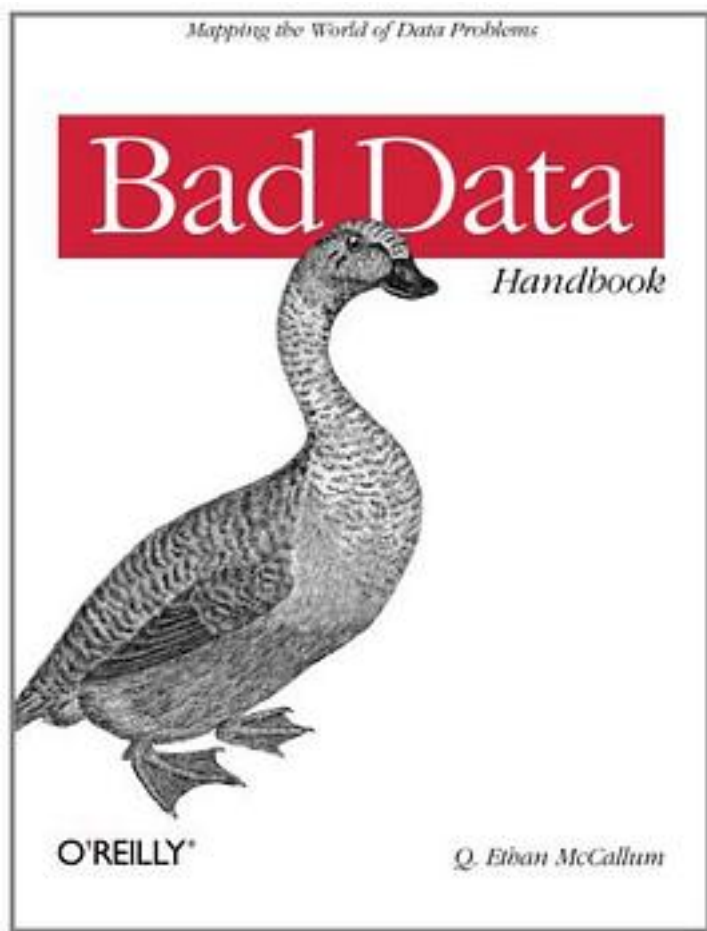


Bad Data Handbook



[Bad Data Handbook_ 下载链接1](#)

著者:Q. Ethan McCallum

出版者:O'Reilly Media

出版时间:2012-11-21

装帧:Paperback

isbn:9781449321888

What is bad data? Some people consider it a technical phenomenon, like missing values or malformed records, but bad data includes a lot more. In this handbook, data expert Q. Ethan McCallum has gathered 19 colleagues from every corner of the data

arena to reveal how they've recovered from nasty data problems.

From cranky storage to poor representation to misguided policy, there are many paths to bad data. Bottom line? Bad data is data that gets in the way. This book explains effective ways to get around it.

Among the many topics covered, you'll discover how to:

Test drive your data to see if it's ready for analysis

Work spreadsheet data into a usable form

Handle encoding problems that lurk in text data

Develop a successful web-scraping effort

Use NLP tools to reveal the real sentiment of online reviews

Address cloud computing issues that can impact your analysis effort

Avoid policies that create data analysis roadblocks

Take a systematic approach to data quality analysis

作者介绍:

Q Ethan McCallum is a consultant, writer, and technology enthusiast, though perhaps not in that order. His work has appeared online on The O'Reilly Network and Java.net, and also in print publications such as C/C++ Users Journal, Doctor Dobbs' Journal, and Linux Magazine. In his professional roles, he helps companies to make smart decisions about data and technology.

目录: Chapter 1 Setting the Pace: What Is Bad Data?

Chapter 2 Is It Just Me, or Does This Data Smell Funny?

Understand the Data Structure

Field Validation

Value Validation

Physical Interpretation of Simple Statistics

Visualization

Keyword PPC Example

Search Referral Example

Recommendation Analysis

Time Series Data

Conclusion

Chapter 3 Data Intended for Human Consumption, Not Machine Consumption

The Data

The Problem: Data Formatted for Human Consumption

The Solution: Writing Code

Postscript

Other Formats

Summary

Chapter 4 Bad Data Lurking in Plain Text

Which Plain Text Encoding?
Guessing Text Encoding
Normalizing Text
Problem: Application-Specific Characters Leaking into Plain Text
Text Processing with Python
Exercises
Chapter 5 (Re)Organizing the Web's Data
Can You Get That?
General Workflow Example
The Real Difficulties
The Dark Side
Conclusion
Chapter 6 Detecting Liars and the Confused in Contradictory Online Reviews
Weotta
Getting Reviews
Sentiment Classification
Polarized Language
Corpus Creation
Training a Classifier
Validating the Classifier
Designing with Data
Lessons Learned
Summary
Resources
Chapter 7 Will the Bad Data Please Stand Up?
Example 1: Defect Reduction in Manufacturing
Example 2: Who's Calling?
Example 3: When "Typical" Does Not Mean "Average"
Lessons Learned
Will This Be on the Test?
Chapter 8 Blood, Sweat, and Urine
A Very Nerdy Body Swap Comedy
How Chemists Make Up Numbers
All Your Database Are Belong to Us
Check, Please
Live Fast, Die Young, and Leave a Good-Looking Corpse Code Repository
Rehab for Chemists (and Other Spreadsheet Abusers)
tl;dr
Chapter 9 When Data and Reality Don't Match
Whose Ticker Is It Anyway?
Splits, Dividends, and Rescaling
Bad Reality
Conclusion
Chapter 10 Subtle Sources of Bias and Error
Imputation Bias: General Issues
Reporting Errors: General Issues
Other Sources of Bias
Conclusions
References
Chapter 11 Don't Let the Perfect Be the Enemy of the Good: Is Bad Data Really Bad?
But First, Let's Reflect on Graduate School ...
Moving On to the Professional World
Moving into Government Work
Government Data Is Very Real

Service Call Data as an Applied Example
Moving Forward
Lessons Learned and Looking Ahead
Chapter 12 When Databases Attack: A Guide for When to Stick to Files
History
Consider Files as Your Datastore
File Concepts
A Web Framework Backed by Files
Reflections
Chapter 13 Crouching Table, Hidden Network
A Relational Cost Allocations Model
The Delicate Sound of a Combinatorial Explosion...
The Hidden Network Emerges
Storing the Graph
Navigating the Graph with Gremlin
Finding Value in Network Properties
Think in Terms of Multiple Data Models and Use the Right Tool for the Job
Acknowledgments
Chapter 14 Myths of Cloud Computing
Introduction to the Cloud
What Is “The Cloud” ?
The Cloud and Big Data
Introducing Fred
At First Everything Is Great
They Put 100% of Their Infrastructure in the Cloud
As Things Grow, They Scale Easily at First
Then Things Start Having Trouble
They Need to Improve Performance
Higher IO Becomes Critical
A Major Regional Outage Causes Massive Downtime
Higher IO Comes with a Cost
Data Sizes Increase
Geo Redundancy Becomes a Priority
Horizontal Scale Isn’t as Easy as They Hoped
Costs Increase Dramatically
Fred’s Follies
Myth 1: Cloud Is a Great Solution for All Infrastructure Components
Myth 2: Cloud Will Save Us Money
Myth 3: Cloud IO Performance Can Be Improved to Acceptable Levels Through
Software RAID
Myth 4: Cloud Computing Makes Horizontal Scaling Easy
Conclusion and Recommendations
Chapter 15 The Dark Side of Data Science
Avoid These Pitfalls
Know Nothing About Thy Data
Thou Shalt Provide Your Data Scientists with a Single Tool for All Tasks
Thou Shalt Analyze for Analysis’ Sake Only
Thou Shalt Compartmentalize Learnings
Thou Shalt Expect Omnipotence from Data Scientists
Final Thoughts
Chapter 16 How to Feed and Care for Your Machine-Learning Experts
Define the Problem
Fake It Before You Make It
Create a Training Set

Pick the Features
Encode the Data
Split Into Training, Test, and Solution Sets
Describe the Problem
Respond to Questions
Integrate the Solutions
Conclusion
Chapter 17 Data Traceability
Why?
Personal Experience
Immutability: Borrowing an Idea from Functional Programming
An Example
Conclusion
Chapter 18 Social Media: Erasable Ink?
Social Media: Whose Data Is This Anyway?
Control
Commercial Resyndication
Expectations Around Communication and Expression
Technical Implications of New End User Expectations
What Does the Industry Do?
What Should End Users Do?
How Do We Work Together?
Chapter 19 Data Quality Analysis Demystified: Knowing When Your Data Is Good Enough
Framework Introduction: The Four Cs of Data Quality Analysis
Complete
Coherent
Correct
aCcountable
Conclusion
• • • • • • ([收起](#))

[Bad Data Handbook_ 下载链接1](#)

标签

数据挖掘

数据分析

数据

计算机

data

统计

Python

O'Reilly

评论

好書啊，早點讀到這書的話，處理數據就不用這麼痛苦了！

技术含量不高，观点虽然多，并不令人“惊奇”

bad data ,bad life

[Bad Data Handbook_ 下载链接1](#)

书评

[Bad Data Handbook_ 下载链接1](#)