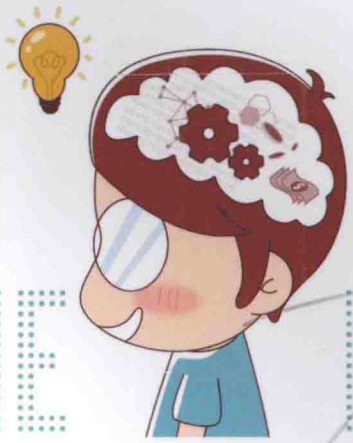
HZ BOOKS
华章IT

白话大数据与机器学习

高扬　卫峥　尹会生◎著
万娟◎插画设计



BIG DATA
MACHINE LEARNING

资深大数据专家多年实战经验总结，拒绝晦涩，开启大数据与机器学习妙趣之旅
以降低学习曲线和阅读难度为宗旨，重点讲解了统计与概率、数据挖掘算法、实际应用案例、数据价值与变现，以及高级拓展技能，清晰勾勒出大数据技术路线与产业蓝图

机械工业出版社
China Machine Press

白话大数据与机器学习

高扬　卫峥　尹会生◎著
万娟◎插画设计

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

白话大数据与机器学习 / 高扬等著 . —北京：机械工业出版社，2016.6（2016.9 重印）

ISBN 978-7-111-53847-9

I. 白… II. 高… III. ①数据处理 ②机器学习 IV. ① TP274 ② TP181

中国版本图书馆 CIP 数据核字（2016）第 115280 号

白话大数据与机器学习

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）

责任编辑：高婧雅　　责任校对：殷　虹

印　　刷：三河市宏图印务有限公司　　版　　次：2016 年 9 月第 1 版第 2 次印刷

开　　本：186mm × 240mm　1/16　　印　　张：21.75（含 0.25 印张彩插）

书　　号：ISBN 978-7-111-53847-9　　定　　价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88379426　88361066　　投稿热线：（010）88379604

购书热线：（010）68326294　88379649　68995259　　读者信箱：hzit@hzbook.com

Preface 前　言

为什么要写这本书

不知从何时开始我们已周身没入大数据时代的潮流，不知不觉被卷入了大数据时代。

无论是每天上网看网页、聊 QQ、聊微信，或者登录银行、网购、买票，或者出行、投宿，甚至是出入任何公众场合、驾车、用水用电……我们无时无刻不在生产着各种数据。而同时我们也在消费着其他人生产的数据，我们使用的众多家电产品，每一个设计细节都融入了设计者对用户体验数据的调查与分析；我们使用的每一部手机、每一台电脑，每一个部件的产出都融入着多得无法想象的指标数据控制下的生产与监控；我们访问的每一个网页、每一个软件，每一次享受到的贴心的产品改动和服务的升级，无不浸透着无数的数据汇集与精细的分析和反馈。这是一场慢慢到来的、贯穿所有产业的革命，这是一次润物细无声的各行业精耕细作的开端。

不管我们是不是愿意，不管我们有没有意识到，我们现在已经身处大数据时代的奇点，而未来要迎接的是大数据奇点爆炸给我们带来的冲击力。我们需要力量来驾驭浪里的航船，我们需要乘风破浪前进的动力。

在这一次远航中，我们不必担心自己的能力水平无法感知数据这种磅礴之力的气魄，不必担心晦涩难懂的公式定理会让我们感到阻力。

请相信我，这是一本通俗易懂的大数据图书，这是一本轻松愉悦的数据挖掘和机器学习的读本，这是一本没有门槛的机器学习实战手册。让我们一起扬帆远航吧！

本书特色

从行为脉络来看，本书基本上是从数据统计、数据指标理解、数据模型、聚类 / 分类与机器学习、数据应用、大数据框架补充知识，以及扩展讨论这样的角度来层层深入完成的。

这种方式会给读者比较好的带入感，让大家——尤其是不擅长数学的读者降低对大数据与机器学习算法的恐惧感。如果读者朋友对排列组合、统计分布这些基础知识比较了解，完全可以考虑跳过这些部分直接去读后面更感兴趣的内容。

为了调节阅读气氛，我们还尝试加入了一些漫画插图。为了让读者朋友能够更快地进行实践，我们几乎在每一个算法讲解后都配有 Python 或者 SQL 语言的实现部分。相信这些能够帮助大家更快、更轻松地阅读本书。

读者对象

（1）对大数据感兴趣但是完全不了解的技术人员。

（2）对机器学习和数据挖掘比较感兴趣的技术人员。

（3）大数据初级从业人员。

如何阅读本书

本书一共分为 18 章。

第 1 章～第 5 章为入门所需基础知识及对数据指标运营的阐述。

第 6 章～第 10 章是对数据挖掘基础知识与算法的介绍。

第 11 章～第 18 章为生产应用与高级扩展。

其中，第 1 章～第 15 章正文内容，以及第 17 章、第 18 章的正文内容由高扬编写。

全书所有的 Python 代码由卫峥编写与补充整理。

第 16 章、附录全部由尹会生编写。

全书所有的漫画插画由万娟创作完成。

勘误和支持

由于水平有限，编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。如果你有更多的宝贵意见，欢迎扫描下方的二维码，关注“奇点大数据”微信公众号和我们进行互动讨论。关注大数据尖端技术发展，关注“奇点大数据”。

同时，你也可以通过邮箱 77232517@qq.com 联系到我，期待能够得到你的真挚反馈，在技术之路上互勉共进。

致谢

特别感谢：万娟女士为本书做的漫画插画内容。

万娟女士现任深圳星盘科技有限公司 UI 设计师，是我在多年工作中遇到过的最敬业的 UI 设计师之一，在 2013 年一起合作的过程中给我留下了非常深刻的印象。

她多次参加全国和国际艺术比赛，曾获得全国青少年绘画大赛铜奖，中国 – 新加坡国际青少年绘画比赛优秀奖，以及全国大学生工业设计大赛三等奖。从小酷爱绘画，理想是开一个属于自己的画室。

她给我留下的最深刻的印象用两个词可以描述：一个词是“敬业”，不管是在过去共事期间的合作，还是在为本书创作插画的过程中，为了保证进度带病坚持创作，都让我非常感动；另一个词是“唯美”，不仅人长得美，作品设计风格也透出现代与时尚的气息。

此外还要对所有支持和关心本书成书的各界朋友表示由衷的感谢：

衷心感谢北京邮电大学软件学院杨谈老师对本书的审校工作。

衷心感谢腾讯公司数据分析师彭瑶女士对本书的审校工作。

衷心感谢重庆工商大学黄辉老师、杨艺老师对本书的大力支持。

衷心感谢机械工业出版社华章公司对本书的支持与帮助。

衷心感谢“奇点大数据”微信群友对本书的关注与支持。

高　扬

目　录 *Contents*

第 1 章 Chapter 1

大数据产业

1.1 大数据产业现状

大数据是近几年来都一直非常火热的一个名词，似乎是伴随着“互联网”的逐渐发展所出现的一个新名词。我们在天天听着“互联网 +”的同时也在听说“大数据 +”。

大数据其实是一个比较抽象和笼统的概念，应该说这个词是为了涵盖性地表达一系列生产和业务行为的一个统称。但是也正是由于这种抽象和过于简略的称谓方式，让每个人都容易对这个词产生见仁见智的不同视角的印象或者看法。

大数据是一个以数据为核心的产业，是一个围绕大数据生命周期不断循环往复的生产过程，同时也是由多种行业分工和协同配合而产生的一个复合性极高的行业。

在我看来，大数据产业生产流程从数据的生命周期的传导和演变上可以分为这样几个部分：**数据收集、数据存储、数据建模、数据分析、数据变现**。

其中每个环节都是非常重要的数据生命环节，每个环节的生产加工行为都是有其价值的，并且每个环节做到极致都可以成就一个伟大的公司。整个完整的产业生态圈就是大数据，它的缩影也渗透在任何一家以数据作为运营基础的公司中。

根据麦肯锡 2011 年发布的一份研究报告，到 2018 年世界范围内将会出现高达 14 万～ 19 万的“大数据”岗位空缺。而艾瑞咨询集团在“2014 年会”上曾指出，全球数据量每 18 个月翻一番，到 2015 年，中国专用数据分析人员预计缺口 1 400 万。

可以看到，在仅仅三四年的时间间隔上，两家咨询公司做出的预测都很大胆，但是两个估算数字相差也确实非常悬殊。究竟哪个数字更贴近“事实”并不好判断，因为大家对“大数据”的概念边界理解可能有很大的偏差，估算出现偏差是必然的，但是有一点可以肯定，大数据人才缺口一定是未来几年非常显著的问题。

2015 年 12 月 21 日，全球第一家大数据交易所——贵阳大数据交易所经过半年多的发展，交易金额已突破 6 000 万元人民币，会员数量超过 300 家，接入贵阳大数据交易所的数据源公司超过 100 家，数据总量超过 10 PB，已发生实际交易的会员超过 70 家。预计在未来 3 ～ 5 年，交易所日交易额将突破 100 亿元[⊖]。

截至目前，中国境内除了贵阳大数据交易所以外，还有长江大数据交易所、武汉东湖大数据交易中心、崇州大数据交易所等十余家大数据交易所挂牌营业。

2016 年 1 月，阿里云的“数加”大数据平台和金山云的 KMR 平台等国内大品牌云产品供应商的重磅产品先后登场，几乎所有有远见的云产品巨擘资本都在向大数据产业链集中。但是这块蛋糕似乎有点太大了，只能边烘焙边分割，谁也没办法一下子全吃掉。

1.2 对大数据产业的理解

“大数据”这个人造词汇其实很容易产生不少误解，尤其是这个“大”字，很容易让人感觉，数据量必须大，而且特别大，越大越能形成产业，也越有价值。其实这真的是“大数据”给人带来的误导。大数据产业的存在其实和其他产业并无二致，本身是为了给其他产业提供服务。

做个假设，假如现在给石油产业冠以“大石油”产业的名字，那么会影响石油行业本身对其他行业的服务样态吗？应该不会。

在“大石油”产业里，同样有人从事着这样的工作内容：**石油勘探、石油开采、石油运输、石油提炼、石油产品销售**等多个细分领域和环节。

最后提供给社会的是由大量人工和智慧凝结在石油产品上的服务，而这些服务极大地方便并满足了社会各领域对于工业能源、建筑材料、食品包装、服装面料、模型器具、日杂用品等多种制造与使用的需求。试想如果没有石油，也就没有廉价汽车与航空动力，尤其是没有乙烯等重要化工原材料的来源，是否存在塑料这样一种廉价的工业制造材料都很难说，那么各个产业则需要用其他造价更为高昂的材料对其进行取代，更不用提家用的天然气和液化石油气了，人们只能再去寻找其他能源：要么不洁净——如柴火和煤炭，要么价格昂贵——如氢气。人们之所以选用石油作为整个产业链的根源，并把它发展成一个完整的产业也是由于这样的原因，大概这个逻辑是比较容易理解的。

类比一下“大数据”产业，数据收集、数据传输、数据存储、数据建模、数据分析、数据交易贯穿了大数据产业的完整产业链。在这个产业链里同样蕴含着和“大石油”一样的东西，这个东西是什么？

数据通过各种软件进行收集，通过网络进行传输，通过云数据中心进行存储，通过数据科学家或者行业专家进行建模和加工，最后数据分析得到的是一种知识，是一种人们通

⊖ 来自《新华网》的报道。