杨方旭◇著

大数据时代背景下
大学生思想政治教育新思路

DASHUJU SHIDAI BEIJINGXIA
DAXUESHENG SIXIANG
ZHENGZHI JIAOYU XINSILU

NORTHEAST NORMAL UNIVERSITY PRESS
WWW.NENUP.COM
东北师范大学出版社

大数据时代背景下
大学生思想政治教育新思路

杨方旭◇著

NORTHEAST NORMAL UNIVERSITY PRESS
WWW.NENUP.COM
东北师范大学出版社

图书在版编目(CIP)数据

大数据时代背景下大学生思想政治教育新思路 / 杨方旭著. -- 长春：东北师范大学出版社，2018.1
ISBN 978-7-5681-4059-1

Ⅰ.①大… Ⅱ.①杨… Ⅲ.①信息技术—应用—大学生—思想政治教育—研究—中国 Ⅳ.①G641-39

中国版本图书馆 CIP 数据核字 (2017) 第 318025 号

□策划编辑：王春彦
□责任编辑：徐小红　刘子齐　□封面设计：优盛文化
□责任校对：张　琪　□责任印制：张允豪

东北师范大学出版社出版发行
长春市净月经济开发区金宝街 118 号（邮政编码：130117）
销售热线：0431-84568036
传真：0431-84568036
网址：http://www.nenup.com
电子函件：sdcbs@mail.jl.cn
河北优盛文化传播有限公司装帧排版
北京一鑫印务有限责任公司
2018 年 4 月第 1 版　2018 年 4 月第 1 次印刷
幅画尺寸：170mm×240mm　印张：13.25　字数：243 千

定价：48.00 元

前言

随着互联网技术的迅猛发展，如今“大数据”正在掀起一场前所未有的信息革命，伴随着数据技术的飞速发展，客观数据作为分析问题的手段，其重要性不言而喻。数据成为21世纪人类工作、生活的得力助手，它的发展深刻影响着人们生活的各个领域。

作为教育重要的一环，传统的大学生思想政治教育受到大数据的猛烈冲击，面临着全新的发展方向和发展格局。作为知识创新的主要阵地，高校更应该花大力气，下大功夫去采集、汇聚与合理利用大数据的相关技术。面对当今的高校大学生，高校思想政治教育者应该高瞻远瞩，充分利用大数据技术带来的机遇，将大学生思想政治教育的主体、客体、教育效果的方方面面进行革新，以便将高校思想政治工作落到实处。

随着信息技术的迅猛发展，大数据时代已经到来并开始影响人类社会的众多领域。对思想政治教育而言，大数据时代更是不可抗拒也无法逃避的新环境。大数据时代的到来为我国大学生思想政治教育工作带来新的契机，促进了对大学生大数据收集、挖掘和分析的积极性。这不仅有利于观察大学生群体的总体特征，实现大学生行为的预警及预测，还有助于探究个体大学生的偏好以及习惯，实现大学生思想政治教育的个性化。如何深刻理解大数据思想政治教育的内涵，把握其对传统的思想政治教育带来的挑战，探索网络大数据背景下高校大数据思想政治教育发展新模式，是现阶段辅导员队伍开展思想政治教育理论研究与实践的重要课题。

目录

此为试读，需要完整PDF请访问：www.ertongbook.com

第一章　大数据时代的来临

第一节　大数据的概念

一、大数据的定义

大数据这个概念是由最先经历信息爆炸的学科，如天文学和基因学创造出来的。如今这个概念已经应用到了几乎所有人类致力于发展的领域中。

大数据并非一个确切的概念。最初，这个概念是指需要处理的信息量过大，已经超出了一般电脑在处理数据时所能使用的内存量，因此工程师们必须改进处理数据的工具。

大数据不仅包含数据的体量，而且强调数据的处理速度。在数据分析领域，大数据是前沿技术，大数据以及数据仓库、数据分析、数据安全、数据挖掘是IT行业时下最火爆的词汇，大数据的商业价值已经成为信息行业争相追逐的焦点。大数据包括各种互联网信息，更包括各种交通工具、生产设备和工业器材上的传感器，随时随地进行测量，不间断地传递着海量的信息数据。利用新处理模式，大数据具有更强的决策力和洞察力，能够优化流程，实现高增长率，处理海量的多样化信息资产。归根结底，大数据技术可以快速处理不同种类的数据，从中获得有价值的信息。

随着网络、传感器和服务器等硬件设施的全面发展，大数据技术促使众多企业

融合自身需求，创造出难以想象的经济效益，实现巨大的社会价值。各行各业利用大数据产生极大增值和效益，表现出前所未有的社会能力，而绝不只是数据本身。所以，大数据可以定义为在合理时间内采集大规模资料，帮助使用者更有效决策的社会过程。

在今天，大数据被认为是一种人们在大规模数据的基础上可以做到的事情，大数据是人们获得新的认知、创造新的价值的源泉，大数据还为改变各种关系服务。

二、大数据的本质

从人类认识史可以发现，对信息的认识史就是人类的认识进步史与实践发展史。人类历史上经历过四次信息革命。第一次是创造语言，语言是即时变换和传递信息的工具，人类通过语言建立相互关系认识世界。语言表明人类要求表达、认识世界并开始作用于世界，通过语言产生思维，将事物的信息抽象表达为声音这个即时载体，但语言的限制和缺点是无法突破个体的时空。第二次是创造文字以及随之而来的造纸与印刷的技术，实现了人类远距离和跨时空的思想传递，人类因此扩大联合。文字虽然突破了时间空间上的限制，但需要耗费太高的交流成本和传播成本。第三次是发明电信通信，电报、广播、电视实现了文字、声音和图像信息的远距离即时传递，为电子计算机与互联网创造奠定了基础。第四次是电子计算机与互联网的创造，是一次空前的伟大综合，其特点是所有信息全部归结为数据，表达形式为数字形式，只要有了 0 和 1 加上逻辑关系就可以构成全部世界。现代通信技术和电子计算机的有效结合，使信息的传递速度和处理速度得到了巨大的提高，人类掌握信息、利用信息的能力达到了空前的高度，人类社会进入了信息社会。在一定意义上，人类文明史是一部信息技术的发展进化历史。

（一）信息

从本体论层次，信息可定义为事物的存在方式和运动状态表现形式。事物泛指存在于人类社会、思维活动和自然界中一切可能的对象，存在方式指事物的内部结构和外部联系，运动状态指事物在时空变化的特征和规律。从认识论层次看，信息是主体所感知或表述的事物存在的方式和运动状态。主体所感知的是外部世界向主体输入的信息，主体所表述的则是主体向外部世界输出的信息。

（二）数据

数据就是指能够客观反映事实的数字和资料，可定义为用意义的实体，它涉及事物的存在形式，是表达知识的字符集合。按性质可分为表示事物属性的定性数据和反映事物数量特征的定量数据。按表现形式可分为数字数据和模拟数据，模拟数

据又可以分为符号数据、文字数据、图形数据和图像数据等。

数据在计算机领域是指可以输入电子计算机的一切字母、数字、符号，具有一定意义，能够被程序处理，是信息系统的组成要素。数据可以记录或传输，并通过外围设备在物理介质上被计算机接受，经过处理而得到结果。计算机系统的每个操作都要处理数据，通过转换、检索、归并、计算、制表和模拟等操作，经过解释并赋予一定的意义之后便成为信息，可以得到人们需要的结果。分析数据中包含的主要特征，就是对数据进行分类、采集、录入、储存、统计检验、统计分析等一系列活动，接收并且解读数据才能获取信息。

（三）数据与信息

数据是信息的载体，信息是有背景的数据，而知识是经过人类的归纳和整理，最终呈现规律的信息。但进入信息时代之后，“数据”二字的内涵开始扩大：不仅指“有根据的数字”，还统指一切保存在电脑中的信息，包括文本、图片、视频等。其中的原因是，20 世纪 60 年代软件科学取得了巨大进步、发明了数据库。此后，数字、文本、图片都不加区分地保存在电脑的数据库中，数据也逐渐成为“数字、文本、图片、视频”等的统称，也即“信息”的代名词。

简单地说，信息是经过加工的数据，或者说，信息是数据处理的结果。信息与数据是不可分离的，数据是信息的表现形式，信息是数据的内涵。数据本身并没有意义，数据只有对实体行为产生影响时才成为信息。信息可以离开信息系统而独立存在，也可以离开信息系统的各个组成和阶段而独立存在；而数据的格式往往与计算机系统有关，并随载荷它的物理设备的形式而改变。大数据可以被看作是依靠信息技术支持的信息群。

三、大数据的分类

（一）依据来源不同

一般分为四类：科研数据、互联网数据、感知数据和企业数据。

1. 科研数据

科研数据在大数据时代前很久就存在，可能来自生物工程、天文望远镜或粒子对撞机，不一而足。这些数据存在于封闭系统中，使用者都是传统上做高性能计算（HPC）的企业，很多大数据技术脱胎于 HPC。

科研数据存在于拥有具有极高计算速度且性能优越的机器的研究机构，包括生物工程研究以及粒子对撞机或天文望远镜，如位于欧洲的国际核子研究中心装备的大型强子对撞机，在其满负荷的工作状态下每秒可以产生 PB 级的数据。

2. 互联网数据

互联网大数据是时代的主流，尤其社交媒体是近年来大数据的主要来源，几乎所有的大数据技术都源于快速发展的国际互联网企业。比如，以搜索著称的百度与谷歌的数据规模都已经达到上千 PB 的级别，而应用广泛影响巨大的脸谱、亚马孙、雅虎、阿里巴巴的数据都突破上百 PB。互联网数据增长的驱动力一是梅特卡夫定律（互联网企业的价值与用户数的平方成正比），二是扎克伯格反复引用的信息分享理论：一个人分享的信息每一到两年翻番。

大型互联网企业的大数据生态系统比较独特，一方面不同程度上参与开源，一方面维护自给自足的生态系统，甚至连硬件都越来越依靠自己了。从谷歌开始，后有“脸书”的“开放计算项目”。大型互联网公司不只是自身产生大体量数据，它还有平台级的带动作用，如“脸书”之于 Zynga，阿里牵头做的数据交换平台。中型互联网公司基本上也能够维持大数据技术团队，只不过与大型互联网公司的核心开发能力和社区贡献能力相比，它们更多部重兵在外围开发、优化和运维。当然，它们多少会有一些绝招，如豆瓣的推荐，暴风的 Hadoop 管理。三线互联网公司有数据但没有大数据能力，这催生了一些大数据技术和服务的机会，如百分点为电商网站做个性化推荐和营销分析，各类广告联盟、移动应用服务平台为网站和移动应用提供统计分析、营销服务等。

3. 感知数据

进入移动互联网时代后，移动平台的感知功能和 LBS 的普及，感知数据基于位置的服务和移动平台的感知功能，与互联网数据逐渐重叠，但感知数据的体量同样惊人，并且总量可能不亚于社交媒体。Teradata 预测感知数据的总量在 2015 年超过社交媒体，并达到后者的 10 ~ 20 倍。重庆曾计划做一个平安城市项目，规划了 50 万个摄像头，数据存储需求要达到百 PB 级别，不亚于世界级的互联网公司。

4. 企业数据

企业数据种类繁杂，企业数据和感知数据本质上并不是 MECE（不重复、不遗漏）的划分，企业同样可以通过物联网收集大量的感知数据，之所以把它们分为两类，是传统上认为企业数据是人产生的，感知数据是物、传感器、标识等机器产生的。企业外部数据日益吸纳社交媒体数据，内部数据不仅有结构化数据，更多是越来越多的非结构化数据，由早期电子邮件和文档文本等扩展到社交媒体与感知数据，包括多种多样的音频、视频、图片、模拟信号等。

可以把企业数据和感知数据放在一起讲，因为它们都涉及传统产业，从经济总量上要比互联网产业大很多，而且传统产业自身的大数据能力有限，所以这是大数

据技术和服务企业的主要目标市场。但目前的现实是，就单个企业而言，具有大数据需求的并不多见。比如，麦肯锡的报告中把制造业列为大数据存量最多的行业，但很少有制造企业上马大数据项目。即使有，如 Zara，只是在市场营销中加入了互联网，以获得来自终端的需求，供应链和生产这方面相比大数据之前没有太多新意。通过数据采集和分析来提升制造业的效率，会是个很大的市场，这是工业物联网，但未必是大数据。

（二）从社会宏观角度

根据其使用主体可分为三类：政府的大数据、企业的大数据、个人的大数据。

1. 政府的大数据

各级政府各个机构拥有海量的原始数据，构成社会发展与运行的基础，包括形形色色的环保、气象、电力等生活数据，道路交通、自来水、住房等公共数据，安全、海关、旅游等管理数据，教育、医疗、信用及金融等服务数据。在具体的政府单一部门里面无数数据固化而没有产生任何价值，如果关联这些数据并使其流动起来综合分析有效管理，这些数据将产生巨大的社会价值和经济效益。

现代城市依托网络智能走向智慧，无论智能电网与智慧医疗，还是智能交通和智慧环保，都离不开大数据的支撑，大数据是智慧城市的核心资本。建设智慧城市，大数据可以在方方面面提供各种决策与智力支持。政府作为国家的管理者应该将数据逐步开放，供给更多有能力的机构组织或个人来分析并加以利用以加速造福人类。奥巴马任期内的一个重要举措是美国政府筹建了一个 data.gov 网站，要求政府公开透明，核心就是政府机构的数据公开。

2. 企业的大数据

企业离不开数据支持有效决策，企业在大数据的帮助下才能为快速膨胀的消费者群体提供差异化的产品或服务，实现精准营销。网络企业应该依靠大数据实现服务升级与方向转型，传统企业面临无处不在的互联网压力，同样必须谋求变革实现融合不断前进。

随着信息技术的发展，数据成为企业的核心资产和基本要素，数据变成产业进而成长为供应链模式，慢慢连接为贯通的数据供应链。互联网时代，互相自由连通的外部数据的重要性逐渐超过单一的内部数据，企业个体的内部数据更是难以和整个互联网数据相提并论。综合提供数据、推动数据应用、整合数据加工的新型公司明显具有竞争优势。大数据时代产生影响巨大的互联网企业，而传统 IT 公司随着网络社会的到来开始进入互联网领域，需要云计算与大数据技术改善产品、提升平台、实现升级，这两类公司互相借鉴，相互合作，彼此竞争。

此为试读，需要完整PDF请访问：www.ertongbook.com

3. 个人的大数据

每人都能通过互联网建立属于自己的信息中心，积累、记录、采集、储存个人的一切大数据信息。根据相关法律规定，经过本人亲自授权，所有个人相关信息将转化为有价值的数据，被第三方采集可以快速处理，获得个性化的数据服务。通过信息技术使得各种可穿戴设备，包括植入的各种芯片都可以通过感知技术获得个人的大数据，包括但不限于体温、心率、视力各类身体数据以及社会关系、地理位置、购物活动等各类社会数据。个人可以选择将身体数据授权提供给医疗服务机构，以便监测出当前的身体状况，制订私人健康计划；还能把个人金融数据授权给专业的金融理财机构，以便制定相应的理财规划并预测收益。当然，国家有关部门还会在法律范围内经过严格程序进行预防监控，实时监控公共安全，预防犯罪。

个人的大数据严格受到法律保护，其他第三方机构必须按法律规定授权使用，数据必须接受公开透明全面监管；采集个人数据应该明确按照国家立法要求，由用户自己决定采集内容与范围；数据只能由用户明确授权才能严格处理。

互联网上的大数据不容易分类，百度把数据分为用户搜索产生的需求数据以及通过公共网络获取的数据；阿里巴巴则根据商业价值将其分为交易数据、社交数据、信用数据和移动数据；腾讯善于挖掘用户关系数据并且在此基础上生成社交数据。通过数据分析人们的许多想法和行为，可以从中发现政治治理、文化活动、社会行为、商业发展、身体健康等各个领域的各种信息，进而可以预测未来。互联网大数据可以分为互联网金融数据以及用户消费产生的行为、地理位置以及社交等大量数据。

四、大数据的技术

大数据技术包括大数据科学、大数据工程和大数据应用。大数据工程指通过规划建设大数据并进行运营管理的整个系统；大数据科学指在大数据网络的快速发展和运营过程中寻找规律，验证大数据与社会活动之间的复杂关系。大数据需要有效地处理大量数据，包括大规模并行处理数据库、分布式文件系统、数据挖掘电网、云计算平台、分布式数据库、互联网和可扩展的存储系统。大量非结构化数据通过关系型数据库处理分析需要大量时间和金钱，大型数据集分析需要大量电脑持续高效地分配工作。大数据分析常和云计算联系到一起，大数据分析相比传统的数据仓库数据量大、查询分析复杂。大数据处理和存储技术源于军事需求，二战期间英国研发了能处理大规模数据的机器，二战后美国致力于数字化处理搜集得到的大量情报信息。“9·11”事件后美国政府在大数据挖掘领域组建了大数据库用于识别可疑

人，通过筛选通信、教育、犯罪、医疗、金融和旅行等记录，之后组建基于网络的信息共享系统。大规模数据分析技术源于社交网络，大数据应用使人们的思维不局限于数据处理机器，重要的是新用途和新见解。对大规模信息的处理需求从根本上推动了大数据相关技术的发展，超级计算机的发明、大数据的存储和处理技术以及大数据分析算法的研发最终导致了大数据教育、金融、医疗等多方面的广泛应用。

第二节　大数据的特征

一、体量巨大，种类繁多

互联网搜索的发展、电子商务交易平台的覆盖和微博等社交网站的兴起，产生了无穷无尽的各种数据内容。传感、存储和网络等计算机科学领域在不断前行，人们在不同领域采集到的数据量达到了前所未有的程度，收集大量数据的原因在于网络数据可以实现同步实时收集，包括电子商务、传感器、智能手机等，还有医疗领域的临床数据和科学研究如基因组研究将 GB 级乃至 TB 级数据输送到数据库。由于占 85% 以上的非结构化数据的增长，数据总量的增速比结构化数据快大概几十倍。数据类型日益繁多，如视频、文字、图片、符号等，发掘这些形态各不相同的数据流之间的相关性是大数据的最大优点。比如，供水系统数据与交通状况比较可以发现清晨洗浴和早高峰的时间密切相关，电网运行数据和堵车时间地点有相关性，交通事故率关联睡眠质量等。

二、开放公开，容易获取

采集大数据不是为了存储而是为了进行分析。大数据不仅存在于特定的政府机构和企业组织，而且在社会生活生产过程中自动产生并存储。电信公司积累客户的电话沟通记录，电子商务网站整合消费者的各种信息，企业通过挖掘海量数据可以增强自身能力，改善运营服务，提供决策支持，实现商业智能进而为企业带来高额经济效益回报，发现企业发展的特殊规律。今天，在一定规则开放性下，依靠应用程序接口技术和爬虫采集技术，越来越多的商业组织和政府机构开始向社会各界和研究机构提供自身采集储存的各种海量数据源，尤其是美国政府走在前列，主动提供具有权威的开放数据源 data.gov 等开源数据。公开容易取得的数据源成为大数据时代的基本特征，产生了巨大的社会影响。

三、重视社会预测

预测是大数据的本质特征。在大数据时代，预见行业未来的能力成为企业追求的目标。最近美国 Netflix 公司推出的《纸牌屋》，即通过采集其 3 000 万用户的播放动作，包括打开、暂停、快进、倒退等动作，分析其注册用户的几百万次评级与搜索，评价受众对不同电视电影节目给予的不同观点，从导演、演员、题材、情节、类型等各个方面了解公众欣赏节目的习惯，通过挖掘海量数据，获得人们的喜好。该公司改变了视频行业的制作方式，用计算方法和逻辑分析替代了以前过时的生产方式，通过大数据能先于受众分析需求，制作节目获得关注。更有意思的案例是，商场居然比父亲更早得知未成年女儿的怀孕信息，商家依据客户的购物行为进而通过大数据分析预测到其有很大的怀孕可能性。人们极为关注大数据预知社会问题的应用功能，在社会科学领域大数据将发挥越来越突出的巨大作用。

四、重视发现而非实证

实证研究强调建立理论假设，设定范围随机抽样，定量调查采集数据，收集相关数据，进而证伪或证实理论假设。大数据则重视数据，创造知识、预测前景、探索未知、关注现象、发现机遇。大数据预见未来依靠的是自下而上的数据收集处理，在不依赖理论假设的前提下去发现知识，洞察趋势，找到规律。例如，沃尔玛超市经过大数据技术分析海量交易数据，发现周末男人买婴儿尿布的同时会顺便买啤酒的独特现象。通常数据挖掘不做刻板假设，具有未知性，但结果有效并且实用，还有就是重视全体忽略抽样。大数据是信息技术自动采集存储的海量数据，可以进行快速分析处理得到结果。随着存储设备成本的不断下降，计算机工具效能日趋先进，处理海量数据的能力快速提升，数据挖掘算法持续加速改进，尤其是机器学习的神经网络建模技术使得抽样调查不再是唯一的方法。大数据理论上可以把握总体数据，更加重视整体的全部数据。

五、非结构化数据的涌现

数据挖掘重视未知的有效信息和实用知识。非结构化数据越来越多，这成为大数据时代的突出特征，现在超过 90% 的数据都是非结构化数据。社交媒体尤其是微博随时产生的无数数据文本，导致有价值的数据隐藏在海量信息中，大数据分析技术从大量文本中挖掘探析人们的态度和行为，呼应舆情监测的社会需求和企业的重

大商机。面对非结构化的大数据采集处理，社会产生了新的需求，技术发生了新的变革，各种 Hadoop 集群、NoSQL 以及 Map Reduce 等非关系型数据库流行，IT 新技术不断涌现。大数据包括数据挖掘、网络挖掘、文本挖掘、机器学习和 NLP 自然语言处理等 IT 和商业智能信息技术和决策支持系统及其在社会科学领域的应用。

第三节　大数据的发展历程

一、数据的“产生及发展”

人类在生产实践中，发明了语言、文字和图形，但仅用这些还无法准确地描述世界，数字作为一项重要的改造世界的工具而产生了。它把抽象的概念具体表达，如“很多”人、“非常”多人可以理解为不同的程度，但如果说 1 千人、1 万人就清清楚楚了。人类的生产、交换等活动都是以数据为基础展开的，如度量衡、货币等的发明和出现，大大地推动了人类文明的发展。

数据的测量产生了最早“有根据的数字”，即数据是对客观世界测量结果的记录，不是随意产生的。测量从一开始产生就是为科学服务的，从古至今，测量都是科学的主要手段，其重要性可以描述为：没有测量，就没有科学。测量得到的数据可以由计算再衍生出新数据，这样看来，一切数据都是人为的产物。但这时的数据还只是传统意义上的数据，它和信息、知识是有严格区别的：数据是信息的载体，信息是数据的背景，知识是经过归纳整理后呈现出来的有规律的信息。

进入信息时代后，巨大的变化产生了，20 世纪 60 年代，软件科学发展，数据库被发明，电脑的数据库用来存储一切数字、文本、图片。这时，数据开始不仅指“有根据的数字”，其内涵扩大到一切保存在电脑中的信息，包括文本、图片、视频等。数据也成了信息的代名词，因为这些信息只是一种对世界的记录，数据因此多了一个来源：记录。

数据库出现以后，信息总量与日俱增，增速也越来越快。20 世纪 90 年代，就有美国人提出了“大数据”概念，这时候还不是真正的大数据时代，这时候数据的重要性在上升，在价值上的重要性已经被预见。21 世纪开始，特别是 2004 年新社交媒体产生以后，数据开始爆炸，大数据的提法又一次出现，这时的大数据既指容量大又指价值大。争议开始了：到底什么算大？多大才是真正的大？

二、大数据的发展

有史以来，处理各种不断增长的数据都是人类社会的难题。

大数据的现代发展历史最早可追溯到美国统计学家赫尔曼·霍尔瑞斯，他被后世称为“数据自动处理之父”。他发明了一台电动“打孔卡片制表机”对卡片特定位置上的孔洞进行识别，并加以自动统计。这一发明被运用在统计1890年的人口普查数据，该机器用两年半时间就完成了预计耗时十三年的工作，这个惊人的速度就是全球进行数据自动处理的新起点。

1943年二战期间，英国为了快速解开纳粹设置的密码，组织工程师发明机器进行大规模数据处理，并采用了第一台可编程的电子计算机实施计算工作。该计算机被命名为“巨人”，为了找出拦截信息中的潜在模式，它以每秒钟5 000字符的速度读取纸卡——将原本需要耗费数周时间才能完成的工作量压缩到了几个小时。

1961年，美国国家安全局——一个刚成立9年就拥有超过12 000个密码学家的情报机构，在间谍饱和的冷战年代，面对超量信息，开始采用计算机自动收集信号处理情报，并努力将仓库内积压的模拟磁盘信息进行数字化处理，仅1961年7月份，该机构就收到了17 000卷磁带。

20世纪40年代以来，人们梦想能拥有一个世界性的信息库。在这个信息库中，信息不仅能被全球的人们存取，而且能轻松地链接到其他地方的信息，使用户可以方便快捷地获得重要的信息。60年代，英国计算机科学家蒂姆·伯纳斯·李发明了一个全球网络资源唯一认证的系统：统一资源标识符。他设计了超文本系统，在这个系统中，每个有用的事物称为一样“资源”，并且由一个全局“统一资源标识符”标识。然后，将超文本嫁接到因特网上，命名为万维网，这些资源通过超文本传输协议传送给用户，用户通过点击链接来获得资源。通过互联网在世界范围内实现了信息共享。

1965年，英特尔的创始人戈登·摩尔通过研究计算机硬件的发展规律得出摩尔定律，该定律认为，同等面积的芯片每过一到两年就可容纳两倍数量的晶体管，能够提高两倍微处理器的性能，或使之价格下降一半。近五十年，硬件的发展基本符合这一定律，到今天，一根头发尖大小的地方就能放上万个晶体管。后来，英特尔公司又发明了22纳米的3D晶体管，比以前的晶体管小了大约1/3，摩尔定律的生命进一步得到延续，导致信息产品功能日趋强大，各种设备体积变小，存储器成本持续缩小了1亿多倍，能以很低的成本保存海量的数据。摩尔定律已经成为描述一切呈指数级增长的事物的代名词，这为大数据时代的到来铺平了硬件道路，打下了物质基础。

除了便宜、功能强大，摩尔定律使计算设备也变得越来越小。1988 年，美国科学家马克·韦泽指出各种各样微型计算设备能随时随地获取并处理数据，这被称为普适计算。普适计算理论指出，计算机发明以后经历三个阶段的发展：一是主机型阶段，一台占据大半个房间的大型机器被很多人共享；二是个人电脑阶段，每个人拥有一台变小了的个人电脑；三是计算机越来越小，甚至从人们视线中消失，日常环境中可以被广泛地部署各种微小计算设备，任何时间地点都可以获得并处理数据，计算融入环境中，即进入普适计算阶段。今天，小巧的智能手机、各种传感器、RFID（射频识别）标签、可穿戴式设备等广泛使用，实现无处不在的数据自动采集，人类收集数据的能力增强，为大数据时代的到来提供了物理基础。

1989 年，英国计算机协会下属的数据挖掘及知识发现专委会举办了第一届数据挖掘学术年会，出版了专门期刊，这是大数据时代一个最重要的里程碑，此后数据挖掘得到了如火如荼的发展。数据挖掘是指通过特定的算法对大量的数据进行自动分析，从而揭示数据当中隐藏的规律和趋势，即在大量的数据当中发现新知识，为决策者提供参考。数据挖掘进步的根本原因是人类能够不断设计出更强大的模式识别算法，这其实是软件的进步。现在的信息技术已经可以把一件产品的流向、每位消费者的情况都记录下来，再通过数据挖掘，为客户量身定制，把消费和服务推向一个高度个性化的时代。基于网络数据的挖掘，不需要制定问卷，也不需要逐一调查，成本低廉。更重要的是，这种分析是实时的，没有滞后性，数据挖掘将成为越来越重要的分析预测工具，抽样技术将下降为辅助工具。数据挖掘的优越性，也集中反映了大数据“量大、多源、实时”的三个特点。

大数据的前沿和热点是机器学习，和数据挖掘相比，其算法并不是固定的，而是带有自调适参数的。也就是说，它能够随着计算、挖掘次数的增多，不断自动调整自己算法的参数，使挖掘和预测的结果更为准确，即通过给机器“喂取”大量的数据，让机器可以像人一样通过学习逐步自我改善、提高，这也是该技术被命名为“机器学习”的原因。除了数据挖掘和机器学习，数据的分析、使用技术已经非常成熟，并且形成了一个谱系，如数据仓库、多维联机分析处理、数据可视化、内存分析都是其体系的重要组成部分。

美国研究员大卫·埃尔斯沃斯和迈克尔·考克斯在 1997 年使用“大数据”来描述超级计算机产生超出主存储器的海量信息，数据集甚至突破远程磁盘的承载能力。

2004 年之前，互联网的主要作用是传播和分享信息，其最主要的组织形式是建立静态的网站；从 2004 年起，以脸谱网（“脸书”）、推特（Twitter）为代表的社交媒体相继问世，一个互联网的崭新时代开始了。互联网开始成为人们实时互动、交

此为试读，需要完整PDF请访问：www.ertongbook.com