HZ BOOKS 华章IT
ELSEVIER

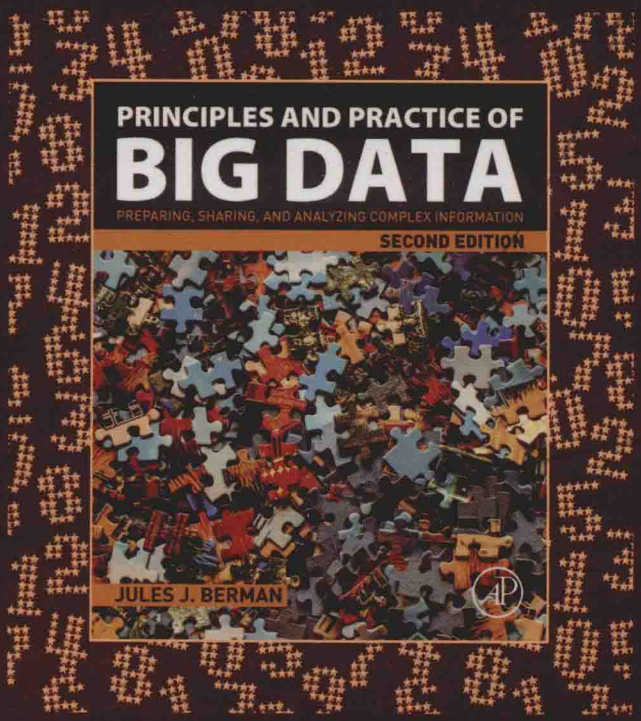
数据科学与工程技术丛书

大数据原理与实践

复杂信息的准备、共享和分析

（原书第2版）

[美] 朱尔斯·J. 伯曼（Jules J. Berman） 著
张桂刚 邢春晓 任广皓 王云 译



PRINCIPLES AND PRACTICE OF BIG DATA

PREPARING, SHARING, AND ANALYZING COMPLEX INFORMATION, SECOND EDITION

机械工业出版社
China Machine Press

数据科学与工程技术丛书

PRINCIPLES AND PRACTICE OF BIG DATA

PREPARING, SHARING, AND ANALYZING COMPLEX INFORMATION, SECOND EDITION

大数据原理与实践

复杂信息的准备、共享和分析

（原书第2版）

[美] 朱尔斯·J. 伯曼（Jules J. Berman） 著

张桂刚 邢春晓 任广皓 王云 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据原理与实践：复杂信息的准备、共享和分析（原书第 2 版）/（美）朱尔斯·J. 伯曼（Jules J. Berman）著；张桂刚等译 . —北京：机械工业出版社，2020.6
（数据科学与工程技术丛书）
书名原文：Principles and Practice of Big Data: Preparing, Sharing, and Analyzing Complex Information, Second Edition

ISBN 978-7-111-65790-3

I. 大… II. ①朱… ②张… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2020）第 095665 号

本书版权登记号：图字 01-2019-0927

大数据原理与实践
复杂信息的准备、共享和分析（原书第 2 版）

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：唐晓琳　　责任校对：马荣敏
印　　刷：大厂回族自治县益利印刷有限公司　　版　　次：2020 年 6 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：23.25
书　　号：ISBN 978-7-111-65790-3　　定　　价：119.00 元

客服电话：（010）88361066　88379833　68326294　　投稿热线：（010）88379604
华章网站：www.hzbook.com　　读者信箱：hzit@hzbook.com

译 者 序

距本书第 1 版中译本的出版已经过去了两年，这期间大数据领域蓬勃发展，许多新知识概念、技术手段层出不穷。虽然本书的第 1 版深入全面地对大数据进行了介绍，但相较于近几年大数据的飞速发展，还是略微有些陈旧。因此，第 2 版应运而生。

第 2 版中，作者在第 1 版的基础上进行了修订，增加了许多新的概念，同时对之前的概念结构以及内容进行了梳理和调整，并结合更加丰富的例子进行讲解。与第 1 版最大的不同在于，在第 2 版中，作者添加了与概念紧密结合的 Python 代码，并配之以注释，让读者在学习理论的同时亦能够了解这些概念的实践方法。

全书从之前的 15 章，增加到了 20 章。对大数据的基本概念以及大数据的分析技巧都进行了详细的讲解。除此之外，与大数据有关的各类其他领域亦有所涉及，如大数据的合法性以及公众问题。全书内容主线逻辑清晰，知识全面，不仅能够作为对大数据领域感兴趣的初学者的入门教材，也能够当作大数据领域工作者、学者的最佳工具书。

本书主要由中科院自动化研究所张桂刚副研究员、任广皓助理研究员以及王云助理研究员翻译完成，清华大学信息技术研究院邢春晓研究员对全书进行了审校。

译者

2020 年 4 月

第 2 版前言

一切都说过了，但因为大家都没有听，我们不得不回归原点，从头开始。

——Andre Gide

优秀的科学作家总是抓住机会为早期的作品写一部第 2 版。无论多么努力地尝试，第 1 版总是会含有不准确，甚至产生误导的言论。随着时间的推移，那些在第 1 版中看起来很精彩的句子也变成了夸大智慧的例子。那些由于太过微不足道而在原始手稿中没有包含进去的部分现在似乎成了需要被完整阐释的深刻内容。第 2 版为懊恼的作者提供了纠正这些的机会。

当 2013 年本书第 1 版出版的时候，这个领域还很年轻，很少有科学家知道大数据能够做什么。在世界各处，无时无刻不断涌入的数据被存储起来，就像小麦被保存在筒仓中一样。对于数据管理者来说，很显然这些被存储起来的数据是没有任何科学价值的，除非它们通过元数据、标识符、时间戳和一组基本描述符恰当地注释。在这种情况下，本书第 1 版指出了在大数据收集、注释、组织并展开过程中那些适当且重要的方法。处理大数据的过程伴随着独特的挑战，第 1 版充满了警告和劝告，旨在引导读者远离灾难。

自第 1 版出版至今已有数年了，此后有数百本关于大数据主题的书籍相继出版。作为一名科学家，我很失望地发现，现今关于大数据的主题都聚焦于营销和预测分析（例如，“谁有可能购买产品 x，由于他们两周前购买了产品 y”），以及机器学习（例如，无人驾驶汽车、计算机视觉、语音识别）等问题上。机器学习在很大程度上依赖于那些大肆宣传的技术，如神经网络和深度学习，这两者并没有简化和拓宽我们理解自然界和物质世界的基本法则和原则。在大多数情况下，这些技术使用的都是相对较新的（即新收集的）、标注较差的（即仅提供一个特定的分析过程所需的最小信息）、不被保存以便于公共评估或重复使用的数据。简而言之，大数据已经走上了阻力最小的道路，避免了本书第 1 版中提出的大多数棘手问题，例如，与公众共享数据的重要性，探索数据对象之间关系（非相似性）的价值，以及创建鲁棒的、不变的且注释良好的数据这一沉重但不可避免的负担。

我当然希望大数据的伟大进步将成为医学、生物学、物理学、工程学和化学领域的根本性突破。为什么大数据的重点从基础科学转向了机器学习？这可能与这样一个事实有关，即包括本书第 1 版在内的所有书籍都没能为读者提供将大数据原理付诸实践的方法。回想起来，光描述理论，然后寄期望于读者开拓出自己的方法是不够的。

因此，在第 2 版中，在介绍理论的同时，亦将提供与之相应的实践手段。读者会发现

用于实现大数据准备和分析的所有方法都非常简单。在大多数情况下，为了理解计算机方法，需要对编程语言有一些基本的了解。尽管会有疑虑，但 Python 将作为本书的首选语言。Python 的优点如下：

- Python 是一种免费的、开源的高级编程语言，易于获取、安装、学习和使用，并且适用于所有主流的计算机操作系统。
- Python 目前非常受欢迎，并且其受欢迎程度似乎越来越高。
- Python 发行版（例如 Anaconda）绑定了数百个非常有用的模块（例如 numpy、matplot 和 scipy）。
- Python 有一个庞大而活跃的社群，这为 Python 方法和模块提供了大量的参考文档。
- Python 支持一些面向对象的技术，这些技术将在第 2 版中有所讨论。

事物都有其两面性，Python 也有缺点：

- 最新版本的 Python 并不向后兼容其早期版本。因此，本书中所包含的脚本和代码块适用于大多数 Python 3.x 版本，但可能不适用于 Python 2.x 版本及更早版本，除非读者准备花费一些时间来进行代码调整。当然，这些简短的脚本和代码块旨在作为概念的简单演示，不能用于构建应用程序的代码。
- 内置的 Python 方法有时优化了速度以利用随机存取存储器（RAM）来保存数据结构，包括通过循环构建的数据结构。对大数据进行迭代可能会耗尽可用的内存，从而导致那些在小数据集上运行良好的 Python 脚本失败。
- Python 面向对象的实现允许多类继承（即，类可以是多个父类的子类）。我们将解释为什么在使用 Python 对大型复杂数据进行编程操作时使用多类继承会引起麻烦，并且给出所对应的补救措施。

本书中描述的每个算法的核心概念都可以在现代计算机上通过任何操作系统、利用主流的编程语言通过几行代码加以实现。本书会提供许多 Python 代码，并讲解主流的操作系统上被广泛使用的开源应用。本书强调，对于大型复杂数据集上的数据分析，大部分都可以通过简单的方法加以实现，而不需要专门的软件系统（例如，并行化的计算过程）或硬件（例如，超级计算机）。同时，完全不熟悉 Python 的读者可能会惊奇地发现，当代码很简短并且附有注释的时候，他们能够轻易地读懂 Python 代码。当然，对于那些主要关注如何掌握大数据原理的读者，可以跳过代码段，不用担心会错过书中的任何理论概念。

第 2 版同时包含了那些被大数据分析领域的其他书籍所忽视的方法论，包括：

- **数据准备**。如何使用元数据注释数据以及如何创建由三元组组成的数据对象。作为计算科学意义的基本延伸，三元组的概念将被全面地讲解。
- **与大数据相关的数据结构**。涵盖诸如 TripleStore、分布式账本、唯一标识符、时间戳、一致性、索引、字典对象、数据持久性、用于数据存储和分布的单向散列的作用以及加密协议等概念。
- **数据对象的分类**。详细讨论如何根据数据对象的共享关系将数据对象分类，以及在大数据分析中分类所起到的计算作用。
- **内省**。如何创建自描述的数据对象，允许数据分析人员对属于同一个类的对象进

行分组，并将方法应用于从其祖先类继承的类对象。

- **在大数据准备和分析中具有特殊效用的算法。**如何使用单向散列、唯一标识符生成器、加密技术、计时方法和时间戳协议来创建不可变（永不改变）的、永久的、私有的唯一数据对象，并且创建便于描述具有许多有用功能的数据结构（例如，区块链和分布式账本，用于安全地共享机密信息的协议，以及在不侵犯隐私的情况下用于协调跨数据集合的标识符的方法）。
- **大数据分析技巧。**如何使用一系列简单的技术（例如，近似、所谓的包络技巧、使用随机数发生器的重复采样、蒙特卡罗模拟以及数据简化方法）来克服大数据规模和维度所施加的诸多分析限制。
- **数据再分析、数据重新利用和数据共享。**为什么对于大数据的第一次分析几乎总是不正确的、有误导的或不完整的？为什么数据再分析已成为每个严肃的大数据分析师必须掌握的关键技能？数据再分析的过程通常会激发大数据资源的重新利用。除非克服数据共享的障碍，否则不能实现数据再分析和数据重新利用。该主题将会被详细讨论。

尽管本书第 2 版希望并且尽最大可能地达到对大数据领域的全面覆盖，但没有办法实现涵盖大数据这一多学科交叉领域的每个概念和方法。为了弥补这些不足，每章都提供扩充的术语表，它定义了文中引入的术语，并对大数据科学家常用的术语提供了进一步的解释。另外，对于所讨论的技术与方法，亦为读者提供了或许有用的参考文献列表，以供进一步阅读。总的来说，第 2 版包含了大约 600 个参考文献，其中大多数能够免费下载，以及 300 多个术语词条，其中许多词条包含了读者可能会觉得有用的简短的 Python 片段。

最后，第 2 版通过案例来向读者展示大数据原理如何付诸实践。虽然案例研究来自多个科学领域，涉及物理学、经济学和天文学，但读者会注意到从生物科学（特别是医学和动物学）中汲取了大量例子。其原因在于，对于陆生生物的分类是现存最古老、最佳的大数据分类。数据组织和数据分析过程中出现的所有经典错误都是在生物学领域中发生的。更重要的是，这些错误已被详细记录下来，并且大多数已得到纠正并公布以供公众参考。如果想了解大数据如何用作帮助科学进步的工具，那么你必须学习这些从生物学世界中获取的案例，该领域很好地记录了过去发生、正在发生以及将要发生的一切。我们已尽最大努力采用同类案例中最简单的，并提供尽可能多的背景解释。

本书第 2 版致力于让读者认识到大数据分析的主要目的是让我们能够提出并回答一系列无法通过小型数据集得到可靠解决方案的问题。我们完全有理由期待本书的读者能够很快取得前几代科学家无法企及的科学突破。祝大家好运！

第 1 版前言

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 300EB（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）[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。我们将评估那些能引导分析人员在不同大数据资源中创建新的合并数据集的方法。

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（数据增长速度快）[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）。

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。

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，形成的形式不同，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。

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添加信息而不改变当前存在的数据，这些方法在本书中进行了详细描述。

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，有的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、为非结构化数据提供结构、数据去标识、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 章是非技术性的，内容上与利用大数据资源的后果相关。这 4 章涉及法律、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 章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中的技术语言和概念，因此我将它们放在了接近尾声的地方。

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、数据领域专家、元数据专家、软件程序员、标准专家、互操作专家、数据统计师、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资源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。

参考文献

[1] Martin Hilbert M, Lopez P. The world's technological capacity to store, communicate, and compute information. Science 2011;332:60–5.

[2] Schmidt S. Data is exploding: the 3V's of Big Data. Business Computing World; 2012. May 15.

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，由马里兰大学医学中心和约翰·霍普金斯医学研究机构共同任命。1998 年，他转入美国国立卫生研究院担任卫生干事，并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。Berman 博士曾任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，在数据科学和疾病生物学领域编写了十余本书籍。最近，由 Elsevier 出版的书包括：

- Taxonomic Guide to Infectious Diseases: Understanding the Biologic Classes of Pathogenic Organisms (2012)
- Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information (2013)
- Rare Diseases and Orphan Drugs: Keys to Understanding and Treating the Common Diseases (2014)
- Repurposing Legacy Data: Innovative Case Studies (2015)
- Data Simplification: Taming Information with Open Source Tools (2016)
- Precision Medicine and the Reinvention of Human Disease (2018)

目　录

第 1 章
引　言

1.1　大数据的定义

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计算机中（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了堆积如山的数据，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，美国国立卫生研究院（NIH）的领导认为，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，在美国国立卫生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，Science，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，当然现在也是。

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，可以通过对特定的关键突变进行基因检测来增强微观研究，但我们并不试图了解人类遗传变异的所有复杂性。我们知道尝试是没有希望的。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沦为处理收集、存储、检索和传送医疗记录等乏味任务的工具，完成这些任务后，计算机会发送服务产生的账单。

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。

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、操作和分析方法也具有特定性（见术语表，Big Data Resource, Data resource）。

“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，网页内容，等等。这些数据量较大的数据集往往美其名曰“列表”；还有一些lotsa data数据集是电子表格（行列二维表），非常大以致我们几乎看不到电子表格的末尾。

大数据资源并不等价于大型的电子表格，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情境下被有效地讨论。

1.2　大数据与小数据

实际上，大科学的主要功能是产生大量可靠且易于获取的数据……洞察力、理解力和科学进步通常都是通过“小科学”来实现的。

——Dan Graur, Yichen Zheng, Nicholas Price, Ricardo Azevedo, Rebecca Zufall, Eran Elhaik[1]

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（见术语表，Databose）。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