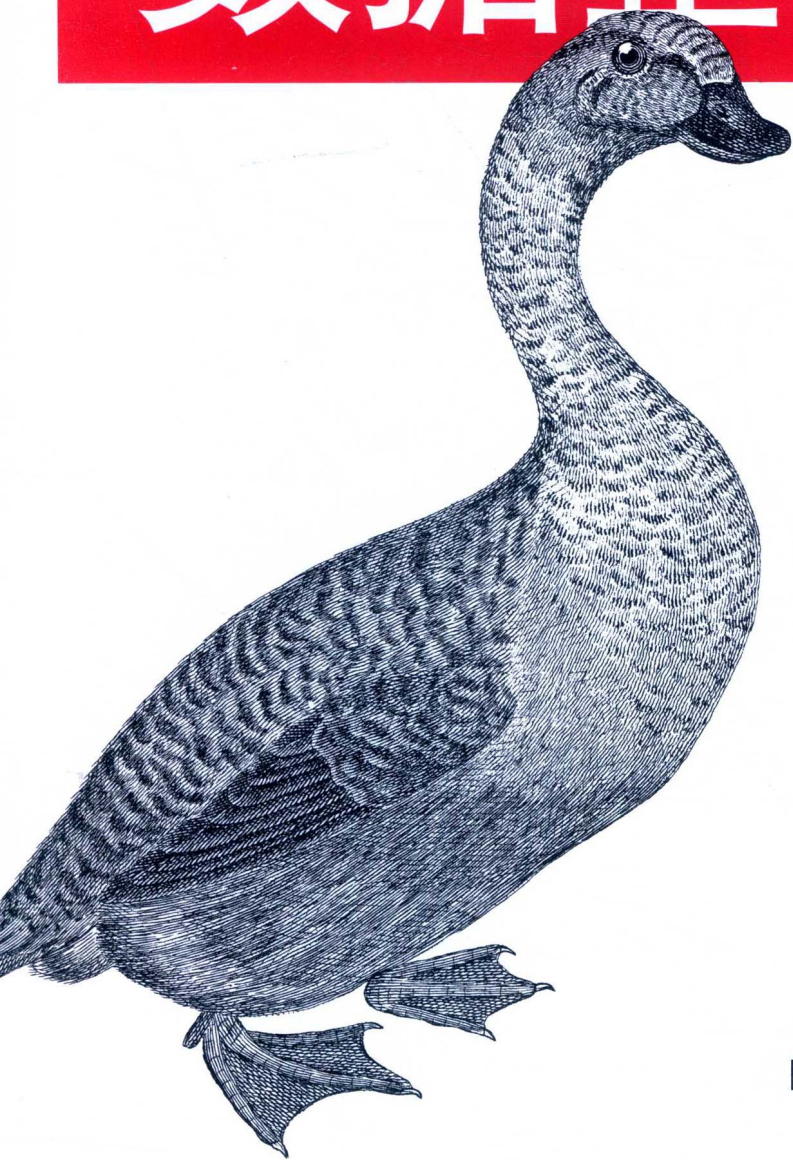


Bad Data Handbook

数据整理

实践指南



[美] Q. Ethan McCallum 著

魏秀丽 李妹芳 译

O'REILLY®



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS

O'REILLY®

数据整理实践指南

[美] Q.Ethan McCallum 著

魏秀丽 李妹芳 译



人民邮电出版社
北京

图书在版编目 (C I P) 数据

数据整理实践指南 / (美) 麦卡伦 (McCallum, Q. E.)
著 ; 魏秀丽, 李妹芳译. — 北京 : 人民邮电出版社,
2016. 3

ISBN 978-7-115-41102-0

I. ①数… II. ①麦… ②魏… ③李… III. ①数据处
理—指南 IV. ①TP274-62

中国版本图书馆CIP数据核字 (2016) 第025032号

版权声明

Copyright© 2013 by O'Reilly Media, Inc.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and Posts & Telecom Press,
2016. Authorized translation of the English edition, 2013 O'Reilly Media, Inc., the owner of all rights
to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

本书中文简体版由 O'Reilly Media, Inc. 授权人民邮电出版社出版。未经出版者书面许可, 对
本书的任何部分不得以任何方式复制或抄袭。

版权所有, 侵权必究。

-
- ◆ 著 [美] Q.Ethan McCallum
译 魏秀丽 李妹芳
责任编辑 陈冀康
责任印制 张佳莹 焦志炜
- ◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 <http://www.ptpress.com.cn>
三河市海波印务有限公司印刷
- ◆ 开本: 787×1000 1/16
印张: 14.25
字数: 261 千字 2016 年 3 月第 1 版
印数: 1-3 000 册 2016 年 3 月河北第 1 次印刷
著作权合同登记号 图字: 01-2013-3682 号
-

定价: 49.00 元

读者服务热线: (010)81055410 印装质量热线: (010)81055316
反盗版热线: (010)81055315

内容提要

随着数据科学的热门，数据的优化、整理以及如何处理不良数据成为人们关注的重点。本书通过处理不良数据，进行数据清理的案例，向读者展示了处理数据的方法。

本书共有 19 章，向读者展示了使用和清理不良数据背后的理论和实践。第 1 部分是 Grubby 的动手实践指南，它向读者介绍了驾驭、提取数据的方法，如何处理文本数据中的数据以及 Web 开发中碰到的数据问题。第 2 部分是让人充满意外的数据，它向读者介绍了数据也会“撒谎”。第 3 部分是方法，它向读者介绍了处理不良数据的一些方法。第 4 部分是数据存储和基础设施，它向读者介绍了如何存储数据。第 5 部分是数据的商业化，它向读者介绍了如何避免数据处理的一些误差。第 6 部分是数据策略，它向读者介绍了如何追踪数据、评估数据质量以及构建数据质量相关平台等。

本书适合数据科学家、数据处理和整理相关开发人员阅读。也适合想要进入数据处理领域的读者阅读。

前言

本书使用的体例



提示

这个图标代表小窍门、建议或一般注解。



警告

这个图标代表警示或警告。

代码范例的使用

本书是帮助你完成工作的。一般说来，你可以在没有我们许可的情况下，在程序和文件里使用本书中的代码，除非你在重复创建代码中极为重要的一部分。比如说，写一段程序要用好几段本书中的代码，不需要经过许可。而销售或发布来自 O'Reilly 书籍中配套的 CD-ROM 是需要经过许可的。通过引用本书内容和代码示例回答问题不需经过许可。但将来自本书的大量代码示例合并到你的产品文档时，必须经过许可。

如果能够标明引用的出处，我们将非常感激，但不做硬性要求。一个出处通常应包含题目、作者、出版社和国际标准图书编号（ISBN, International Standard Book Number）。例如：“Bad Data Handbook, Q. Ethan McCallum (Q'Reilly). Copyright 2013 Q. McCallum, 978-1-449-32188-8.”

如果你认为代码示例的使用已超出正常使用的范围或者超出了上面的许可范围，请使用 permissions@oreilly.com 随时联系我们。

Safari® 图书在线



Safari 图书在线 (www.safaribooksonline.com) 是一种按需数字化图书馆。它提供的专业内容都来自科技和商业领域的世界顶级作者。既有书籍也有视频形式。

技术专业人员、软件开发人员、网页设计师以及商业和创新型人才将 Safari 图书在线作为他们研究、处理问题、学习和认证培训的首选。

Safari 图书在线为各种组织、政府机构和个人提供一系列的产品组合和计价程序。订阅者可以在完全可搜索的数据库中访问成千上万的书籍、培训视频以及出版前的手稿。这些数据库来自诸如 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Focal Press、Cisco Press、FT Press、Apress、Manning、New Riders、McGraw-Hill、Jones & Bartlett、Course Technology 等的很多其他来源。如想了解更多关于 Safari 图书在线的内容，请在线访问。

如何联系我们

请将您对此书的评论及问题邮寄到出版社：

美国：

O'Reilly Media, Inc.

1005 Gravenstein Highway North

Sebastopol, CA 95472

中国：

北京市西城区西直门南大街 2 号成铭大厦 C 座 807 室（100035）

奥莱利技术咨询（北京）有限公司

本书有用来提供勘误表和其他相关信息的网页，网址为 http://oreil.ly/bad_data_handbook。

如有技术问题或对本书进行评论，请发至电子邮箱 bookquestions@oreilly.com。

如想了解关于本书更多的信息、课程、会议和新闻，请访问网站 <http://www.oreilly.com>。

我们在 Facebook 上的网址：<http://facebook.com/oreilly>。

我们在 Twitter 上的网址：<http://twitter.com/oreillymedia>。

我们在 YouTube 上的网址：<http://www.youtube.com/oreillymedia>。

致谢

出版商经常将致谢部分放在一个小角落里，真是很奇怪。这使读者很容易轻视成书

的过程。如果你自己没有写过书或者参与出版，在你知道如何将思路转变成完整的图书时，会感到十分惊讶！

恕我直言，一本书是一件重要的事。要出版一本书意味着将要在一段时间里，有时按月计算有时甚至按年来计算，集中协调很多人。在这里，我希望清楚地阐明我为此书的成功做过贡献的人们的感谢之情。

Mike Loukides: 这一切都始于与 Mike 的一次随意的闲谈。我们的聊天发展成了头脑风暴，形成了一种理念，并最终成就了这本书。（我们要感谢这些意外的碰撞。我想知道如果我和 Mike 在不同的日期、不同的时间谈话，我们是否会有本完全不同的书了呢？）

Meghan Blanchette: 作为本书的编辑，Meghan 使一切事情都秩序井然，保持正常工作。她不知疲倦地为我们提供了源源不断的好主意和反馈信息。一想到《数据整理实践指南》里只有一个主题是在她的监督之下完成的，就给人无比深刻的印象。我期待着下个项目能和她再次合作，不管是什么项目或在什么时候。

编著者和那些帮助我找到他们的人：我同其他 18 位作者共同完成了这本书，他们 18 位编者的部分涵盖了丰富的主题内容和故事。对于他们所付出的时间、努力和对此的适应，特别是他们对待我的反馈时的态度要表示深深的感谢。同时，我也要感谢帮助我联系可能参与的作者的每个人，没有他们，这本书不会有这么长，涵盖的范围没有这么广。

整个 O'Reilly 团队：很高兴在写书的时候有 O'Reilly 团队的支持。整个过程是天衣无缝的：事情就这么成功了，这说明我得开始专注于写作了。感谢大家！

作者简介

作者简介（按出现的先后顺序排列）

Kevin Fink 是一位经验丰富的商业科技经理人。他热衷于将数据转变成商业价值。他除了帮助一些公司成长外（在 WhitePages.com 做过 4 年首席技术官[Chief Technology Officer, CTO]），还曾帮助两家公司上市（1999 年在 N2H2 做 CTO，2011 年在 Demand Media 做业务处理程序[Service Processor, SVP]工程）。另外，他和夫人还在西雅图东部经营了 Traumhof，一个进行马术训练和寄养的马场，作为夫妻共同的财产。在空余时间里，他喜欢远足，喜欢和儿子一起骑双人自行车，喜欢玩寻宝游戏。

Paul Murrell 是位于新西兰的奥克兰大学统计系的一名高级讲师。他的研究领域是统计计算和制图学，也是 R 项目核心开发团队里的成员。他是 *R Graphics* 和 *Introduction to Data Technologies* 的作者，还是美国统计协会的会员。

Josh Levy 是德克萨斯州奥斯丁的一位数据科学家。他从事内容推荐和文本挖掘系统的工作。他在北卡罗来纳大学获得博士学位。在那里他为医学图像分割研究了统计形状模型。他最喜欢的桌上足球射门方向是从守卫区斜飞着进球。

Adam Laiacano 在美国东北大学获得了电气工程理学学士学位。在加入有名的创建于纽约的新公司之前，他曾经花了多年时间为原子钟设计信号检测系统。

Jacob Perkins 是 Weotta 的 CTO，是自然语言工具包（Natural Language Toolkit, NLTK）的贡献者，是 *Python Text Processing with NLTK Cookbook* 的作者。他也创建了 NLTK 的示范产品和应用程序接口（Application Program Interface, API）站点 text-processing.com，并且定期在 streamhacker.com 发表博客。之前，他曾发明了冰箱。

Spencer Burns 是生活在旧金山的一位数据科学家/工程师。在过去的 15 年里他一直在进行着从混乱的数据中提取信息的工作，这些数据的范围从情报到计量金融和社交媒体。

Richard Cotton 是一位有着化学健康与安全背景的数据科学家。他对让非技术型

用户有权使用统计模型的工具有着广泛的研究。他提出 R 包，可以为了确保函数有合理的 API，检查变量和“签名”状态。他经营着一家名为 The Damned Liars 的统计咨询公司。

Philipp K. Janert 在德国出生并长大，他于 1997 年在华盛顿大学获得理论物理学博士学位，并且从那时起就一直从事科技产业。他曾在 Amazon.com 工作 4 年，在那里他发起并主导了多个项目来完善亚马逊的订单处理流程。他写过两本关于数据分析方面的书，其中一本是畅销书 *Data Analysis with Open Source Tools* (O'Reilly, 2010)，而且他的作品也出现在 Perl.com、IBM developerWorks、IEEE Software 以及 Linux 杂志上。他对 Perl 综合典藏网 (Comprehensive Perl Archive Network, CPAN) 和其他开放源代码项目也做过贡献。他现在住在太平洋西北地区。

Jonathan Schwabish 是美国国会预算办公室的经济学家。他进行了关于不平等、移民、退休保障、数据测量、食品券和美国国家政策和其他方面的研究。他的作品在 *Journal of Human Resource* 和 *National Tax Journal* 等报刊出版。他也是数据可视化的创建者，他对各种类型的主题进行设计，从食品券到卫生保健，再到教育。他的可视化作品曾在 visualizaing.org 和 visul.ly 两个网站特载。他也曾针对数据可视化策略和最佳实践在许多政府机构和政策机构发言。他在麦迪逊的威斯康星大学获得经济学学士学位，在美国锡拉丘兹大学获得经济学博士学位。

Brett Goldstein 从 2012 年开始在芝加哥创新科技部做专员。在这之前他是芝加哥的首席数据官。在此期间，他引领这个城市利用数据改善政府为居民服务的方式。在来到市政厅做首席数据官之前，他成立并指挥芝加哥警察局预测分析小组。目的是预测在什么时候什么地点会发生犯罪活动。在进入公共部门之前，他是 OpenTable 的早期雇员，已经帮助公司运营 7 年之久。他在康涅狄格学院获得文学学士学位，在萨福克大学获得刑事司法的理学硕士学位，在芝加哥大学获得计算机科学的理学硕士学位。Brett 目前正在伊利诺伊大学芝加哥校区向他的犯罪学、法律和司法学博士学位努力。他与妻子和 3 个孩子居住在芝加哥。

Bobby Norton 是 Tested Minds 的联合创始人。Tested Minds 是一家以社交学习和快速反馈产品为重点的创新型公司。他已经为多家公司开发软件达十多年之久，比如洛克希德·马丁（美国一家航空航天制造商）、美国国家航空航天局、通用电气的全球研究部、ThoughtWorks（一家全球软件设计与定制领袖企业）、DRW Trading Group（一家自营性交易公司）以及 Aurelius（一个具有长期投资意愿的工业投资公司，总部位于德国慕尼黑）。他的数据科学选择工具包括 Java、Clojure、Ruby、Bash 和 R。Bobby 在佛罗里达州立大学获得计算机科学的理学硕士学位。

Steve Francia 是 10gen（美国一家软件公司，现已更名为 MongoDB Inc.）的首席宣传官。他负责 MongoDB 数据库的用户体验。在去 10gen 之前，他在 Opensky（一家提供技术信息服务的公司）、Portero（美国奢侈品网站）、Tackle（美国为高中运动员服务的网站）和 Supernerd（一家传媒公司）做执行工程。他是一位非常受欢迎的演说家，他所涉及的主题广泛，包括云计算、大数据、电子商务、开发与数据库。他还是作家，曾发表过作品。他是辛迪加博主并且经常向行业刊物投稿。他的作品曾在纽约时报、英国卫报、Mashable（互联网新闻博客）、Read-WriteWeb（读写网）等报纸和网站受到关注。Steve 长期为开放源码做着贡献。他享受用 Vim（一种文本编辑器）编码并坚持流行的 Vim 发布。他与妻子和 4 个孩子在康涅狄格州生活。

Tim McNamara 是一个提着手提电脑，怀揣着美好愿望的新西兰人。他是本地和全球开放数据社区活跃的参与者。他奔走于组织当地集会，从而协助全球危机公关运动。他作为程序员的技术是在帮助开发撒哈拉沙漠灾难管理系统时开始学习的，在帮助运行 One Laptop Per Child XO 的 Sugar Lab 软件时取得了很大的进步。Tim 最近刚刚转向了 escience 领域。在那里他的工作是支持研究社团对技术的理解与吸收。

Marck Vaisman 是一位数据科学家，也是一位顾问、企业家、投资大师和黑客。Marck 是 DataXtract 和 LLC 网站主要的数据学家。他帮助的客户非常多，从新创立的公司到世界 500 强，涵盖了各种各样的数据科学项目。他的工作经验横跨管理咨询、电信、网络、技术产业。他是 Data Community DC（数据专业人和科学家社区）的联合创始人。Data Community DC 是一个通过组织聚会（包括 Data Science DC 和 R Users DC）和其他举措来着力建立华盛顿数据社区并推动数据和统计科学的组织。他在范德堡大学获得 MBA 学位，在波士顿大学获得机械工程的理学学士学位。在他不做与数据相关的事情时，他会与他的家人、朋友一起学习、一起游泳、寻找新鲜有趣的游乐场所亦或是享受美味的啤酒。

Pete Warden 是前苹果软件工程师，为 O'Reilly 写过 *Big Data Glossary* 和 *Data Source Handbook*。创建了开放源代码项目 Data Science Toolkit 和 OpenHeatMap，并且发布了苹果的 iPhone 位置跟踪文件。他是 Jetpac 的 CTO 和创始人。Jetpac 是依照数据处理社交图片的应用商。至今，Jetpac 已经分析了来自 300 万人的超过十亿张照片。

Jud Valeski 是 Gnip 的 CEO 和联合创始人。Gnip 是为企业应用程序提供社交数据的领头羊。从面对产品的客户端消费者到大规模的后台基础设施项目，他已经从事技术工作 20 多年了。他曾经是 IBM、美国网景公司（Netscape）、onebox.com、美

国在线公司（AOL, American On Line）和 me.dium 的工程、产品以及企业并购团队中的一员。他曾在世界上数以千万的人们使用的产品中发挥了核心的作用。

Reid Draper 是一位函数程序员。他对分布式系统、编程语言，还有咖啡感兴趣。目前他正在为 Basho 做分布式数据库：Riak。

Ken Gleason 拥有 20 多年的技术工作经验，包括实时交易系统软件架构和开发以及零售金融服务应用程序设计。在过去的 10 年里，他在电子交易的数据驱动领域工作。在此期间他管理产品研发和高频率的交易策略。Ken 在芝加哥大学布斯商学院获得 MBA 学位，在美国西北大学获得理学学士学位。

Q. Ethan McCallum 是一位专业的服务顾问。他的技术兴趣广泛，涵盖了从数据分析到软件，还有基础设施。他专业专注于通过实际的技术应用，帮助企业在减少风险、增加利润和进行明智的决策方面提高他们的能力。他的文字作品已在线上发表并且出版发行，其中包括 *Parallel R: Data Analysis in the Distributed World* (O'Reilly, 2011)。

Q. Ethan McCallum 是一位顾问、作家和科技爱好者（顺序不论先后）。他的作品已经在 The O'Reilly Network 和 Java.net 上发表，并且发表过 *C/C++ Users Journal*、*Doctor Dobbs's Journal* 和 *Linux Magazine*。他以专业人士的角色帮助企业做出明智的数据和技术决定。

出版说明

本书的封面是一只细嘴雁，叫 Ross's goose (Chen rossii 或 Anser rossii)。它是一种北美的鸟类，得名于一个在加拿大西北地区雷索卢申堡工作的名叫 Bernard R. Ross 的哈德逊湾公司的代理人。还有北方人称其为“galoot”和“scabby-nosed wavey”。关于这种雁是属于 Chen “白”雁类还是属于传统的 Anser “灰”雁类还存在争议。它们的羽毛以白色为主，黑色翅尖，这令人想起雪雁，但比雪雁小 40%。

不论它们属于哪个种类，这种鸟都在加拿大北部和北极中部进行繁殖，然后到远至美国南部、加州中部越冬，有时也去墨西哥北部越冬。在欧洲西部，这些鸟大部分在野生的鸟群里，但是这些逃离的或野生的鸟经常会遇到其它种类的野生雁（如加拿大雁、灰雁、黑雁）。

目录

第 1 章	从头说起：什么是噪音数据	1
第 2 章	是我的问题还是数据的问题	4
2.1	理解数据结构	5
2.2	校验	8
2.2.1	字段校验	8
2.2.2	值校验	9
2.2.3	简单统计的物理解释	10
2.3	可视化	11
2.3.1	关键词竞价排名示例	13
2.3.2	搜索来源示例	18
2.3.3	推荐分析	19
2.3.4	时间序列数据	22
2.4	小结	27
第 3 章	数据是给人看的不是给机器看的	28
3.1	数据	28
3.1.1	问题：数据是给人看的	29
3.1.2	对数据的安排	29
3.1.3	数据分散在多个文件中	32
3.2	解决方案：编写代码	34
3.2.1	从糟糕的数据格式中读取数据	34
3.2.2	从多个文件中读取数据	36
3.3	附言	42
3.4	其他格式	43
3.5	小结	45
第 4 章	纯文本中潜在的噪音数据	46
4.1	使用哪种纯文本编码	46
4.2	猜测文本编码格式	50
4.3	对文本规范化处理	53
4.4	问题：在纯文本中掺入了特定应用字符	55

4.5	通过 Python 处理文本	59
4.6	实践练习题	60
第 5 章	重组 Web 数据	62
5.1	你能获得数据吗	63
5.1.1	一般工作流程示例	64
5.1.2	Robots 协议	65
5.1.3	识别数据组织模式	66
5.1.4	存储离线版本	68
5.1.5	网页抓取信息	69
5.2	真正的困难	73
5.2.1	下载原始内容	73
5.2.2	表单、对话框和新建窗口	73
5.2.3	Flash	74
5.3	不利情况的解决办法	75
5.4	小结	75
第 6 章	检测撒谎者以及相互矛盾网上评论的困惑	76
6.1	Weotta 公司	76
6.2	获得评论	77
6.3	情感分类	77
6.4	极化语言	78
6.5	创建语料库	80
6.6	训练分类器	81
6.7	分类器验证	82
6.8	用数据设计	84
6.9	经验教训	84
6.10	小结	85
6.11	信息资源	86
第 7 章	请噪音数据站出来	87
7.1	实例 1: 在制造业中减少缺陷	87
7.2	实例 2: 谁打来的电话	90
7.3	实例 3: 当“典型”不等于“平均”	92
7.4	经验总结	95
7.5	到工厂参观能成为试验的一部分吗	96

第 8 章 血、汗和尿	97
8.1 书呆子戏剧性工作交换	97
8.2 化学家如何整理数字	98
8.3 数据库都是我们的	99
8.4 仔细检查	102
8.5 生命短暂的漂亮代码库	103
8.6 改变化学家（和其他电子表单滥用者）	104
8.7 传递线（tl）和数据记录器（dr）	105
第 9 章 当数据与现实不匹配	107
9.1 到底是谁的报价机	108
9.2 股票分割、股利和调整	110
9.3 糟糕的现实	112
9.4 小结	114
第 10 章 偏差和误差的来源	115
10.1 估算上的偏差：一般性的问题	117
10.2 报告上的误差：一般性的问题	118
10.3 其他偏差来源	121
10.3.1 顶层编码/底部编码	121
10.3.2 Seam 偏差	122
10.3.3 代理报告	123
10.3.4 样本选择	123
10.4 小结	124
参考文献	124
第 11 章 不要把完美和正确对立起来：噪音数据真是噪音吗	128
11.1 回忆学校生活	128
11.2 向着专业领域前进	129
11.2.1 政府工作	130
11.2.2 政府数据非常真实	131
11.3 应用实例——服务电话	132
11.4 继续前进	133
11.5 经验与未来展望	134
第 12 章 数据库攻击：什么时候使用文件	135
12.1 历史	135

12.2	建立我的工具箱	136
12.3	数据存储——我的路障	136
12.4	将文件作为数据存储器	137
12.4.1	文件简单	138
12.4.2	文件处理一切	138
12.4.3	文件可包含任何数据形式	138
12.4.4	局部数据破坏	139
12.4.5	文件拥有很棒的工具	139
12.4.6	没有安装税	139
12.5	文件的概念	140
12.5.1	编码	140
12.5.2	文本文件	140
12.5.3	二进制数据	140
12.5.4	内存映射文件	140
12.5.5	文件格式	140
12.5.6	分隔符	142
12.6	文件支持的网络框架	143
12.6.1	动机	143
12.6.2	实现	145
12.7	反馈	145
第 13 章	卧库表，隐网络	146
13.1	关系成本分配模型	147
13.2	组合展开微妙的作用	150
13.3	隐藏网络的浮现	151
13.4	存储图表	151
13.5	利用 Gremlin 遍历图表	152
13.6	在网络属性里寻找价值	154
13.7	从多重数据模型角度考虑并使用正确的工具	155
13.8	致谢	155
第 14 章	云计算神话	156
14.1	关于“云”的介绍	156
14.2	何谓“云”	156
14.3	云和大数据	157
14.4	Fred 的故事	157

14.4.1	起初一切都好	157
14.4.2	基础结构全部放在云端	158
14.4.3	随着规模增长, 最初的扩展很轻松	158
14.4.4	麻烦出现了	158
14.4.5	需要提高性能	158
14.4.6	关键要提高 RAID 10 性能	158
14.4.7	重要的局部运行中断引发长期停机	159
14.4.8	有代价的 RAID 10	159
14.4.9	数据规模增大	160
14.4.10	地理冗余成为首选	160
14.4.11	水平扩展并不像想像得那么简单	160
14.4.12	成本显著增长	160
14.5	Fred 的荒唐事	161
14.5.1	神话 1: 云是所有基础设施组件的解决方案	161
14.5.2	神话 2: 云可以节约成本	161
14.5.3	神话 3: 通过 RAID 可以将 cloud 10 的性能提高至可接受的水平	163
14.5.4	神话 4: 云计算使水平扩展轻松	163
14.6	结论和推荐	164
第 15 章	数据科学的阴暗面	165
15.1	避开这些陷阱	165
15.1.1	对数据一无所知	166
15.1.2	应该只为数据科学家提供一种工具来解决所有问题	167
15.1.3	应该为了分析而分析	169
15.1.4	应该学会分享	169
15.1.5	应该期望数据科学家无所不能	170
15.2	数据学家在机构中的位置	170
15.3	最后的想法	171
第 16 章	如何雇佣机器学习专家	172
16.1	确定问题	172
16.2	模型测试	173
16.3	创建训练集	174
16.4	选择特征	175
16.5	数据编码	176
16.6	训练集、测试集和解决方案集	176

16.7	问题描述	177
16.8	回答问题	178
16.9	整合解决方案	178
16.10	小结	179
第 17 章	数据的可追踪性	180
17.1	原因	180
17.2	个人经验	181
17.2.1	快照	181
17.2.2	保存数据源	181
17.2.3	衡量数据源	182
17.2.4	逆向恢复数据	182
17.2.5	分阶段处理数据并保持各阶段的独立性	182
17.2.6	识别根源	183
17.2.7	寻找要完善的区域	183
17.3	不变性：从函数程序设计借来的理念	183
17.4	案例	184
17.4.1	网络爬虫	184
17.4.2	改变	185
17.4.3	聚类	185
17.4.4	普及度	185
17.5	小结	186
第 18 章	社交媒体：是可抹去的印记吗	187
18.1	社交媒体：到底是谁的数据	188
18.2	管控	188
18.3	商业重组	190
18.4	对沟通和表达的期望	190
18.5	新的最终用户期望的技术含义	192
18.6	这个行业是做什么的	194
18.6.1	验证 API	195
18.6.2	更新通知 API	195
18.7	最终用户怎么做	195
18.8	我们怎样一起工作	196