# 大数据背景下高校教育管理信息化发展与创新研究

林 榕 著



吉林大学出版社

# 大数据背景下高校教育管理信息化发展与创新研究

◎ 林 榕 著

吉林大学出版社

**图书在版编目(CIP)数据**

大数据背景下高校教育管理信息化发展与创新研究 / 林榕著.—长春 : 吉林大学出版社, 2018.12
ISBN 978-7-5692-4056-6

Ⅰ. ①大… Ⅱ. ①林… Ⅲ. ①高等教育—教育管理—信息化—研究—中国 Ⅳ. ① G649.2

中国版本图书馆 CIP 数据核字 (2019) 第 008633 号

**书　　名：大数据背景下高校教育管理信息化发展与创新研究**
DASHUJU BEIJING XIA GAOXIAO JIAOYU GUANLI XINXIHUA FAZHAN YU CHUANGXIN YANJIU

作　　者：林　榕　著
策划编辑：邵宇彤
责任编辑：邵宇彤
责任校对：张瑞伦
装帧设计：优盛文化
出版发行：吉林大学出版社
社　　址：长春市人民大街 4059 号
邮政编码：130021
发行电话：0431-89580028/29/21
网　　址：http://www.jlup.com.cn
电子邮箱：jdcbs@jlu.edu.cn
印　　刷：三河市华晨印务有限公司
开　　本：170mm × 240mm　1/16
印　　张：10.75
字　　数：200 千字
版　　次：2019 年 3 月第 1 版
印　　次：2019 年 3 月第 1 次
书　　号：ISBN 978-7-5692-4056-6
定　　价：45.00 元

# 前　言

当下是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个时代的产物。大数据是对大量、动态、能持续的数据，运用新系统、新工具、新模型的挖掘，从而获得具有洞察力和新价值的信息。以前，面对庞大的数据，我们可能会一叶障目、只见一斑，不能了解到事物的真正本质，导致在科学工作中得到错误的推断，而大数据时代的来临，一切真相将会展现在大家面前。

教育管理在高校管理工作中处于核心地位，贯穿于教学工作的各个阶段，帮助实现资源最佳配置，科学安排教学计划，使教学工作得以顺利进行。高校教育管理工作可以发挥培养优秀人才的重要作用，促进高校发展，提高教育水平，在当今社会万众创新的发展格局下，高校教育管理的创新已成必然。

在知识经济时代，行业市场中的竞争等同于人力资源的竞争，而优秀人才的培养则要依赖于科学高效的教育。纵观我国高校的教育现状，虽然教育水平有了比较明显的提升，但是有关教育管理信息化的建设仍然存在很多问题。教育信息化最主要的目标是将信息技术巧妙地应用到高校的教育管理工作中，借助于网络的力量来丰富高校教育内容，以便更好地满足现代创新人才的需求。高校的教育管理水平想要达到新高度，就必须积极改进传统教育管理方法，不仅要合理应用先进大数据信息技术，而且还要整合一切现有资源，将其实际价值淋漓尽致地发挥出来。

# 目 录

# 第一章　大数据概述

## 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。大数据这个术语最早应用于 apacheorg 的开源项目 Nutch，用来表达批量处理或分析网络搜索索引产生的大量数据集。

谷歌公开发布 Map Reduce 和 Google File System（GFS）之后，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、工业器材上的传感器，随时随地进行测量，不间断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，并快速处理。只有快速才能发挥实际作用。

随着网络、传感器和服务器等硬件设施全面发展，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和商业价值。

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、处理大规模资料，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

观察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。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

哈特莱（R.V.Hartley）1928 年在《信息传输》一文中对信息的定义是有新内容、新知识的消息。信息的奠基人香农（C.E.Shannon）1948 年认为信息是消除随机不确定性、是肯定性的确认和确定性的增加，并提出信息量的概念和信息熵的计算方法，从而奠定了信息论的基础。美国数学家诺伯特·维纳在《控制论——动物和机器中的通信与控制问题》中指出，信息是适应控制外部世界的过程中，同外部世界交换的内容，信息就是信息，既非物质，也非能量。

从本体论层次看，信息可定义为事物的存在方式和运动状态的表现形式。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表达事物的存在形式，是表达知识的字符集合。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：不仅指代“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。

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部分和阶段而独立存在；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）依据来源不同分类

大数据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，例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就可以产生 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 的规模级别，而应用广泛、影响巨大的脸谱、亚马逊、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，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 Facebook 的 Open Compute Project，国内则有 TAB 主导的天蝎计划。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 Facebook 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，基于位置的服务和移动平台的感知功能，感知数据逐渐与互联网数据越来越重叠，但感知数据的体量同样惊人，并且总量或许可能不亚于社交媒体。Teradata 曾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也并不是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，增长极其迅猛。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。企业外部数据则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，是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。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。

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；阿里巴巴则根据其商业价值分为交易数据、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。通过数据进行分析人们的许多想法和行为，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。

### （二）依据使用主体分类

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。到2012年底

已经有180个国内城市开始投资建设智慧城市，总的投资规模包括数据平台的投入和通信网络的各种基础设施，大约6000亿元人民币。根据十二五规划，各地建设智慧城市仅基础设备的投资拉动规模总和就大约有1万亿元人民币。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，供更多有能力的机构组织或个人来分析并加以利用，以造福人类。

2. 企业的大数据

企业离不开数据支持有效决策。只有通过数据才能快速发展，实现利润，维护客户，传递价值，支撑规模，增加影响，撬动杠杆，带来差异，服务买家，提高质量，节省成本，扩大吸引力，打败对手，开拓市场。企业需要大数据的帮助，才能对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。网络企业更应该依靠大数据实现服务升级与方向转型，而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。

3. 个人的大数据

每个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，被第三方采集及快速处理，获得个性化的数据服务。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包括但不限于体温、心率、视力等各类身体数据，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，以便制订相应的理财规划并预测收益。国家有关部门还会在法律允许的范围内，经过严格程序，实时监控公共安全，预防犯罪。

个人的大数据严格受法律保护，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的内容与范围；数据只能由用户明确授权才能处理。

### 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。大数据可有效地处理大量数据，包括大规模并行处理（MPP）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。当前用于分析大数据的工具主要有开源与商用两个生态圈，开源大数据生态圈主要包括 Hadoop HDFS、Hadoop Map Reduce、HBase 等，商用大数据生态圈包括一体机数据库、数据仓库及数据集市。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，因为大型数据集分析需要大量电脑持续高效分配工作。大数据分析常和云计算联系在一起，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到的大量情报信息。计算机和互联网技术导致大数据处理出现困难，“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，筛选通信、教育、犯罪、医疗、金融和旅行等记录，之后组建了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，最终导致了教育、金融、医疗等多方面大数据的广泛应用。

## 第二节　大数据的特征

### 一、体量巨大，种类繁多

互联网搜索技术的进步、电商平台的全面覆盖以及社交平台的快速兴起，促进了多元化数据的产生，而且这些数据在未来甚至会呈指数增长。互联网、存储

等计算机科学领域正在迅猛进步，人们从多元化领域获得的数据资料成倍增加，搜集海量数据的根源是网络数据能够同步实时收集，医疗领域的数据资料与科研领域的研究数据都会成倍增加。占数据总量比重高达 85% 以上的非结构数据的增速远远高于结构化数据。就网络企业等相关投资者而言，这样的数据预测能够有效提升自信心。美国咨询公司麦肯锡对大数据进行了定义，指出大数据是传统数据库以及软硬件无法收集、储存和分析的巨大数据集。随着数据种类不断增多，如视频图片等信息增速的扩大，挖掘多元形式数据流间的关系成了大数据最为显著的优势。例如，对供水系统数据和交通状况的数据资料进行关联分析，得到清晨洗浴与早高峰时间存在着密切关联的结论；将堵车地点时间的数据资料和电网运行的数据资料进行分析，得到的结论是睡眠质量与交通事故的发生率存在内部关联。

## 二、开放公开，容易获得

人们之所以重视收集大数据，主要的目的是要开展数据分析。大数据并非只是在政府、企业等组织机构当中存在，还存在于社会生产、生活之中，具备自动性的特征。如，电信企业累积客户的电话记录，电商网站整合消费者信息，企业通过对大数据进行充分的分析与挖掘，能够全面提高企业的综合实力，优化企业运营，提升企业决策准确度，推动商业智能的长效发展，为企业经济效益最大化目标的实现创造良好条件。在一定规则开放性的背景之下，借助应用程序接口与爬虫采集等技术手段，大量企业组织与政府部门能够为社会各界以及科研等机构提供海量数据资源。开放公开容易获取的数据源，是大数据时代的基本特点，因此而对整个社会产生巨大影响。

## 三、重视社会预测

从本质上进行分析预测，是大数据特点的体现。在大数据背景下，预见行业未来前景的能力，成了企业不懈追求的目标。美国 Netflix 公司推出《纸牌屋》，通过收集 3000 万用户播放动作，研究用户几百万次评级和搜索，评估受众面对差异化节目给出的不同观点，从多个角度掌握观众在节目欣赏方面的实际习惯，利用对海量数据的挖掘与分析获知人们的兴趣爱好和节目偏好信息。这个公司采集用户的多元化具体数据资料，为视频行业的制作方法改革创造了良好条件，使得视频行业开始运用算法与逻辑分析的方式替代以往的生产方式。对大数据手段的应用能够预先分析受众情况，了解他们青睐的节目类型。更为有趣的一个案例是，商场比父亲更早获知未成年女儿怀孕的消息，因为商家通过对客户购买行为进行

大数据的挖掘与预测，能够获知怀孕的可能性。人们越来越重视大数据在预知社会多元问题方面的作用，同时也开始将其广泛推广应用到社会科学领域。

### 四、重视发现而非实证

实证研究特别关注构建理论假设，设定范围，并进行随机抽样，展开数据的定量调查与收集，从而证伪或证实理论假设。连续线性决策需要缜密的逻辑思维。大数据把关注点放在了数据方面，强调对数据的运用，创造知识，预测未来，挖掘本质，发现机遇。要实现对未来前景的预测，主要借助自下而上数据收集处理的方法，而不是依靠以理论假设为根基发现知识，预知未来，探寻规律。比方说，沃尔玛超市利用大数据技术对大量交易数据资料进行分析，获得的一个重要结论是，假如周末男人在购买婴儿尿布的同时，通常会顺便购买啤酒。利用大数据获得的结果，通常情况下是极具实用价值的，这也是很多超市在货物安排和摆放当中常常会遵循的规律。除此以外，大数据理论更能够从整体上进行数据的分析和把握，所以获得的分析结论价值极大，可以用于做出相关决策和获知规律的重要根据。

### 五、非结构化数据的涌现

数据挖掘关注的是未知有效信息与实用性强的知识，更多的属于非结构化数据，这是大数据时代非常突出的一个特点。如今 90% 甚至以上的数据均属于非结构化数据。社交媒体会随时产生无数数据文本，造成大量具有价值的数据资料被隐藏在了信息海洋当中。大数据技术从海量文本资料当中挖掘信息，获知人们的态度与行为的相关信息，呼应舆情监测的社会需要与企业商机。在对大量非结构化数据进行收集处理和分析时，社会出现了大量新需要，技术领域产生了极大的变革，同时也让很多非关系型数据库得到发展，大量计算机新技术持续不断地产生。大数据涵盖数据挖掘、网络挖掘、文本挖掘、IT 和商业智能信息技术、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、数据的开始

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，因此数字作为一项重要的改造世界的工具产生了。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1000人、10000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，没有测量，就没有科学。测量出来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。但这时的数据还只具有传统意义，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，虽还不是真正的大数据时代，但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、大数据的开始

有史以来，处理各种不断增长的数据一直是人类社会的难题。

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”。他发明了电动“打孔卡片制表机”来对卡片特定位置上的孔洞进行识别，并加以自动统计。这一发明被运用于统计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13年的人工统计工作量，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（Colossus）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5000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（NSA）——一个刚成立9年就拥有超过12000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，首先应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17000卷磁带）。

从20世纪40年代起，人们就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。20世纪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，设计超文本系统。在这个系统中，每个有用的事物，称为一样“资源”，并且有一个全局“统一资源标识符”（Uniform Resource Identifier，URI）标识，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（Hypertext Transfer Protocol）传送给用户，而后者通过点击链接来获得资源。人们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（Gordon Moore）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，使微处理器的性能提高两倍，或使价格下降一半。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（Mark Weiser）指出，各种各样的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、可穿戴设备等广泛使用，实现了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（ACM）下属的数据挖掘及知识发现专委会（SIGKDD）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