

数据天才：数据科学家修炼之道



[数据天才：数据科学家修炼之道_下载链接1](#)

著者:【美】 Vincent Granville

出版者:电子工业出版社

出版时间:2017-5

装帧:平装

isbn:9787121308833

这是一本跟数据科学和数据科学家有关的“手册”，它还包含传统统计学、编程或计算机科学教科书中所没有的信息。

《数据天才：数据科学家修炼之道》有3个组成部分：一是多层次地讨论数据科学是什么，以及数据科学涉及哪些其他学科；二是数据科学的技术应用层面，包括教程和案例研究；三是给正在从业和有抱负的数据科学家介绍一些职业资源。《数据天才：数据科学家修炼之道》中有很多职业和培训相关资源（如数据集、网络爬虫源代码、数据视频和如何编写API），所以借助《数据天才：数据科学家修炼之道》，你现在就可以开始数据科学实践，并快速地提升你的职业水平。

《数据天才：数据科学家修炼之道》是写给数据科学家和相关专业人士的（如业务分析师、计算机科学家、软件工程师、数据工程师和统计学家），也适合有兴趣转投大数据科学事业的人阅读。

作者介绍:

Vincent Granville博士是一名富有远见的数据科学家，有15年大数据、预测建模、数字分析和业务分析的经验。Vincent在评分技术、欺诈检测和网络流量优化及增长等领域，是举世公认的权威专家。在过去的10年中，他曾与Visa一起研究实时信用卡欺诈检测，与CNET一起研究广告组合优化，与Microsoft（微软公司）一起研究“改变点检测”，与Wells Fargo（富国银行）一起研究在线用户体验，与InfoSpace一起研究搜索智能，与eBay一起研究自动竞价，与各大搜索引擎、广告网络和大型广告客户一起研究点击欺诈检测。Vincent也管理着LinkedIn上最大的“大数据及分析数据科学家”小组，该小组拥有超过100000名成员。

最近，Vincent推出了数据科学中心（Data Science Center）这个大数据、业务分析和数据科学界的领先社区。Vincent曾是剑桥大学和美国国家统计科学学院的博士后。他曾入围沃顿商业计划竞赛和比利时数学奥林匹克的决赛。Vincent已经在统计期刊上发表了40篇论文，并且是许多国际会议的受邀演讲嘉宾。他还开发了一种新的数据挖掘技术，被称为隐性决策树，他还拥有多项专利，是发表数据科学书籍的第一人，并筹集了600万美元的创业启动资金。根据福布斯的排名，Vincent是大数据领域前20位有影响力的人物之一，被VentureBeat、MarketWatch和美国有线新闻网（CNN）专门报道。Vincent的Twitter账号为@Analyticbridge。

关于译者

吴博：利兹大学博士后，具备多年机器学习研发、数据科学从业经验。曾任爱立信大数据高级研究员，多家公司数据科学家及数据变现业务负责人。现任深圳市宜远智能科技有限公司创始人。

张晓峰：哈尔滨工业大学深圳研究生院计算机科学与技术学院副教授、博士生导师，主要研究方向为数据挖掘、隐私保护和机器学习等。曾在北大方正研究院、香港大学电子技术研究所工作。主持包括国家自然科学基金面上项目，以及其他省/市纵向、横向课题十余项。已在国内外重要学术刊物与会议上发表SCI/EI索引论文40余篇。

季春霖：深圳光启高等理工研究院联合创始人，副院长；深圳市统计学会副会长；哈佛大学博士后，杜克大学统计学博士，剑桥大学硕士；广东省自然科学基金杰青项目获得者；发表包括Science在内的论文60余篇，授权专利400余项。热衷于贝叶斯统计学及其应用。

目录: 第1章 数据科学是什么 1

- 真伪数据科学对比 2
- 伪数据科学的两个例子 5
- 新大学的面貌 7
- 数据科学家 10
- 数据科学家与数据工程师 10
- 数据科学家与统计学家 12
- 数据科学家与业务分析师 13
- 13个真实世界情景中的数据科学应用 14
- 情景1：国家对烈性酒销售的垄断结束后，DUI（酒后驾驶）逮捕量减少 15
- 情景2：数据科学与直觉 17
- 情景3：数据故障将数据变成乱码 19
- 情景4：异常空间的回归 21
- 情景5：分析与诱导在提升销量上有何不同价值 22
- 情景6：关于隐藏数据 24
- 情景7：汽油中的铅会导致高犯罪率。真的吗 25
- 情景8：波音787（梦幻客机）问题 26
- 情景9：NLP的7个棘手句子 27
- 情景10：数据科学家决定着我们所吃的食品 28
- 情景11：用较好的相关性增加亚马逊的销售量 30
- 情景12：检测Facebook上的假档案或假“喜欢”数 32
- 情景13：餐厅的分析 33
- 数据科学的历史、开拓者和现代趋势 33
- 统计学将会复兴 34
- 历史与开拓者 36
- 现代的趋势 38
- 最近的问答讨论 40
- 总结 44
- 第2章 大数据的独特性 45
- 两个大数据的问题 45
- 大数据“诅咒” 45
- 数据快速流动问题 50
- 大数据技术示例 56
- 大数据问题是数据科学所面临挑战的缩影 56
- 大规模数据集的聚类和分类 58
- 1亿行的Excel 63
- MapReduce不能做什么 67
- 问题 67
- 3种解决方案 68
- 结论：何时使用MapReduce 69
- 沟通问题 70
- 数据科学：统计学的终结 72
- 8种最差的预测建模技术 72
- 把计算机科学、统计学和行业专业知识结合在一起 74
- 大数据生态系统 78
- 总结 79
- 第3章 成为一名数据科学家 80
- 数据科学家的主要特征 80
- 数据科学家的职能 80
- 横向与纵向数据科学家 83
- 数据科学家的类型 86
- 伪数据科学家 86
- 自学成才的数据科学家 86
- 业余数据科学家 87
- 极限数据科学家 89

- 数据科学家人群特征 90
- 数据科学方面的培训 91
 - 大学课程 91
 - 公司和协会培训项目 95
 - 免费培训项目 96
- 数据科学家职业道路 98
 - 独立顾问 98
 - 创业者 105
- 总结 118
- 第4章 数据科学的技术 (I) 119
 - 新型指标 120
 - 优化数字营销活动的指标 121
 - 欺诈检测的指标 122
 - 选择合适的分析工具 124
 - 分析软件 124
 - 可视化工具 125
 - 实时产品 126
 - 编程语言 128
 - 可视化 128
 - 用R生成数据视频 129
 - 更复杂的视频 133
 - 无模型的统计建模 134
 - 无模型的统计建模是什么 135
 - 该算法是如何工作的 135
 - 源代码生成数据集 137
 - 三类指标：中心性、波动性、颠簸性 137
 - 中心性、波动性和颠簸性之间的关系 138
 - 定义颠簸性 138
 - 在Excel中计算颠簸性 139
 - 使用颠簸系数 141
 - 大数据的统计聚类 141
 - 大数据的相关性和拟合度 143
 - 一系列新的秩相关性 146
 - 渐近分布与归一化 148
 - 计算复杂度 152
 - 计算 $q(n)$ 152
 - 理论上的解决方案 155
 - 结构系数 156
 - 确定簇的数量 157
 - 方法 157
 - 例子 158
 - 网络拓扑映射 159
 - 安全通信：数据加密 163
 - 总结 166
- 第5章 数据科学的技术 (II) 167
 - 数据字典 168
 - 什么是数据字典 168
 - 建立数据字典 169
 - 隐性决策树 169
 - 实现方法 171
 - 示例：互联网流量打分 173
 - 结论 175
 - 与模型无关的置信区间 175
 - 方法 175

- 分析桥第一定理 176
- 应用 177
- 源代码 178
- 随机数 179
- 解决问题的4个办法 181
- 拥有超强直觉能力的业务分析师的直观法 182
- 软件工程师的蒙特卡洛模拟法 182
- 统计学家的统计建模方法 183
- 计算机科学家的大数据方法 183
- 因果关系和相关性 183
- 怎样检测因果关系 184
- 数据科学项目的生命周期 186
- 预测模型的错误 189
- 逻辑相关回归 191
- 变量之间的相互作用 191
- 一阶近似 191
- 二阶近似 193
- 用Excel进行回归分析 195
- 实验设计 196
- 有趣的指标 196
- 把患者分成不同的人群进行治疗 196
- 私人定制的治疗 197
- 分析即服务和应用程序接口 198
- 工作原理 199
- 实施案例 199
- 关键词相关的API的源代码 200
- 其他主题 204
- 当数据库改变时，保存好数值 204
- 优化网络爬虫 205
- 哈希连接 206
- 用于模拟簇的简单源代码 207
- Hadoop和大数据的新型合成方差 208
- Hadoop和MapReduce的介绍 208
- 综合指标 209
- Hadoop、数值的和统计的稳定性 210
- 方差的抽象概念 211
- 一个新的大数据定理 213
- 平移不变性的度量标准 214
- 实现：通信和计算成本 214
- 最终意见 215
- 总结 215
- 第6章 数据科学应用案例研究 217
- 股票市场 217
- 使回报率提高500%的模式 217
- 优化统计交易策略 220
- 股票交易的API：统计模型 222
- 股票交易的API：具体实现 225
- 股票市场模拟 226
- 些许数学知识 229
- 新趋势 231
- 加密 232
- 数据科学应用：隐写术 232
- 好的电子邮件加密 236
- 验证码破解 239

- 欺诈检测 240
 - 点击欺诈 241
 - 连续点击评分与二进制欺诈/非欺诈 242
 - 数学模型与基准 244
 - 虚假转化产生的偏差 245
 - 一些误解 246
 - 统计面临的挑战 246
 - 点击评分优化关键词出价 247
 - 组合优化自动快速的特征选择 249
 - 特征的预测能力：交叉验证 250
 - 勾连检测和僵尸网络的关联规则检测 254
 - 模式检测的极值理论 255
- 数字分析 256
 - 在线广告：到达率和频率的计算公式 256
 - 电子邮件营销：提高300%的性能 257
 - 在7天内优化关键词广告宣传活动 258
 - 自动新闻提要优化 260
 - 用bit-ly进行竞争情报分析 261
 - 测量 Twitter 哈希标签 (hashtag) 的收益 263
 - 用3个修补方法提升谷歌搜索 267
 - 改进相关性的算法 270
 - 广告循环问题 272
- 杂项 273
 - 简单模型会获得更好的销售预测 273
 - 更好的医疗欺诈检测 275
 - 归因模型 276
 - 预测陨石撞击 277
 - 在路口停车场收集数据 281
 - 数据科学的其他应用 282
- 总结 282
- 第7章 踏上你的数据科学职业之路 283
 - 面试问题 283
 - 关于工作经验的问题 283
 - 技术问题 285
 - 一般性问题 286
 - 关于数据科学项目的问题 288
 - 测试你自己的视觉和分析思维 291
 - 通过肉眼的检测模式 292
 - 识别偏差 294
 - 误导性的时间序列和随机游走 295
 - 从统计学家到数据科学家 296
 - 数据科学家也是统计从业人员 297
 - 谁应该给数据科学家教统计学 298
 - 雇佣问题 298
 - 数据科学家与数据架构师密切合作 299
 - 谁应该参与战略思考 299
 - 两种类型的统计学家 300
 - 大数据与取样 301
 - 数据科学家的分类 302
 - 数据科学最流行的技能集合 302
 - LinkedIn上的顶级数据科学家 306
 - 400个数据科学家职位头衔 309
 - 薪酬调查 311
 - 根据技能和位置的薪酬分类 312

- 创建自己的薪酬调查表 316
- 总结 317
- 第8章 数据科学资源 318
- 专业资源 318
- 数据集 318
- 书籍 319
- 会议与组织 322
- 网站 324
- 概念定义 324
- 职业建设资源 327
- 招聘数据科学家的公司 328
- 数据科学招聘广告的样本 329
- 简历样本 329
- 总结 331
- • • • • (收起)

[数据天才：数据科学家修炼之道_下载链接1](#)

标签

数据科学

机器学习

数据挖掘

计算机

计算科学

大数据

data.mining

AI

评论

概括性地描述了数据科学家的方方面面，介绍了不少网络资源，需要的技能。不是算法介绍的书，而是一本成长道路的指导书。从中的收货：自己缺少NLP和数据可视化的技能；具备的能力：大数据、数据分析、机器学习、神经网络。

很有意思的一本书，非常接地气的数据科学家

导引类的工具书，看了差不多三整天。聚类分析/交叉验证/数据字典/决策树/蒙特卡洛模拟

一般

不推荐，有些观点不能接受，翻译得也看不懂。

有的过于追求细节 例子也比较啰嗦 能把数据科学说清楚就够了其他的内容不太需要

对实用派来讲太高深，很多理论可能还没实现，想成为数据科学家的可以阅读

太高深了

高屋建瓴的概括了数据科学家的技能领域、工作内容和职业要求，对希望迈入数据科学行业的人设计自己的学习路径和职业路径有指导意义，推荐一读

在鹅厂工作的学姐推荐的书。读下来的感受是，作为一个门外汉，书中数据科学家的定义完全打破了自己原来的认知。数据科学家更多的是在数据中挖掘到有用的信息，这些信息可能有益于提高收入/避免损失/解释现象等。一些技术确实需要掌握，但更注重想

法。目前在一个整本书通读的状态，很多地方还云里雾里不太清楚，希望能有人指点一二或者一起讨论。还是得二刷。

内容不系统，但对很多具体的问题有细致的讨论，是我个人喜欢的风格。

[数据天才：数据科学家修炼之道_下载链接1](#)

书评

之前在IBM工作时，北美大神们流传过来一份书单，其中包含本书，然后就把其中文版放在购物车中，由于书名我不是很喜欢所以直到最近才下单并且一气读完。内容很好，书名很烂。
本书围绕数据科学家这个新兴职业展开，内容非常庞杂，看得出作者对这个职业的思考是很广博的。干货多...

[数据天才：数据科学家修炼之道_下载链接1](#)