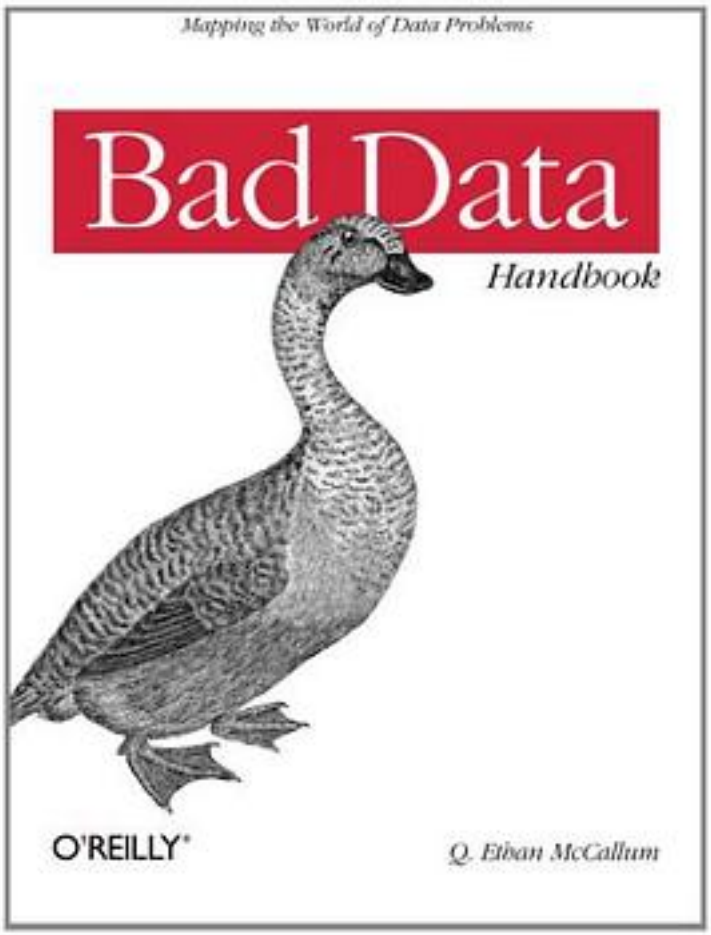
Bad Data Handbook



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 Dobb’ s 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

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_