Broadview
www.broadview.com.cn

# 大数据智能核心技术入门

## 从大数据到人工智能

杜圣东 著

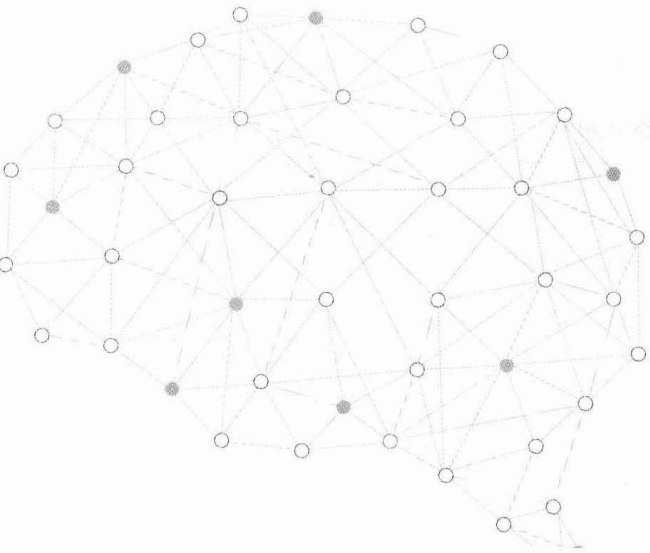

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

# 大数据智能核心技术入门

## 从大数据到人工智能

杜圣东 著





電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

## 内 容 简 介

本书跟从大数据和人工智能应用的融合之路，通过分析和解读整个数据驱动智能核心技术，希望能给读者提供一个大数据智能核心技术体系的入门学习和应用参考指南。本书前半部分内容重在核心技术解读：包括大数据智能的概论、大数据智能核心技术体系的多维解读、深度学习关键技术要点的分析，大数据智能应用三段论和敏捷大数据方法论的提出等内容。后半部分内容重在应用实践的探讨，深入分析了当前大数据智能独角兽 Palantir、AlphaGo、Watson 等核心产品和技术，并从个人学习到工程实践，从企业应用到政府治理，从业务理解到技术选型等多个层面，逐一解读大数据智能技术在学习、应用过程中面临的关键问题、陷阱，并给出参考意见。

本书通过核心技术解读帮助读者学习、理解、应用大数据智能，具有重要的参考价值。本书适合的读者包括关注大数据和人工智能相关技术领域的在校学生、个人学习者和研发工程师、技术主管、企业高管、政府管理人员等。



**图书在版编目（CIP）数据**

大数据智能核心技术入门：从大数据到人工智能 /杜圣东著. —北京：电子工业出版社，2019.4
ISBN 978-7-121-35684-1

Ⅰ. ①大… Ⅱ. ①杜… Ⅲ. ①数据处理②人工智能 Ⅳ. ①TP274②TP18

中国版本图书馆 CIP 数据核字(2018)第 280890 号

责任编辑：石 倩
印　　刷：涿州市京南印刷厂
装　　订：涿州市京南印刷厂
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：720×1000　1/16　印张：15.5　字数：285.2 千字
版　　次：2019 年 4 月第 1 版
印　　次：2019 年 4 月第 1 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

# 自序

我们所生活的世界，就像一片数据混沌（Data Chaos），大数据爆炸式增长并以惊人的速度进行传播，社交网络的实时性打破了数据发布的时空限制，信息流动的速度、广度和深度让传统管理决策模式面临挑战，还有科技高速发展所带来的冲击都在加大未来的不确定性。如何从数据混沌中发现规律，成为预测未来的“先知”，是历代人类的梦想。不管是古人的占卜，前几年的专家系统、数据挖掘、商业智能，还是当下的机器学习、人工智能、深度学习等技术应用，都源于降低未来的不确定性这一本源需求。软件在加速吞噬物理世界，大数据在淹没我们有限的大脑认知，而大部分人对其技术原理和特性却知之甚少。我们寄希望于大数据智能技术这根“救命稻草”，帮助我们面向过去，发现数据规律，归纳已知；面向未来，学习数据趋势，预测未知，从而提升对事物的理解和决策处置能力。然而，要达到这一目标并不容易，人工智能发展 60 年，进展缓慢，直到现在才出现些许曙光。

DT（Data Technology）时代，大数据驱动的人工智能技术生逢其时，从战胜人类顶尖棋手、帮助发现引力波到自动驾驶、精准医疗、安全防控等，就像望远镜改变了我们对宇宙的看法，显微镜改变了我们对微观世界的认知，而通过大数据智能技术来解构我们亲手塑造的数字世界，代表了一种新的认知范式。可以预见的是，随着深度智能技术的高速发展和这一拨“猫”“狗”AI 工程的野蛮生长，人类正在大踏步迈入大数据智能时代。面对兴起的大数据智能热潮，如何应对新兴技术应用带来的挑战？我们又有多少深入的理解？对于技术、架构、算法、伦理、趋势知多少？甲骨文 CEO Larry Ellison（拉里·埃里森）曾说过：“信息科技是唯一能媲美好莱坞的产业，技术明星可能比荧幕明星陨落得更快。”前沿信息技术从来不缺流行词，从 IT 到 DT，从移动互联网到物联网，从云计算到框计算，从数据库到数据湖，从云存储到区块链，从大数据到大数

据智能。当谈及人工智能时，更是这样，有机器智能，还有计算智能；有机器学习，还有深度学习；有感知计算，还有认知计算；有 Watson 还有 AlphaGo；有 TensorFlow 还有 Pytorch。一堆眼花缭乱的技术名词和系统框架，让人云里雾里，内行要全面掌握已是困难重重，外行要摸出门道，可谓难上加难。

面对庞杂的大数据智能科学与技术生态体系，还有眼花缭乱的技术热词，有的是新瓶装旧酒，有的是全新的技术理念和架构，有的是一阵风突然冒出又很快散去，有的是三起两落几十年而不倒……上述种种技术我们如何快速入门，把握重点，理解本质，并有效学习和应用？这是一大挑战！市面上的同领域作品要么偏社科解读，浮于概念介绍；要么偏技术细节，局限于某个技术点。

如何从业务到技术，从学习到实践，从治理到应用，抽丝剥茧、拨开数据驱动智能科学与技术迷雾，洞察大数据和人工智能相关技术生态的全景视图，构建完整的大数据智能知识结构与技术体系？这是一大难题！笔者跟从大数据和人工智能应用的融合之路，通过分析和解读整个数据驱动智能科学技术，希望能给读者提供一个大数据智能核心技术体系的入门学习和应用参考指南。本书写作初衷即基于此，无奈水平有限、涉猎有限、精力有限，难免有遗误的地方，还望同行批评指正。如能为读者提供一点点有益的参考，则心愿足矣。

杜圣东

2018 年 10 月

---

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- 提交勘误：您对书中内容的修改意见可在 *提交勘误* 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- 交流互动：在页面下方 *读者评论* 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：http://www.broadview.com.cn/35684

# 目录

# 第 1 章

# 大数据智能概论

## 1 大数据现象的本质

### 大数据源起：先知的诱惑

大数据（Big Data）[1]时代，我们周围充斥着各种不同的信息、理论、观点，还有“噪声”。软件在加速吞噬物理世界，数字化、信息化、网络化之后正在走向虚拟化、智能化，我们所生活的世界，就像一片信息混沌（Information Chaos）[2]。大数据一直在爆炸式增长，并以惊人的速度进行传播，前沿信息技术高速发展所带来的冲击加大了未来的不确定性。不管是线上还是线下，当我们接收的数据和信息越来越多时，面临的选择就越多，如若不善于过滤、挖掘和分析，进行各种决策时就可能会造成负面影响，从而放大我们对未来不确定性的焦虑。小到个人选择，大到国家决策，都在这样一片混沌中煎熬着。如何

---

1 大数据参考定义：大数据是指一些使用目前现有数据库管理工具或传统数据处理技术很难处理的大型而复杂的数据集。其挑战包括采集、管理、存储、搜索、共享、分析和可视化。整合更大数据集的目的是为了通过数据分析来挖掘出更大价值。

2 混沌：1963 年，美国气象学家爱德华·罗伦兹提出混沌理论（chaos），即非线性系统具有的多样性和多尺度性。混沌理论认为，在混沌系统中，初始条件十分微小的变化，经过不断放大，其未来状态会产生极其巨大的差别。

从信息混沌中发现有价值的规律，成为预测未来的“先知”[1]，抑或是少出几只“黑天鹅”[2]。这是历代人类的梦想，不管是古人的占卜、算命，还是前些年的专家系统、数据挖掘、商业智能，还是当下的机器学习、人工智能、深度学习等技术和应用，都源于我们对未来不确定性的担忧，当然还有应对庞杂信息管理时的失控状态。物理世界正在信息化、软件化，而大部分人对其原理和特性却知之甚少，就像我们的金融交易系统一样，一旦出现黑天鹅事件，系统越复杂，造成的冲击就越大，而机器学习、深度学习等黑箱[3]算法应用正在加剧这一趋势。另外，高速网络和社交软件的实时性打破了数据生成和发布的时空限制，信息流动的速度和广度让传统管理与决策面临挑战。

随着舍恩伯格教授的《大数据时代：生活、工作与思维的大变革》[4]一书的面世，让我们认识到了大数据的重要性。只要抓住大数据这根救命稻草，我们就有机会做“先知”吗？从而更有能力把自己和周遭的信息世界管理得更好吗？在一定程度上讲是这样的，但我们也要知道，任何技术都是把双刃剑，**大数据的可预测性、大数据的迭代性本质和应用闭环特征，创造了一种新的认知范式和管理、决策思维。**但数据分析模型的黑箱和操作的自动化，却削弱了我们对数据问题本身的理解和深度探索能力（在没有大数据工具的条件下），机器的量化分析、智能学习能力与人的主观决策判断在短时间内还难以有机融合。大数据应用为什么难以落地，虽然我们已经不缺模型算法、计算和数据资源，但还是缺乏提出正确问题和有效利用大数据分析工具解决问题的能力，就好比用大炮没有打到蚊子，我们不能说大炮没用，而会说这个人的方法搞错了。

---

1 先知：宗教里的一个概念，一般指对宇宙、人类社会或自然科学方面的大事能提前了解或能准确预言的人。

2 黑天鹅：代表不可预测的重大稀有事件，意料之外却又改变一切。人们总是对一些事物视而不见，并习惯于以有限的生活经验和信念来解释这些意料之外的事件。

3 黑箱算法：常指一些机器学习算法，一般由程序员编写模型，自动从数据中学习出模式和规律，而如何学习这个过程却难以或无法解释。

4 维克托·迈尔-舍恩伯格，肯尼思·库克耶，ViktorMayer-Schonberger,等. 大数据时代:生活、工作与思维的大变革[M]. 浙江人民出版社, 2013.

**案例分析** 《大数据时代：生活、工作与思维的大变革》提出的三大核心观点

《大数据时代：生活、工作与思维的大变革》一书提出了几个颠覆传统认知的核心观点：

一是不需要抽样的样本，而是要全体数据。比如传统的人口调查或产品品牌评价分析，多是基于小数据抽样的统计方法，因为当时条件下很难有获取大数据的渠道，而且如果要做全样本分析，则人力、物力的投入极其巨大。大数据时代，基于移动互联网、物联网、社交网络等技术，在一定程度上解决了全体数据源采集的问题，但也不能一概而论，很多情况下要获取全体数据是不现实的，需要抽样小数据进行辅助分析。

二是大数据并不精确，而是混杂的。这是相对于大数据的多源异构性特点来讲的，全体数据一定是来源于多个渠道的，数据格式多样化，再加上大规模的量级，大数据集中包含“噪声”、错误或偏差数据项都很正常，这就需要大数据处理技术能包容这一问题。基于大数据的简单模型预测往往比基于小数据的复杂模型预测更有效，这个观点首次由 Google 提出，深度学习的广泛应用验证了这一观点。

三是大数据分析要解决的关键问题不是因果关系，而是重在相关性分析。这个说法存在较大争议，理想的大数据技术不应该只解决相关性分析，还要能解决因果的推理，只是相关性分析相对更容易实现。

上述观点的提出可以说是对统计学时代的传统分析方法提出了质疑，大数据时代 DT 技术的变革，其核心理论基础多是基于上述几点。

## 从《琅琊榜》看大数据本质

从前些年的物联网、云计算到现在的大数据和人工智能，为什么这些信息技术能够兴起并备受各方关注？大数据现象的本质是什么？怎么认识和理解大

数据智能？笔者不想再向大家啰唆 4V 还是 5V[1]，而是来谈谈一部武侠剧《琅琊榜》。为什么叫《琅琊榜》，因为有一个高端神秘的大数据公司——琅琊阁，每年都会发布武林高手排行榜，并为各方提供及时的情报服务。

要理解大数据技术，我们可以分析一下琅琊阁的这些榜单到底是怎么排出来的，为什么琅琊阁的情报服务让皇家也趋之若鹜。我们都知道，现代的各种排行榜，都是以海量数据作为基础进行深度分析的。影片开头青山绿水之间的琅琊阁地宫就是座海量“大数据中心”（分布式存储，见图 1-1），江左盟广布天下的分站和盟员就是数据采集端（手机 APP、网站、传感器等设备终端），而飞鸽传书就是古时候的高速信息传输通道（物联网、移动互联网），当然琅琊阁还有帮隐秘的数据科学家（负责数据挖掘分析和智能预测建模），所以才能成就广为人知的麒麟之才——梅长苏，“麒麟才子，得之可得天下”的关键不在于梅长苏个人，而在于他背后的神秘大数据公司琅琊阁。

**图 1-1　琅琊阁的“大数据中心”**

不管是书家笔下的军师诸葛亮、刘伯温，还是抗战轶事中的林彪将军，都是善于收集和分析情报的数据科学家。只要掌握足够的数据和信息，就能对事物的本质，对时局和对手有深入的认识，足不出户而知天下事。大数据时代更是这样，我们每个人的一切都在加速数字化，吃穿住行用，甚至我们的身体和思想本身在各大 IT 巨头的数据中心里都能找到对应的数字副本，只要能集中分析这些大数据，就能从多个层面解码任何一个人。在万物互联和数字化、网络

1　4V、5V：一般指 Volume，数据总量大；Variety，多源异构——数据种类和来源多样化；Value，数据价值密度相对较低；Velocity，数据增长速度快，处理速度快，时效性要求高。5V 就是在 4V 的基础上加上 Veracity，强调数据质量，如准确性、可信度。