HZ BOOKS
华章IT

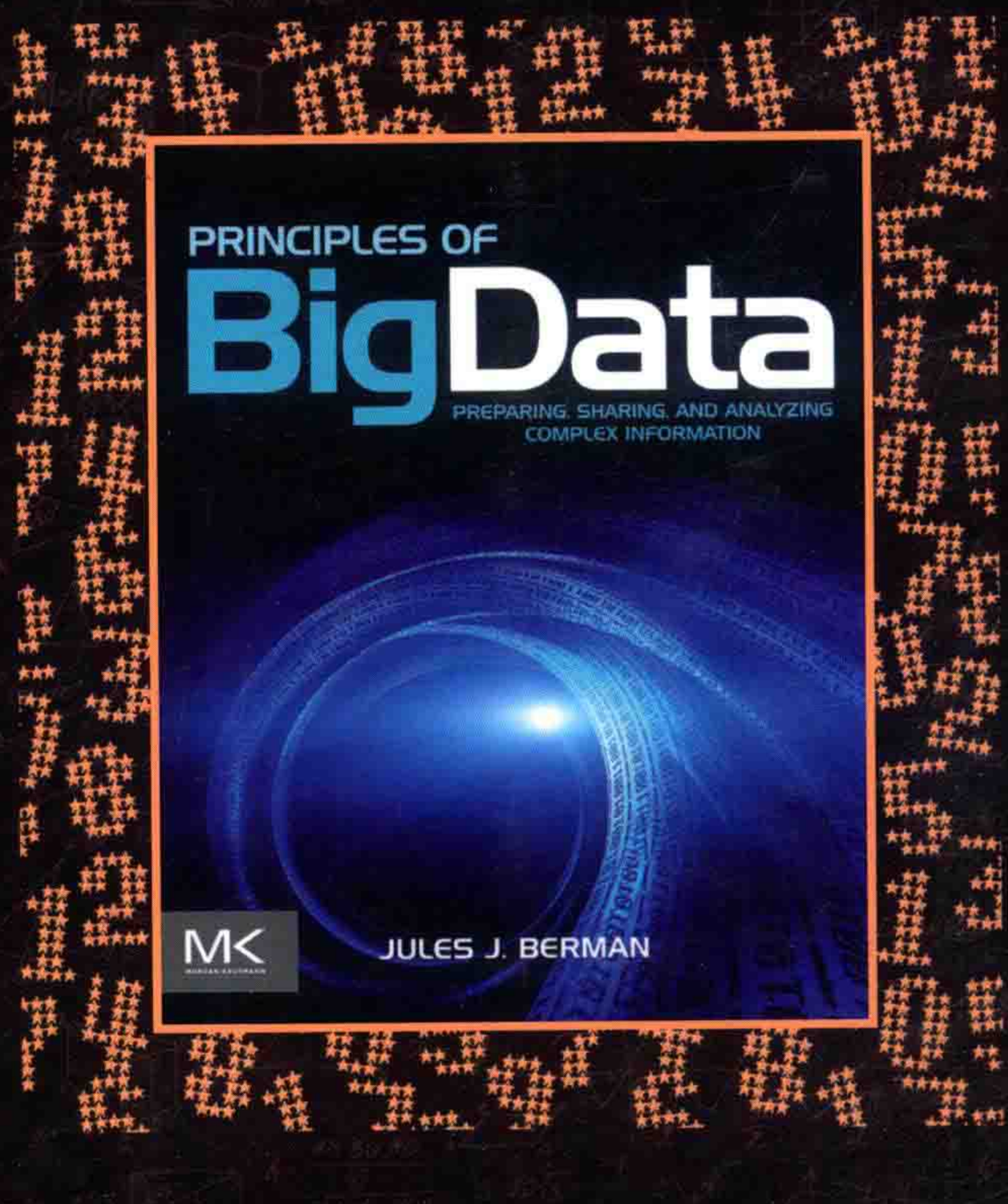
ELSEVIER
爱思唯尔

数据科学与工程技术丛书

大数据原理

复杂信息的准备、共享和分析

[美] 朱尔斯 J. 伯曼(Jules J. Berman) 著
邢春晓 张桂刚 张勇 译



PRINCIPLES OF BIG DATA

PREPARING, SHARING, AND ANALYZING COMPLEX INFORMATION

机械工业出版社
China Machine Press

数据科学与工程技术丛书

PRINCIPLES OF BIG DATA

PREPARING, SHARING, AND ANALYZING COMPLEX INFORMATION

大数据原理

复杂信息的准备、共享和分析

[美] 朱尔斯 J. 伯曼(Jules J. Berman) 著

邢春晓 张桂刚 张勇 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据原理：复杂信息的准备、共享和分析 /（美）朱尔斯 J. 伯曼（Jules J. Berman）著；邢春晓，张桂刚，张勇译 . —北京：机械工业出版社，2017.6
（数据科学与工程技术丛书）
书名原文：Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information

ISBN 978-7-111-57216-9

I. 大… II. ① 朱… ② 邢… ③ 张… ④ 张… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2017）第 146715 号

本书版权登记号：图字：01-2013-7853

大数据原理：复杂信息的准备、共享和分析

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：曲　熠　　责任校对：殷　虹
印　　刷：三河市宏图印务有限公司　　版　　次：2017 年 7 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：13.5
书　　号：ISBN 978-7-111-57216-9　　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

译 者 序

大数据在国民经济的各个领域中都扮演着越来越重要的角色。美国前任总统奥巴马称大数据为“未来的新石油”，可见大数据的重要作用。目前大数据在健康医疗、互联网用户行为分析、社交网络、交通规划、在线教育以及互联网金融等领域均有着重要的应用。

但是在市场上，系统性地介绍大数据原理的书籍并不多，要么浅尝辄止，要么分析不够全面。本书对大数据原理的介绍比较全面和深入，而且具有鲜明的特色，尤其是在大数据应用如此广泛的环境下，可谓是一本不可多得的佳作。

全书 15 章内容逻辑清晰，系统阐述了大数据的各个方面，如为非结构化数据提供结构、数据识别、数据语义、数据集成、数据分析等，可以帮助读者由浅入深地掌握数据架构的精髓。

本书的翻译工作主要由中国科学院自动化研究所张桂刚副教授和清华大学信息技术研究院张勇副研究员完成，清华大学信息技术研究院邢春晓研究员对全书进行了审核。

译者

2017 年 5 月

前　言

我们不能用导致问题的方法去解决问题。

——Albert Einstein

数以百万计的电脑每时每刻都有数据注入。在全球范围内，所有计算机上存储的数据总量约为 3000EB（约 3000 亿 GB），并正以每年 28% 的速度增加。尽管如此，与未被存储的数据量相比，存储下来的数据量仍是微不足道的。据统计，每年约有 1.9ZB 的数据传输量（约 19 000 亿 GB；见术语表，Binary sizes）[1]。日益纷繁复杂的数字化信息将引发新一代数据资源的涌现。

现在，我们有能力从各类资源中得到众多不同类型的数据对象，也能够获取来自未来或遥远过去的数据，这要求我们找到能够准确描述每个数据片段的方法，这样就不至于将数据项混淆，进而能够在需要的时候搜索和追踪对应的数据项。精明的信息学专家明白一个道理：如果要在我们的星球上精确地描述每一件事，必然需要一个“辅助星球”来掌控所有信息，同时后者也必然要比我们的物理星球大很多。

急于获取和分析数据时，往往容易忽视数据的准备工作。如果大数据资源中的数据没有得到有效的组织、综合和准确的描述，那么这些数据资源将毫无价值。本书的首要目标是解释大数据资源建立的原理。大数据资源中的所有数据必须具备某种形式以支持搜索、检索和分析，分析方法必须可再现，分析结果必须可验证。

大数据潜在的最大益处也许是它能够连接一些看似无关的学科，从而开发和测试那些无法通过单个学科领域知识完成的假设性想法。

大数据到底是什么？大数据的特征可以通过三个 V 来描述：Volume（数据体量大）、Variety（数据类型多）和 Velocity（处理速度快）[2]。大数据相关人士常常也会提出其他 V，例如 Vision（有目的和计划）、Verification（确保数据符合规范）和 Validation（核实目标已完成；见术语表，Validation）。

在有关元数据的文献中已对很多大数据的基本原理进行了描述。这类文献讨论了数据描述形式（即如何描述数据）、数据描述语法（例如各种标记语言，如 XML 等）、语义（即如何用计算机可理解的陈述方式传达数据的含义）、语义的表达语法（例如架构规范，如资源描述框架（RDF）和 Web 本体语言（OWL））、包含数据价值和自描述信

息的数据对象的建立、本体的调度以及以数据对象为成员的类层次体系（见术语表，Specification，Semantics，Ontology，RDF，XML）。

对于在数据密集型领域已经取得成功的专家而言，研究元数据似乎是在浪费时间，因为他们对元数据的形式化没有诉求。许多计算机科学家、统计学家、数据库管理员和网络专家可以毫不费力地处理大量的数据，也许他们不认为有必要为大数据资源创造一个“奇怪”的新数据模型。他们觉得自己真正需要的是更大的存储容量和更强大的分布式计算机系统，凭借这些，他们就能存储、检索和分析体量越来越大的数据。然而，这种想法只有在系统使用的数据相对简单或者具有统一标准格式时才适用。一旦大数据资源中的数据变得非常复杂多样，元数据的重要性就会凸显。我们将重点讨论元数据中与大数据息息相关的思想和概念，并重点解释这些思想和概念的必要性以及它们之间的相关性，但不会过于深究细节。

当数据的来源不同，形成许多不同的形式，大小还在增长，价值也在改变，那么当时间延伸到过去和未来时，这场比赛将从数据计算领域转移到数据管理领域。希望本书能说服读者，更快、更强大的计算机是很不错，但这些设备不能弥补在数据准备工作中的不足之处。可以预见，大学、联邦机构和公司将投入大量资金、时间和人力来尝试研究大数据。但如果忽视基础层面的事情，那么他们的项目很可能失败。相反，如果重视大数据的基础知识，则会发现大数据分析能够在普通的计算机上较容易地执行。简单来说，数据本身胜于计算，这也是整本书不断重复的观点。

在其他书籍中，一般会忽略与数据准备过程相关的三个至关重要的主题：**标识符**、**不变性**和**内省**。

完善的标识符系统可以确保属于某个特定数据对象的所有数据能够通过标识符被正确地赋给该对象，而不是其他对象。这看起来很简单，事实也确实如此，但多数大数据资源总是杂乱无章地分配标识符，致使与某个特定对象相关的信息分散在数据源的各个角落，甚至直接被错误地附加到其他对象中，于是当我们需要追踪这些数据的时候已无能为力。对象标识的概念最为重要，因为在面对复杂的大数据资源时，该资源需要被有效地假设为一个唯一标识符集合。本书第 2 章讨论了数据的标识符。

不变性是指被收集到大数据资源中的数据是永久的、不能被篡改的。乍一看，不变性是一个荒诞的和不可能的限制条件。在现实世界中，常有错误发生，信息会发生改变，而且描述信息改变的方法也会发生变化。但一个精明的数据管理员总是知道如何向数据对象中增加信息而不改变当前存在的数据，这些方法在本书第 6 章进行了详细描述。

内省这个词借用了面向对象的程序设计用语，在大数据的相关文献中并不常见。它是指当数据对象被访问时其自我描述的能力。借助内省，大数据资源的使用者能够快速确定数据对象的内容和该对象的层次结构。内省允许使用者查看那些可被分析的数据关系类型，并弄清楚不同数据资源之间是如何交互的。本书第 4 章对内省进行了详细讲解。

本书的另一个主题是数据索引，这也是在大数据相关文献中常被忽视的内容。尽管有很多书籍是基于所谓的书后索引编写而成的，但是为大而杂的数据资源准备索引却需要花费大量精力。因此，多数大数据资源根本没有正式的索引。也许会有一个网页来链接解释性文件，又或者有一个简短且粗糙的“帮助”索引，但很少能找到一个包含完善的、更新过的词条列表和链接的大数据资源。在没有合理索引的情况下，除了少部分行家外，大部分大数据资源对我们根本毫无用处。我很奇怪，有组织愿意花费数亿美元在大数据资源上，却不愿意投资数千美元来建立合理的索引。

在现有的关于大数据的文献中很难找到上述四个主题，除此之外，本书也涵盖了常见的与大数据设计、架构、操作和分析相关的其他主题，包括数据质量、数据标识、数据标准和互操作性问题、遗留数据、数据简化和交换、数据分析和软件问题等。针对这些主题，本书将重点讨论其背后的基本原理，而并不关注编程和数学公式。本书给出了一个全面的术语表，涵盖了书中出现的所有技术词汇和专有词汇。该术语表对与大数据实际相关的词条进行了解释说明，读者可以视该术语表为一个独立的文档。

最后 4 个章节是非技术性的，当然内容上仍与我们讨论的大数据资源的开发一致。这 4 个章节涉及法律、社会和伦理问题。本书最后以我个人对大数据未来及其对世界的影响的观点作为结束。在准备本书时，我在想这 4 个章节放在本书的最前面是不是更合适，因为也许这样能够激发读者对其他技术章节的兴趣。最终，考虑到有些读者不熟悉这些章节的技术语言和概念，因此我将它们放在了接近尾声的地方。具有较强信息学背景的读者从本书第 12 章开始阅读也许更能体会到乐趣。

读者也许会注意到本书中所描述的多数案例来自医学信息学。当前，讨论这一领域的时机已经成熟，因为每一个读者在经济和个人层面都深受来自医学领域所产生的大数据政策和行为的影响。除此之外，关于医疗健康的大数据项目的文献十分丰富，但其中很多文献的成果存在争议，我认为选择那些我可以引证的、可靠的素材是非常重要的。因此，本书参考文献非常多，有超过 200 篇来自期刊、报纸以及书籍的文章，多数文章可从网上下载。

谁应该读这本书？本书是为那些管理大数据资源的专业人士和计算机及信息学领域的学生而写的。专业人士包括：企业和投资机构的领导者，他们必须为项目投入资源；项目主管，他们必须制定一系列可行的目标并管理一个团队，这个团队中的每个人都有一些技能和任务，包括网络专家、元数据专家、软件程序员、标准专家、互操作专家、数据统计分析师以及来自预期用户社区的代表等。来自信息学、计算机科学以及统计学专业的学生会发现，在大学课程中很少讨论大数据面临的挑战，而这些挑战往往是令人惊讶的，有时甚至称得上是令人震惊的。

通过掌握大数据设计、维护、增长和验证的基础知识，读者可以学会如何简化大数据产生的无穷无尽的任务。如果数据准备合理，经验老到的分析师就能够发现不同大数

据资源中数据对象之间的关系。读者会找到整合大数据资源的方法，这比独立的数据库能够提供的好处多得多。

致谢

感谢 Roger Day、Paul Lewis 为书稿的每一章给出了深刻和有价值的评论。感谢 Stuart Kramer 在本书写作初期对文字内容和组织结构给出的宝贵建议。特别感谢 Denise Penrose 在 Elsevier 工作到最后一天以使这本书得以顺利发行。感谢 Andrea Dierna、Heather Scherer 以及 Morgan Kaufmann 所有为本书的出版和营销做出努力的员工们。

作者简介

Jules J. Berman 本科毕业于麻省理工学院，在获得了该校的两个科学学士学位（数学、地球与行星科学）后，他又获得了天普大学的哲学博士学位以及迈阿密大学的医学博士学位。他的博士研究工作是在天普大学的费尔斯癌症研究所和位于纽约瓦尔哈拉的美国健康基金会完成的。Berman 博士在美国国家健康研究院完成了他的博士后研究工作，并曾在华盛顿特区的乔治·华盛顿大学医学中心实习过一段时间。Berman 博士曾在马里兰州巴尔的摩市退伍军人管理局医疗中心担任解剖病理学、外科病理学和细胞病理学的首席专家，在那里他被任命为马里兰大学医学中心和约翰·霍普金斯医学研究机构的主任。1998 年，他在美国国家癌症研究所癌症诊断计划中任病理信息学项目主管，在那里他从事大数据项目工作。2006 年，Berman 博士成为病理信息学协会主席。2011 年，他获得了病理信息学协会终身成就奖。他是数百部科学出版物的作者之一。如今，Berman 博士是一名自由作家，专注于信息科学、计算机程序设计和病理学三个专业领域的书籍写作。

目　录

第0章
引　言

这是数据，笨蛋。

——Jim Gray

回到20世纪60年代，我的高中学校在重要比赛之前都会召开动员大会。在一次动员大会中，橄榄球队的教练扛着一大箱的电脑纸走到舞台中央，每张纸折叠着与下一张相接，并打上孔串了起来。这位教练宣布校队所有成员的竞技能力已经被存储到学校的电脑中（很幸运，当时我们有自己的IBM-360主机），同样，竞争对手的数据也被存储到这台计算机中。我们指示这台计算机消化这些信息，并给出能赢下当年感恩节比赛的队名。于是这台计算机就吐出了前面提到的那一箱电脑纸，最后一张纸显示我们将赢得比赛。第二天，我们遭遇了在年复一年的竞争中的又一次可耻的失败。

让时间快进到大约50年前，马里兰州贝塞斯达国家癌症研究中心会议室，我正在听取一位女性顶级科学管理员讲述过去十年癌症研究的快速发展。她表明，当时最好的研究计划是多机构的和数据密集型的。那些受到资助的研究人员当时使用高通量分子方法，在短短几分钟内就能为每个组织样本产生堆积如山的数据，而当时能想到的只有一种解决方法，就是依靠超级计算机和一批聪明的程序员，他们可以分析这些数据并告诉我们这些数据背后的含义。

与我高中那位教练想的一样，美国国家健康研究院（NIH）的领导们认为，只要计算机足够“大”，无论输入多少信息量，它都能够输出结果。

然而在大约2003年的一天，在美国国家健康研究院的一间会议室里，我表明了自己的想法，指出不能只是单纯地向计算机输入数据，然后等待给出预期的结果。从古至今，任何一门科学都是一个约简的过程，即从复杂的、描述性的数据集到简化的概括。让那种昂贵的超级计算机来处理数据量越来越大、越来越复杂的生物数据几乎是不现实的，也没这个必要（见术语表，Supercomputer）。那天，我的想法没有被接受，研制高性能超级计算机当时仍是一个非常热门的课题，当然现在仍然是。

自基于超级计算机的癌症诊断方法提出以来已过去十年之久，那台诊断用的超级计算机设备仍没有制造出来。医院实验室用的诊断工具还是1590年研制出来的微电子显微镜。如今，我们从报刊中了解到科学家能够通过窥探组成我们基因的DNA的全部序列来给出重要的诊断结果。尽管如此，医生很少能对全基因组扫描排列，也没有人知道如何有效地使用基

因数据。你也许会说医院和诊所有很多计算机，但这些计算机并非用来“计算”你的诊断结果。在医疗场所的计算机大部分仍是收集、存储、检索数据和传送医疗记录的工具。

在我们能够充分利用大量且复杂的数据资源之前，需要深入思考大数据的意义和命运。

0.1 大数据的定义

大数据可以用三个“V”来定义：

1. Volume——数据体量大。

2. Variety——数据的来源多种多样，包括传统数据库、图像、文件和其他复杂的记录。

3. Velocity——通过吸收来自补充数据集的数据，引入已存档的数据或遗留的数据集，以及来自多种数据源的流数据，数据一直在变。

大数据（big data）不是很多数据（lotsa data），也不是海量数据（massive data），理解这一点很重要。在大数据资源中，上述三个“V”必须都适用。大数据资源独有的数据量大、复杂程度高和数据无穷无尽的特点决定了其数据设计、操作和数据分析方法也具有特定性。

“lotsa data”常用来表示大量格式简单的记录数据的集合，例如：每颗可观测到的星星的大小和位置；每个在美国的人和他们的电话号码；每个现存物种及其谱系；等等。这些数据量较大的数据集往往美其名曰“列表”，其中有一些是目录，其目的是存储和检索信息；还有一些 lotsa data 数据集是电子表格（行列二维表），数学上等价于一个巨大的矩阵。出于科学研究的目的，有时同时分析一个矩阵中的所有数据是非常必要的。矩阵分析强调计算，也许需要一台超级计算机的协助。这种对于大型矩阵的全局分析不是本书的主题。

大数据资源并不等价于一个大型的电子表格，也不意味着从总体上进行分析。大数据分析是一个多步骤的过程，在此过程中数据经过提取、过滤和转换，然后进行逐个分析或递归分析。在你读这本书时，会发现“lotsa data”与大数据之间的区别非常之大，这两者几乎不能在同一场所被有效地讨论。

0.2 大数据 VS 小数据

大数据不是已经膨胀到一个电子表格无法装下的小数据，也不是碰巧变得非常大的数据库。然而，一些习惯于处理小数据集的专业人士认为他们的电子表格和数据库技巧也适用于大数据资源，不需要掌握新的技巧或使用新的分析范式。从他们的角度，当数据变得越来越大时，只需要计算机去适应（计算速度更快、信息获取更多、存储容量更大等），大数据并没有摆出一些特殊难题以致于一台超级计算机都无法解决。

这种看待大数据的态度在数据库管理员、程序员和统计学家中普遍存在，但这是反生产力的。长此以往，将导致软件缓慢甚至无效，高投入低回报，数据分析能力不佳，甚至产生无用且不可逆的大数据资源缺陷。

让我们来看几个一般性差异，这些可以帮助我们区分大数据和小数据。

1. 目标

小数据——常用来回答某个特定问题或服务于某个特定目标。

大数据——通常在思想上围绕一个目标而设计，但这个目标是可变的，摆出的问题也是千变万化的。这里有一个简短的、虚构的大数据资助基金，其目标是把来自渔业、海岸警卫队、商业航运、沿海管理机构的持续增长的数据收集起来，以支持下半岛的各种政府和商业

管理的学习研究。在这个虚构的事件中，有一个模糊的目标，但这个目标显然没有办法指明大数据资源具体包含哪些内容，也无法完全解释大数据资源中的那些多种多样的数据以何种组织形式存在，如何与其他数据资源发生联系以及如何利用其进行数据分析。无论是谁都不能详述大数据的最终命运，通常来讲，大数据总是给我们带来惊喜。

2. 地点

小数据——通常，小数据属于某个机构，常常存储在某台电脑中，有时也会存储在某个文件夹中。

大数据——通常通过电子空间传输，被分配到多个网络服务器上，存在于地球的任何地方。

3. 数据结构和内容

小数据——通常包含高度结构化的数据，数据域被限制在某个单一的学科或分支学科之内。这些数据通常来自一个顺序电子表格，其记录格式是统一的。

大数据——必须有吸收非结构化数据的能力（如自由文本、图像、视频、音频、实体对象等）。数据源的内容也许跨多个学科，而其中每个独立的数据对象又有可能与其他大数据资源的数据相关联。

4. 数据准备

小数据——在很多情况下，数据使用者从其个人的目的出发准备数据。

大数据——数据来自众多多样化的数据源，并由很多人来准备。数据的使用者很少是该数据的准备者。

5. 寿命

小数据——当数据项目结束时，小数据保存的时间有限（很少超过研究数据的传统学术寿命，即大概 7 年），然后被擦除。

大数据——大数据项目使用的数据通常需要永久保存。理想情况下，当原始资源寿命结束时，存储在大数据资源中的数据将被吸收到另一个资源池中。很多大数据项目累积的数据会延伸到未来和过去（例如遗留数据）。

6. 测量

小数据——通常小数据使用一个实验协议来进行测量，且该数据可由某个标准单元集描述。

大数据——众多不同类型的数据以多种不同的电子格式传输着。当数据可测量时，测量结果可通过多种协议获取。对数据管理者而言，确定大数据的质量是最困难的任务之一。

7. 再现性

小数据——小数据项目通常情况下是重复的。如果有关于数据质量的问题，或对数据再现性、从数据中得到的结论的正确性有疑问，那么整个项目可被重现，并产生新的数据集。

大数据——通常复制大数据几乎是不可行的。在多数情况下，人们希望能够在大数据资源中发现坏数据并进行标记等。

8. 风险

小数据——小数据项目的开销是有限的，实验室和研究机构往往能够从偶然的小数据失败中恢复过来。

大数据——大数据项目会非常昂贵。一个大数据项目的失败会导致公司破产、机构崩塌、大规模解雇员工以及存储在资源中的所有数据的瞬间瓦解。举个例子，NIH 大数据项目，全称为“NCI cancer Biomedical Informatics Grid”，即“癌症生物医学信息网格”（见术语表，Grid），该项目从 2004 年到 2010 年花费了至少 3.5 亿美元。审查资源的一个专设委员会发现尽管项目组投入了数百名癌症研究人员和信息专家的努力，但项目基本没有完成且资金投入巨大，最终该项目被废止[3]。自那以后，这些数据资源很快被终止了[4]。虽然以金钱、时间和工作量来衡量该项目，其开销无疑是巨大的，但大数据的失败也许仍有一些可取的价值，毕竟失败是成功之母。

9. 内省

小数据——独立的数据点由它们在数据表或数据库中的行和列的位置识别（见术语表，Data point）。如果知道行和列的表头，那么就可以找到和列举其中包含的全部数据。

大数据——除非大数据资源可以如预期的那样设计良好，否则即便是数据管理员也难以理解大数据资源的内容和组织形式（见术语表，Data manager）。要获取数据、掌握数据价值信息和数据组织信息，需通过内省技术才能达成（见术语表，Introspection）。

10. 分析

小数据——大多数情况下，项目中的所有数据可同时进行全部分析。

大数据——无论是在超级计算机中还是在多个计算机中并行进行的大数据分析几乎都需要一步步递增式完成（见术语表，Parallel computing，MapReduce）。这些数据需经过多种方法进行提取、查看、删减、标准化、转换、可视化、释义和再分析等操作。

0.3 大数据在哪里

一般来讲，大数据的推动力是一种被动刺激。各个公司和一些专业行政机构，无论他们是否愿意，都不得不存储和检索大量收集到的数据。

大数据往往通过多种不同的机制出现。

1. 企业在其正常的业务活动过程中，收集了大量数据并试图组织这些数据，以期可以根据需要检索资料。大数据致力于简化这个实体的正常活动。数据等待着被使用，这个组织不是寻求发现什么或开展其他新的业务活动，而只是简单地想更好地使这些数据为其现有业务服务。一个典型的医疗中心就是一个“意外的”大数据资源的好例子，医生和护士日复一日地照顾病人并将数据记录到医院信息系统，使得收集到的数据达到 TB 级，而这些数据以多种形式存在，如实验报告、处方单、临床案例和收费数据。这些信息大部分是为了某个一次性的特定用途而产生的（例如，支持某个临床决策，确定某个疗程该如何收费）。行政人员根据收集到的数据来达到一些目的，如提高服务质量、提高员工效率或降低运营成本。

2. 企业在其正常的业务活动过程中已经收集了大量数据，并确信凭借这些数据可以开发新的业务活动。如果是一些现代化企业——这些企业不会将其业务限定在某种制造工艺或仅面向某个客户群体。他们一直在寻找新的机遇，他们收集的数据也许恰好可以帮助这些企业基于客户的喜好来开发新的产品，从而开辟新的市场或通过网络销售产品。这些企业将成为受益于大数据的制造企业。

3. 企业制定一个基于大数据资源的商业模型。和以往的企业不同，这个企业以大数据起步，然后加入实体成分。亚马逊和联邦快递应该划入这一类，因为他们是从提供一种数据