

УДК 004.6  
ББК 32.972  
Х99

Переводчик: М. Белоголовский  
Редактор: В. Ионов

**Хэнд Д.**

**Х99** Темные данные: Практическое руководство по принятию правильных решений в мире недостающих данных / Дэвид Хэнд ; Пер. с англ. — М. : Альпина Пабlishер, 2021. — 366 с.

ISBN 978-5-9614-4143-7

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

УДК 004.6  
ББК 32.972

*Все права защищены. Никакая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами, включая размещение в сети интернет и в корпоративных сетях, а также запись в память ЭВМ для частного или публичного использования, без письменного разрешения владельца авторских прав. По вопросу организации доступа к электронной библиотеке издательства обращайтесь по адресу [tylib@alpina.ru](mailto:tylib@alpina.ru).*

© 2020 by David J. Hand  
This edition published by arrangement  
with the Science Factory, Louisa Pritchard  
Associates and The Van Lear Agency LLC.  
© Издание на русском языке, перевод, оформление.  
ООО «Альпина Пабlishер», 2021

ISBN 978-5-9614-4143-7 (рус.)  
ISBN 978-0691182377 (англ.)

# СОДЕРЖАНИЕ

Предисловие.....	9
------------------	---

## ЧАСТЬ I ТЕМНЫЕ ДАННЫЕ

### Происхождение и последствия

ГЛАВА 1. Темные данные: Незримая сила, которая формирует наш мир.....	13
Призрак данных.....	13
Так вы думаете, у вас есть все данные?.....	23
Не было ничего необычного, поэтому мы не придали этому значения.....	28
Сила темных данных.....	33
Всюду вокруг нас.....	36
ГЛАВА 2. Обнаружение темных данных:	
Что мы собираем, а что нет.....	41
Темные данные со всех сторон.....	41
Извлечение, отбор и самоотбор данных.....	44
От нескольких ко многим.....	58
Экспериментальные данные.....	72
Остерегайтесь человеческих слабостей.....	84
ГЛАВА 3. Определения и темные данные:	
Что именно вы хотите узнать?.....	89
Ошибки в определениях и измерениях.....	89
Невозможно измерить все.....	98
Скрининг.....	108
Выбор на основе прошлого.....	113

<b>ГЛАВА 4. Непреднамеренные темные данные:</b>	
Видим одно, регистрируем другое .....	117
Общая картина .....	117
Обобщение .....	121
Человеческий фактор .....	123
Недостатки приборов .....	129
Объединение наборов данных .....	132
<b>ГЛАВА 5. Стратегические темные данные:</b>	
Уловки, обратная связь и информационная асимметрия .....	135
Уловки .....	135
Обратная связь .....	144
Информационная асимметрия .....	151
Неблагоприятный отбор и алгоритмы .....	154
<b>ГЛАВА 6. Умышленно затемненные данные:</b>	
Мошенничество и обман .....	164
Мошенничество .....	164
Кража идентификационных данных и интернет-мошенничество .....	169
Мошенничество с личными финансами .....	174
Мошенничество на финансовых рынках и инсайдерская торговля .....	179
Страховое мошенничество .....	184
И прочее .....	190
<b>ГЛАВА 7. Наука и темные данные: Природа познания .....</b>	<b>194</b>
Сущность науки .....	194
Если бы я знал заранее. . . ..	199
Натыкаясь на темные данные .....	210
Темные данные и общая картина .....	212
Соккрытие фактов .....	230
Отзыв публикаций .....	249
Источник и его достоверность: кто вам такое сказал? .....	251

## ЧАСТЬ II

# ОСВЕЩЕНИЕ И ИСПОЛЬЗОВАНИЕ ТЕМНЫХ ДАННЫХ

<b>ГЛАВА 8. Принцип работы с темными данными:</b>	
Проливаем свет.....	259
Надежда!.....	259
Связываем наблюдаемые и недостающие данные.....	261
Определение механизма появления недостающих данных.....	271
Работа с имеющимися данными .....	273
Выход за пределы данных: что, если ты умрешь первым? .....	279
Выход за пределы данных: вменение.....	283
Неправильное число!.....	295
<b>ГЛАВА 9. Полезные темные данные:</b>	
Переосмысление вопроса .....	302
Соккрытие данных.....	302
Соккрытие данных от самих себя:	
рандомизированные контролируемые исследования .....	304
Что могло бы быть.....	306
Репликация данных .....	311
Мнимые данные: байесовское априорное распределение .....	318
Частная жизнь и защита конфиденциальности .....	321
Сбор данных в темноте.....	331
<b>ГЛАВА 10. Классификация темных данных:</b>	
Путь в лабиринте.....	335
Систематика темных данных .....	335
Подведение итогов.....	344
<b>Примечания .....</b>	<b>353</b>