообщают о том, что на самом деле делают люди, а это может быть полезным только тогда, когда вы не пытаетесь исследовать, что люди думают и чувствуют. Например, своевременное обнаружение недовольства сотрудников тем, как идут дела, может быть не менее важным для корпорации, как и наблюдение за их поведением в жестких рамках повседневной работы, когда начальник буквально стоит за спиной. Но, чтобы узнать, что чувствуют люди, нам придется активно допытываться этого, например с помощью опроса. Для решения разных задач требуются и разные стратегии сбора данных, при этом каждая из них грозит своими особыми проблемами, связанными с темными данными.

Мое первое настоящее знакомство с темными данными состоялось в сфере банковских услуг для потребительского сектора: кредитные и дебетовые карты, персональные займы, автокредиты, ипотека и прочие подобные вещи. Данные о транзакциях по кредитным картам представляют собой гигантские наборы данных, поскольку миллионы клиентов ежегодно совершают миллиарды операций. Так, с июня 2014 г. по июнь 2015 г. было совершено около 35 млрд транзакций по картам Visa[14]. Каждый раз, когда покупка оплачивается кредитной картой, регистрируется потраченная сумма, валюта, продавец, дата и время транзакции, а также многие другие детали, общий список которых включает 70–80 пунктов. Большую часть этой информации составляют данные, необходимые для совершения транзакции и списывания суммы с соответствующего счета – это обязательная часть операции, поэтому пропуск таких деталей маловероятен или даже невозможен. Например, операция не может быть выполнена без информации о том, сколько взимать или с кого взимать. Но есть и такие данные, которые не критичны для проведения операции, поэтому существует вероятность того, что они не будут собраны. В частности, номер партии товара, его идентификационный код или цена за единицу не являются обязательной информацией для проведения транзакции. Очевидно, что это *DD-тип 1: данные, о которых мы знаем, что они отсутствуют*.

Что еще хуже, во всяком случае в отношении темных данных, клиенты рассчитываются за покупки не только кредитными картами, но и наличными. Это означает, что реестр *всех* покупок и транзакций, созданный на основе данных по кредитным картам, будет содержать невидимые массивы темных данных – *DD-тип 4: самоотбор*. Вдобавок существует несколько операторов кредитных карт. Данные одного оператора не могут считаться репрезентативными для всей совокупности держателей кредитных карт и уж тем более для населения в целом. Таким образом, несмотря на многообещающие перспективы, административные данные имеют скрытые недостатки, связанные с темными данными.

Конкретной проблемой, с которой мне пришлось столкнуться, был заказ на создание «системы показателей» – статистической модели для прогнозирования вероятности неплатежей, которая могла бы использоваться при принятии решений о предоставлении кредитов. Мне был открыт доступ к большому набору данных, содержащему информацию из заявок предыдущих клиентов, а также их кредитные истории, показывающие действительную картину того, платили они или нет по своим обязательствам.

По сути ничего сложного в этом заказе не было. Я должен был выяснить, какие сочетания характеристик отличают клиентов, выполнивших свои обязательства, от тех, кто допустил дефолт. Это позволило бы классифицировать будущих заявителей как «добросовестные заемщики» или «потенциальные неплательщики».

[14] https://www.quora.com/How-many-credit-and-debit-card-transactions-are-there-every-year, accessed 24 August 2018.

Проблема заключалась в том, что банк хотел получить модель, позволяющую делать прогнозы в отношении *всех* будущих заявителей. Предоставленные мне данные, безусловно, не были генеральной совокупностью, отражавшей всех заявителей – они касались лишь тех, кто уже прошел процесс отбора. Надо полагать, состоявшиеся клиенты получили кредиты, потому что им был присвоен статус приемлемого риска в соответствии с каким-то более ранним механизмом отбора – на основе либо предыдущей статистической модели, либо субъективной оценки менеджеров банка. Те, кого сочли слишком рискованными, не получили ссуду, поэтому я не мог знать о том, насколько добросовестно они выполнили бы свои обязательства. Я даже не имел понятия, сколько заявителей было отклонено ранее и не попало в мой набор данных. Короче говоря, данные, предоставленные мне, были искаженной выборкой с неизвестными критериями отбора (или смещением выборки), и любая статистическая модель, построенная на этом наборе данных, вводила бы в заблуждение в случае применения ко всем потенциальным кандидатам.

Конец ознакомительного фрагмента.