数据即未来

[美] 布瑞恩·戈德西（Brian Godsey）著
陈斌 译

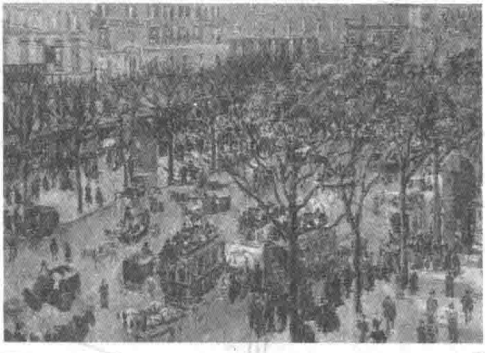
大数据王者之道

THINK LIKE A DATA SCIENTIST

Tackle the Data Science Process Step-by-Step

余晓芒 涂子沛 唐彬 张瑞海 向江旭 逄伟 郭大刚
| 联袂力荐 |

机械工业出版社
China Machine Press



数据即未来

[美] 布瑞恩·戈德西（Brian Godsey）著
陈斌 译

大数据王者之道

THINK LIKE A DATA SCIENTIST

Tackle the Data Science Process Step-by-Step

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数据即未来：大数据王者之道 /（美）布瑞恩·戈德西（Brian Godsey）著；陈斌译．
—北京：机械工业出版社，2018.1
书名原文：Think Like a Data Scientist：Tackle the Data Science Process Step-by-Step

ISBN 978-7-111-58926-6

I. 数… II. ① 布… ② 陈… III. 数据处理 IV. TP274

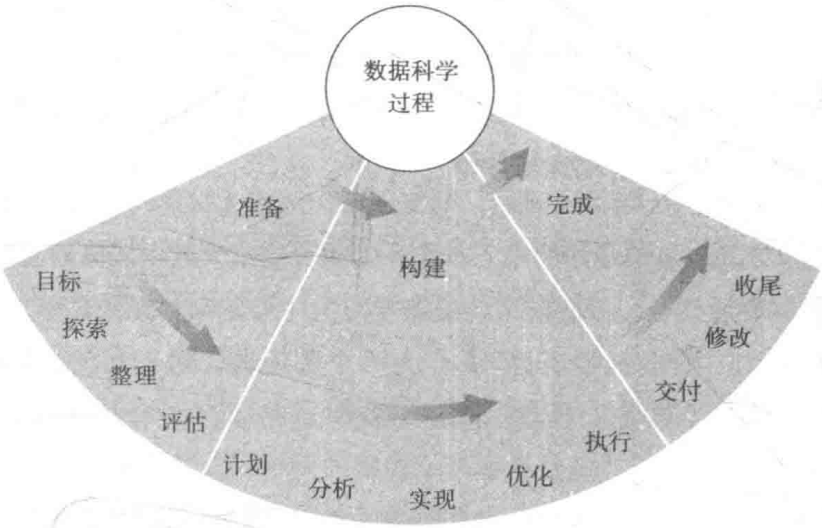
中国版本图书馆 CIP 数据核字（2018）第 008611 号

数据即未来：大数据王者之道

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：张志铭
责任校对：李秋荣
印　　刷：中国电影出版社印刷厂
版　　次：2018 年 3 月第 1 版第 1 次印刷
开　　本：147mm × 210mm 1/32
印　　张：13.75
书　　号：ISBN 978-7-111-58926-6
定　　价：79.00 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

数据科学项目的生命周期



本书围绕着数据科学项目的三个阶段组织：

- 第一阶段是准备，把时间和精力花在项目初期的信息收集上，以避免事后的麻烦。
- 第二阶段是构建，利用在准备阶段所获得的信息，以及统计和软件可提供的所有工具构建产品，把计划付诸行动。
- 第三阶段是完成，交付产品、获得反馈、进行修改、支持产品和结束项目。

•• 本书赞誉 ••

陈斌先生从硅谷回国后，笔耕不辍，这本《数据即未来：大数据王者之道》是他翻译的第三本作品。随着国家大数据战略的推进，以数据联通整合、分析应用、机器学习为中心的项目越来越多，本书称之为“数据科学项目”。作为一个管理者，对于如何成功地准备、组织、规划、构建、实施、交付这些项目，本书提供了很多见解。

——涂子沛，阿里数据副总裁，《大数据》和《数据之巅》作者

宇宙万物不断演变，数据记载了万物变化的过程。数据工程为我们搜集、存储和管理数据奠定了基础，数据科学为我们探索数据世界的未知提供了思考和研究的框架。深刻领悟《数据即未来：大数据王者之道》书里所论述的数据科学探索过程、方法和理论，将有助于您深刻掌握数据世界发展变化的规律。

——张瑞海，北京百悟科技有限公司董事长

人工智能的核心是数据，如何准备、构建和交付高质量的数据产品至关重要，愿这本书成为大数据、人工智能学习者和从业者的

良师益友！

——向江旭，苏宁云商 IT 总部执行副总裁，苏宁技术研究院院长

我们正处在一个新的时代，这个时代里数据是最新的燃料，数据和人工智能正在影响人类生活的方方面面。不只是数据科学家才要懂数据，每一个人，每一种职业，都需要一定的数据思维能力，把数据变成助推自己工作和生活的燃料。本书可以帮助读者掌握数据相关的基础知识，培养初步的数据思维，是一本非常好的入门书！

——逄伟，携程旅行网 CDO 首席数据官

以近 20 万字的内容讲述数据科学这个类似方法论的话题，确实是一个非常具有挑战性的任务。在数据库架构与管理、大数据工程、机器学习、高性能计算、AI 等工程能力全面改善的今天，我们依然需要解决一个问题，那就是如何高效、正确地使用这些能力，在实践中更有效地解决问题呢？技术本身是有效能边界的，技术带来的结果也有不确定性，甚至会导致某些风险。如何解决这些新的问题，规避这些风险和冲突呢？比如，在数据日益资源化的今天，如何提高数据提取出有效信息的效率，通过有效信息累积构架知识体系，并且能够使知识体系更加有效地收敛和自洽呢？数据科学正着手解决这些问题。这已经超出了技术本身，更接近于哲学的范畴了。本书细致入微地讨论了数据科学解决问题的过程，始终聚焦在数据科学项目中所特有的概念和挑战，不是停留在解释如何使用各种最新的工具和技术的炫技层次，而是组织与利用现有资源和信息实现项目目标的过程，为数据科学的过程引领导航。

——郭大刚，北京市互联网金融行业协会秘书长

•• 中文版序一 ••

大数据与太极相生新科学范式

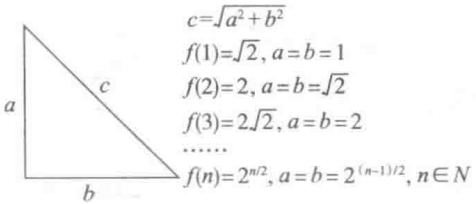
正当我国大力推动实施大数据战略、加快建设数字中国之际，《数据即未来：大数据王者之道》一书的出版发行，对于培育造就一批大数据领军企业，进行大数据人才队伍培养的前瞻性布局，具有十分积极的意义。

本书作者布瑞恩提出“数据科学家专注于创建依赖于数据和结果的概率陈述系统”；“软件研发人员和数据科学家的系统，可以分别与数学概念的‘逻辑’和概率类比”；“‘若 A 则可能 B’这样的概率语句并没有那么简单”等思想以及书中所介绍的具体的数据科学方法，颠覆了预先假设，用过去推断将来的重复性实验，以及具有可逆性、确定性的传统科学范式。

本书强调了“数据科学是指导数据项目开展和决策的一系列过程和概念”；大数据项目的路线图重要的是“面向过程、与客户互动提出问题”。当今，社会与科学技术飞速发展，突出了对一切事物发展的数据分析都要重视时间思维，聚焦到事物发展的关键时空

点上。本书以科学的流程分析，展开大数据项目应该如何实现产品或服务的再造。

布瑞恩的上述思想与中国的太极和道德经一表一里互补的哲学思想有着很大的相似性。道德经阐述“道生一，一生二，二生三，三生万物”。道生一，一是阴；一生二，二是不一的阳；二生三，三是一不一的和，是太极；三生万物，是指一切存在物都是由阴、阳、和三态构成的，“一不一同一”的机理生成了宇宙的一切。1964 年美国科学家盖尔曼提出中子、质子这一类强子是由三个更基本的单元夸克构成的，验证了道德经中《三生万物》的物理存在原理。



本书阐明了软件与软件研发人员、数据与数据科学家、专业知识与专业人员对应的基本三要素之间的关系。起主导作用的数据科学家用大数据在概率世界中探索、研究客观对象得到相对确定的整体结构的序参数和发展态势，从而更准确地把握了事物的发展规

律。参看本文插图中勾股定理的两次迭代运算，可以深刻地理解大数据是太极或和，它能够不断发现或发明新的算法或模式，将人类从阳的已知“有理世界”带入一个阴的未知“无理世界”，显然，这是科学方法论的一次伟大的变革。

道德经说“道可道也，非恒道也”。宇宙的基本规律是不规律，是规律与不规律的同一。热力学第二定律的熵增表述的是有序从无序中产生；而牛顿第三定律相互作用表明一切存在只要因相互作用而链接，它们就共存于同一个世界，存在着同一共性即“一不一同一”。以往的科学范式是基于牛顿机械力学的二元逻辑，大数据科学方法论是基于阴、阳、和（太极）的三元逻辑，这同样也是对科学范式的一项重要创新。

从 2016 年 6 月至今，本书译者陈斌为大力推广大数据，不仅在全国多处留下了他宣传演讲大数据的足迹，而且还在此期间完成了《架构即未来》《架构真经》《数据即未来：大数据王者之道》三本译作。他这种艰苦奋斗、坚持不懈的奉献精神感人至深、令人钦佩。

余晓芒，中国联通前副总裁，中国信息大学校长

•• 中文版序二 ••

大数据在我们这个时代的重要性愈渐凸显，大数据战略也已然高屋建瓴地提到了国家层面。2017 年 12 月，中共中央总书记习近平在主持就实施国家大数据战略第二次集体学习时发表了讲话，习总书记强调，要坚持数据开放、市场主导，以数据为纽带促进产学研深度融合，形成数据驱动型创新体系和发展模式，培育造就一批大数据领军企业，打造多层次、多类型的大数据人才队伍。

身处滚滚洪流奔腾向前的商业时代，我们切身地感受到大数据发展的欣欣向荣。不过，恰如这本《数据即未来：大数据王者之道》所说，数据科学仍然是一个新领域，虽然它的组成部分：比如统计学、软件研发、基于证据解决问题等，都是我们非常熟悉的传统领域，但数据科学却不能简单地等同于数据库架构与管理、大数据工程、机器学习或者高性能计算。作者布瑞恩·戈德西认为，数据科学的核心在于数据内容之间的相互作用，给定项目的目标以及实现这些目标的数据分析方法。所以本书并不是要老生常谈地讲述数据库软件或者统计学方法，尽管在谈到数据科学时会不可避免地提及

这些内容，但它的核心仍在于将科学方法应用于数据集从而实现项目的目标。

所以，本书并不是一本简单的操作手册，而是让你可以按图索骥地学会数据科学。确切地说，它更适合中高层的数据科学家，帮助他们明晰要解决的问题，并找到问题解决的策略。本书不会过多纠结于细节的战术，而是更注重思维方式的梳理，以及对数据科学的深刻洞察。这就是译者易宝 CTO 陈斌把它翻译为《数据即未来：大数据王者之道》的原因，这个名字昭示了数据科学的重要性，以及本书的读者在数据之路上不断追求，力争上游的雄心。

确实，无论作为国家的发展，还是作为企业的生存之路，我们都要养成以解决问题为导向的好习惯。就易宝支付而言，成立至今已经有 15 年的历史。支付离不开货币，在这 15 年中，由于移动互联网的全面到来，联接已无处不在，数据正成为新时代的货币，因此支付公司也必须升级为经营数据的公司。在此过程中，我们发现，真正优秀的技术人才都是以解决问题为导向的，而不是沉迷于技术本身，因此，我们不仅强调客户，还强调了“内部客户”，以及基于客户的数据驱动，以推动技术团队更积极地站到对方的角度思考，即他们交付的成果能不能为别人解决实际问题。值得高兴的是，正是在以本书译者 CTO 陈斌为代表的一群骨干的努力下，整个技术团队越来越有解决问题意识，越来越有紧贴业务感觉，共同推动问题的解决，促使易宝支付取得了今天的成绩。我相信，这也是中国无数野蛮生长起来的企业的缩影。

数据科学是一门日新月异的科学，数据库常变，软件常变，硬

件常变……不变的只有洞察本质的思维方式和对问题解决之道的不懈追求，因此，期待这本《数据即未来：大数据王者之道》能给数据工作者深度启发，在大数据的路上愈走愈远，拥抱大数据时代，拥抱未来！

唐彬，易宝支付 CEO

•• 译者序 ••

汹涌的数字瀑布闪烁着神秘的光彩，密密麻麻地排满了整个屏幕，作为影史经典之作《黑客帝国》的片头，这一幕早已深入人心。而正如这一片头所显示的，今天的世界已然变成了一个数据的世界。阿里研究院甚至提出了从 IT（信息科技）转向 DT（数据科学）的战略方向。大数据（Big Data）也和人工智能（AI）、云计算（Cloud Computing）、区块链（DLT，分布式记账技术）合称为了 ABCD 四大新锐技术。

为什么会产生数据科学呢？首先，随着社会的发展，人类的社会实践、生产实践和科学实验产生了大量的数据。近年来，由于移动互联网的快速发展，数据产生的速度也随之激增。技术的进步，也使得数据的记录和整理变得越来越便利。在这一背景下，数据的海量增加使得人们对于数据采集、清洗、过滤、分析、建模和表达的需求也越来越殷切。人们的聚焦点也从如何生产、收集和管理数据，转向如何更好地建立模型和分析数据。由此，数据科学应运而生。

其实，如今在互联网行业里，也有很多从事与数据相关工作

的人，包括最基础的数据库管理员（DBA）、维护大数据技术基础（Hadoop/Spark）的系统管理员、研发分布式数据处理程序的程序员、从事数据结构分析与管理的数据架构师、聚焦数据建模的工程师以及负责以可视化手段展示数据的工程师等等。虽然这些人的工作都与数据相关，其中有些人是数据的搬运工，有些人是数据的处理工，有些人是数据库的管理员，但是他们都不能称为数据领域的王者。这就像铁匠每天都在与铁打交道，但是我们从来不把铁匠称为金属学家；农民每天都在和土地打交道，但是我们从来不把农民称为土壤学家，我们每个人天天都在做各种计算，但是我们从来不把自己称为数学家。

那么，在数据的王国里，究竟谁是数据之王？我认为只有那些真正掌握数据科学项目的过程，知道如何探索数据、深入分析数据、用数据解决现实中问题的人才是数据世界里真正的王者，即数据科学家。

那么，如何从搬砖的数据民工变成一个指点江山的数据科学家呢？这需要行业的积淀，个人的努力，还有科学的指导。

本书作者布瑞恩·戈德西结合自己的亲身经历，讲述了数据科学中从项目准备、解决方案构建到项目交付的全部过程，系统地论述了数据科学的完整过程。特别是作者结合自己的成长过程以及工作经历，以案例的形式深入浅出地讲解了在开展数据科学项目的过程中可能遇到的各种问题，使本书成为有志于从事数据科学相关工作的初学者的极佳入门指南，并且对已经拥有数据科学项目经验的人来说，本书也非常实用和有借鉴价值。

数据科学作为一门独立的科学仅仅是近两三年的事情，因此，

这个领域是神秘的，令人向往的，这里充满了荆棘，也蕴含着无数的机会，需要大批有志从事数据科学探索的人加入其中。如果你也想了解数据科学，走进数据科学，甚至成为该领域的王者，那么本书将是你最好的敲门砖。

陈斌

2017 年 11 月

前 言

2012年，《哈佛商业评论》中的一篇文章将数据科学家誉为“21世纪最性感的职业”。公平地说，在本世纪剩下的87年里，这个说法可能会有所改变。虽然现在数据科学家的确得到了很多关注，介绍数据科学的书籍也正在激增。但是，仅仅把从别处能够找到的文字重复一下或者将其重新包装成另一本书毫无意义。在研究了数据科学的新文献后，我发现大多数作者愿意解释如何使用各种最新的工具和技术，却不愿意详细地讨论数据科学中解决问题的过程。有抱负的数据科学家在熟读了几本书并掌握了最新算法和数据存储知识后，仍然会问同一个问题：应该从哪里开始？

所以，虽然这也是一本介绍数据科学的书，但本书试图引导读者通过存在很多歧路、陷阱并且目的地未知的数据科学之路，对可能发生的意外提出警告，让读者做好准备，并给出如何应对意外的建议。虽然本书将会讨论哪些工具可能最有用及其原因，但主要目标始终是为学习数据科学的过程引路导航，以便在现实生活中智慧、高效、成功地找到以数据为中心的问题的实际解决方案。

•• 致　谢 ••

感谢 Manning 出版社中每一位曾帮助我让本书成为现实的人，也感谢 Manning 出版人 Marjan Bace 给我这个机会。

我还要感谢 Mike Shepard 对本书技术方面所做的评估，感谢在手稿写作过程中提供了有益反馈的审稿人。这些人包括 Casimir Saternos、Clemens Baader、David Krief、Gavin Whyte、Ian Stirk、Jenice Tom、Łukasz Bonenberg、Martin Perry、Nicolas Boulet-Lavoie、Pouria Amirian、Ran Volkovich、Shobha Iyer 和 Valmiky Arquissandas。

最后，我要特别感谢在 Unoceros 和 Panopticon 实验室的现任及前任队友，他们以多种形式为本书提供了充足的素材，包括：软件研发和数据科学的经验与知识、富有成果的对话、疯狂的想法、有趣的故事、尴尬的错误，最重要的是愿意满足我的好奇心。