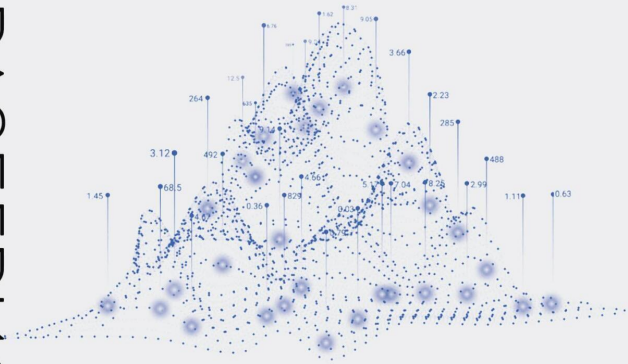
РАЗБЕРИСЬ В



DATA SCIENCE

КАК ОСВОИТЬ НАУКУ О ДАННЫХ И НАУЧИТЬСЯ ДУМАТЬ КАК ЭКСПЕРТ

АЛЕКС ДЖ. ГАТМАН
ДЖОРДАН ГОЛДМЕЙЕР

- МЕТОДЫ АНАЛИТИКИ ДАННЫХ
- ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ
- РАСПРОСТРАНЕННЫЕ ОШИБКИ ПРИ РАБОТЕ С ДАННЫМИ

БОМБОРА
ИЗДАТЕЛЬСТВО

Алекс Дж. Гатман
Джордан Голдмейер

Разберись в Data Science. Как освоить науку о данных и научиться думать как эксперт

Серия «Мировой компьютерный бестселлер»

indd предоставлен правообладателем
http://www.litres.ru/pages/biblio_book/?art=69106330
ISBN 978-5-04-184971-9

Аннотация

Перед вами исчерпывающее руководство по основам Data Science. С помощью него вы сможете научиться мыслить статистически и понимать, какую роль в вашей работе играет аналитика, пользоваться языком науки о данных, избегать распространенных ошибок при работе с ними и, наконец, разобраться в полезных инструментах, которые используют эксперты.

В формате PDF A4 сохранен издательский макет книги.

Содержание

Алекс Дж. Гатман, Джордан Голдмейер Разберись в Data Science

Как освоить науку о данных и научиться думать как эксперт

Jordan Goldmeier, Alex J. Gutman

BECOMING A DATA HEAD: How to Think, Speak and Understand Data Science, Statistics and Machine Learning

* * *

Посвящается моим детям Элли, Уильяму и Эллен.

Элли было три года, когда она узнала, что ее папа – «доктор».

Озадаченно посмотрев на меня, она сказала: «Но ведь ты не помогаешь людям…»

Памятуя об этом, я также посвящаю эту книгу вам, читатель.

Надеюсь, что она вам поможет.

– Алекс

Посвящается Стивену и Мелиссе.

– Джордан

Предисловие

Книга «Разберись в Data Science» вышла очень своевременно, учитывая текущую ситуацию с данными и аналитикой в организациях. Давайте кратко пробежимся по последним событиям. Начиная с 1970-х годов лишь немногие передовые компании эффективно использовали данные и аналитику для принятия решений и обоснования своих действий. Большинство игнорировало этот ценный ресурс или не придавало ему особого значения.

В 2000-х годах ситуация стала меняться, и компании начали понимать, как они могут изменить свою ситуацию с помощью данных и аналитики. К началу 2010-х годов интерес стал смещаться в сторону «больших данных», которые изначально появились в интернет-компаниях, а затем распространились по всей экономике. В связи с возросшим объемом и сложностью данных в компаниях возникла роль «дата-сайентиста», опять же, сначала в Силиконовой долине, а затем повсюду.

Однако как только фирмы начали приспосабливаться к большим данным, в период с 2015 по 2018 год акцент во многих фирмах снова сместился, на этот раз в сторону искусственного интеллекта. Сбор, хранение и анализ больших данных уступили место машинному обучению, обработке естественного языка и автоматизации.

В основе этих быстрых сдвигов фокуса лежал ряд допущений относительно данных и аналитики, распространенных внутри организаций. Я рад сообщить, что книга «Разберись в Data Science» разрушает многие из них и делает это весьма своевременно. Многие люди, внимательно наблюдающие за этими тенденциями, уже начинают признавать, что эти допущения направляют нас по непродуктивному пути. В оставшейся части этого предисловия я опишу пять взаимосвязанных допущений и то, как изложенные в этой книге идеи обоснованно опровергают их.

Допущение 1. Аналитика, большие данные и ИИ – совершенно разные явления.

Многие полагают, что «традиционная» аналитика, большие данные и ИИ – это отдельные явления. Однако авторы книги «Разберись в Data Science» справедливо считают, что эти вещи тесно связаны друг с другом. Все они требуют статистического мышления, использования традиционных аналитических подходов, вроде регрессионного анализа, а также методов визуализации данных. Предиктивная аналитика – это, по сути, то же самое, что и контролируемое машинное обучение. Кроме того, большинство методов анализа данных работают с наборами данных любого размера. Короче говоря, главный по данным может эффективно работать во всех трех областях, так что заострять внимание на различиях между ними не очень продуктивно.

Допущение 2. В этой песочнице могут играть только дата-сайентисты.

Мы часто прославляли дата-сайентистов, полагая, что только они способны эффективно работать с данными и аналитикой. Тем не менее в настоящее время зарождается важная тенденция к демократизации этих идей, и все больше организаций расширяют полномочия «гражданских специалистов по работе с данным». Автоматизированные инструменты машинного обучения упрощают создание моделей, которые отлично справляются с прогнозированием. Разумеется, нам все еще нужны профессиональные дата-сайентисты для разработки новых алгоритмов и проверки работы гражданских специалистов, занимающихся сложным анализом. Однако организации, которые демократизируют занятие аналитикой и наукой о данных, привлекая к этому «любителей», способны значительно расширить использование этих важных возможностей.

Допущение 3. Дата-сайентисты – это единороги, обладающими всеми необходимыми навыками.

Мы привыкли полагать, что дата-сайентисты, умеющие разрабатывать модели, также способны решать все остальные задачи, связанные с внедрением этих моделей. Другими словами, мы считаем их своеобразными «единорогами», которые могут все. Но таких «единорогов» нет вообще, или

они существуют лишь в небольшом количестве. Главные по данным, которые понимают не только основы науки о данных, но и особенности бизнеса, а также способны эффективно управлять проектами и выстраивать деловые отношения, будут чрезвычайно ценны как участники проектов по работе с данными. Они могут стать продуктивными членами команд дата-сайентистов и повысить вероятность того, что проекты по работе с данными принесут бизнесу пользу.

Допущение 4. Чтобы преуспеть в работе с данными и аналитикой, вам необходимы выдающиеся математические способности и много тренировок.

Еще одно похожее допущение сводится к тому, что для работы с данными человек должен быть очень хорошо подготовлен в этой области, а также хорошо разбираться в математике. Математические способности и подготовка, безусловно, очень важны, но авторы книги «Разберись в Data Science» утверждают (и я с ними согласен), что мотивированный ученик способен освоить необходимые навыки в достаточной степени для того, чтобы стать полезным участником проектов по работе с данными. Во-первых, общие принципы статистического анализа далеко не так сложны, как может показаться. Во-вторых, для того, чтобы «быть полезным» участником проектов по работе с данными, ваш уровень владения аналитикой не обязательно должен быть чрезвычайно высоким. Работа с профессиональными дата-сайентистами или

автоматизированными ИИ-программами требует лишь любознательности и умения задавать хорошие вопросы, находить взаимосвязи между бизнес-проблемами и количественными результатами, а также обращать внимание на сомнительные предположения.

Допущение 5. Если в колледже или аспирантуре вы не занимались в основном количественными предметами, вам слишком поздно осваивать навыки, необходимые для работы с данными и аналитикой.

Это предположение подтверждается данными опросов. Согласно результатам опроса, проведенного компанией Splunk в 2019 году, в котором приняли участие около 1300 руководителей по всему миру, практически каждый респондент (98 %) согласен с тем, что навыки работы с данными важны для специалистов будущего[1]. А 81 % респондентов считает, что навыки работы с данными необходимы для того, чтобы стать старшим руководителем в их компаниях, а 85 % согласны с тем, что ценность таких навыков в их фирмах будет расти. Тем не менее 67 % респондентов заявили, что им неудобно получать доступ к данным или использовать их самостоятельно, 73 % считают, что навыки работы с данными труднее освоить, чем другие бизнес-навыки, а 53 % – что они слишком стары для освоения навыков работы с данны-

[1] Splunk Inc., "The State of Dark Data," 2019, **www.splunk.com/en_us/form/the-state-of-dark-data.html**.

ми. Подобное пораженчество наносит ущерб как отдельным лицам, так и организациям в целом, и ни авторы этой книги, ни я не считаем его оправданным. В ходе чтения этой книги вы увидите, что в этом нет ничего сложного!

Итак, отбросьте эти ложные допущения и станьте главным по данным. Это позволит вам повысить свою ценность как сотрудника и сделать свою организацию более успешной. Именно по этому пути движется мир, так что пришло время узнать больше о данных и аналитике. Я уверен, что процесс чтения книги «Разберись в Data Science» окажется гораздо более полезным и приятным, чем вы можете себе представить.

Томас Х. Дэвенпорт
Заслуженный профессор Бэбсон-колледжа, приглашенный профессор Бизнес-школы Саида при Оксфордском университете, научный сотрудник инициативы Массачусетского технологического института в сфере цифровой экономики, автор книг «Аналитика как конкурентное преимущество», «Внедрение искусственного интеллекта в бизнес-практику: Преимущества и сложности» и «Big Data @ Work»

Введение

Данные – это, пожалуй, важнейший аспект вашей работы, нравится вам это или нет. И, скорее всего, вы решили прочитать эту книгу, чтобы лучше в них разобраться.

Для начала стоит констатировать то, что уже почти превратилось в клише: в настоящее время мы создаем и потребляем больше информации, чем когда-либо прежде. Мы, без сомнения, живем в эпоху данных, которая породила массу обещаний, модных словечек и продуктов, многие из которых вы, ваши менеджеры, коллеги и подчиненные уже используете или будете использовать. Однако, несмотря на распространение этих обещаний и продуктов, проекты по работе с данными терпят неудачу с пугающей регулярностью [2].

Разумеется, мы не утверждаем, что все обещания пусты, а продукты – ужасны. Скорее, чтобы по-настоящему разобраться в этой области, вы должны принять фундаментальную истину: работа с данными очень сложна и сопряжена с нюансами и неопределенностью. Данные, безусловно, важны, но работать с ними совсем не просто. И все же существует целая индустрия, которая заставляет нас думать иначе, обещает определенность в мире неопределенности и иг-

---

[2] Venture Beat. "87 % of data science projects failing": **venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production**

рает на страхе компаний упустить выгоду. Мы называем это промышленным комплексом науки о данных.

Промышленный комплекс науки о данных

Эта проблема касается всех. Компании бесконечно ищут продукты, которые думали бы за них. Менеджеры нанимают профессионалов в области аналитики, которые на самом деле таковыми не являются. Дата-сайентистов нанимают для работы в компаниях, которые к ним не готовы. Руководители вынуждены слушать техническую болтовню и делать вид, что понимают, о чем идет речь. Работа над проектами стопорится. Деньги тратятся впустую.

Тем временем промышленный комплекс науки о данных штампует новые концепции быстрее, чем мы можем определить и сформулировать порождаемые ими возможности (и проблемы). Стоит моргнуть, и обязательно что-нибудь пропустишь. Когда авторы этой книги начали работать вместе, все говорили о больших данных. Со временем популярной новой темой стала наука о данных. Затем внимание общественности сосредоточилось на машинном обучении, глубоком обучении и искусственном интеллекте.

Но самых любознательных и критически мыслящих из нас что-то не устраивает. Действительно ли эти проблемы новые? Или они просто переосмысление старых?

Ответ на оба вопроса утвердительный.

Однако мы надеемся, что вы задаетесь более важным во-

просом – «Как научиться критически мыслить и говорить о данных?»

Мы вас этому научим.

В этой книге вы познакомитесь с инструментами, терминами и образом мышления, необходимыми для навигации по промышленному комплексу науки о данных. Вы научитесь понимать данные и связанные с ними проблемы на более глубоком уровне, критически относиться к данным и результатам, с которыми сталкиваетесь, а также разумно говорить обо всем, что касается данных.

Короче говоря, вы станете главным по данным.

Почему нам это важно

Прежде чем мы начнем, стоит сказать, почему авторов этой книги, Алекса и Джордана, так волнует эта тема. В этом разделе мы опишем два важных примера того, как данные повлияли на общество в целом и на нас лично.

Кризис субстандартного ипотечного кредитования

Мы едва закончили колледж, когда разразился кризис субстандартного ипотечного кредитования. Мы оба устроились на работу в BBC в 2009 году, когда найти работу было очень трудно. Нам повезло, поскольку мы обладали востребованным навыком – мы умели работать с данными. Мы каждый день работали над преобразованием результатов исследований, проведенных аналитиками и учеными BBC, в продукты, которые могло бы использовать правительство. Наш прием на работу стал предвестником грядущего роста важности тех ролей, которые мы исполняли. Будучи специалистами по работе с данными, мы наблюдали за развитием ипотечного кризиса с интересом и любопытством.

У кризиса субстандартного ипотечного кредитования бы-

ло множество причин[3]. Приводя его здесь в качестве примера, мы не отрицаем прочие факторы, однако, по нашему мнению, важнейшим из них была серьезная проблема с данными. Банки и инвесторы создали модели для оценки ценности обеспеченных ипотекой долговых обязательств (CDO) – инвестиционных инструментов, ставших причиной обвала рынка США.

Облигации с ипотечным покрытием считались безопасными инструментами, поскольку распределяли риск дефолта по кредиту между несколькими инвестиционными единицами. Идея заключалась в том, что если лишь некоторые активы в портфеле ипотечных кредитов окажутся убыточными, это не окажет существенного влияния на стоимость всего портфеля.

И все же, если поразмыслить, становится очевидно, что некоторые фундаментальные предположения были неверны. В первую очередь речь идет о допущении независимости между возможными дефолтами, то есть предположении о том, что если заемщик А не выполнит обязательства по кредиту, это не повлияет на риск неплатежа заемщика Б. Впоследствии мы узнали о том, что дефолты происходят по принципу домино, то есть предыдущий дефолт может предсказать вероятность дальнейших дефолтов. Дефолт по одному ипотечному кредиту приводил к снижению стоимости

---

3 **www.brookings.edu/wp-content/uploads/2016/06/11_origins_crisis_baily_litan.pdf**

находящейся поблизости недвижимости, что способствовало росту риска дефолта по соответствующим кредитам. По сути, один дом утягивал за собой соседние.

Допущение независимости фактически связанных между собой событий – распространенная ошибка в статистике.

Но давайте углубимся в эту историю. Инвестиционные банки создали модели, которые переоценили эти инвестиции. Модели, о которых мы поговорим далее в книге, – это упрощенные версии реальности. Они используют предположения о реальном мире для понимания и предсказания определенных явлений.

А кто создавал эти модели? Это были люди, которые заложили основы будущей профессии дата-сайентиста. Люди вроде нас. Статистики, экономисты, физики – люди, которые занимались машинным обучением, искусственным интеллектом и статистикой. Они работали с данными. И они были умны. Невероятно умны.

И все же что-то пошло не так. Может быть, они не сумели задать правильные вопросы? Или информация о риске и неопределенности не была должным образом донесена до лиц, принимающих решения, в результате чего у них возникла иллюзия совершенно предсказуемого рынка недвижимости? А может быть, кто-то откровенно соврал о результатах?

Но больше всего нас интересовало то, как избежать подобных ошибок в нашей собственной работе?

У нас было много вопросов, и об ответах мы могли лишь

гадать, но одно было ясно – это была крупномасштабная катастрофа с данными. И она обещала быть не последней.

Всеобщие выборы в США 2016 года

8 ноября 2016 года кандидат от республиканцев Дональд Дж. Трамп победил на всеобщих выборах в Соединенных Штатах, обойдя предполагаемого лидера и кандидата от демократической партии Хиллари Клинтон. Для политических социологов это стало настоящим шоком, поскольку их модели не предсказывали его победу. А год был самым подходящим для подобных предсказаний.

В 2008 году Нейт Сильвер, автор блога FiveThirtyEight, тогда бывшего частью газеты *The New York Times*, проделал фантастическую работу и предсказал победу Барака Обамы. В то время эксперты скептически относились к способности его алгоритма прогнозирования точно предсказывать результаты выборов. В 2012 году Нейт Сильвер снова оказался в центре внимания, предсказав очередную победу Обамы.

К этому моменту деловой мир уже начал осваивать работу с данными и нанимать дата-сайентистов. Успешное предсказание переизбрания Барака Обамы Нейтом Сильвером лишь подчеркнуло важность и оракулоподобные возможности прогнозирования на основе данных. Статьи в деловых журналах предостерегали руководителей о том, что если они не освоят работу с данными, то проиграют в конкурентной

борьбе. Промышленный комплекс науки о данных заработал в полную силу.

К 2016 году каждое крупное новостное издание вложило средства в алгоритм предсказания исхода всеобщих выборов. Подавляющее большинство из них прогнозировали сокрушительную победу кандидата от демократической партии Хиллари Клинтон. Как же они ошибались.

Давайте сравним эту ошибку с кризисом субстандартного ипотечного кредитования. Можно было бы утверждать, что мы многому научились и что интерес к науке о данных должен был бы позволить избежать ошибок прошлого. Действительно, начиная с 2008 года, новостные организации стали нанимать дата-сайентистов, вкладывать средства в проведение опросов общественного мнения, формировать команды аналитиков и тратить большое количество денег на сбор качественных данных.

Что же произошло, учитывая все это время, деньги, усилия и образование?[4]

Наша гипотеза

Почему возникают подобные проблемы с данными? Мы видим три причины: сложность проблемы, недостаток кри-

[4] Нейт Сильвер написал по этому поводу целую серию статей (**fivethirtyeight.com/tag/the-real-story-of-2016**). Одна из ошибок социологов заключалась в допущении независимости событий, как и в случае с ипотечным кризисом.

тического мышления и плохая коммуникация.

Во-первых (как мы уже говорили), работа с данными зачастую очень сложна. Даже при наличии большого количества данных, подходящих инструментов, методик и умнейших аналитиков случаются ошибки. Прогнозы могут и будут оказываться ошибочными. И это не критика данных и статистики. Такова реальность.

Во-вторых, некоторые аналитики и заинтересованные стороны перестали критически относиться к проблемам данных. Промышленный комплекс науки о данных в своем высокомерии нарисовал картину уверенности и простоты, и некоторые люди на нее купились. Возможно, такова человеческая природа: люди не хотят признавать, что не знают будущего. Однако ключевым аспектом правильного осмысления и использования данных является признание возможности принятия неверного решения. Это означает понимание и распространение информации о рисках и неопределенностях. Но эта идея где-то затерялась. Мы надеялись, что колоссальный прогресс в исследованиях и методах анализа и работы с данными обострит критическое мышление каждого человека, но, судя по всему, некоторые люди его, наоборот, отключили.

Третья причина возникновения проблем с данными, по нашему мнению, – плохая коммуникация между дата-сайентистами и лицами, принимающими решения. Даже при наличии самых лучших намерений результаты зачастую доно-

сятся с искажениями. Лица, принимающие решения, не говорят на языке данных, потому что никто не удосужился их этому научить. Кроме того, специалисты по работе с данными далеко не всегда способны понятно объяснить те или иные вещи. Итак, существует пробел в общении.

Данные на рабочем месте

Ваши проблемы с данными, скорее всего, не грозят обрушением мировой экономики или неправильным предсказанием результатов следующих президентских выборов в США, но контекст этих историй имеет значение. Если недопонимание и ошибки в критическом мышлении случаются на глазах у всего мира, то, вероятно, это происходит на вашем рабочем месте. В большинстве случаев эти микросбои укрепляют культуру безграмотности в отношении данных.

Это происходило и на нашем рабочем месте и отчасти по нашей вине.

Сцена в зале заседаний

Поклонникам научной фантастики и приключенческих фильмов хорошо знакома такая сцена: герой сталкивается, казалось бы, с нерешаемой задачей, и мировые лидеры и ученые собираются вместе, чтобы обсудить ситуацию. Один из ученых, самый занудный среди всей группы, предлагает идею, используя непонятный жаргон, а генерал обрывает его, требуя «говорить по-человечески». После этого зритель получает некоторое объяснение того, что имелось в виду. Суть этого момента – преобразование критически важной для миссии информации в то, что способен понять не

только наш герой, но и зритель.

Мы часто обсуждали этот сюжет в контексте нашей роли исследователей для федерального правительства. Почему? Потому что нам казалось, что ситуация никогда не разворачивалась таким образом. На ранних этапах нашей карьеры мы часто наблюдали нечто противоположное.

Мы представляли нашу работу людям, смотревшим на нас пустыми глазами, которые вяло кивали, а иногда почти засыпали. Мы наблюдали за тем, как сбитые с толку зрители воспринимали все, что мы говорили, без единого вопроса. Их либо впечатляло то, какими умными мы казались, либо им было скучно, потому что они ничего не понимали. Никто не просил повторить сказанное на понятном всем языке. Очень часто ситуация разворачивалась следующим образом:

Мы: «Проведя анализ бинарной переменной отклика методом контролируемого обучения с использованием множественной логистической регрессии, мы получили вневыборочную производительность со специфичностью 0,76 и несколько статистически значимых независимых переменных с использованием значений альфа равных 0,05».

*Бизнес-профессионал: *неловкое молчание**

Мы: «Это понятно?»

*Бизнес-профессионал: *снова тишина**

Мы: «Есть вопросы?»

Бизнес-профессионал: «В данный момент вопросов нет».

Внутренний монолог бизнес-профессионала: «О чем, черт возьми, они говорят?»

Увидев подобную сцену в кино, вы могли бы подумать: надо перемотать назад, возможно, я что-то упустил. Но в реальной жизни, когда принимаемые решения имеют огромное влияние на результат миссии, такое случается редко. Мы не перематываем. Мы не просим разъяснений.

Оглядываясь назад, мы понимаем, что наши презентации были слишком техническими. Отчасти причина заключалась в банальном упрямстве: до ипотечного кризиса технические детали чрезмерно упрощались; аналитиков приглашали для того, чтобы они говорили руководителям то, что те хотели услышать, но мы не собирались играть в эту игру. Мы хотели, чтобы наши зрители понимали нас.

Но мы перестарались. Наша аудитория не могла критически осмыслить результаты нашей работы, потому что не понимала, о чем мы говорили.

Мы подумали, что должен быть способ получше. Мы хотели повлиять на ситуацию с помощью своей работы, поэтому начали практиковаться в объяснении сложных статистических концепций друг другу и нашим зрителям, а также исследовать то, как наши объяснения воспринимают другие люди.

Нам удалось обнаружить точку соприкосновения между специалистами по работе с данными и бизнес-профессионалами, в которой могут иметь место честные дискуссии о дан-

ных, не будучи при этом слишком техническими или слишком упрощенными. Это предполагает более критическое отношение обеих сторон к проблемам данных вне зависимости от их масштаба. Именно об этом и пойдет речь в этой книге.

Вы можете понять общую картину

Для лучшего понимания данных и работы с ними вам необходимо быть готовым к изучению сложных концепций. И даже если вы уже знакомы с ними, мы научим вас тому, как донести их до вашей аудитории.

Вам также предстоит принять такой редко обсуждаемый факт, что во многих компаниях работа с данными оказывается неэффективной. Вы разовьете интуицию, понимание и здоровый скептицизм в отношении чисел и терминов, с которыми сталкиваетесь. Эта задача может показаться сложной, но эта книга поможет вам ее решить. И для этого вам не понадобятся ни навыки программирования, ни докторская степень.

С помощью четких объяснений, мысленных упражнений и аналогий вы сможете выстроить ментальную модель для понимания науки о данных, статистики и машинного обучения.

В следующем примере мы сделаем именно это.

Классификация ресторанов

Представьте, что вы идете по улице и видите пустую витрину с вывеской «Новый ресторан: скоро открытие». Вы устали питаться в сетевых ресторанах и постоянно ищете но-

вые местные заведения, поэтому задаетесь вопросом: «Появится ли здесь новый независимый ресторан?»

Давайте поставим этот вопрос более формально: как вы думаете, будет ли новый ресторан сетевым или независимым?

Угадайте. (Серьезно, подумайте об этом, прежде чем двигаться дальше.)

В реальной жизни вы сделали бы довольно хорошее предположение за доли секунды. Находясь в модном районе с множеством местных пабов и закусочных, вы бы предположили, что ресторан будет независимым. А если бы речь шла о межштатной автомагистрали с расположенным рядом торговым центром, вы бы предположили, что ресторан будет сетевым.

Но когда мы задали вопрос, вы заколебались. Вы подумали, что мы предоставили недостаточно информации. И вы были правы. Мы не предоставили вам никаких данных для принятия решения.

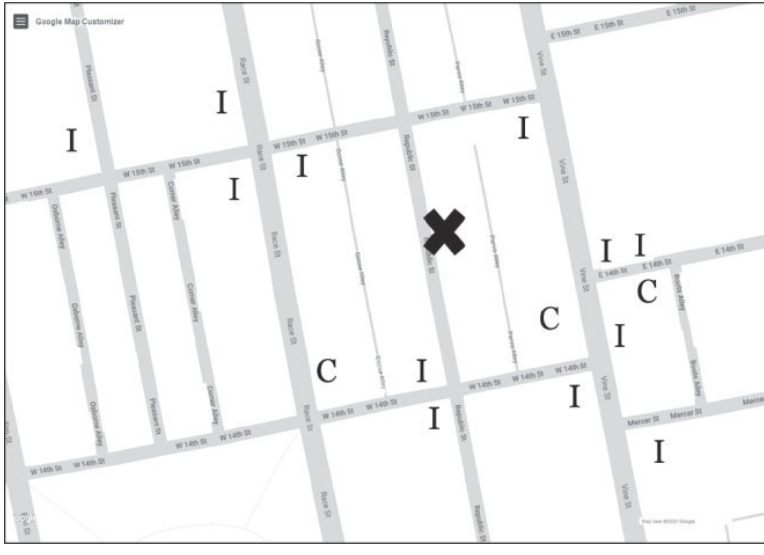
Мораль: для принятия обоснованных решений требуются данные.

Теперь посмотрите на первое изображение на следующей странице. Новый ресторан отмечен крестиком (X), буквой *C* обозначены сетевые рестораны (chain), а буквой *I* – независимые (independent) местные закусочные. Какое предположение вы сделали бы на этот раз?

Большинство людей предполагает, что ресторан будет

независимым (I), потому что такова большая часть близлежащих ресторанов. Однако обратите внимание на то, что независимыми являются далеко не все из них. Если бы мы попросили вас оценить уровень достоверности[5] вашего прогноза в диапазоне от 0 до 100, то она, скорее всего, была бы высокой, но не равной 100, поскольку по соседству вполне может появиться еще один сетевой ресторан.

Мораль заключается в следующем: предсказания никогда не могут быть на 100 % достоверными.



[5] Примечание для коллег-статистиков: мы имеем в виду обычную, а не статистическую достоверность.

Район Овер-Райн, Цинциннати, штат Огайо

Теперь взгляните на следующее изображение. В этом районе есть большой торговый центр, и большинство ресторанов здесь – сетевые. Когда людям предлагается предсказать, каким будет новый ресторан в этом районе – сетевым или независимым, большинство выбирает вариант (C). Но нам нравится, когда кто-то выбирает вариант (I), потому что это подчеркивает несколько важных моментов.

Кенвуд Таун Центр, Цинциннати, штат Огайо

В ходе этого мысленного эксперимента каждый участник создает в своей голове слегка отличающийся алгоритм. Разумеется, все смотрят на маркеры, окружающие интересующую нас точку X, чтобы понять особенности района, но в какой-то момент необходимо решить, что ресторан находится слишком далеко, чтобы повлиять на прогноз. Иногда человек видит единственный ближайший ресторан, в данном случае – независимый (I), и основывает на этом свой прогноз: «Ближайшим соседом ресторана X является независимый ресторан (I), поэтому мой прогноз – (I)».

Однако большинство людей учитывают несколько соседних ресторанов. На втором изображении вокруг нового ресторана нарисована окружность, включающая семь его ближайших соседей. Вероятно, вы выбрали другое число, но мы выбрали 7. Шесть из семи ресторанов сетевые (C), поэтому мы прогнозируем, что новый ресторан тоже будет сетевым.

Что дальше?

Если вы поняли пример с рестораном, значит, вы уже на пути становления главным по данным. Давайте пройдемся по тому, что вы узнали.

– Вы выполнили классификацию, предсказав метку для нового ресторана (сетевой или независимый), обучив алгоритм на наборе данных (содержащем местоположения ресто-

ранов и соответствующие метки).

– В этом состоит суть машинного обучения! Просто для разработки алгоритма вы использовали не компьютер, а собственную голову.

– Данный тип машинного обучения называется контролируемым обучением, потому что вы знали, что существующие рестораны были сетевыми (C) или независимыми (I). Эти метки направляли (то есть контролировали) ход ваших мыслей при размышлении о том, как расположение ресторана связано с его типом (сетевой или независимый).

– Если еще конкретнее, то вы использовали алгоритм контролируемой классификации под названием метод *k-ближайших соседей*[6]. Если K = 1, посмотрите на ближайший ресторан и получите свой прогноз. Если K = 7, посмотрите на 7 ближайших ресторанов и сделайте предсказание на основе их большинства. Это интуитивно понятный и мощный алгоритм. И в нем нет никакого волшебства.

– Вы также узнали о том, что для принятия обоснованных решений вам нужны данные. Однако помимо них вам необходимо кое-что еще. В конце концов, в этой книге много внимания уделяется критическому мышлению. Мы хотим показать не только то, как работают те или иные вещи, но и то, почему иногда они не срабатывают. Если бы мы попро-

---

[6] Метод *k*-ближайших соседей можно использовать для предсказания не только классов, но и чисел. Эти так называемые задачи регрессии мы рассмотрим далее в книге.

сили вас спрогнозировать, опираясь на приведенные в этом разделе изображения, будет ли новый ресторан ориентирован на детей, вы бы не смогли ответить. Для принятия обоснованных решений подходят далеко не любые данные. Для этого нужно достаточное количество точных и релевантных данных.

– Помните технические термины, которые мы упоминали ранее, говоря об «…анализе бинарной переменной отклика методом контролируемого обучения?..» Поздравляем, вы только что выполнили такой анализ. Переменная отклика – это просто еще одно название метки, и она является бинарной, потому что в нашем примере их было две – (C) и (I).

В этом разделе вы многое узнали, причем даже не осознавая этого.

Для кого написана эта книга?

Как говорится в начале этой книги, данные затрагивают жизни многих сотрудников современных корпораций. Мы придумали нескольких аватаров, представляющих людей, которые могут выиграть от становления главными по данным.

Мишель – специалист по маркетингу, которая работает бок о бок с аналитиком данных. Она разрабатывает маркетинговые инициативы, а ее коллега собирает данные и измеряет влияние, оказываемое этими инициативами. Мишель считает, что их работа должна быть более инновационной, но не может донести до коллеги свои потребности в данных и их анализе. Общение между ними затруднено. Она поискала в Google некоторые специальные термины (машинное обучение и прогностическая аналитика), но в большинстве найденных ею статей использовались чрезмерно технические определения, неразборчивый компьютерный код, реклама аналитического программного обеспечения или консультационных услуг. В результате поисков она почувствовала еще большую тревогу и растерянность, чем раньше.

Даг имеет докторскую степень в области наук о жизни и работает в отделе исследований и разработок крупной корпорации. Скептик по натуре, он задается

вопросом о том, не является ли шумиха вокруг данных очередным хайпом. Однако Даг старается не демонстрировать свой скептицизм на рабочем месте (особенно в присутствии нового директора, который носит футболку с надписью «Данные – это новая нефть»), поскольку не хочет, чтобы его считали дата-луддитом. В то же время он чувствует себя не у дел и решает узнать, из-за чего весь этот шум.

Реджина – топ-менеджер компании и хорошо осведомлена о последних тенденциях в области науки о данных. Она курирует новое подразделение своей компании, занимающееся наукой о данных, и регулярно взаимодействует со старшими дата-сайентистами. Реджина доверяет своим специалистам, но ей хотелось бы иметь более глубокое понимание сути их деятельности, потому что ей часто приходится представлять и отстаивать результаты работы своей команды перед советом директоров компании. Реджине также поручена проверка нового технологического программного обеспечения. Она подозревает, что некоторые заявления поставщиков относительно «искусственного интеллекта» слишком хороши, чтобы быть правдой, и хочет получить дополнительные технические знания, чтобы отделить маркетинговые заявления от реальности.

Нельсон руководит работой трех дата-сайентистов в рамках своей новой должности. Будучи специалистом по компьютерным наукам, он знает, как писать программы и работать с данными, но плохо

разбирается в статистике (поскольку прошел в колледже только один курс) и машинном обучении. Учитывая наличие технического образования, он хочет и может разобраться в деталях, но просто не находит на это времени. Руководство также побуждает его команду «больше заниматься машинным обучением», но на данный момент это кажется ей волшебным черным ящиком. Нельсон приступает к поиску материала, который поможет ему завоевать доверие команды и понять, какие проблемы можно решить с помощью машинного обучения, а какие – нет.

Мы надеемся, вы узнали себя в одном или нескольких из этих персонажей. Общим для них и, вероятно, для вас является желание стать лучшим «потребителем» данных и аналитики, с которыми вы сталкиваетесь.

Мы также создали аватар, представляющий людей, которым следует прочитать эту книгу, но которые, скорее всего, не станут этого делать (потому что в каждой истории должен быть злодей).

Джордж – менеджер среднего звена, читает последние деловые статьи об искусственном интеллекте и рассылает понравившиеся вверх и вниз по своей цепочке управления, как доказательство своей технической подкованности. Однако в зале заседаний он предпочитает «прислушиваться к своей интуиции». Джорджу нравится, когда его датасайентисты представляют ему цифры с помощью

одного или двух слайдов. Когда результаты анализа согласуются с тем, что подсказала его интуиция, прежде чем он заказал исследование, он передает их вверх по цепочке и хвастается перед коллегами «внедрением искусственного интеллекта». Если результаты анализа не согласуются с его интуицией, он задает своим дата-сайентистам ряд туманных вопросов и отправляет их на поиски «доказательств», необходимых для продвижения его проекта.

Не будьте такими, как Джордж. Если вы знаете «Джорджа», порекомендуйте ему эту книгу и скажите, что он похож на «Реджину».

Зачем мы написали эту книгу

Мы считаем, что многие люди, похожие на описанные выше аватары, хотят больше узнать о данных, но не знают, с чего начать. Существует широкий спектр книг, посвященных науке о данных и статистике. На одном конце этого спектра находятся нетехнические книги, превозносящие достоинства и перспективы работы с данными. Какие-то из них лучше, чем другие. Самые лучшие из них напоминают современные бизнес-книги. Однако многие написаны журналистами, которые стремятся драматизировать начало эпохи данных.

В этих книгах описывается то, как те или иные бизнес-проблемы были решены путем их рассмотрения через призму данных. И в них даже встречаются такие понятия, как искусственный интеллект, машинное обучение и тому подобное. Не поймите нас неправильно, эти книги способствуют созданию осведомленности. Однако они не позволяют глубоко вникнуть в соответствующие темы, вместо этого сосредотачиваясь на высокоуровневом описании конкретной проблемы и ее решения.

На другом конце спектра находятся узкотехнические книги – 500-страничные тома в твердом переплете, пугающие как своим объемом, так и содержанием.

На противоположных сторонах этого спектра сосредото-

чены горы книг, что усугубляет разрыв в общении, – большинство людей читают либо только бизнес-книги, либо только технические книги, а не то и другое.

К счастью, между этими двумя крайностями есть много отличных книг. Нашими любимыми являются следующие:

– «Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking», Фостер Провост и Том Фосетт (Издательство: O'Reilly Media, 2013 год);

– «Много цифр. Анализ больших данных при помощи Excel», Джон Форман (Издательство: Альпина Паблишер, 2016 год).

Мы хотим добавить к этому списку еще одну книгу, которую вы сможете прочитать, не имея под рукой ни компьютера, ни блокнота. Если вам понравится наша книга, мы настоятельно рекомендуем прочитать одну или обе из указанных выше книг, чтобы углубить свое понимание. Вы не пожалеете.

Кроме того, мы очень любим эту тему. Если с помощью своей книги нам удастся побудить вас узнать больше о данных и аналитике, мы будем считать, что достигли успеха.

Что вы узнаете

Эта книга поможет вам построить ментальную модель для понимания науки о данных, статистики и машинного обучения. Ментальная модель – это «упрощенное представление наиболее важных элементов некоторой предметной области, достаточное для решения проблем»[7]. Думайте о ней как о хранилище в вашем мозгу, в которое вы можете поместить информацию.

Некоторые книги и статьи начинаются со списка определений: «Машинное обучение – это…», «Глубокое обучение – это…» и так далее. Чтение технических определений в отсутствие ментальной модели, в которую эту информацию можно было бы вписать, похоже на скупку одежды, которую вам негде хранить. Рано или поздно вся она окажется на свалке.

Однако с помощью ментальной модели вы научитесь понимать, думать и говорить на языке данных. Вы станете главным по данным.

В частности, прочитав эту книгу, вы сможете:

– Думать статистически и понимать, какую роль вариации играют в вашей жизни и процессе принятия решений.

[7] Эта идея обсуждается в чрезвычайно полезной книге Г. Уилсона «Teaching tech together» (CRC Press, 2019).

– Разбираться в данных – разумно говорить и задавать правильные вопросы о статистике и результатах, с которыми сталкиваетесь на рабочем месте.

– Осознавать истинное положение вещей в сфере машинного обучения, текстовой аналитики, глубокого обучения и искусственного интеллекта.

– Избегать распространенных ловушек при работе с данными и их интерпретации.

Как организована эта книга

Главный по данным – это тот, кто способен критически осмыслять данные вне зависимости от своей официальной роли. Это может быть аналитик, сидящий за компьютером, или топ-менеджер, наблюдающий за работой других. В этой книге вам как главному по данным предстоит сыграть разные роли.

Хотя «сюжет» книги выстроен в хронологическом порядке, каждая глава – это отдельный урок, который может быть изучен сам по себе. Однако мы рекомендуем прочитать книгу от начала до конца, чтобы выстроить свою ментальную модель и перейти от основ к более глубокому пониманию.

Книга состоит из четырех частей.

Часть I. Думайте как главный по данным. В этой части вы научитесь мыслить критически и задавать правильные вопросы о проектах по работе с данными, реализуемых в вашей организации; вы узнаете, что такое данные, а также освоите специальную терминологию и научитесь смотреть на мир через призму статистики.

Часть II. Говорите как главный по данным. Главные по данным – активные участники важных обсуждений. Эта часть научит вас «спорить» с данными и задавать правильные вопросы для понимания статистики, с которой вы сталкиваетесь. В ней вы

познакомитесь с основными понятиями статистики и теории вероятностей, необходимыми для понимания и оспаривания предоставляемых вам результатов.

Часть III. Освойте набор инструментов дата-сайентиста. Главные по данным должны понимать фундаментальные концепции, лежащие в основе работы статистических моделей и моделей машинного обучения. В этой части вы получите интуитивное представление о неконтролируемом обучении, регрессии, классификации, текстовой аналитике и глубоком обучении.

Часть IV. Гарантируйте успех. Главные по данным знают о распространенных ошибках, допускаемых при работе с данными. В этой части вы узнаете о технических ловушках, которые приводят к провалу проектов, а также о людях и типах личностей, участвующих в соответствующих проектах. Наконец, мы дадим вам несколько рекомендаций о том, как добиться успеха в качестве главного по данным.

Прежде чем мы начнем

Мы не раз отмечали, что объем данных растет гораздо быстрее, чем наша способность формулировать порождаемые этим проблемы и возможности. Мы показали, что прошлое как всего общества, так и авторов этой книги наполнено неудачами, связанными с данными. И только поняв это прошлое, мы можем понять будущее. Для начала мы познакомили вас с несколькими важными концепциями в примере с классификацией ресторанов.

Для более глубокого понимания данных вам необходимо прорваться сквозь шум, критически осмыслить связанные с данными проблемы и научиться эффективно взаимодействовать с соответствующими специалистами. Мы уверены, что, вооружившись этими знаниями, вы добьетесь успеха.

Готовы? Ваш путь становления главным по данным начинается на следующей странице.

Часть I
Думайте как главный по данным

Многие компании спешат попробовать «что-нибудь новенькое», не останавливаясь для того, чтобы задать правильные бизнес-вопросы, изучить базовую терминологию или научиться смотреть на мир сквозь призму статистики.

У главных по данным не будет такой проблемы. Часть I, «Думайте как главный по данным», подготовит вас к предстоящему пути и поможет сформировать правильный настрой для размышлений о данных и их понимания. Эта часть состоит из следующих глав:

Глава 1
В чем суть проблемы?

«Хорошо сформулированная проблема – это наполовину решенная проблема»
– Чарльз Кеттеринг, изобретатель и инженер

Первый шаг на пути становления главным по данным заключается в том, чтобы помочь своей организации выбрать для решения те проблемы, которые действительно важны.

Это может показаться очевидным, однако многие из вас наверняка были свидетелями того, как компании говорили, насколько замечательные у них данные, а затем преувеличивали их влияние, неправильно интерпретировали результаты или инвестировали в технологии работы с данными, которые не создавали ценности для бизнеса. Часто кажется, что компании запускают проекты по работе с данными просто потому, что им нравится, как это звучит, не вполне понимая важность самих проектов.

Такой подход оборачивается напрасной тратой времени и денег и может породить негативное отношение к будущим проектам. Действительно, стремясь найти скрытую ценность в имеющихся данных, многие компании часто терпят неудачу на самом первом этапе процесса, связанном с определе-

нием стоящей перед бизнесом проблемы[8]. Итак, в этой главе нам предстоит вернуться к началу.

В следующих разделах мы рассмотрим полезные вопросы, которые следует задать главному по данным, чтобы убедиться в важности его работы. Затем мы рассмотрим примеры того, как игнорирование этих вопросов оборачивается провалом проекта. Наконец, мы обсудим некоторые скрытые издержки, связанные с недостатком четкости в исходном определении проблемы.

Вопросы, которые должен задать главный по данным

Мы по опыту знаем, что вернуться к основным принципам и задать фундаментальные вопросы гораздо сложнее, чем кажется на первый взгляд. Каждая компания имеет уникальную культуру, и командная динамика не всегда позволяет открыто задавать вопросы – особенно те, которые могут заставить других почувствовать свою несостоятельность. Многие главные по данным не могут даже начать задавать важные вопросы, способствующие реализации про-

[8] Надежная стратегия работы с данными способна смягчить эти проблемы. Разумеется, важным компонентом любой подобной стратегии является решение значимых проблем, и именно на этом мы сосредоточим внимание в этой главе. Если вы хотите узнать больше о высокоуровневой стратегии работы с данными, обратитесь к книге *Jagare, U.* Data science strategy for dummies. (John Wiley & Sons, 2019).

ектов. Вот почему иметь культуру, которая поощряет постановку таких вопросов, так же важно, как и сами вопросы.

Не существует универсальной формулы, подходящей для всех компаний и главных по данным. Если вы руководитель, мы призываем вас создать открытую среду, позволяющую задавать такие вопросы. (Начните с привлечения к обсуждению технических экспертов.) И задавайте вопросы сами. Это позволит вам продемонстрировать такую ключевую черту лидерства, как смирение, а также побудит других включиться в процесс. Если вы не руководитель, мы все равно рекомендуем вам задавать эти вопросы, не боясь нарушить статус-кво. Наш совет – просто делать все от себя зависящее. Исходя из опыта, мы считаем, что задавание правильных вопросов всегда позволяет получить гораздо больше, чем отказ от этого.

Мы хотим научить вас вовремя замечать предупреждающие знаки и сообщать о возникающих проблемах. Вот пять вопросов, которые вам следует задать, прежде чем приступать к работе с данными:

1. Почему эта проблема важна?
2. Кого затрагивает эта проблема?
3. Что, если у нас нет нужных данных?
4. Когда проект будет завершен?
5. Что, если нам не понравятся результаты?

Давайте подробно разберем каждый из них.

Почему эта проблема важна?

Несмотря на кажущуюся простоту, этот фундаментальный вопрос часто упускают из виду. Зачастую еще до начала реализации проекта мы сосредоточиваем внимание на способах решения проблемы и на потенциальных выгодах от этого. В конце главы мы поговорим об истинных последствиях оставления этот вопроса без ответа. Как минимум этот вопрос позволяет определиться с ожиданиями относительно результатов проекта. Это важно, поскольку проекты по работе с данными требуют затрат времени и сил, а зачастую и дополнительных инвестиций в технологии и данные. Простое определение важности проблемы до запуска проекта поможет повысить эффективность использования ресурсов компании.

Вы можете задать этот вопрос по-разному:

– Что мешает вам (нам) спокойно спать по ночам?
– Почему это важно?
– Эта проблема новая или она уже была решена ранее?
– Какой приз на кону? (Какова отдача от инвестиций?)

Вам нужно понять, как эту проблему видят другие. Это, в свою очередь, поможет понять, как разные люди будут под-

держивать проект и согласятся ли они на его запуск.

Во время первоначальных обсуждений вам следует сосредоточиться на центральной бизнес-проблеме и пристально следить за разговорами о последних технологических тенденциях: они могут легко отвлечь участников от основной темы совещания. Обращайте особое внимание на два предупреждающих знака:

– Фокус на методологии. Это когда компании кажется, будто использование какого-то нового метода анализа данных или технологии даст ей некое преимущество. Вы наверняка сталкивались с маркетинговыми уловками наподобие: «Если вы не используете искусственный интеллект (ИИ), то вы отстаете…» Или когда компания привязывается к какому-то понравившемуся ей модному термину (вроде «анализа настроений»).

– Фокус на конечном результате. Некоторые проекты сбиваются с пути, потому что компании уделяют слишком много внимания тому, каким должен быть конечный результат. Например, они говорят о необходимости создания в рамках проекта интерактивной информационной панели. Вы приступаете к реализации проекта и оказываетесь перед выбором между созданием новой информационной панели и установкой системы бизнес-аналитики. Проектные группы должны быть готовы сделать шаг назад и понять, как именно то, что они собираются создать, принесет пользу органи-

зации.

То, что оба предупреждающих знака касаются технологии, а также то, что ее не следует упоминать на этапе определения проблемы, может показаться неожиданностью или облегчением. На более позднем этапе реализации проекта методологиям и результатам, безусловно, придется уделить внимание. Однако в самом начале проблема должна быть изложена в ясных и понятных каждому терминах. Вот почему мы рекомендуем вам отказаться от технической терминологии и маркетинговой риторики. Начните с описания проблемы, которую требуется решить, а не технологии, которую планируется использовать.

Почему это важно? Дело в том, что проектные команды обычно состоят из тех, кто обожает данные, и тех, кто их боится. Как только в ходе обсуждения проблемы разговор заходит о методах анализа или технологиях, могут произойти две вещи. Люди, которых пугают данные, перестают участвовать в определении бизнес-проблемы. А те, кто их обожает, быстро разбивают проблему на технические подзадачи, которые могут соответствовать или не соответствовать реальной бизнес-цели. После превращения бизнес-проблемы в набор подзадач, связанных с обработкой данных, на обнаружение допущенной ошибки могут уйти недели и даже месяцы, потому что после начала работы над проектом никто не захочет пересматривать формулировку основной проблемы.

По сути, команды должны ответить на вопрос: «Действительно ли это реальная бизнес-проблема, которую необходимо решить, или мы занимаемся анализом данных ради него самого?» Это хороший и прямолинейный вопрос, который следует задавать именно сейчас, когда вокруг науки о данных и смежных областей такой ажиотаж и путаница.

Кого затрагивает эта проблема?

В данном случае важно понять не только то, кого затрагивает проблема, но и то, как может измениться работа соответствующих специалистов в будущем.

Вы должны подумать обо всех уровнях организации (а также о ее клиентах, если таковые имеются). Мы не имеем в виду дата-сайентиста, работающего над проблемой, или команду инженеров, которым придется поддерживать программное обеспечение. Речь идет об установлении конечных пользователей. Зачастую это не только те люди, которые участвуют в определении проблемы. Поэтому очень важно понять, чья повседневная работа будет затронута в случае реализации проекта, и привлечь этих людей *к* его обсуждению.

Мы рекомендуем перечислить имена тех, чья работа изменится в случае решения поставленной проблемы. Если таких людей много, соберите небольшую группу из их представителей. Составьте список этих людей и поймите, как на

них повлияет результат проекта – а затем свяжите полученные ответы с последним вопросом.

Вы можете выполнить пробный запуск решения в рамках мысленного эксперимента. Допустите возможность ответа на вопрос, а затем спросите свою команду:

– Можем ли мы использовать полученный ответ?
– Чья работа от этого изменится?

Разумеется, это предполагает, что у вас есть нужные данные для ответа на вопрос. (Как мы увидим в главе 4, это предположение может оказаться чрезмерно оптимистичным.) Тем не менее вы должны ответить на эти вопросы и рассмотреть несколько сценариев, предполагающих успешное решение проблемы. Во многих случаях ответы на эти вопросы позволяют либо усилить влияние предложенного проекта, либо установить тот факт, что его реализация не предвещает коммерческой выгоды.

Конец ознакомительного фрагмента.