

Алексей Благирев

BIG DATA

ПРОСТЫМ ЯЗЫКОМ



Издательства АСТ
Москва

УДК 004.67
ББК 32.973.233-018
Б68

Благирев, Алексей.

Б68 **Big data простым языком / А. Благирев.** – Москва: Издательство АСТ, 2019. – 256 с.: ил. – (Бизнес-бук).

ISBN 978-5-17-111829-7.

Наш телефон знает о нас больше, чем мы думаем. Он умеет собирать и анализировать информацию о том, как мы передвигаемся по городу, какие посты лайкаем и какими приложениями пользуемся. Он сообщит о пробках и поторопит на работу, чтобы мы не опоздали; подберет музыку под наше настроение и составит список персональных рекомендаций, чем можно занять себя в течение дня. Телефон – больше не устройство, по которому звонят, это уже средство управления окружающим нас миром. Незаметно мы окружили себя такими интерфейсами, которые создают невидимый барьер между человеком и окружающей средой. Планирование, управление, коммуникация, все теперь строится через эти программы и девайсы. Даже человеческие отношения.

Но насколько глубока кроличья нора? Каждому предстоит разобраться в этом самому. Эта книга поможет донести основные принципы проектирования и создания таких интерфейсов управления бизнесом, обществом и окружающим нас миром посредством Больших данных. Читайте, наслаждайтесь и помните: сожжение книг противозаконно.

УДК 004.67
ББК 32.973.233-018

ISBN 978-5-17-111829-7.

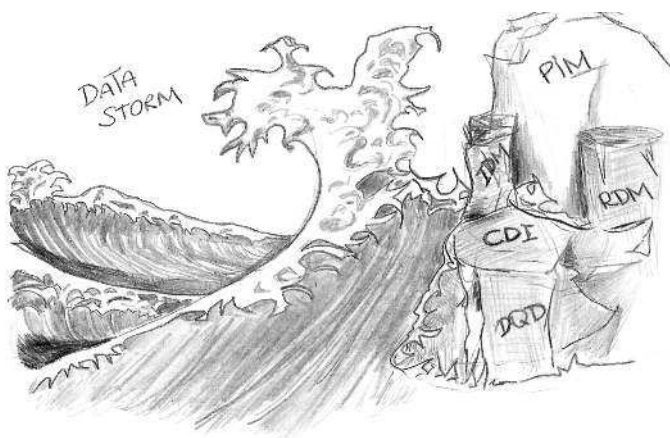
© Благирев А., текст, иллюстрации
© ООО «Издательство АСТ»

ПРЕДИСЛОВИЕ

Люблю людей.

Именно такие мысли остаются в голове, когда тебе предлагают полностью переписать книгу. А если вы читаете это, значит, мне удалось, и я все-таки ее переписал.

Началось все с того, что один мой друг спросил, не знаю ли я людей, которые могли бы простым языком написать про Большие данные. Тогда я сразу представил бесконечное количество писем от издательства, разговоров,



уточнений, переписываний, — всей этой суеты, и первое, что мне хотелось ответить: «Нет, таких разумных существ я не знаю».

Да и смысл писать про Большие данные, если про них уже столько всего написано и рассказано? Вероятность написать что-то умное — минимальна.

И вот я начал писать... Я даже уже представлял себе, как героически заканчиваю эту книгу, становлюсь миллионером и на все деньги с продаж бесконечных тиражей иду погашать ипотеку.

План был гениален, оставалось только его воплотить.

Но, когда я показал плод своих трудов редактору, он сказал, что книга сложна для восприятия, иными словами, подходит только для ботанов. Я честно писал ее с использованием книжной лексики, сложных эвфемизмов, деепричастных оборотов и кропотливо вставлял в текст ссылки на источники, если вдруг упоминал материалы других авторов.

Один раз даже пришлось взять DMBOK, такую специальную «поваренную книгу» с инструкциями и стандартами организации работы с данными. Я перевел из нее целую главу на русский, но мне сказали, что это точно «слишком мощно» для читателя, как и попытка проанализировать существующее регулирование данных.

Итак, в поисках правды, баланса и закрытия личных гештальтов — как сейчас популярно говорить, мне дали книгу «Хулиномика» как пример образцовой книги жанра нон-фикшн.

Когда я взял в руки «Хулиномику», помимо ненормативной лексики в мыслях у меня появились смелые очертания нового эксперимента, поэтому вы держите в руках книгу про Большие данные, изданную под влиянием уникальной простоты и творческой логики изложения.

Мир данных — это компот, из которого трудно отделить то, что нужно знать, а что нет. И вроде бы все интересно, про все можно рассказать, но как понять, что из этого важно, например, учителю физкультуры, который на досуге решил погрузиться в данные?

Задача оказалась сложнее, чем я думал.

Если вы пишете, скажем, про физику, то план изложения поправит научный редактор. А тут — технологии, англицизмы, и людей, знающих ключевые понятия, широту и многогранность Больших данных в издательстве просто-напросто нет.

И я взялся за дело. Сам.

Для начала я решил, что в каждой главе будет два уровня сложности. Первый — для тех, кто собрался почитать про данные, сидя на белом диване в тихой комнате, второй — для тех, чья сфера деятельности связана с данными.

Я написал большую главу про стратегию данных для тех, кто вынужден проектировать стратегию с нуля; попытался разобраться, как данные влияют на корпоративное управление компаниями; показал на ошибках людей, рисующих сложные, малопонятные графики, что формат изложения информации не менее важен, чем сам процесс получения знания.

Конечно, то, что вы держите в руках, — сильно переработанный вариант, но не менее достойный. Наверное.

Сегодня этот компот под названием «мир данных» — уже целая экономика, которая сильно повлияла на все вокруг, включая людей. Теперь нашими данными располагают голосовые помощники, а банки и компании, с которыми мы когда-либо имели дело, все чаще напоминают о себе и требуют внимания. Наш телефон знает, когда мы собираемся на работу, и заранее подгоняет нас к выходу, чтобы мы не опоздали из-за пробок, а когда мы выбираем песню, которую хотим послушать в машине, он выдает нам подходящий плейлист.

Важно знать, что за данные, а точнее за искусственный интеллект, начали активно «топить» в обществе и бизнесе, поднимая проблемы этики их использования.

Просто задумайтесь, вся цифровая среда уже оперирует такими понятиями как «лайки», «репосты», «конверсии». Люди уже обсуждают, как и где подешевле купить трафик себе на сайт, а накруткой подписчиков в Инстаграме не пользуется только ленивый.

Мы оставили позади (в первой версии книги) весь романтизм и большие надежды, поместив в новую версию экспертное мнение по основным блокам работы с данными.

Читайте, наслаждайтесь и помните: сожжение книг противозаконно.

Алексей Благирев

ГЛАВА 1. ЧТО ТАКОЕ BIG DATA?

МАРСИАНСКИЕ ДИАЛЕКТЫ

*О Больших данных, или Big Data
сегодня знают все.*

Или еще нет?
Регулярно данные обсуждаются на сложных конференциях, где популярные компании собирают под своими тентами от дождя пару тысяч молодых людей, размещают роботов и плюшевые пуфики, предлагают даже сыграть в игру с ботом, чтобы посетители могли поучаствовать в машинном обучении. Происходит это примерно так: за ограниченное количество ходов игроку необходимо как можно быстрее спойть девушку-робота.

В общем, кто чем пытается покорить свою аудиторию, рассказывая о работе сервисов

с данными. Вот только ни у кого нет единой картины.

Одни компании говорят про конфиденциальность, другие — про машинное обучение, перечислять можно бесконечно. Есть даже гипотеза о том, что общая картина больше никому не нужна.

«Как это не нужна?» — спросите вы и поспешите на ее поиски.

Выйдете вы из зоны комфорта, пройдётесь по ключевым конференциям, связанным с данными, прочтете статейки известных умных авторов, но все равно толком ничего не соберется вместе.

Чтобы погрузиться в эту тему, надо взять лопату и копать, копать, копать: по кусочкам собирать смыслы, общаться с разными людьми. Администраторы баз данных могут рассказать вам о том, как настраивать кластеры, а ребята, которые копаются в аналитике, помогут разобрать общую логику процесса.

Только вот почему-то каждый эксперт понимает один и тот же термин по-своему. Будто люди строили Вавилонскую башню из данных, чтобы достучаться до небес, а в конце концов все равно заговорили на разных языках, как написано в Ветхом завете. И эти эксперты вкладывают в, казалось бы, обычные слова, какое-то свое понимание, близкое только им.

Конечно, всех бы мог спасти робот-переводчик, который знает тридцать три наречия межпланетных иезуитов. Но, боюсь, пока его функционал не вырос до такого уровня, придется прикидываться

оленеводами, которые впервые слышали о Больших данных. Надо признать, что в некоторых историях мне пришлось разбираться прям с самого что ни на есть нуля, так что расслабьтесь и получайте удовольствие. Будет весело!

А начнем с того, что познакомимся с народом.

#1.

Есть такие важные и бессмертные инженеры по машинному обучению. Задача их проста — проектировать логику и обучать алгоритмы, известные как нейронные сети, заводя в них все новые и новые данные. Если спросить этих инженеров о чем-нибудь другом из области данных, то в большинстве случаев они понятия не будут иметь, о чем их спрашивают — например, кто такие дата-стюарды?

#2.

Дата-стюарды и инженеры качества данных — это такие человечки, которые все правят, чинят и спасают, как Мастер Феликс-младший из игры Fix-It Felix Jr, по ней еще несколько лет назад сняли мультфильм «Ральф». Миссия стюардов и инженеров велика и необъятна. В данных всегда происходит переполох, и нужны те самые brave ребята, которые прибегут со словами «я починю!». Они измеряют искажения в данных и исправляют те самые ошибки, которые допускают пользователи, работая с информацией.

Если спросить у них, в чем роль инженеров по машинному обучению и почему они вообще так называются, то, очень вероятно, что ответа мы не получим. И это нормально.

Разные бригады экспертов занимаются разной работой.

#3.

Архитекторы и аналитики данных — это олицетворение разума. Они опираются на различные правила и методологию, чтобы структурировать данные внутри организации. Например, вместо обозначения таблички «N45» они напишут какое-нибудь гордое «Контрагент» и определят, что в этой табличке должна содержаться информация, касающаяся только контрагента, — например «ИМЯ» / «НАЗВАНИЕ», «ПАСПОРТ» / номер регистрации компании и так далее.

Суть архитекторов и аналитиков — стандартизировать взаимоотношения пользователей с данными и сделать самое главное: навести в этих данных порядок.

Результаты работы этих незаурядных личностей влияют через данные на управление организациями. По-умному их называют data-driven организациями. Они бывают разных типов и устроены все по-разному, но описать data-driven организации или отличить их друг от друга сможет далеко не каждый из описанных специалистов. И это еще один большой вызов.

Разные профессии работы с данными разговаривают на разных языках и формируют собой организации нового типа, где люди не имеют единого представления о том, как ими управлять. Вопрос «чем отличается data-driven организация от data-informed организации?» введет в дичайший ступор не только читателя, но и экспертов, которые работают с данными каждый день.

Перспектива восприятия нового во многом касается наличия практических навыков. Конечно, сегодня мало кто из экспертов имеет руководящий опыт и был тем самым директором по данным, который пытался изменить мир, запуская трансформационные процессы в своей организации для того, чтобы повысить значение использования данных. Это прерогатива людей, которые стоят у руля, а они обычно не разбираются в технике, считая, что она не влияет на принимаемые с точки зрения развития бизнеса решения.

А это все не так. Свойства информационной среды, которые заложены в ней при ее проектировании, оказывают непосредственное влияние на объем и качество принимаемых решений в этой среде.

Когда люди учатся писать на таком языке программирования как Python, им не рассказывают, какие фреймворки проектирования хранилища данных существуют, и что работает, а что уже устарело. Не важно, откуда специалист, интересуется его бизнес или IT, картина везде одна.

Получается, что знание сегментировано, утрировано и преподносится как тайное сокровище, хотя это не так.

Даже разработка на Python проста и похожа на обыкновенную разработку макросов в Excel.

Разбирая управленческие вопросы в организации, в части управления данными, стоит отметить самое важное и, наверное, самое главное. Гештальт, где должно определиться место функции управления данными или так называемого «директора по данным», до сих пор не закрыт и полон споров и противоречий.

IT-сфера активно определяет себя как поставщика данных и, соответственно, хочет играть в них ключевую роль, хотя большинство директоров в IT-сфере понятия не имеют, как правильно проектировать хранилища данных или функцию управления ими. Все ждут постановки от бизнес-подразделений.

Но сейчас ситуация, конечно, намного лучше, чем несколько лет назад, когда бюджеты заливались в бессмысленные проекты, обреченные на смерть еще в пубертатном периоде использования технологии. Тогда пожилые дядечки в возрасте, которые рулили IT-департаментами, с большой долей вероятности были поклонниками Билла Инмона (автора первой книги по созданию хранилища данных) или Ральфа Кимбалла (антагониста Билла). Конечно, согласия между этими концептами мало, и все споры всегда превращаются в дедовские войны на лазерных мечях. Причем, у них разное мнение даже на счет того, как и какими инструментами правильно обрабатывать данные в этих хранилищах.



Например, основной подход — это обрабатывать данные по расписанию, используя специальные инструменты — программы (ETL или ELT) для этой задачи.

Современные эксперты запустили уже свою собственную религию о том, как правильно использовать данные и собирать их в специальную штуку под названием Data Lake. Некоторые из этих экспертов пошли так далеко, что даже отказались от привычных инструментов обработки данных (ETL или ELT), заменив их малопонятной парадигмой, — разбивая все алгоритмы обработки на одинаковые шаги и превращая эти шаги в отдельные программы (сервисы) для создания сложных алгоритмов обработки данных.

Я вам скажу так: все, что можно было когда-либо сделать в Больших данных и машинном обучении — уже сделано. Теперь нужно просто брать существующие методы и сервисы и показывать им новые данные, обучая тем самым алгоритмы адаптироваться.

Перевожу на отечественный. Все, что осталось большинству специалистов — это участвовать в решении только одной задачи, загружать все больше данных для обучения уже существующих алгоритмов. Так ли это? Еще разберемся. Но такие мировые компании как Gartner, уже признают, что роль человека в кооперации с искусственным интеллектом отходит на задний план: необходимо предоставить искусственному интеллекту возможность учиться решать ежедневные задачи. Называется этот подход *Augmented Intelligence*.

В этой книге вместе представлены различные подходы и методы, которые в совокупности с заумной точкой зрения ведут читателя по новым путям работы с данными. Разобщенность терминологии и понятий, собственно, и подтолкнула меня к идее описать практический опыт тех решений, которые можно использовать для получения практического результата. Это должно помочь определить и выявить новые перспективы в работе с данными, чтобы освоить те дальние рубежи экономики, куда еще не проникла цифровизация.

ЧТО ЖЕ ЭТО ВСЕ-ТАКИ ТАКОЕ И ОТКУДА ВЗЯЛОСЬ?

Начну со сложного. Понятие *Big Data* — это такое облако тегов, которое имеет несколько из-

мерений, то есть зависит от ракурса, с которого смотрят.

Пространство координат, благодаря которым можно легко разобраться в том, что такое Big Data, постоянно меняется, создавая отдельные группы понятий, практически не связанных друг с другом. Трудно представить, да?

В Интернете есть известный мем о том, что в одном сперматозоиде содержится 37,5 мегабайт информации ДНК¹. А в результате генерального «салюта» выдается порядка 1500 терабайт.



К слову, в 2013 году мне удалось стать участником крупнейшего внедрения в банковском секторе размером в 51 терабайт. Я внедрял хранилище данных Vertica от Hewlett-Packard. Когда моя команда поместила все транзакции одного крупного банка в это хранилище, у нас получилось немногим больше десяти терабайт. А тут почти в 30 раз больше. В 30!

Так что самые «большие» данные еще впереди.

А теперь просто. Понятие Big Data можно сравнить с термином «инди-рок», который появился в 80-х годах. Так называли стиль, напоминающий гаражный рок или брит-поп, который играли группы в колледжах или университетах. Благодаря журналистам этот термин обрел мно-

¹ По некоторым оценкам используется цифра 760,6 мегабайт для XX-хромосом и 735,9 мегабайт для XY-хромосом, или используется оценка в 400 мегабайт на один сперматозоид, что, в принципе, еще больше.