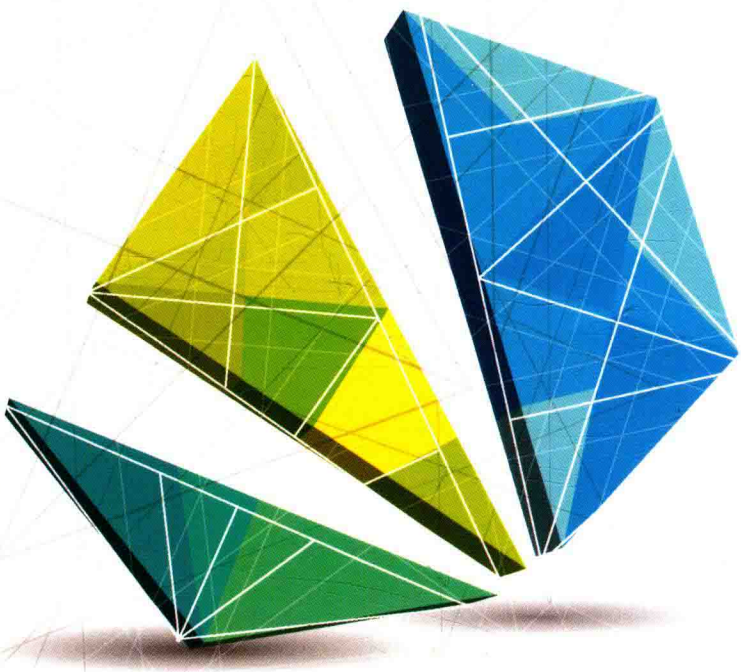
高等学校大数据技术与应用规划教材

大数据及其可视化

DASHUJU JIQI KESHIHUA

周苏　王文　等编著



中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

高等学校大数据技术与应用规划教材

大数据及其可视化

周　苏　王　文　等编著

中国铁道出版社
CHINA RAILWAY PUBLISHING HOUSE

内 容 简 介

大数据及其可视化是一门理论性和实践性都很强的课程。本书针对计算机、信息管理、经济管理和其他相关专业学生的发展需求，系统、全面地介绍了关于大数据及其可视化技术的基本知识和技能，详细介绍了大数据与大数据时代、数据可视化之美、Excel 数据可视化方法、Excel 数据可视化应用、大数据的商业规则、大数据激发创造力、大数据预测分析、支撑大数据的技术、数据引导可视化、Tableau 可视化初步、Tableau 数据管理与计算、Tableau 可视化设计、Tableau 地图与预测分析和 Tableau 分享与发布等内容，具有较强的系统性、可读性和实用性。

本书适合作为普通高等院校相关专业“大数据基础”“大数据导论”“大数据可视化”等课程的教材，也可供有一定实践经验的软件开发人员、管理人员学习参考。

图书在版编目（CIP）数据

大数据及其可视化/周苏，王文等编著. —北京：中国铁道出版社，2016.8

高等学校大数据技术与应用规划教材

ISBN 978-7-113-22079-2

Ⅰ. ①大… Ⅱ. ①周… ②王… Ⅲ. ①数据处理－高等学校－教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2016）第 177191 号

书　　名：大数据及其可视化
作　　者：周　苏　王　文　等编著

策　　划：周海燕　　**读者热线：**（010）63550836
责任编辑：周海燕　冯彩茹
封面设计：穆　丽
责任校对：汤淑梅
责任印制：郭向伟

出版发行：中国铁道出版社（100054，北京市西城区右安门西街 8 号）
网　　址：http://www.51eds.com
印　　刷：北京海淀五色花印刷厂
版　　次：2016 年 8 月第 1 版　　2016 年 8 月第 1 次印刷
开　　本：787 mm×1 092 mm　1/16　印张：20　字数：448 千
书　　号：ISBN 978-7-113-22079-2
定　　价：46.00 元

前　言

大数据（Big Data）的力量正在积极地影响着社会的方方面面，它冲击着许多主要的行业，包括零售业、电子商务和金融服务业等，同时，也正在彻底地改变人们的教育方式、生活方式、工作方式。如今，通过简单、易用的移动应用和基于云端的数据服务，人们能够追踪自己的行为以及饮食习惯，还能提升个人的健康状况。因此，有必要真正理解大数据这个极其重要的议题。

中国是大数据最大的潜在市场之一。据估计，中国有近 6 亿网民，这就意味着中国的企业拥有绝佳的机会来更好地了解其客户并提供更个性化的体验，同时，为企业增加收入并提高利润。阿里巴巴就是一个很好的例子，其不但在商业模式上具有颠覆性，而且还掌握了与购买行为、产品需求和库存供应相关的海量数据。除了阿里巴巴高层的领导能力之外，大数据必然是其成功的一个关键因素。

然而，仅有数据是不够的。对于身处大数据时代的企业而言，成功的关键还在于找出大数据所隐含的真知灼见。“以前，人们总说信息就是力量，但如今，对数据进行分析、利用和挖掘才是力量之所在。”

很多年前，人们就开始对数据进行利用。例如，航空公司利用数据为机票定价，银行利用数据搞清楚贷款对象，信用卡公司则利用数据侦破信用卡诈骗等。但直到最近，数据才真正成为人们日常生活的一部分。随着谷歌（Google）以及 QQ、微信、淘宝等的出现，大数据游戏被永远改变了。你和我，或者任何一个享受这些服务的用户都生成了一条数据足迹，它能够反映出人们的行为。每次进行搜索时，如查找某个人或者访问某个网站，都加深了这条足迹。互联网企业开始创建新技术来存储、分析激增的数据——结果就迎来了被称为“大数据”的创新爆炸。

进入 2012 年以来，由于互联网和信息行业的快速发展，大数据越来越引起人们的关注，已经引发云计算、互联网之后 IT 行业的又一大颠覆性的技术革命。人们用大数据来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。云计算主要为数据资产提供保管、访问的场所和渠道，而数据才是真正有价值的资产。企业内部的经营信息、互联网世界中的商品物流信息，互联网世界中的人与人交互信息、位置信息等，其数量将远远超越现有企业 IT 架构和基础设施的承载能力，实时性要求也将大大超越现有的计算能力。如何盘活这些数据资产，使其为国家治理、企业决策乃至个人生活服务，是大数据的核心议题，也是云计算内在的灵魂和必然的升级方向。

对于在校大学生来说，大数据及其可视化的理念、技术与应用是一门理论性和实践性都很强的“必修”课程。在长期的教学实践中，我们体会到坚持“因材施教”的重要原则，把实践环节与理论教学相融合，抓实践教学促进理论知识的学习，是有效改善教学效果和提高教学水平的重要方法之一。本书的主要特色是：理论联系实际，

结合一系列了解和熟悉大数据理念、技术与应用的学习和实践活动，把大数据及其可视化的相关概念、基础知识和技术技巧融入实践中，使学生保持浓厚的学习热情，加深对大数据技术的兴趣、认识、理解和掌握。

本书系统、全面地介绍了大数据及其可视化的基本知识和应用技能，详细介绍了大数据与大数据时代、数据可视化之美、Excel 数据可视化方法、Excel 数据可视化应用、大数据的商业规则、大数据激发创造力、大数据预测分析、支撑大数据的技术、数据引导可视化、Tableau 可视化初步、Tableau 数据管理与计算、Tableau 可视化设计、Tableau 地图与预测分析，以及 Tableau 分享与发布等内容，具有较强的系统性、可读性和实用性。

本课程的教学评测可以从这样几个方面入手，即：

（1）每章课前【案例导读】（14 次）。

（2）每章课后【实验与思考】（14 次）。

（3）课程设计（附录）。

（4）课程实验总结（附录）。

（5）结合平时考勤。

（6）任课老师认为必要的其他考核方法。

与本书配套的教学 PPT 课件等文档可从中国铁道出版社教学资源网站（www.tdpress.com\51eds）的下载区下载，欢迎教师与作者交流并索取为本书教学配套的相关资料并交流：zhousu@qq.com，QQ：81505050，个人博客：http://blog.sina.com.cn/zhousu58。

本书由周苏、王文等编著，并得到浙江大学城市学院、浙江商业职业技术学院、温州安防职业技术学院等多所院校师生的支持，王硕苹、张丽娜、张健、吴林华等参与了本书的部分编写工作，在此一并表示感谢！

由于编者水平有限，加之时间仓促，书中难免存在疏漏和不足之处，恳请读者批评指正。

周　苏

2016 年初夏于西子湖畔

目　录

第1章 大数据与大数据时代

【案例导读】亚马逊推荐系统

虽然亚马逊[①]的故事大多数人都耳熟能详，但只有少数人知道它早期的书评内容最初是由人工完成的。当时，亚马逊公司（见图 1-1）聘请了一个由 20 多名书评家和编辑组成的团队，他们写书评、推荐新书，挑选非常有特色的新书标题放在亚马逊的网页上。这个团队创立了“亚马逊的声音”版块，成为当时公司皇冠上的一颗宝石，是其竞争优势的重要来源。《华尔街日报》的一篇文章中热情地称他们为全美最有影响力的书评家，因为他们使得书籍销量猛增。

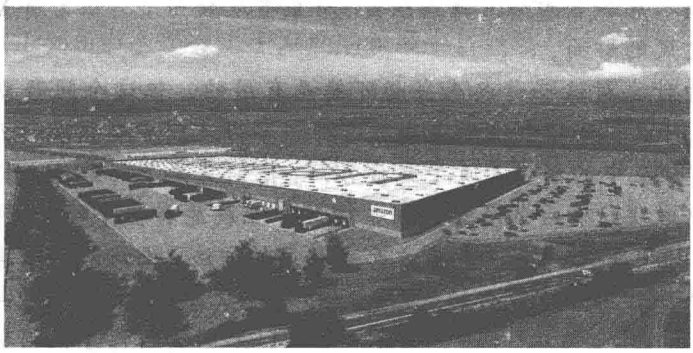


图 1-1　亚马逊公司

亚马逊公司的创始人及总裁杰夫·贝索斯决定尝试一个极富创造力的想法：根据客户个人以前的购物喜好，为其推荐相关的书籍。

从一开始，亚马逊就从每一个客户那里搜集了大量的数据。比如说，他们购买了什么书籍？哪些书他们只浏览却没有购买？他们浏览了多久？哪些书是他们一起购买的？客户的信息数据量非常大，所以亚马逊必须先用传统的方法对其进行处理，通过样本分析找到客户之间的相似性。但这些推荐信息是非常原始的，就如同你在买一件婴儿用品时，会被淹没在一堆差不多的婴儿用品中一样。詹姆斯·马库斯回忆说：“推荐信息往往为你提供与你以前购买物品有微小差异的产品，并且循环往复。”

亚马逊的格雷格·林登很快就找到了一个解决方案。他意识到，推荐系统实际上

① 亚马逊公司（Amazon）：是美国最大的网络电子商务公司之一，也是“财富 500 强”公司，位于华盛顿州的西雅图，成立于 1995 年 7 月，已成为全球商品种类最多的网上零售商。亚马逊致力于成为全球最“以客户为中心”的公司，使客户能在公司网站上找到和发现任何他们想在线购买的商品，并努力为客户提供最低的价格。

并没有必要把顾客与其他顾客进行对比，这样做在技术上也比较烦琐，需要做的是找到产品之间的关联性。1998 年，林登和他的同事申请了著名的“item-to-item”协同过滤技术的专利。方法的转变使技术发生了翻天覆地的变化。

因为估算可以提前进行，所以推荐系统不仅快，而且适用于各种各样的产品。因此，当亚马逊跨界销售除书以外的其他商品时，也可以对电影或烤面包机这些产品进行推荐。由于系统中使用了所有的数据，推荐会更理想。林登回忆道：“在组里有句玩笑话，说的是如果系统运作良好，亚马逊应该只推荐你一本书，而这本书就是你将要买的下一本书。”

现在，公司必须决定什么应该出现在网站上，是亚马逊内部书评家写的个人建议和评论，还是由机器生成的个性化推荐和畅销书排行榜？

林登做了一个关于评论家所创造的销售业绩和计算机生成内容所产生的销售业绩的对比测试，结果他发现两者之间相差甚远。他解释说，通过数据推荐产品所增加的销售远远超过书评家的贡献。计算机可能不知道为什么喜欢海明威[①]作品的客户会购买菲茨杰拉德[②]的书。但是这似乎并不重要，重要的是销量。最后，编辑们看到了销售额分析，亚马逊也不得不放弃每次的在线评论，最终，书评组被解散。林登回忆说：“书评团队被打败、被解散，我感到非常难过。但是，数据没有说谎，人工评论的成本是非常高的。”

如今，据说亚马逊销售额的 1/3 都来自于它的个性化推荐系统。有了它，亚马逊不仅使很多大型书店和音乐唱片商店歇业，而且当地数百个自认为有自己风格的书商也难免受转型之风的影响。

知道人们为什么对这些信息感兴趣可能是有用的，但这个问题目前并不是很重要，而知道“是什么”可以创造点击率，这种洞察力足以重塑很多行业，不仅仅只是电子商务。所有行业中的销售人员早就被告知，他们需要了解是什么让客户做出了选择，要把握客户做决定背后的真正原因，因此专业技能和多年的经验受到高度重视。大数据却显示，还有另外一个在某些方面更有用的方法。亚马逊的推荐系统梳理出了有趣的相关关系，但不知道背后的原因——知道是什么就够了，没必要知道为什么。

（本案例由作者根据相关资料改写）

阅读上文，请思考、分析并简单记录：

（1）你了解亚马逊等电商网站的推荐系统吗？请列举一个这样的实例（你选择购买什么商品，网站又给你推荐了其他什么商品）。

答：_____

_____

_____

（2）亚马逊书评组和林登推荐系统各自成功的基础是什么？

答：_____

_____

① 欧内斯特·米勒尔·海明威（1899 年 7 月 21 日—1961 年 7 月 2 日），美国小说家。被誉为美利坚民族的精神丰碑。出生于美国伊利诺伊州芝加哥市郊区的奥克帕克，代表作有《老人与海》《太阳照常升起》《永别了，武器》《丧钟为谁而鸣》等，凭借《老人与海》获得 1953 年普利策奖及 1954 年诺贝尔文学奖。

② 菲茨杰拉德，美国小说家。1920 年出版了长篇小说《人间天堂》，从此出名。1925 年《了不起的盖茨比》问世，奠定了他在现代美国文学史上的地位，成了 20 世纪 20 年代“爵士时代”的发言人和“迷惘的一代”的代表作家之一。

（3）为什么书评组最终输给了推荐系统？请阐述你的观点。

答：

（4）简单描述你所知道的上一周内发生的国际、国内或者身边的大事。

答：

1.1 大数据概述

信息社会所带来的好处是显而易见的：每个人口袋里都揣有一部手机，每台办公桌上都放着一台计算机，每间办公室内都连接到局域网甚至互联网。半个世纪以来，随着计算机技术全面和深度地融入社会生活，信息爆炸已经积累到了一个开始引发变革的程度。信息总量的变化导致了信息形态的变化——量变引起质变。最先经历信息爆炸的学科，如天文学和基因学，创造出了“大数据”（Big Data）这个概念。如今，这个概念几乎应用到所有人类致力于发展的领域中。

1.1.1 数据与信息

数据是反映客观事物属性的记录，是信息的具体表现形式。数据经过加工处理之后，就成为信息；而信息需要经过数字化转变成数据才能存储和传输。所以，数据和信息之间是相互联系的。

数据和信息也是有区别的。从信息论的观点来看，描述信源的数据是信息和数据冗余之和，即数据=信息+数据冗余。数据是数据采集时提供的，信息是从采集的数据中获取的有用信息，即信息可以简单地理解为数据中包含的有用的内容。

一个消息越不可预测，它所含的信息量就越大。事实上，信息的基本作用是消除人们对事物了解的不确定性。信息量是指从 N 个相等的可能事件中选出一个事件所需要的信息度量和含量。从这个定义看，信息量与概率是密切相关的。

1.1.2 天文学——信息爆炸的起源

综合观察社会各个方面的变化趋势，我们能真正意识到信息爆炸或者说大数据的时代已经到来。以天文学为例，2000 年斯隆数字巡天[①]项目（见图 1-2）启动时，位于新墨西哥州的望远镜在短短几周内搜集到的数据，就比世界天文学历史上总共搜集的数据还要多。截至 2010 年，信息档案已经高达 1.4×2^{42} B。不过，预计 2016 年底，在智利投入使用的大型视场全景巡天望远镜在 5 天之内即可获得同样多的信息。

① 斯隆数字巡天：位于新墨西哥州阿帕奇山顶天文台的 2.5 m 口径望远镜红移巡天项目。计划观测 25%的天空，获取超过一百万个天体的多色测光资料和光谱数据。2006 年，斯隆数字巡天进入了名为 SDSS-II 的新阶段，进一步探索银河系的结构和组成，而斯隆超新星巡天计划搜寻 I a 型超新星爆发，以测量宇宙学尺度上的距离。

天文学领域发生的变化在社会各个领域都在发生。2003 年，人类第一次破译人体基因密码时，辛苦工作了十年才完成了三十亿对碱基对的排序。大约十年之后，世界范围内的基因仪每 15 min 就可以完成同样的工作。在金融领域，美国股市每天的成交量高达 70 亿股，而其中 2/3 的交易都是由建立在数学模型和算法之上的计算机程序自动完成的，这些程序运用海量数据来预测利益和降低风险。

图 1-2　美国斯隆数字巡天望远镜

互联网公司更是要被数据淹没。谷歌公司每天要处理超过 24 拍字节（PB，2^{50}B）的数据，这意味着其每天的数据处理量是美国国家图书馆所有纸质出版物所含数据量的上千倍。

从科学研究到医疗保险，从银行业到互联网，各个不同的领域都在讲述着一个类似的故事，那就是爆发式增长的数据量。这种增长超过了人们创造机器的速度，甚至超过了人们的想象。人类存储信息量的增长速度比世界经济的增长速度快 4 倍，而计算机数据处理能力的增长速度则比世界经济的增长速度快 9 倍，每个人都受到了这种极速发展的冲击。

以纳米技术为例。纳米技术专注于把东西变小而不是变大。其原理就是当事物到达分子级别时，它的物理性质就会发生改变。一旦知道这些新的性质，就可以用同样的原料做以前无法做的事情。铜本来是用来导电的物质，但它一旦到达纳米级别就不能在磁场中导电了。银离子具有抗菌性，但当它以分子形式存在时，这种性质就会消失。一旦到达纳米级别，金属可以变得柔软，陶土可以具有弹性。同样，当人们增加所利用的数据量时，也就可以做很多在小数据量的基础上无法完成的事情。

大数据的科学价值和社会价值正是体现在这里。一方面，对大数据的掌握程度可以转化为经济价值的来源。另一方面，大数据已经撼动了世界的方方面面，从商业科技到医疗、政府、教育、经济、人文以及社会的其他各个领域。尽管人们还处在大数据时代的初期，但人们的日常生活已经离不开它。

1.1.3　大数据的定义

所谓大数据，狭义上可以定义为：用现有的一般技术难以管理的大量数据的集合。对大量数据进行分析，并从中获得有用观点，这种做法在一部分研究机构和大企业中早已存在。现在的大数据和过去相比，主要有 3 点区别：第一，随着社交媒体和传感器网络等的发展，人们身边正产生出大量且多样的数据；第二，随着硬件和软件技术的发展，数据的存储、处理成本大幅下降；第三，随着云计算的兴起，大数据的存储、处理环境已经没有必要自行搭建。

所谓“用现有的一般技术难以管理”，是指用目前在企业数据库占据主流地位的关系型数据库无法进行管理的、具有复杂结构的数据。或者也可以说，是指由于数据量的增大，导致对数据的查询（Query）响应时间超出允许范围的庞大数据。

研究机构 Gartner 给出了这样的定义：“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

麦肯锡[①]说："大数据指的是所涉及的数据集规模已经超过了传统数据库软件获取、存储、处理和分析的能力。这是一个被故意设计成主观性的定义，并且是一个关于多大的数据集才能被认为是大数据的可变定义，即并不定义大于一个特定数字的TB才叫大数据。因为随着技术的不断发展，符合大数据标准的数据集容量也会增长；并且定义随不同的行业也有变化，这依赖于在一个特定行业通常使用何种软件和数据集有多大。因此，大数据在今天不同行业中的范围可以从几十TB到几PB。"

随着"大数据"的出现，数据仓库、数据安全、数据分析、数据挖掘等围绕大数据商业价值的利用正逐渐成为行业人士争相追捧的利润焦点，在全球引领了又一轮数据技术革新的浪潮。

1.1.4 用3V描述大数据特征

从字面来看，"大数据"这个词可能会让人觉得只是容量非常大的数据集合而已。但容量只不过是大数据特征的一个方面，如果只拘泥于数据量，就无法深入理解当前围绕大数据所进行的讨论。因为"用现有的一般技术难以管理"这样的状况，并不仅仅是由于数据量增大这一个因素所造成的。

IBM说："可以用3个特征相结合来定义大数据：数量（Volume，或称容量）、种类（Variety，或称多样性）和速度（Velocity），或者就是简单的3V，即庞大容量、极快速度和种类丰富的数据"如图1-3所示。

1. Volume（数量）

用现有技术无法管理的数据量，从现状来看，基本上是指从几十TB到几PB这样的数量级。当然，随着技术的进步，这个数值也会不断变化。

如今，存储的数据数量正在急剧增长中，人们存储所有事物包括：环境数据、财务数据、医疗数据、监控数据等。有关数据量的对话已从TB级别转向PB级别，并且不可避免地会转向ZB级别。但是，随着可供企业使用的数据量不断增长，可处理、理解和分析的数据的比例却不断下降。

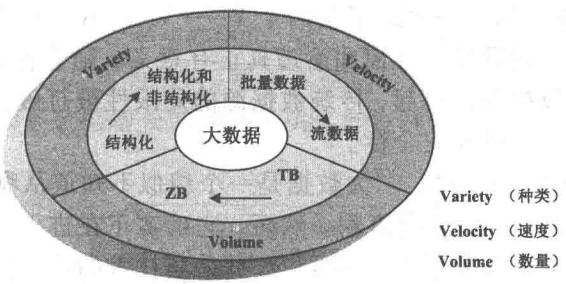


图1-3 按数量、种类和速度来定义大数据

① 麦肯锡公司：是世界级领先的全球管理咨询公司。自1926年成立以来，公司的使命就是帮助领先的企业机构实现显著、持久的经营业绩改善，打造能够吸引、培育和激励杰出人才的优秀组织机构。

麦肯锡在全球52个国家有94个分公司。在过去十年中，麦肯锡在中国地区完成了800多个项目，涉及公司整体与业务单元战略、企业金融、营销/销售与渠道、组织架构、制造/采购/供应链、技术、产品研发等领域。

麦肯锡的经验是：关键是找那些企业的领导们，他们能够认识到公司必须不断变革以适应环境变化，并且愿意接受外部的建议，这些建议在帮助他们决定做何种变革和怎样变革方面大有裨益。

2. Variety（种类、多样性）

随着传感器、智能设备以及社交协作技术的激增，企业的数据也变得更加复杂，因为它不仅包含传统的关系型数据，还包含来自网页、互联网日志文件（包括单击流数据）、搜索索引、社交媒体论坛、电子邮件、文档、主动和被动系统的传感器数据等原始、半结构化和非结构化数据。

种类表示所有的数据类型。其中，爆发式增长的一些数据，如互联网上的文本数据、位置信息、传感器数据、视频等，用企业中主流的关系型数据库是很难存储的，它们都属于非结构化数据。

当然，在这些数据中，有一些是过去就一直存在并保存下来的。和过去不同的是，除了存储，还需要对这些大数据进行分析，并从中获得有用的信息，例如监控摄像机中的视频数据。近年来，超市、便利店等零售企业几乎都配备了监控摄像机，最初目的是为了防范盗窃，但现在也出现了使用监控摄像机的视频数据来分析顾客购买行为的案例。

例如，美国高级文具制造商万宝龙（Montblane）过去是凭经验和直觉来决定商品陈列布局的，现在尝试利用监控摄像头对顾客在店内的行为进行分析。通过分析监控摄像机的数据，将最想卖出去的商品移动到最容易吸引顾客目光的位置，使得销售额提高了 20%。

3. Velocity（速度）

数据产生和更新的频率，也是衡量大数据的一个重要特征。就像搜集和存储的数据量和种类发生了变化一样，生成和需要处理数据的速度也在变化。不要将速度的概念限定为与数据存储相关的增长速率，应动态地将此定义应用到数据，即数据流动的速度。有效处理大数据需要在数据变化的过程中对它的数量和种类进行分析，而不只是在它静止后执行分析。

例如，遍布全国的便利店在 24 h 内产生的 POS 机数据、电商网站中由用户访问所产生的网站点击流数据、高峰时达到每秒近万条的微信短文、全国公路上安装的交通堵塞探测传感器和路面状况传感器（可检测结冰、积雪等路面状态）等，每天都在产生着庞大的数据。

IBM 在 3V 的基础上又归纳总结了第四个 V——Veracity（真实和准确）。只有真实而准确的数据才能让对数据的管控和治理真正有意义。随着社交数据、企业内容、交易与应用数据等新数据源的兴起，传统数据源的局限性被打破，企业愈发需要有效的信息治理以确保其真实性及安全性。

IDC（互联网数据中心）说："大数据是一个貌似不知道从哪里冒出来的大的动力。但实际上，大数据并不是新生事物。然而，它确实正在进入主流，并得到重大关注，这是有原因的。廉价的存储、传感器和数据采集技术的快速发展、通过云和虚拟化存储设施增加的信息链路，以及创新软件和分析工具，正在驱动着大数据。大数据不是一个'事物'，而是一个跨多个信息技术领域的动力/活动。大数据技术描述了新一代的技术和架构，其被设计用于：通过使用高速（Velocity）的采集、发现和/或分析，从超大容量（Volume）的多样（Variety）数据中经济地提取价值（Value）。"

这个定义除了揭示大数据传统的 3V 基本特征，还增添了一个新特征：Value（价值）。总之，大数据是个动态的定义，不同行业根据其应用的不同有着不同的理解，