Содержание

ЧАСТЬ ТРЕТЬЯ
«Как я могу это показать?»
Представление данных

Бонус для читателей

Спасибо, что выбрали эту книгу. Вы сделали огромный шаг на пути в науку о данных.

Получите бесплатный доступ к моему курсу A-Z Data Science. Просто зайдите на сайт www.superdatascience.com/bookbonus и используйте пароль datarockstar.

Удачи в анализе данных!

Введение

«Наверное, вы всегда хотели стать аналитиком данных — с самого детства?»

Мне приятно, что меня об этом спрашивают. Да, я люблю свою работу. Я с большим удовольствием обучаю студентов основам науки о данных. И здорово, что люди, похоже, думают, что энтузиазм по отношению к данному предмету возник во мне еще в молодом возрасте. Но это абсолютно не соответствует действительности. Скажем честно, ни один ребенок не мечтает о том, чтобы стать ученым — аналитиком данных. Дети хотят быть космонавтами. Танцорами. Врачами. Пожарными. И если вы грезите о спасении жизней или о полетах в космическом пространстве, вы вряд ли остановите свой выбор на столь приземленном занятии.

Когда люди спрашивают меня, всегда ли я хотел построить карьеру в области науки о данных, я возвращаюсь к своему детству и вижу маленького русского мальчика, выросшего в Зимбабве. Запах тлеющих углей, брачные вопли африканских красных жаб, незабываемый уют зимнего вечера, кончики пальцев, переворачивающие страницу за страницей сборника историй для детей, — это фрагменты воспоминаний о множестве прекрасных вечеров, когда я слушал русские сказки, которые читала мне мама.

Моя мать хотела, чтобы я, мои братья и сестры любили Зимбабве, но она также заботилась о том, чтобы мы знали свои культурные корни. Она подумала, как наилучшим образом передать нам эту информацию, и решила, что самый действенный способ — сказки. Когда я в конце концов вернулся в Москву — в город, который едва помнил, — то почувствовал, что возвращаюсь домой, благодаря крупицам информации о России, вплетенным в затейливые сюжеты.

Такова сила повествования. И все множество услышанных сказок я хотел разбить на составляющие их компоненты. Мне нужно было

увидеть большую картину, но я хотел видеть ее сквозь призму маленьких деталей. Я был очарован каждой частью механизма, создающего что-то настолько прекрасное. Я интуитивно знал: для того чтобы самому рассказать хорошую историю, сначала нужно собрать эти маленькие единицы информации. Именно так сформировалось мое отношение к данным.

В сегодняшнюю цифровую эпоху данные используются для создания историй о том, кто мы такие, как мы себя представляем, что нам нравится и когда мы хотим чего-то. Для того, чтобы проложить тропинку с уникальными виртуальными следами. Машины теперь знают о нас больше, чем мы сами, благодаря всем доступным им данным. Они читают наши личные данные как сборник рассказов о нас. И в науке о данных замечательно то, что любая дисциплина сегодня записывает свои данные, а это значит, что, освоив профессию аналитика данных, мы также можем стать космонавтами, танцорами и врачами, о чем так сильно мечтали.

Мало кто знает, что работать с данными в конечном итоге означает быть рассказчиком, передающим информацию. Так же, как и структурные компоненты историй, проекты по анализу и обработке данных тоже организованы логически. В книге «Работа с данными в любой сфере» четко выделяются пять этапов, которые составляют то, что я называю процессом обработки и анализа данных. Это не единственный подход, который можно использовать, но он обеспечит нашему проекту связь с практикой и продвижение к логическому завершению. И он четко и ясно структурирован, что мне так нравилось в детстве.

И вот я решил рассказать историю данных…

Но я абсолютный новичок

Наука о данных фактически является одной из тех областей, которые извлекают выгоду из опыта других сфер. Я надеюсь, что многие мои читатели уже весьма преуспели в той или иной профессии. Хорошо. Вы *ничего* не потеряете, если обратитесь к науке о данных, работая в другой области. Отнюдь не вредно для начала разбираться в чем-то

еще. Это своего рода фундамент, который вам пригодится, чтобы стать хорошим аналитиком данных.

Начав работать в транснациональной консалтинговой компании Deloitte, я не знал ни одного из алгоритмов, которые мы рассмотрим в этой книге. Да никто от меня этого и не ожидал. Совсем немногие начали свою карьеру с науки о данных. Прочитав книгу, вы обнаружите, что те, кто добился успеха в этой сфере, даже не думали о ней, пока находились в начале своей карьеры. Итак, отбросьте страх перед цифровой неграмотностью — взяв эту книгу, вы сделали первый шаг на пути в мир науки о данных.

Эй, а где код?

Если вы, как и я, пролистываете книгу, прежде чем приступить к чтению, то, возможно, заметили, что вам не встретилось ни одной строки кода. Я слышу, как вы говорите: «Но это ведь книга о науке о данных, так что же происходит?» Наука о данных — чрезвычайно широкий предмет. «Работа с данными в любой сфере» погружает вас в тему и вдохновляет на размышления о том, как эта дисциплина может быть включена в вашу текущую или будущую деловую практику. Вы узнаете *методы* науки о данных — потому что ее «ингредиенты» (код) легко доступны онлайн. Если воспользоваться аналогией с приготовлением пищи, перед вами в меньшей степени просто книга рецептов и в большей — подробная информация об основных методах, используемых в науке о данных. Изучите их тщательно, и вы начнете интуитивно понимать, *почему* вам нужно применять определенные коды и методы, — гораздо более эффективный подход к обучению, чем просто предоставление строк кода для подключения к вашему проекту.

Как пользоваться этой книгой

Я написал эту книгу специально для того, чтобы вы могли обратиться к ней, где бы вы ни находились — в поезде, в ванне, в ожидании человека своей мечты. Читайте ее по частям или в один присест, по главам,

выбирая самое лучшее, выделяя нужное желтым маркером, наклейками. В начале каждой части вы найдете краткое введение, помогающее быстро определить, какая глава окажется для вас наиболее интересной. Часть первая более объемна, она дает общее представление о науке о данных. Вторая и третья части сосредоточены на процессах анализа и обработки данных, интуиции, стоящей за некоторыми из самых мощных на сегодняшний день аналитических моделей, и на том, как повысить ваши шансы на успех, совершая первые шаги в направлении цели.

Если вы новичок, то получите максимальную отдачу от книги, прочитав ее от корки до корки. Если вы знакомы с наукой о данных как с дисциплиной и хотите добраться до сути того, как применять ее методы, не стесняйтесь обратиться к главе, которая вам больше всего поможет.

ЧАСТЬ ПЕРВАЯ
«Что это?» Ключевые принципы

Учитывая очевидно безграничный потенциал технических и прикладных наук и связанные с ними широкие возможности для умелых предпринимателей, некоторые могут спросить, почему они вообще должны заниматься наукой о данных — почему бы просто не изучить технологические принципы? В конце концов, технологии управляют миром и не выказывают никаких признаков сдачи позиций. Любой читатель, заботящийся о своей карьере, может подумать, что научиться разрабатывать новые технологии, несомненно, будет наилучшим способом двигаться вперед.

Легко расценивать технологии как фактор, который меняет мир, — они дали нам персональный компьютер, интернет, искусственные органы, беспилотные автомобили, глобальную систему позиционирования (GPS), — но мало кто думает о науке о данных как о движущей силе многих из этих изобретений. Вот почему вам стоит прочитать именно *эту* книгу, а не книгу о технологиях: вам нужно понять, как работает система, чтобы внести в нее изменения.

Мы не должны рассматривать данные только как скучных, но готовых помочь родителей, а технологии — как стильных подростков. Важность науки о данных не начинается и не заканчивается объяснением того, что технологии нуждаются в данных как одном из многих других функциональных элементов. Это было бы отрицанием прелести данных и множества интересных приложений, которые они

предлагают для работы и игры. Короче говоря, невозможно иметь одно без другого. Это означает, что, если у вас есть основа для науки о данных, перед вами будет открыта дверь к широкому кругу других областей, в которых нужен аналитик данных. Это делает науку о данных необычной и благоприятной областью исследований и практики.

В первой части приводится информация о вездесущности данных, а также о развитии и ключевых принципах науки о данных. Эти сведения полезны для начального погружения в предмет. Вы получите четкое представление о том, какое отношение данные имеют к вам, и задумаетесь не только о том, как данные могут непосредственно принести пользу вам и вашей компании, но и как вы можете в течение длительного времени использовать их в профессиональной и прочих сферах.

Начало пути

Глава 1 станет началом нашего путешествия в науку о данных. Сначала в ней будет продемонстрировано, насколько велики масштабы распространения данных и то, каким образом мы все вносим вклад в их производство в наш компьютерный век. Затем я расскажу, как люди собирают данные, работают с ними и, что очень важно, как данные можно использовать для поддержки большого количества проектов и методов внутри и вне самой дисциплины.

Мы установили, что проблемы с наукой о данных частично связаны не с ее относительной сложностью, а скорее с тем, что эта область знаний для многих по-прежнему покрыта туманом. Только когда мы точно понимаем, сколько данных имеется и как они собраны, мы можем начать рассматривать различные способы работы с ними. Мы достигли той точки в нашем технологическом развитии, когда информацию можно эффективно собирать и хранить на благо всех отраслей промышленности и научных дисциплин, о чем свидетельствует количество общедоступных баз данных и финансируемых правительством проектов по агрегированию данных культурными и политическими институтами. Вместе с тем сравнительно немногие знают, как получить доступ к данным и как их проанализировать. Если же люди

не осознают пользу данных для своей профессиональной деятельности, все красивые массивы данных только собирают пыль. В этой главе объясняется, почему наука о данных крайне важна *именно сейчас*, почему это не просто тенденция, которая скоро выйдет из моды, и почему вы должны рассмотреть возможность внедрения ее практик в качестве ключевого компонента решения ваших рабочих задач.

Наконец, в этой главе описывается, как стремительная траектория развития технологий не позволяет нам даже на время отвернуться от науки о данных. Каковы бы ни были представления о мире, к которому мы стремимся, невозможно остановить сбор данных, их обработку и использование. Тем не менее нельзя игнорировать тот факт, что сами по себе данные не касаются вопросов морали, и это обусловливает возможность их нечестного или неправильного использования. Те из вас, кто обеспокоен такого рода злоупотреблениями, могут принять участие в противостоянии им и вступить в дискуссию с глобальными институтами, которые занимаются проблемами, связанными с этикой данных — аспектом, который я нахожу настолько существенным, что отвел ему специальный подраздел в главе 3.

Будущее принадлежит данным

Все — каждый процесс, каждый датчик — скоро будет управляться данными. Это резко изменит способ ведения бизнеса. Я предсказываю, что через десять лет от каждого сотрудника любой организации в мире будет требоваться обладание определенным уровнем грамотности в сфере данных и умение работать с ними, получая на их основе некоторые идеи для повышения ценности бизнеса. Не такая уж дикая мысль, если учесть, что на момент публикации этой книги предполагается, что многие люди знают, как пользоваться цифровым кошельком Apple Pay, выведенным на рынок только в 2014 г.

Глава 2 — «Как данные удовлетворяют наши потребности» — наглядно демонстрирует, что данные являются эндемичными для каждого аспекта нашей жизни. Они управляют нами, накапливая силу в цифрах. Данные всегда играли важную роль в нашем существовании. Наша ДНК несет в себе основные данные о нас, и эти базовые формы

данных руководят нами: отвечают за то, как мы выглядим, за форму наших конечностей, за структуру нашего мозга и его способность обрабатывать информацию, а также за диапазон эмоций, которые мы испытываем. Мы — хранилища этих данных, шагающие флеш-накопители биохимической информации; вместе с данными нашего партнера мы передаем их нашим детям и «кодируем». Не интересоваться данными означает не интересоваться самыми фундаментальными принципами жизни.

В этой главе объясняется, как данные используются во многих областях, и для иллюстрации я использую примеры, которые непосредственно перекликаются с пирамидой потребностей Абрахама Маслоу, теорией, хорошо знакомой многим ученым и практикам в области бизнеса и управления. Если эта иерархия является для вас новинкой, не беспокойтесь — я объясню ее суть и то, как она применима к нам, в главе 2.

Приостановка развития

Последняя глава первой части покажет, как новички в науке о данных могут изменить свое мышление, чтобы погрузиться в нее, и поможет выявить те области, где уже сейчас возможно применить анализ данных. Многие достижения науки о данных основательно затронули другие сферы и поставили вопросы о будущем перед самыми разными специалистами и учеными. Если вы хотите развивать свою карьеру как аналитик данных, эта глава подскажет некоторые идеи для сфер, в которых вы, возможно, уже работаете.

В главе 3 я также представлю некоторые наиболее важные подходы, которые вы можете использовать, чтобы начать работу как практик. Наука о данных намного проще, чем многие другие научные дисциплины. Вам не нужно быть прирожденным ученым, чтобы овладеть принципами науки о данных. Что вам *действительно* необходимо — это умение придумывать различные способы извлекать пользу из данных тогда, когда дело касается бизнес-операций или личной мотивации. Ведь ученые — исследователи данных изучают *возможности* предоставленной информации. Вы можете удивиться, узнав, что у вас

уже есть некоторые навыки и опыт, которые вы можете использовать на своем пути к освоению этой дисциплины.

Разумеется, новичкам необходима разумная осторожность. Любой, кто использовал Excel, работал в офисной среде или изучал в университете предмет, имеющий научную составляющую, вероятно, уже встречался с данными. Но некоторые из методов использования данных, которые вы, возможно, усвоили, будут неэффективными, и приверженность тому, что вы уже знаете, может помешать вам изучить наиболее действенные способы использования массивов данных: мы обсудим это подробно во второй и третьей частях.

Несмотря на явный положительный эффект использования данных, важно не обольщаться. Поэтому в главе 3 рассматриваются и различные угрозы безопасности, которые данные могут представлять для своих пользователей, и то, как работают аналитики данных для решения текущих и потенциальных проблем. Этика данных является особенно привлекательной и заслуживающей внимания областью, поскольку она способна изменять и направлять будущие разработки в области науки о данных. Учитывая то, что мы знаем о сборе информации, этика данных — в той мере, в какой ее можно использовать в машинах и онлайн, — создает основу для общения людей и технологий. Когда вы прочитаете эту главу, подумайте о том, как каждая из областей может быть связана с тем, как вы работаете, и насколько полезны для вашего бизнеса дальнейшие инвестиции в эту сферу.

Определение данных

01

Подумайте о последнем фильме, который вы видели в кинотеатре. Как вы впервые узнали о нем? Возможно, вы кликнули на трейлер, когда YouTube рекомендовал его вам, или же ролик появился в качестве рекламы, прежде чем YouTube показал вам видео, которое вы действительно хотели посмотреть. Может быть, вы прочитали в социальной сети, что ваш друг хвалит картину, или в вашей новостной ленте появился увлекательный клип из фильма. Если вы любитель кино, сайт-агрегатор мог подобрать его для вас как фильм, который вам может понравиться. Вы, не исключено, нашли анонс фильма за пределами интернета — в своем любимом журнале либо же могли обратить внимание на афишу по дороге в кофейню, где лучше работает Wi-Fi.

Ни один из этих источников информации не был случайным. Звезды не просто сошлись для вас и фильма в нужный момент. Оставим идеалистические совпадения неожиданным экранным встречам. То, что привело вас в кино, было в меньшей степени желанием увидеть фильм и в гораздо большей — мощной смесью основанных на данных признаков, которые выделили вас в качестве вероятного зрителя, *прежде чем вы сами поняли, что хотите посмотреть фильм*.

Когда вы взаимодействовали с каждым из этих источников информации, вы оставили немного сведений о себе. Мы называем их выхлопными данными. Этот процесс не ограничивается вашим присутствием в онлайне и важен не только для создания социальных сетей. Независимо от того, используете ли вы социальные медиаплатформы, *нравится* вам это или нет, вы делитесь своими данными.

Так было всегда — мы просто научились лучше записывать и собирать их. Любое количество ваших ежедневных взаимодействий может способствовать этому «выхлопу». По дороге в лондонское метро вас запечатлевают камеры видеонаблюдения. Сев на поезд, вы добавляете

информацию в базу «Транспорт» статистических данных Лондона об использовании метро в час пик. Когда вы делаете закладки или выделяете страницы романа на своем устройстве для чтения Kindle, вы помогаете дистрибьюторам понять, что особенно понравилось читателю, и что они могли бы разместить в будущих маркетинговых материалах, и как глубоко читатели склонны погрузиться в роман, прежде чем остановиться.

Если вы наконец решите отказаться от испытаний в общественном транспорте и вместо этого поедете в супермаркет на автомобиле, выбранная вами скорость поможет GPS-сервисам показывать своим пользователям в режиме реального времени, насколько напряженный трафик в районе, и также позволит вашему автомобилю оценить, сколько еще времени остается, прежде чем вам стоит искать автозаправочную станцию.

И сегодня, когда вы выходите из этих точек соприкосновения, оставленные вами данные уже собраны и добавлены в «проект» о вас, который детализирует ваши интересы, действия и желания.

Но это только начало истории данных. Я расскажу вам о том, насколько действительно распространены данные. Вы узнаете основные понятия, которые пригодятся на пути к овладению наукой о данных, а также ключевые определения, инструменты и методы — они позволят вам применить навыки работы с данными к своей собственной деятельности. Эта книга расширит ваши горизонты, показывая, как наука о данных может использоваться в разных областях такими способами, которые прежде казались вам невозможными. Я опишу, как умение работать с данными может дать толчок вашей карьере и изменить ваш бизнес — будь то посредством идей, которыми вы впечатлите топ-менеджеров, или даже благодаря запуску стартапа.

Данные повсеместны

Прежде чем двигаться дальше, нужно уточнить, что подразумевается под данными. Когда люди размышляют о данных, они думают о том, как те активно собираются, хранятся в базах данных на непостижимых

корпоративных серверах и направляются на исследования. Но это устаревший взгляд. Сегодня данные гораздо более вездесущи*.

Все весьма просто: данные — это любая единица информации. Это побочный продукт любых действий, пронизывающих каждую часть нашей жизни не только в сфере интернета, но также в истории, географии и культуре. Наскальные изображения — данные. Музыкальный аккорд — данные. Скорость автомобиля, билет на футбольный матч, ответ на вопрос анкеты — все это данные. Книга — это тоже данные, как и глава в этой книге, как слово в главе, а также буква в слове. Им не нужно *быть собранными*, чтобы считаться данными. Их не нужно хранить в архиве организации, чтобы они считались данными. Значительная часть данных в мире, вероятно, пока не объединены в какой-либо базе данных.

Предположим, что в этом определении данных как единицы информации данные являются *осязаемым прошлым*. Весьма мудро, если задуматься. Данные — это прошлое, а прошлое — это данные. Запись всего, что можно отнести к данным, называется базой данных. И аналитики данных могут использовать их для лучшего понимания наших нынешних и будущих действий. Они применяют тот же принцип, что веками использовали историки: мы можем учиться на опыте истории. Мы можем учиться на наших успехах — и на наших ошибках, чтобы улучшить настоящее и будущее.

Единственный аспект данных, который в последние годы резко изменился, — наша способность собирать, организовывать, анализировать и визуализировать их в контекстах, которые ограничены только нашим воображением. Куда бы мы ни пошли, что бы мы ни покупали, какими бы ни были наши интересы, все эти данные собираются и систематизируются в тренды, которые помогают рекламодателям и маркетологам продвигать свои продукты к тем, кто в них заинтересован;

---

* Теперь вы, вероятно, привыкли к тому, что люди используют слово «данные» как множественную форму слова «данное» и что на самом деле правильно употреблять его с глаголами во множественном, а не в единственном числе. Вы можете упомянуть, что «данное» было впервые зафиксировано в 1645 г. как используемое в единственном числе Томасом Уркхартом и что только 60 лет спустя, в 1702-м, это слово стало использоваться как существительное во множественном числе. — *Здесь и далее, за исключением особо оговоренных случаев, прим. автора.*

которые показывают политические предпочтения членов правительства в соответствии с их происхождением или возрастом и которые помогают ученым создавать искусственный интеллект (ИИ), реагирующий не только на простые запросы, но и на сложные эмоции, этику и идеологию.

С учетом всех обстоятельств вы можете спросить: «Каковы же ограничения: что мы называем данными, а что — нет? Считаются ли фактические сведения о цикле цветения растения (количественные данные) такими же данными, как фиксация ученым культурного обычая, связанного с передачей умирающему родственнику букета цветов из родной страны (качественные данные)?» Ответ — да. Данные не дискриминируются. Не имеет значения, является ли рассматриваемая единица информации количественной или качественной. Качественные данные, возможно, были менее полезными в прошлом, когда не была достаточно сложной технология их обработки, но благодаря достижениям в алгоритмах, способных обрабатывать такие данные, этот недостаток быстро уходит в прошлое.

Говоря об ограничениях понятия «данные», еще раз вспомните, что данные — это прошлое. Вы не можете получать данные из будущего, если только вам не удалось создать машину времени. Но в то время как данные нельзя получить из будущего, с их помощью *можно* получить представление о грядущем и прогнозировать его. И именно способность данных восполнить пробелы в наших знаниях делает их настолько увлекательными.

Большие данные прекрасны

Теперь, когда мы разобрались, что такое данные, нужно по-другому взглянуть на то, где и как они фактически хранятся. Мы уже продемонстрировали наш широкомасштабный потенциал создания данных (это «выхлопные данные») и пояснили, что, трактуя их как единицу информации, мы создаем очень широкую концепцию того, что понимается под данными. Итак, если они где-то рядом, где все это *происходит?*

К настоящему времени вам, вероятно, доводилось слышать термин «большие данные». Проще говоря, большие данные — это название,

присвоенное массивам данных со столбцами и строками, которых настолько много, что они не могут быть обработаны обычным аппаратным и программным обеспечением в течение разумного промежутка времени. По этой причине сам термин является динамичным — то, что расценивалось как большие данные в 2015 г., уже не будет считаться большими данными в 2020-м, поскольку к тому времени будут разработаны технологии, легко справляющиеся с подобными объемами.

Три V

Чтобы можно было считать массив данных большими данными, должно быть выполнено хотя бы одно из трех условий:

1. Объем данных — то есть размер массива данных (например, количество строк) — должен исчисляться миллиардами.
2. Скорость, то есть то, как быстро собираются данные (например, потоковое видео в интернете), предполагает, что скорость генерируемых данных слишком высока для адекватной обработки с использованием обычных методов.
3. Разнообразие. Это подразумевает либо разнородность типов информации, содержащейся в массиве данных, таком как текст, видео, аудио или файлы изображений (известные как неструктурированные данные), либо таблицы, содержащие значительное количество столбцов, которые представляют разные свойства данных.

Мы пользуемся большими данными в течение многих лет для всех видов дисциплин и гораздо дольше, чем вы могли бы ожидать, — просто до 1990-х гг. не было термина для их обозначения. Так что я вас шокирую: большие данные — это не большая новость. Это, конечно, не новая концепция. Многие, если не все, крупнейшие корпорации располагают огромными хранилищами данных об их клиентах, продуктах и услугах, которые собирались в течение длительного времени. Правительства хранят данные о людях, полученные в результате переписей и регистрации по месту проживания. Музеи хранят культурные

данные — от артефактов и сведений о коллекционере до выставочных архивов. Даже наши собственные тела хранят большие данные в виде генома (подробнее об этом в главе 3 «Мышление, необходимое для эффективного анализа данных»).

Короче говоря, если вы просто не в состоянии работать с данными, то можете назвать их большими данными. Когда ученые используют термин, они делают это не просто так. Он применяется, чтобы привлечь внимание к тому, что стандартных методов для анализа данных, о которых идет речь, недостаточно.

Почему такая суета вокруг больших данных?

Вам может показаться странным, что мы только начали понимать, насколько значимыми могут быть данные. Но когда мы в прошлом собирали данные, единственное, что мешало нам превратить их во что-то полезное, было отсутствие технологий. В конце концов, важно не то, насколько огромны данные; важно, что вы с ними делаете. Любые данные, «большие» или иные, полезны, только если из них можно извлечь информацию, и до того, как была разработана соответствующая технология, чтобы помочь нам проанализировать и масштабировать эти данные, их полезность могла быть измерена только интеллектуальными возможностями человека, пытавшегося с ними совладать. Но для сортировки больших данных требуется более быстрый и мощный процессор, чем человеческий мозг. До технологических разработок XX в. данные хранились на бумаге, в архивах, библиотеках и хранилищах. Теперь почти все новые данные, которые мы собираем, хранятся в цифровом формате (и даже старые данные активно преобразуются в цифровые, о чем свидетельствует огромное количество ресурсов, сосредоточенных в таких цифровых собраниях, как Europeana Collections и Google Books).

Хранение и обработка данных

С изобретением компьютера появилась возможность автоматизации процесса хранения и обработки данных. Но большие массивы данных

увязли в первых машинах; ученым, работавшим с электронными массивами данных в 1950-х гг., приходилось ждать решения простой задачи несколько часов. Вскоре пришли к выводу, что для *правильной* обработки больших массивов данных — для установления связей между элементами и использования этих связей с целью получения точных и значимых прогнозов — нужно создавать информационные носители, которые могли бы управлять данными и справляться с их хранением. Разумеется, по мере совершенствования технологий, основанных на вычислениях, менялись и возможности компьютеров по хранению и обработке данных. И за последние 70 лет мы не только научились эффективно хранить информацию, но и смогли сделать эту информацию переносимой. Те же самые данные, которые в 1970-х гг. помещались только на 177 778 гибких дисках, к 2000-му могли поместиться на *одном флеш-накопителе*. Сегодня вы можете хранить все это и многое другое в облаке (хранилище с виртуализированной инфраструктурой, которая позволяет просматривать ваши личные файлы из любой точки мира)*. Когда вы в следующий раз обратитесь к личным документам, хранящимся в местной библиотеке, у вас на работе или просто в вашем мобильном устройстве, имейте в виду: вы фактически делаете то, что в 1970-х гг. потребовало бы использования более 100 000 дискет.

Когда новые технологии облегчили хранение данных, исследователи начали обращать внимание на то, как эти сохраненные данные могут быть использованы на практике. Как мы начали создавать порядок из хаоса? Вернемся к нашему предыдущему примеру — фильму, который вы в последний раз смотрели в кинотеатре. Вероятно, вы были выбраны, чтобы увидеть этот фильм, не проницательным маркетологом, сосредоточенно изучавшим соответствующие критерии, а умной машиной, которая изучила ваши «выхлопные данные» и сопоставила их с найденными ею демографическими сведениями о тех, кто увидел этот фильм и получил от него удовольствие. Это может казаться новинкой, но, как мы уже установили, данные и их (ручная)

---

* Облачные данные хранятся за пределами сайта и в основном перемещаются по подводным кабелям, которые укладываются на дно океана. Так что облако находится не в воздухе, как мы могли подумать, а под водой. Карту расположения этих кабелей можно найти на www.submarinecablemap.com.

обработка уже давно существуют. Некоторые из киностудий Голливуда еще в 1950-х гг. собирали данные о том, что конкретно — от актера до режиссера и жанра — хотела увидеть их аудитория, а потом преобразовывали эту информацию в демографические характеристики респондентов, включавшие в себя возраст, местожительство и пол. Даже в то время люди принимали способные изменить ход событий решения в соответствии с информацией, извлеченной из данных.

RKO Pictures

Почему RKO Pictures, одна из голливудских студий «Большой пятерки» в 1950-х гг., продолжала снимать Кэтрин Хепберн в своих фильмах? Потому что данные показывали, что это был беспроигрышный выбор, способный привлечь внимание людей и в конечном итоге заставить их пойти в кинотеатры.

Конечно, есть место и для интуиции. На первом кастинге режиссер Джордж Кьюкор нашел актрису странной, но также признал, что «она обладала огромным чувством, которое проявлялось даже в том, как она брала стакан. Я подумал, что она очень талантлива...» (Fowles, 1992). Вот пример интуиции.

Опираясь на данные о положительном восприятии Хепберн зрительской аудиторией, RKO позже смогла воспользоваться и интуитивными предположениями Кьюкора относительно таланта актрисы и превратить их в надежные прогнозы о том, что студия сможет и дальше зарабатывать свои миллионы.

Это произошло благодаря Джорджу Гэллапу — первому человеку, который рассказал руководителям Голливуда о возможности использовать данные для принятия решений и прогнозирования, включая подбор актеров на главные роли и определение того, в какой жанр наиболее целесообразно вкладывать деньги*.

---

* Гэллап был статистиком, впервые ставшим известным публике, когда разработал метод, с помощью которого он точно предсказал переизбрание Франклина Д. Рузвельта в 1936 г.

Чтобы помочь RKO сделать это, Гэллап собрал, объединил и проанализировал качественные и количественные данные, которые охватывали демографическую информацию о зрительской аудитории RKO и ее мнение о фильмах, выпускаемых киностудией. Собирая эти данные, Гэллап создал модель, которая в первый раз сегментировала аудиторию кинозрителей демографически, выделив тех, кто благоприятно реагировал на определенные жанры, — модель, которая может и будет использоваться в дальнейшем для выборки и анализа данных.

Разрекламированный как предсказатель, помогающий студиям разбогатеть, Гэллап быстро стал любимцем многих лидеров киноиндустрии США, проверяя по данным опросов и интервью отношение аудитории к персонажам различных лент, от мультиков Уолта Диснея до фильмов Орсона Уэллса*.

Своим успехом Гэллап был обязан только данным (возможно, его можно назвать первым высокооплачиваемым аналитиком данных в мире). Его усилия в области статистики привели к тому, что этот ресурс по-прежнему имеет ценность за пределами своего первоначального замысла, обладая потенциалом охвата *неструктурированных* данных: записанных интервью представителей аудитории, отражающих культурные и социальные ценности того времени. Возможно, Гэллап подозревал, что потенциал анализа данных может только расти.

Данные могут генерировать контент

Итак, что если после всех умных свидетельств, основанных на данных, вы возненавидели фильм, который недавно видели в кинотеатре? Ну, данные, возможно, не могут предсказать все, но они, безусловно, заставили вас занять место перед экраном. Иногда данные могут получить тройку за достижения, но они всегда получают отлично за усилия. И над первым уже работают. Вместо того чтобы привязывать

* Более подробно о новаторской работе Джорджа Гэллапа см. Ohmer (2012).

нужные демографические показатели аудитории к новому фильму или телевизионному сериалу, кинокомпании теперь находят способы использовать данные об аудитории, чтобы принимать обоснованные решения о предлагаемых публике развлечениях.

Но эта перемена влечет за собой необходимость в большем количестве данных. По этой причине сбор данных не прекращается, как только вы посмотрели выбранный для вас фильм; любые последующие комментарии, которые вы оставляете в социальных сетях или шлете по электронной почте, изменение ваших привычек просмотра фильмов в интернете генерируют о вас как о «кинозрителе» свежий массив данных, который учитывается в любых будущих рекомендациях, прежде чем наконец вы станете частью какой-либо демографической группы. Таким образом, по мере того как из подростка-эмо, интересующегося только демоническим пением, вы превращаетесь в любителя сложной сюрреалистической буффонады, которого все избегают на коктейльных вечеринках, ваши данные будут меняться вместе с вами и адаптироваться к этим колеблющимся предпочтениям.

В качестве примечания: еще более приятная новость состоит в том, что данные не отрицают ваших интересов. Если вы только *прикидываетесь* знатоком, но в действительности, как только опускаете шторы, до сих пор наслаждаетесь дрянными фильмами о зомби, ваши данные сохранят этот тайный вскормленный вами энтузиазм.

Конечно, оборотная сторона медали в том, что ваши данные могут выдавать секреты, касающиеся ваших предпочтений. Имейте в виду, что данные — это запись действий, они не будут лгать на ваш счет. Некоторые даже тратят недюжинные усилия, чтобы скрыть свой «фактический» след на сайтах цифровых музыкальных сервисов, теша собственное тщеславие: они запускают альбом музыки, которая, по их мнению, служит в обществе признаком хорошего вкуса, но не слушают ее, так что их накопленные данные представят искаженную версию того, что им нравится. На мой взгляд, у этих людей слишком много свободного времени, но манипулирование данными тем не менее является важной темой, и со временем мы вернемся к ней.

Кейс: Netflix

Сериал «Карточный домик», выпущенный развлекательной компанией Netflix, впервые доказал индустрии, насколько сильны могут быть данные не только в том, что касается охвата нужной аудитории определенными разновидностями контента, но и в управлении фактическим *производством* контента.

Сериал — политическая драма — выпуска 2013 г. был первой проверкой того, как данные могут быть применены в производстве хитов. В преддверии создания «Карточного домика» Netflix собирала данные о своих пользователях. Полученные сведения о зрительских привычках позволили Netflix группировать свой видеоконтент в разнообразные и даже удивительные категории. Интерфейс скрывал от пользователей эти категории, но тем не менее они были использованы компанией, чтобы представить нужный фильм нужной аудитории.

Когда информация об этих подкатегориях появилась в интернете несколько лет назад, люди были ошеломлены. Чтобы вы могли получить представление о том, насколько точно действовала Netflix, вот некоторые варианты подкатегорий: «Захватывающие фильмы ужасов 1980-х», «Хорошее образование и воспитание с участием героев “Маппет-шоу”», «Драмы шоу-бизнеса», «Глуповатая независимая сатира», «Откровенные фильмы о реальной жизни», «Умные фильмы о заграничных войнах», «Бросающие в дрожь триллеры» и «Признанные критиками мрачные фильмы-экранизации». Таковы весьма специфические предпочтения зрителей. Но Netflix нашла значительную аудиторию для каждой из этих категорий и для многих других.

В конце концов исследователи данных в Netflix начали видеть совпадения в зрительских моделях их аудитории. Оказалось, что существует значительное число подписчиков Netflix, которые наслаждались и работой Кевина Спейси, и серьезными политическими драмами. Остальное — перезапуск оригинального «Карточного домика» 1990-х гг. с Кевином Спейси в главной роли — это история (или это данные?).

Оседлав волну успеха

Netflix оказалась права, высоко оценив возможности данных: сериал «Карточный домик» был отмечен наградами и получил высокие оценки критиков. Поэтому неудивительно, что многие конкуренты Netflix попытались скопировать эту выигрышную модель. Хейделин де Понтевес, предприниматель в области данных и мой бизнес-партнер, работал на конкурента Netflix в целях создания подобной системы.

> «Мы знали, что у Netflix уже есть мощная система рекомендаций, и поэтому от нас как разработчиков баз данных и операционных систем требовалось не создать то же самое для нашей компании, а найти, где можно добиться разницы. Мы поняли, что для разработки действительно интересной системы нам нужно сделать больше чем просто инструмент для рекомендаций фильмов, соответствующих определенным демографическим сегментам. Мы также хотели создать алгоритм, позволяющий предлагать фильмы, которые могли бы вывести пользователей из их зоны комфорта, но в то же время доставить им удовольствие. Мы действительно стремились к тому, чтобы появился некий элемент неожиданности».
>
> (Де Понтевес, 2017 г.)

Хейделин понимал, что для достижения этой цели потребуется сложная система, способная проникнуть в головы пользователей и понять их предпочтения лучше, чем те сами понимали это. Он достиг цели, извлекая все имевшиеся у компании данные по клиентам и применяя правильное сочетание моделей, чтобы найти связи между зрительскими привычками. Помните, что этот подход почти такой же, как был у Джорджа Гэллапа многие годы назад; благодаря доступным технологиям и воображению аналитика данных мы теперь можем получить доступ к данным гораздо более хитроумным (и автоматизированным) способом.

Использование данных

Некоторые могут посетовать, что такой подход к использованию данных для творческого контента фактически убивает творчество. На это я бы ответил им, что данные всего лишь следуют за тем, чего хотят люди. Для любой отрасли желательно показать нужной аудитории в нужное время и в нужном месте соответствующий контент, чтобы побудить клиентов покупать их услуги. Таким образом, данные сделали индустрию более демократичной, потому что, хотя машины могут начать влиять на наши предпочтения в покупках, мы по-прежнему сохраняем самую ценную информацию: человеческое желание. Машины не говорят нам, чего мы хотим; они создают для нас связи, о которых мы, возможно, не знали.

Данные не приказывают людям идти и смотреть фильмы о супергероях и не смотреть французские сюрреалистические фильмы; они прислушиваются к тому, чего люди хотят и от чего получают удовольствие*. Если вы считаете, что существует проблема удушения творчества, то это не вина данных — это вина нашего общества. Я не устану подчеркивать, что данные *являются* прошлым. Это всего лишь запись информации. Если вы *хотите* видеть больше французских сюрреалистических фильмов, то просто идите и смотрите их — и убедитесь, что после просмотра вы о них говорите**. Может показаться, что вы просто добавляете шума в интернете, но этот «шум» быстро обрабатывается и становится доступным для использования повсюду. Благодаря данным в нынешнюю эпоху наши голоса действительно могут быть услышаны и иметь реальную власть — так почему бы не воспользоваться этим?

Кроме того, модели для использования данных еще несовершенны. В случае с медиаиндустрией другие корпорации приняли концепцию Netflix, и некоторые могут отметить, что одни преуспели больше, а другие — меньше. Но опять же, в этом нет заслуги данных, это творческий вклад людей. В конце концов, *именно здесь* находится нынешний предел нашей способности использовать данные для создания контента. Наверное, мы сможем оценить *вероятное* число людей, заинтересованных в концепции, но на карту поставлено гораздо больше, так как конечный успех любой формы развлечений будет обусловлен талантом ее создателя. Пусть это станет предупреждением для писателей и режиссеров, которые надеются получить легкие результаты, полагаясь исключительно на данные: базы

---

* Пример того, какие проблемы и возможности связаны с аналитикой данных в киноиндустрии, см. у Mishra and Sharma (2016), в докладе которых анализируется кинопроизводство и продюсирование в Индии.

** Естественно, на пути этого подхода есть препятствия. Вы не сможете победить миллионы поклонников супергероев в Китае, которые в значительной степени отвечают за то, что Голливуд продолжает наращивать выпуск фильмов о мужчинах (и женщинах) в колготках, спасающих мир от зла. Вопросы о том, как данные влияют на творчество, возможно, выходят за рамки этой книги, но я бы сказал, что всегда существовало и всегда будет существовать пространство для творчества, даже в мире, управляемом данными. Мы не становимся тупее; мы просто делаем промышленность более эффективной.

данных, которые показывают успех фильмов разных жанров, могут быть полезным руководством для последующих действий, но будут оставаться только руководством, поскольку результат работы зависит от таланта человека.

Почему данные важны сейчас

Многие уже в курсе того, что технологии в будущем могут существенно повлиять на рабочие места. Если вы чувствуете себя достаточно смелым, введите в поисковую строку Google «технологическое воздействие на рабочие места» /«technological impact on jobs» — и вы увидите, что несметное количество статей посвящено вероятности автоматизации в сфере вашей деятельности*. Хотя эта информация подкреплена данными, я бы сказал, что, возможно, мнение исследователей в некоторой степени субъективно, если принять во внимание задачи, которые необходимо выполнять на конкретных рабочих местах. Так, я бы, конечно, не рекомендовал учиться на спортивного арбитра по той причине, что эта работа зависит от *данных* об игре, — машины неизбежно будут поставлять более точные данные, чтобы подтвердить или опровергнуть любые заявления соперников. Судья может быть данью традиции, которая делает опыт более личностным или захватывающим *прямо сейчас*, но, на мой взгляд, ностальгия, связанная с профессией, не означает, что она будет востребована вечно.

Даже после того, как выяснилось, насколько всепоглощающими являются данные, некоторые все еще могут надеяться на то, что наука о данных не повлияет на их бизнес в ближайшее время. В конце концов, нужно время, чтобы что-то произошло. Но думать таким образом было бы большой ошибкой, потому что это отрицало бы принцип закона Мура.

---

* Опасения по поводу технологической безработицы не новы — Джон Мейнард Кейнс писал об этом в 1930-х гг.: «Мы страдаем от новой болезни, названия которой некоторые читатели, возможно, еще не слышали, но о которой они многое услышат в ближайшие годы, а именно — о технологической безработице» (Кейнс, 1963).

Закон Мура

Закон Мура — это закон прогнозирования. Предложенный соучредителем Intel Гордоном Муром в 1965 г., он в первую очередь касался ожидаемого со временем увеличения числа транзисторов (устройств, используемых для управления электрическим током) на квадратный дюйм в интегральных схемах (например, компьютерных микросхемах, микропроцессорах, материнских платах). Было замечено, что число этих транзисторов примерно удваивается каждые два года, и закон утверждал, что тенденция будет продолжаться. На сегодняшний день это подтвердилось*.

В восприятии непрофессионала это означает, что, если вы пойдете в свой местный компьютерный магазин сегодня и купите компьютер за £1000, а через два года приобретете еще один тоже за £1000 в том же магазине, вторая машина будет в два раза мощнее, хотя она стоит столько же.

Многие применили этот закон к растущему как грибы количеству достижений в области науки о данных. Она является одной из самых быстроразвивающихся академических дисциплин, и занимающиеся ею профессионалы используют все более изощренные способы, чтобы найти новые средства для сбора данных, построения экономичных систем их хранения и разработки алгоритмов, которые превращают все эти порции больших данных в ценные идеи. Доводилось ли вам когда-либо чувствовать, что технологии движутся вперед так быстро, что вы не успеваете идти в ногу со временем? Тогда подумайте об аналитиках данных. Они играют в салочки с технологией, которая *еще даже не изобретена*.

Кейс: Siri

В качестве примера рассмотрим развитие технологии распознавания речи. Создатели Siri Даг Киттлаус, Адам Чейер и Том Грубер разработали умного личного

* Относительно транзисторной инфраструктуры у закона Мура есть ограничения. При размере около 1 нм свойства полупроводникового материала нарушаются такими квантовыми эффектами, как квантовое туннелирование. Кроме того, дальнейшее развитие инфраструктуры потребует альтернативы кремнию, который сейчас используется в качестве основного материала. — *Прим. науч. ред.*

помощника задолго до того, как технология стала достаточно зрелой, чтобы можно было реализовать идеи и вывести их на рынок. Авторы Siri создали инструменты и алгоритмы для работы с имевшимися у них данными, чтобы поддерживать технологию распознавания речи, которая тогда еще не была изобретена.

Однако они знали, что, хотя было невозможно использовать программное обеспечение с имевшейся в то время технологией, в конечном итоге запуск Siri *станет возможным*, нужно лишь подождать, пока технология выкристаллизуется. Короче говоря, они уловили технологические тенденции.

Концепцией, которую создатели Siri использовали для своих прогнозов, служил закон Мура. И это невероятно важно для науки о данных. Закон Мура применяется к многим технологическим процессам и является необходимым правилом при рассмотрении и принятии деловых решений и реализации проектов; мы вернемся к его обсуждению в главе 3 «Мышление, необходимое для эффективного анализа данных».

Беспокойство ни к чему не приводит

Голливуд и индустрия развлечений в целом долгое время придерживались мрачной идеи, что использование данных и связанные с ними злоупотребления угрожают человечеству. Нам стоит задуматься над этой не предвещающей ничего хорошего фразой из фильма «2001: Космическая одиссея»: «Открой дверь модульного отсека, ЭАЛ», где ЭАЛ — технология искусственного интеллекта (ИИ) космического корабля — настолько усовершенствован, что решает не подчиняться команде человека и действовать согласно своим (превосходящим) суждениям. «Из машины», «Она», «Бегущий по лезвию», «Призрак в доспехах» — все эти фильмы посвящены воображаемым проблемам, с которыми могут столкнуться люди, когда технологии начнут развивать собственное сознание и предвидеть наши действия.

Но есть, с моей точки зрения, еще одна область, где злонамеренное применение данных — имеющее значительно больше общего

с злоупотреблениями *людей*, чем роботов, — гораздо более вероятно и неотвратимо. Речь идет о конфиденциальности. С вопросами конфиденциальности связаны многие наши взаимодействия в интернете. Люди могут оставаться анонимными, но информация о них всегда будет где-то собираться — и использоваться. Даже если эти данные лишены характерных индикаторов, отсылающих к тому или иному индивидууму, некоторые могут спросить: «Правильно ли, что такие данные вообще собирают?»

Ваш онлайн-след

Читатели, которые пользовались интернетом в 1990-х гг., знакомы со словом «аватар» — довольно безобидное изображение, которое мы выбирали для представления себя на онлайн-форумах. Сегодня термин «аватар» используется для описания чего-то гораздо более широкого. Теперь он означает нашего неосязаемого двойника в виртуальном мире, массив данных о нас, составленный на основе наших заданных поисков, выбора и покупок, которые мы делаем в интернете, и всего, что мы публикуем в Сети, от текста до изображений. Такие данные являются потенциальным золотым дном, неиссякаемым источником информации для кредитных агентств и компаний-агрегаторов, которые затем могут использовать эти сведения для продажи другим.

Ввиду развития науки о данных встают вопросы этики и безопасности, касающиеся проницаемости, искажения и захвата данных (а этика — это область, которую мы рассмотрим в главе 5 «Подготовка данных»). У нас есть очень веские основания беспокоиться о доступах, которые открывает наука о данных, и о том, что она не делает различий в том, кто — или что — обращается к этой информации. Хотя переход от бумажного к цифровому документообороту позитивно сказался на практике ведения дел в компаниях, данные все еще могут пропадать или приходить в негодность, а также на них может существенно повлиять человек (это касается неверной информации, потери баз данных и шпионажа), что будет иметь разрушительные последствия.