



《大数据导论（通识课版）》

教材官网: <http://dbllab.xmu.edu.cn/post/bigdataintroduction/>

温馨提示: 编辑幻灯片母版, 可以修改每页PPT的厦大校徽和底部文字

第1章 大数据概述

(PPT版本号: 2019年秋季学期)



扫一扫访问教材官网

林子雨

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn ▶▶

主页: <http://www.cs.xmu.edu.cn/linziyu>





课程教材

- 林子雨 编著 《大数据导论——数据思维、数据能力和数据伦理（通识课版）》
- 高等教育出版社，2019年11月



提纲

- 1.1 数据
- 1.2 大数据时代
- 1.3 大数据的发展历程
- 1.4 世界各国的大数据发展战略
- 1.5 大数据的概念
- 1.6 大数据的影响
- 1.7 大数据的应用
- 1.8 大数据产业



高校大数据课程

公共服务平台

百度搜索厦门大学数据库实验室网站访问平台





主讲教师



2018年国家精品在线开放课程

主讲教师：林子雨

中国高校首个“数字教师”提出者和建设者

2009年7月从事教师职业以来

累计**免费**网络发布超过**500万**字高价值教学和科研资料

网络浏览量超过**500万**次



数字教师LOGO



1.1 数据

- 1.1.1 数据的概念
- 1.1.2 数据类型
- 1.1.3 数据组织形式
- 1.1.4 数据生命周期
- 1.1.5 数据的使用
- 1.1.6 数据的价值性
- 1.1.7 数据爆炸



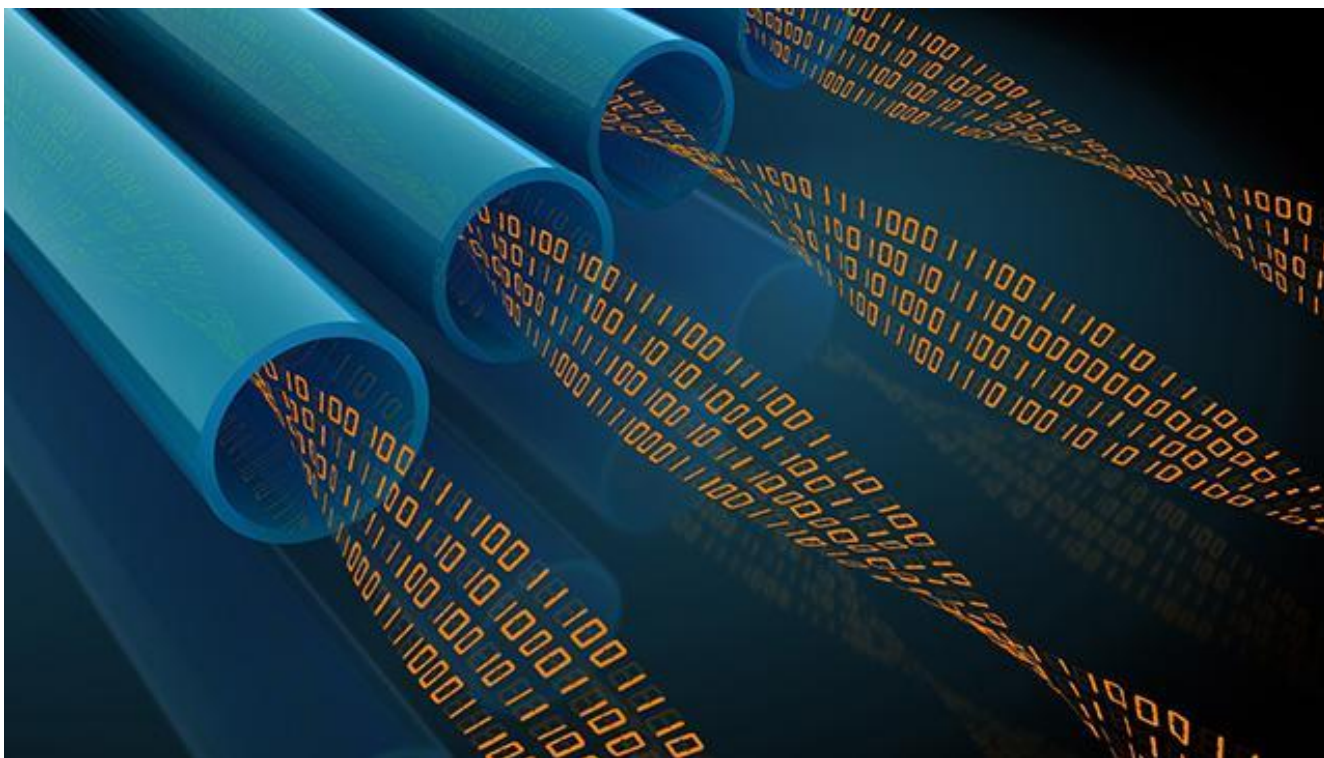
1.1.1 数据的概念

- 数据是指对客观事件进行记录并可以鉴别的符号，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合，是可识别的、抽象的符号
- 数据和信息是两个不同的概念，信息是较为宏观的概念，它由数据的有序排列组合而成，传达给读者某个概念方法等，而数据则是构成信息的基本单位，离散的数据没有任何实用价值。



1.1.1 数据的概念

数据也被称为“未来的石油”





1.1.2 数据类型





1.1.3 数据组织形式

计算机系统中的数据组织形式主要有两种，即文件和数据库。

(1) 文件：计算机系统中的很多数据都是以文件形式存在的，比如一个**WORD**文件、一个文本文件、一个网页文件、一个图片文件等等

(2) 数据库：计算机系统中另一种非常重要的数据组织形式就是数据库，今天，数据库已经成为计算机软件开发的基础和核心



1.1.4 数据生命周期

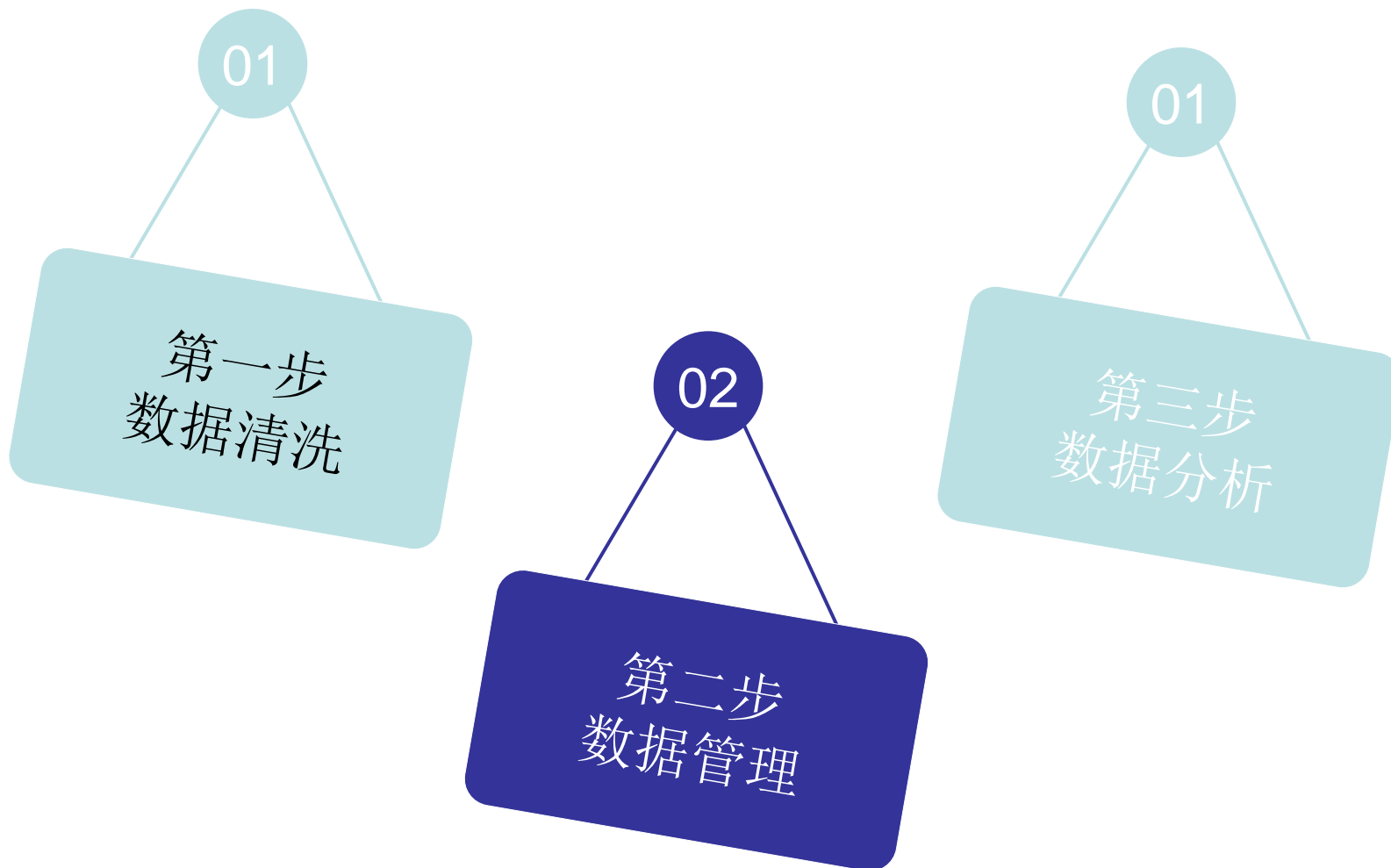
数据都存在一个生命周期，数据生命周期是指数据从创建、修改、发布利用到归档/销毁的整个过程。

数据生命周期管理工作包括以下几个方面：

- 对数据进行自动分类，分离出有效的数据，对不同类型数据制定不同的管理策略，并及时清理无用的数据。
- 构建分层的存储系统，满足不同类型的数据对不同生命周期阶段的存储要求，对关键数据进行数据备份保护，对处于生命期末期的数据进行归档并保存到适合长期保存数据的存储设备中。
- 根据不同的数据管理策略，实施自动分层数据管理，即自动把不同生命周期阶段的数据存放在最合适的存储设备上，提高数据可用性和管理效率。



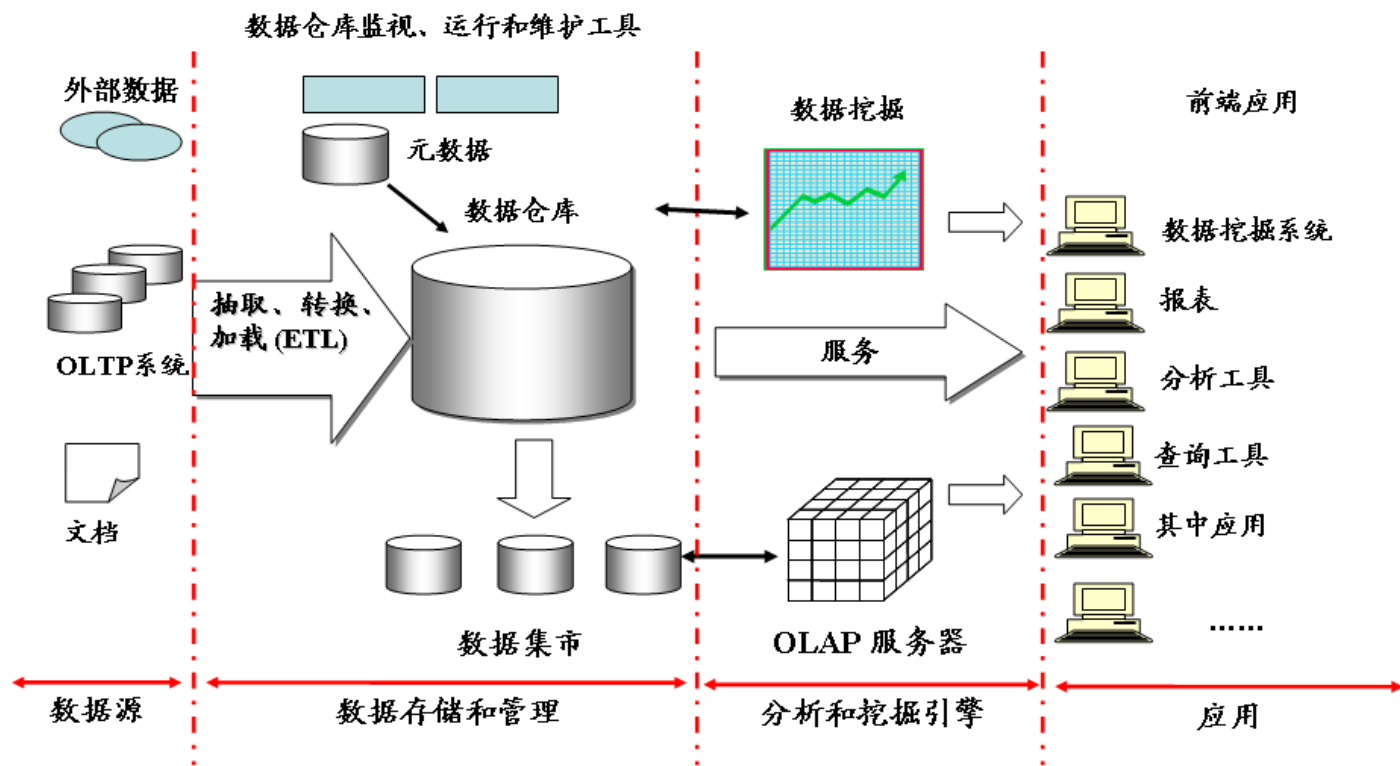
1.1.5 数据的使用





1.1.5 数据的使用

数据使用的实例：数据仓库





1.1.6 数据的价值性

在过去，一旦数据的基本用途实现了，往往就会被删除，一方面是由于过去的存储技术落后，人们需要删除旧数据来存储新数据，另一方面则是人们没有认识到数据的潜在价值。

数据的价值不会因为不断被使用而削减，反而会
因为不断重组而产生更大的价值

各类收集来的数据都应当被尽可能长时间地保存下来，同时也应当在一定条件下与全社会分享，
并产生价值



1.1.7 数据爆炸

人类进入信息社会以后，数据以自然方式增长，其产生不以人的意志为转移

从1986年开始到2010年的20年时间里，全球数据的数量增长了100倍，今后的数据量增长速度将更快，我们正生活在一个“数据爆炸”的时代





1.2 大数据时代

1.2.1 第三次信息化浪潮

1.2.2 信息科技为大数据时代提供技术支撑

1.2.3 数据产生方式的变革促成大数据时代的来临



1.2.1 第三次信息化浪潮

- 根据IBM前首席执行官郭士纳的观点，IT领域每隔十五年就会迎来一次重大变革

表1-1 三次信息化浪潮

信息化浪潮	发生时间	标志	解决问题	代表企业
第一次浪潮	1980年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010年前后	物联网、云计算和大数据	信息爆炸	将涌现出一批新的市场标杆企业



1.2.2 信息科技为大数据时代提供技术支撑

1. 存储设备容量不断增加

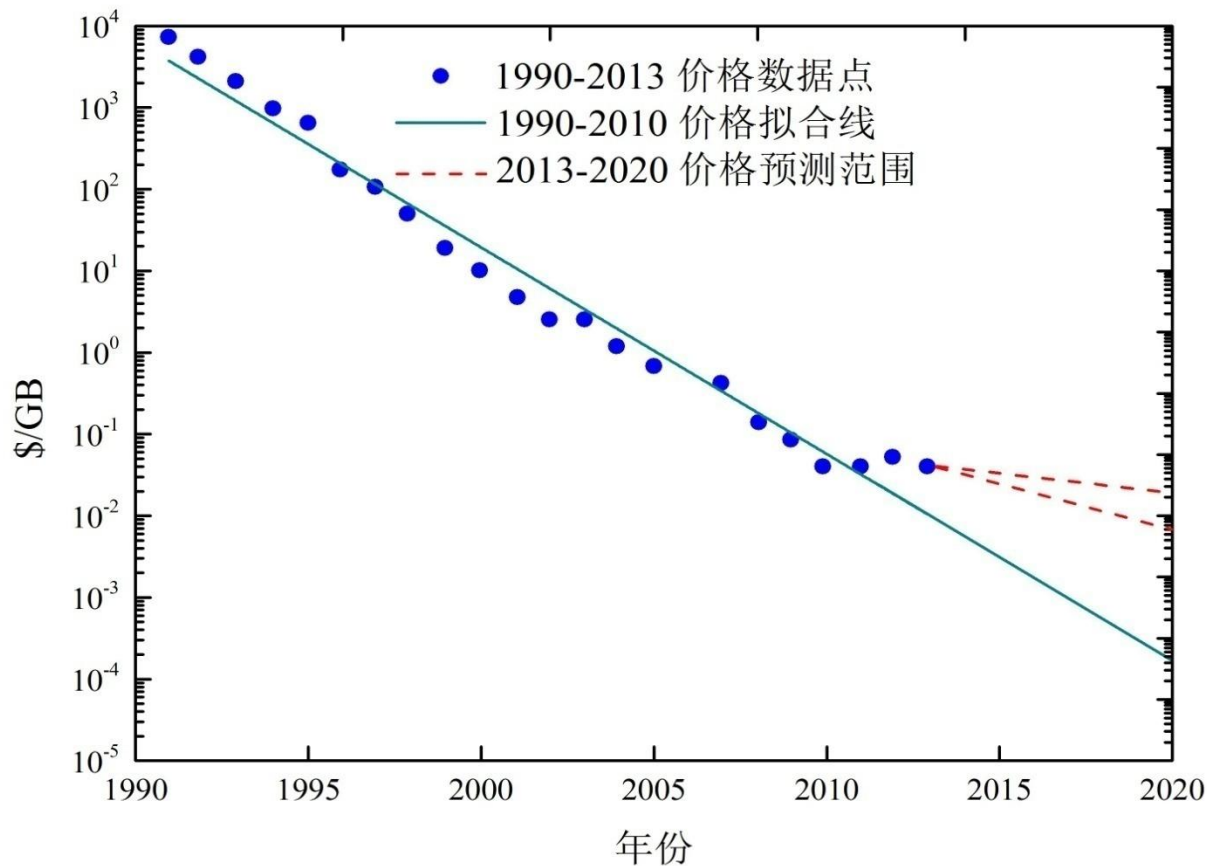


图 存储价格随时间变化情况



1.2.2 信息科技为大数据时代提供技术支撑

2. CPU处理能力大幅提升

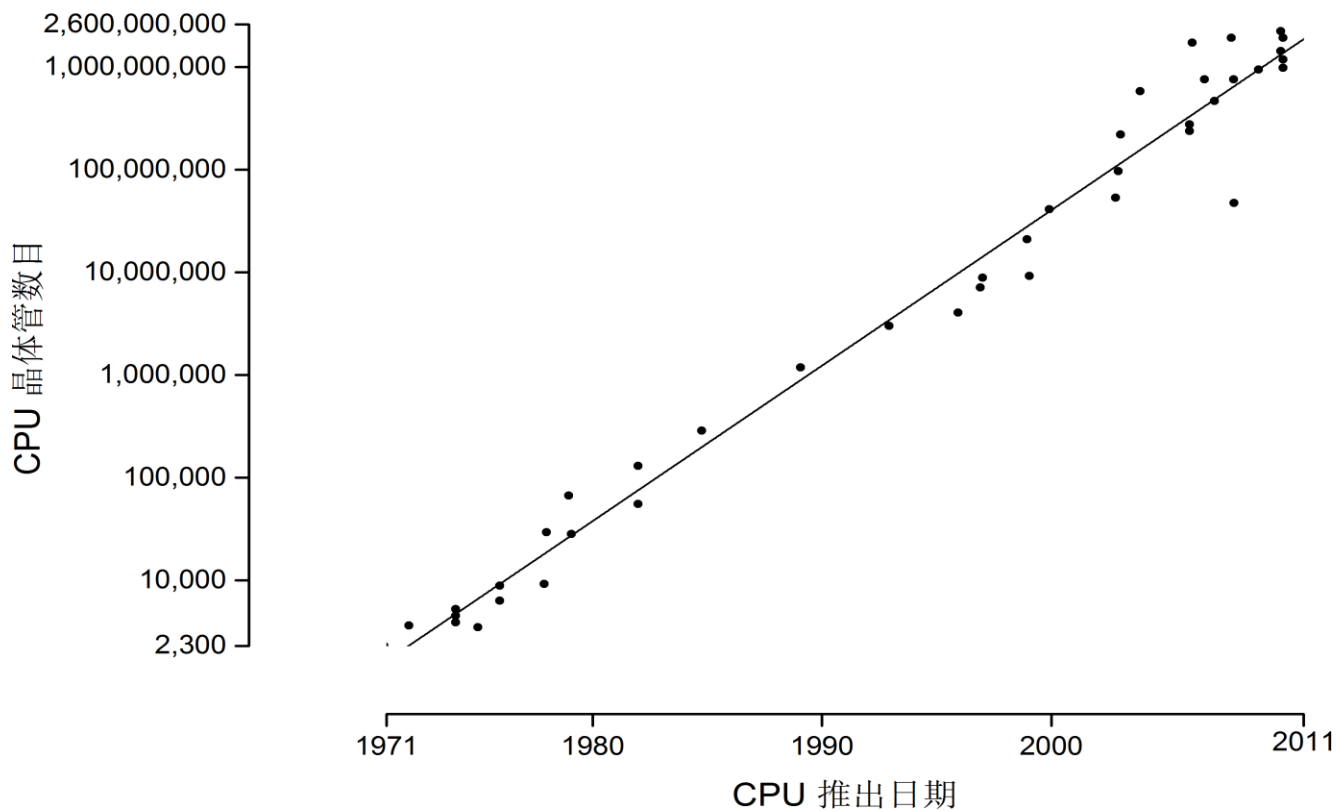


图 CPU晶体管数目随时间变化情况



1.2.2 信息技术为大数据时代提供技术支撑

3. 网络带宽不断增加

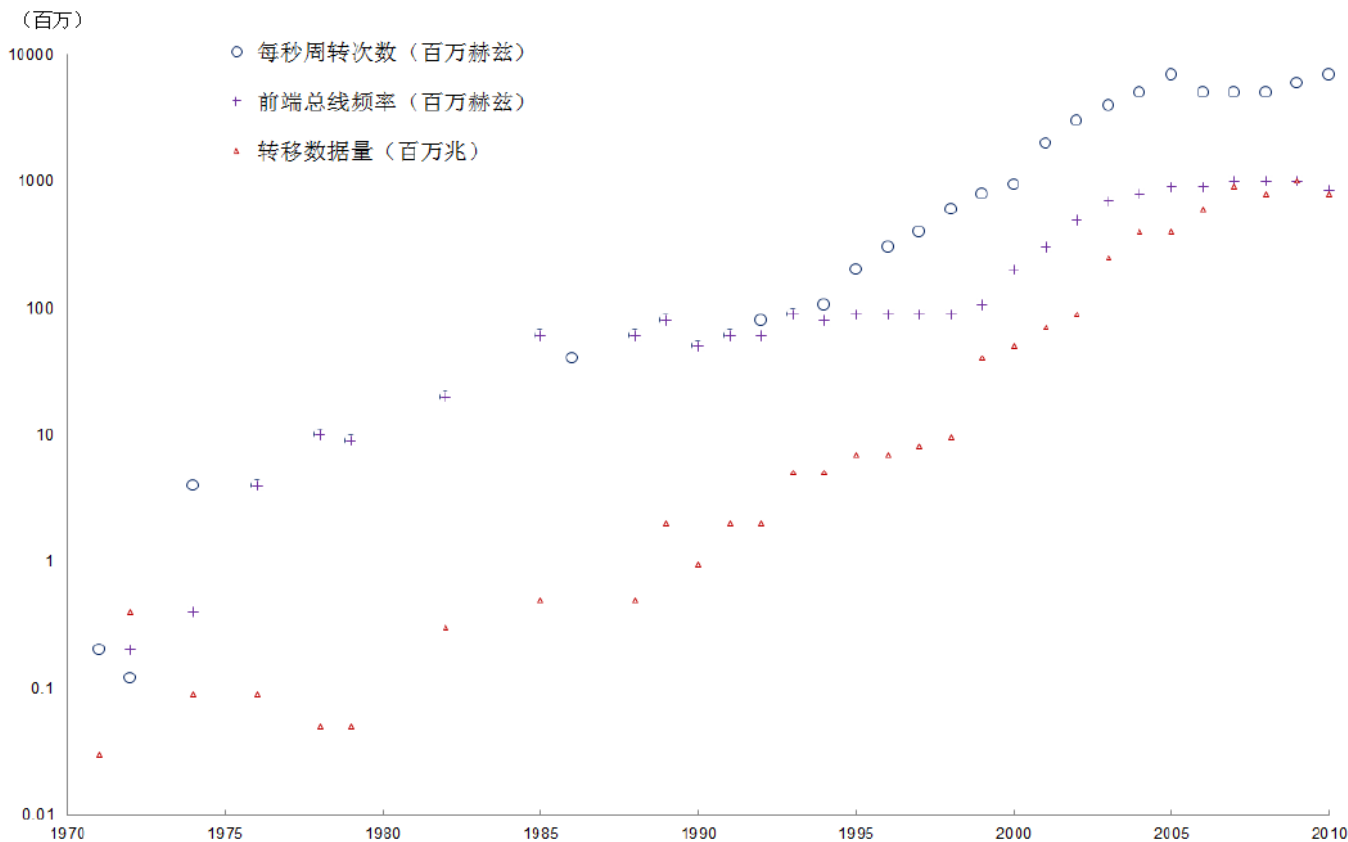


图 网络带宽随时间变化情况



1.2.3 数据产生方式的变革促成大数据时代的来临



图 数据产生方式的变革



1.3 大数据的发展历程

表 大数据发展的三个阶段

阶段	时间	内容
第一阶段：萌芽期	上世纪90年代至本世纪初	随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术开始被应用，如数据仓库、专家系统、知识管理系统等。
第二阶段：成熟期	本世纪前十年	Web2.0应用迅猛发展，非结构化数据大量产生，传统处理方法难以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统两大核心技术，谷歌的GFS和MapReduce等大数据技术受到追捧，Hadoop平台开始大行其道
第三阶段：大规模应用期	2010年以后	大数据应用渗透各行各业，数据驱动决策，信息社会智能化程度大幅提高



1.4 世界各国的大数据发展战略

1.4.1 美国

1.4.2 英国

1.4.3 法国

1.4.4 韩国

1.4.5 日本

1.4.6 中国



1.4世界各国的大数据发展战略

国家	战略
美国	稳步实施“三步走”战略，打造面向未来的大数据创新生态
英国	紧抓大数据产业机遇，应对脱欧后的经济挑战
法国	通过发展创新性解决方案并应用于实践来促进大数据发展
韩国	以大数据等技术为核心应对第四次工业革命
日本	开放公共数据，夯实应用开发
中国	实施国家大数据战略，加快建设数字中国



1.4.1 美国

美国是率先将大数据从商业概念上升至国家战略的国家，通过稳步实施“三步走”战略，在大数据技术研发、商业应用以及保障国家安全等方面已全面构筑起全球领先优势。

- 第一步是快速部署大数据核心技术研究，并在部分领域积极开发大数据应用。
- 第二步是调整政策框架与法律规章，积极应对大数据发展带来的隐私保护等问题。
- 第三步是强化数据驱动的体系和能力建设，为提升国家整体竞争力提供长远保障。



1.4.2 英国

- 英国政府于2010上线政府数据网站Data.gov.uk, 同美国的Data.gov平台功能类似, 但主要侧重于大数据信息挖掘和获取能力的提升
- 在2012年发布了新的政府数字化战略, 实现大数据驱动的社会经济增长
- 2013年英国政府加大了对大数据领域研究的资金支持



1.4.3 法国

- 2011年7月，法国启动了开放数据项目，通过实现公共数据在移动终端上的使用，最大限度地挖掘数据的应用价值。项目内容涉及交通、文化、旅游和环境等领域。
- 2013年12月，法国政府发布《数字化路线图》，明确了大数据是未来要大力支持的高新技术。
- 此外，法国中小企业、创新和数字经济部推出大数据规划，在2013年至2018年在法国巴黎等地创建大数据孵化器



1.4.4 韩国

- 韩国的智能终端普及率以及移动互联网接入速度一直位居世界前列，这使得其数据产出量也达到了世界先进水平
- 在朴槿惠政府倡导的“创意经济”国家发展方针指导下，韩国多个部门提出了具体的大数据发展计划
- 2016年年底，韩国发布以大数据等技术为基础的《智能信息社会中长期综合对策》，以积极应对第四次工业革命的挑战



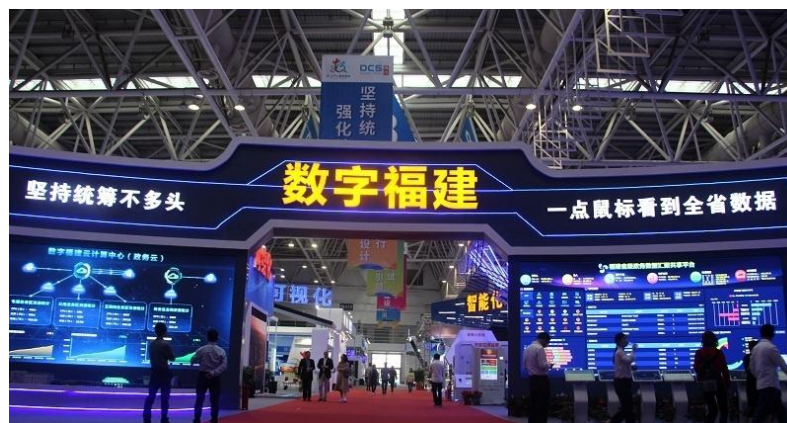
1.4.5 日本

- **2010年5月**，日本发达信息通信网络社会推进战略本部发布了以实现国民本位的电子政府、加强地区间的互助关系等为目的的《信息通信技术新战略》。
- **2012年6月**，日本IT战略本部发布电子政务开放数据战略草案
- **2012年7月**，日本政府推出了《面向2020年的ICT综合战略》，大数据成为发展的重点
- **2013年6月**，日本公布新IT战略——创新最尖端IT国家宣言，明确了**2013-2020年**期间以发展开放公共数据为核心的日本新IT国家战略



1.4.6 中国

- 2015年8月，国务院印发了《促进大数据发展行动纲要》。党的十八届五中全会将大数据上升为国家战略。在党的十九大报告中，习近平总书记明确指出：“推动互联网、大数据、人工智能和实体经济深度融合”。
- 2018年4月22日-24日，首届“数字中国”建设峰会在福建省福州市举行





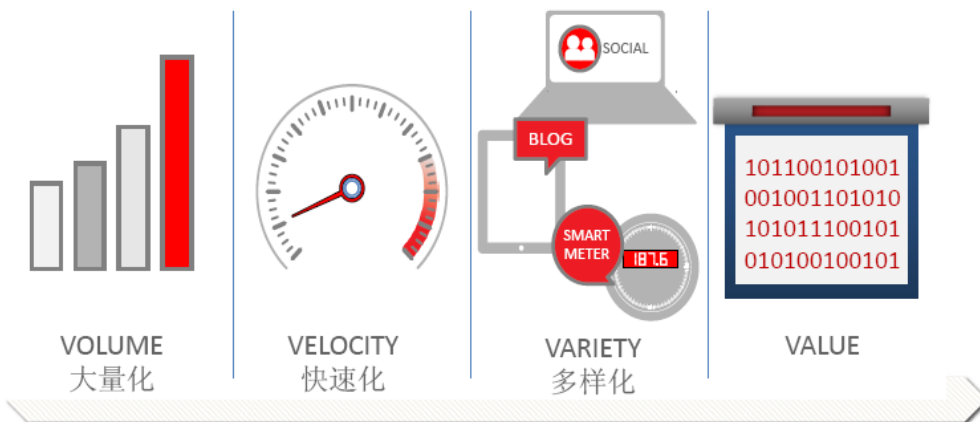
1.5 大数据的概念

1.5.1 数据量大

1.5.2 数据类型繁多

1.5.3 处理速度快

1.5.4 价值密度低

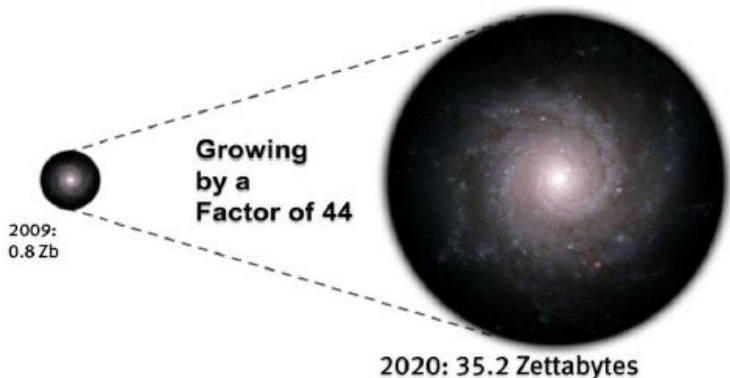


大数据不仅仅是数据的“大量化”，而是包含“快速化”、“多样化”和“价值化”等多重属性。



1.5.1 数据量大

- 根据IDC作出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（大数据摩尔定律）
- 人类在最近两年产生的数据量相当于之前产生的全部数据量
- 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍

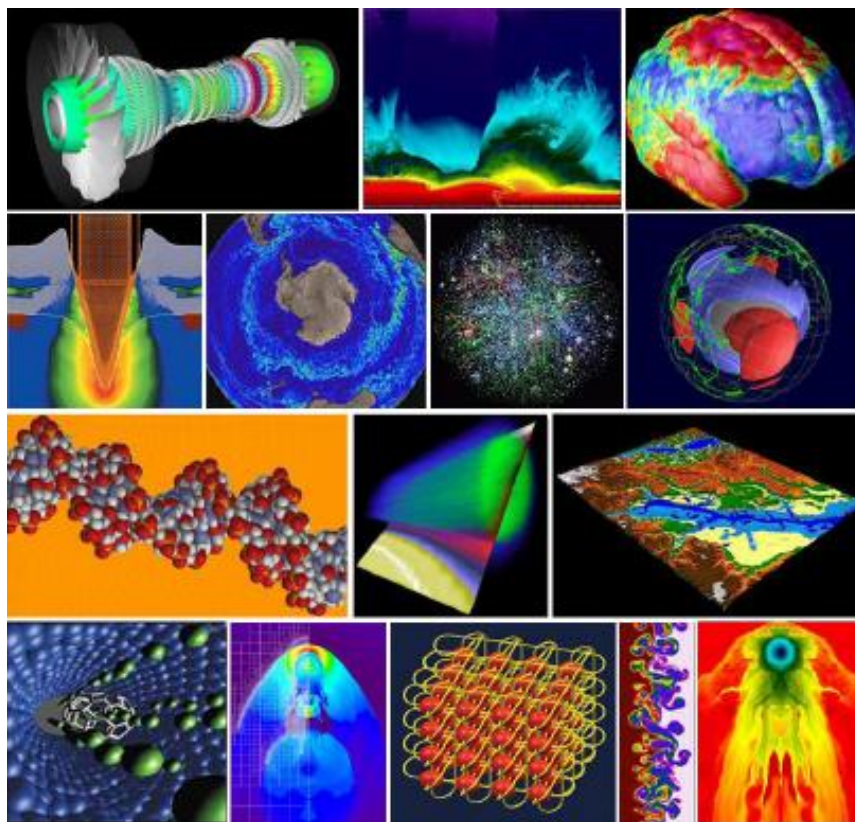


TERABYTE	10 的 12 次方	一块 1TB 硬盘		200,000 照片或 mp3 歌曲
PETABYTE	10 的 15 次方	两个数据中心机柜		16 个 Blackblaze pod 存储单元
EXABYTE	10 的 18 次方	2,000 个机柜		占据一个街区的 4 层数据中心
ZETTABYTE	10 的 21 次方	1000 个数据中心		纽约曼哈顿的 1/5 区域
YOTTABYTE	10 的 24 次方	一百万个数据中心		特拉华州和罗德岛州



1.5.2数据类型繁多

- 大数据是由结构化和非结构化数据组成的
 - 10%的结构化数据，存储在数据库中
 - 90%的非结构化数据，它们与人类信息密切相关



- 科学研究
 - 基因组
 - LHC 加速器
 - 地球与空间探测
- 企业应用
 - Email、文档、文件
 - 应用日志
 - 交易记录
- Web 1.0数据
 - 文本
 - 图像
 - 视频
- Web 2.0数据
 - 查询日志/点击流
 - Twitter/ Blog / SNS
 - Wiki



1.5.3处理速度快

- ❑ 从数据的生成到消耗，时间窗口非常小，可用于生成决策的时间非常少
- ❑ 1秒定律：这一点也是和传统的数据挖掘技术有着本质的不同

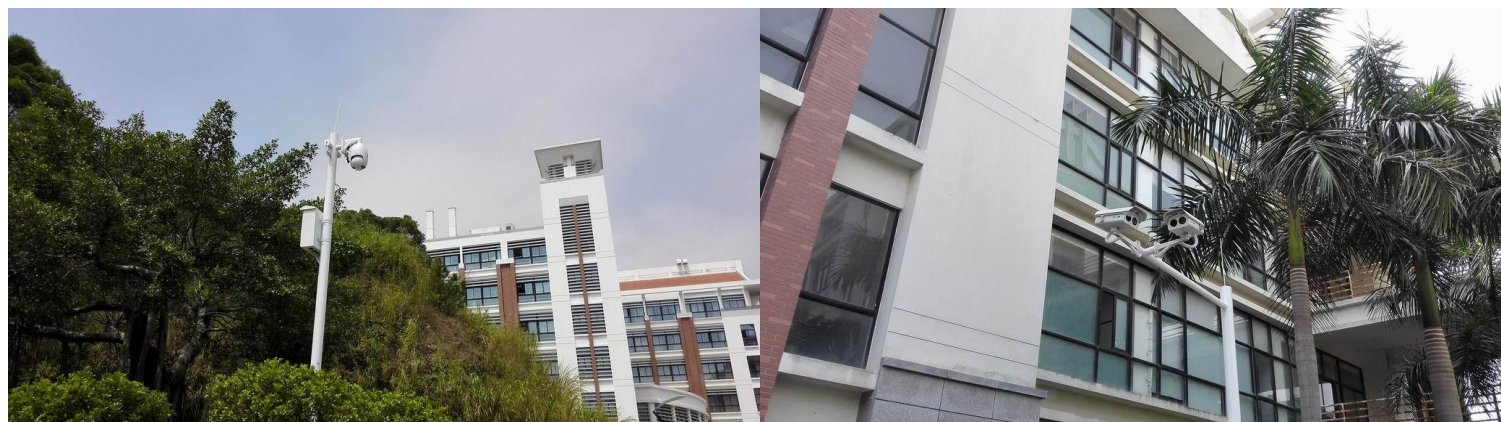




1.5.4价值密度低

价值密度低，商业价值高

以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，但是具有很高的商业价值





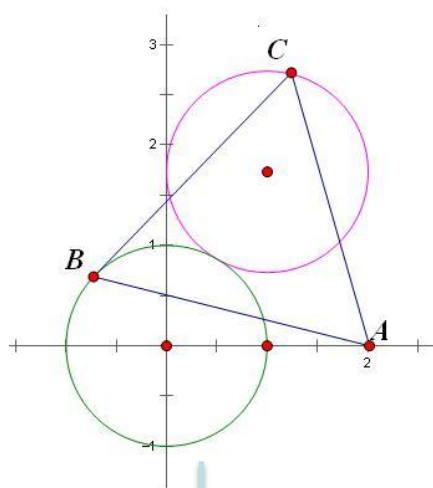
1.6 大数据的影响

- 1.6.1 大数据对科学研究的影响
- 1.6.2 大数据对社会发展的影响
- 1.6.3 大数据对就业市场的影响
- 1.6.4 大数据对人才培养的影响



1.6.1 大数据对科学研究的影响

图灵奖获得者、著名数据库专家Jim Gray 博士观察并总结人类自古以来，在科学研究上，先后历经了实验、理论、计算和数据四种范式



实验

理论

计算

数据

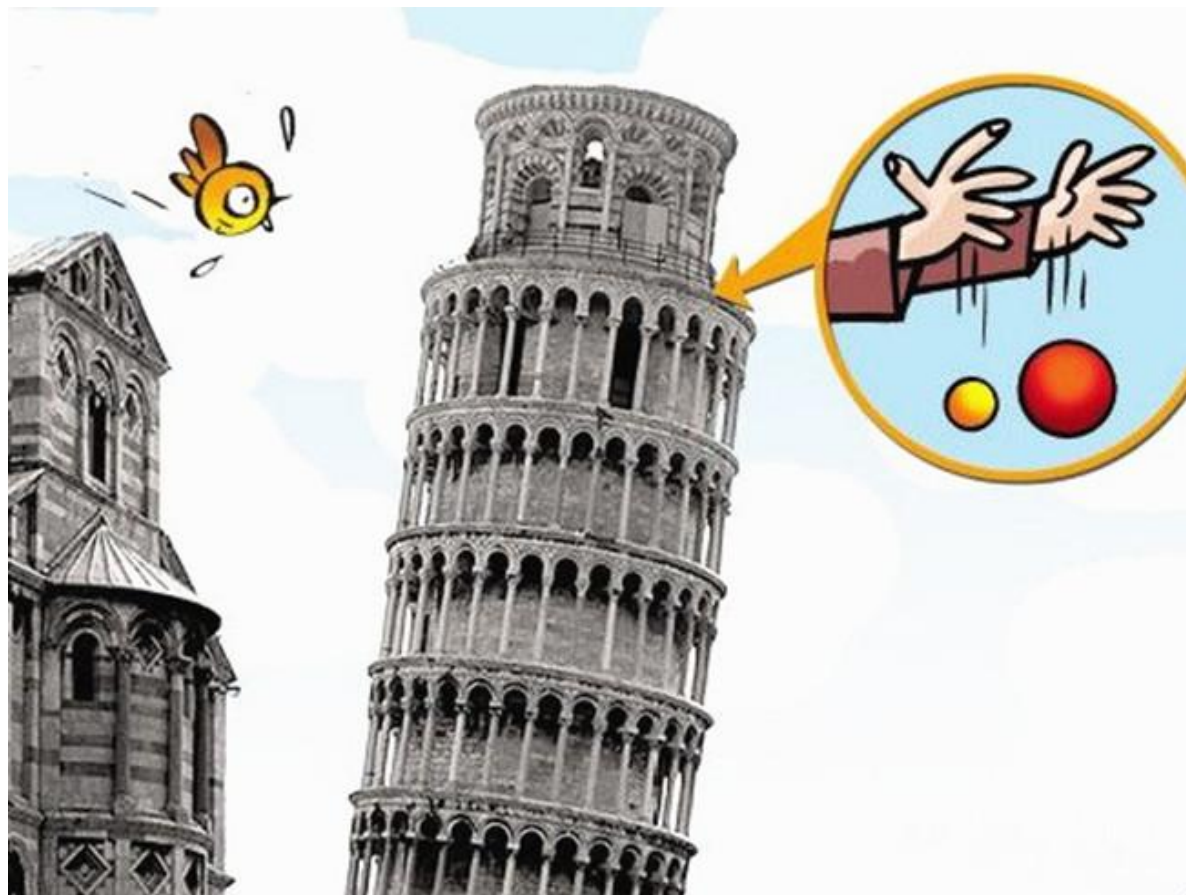


1.6.1 大数据对科学研究的影响

科学研究第一种范式：实验



伽利略

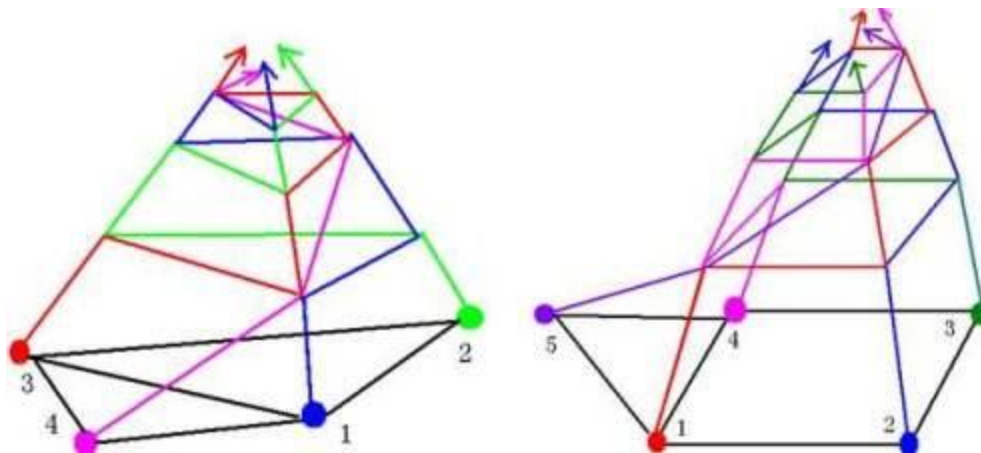


伽利略在比萨斜塔做两个铁球同时落地实验



1.6.1 大数据对科学研究的影响

科学研究第二种范式：理论



几何理论



牛顿三大定律



1.6.1 大数据对科学研究的影响

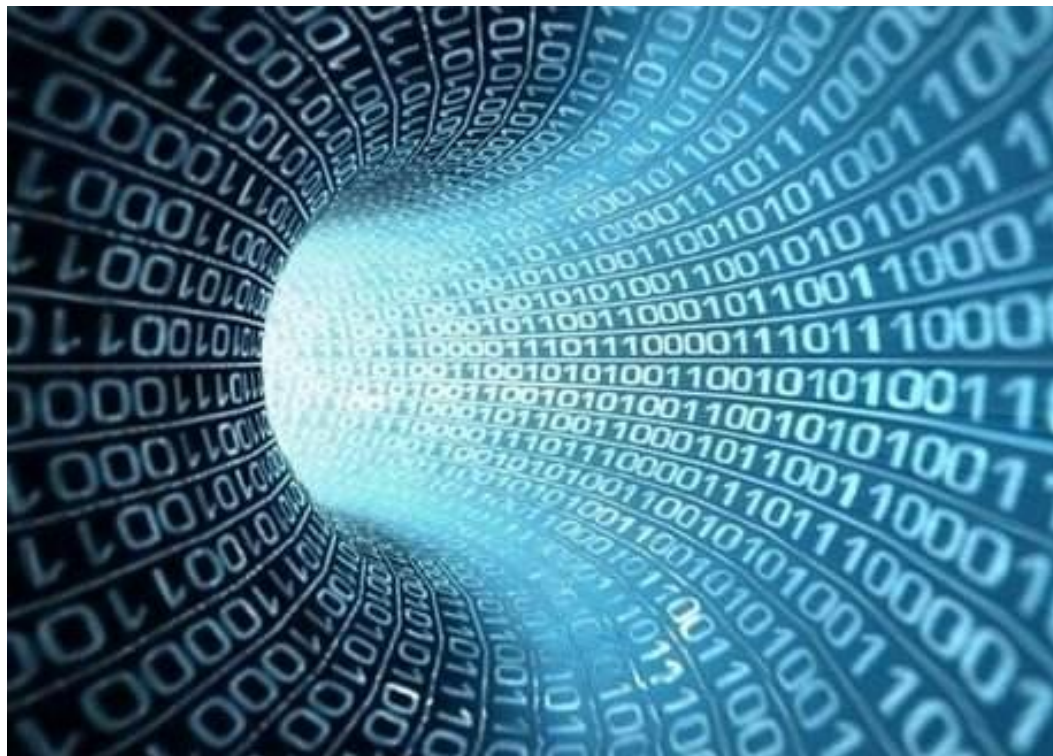
科学研究第三种范式：计算





1.6.1 大数据对科学研究的影响

科学研究第四种范式：数据



大数据时代，以数据为中心



1.6.2 大数据对社会发展的影响

- 大数据决策逐渐成为一种新的决策方式
- 大数据成为提升国家治理能力的新途径
- 大数据应用有力促进了信息技术与各行业的深度融合
- 大数据开发大大推动了新技术和新应用的不断涌现



1.6.3 大数据对就业市场的影响

大数据的兴起使得数据科学家成为热门职业



- 麦肯锡报告，到2018年，在“具有深入分析能力的人才”方面，美国面临着14万到19万的缺口，“可以利用大数据分析来做出有效决策的经理和分析师”缺口则会达到150万

- 国内有大数据专家估算过，5年内国内的大数据人才缺口会达到130万，以大数据应用较多的互联网金融为例，这一行业每年增速达到4倍，届时，仅互联网金融需要的大数据人才就是现在需求的4倍以上

- 根据第四届中国贵州人才博览会发布《全国大数据人才需求指数报告》，2016年2月份，贵阳大数据人才月薪已逼近8000元



1.6.4 大数据对人才培养的影响

大数据时代到底需要什么样的人才？

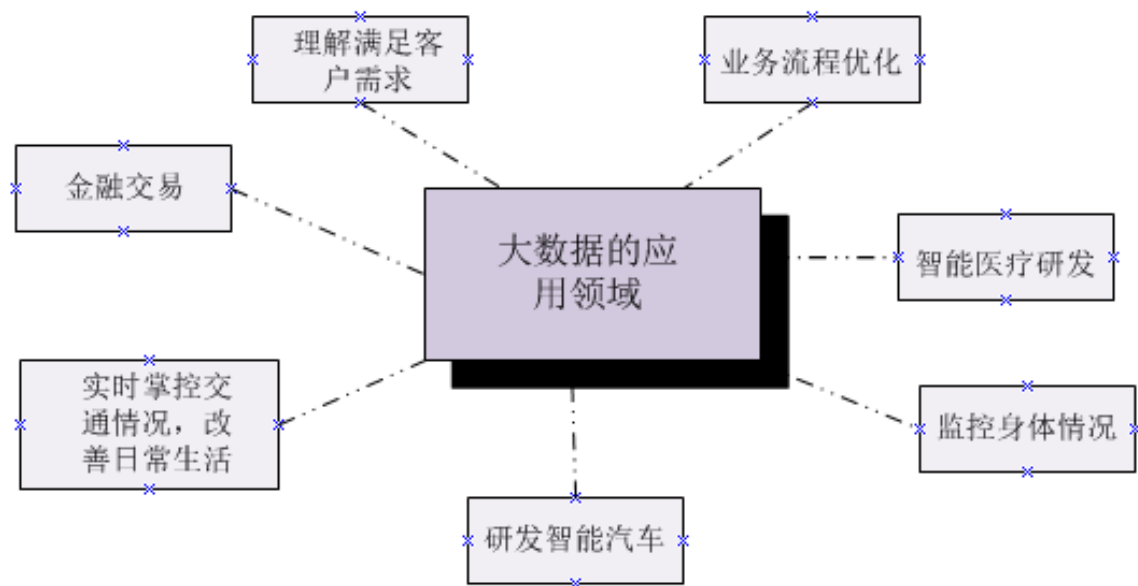
- 一是计算机技术相关人才，包括平台搭建和应用开发
- 二是统计学相关人才，包括数学、建模、算法
- 三是业务人才，就是要有一定的专业领域知识，只有明白目标领域知识的人才能了解数据的意义以及指导数据分析的方向并判断数据分析结果的可信性





1.7大数据的应用

- 大数据无处不在，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业都已经融入了大数据的印迹





1.7大数据的应用

- 就企业而言，对大数据的掌握程度可以转化为经济价值的源泉
- 就政府而言，大数据的发展将会提高政府科学决策水平，改变政府传统“拍脑袋”式决策，变为用数据说话，利用大数据分析社会、经济、人文生活等规律，从而为国家宏观调控、战略决策、产业布局等夯实根基
- 在医疗领域，大数据也有不俗表现
- 大数据也悄然地影响着绿茵场上强弱的较量



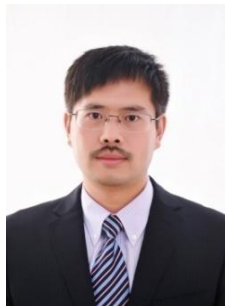
1.8 大数据产业

- 大数据产业是指一切与支撑大数据组织管理和价值发现相关的企业经济活动的集合

产业链环节	包含内容
IT基础设施层	包括提供硬件、软件、网络等基础设施以及提供咨询、规划和系统集成服务的企业，比如，提供数据中心解决方案的IBM、惠普和戴尔等，提供存储解决方案的EMC，提供虚拟化管理软件的微软、思杰、SUN、Redhat等
数据源层	大数据生态圈里的数据提供者，是生物大数据（生物信息学领域的各类研究机构）、交通大数据（交通主管部门）、医疗大数据（各大医院、体检机构）、政务大数据（政府部门）、电商大数据（淘宝、天猫、苏宁云商、京东等电商）、社交网络大数据（微博、微信、人人网等）、搜索引擎大数据（百度、谷歌等）等各种数据的来源
数据管理层	包括数据抽取、转换、存储和管理等服务的各类企业或产品，比如分布式文件系统（如Hadoop的HDFS和谷歌的GFS）、ETL工具（Informatica、Datastage、Kettle等）、数据库和数据仓库（Oracle、MySQL、SQL Server、HBase、GreenPlum等）
数据分析层	包括提供分布式计算、数据挖掘、统计分析等服务的各类企业或产品，比如，分布式计算框架MapReduce、统计分析软件SPSS和SAS、数据挖掘工具Weka、数据可视化工具Tableau、BI工具（MicroStrategy、Cognos、BO）等等
数据平台层	包括提供数据分享平台、数据分析平台、数据租售平台等服务的企业或产品，比如阿里巴巴、谷歌、中国电信、百度等
数据应用层	提供智能交通、智慧医疗、智能物流、智能电网等行业应用的企业、机构或政府部门，比如交通主管部门、各大医疗机构、菜鸟网络、国家电网等



附录A：主讲教师林子雨简介



主讲教师：林子雨

单位：厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

个人网页: <http://dblab.xmu.edu.cn/post/linziyu>

数据库实验室网站: <http://dblab.xmu.edu.cn>



扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革委员会副局长。中国计算机学会数据库专业委员会委员，中国计算机学会信息系统专业委员会委员。国内高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度和2017年度厦门大学教学类奖教金获得者，荣获2017年福建省精品在线开放课程、2018年厦门大学高等教育成果特等奖、2018年福建省高等教育教学成果二等奖、2018年国家精品在线开放课程。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，并以第一作者身份在《软件学报》《计算机学报》和《计算机研究与发展》等国家重点期刊以及国际学术会议上发表多篇学术论文。作为项目负责人主持的科研项目包括1项国家自然科学基金青年基金项目(No.61303004)、1项福建省自然科学基金项目(No.2013J05099)和1项中央高校基本科研业务费项目(No.2011121049)，主持的教改课题包括1项2016年福建省教改课题和1项2016年教育部产学研协作育人项目，同时，作为课题负责人完成了国家发改委城市信息化重大课题、国家物联网重大应用示范工程区域试点泉州市工作方案、2015泉州市互联网经济调研等课题。中国高校首个“数字教师”提出者和建设者，2009年至今，“数字教师”大平台累计向网络免费发布超过500万字高价值的研究和教学资料，累计网络访问量超过500万次。打造了中国高校大数据教学知名品牌，编著出版了中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》，并成为京东、当当网等网店畅销书籍；建设了国内高校首个大数据课程公共服务平台，为教师教学和学生学习大数据课程提供全方位、一站式服务，年访问量超过100万次。



附录B：大数据学习路线图



大数据学习路线图访问地址：<http://dblab.xmu.edu.cn/post/10164/>



附录C： 《大数据技术原理与应用》 教材

《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社 ISBN:978-7-115-44330-4 定价：49.80元



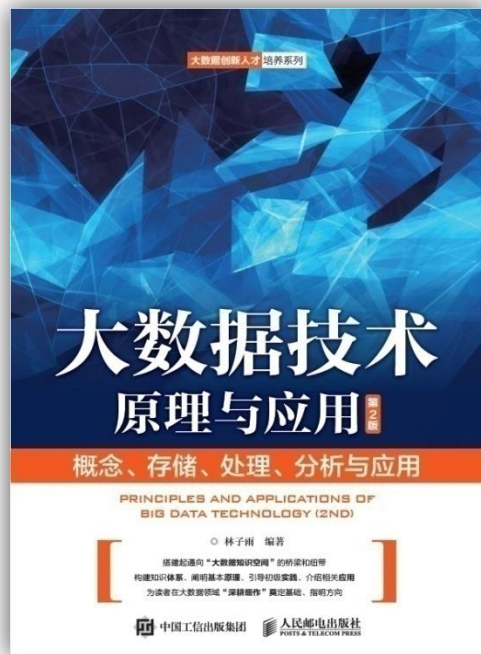
扫一扫访问教材官网

全书共有15章，系统地论述了大数据的基本概念、大数据处理架构Hadoop、分布式文件系统HDFS、分布式数据库HBase、NoSQL数据库、云数据库、分布式并行编程模型MapReduce、Spark、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：

<http://dbl原因.xmu.edu.cn/post/bigdata>





附录D：《大数据基础编程、实验和案例教程》

本书是与《大数据技术原理与应用（第2版）》教材配套的唯一指定实验指导书

大数据教材



1+1黄金组合
厦门大学林子雨编著

配套实验指导书



- 步步引导，循序渐进，详尽的安装指南为顺利搭建大数据实验环境铺平道路
- 深入浅出，去粗取精，丰富的代码实例帮助快速掌握大数据基础编程方法
- 精心设计，巧妙融合，五套大数据实验题目促进理论与编程知识的消化和吸收
- 结合理论，联系实际，大数据课程综合实验案例精彩呈现大数据分析全流程

清华大学出版社 ISBN:978-7-302-47209-4 定价：59元



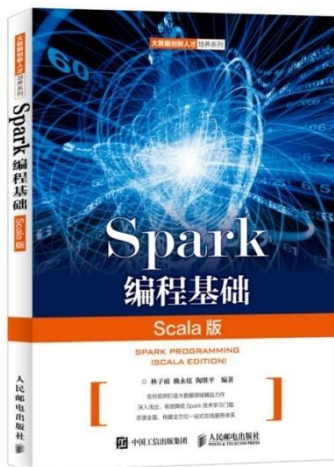
附录E：《Spark编程基础（Scala版）》

《Spark编程基础（Scala版）》

厦门大学 林子雨，赖永炫，陶继平 编著

披荆斩棘，在大数据丛林中开辟学习捷径
填沟削坎，为快速学习Spark技术铺平道路
深入浅出，有效降低Spark技术学习门槛
资源全面，构建全方位一站式在线服务体系

人民邮电出版社出版发行，ISBN:978-7-115-48816-9
教材官网：<http://dmlab.xmu.edu.cn/post/spark/>



本书以Scala作为开发Spark应用程序的编程语言，系统介绍了Spark编程的基础知识。全书共8章，内容包括大数据技术概述、Scala语言基础、Spark的设计与运行原理、Spark环境搭建和使用方法、RDD编程、Spark SQL、Spark Streaming、Spark MLlib等。本书每个章节都安排了入门级的编程实践操作，以便读者更好地学习和掌握Spark编程方法。本书官网免费提供了全套的在线教学资源，包括讲义PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。



附录F：高校大数据课程公共服务平台



高校大数据课程

公 共 服 务 平 台

<http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/>



扫一扫访问平台主页



扫一扫观看3分钟FLASH动画宣传片



附录G：高校大数据实训课程系列案例教材

为了更好地满足高校开设大数据实训课程的教材需求，厦门大学数据库实验室林子雨老师团队联合企业共同开发了《高校大数据实训课程系列案例》，目前已经完成开发的系列案例包括：

《基于协同过滤算法的电影推荐》

《电信用户行为分析》

《实时日志流处理分析》

《微博用户情感分析》

《互联网广告预测分析》

《网站日志处理分析》

系列案例教材将于2019年陆续出版发行，教材相关信息，敬请关注网页后续更新！

<http://dblab.xmu.edu.cn/post/shixunkecheng/>



扫一扫访问大数据实训课程系列案例教材主页

The background of the slide features several faint, light-blue silhouettes of people. At the top, there are two groups of people standing and holding hands. On the right side, a person is shown in profile, looking towards the center. On the left side, two people are shown in profile, one appearing to be speaking or gesturing towards the other. The overall scene suggests a group of people in a meeting or presentation setting.

Thank You!

Department of Computer Science, Xiamen University, 2019