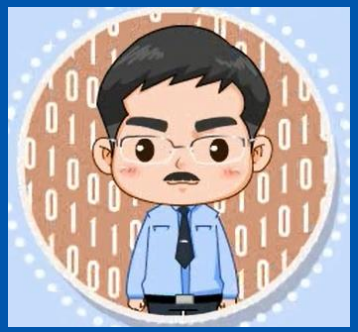


《大数据导论》

教材官网： http://dblab.xmu.edu.cn/post/bigdata-introduction/

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：2020年秋季学期）

扫一扫访问教材官网

林子雨 博士/副教授

厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

主页： **http://dblab.xmu.edu.cn/post/linziyu**

课程教材

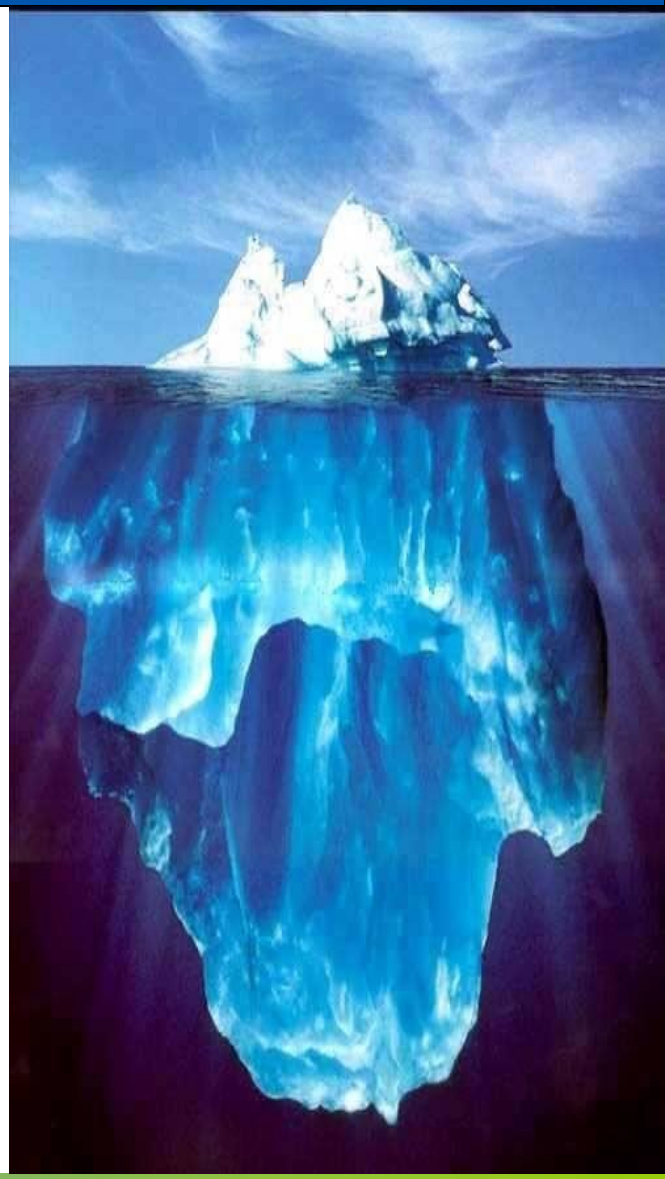
- 林子雨 编著《大数据导论》
- 人民邮电出版社，2020年9月第1版
- ISBN:978-7-115-54446-9 定价：49.80元

教材官网：http://dblab.xmu.edu.cn/post/bigdata-introduction/

扫一扫访问教材官网

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
1.9 高校大数据专业



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

数字教师LOGO

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1000万字高价值教学和科研资料
网络浏览量超过500万次

1.1 数据

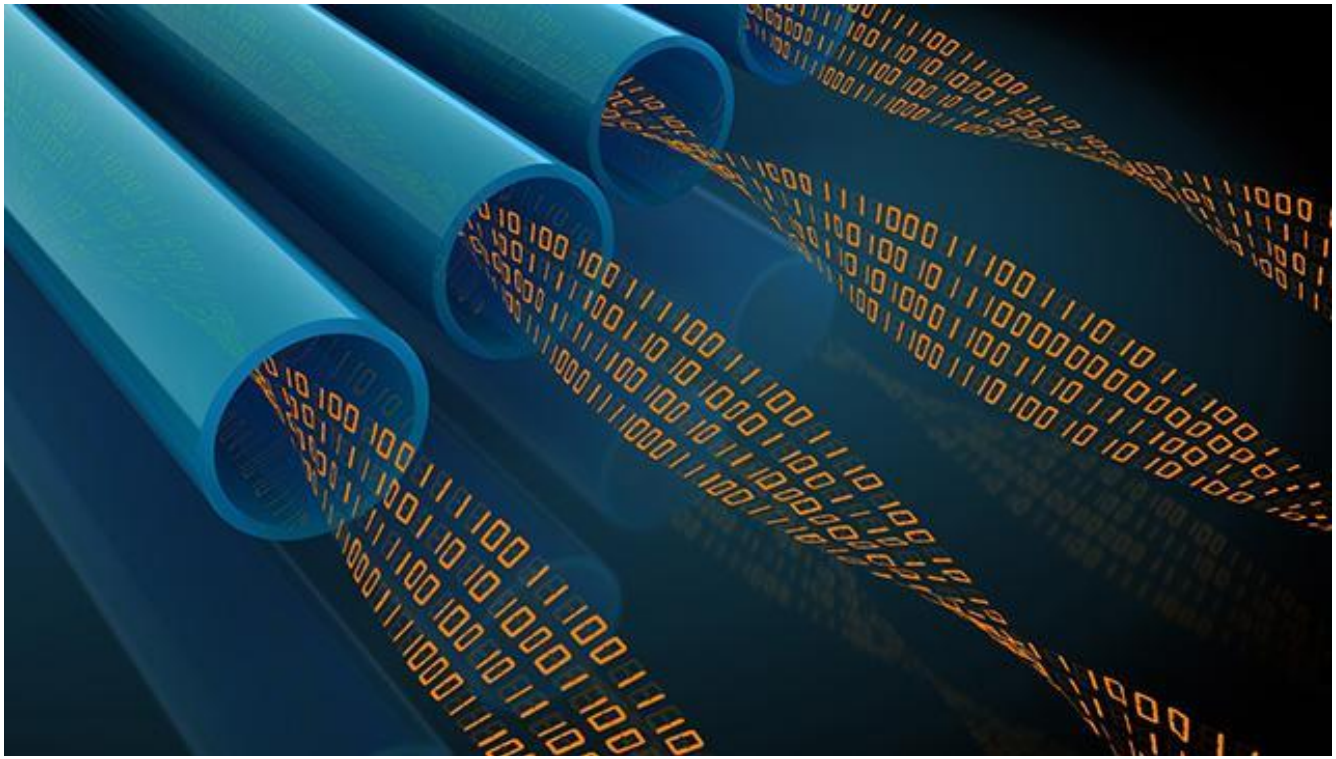
1.1.1 数据的概念
1.1.2 数据类型
1.1.3 数据组织形式
1.1.4 数据的使用
1.1.5数据的价值性
1.1.6数据爆炸

1.1.1 数据的概念

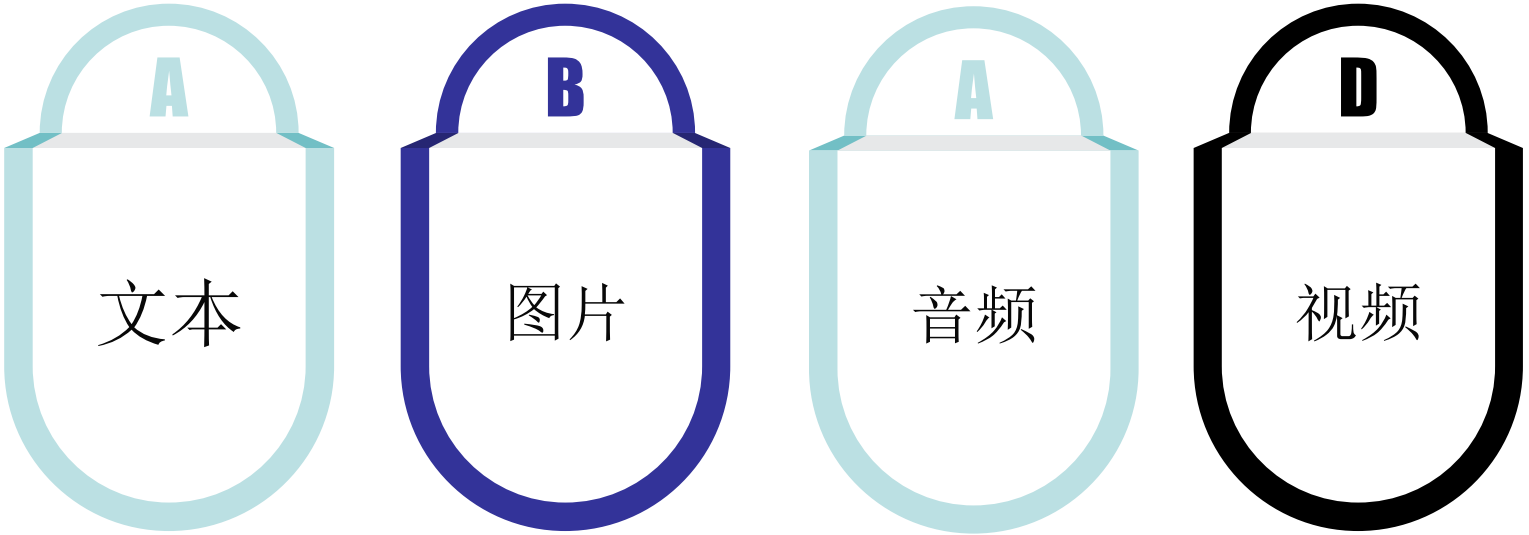
- 数据是指对客观事件进行记录并可以鉴别的符号，是对客观事物的性质、状态以及相互关系等进行记载的物理符号或这些物理符号的组合，是可识别的、抽象的符号
- 数据和信息是两个不同的概念，信息是较为宏观的概念，它由数据的有序排列组合而成，传达给读者某个概念方法等，而数据则是构成信息的基本单位，离散的数据没有任何实用价值。

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



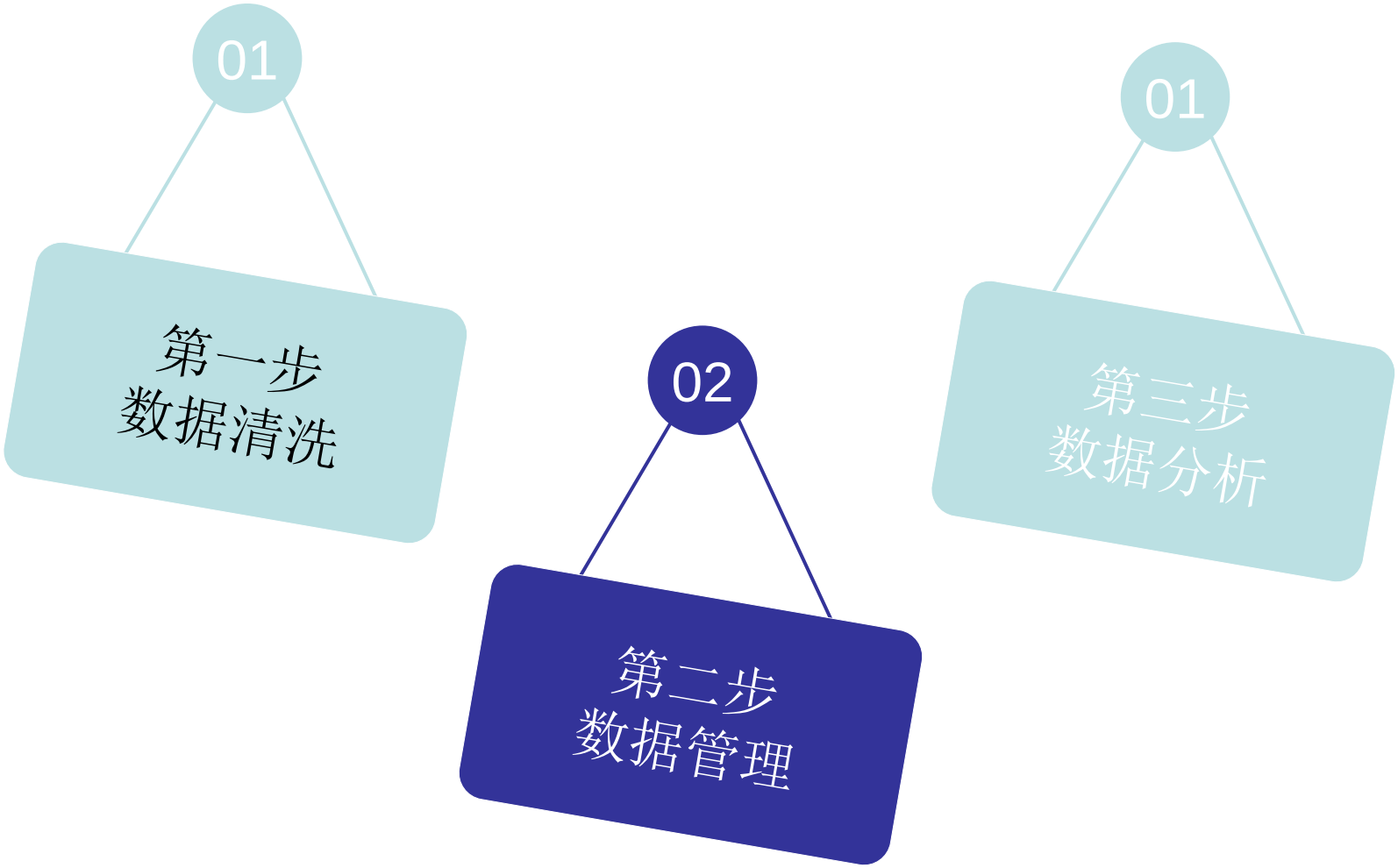
1.1.3 数据组织形式

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

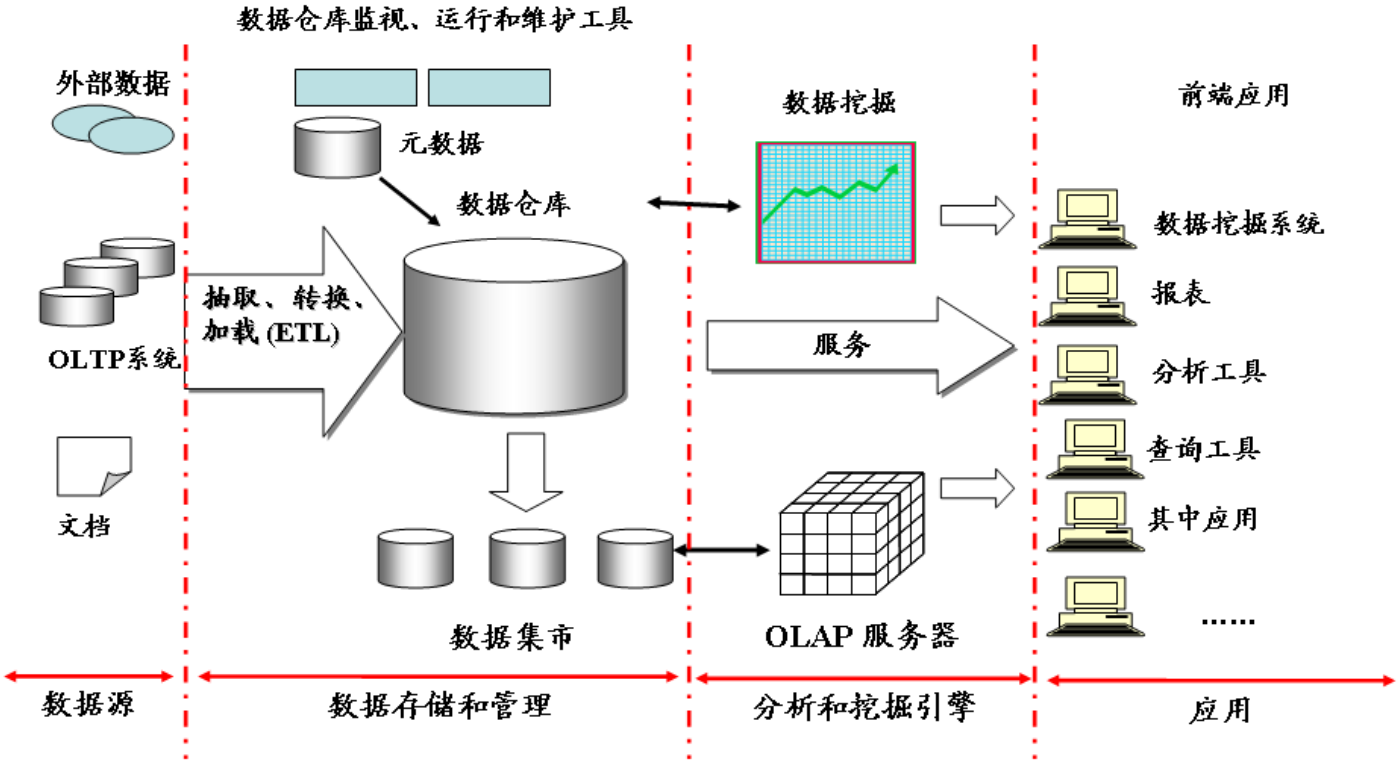
（2）数据库：计算机系统中另一种非常重要的数据组织形式就是数据库，今天，数据库已经成为计算机软件开发的基础和核心

1.1.4 数据的使用



1.1.4 数据的使用

数据使用的实例：数据仓库



1.1.5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

1.1.6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

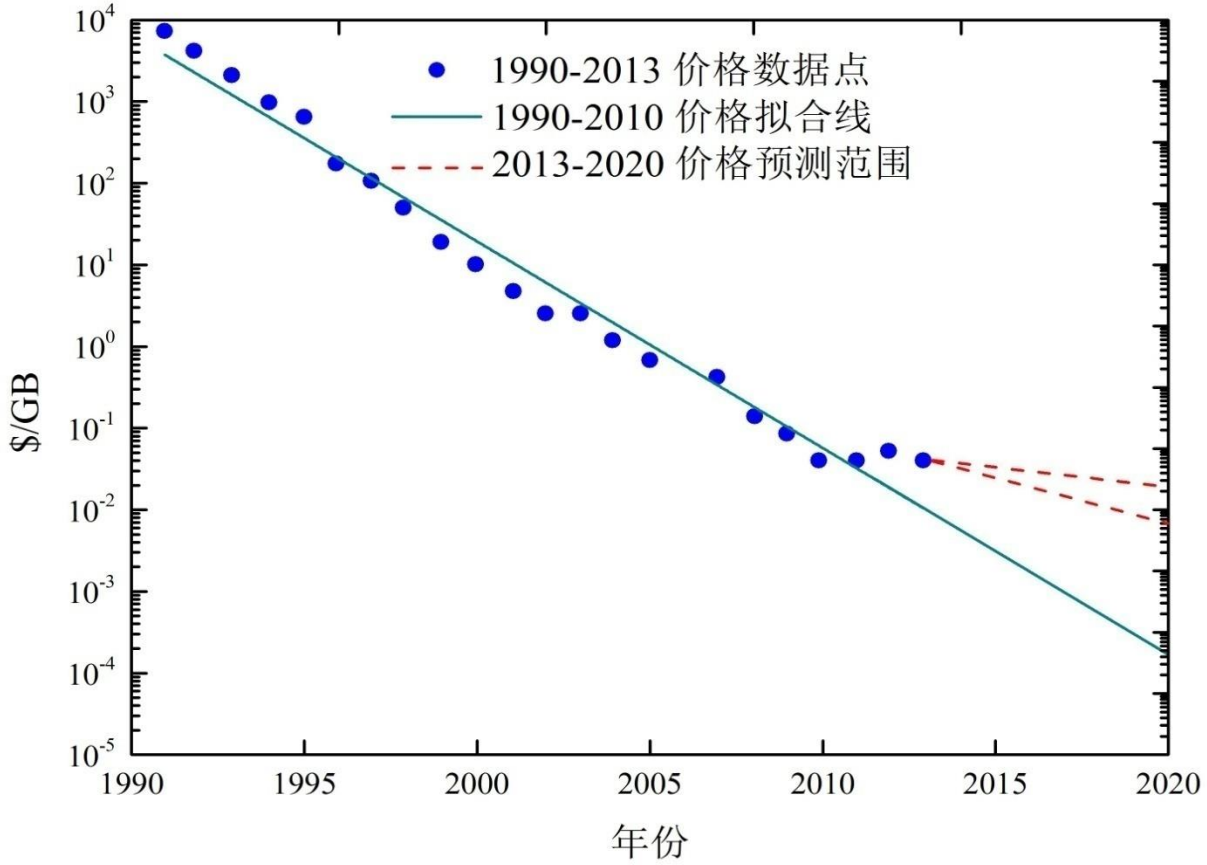


图 存储价格随时间变化情况

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

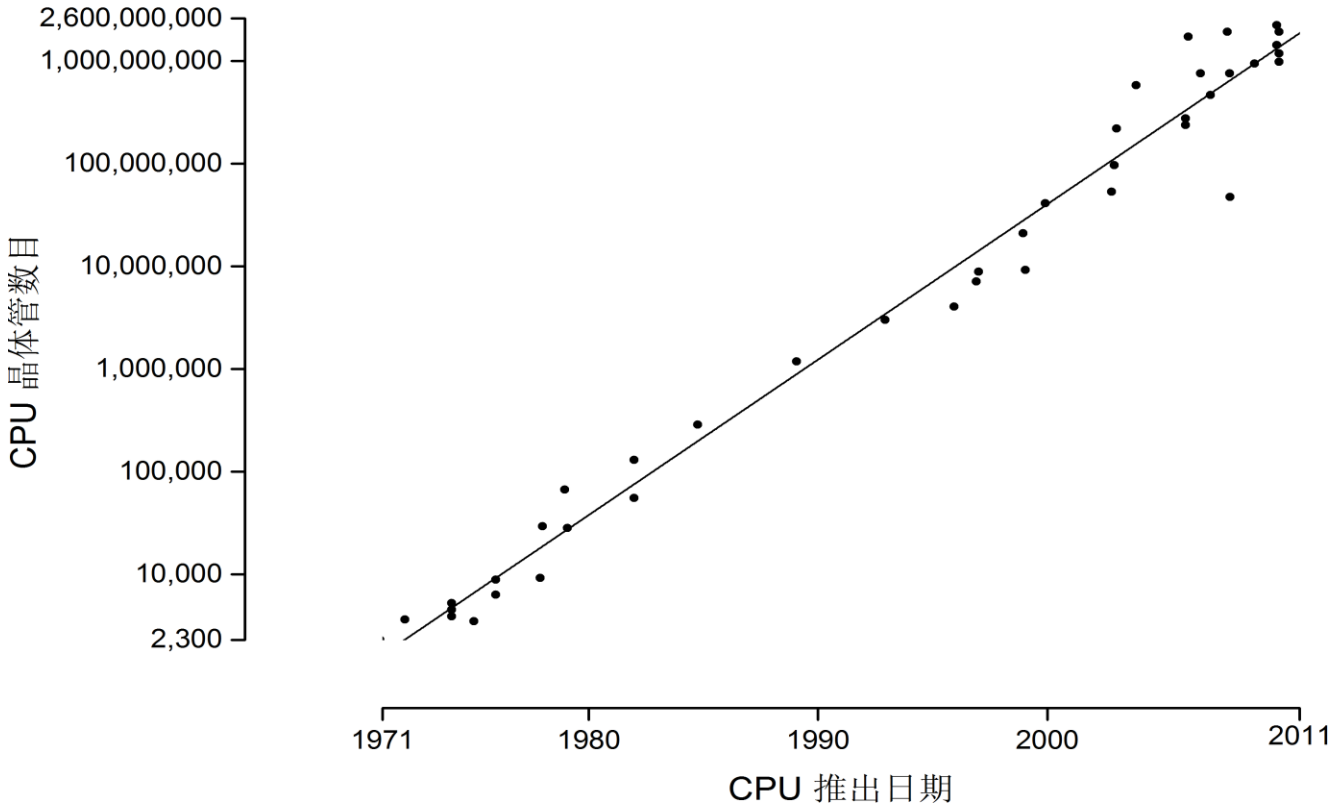


图 CPU晶体管数目随时间变化情况

1.2.2 信息科技为大数据时代提供技术支撑

在信息化基础设施方面，据工业和信息化部官网消息，截至2019年12月底，我国互联网宽带接入端口数量达9.16亿个，其中，光纤接入端口占互联网接入端口的比重达91.3%；光缆线路总长度已达4750万公里，相当于在京沪高铁线上往返1.8万余次。同时，近五年来固定宽带和移动宽带资费平均下降90%，速率提升6倍。目前，我国已基本实现“城市光纤到楼入户，农村宽带进乡入村”。

据中国信息通信研究院（简称中国信通院）数据，截至2020年2月底，全国建设开通5G基站达16.4万个，5G网络建设基础不断夯实。2020年中国将建设60万～80万个5G基站。

1.2.2 信息科技为大数据时代提供技术支撑

3. 网络带宽不断增加

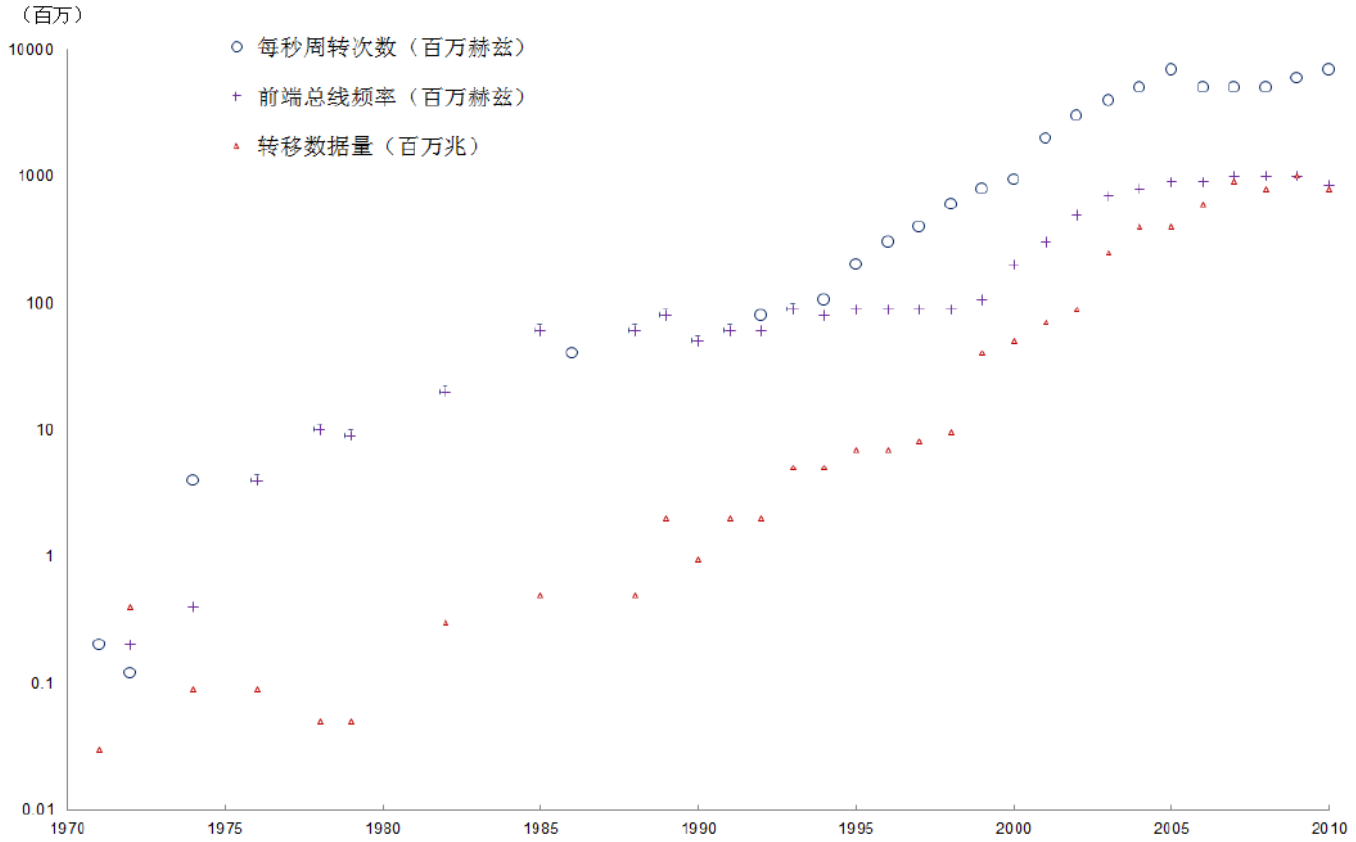


图 网络带宽随时间变化情况

1.2.3 数据产生方式的变革促成大数据时代的来临

1 运营式系统阶段

- 数据库的出现使得数据管理的复杂度大大降低，数据往往伴随着一定的运营活动而产生并记录在数据库中，数据的产生方式是被动的

2 用户原创内容阶段

- 数据爆发产生于 Web 2.0 时代，而 Web 2.0 的最重要标志就是用户原创内容
- 智能手机等移动设备加速内容产生
- 数据产生方式是主动的

3 感知式系统阶段

- 感知式系统的广泛使用
- 人类社会数据量第三次大的飞跃最终导致了大数据的产生

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战略性高新技术。
- 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

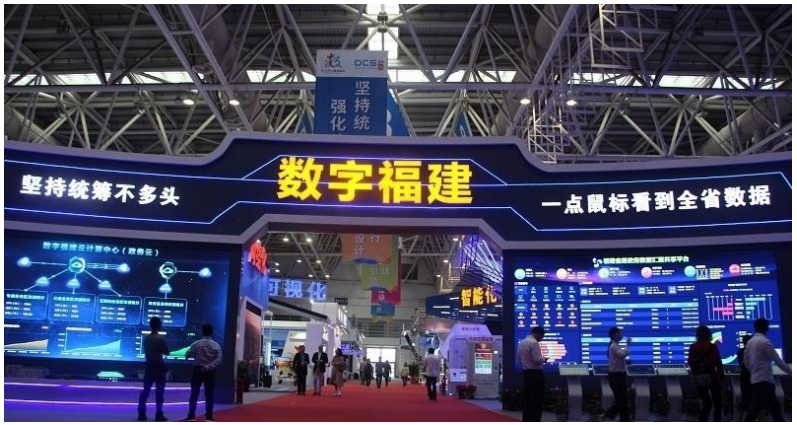
1.4.5 日本

•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标的《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

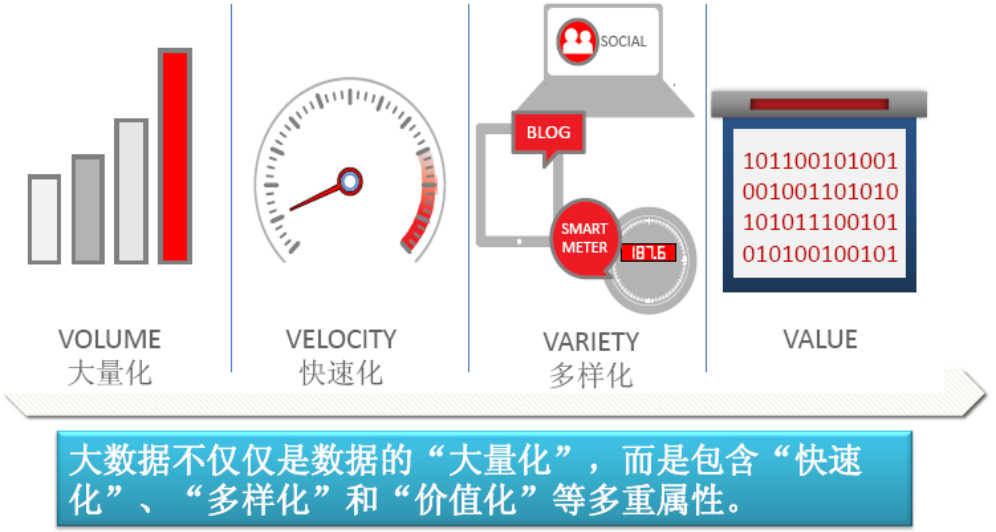
•2015年8月，国务院印发了《促进大数据发展行动纲要》。党的十八届五中全会将大数据上升为国家战略。在党的十九大报告中，习近平总书记明确指出：“推动互联网、大数据、人工智能和实体经济深度融合”。

•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



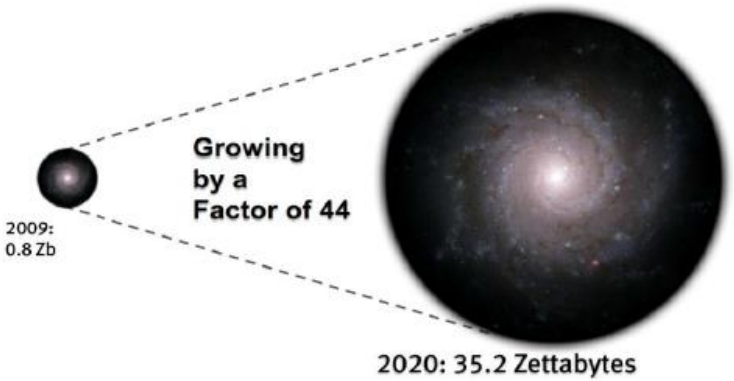
1.5大数据的概念

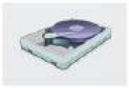
1.5.1数据量大
1.5.2数据类型繁多
1.5.3处理速度快
1.5.4价值密度低



1.5.1数据量大

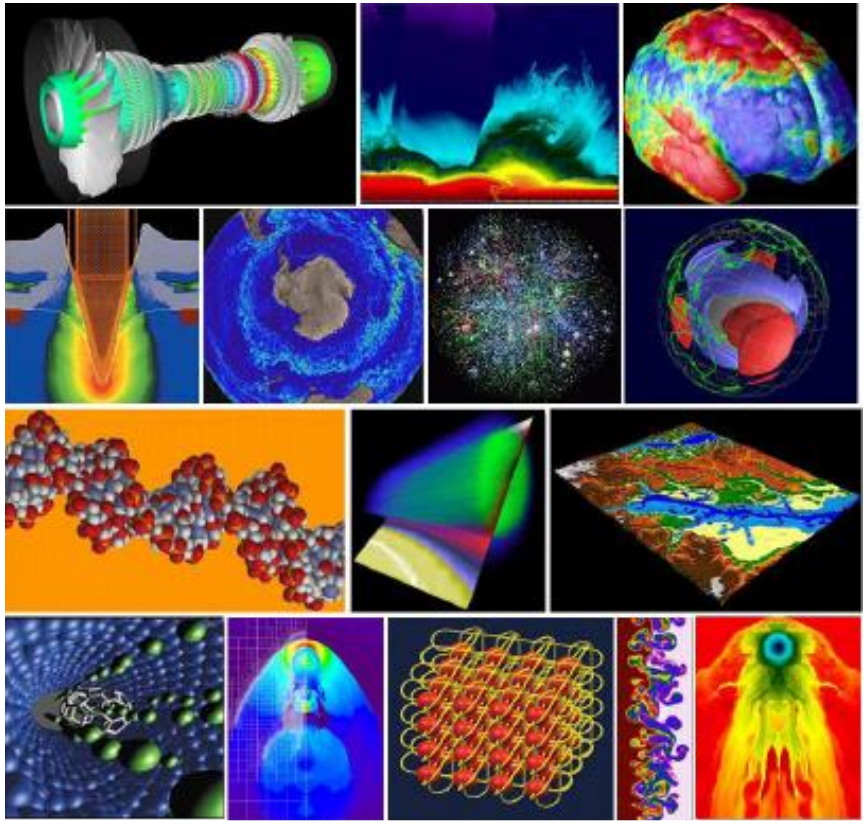
- 根据IDC作出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（大数据摩尔定律）
- 人类在最近两年产生的数据量相当于之前产生的全部数据量
- 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍



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

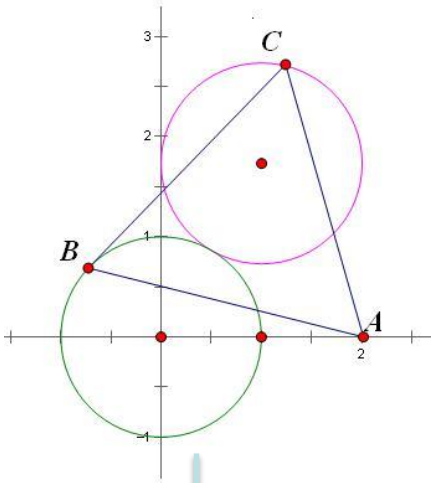



1.6大数据的影响

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

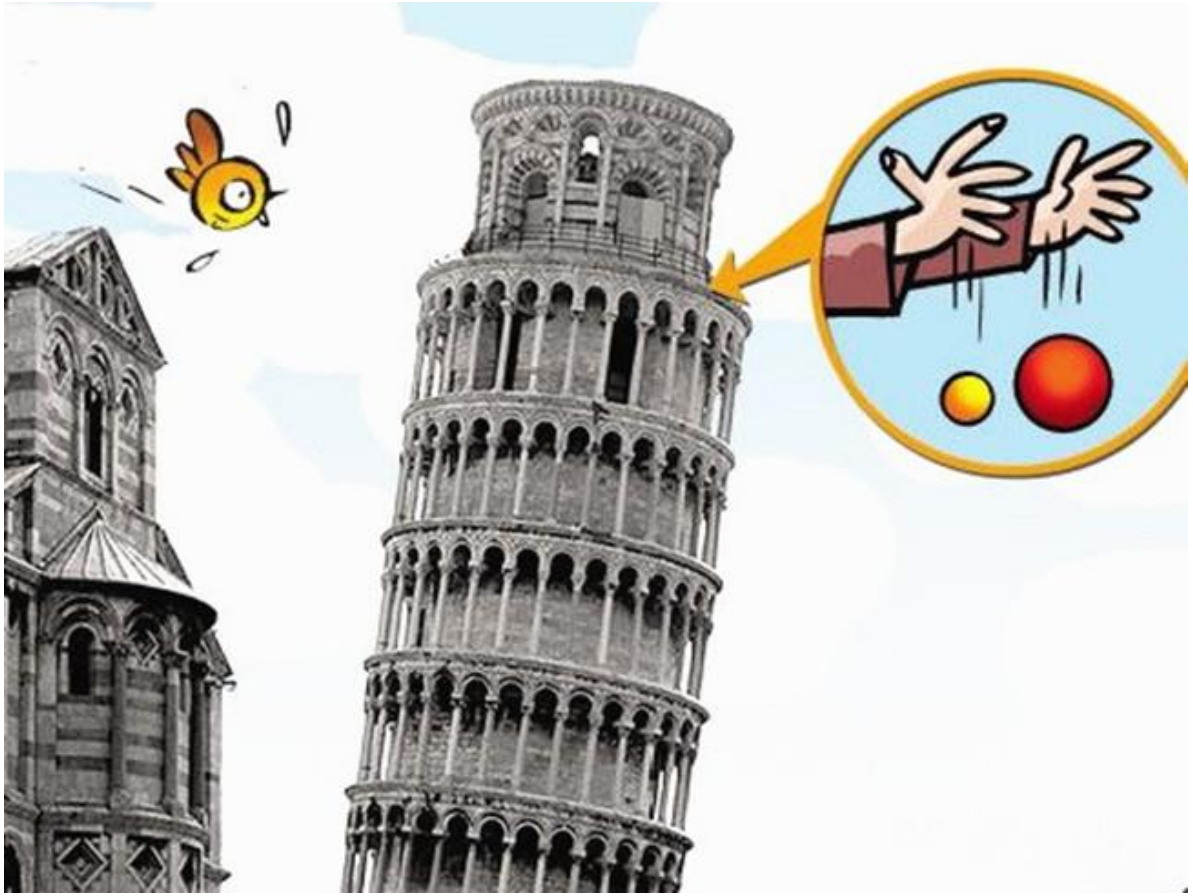


实验　　理论　　计算　　数据

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

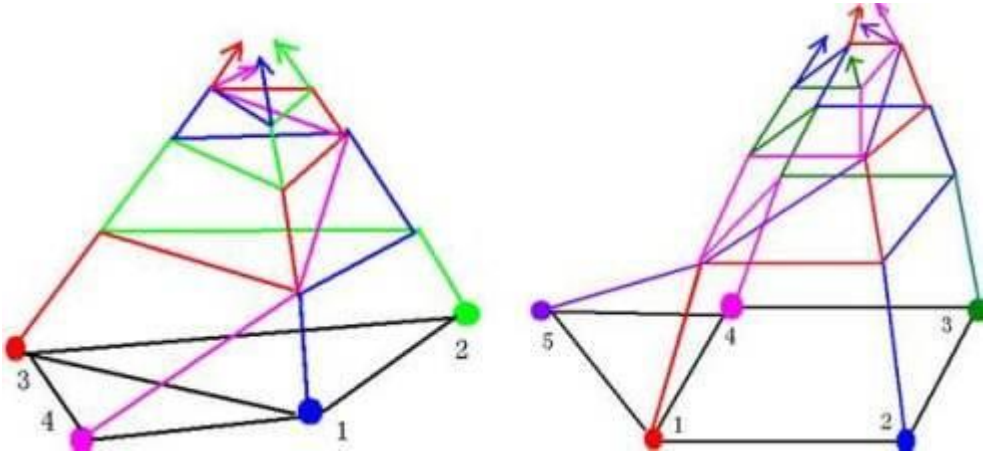
伽利略



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

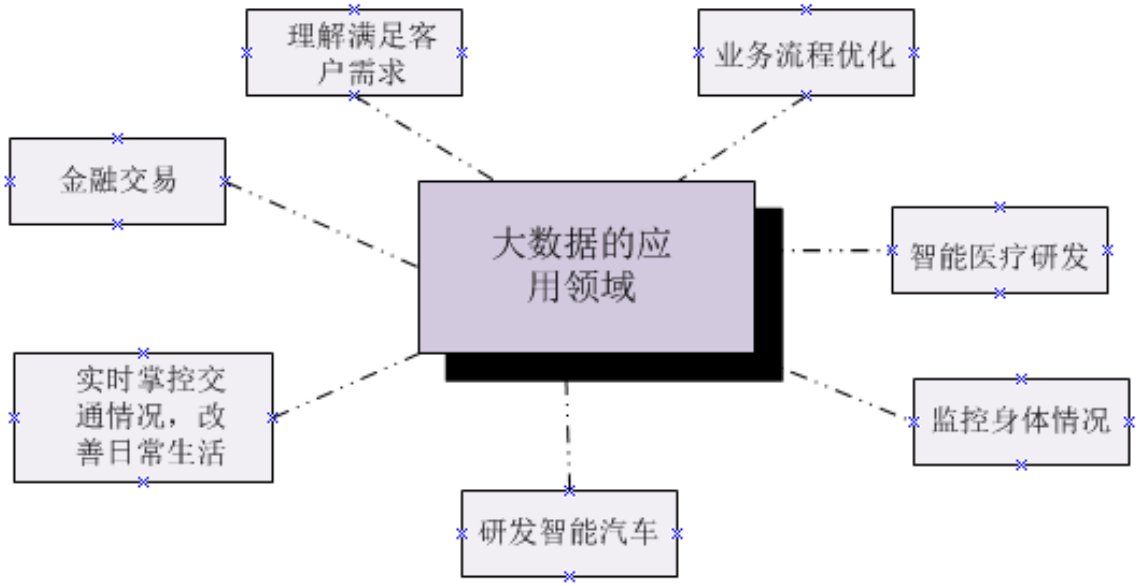
1.6.4大数据对人才培养的影响

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

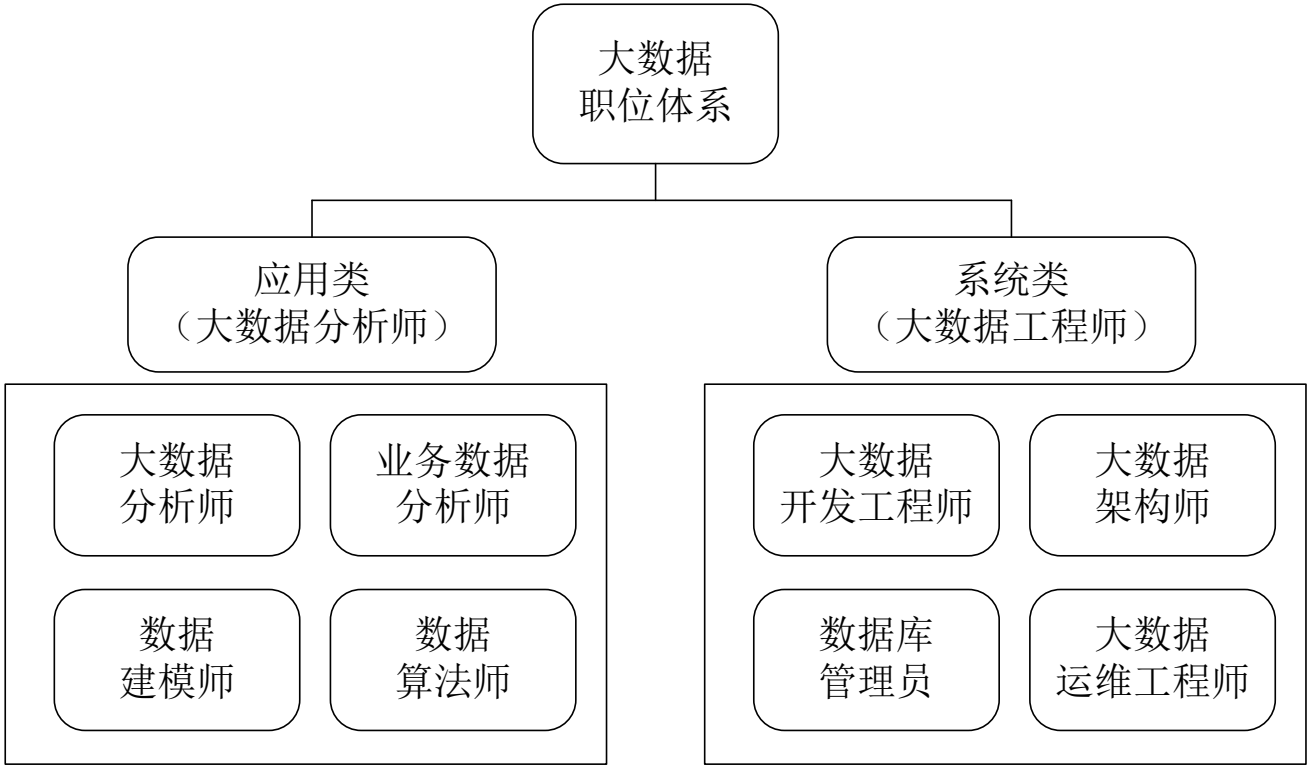
1.9 高校大数据专业

1.9.1 大数据专业的人才培养目标
1.9.2 毕业生就业岗位
1.9.3 大数据专业知识体系
1.9.4 大数据专业课程体系
1.9.5大数据专业的编程语言

1.9.1 大数据专业的人才培养目标

大数据专业致力于培养符合国家战略及大数据产业发展需求，具备较好的数据素养和数理基础、扎实的编程基础以及大数据基础知识与技能，熟练掌握大数据采集、预处理、存储、处理、分析、应用技术，能够运用大数据思维、模型和工具解决实际问题的高级复合型人才。大数据专业的毕业生能在互联网企业、金融机构、科研院所、高等院校等从事大数据分析、挖掘、处理、服务、应用和研究工作，亦可从事各行业大数据系统的集成、设计、开发、管理、维护等工作，也适合在高等院校及科研院所的相关交叉学科继续深造。

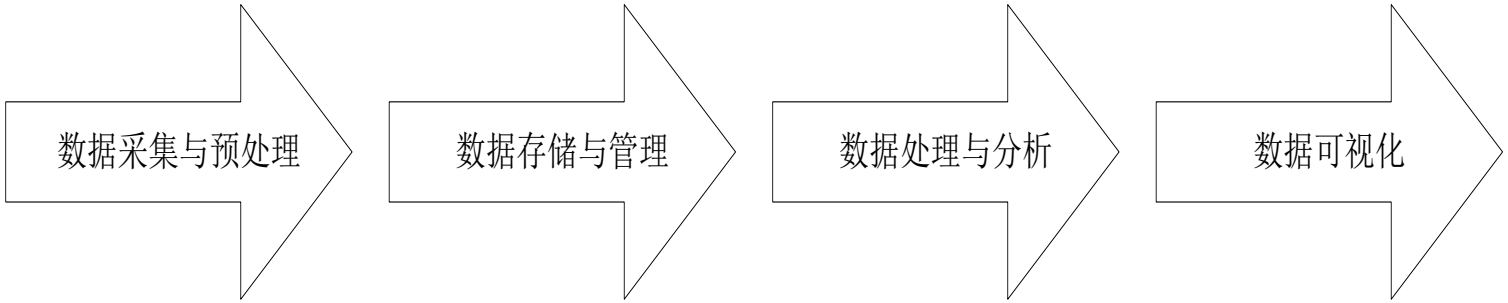
1.9.2 毕业生就业岗位


大数据
职位体系
应用类
（大数据分析师）
系统类
（大数据工程师）
大数据
分析师
业务数据
分析师
数据
建模师
数据
算法师
大数据
开发工程师
大数据
架构师
数据库
管理员
大数据
运维工程师

1.9.3 大数据专业知识体系

从学科角度而言，大数据可以理解为一个跨多学科领域的，从数据中获取知识的科学方法、技术和系统的集合。因此，大数据专业知识体系涵盖了计算机、数学、统计学等多个学科领域，结合了诸多领域中的理论和技术，包括应用数学、统计学、模式识别、机器学习、人工智能、深度学习、数据可视化、数据挖掘、数据仓库、分布式计算、云计算、系统架构设计等。

从大数据分析角度而言（如图所示），典型的大数据分析过程包括：数据采集与预处理、数据存储与管理、数据处理与分析、数据可视化等。因此，大数据专业知识体系涵盖了数据采集与预处理技术、数据存储与管理技术、数据处理与分析技术、数据可视化技术等。同时，在分析过程中，对商业领域的业务知识也需要一定的理解。



1.9.4 大数据专业课程体系

大数据专业课程体系涵盖通识教育课、学科基础课、专业基础课、专业核心课和专业课，具体如下：

（1）通识教育课：思政类课程、军体类课程、外语课、创新创业课等；

（2）学科基础课：高等数学、线性代数、概率论与数理统计等；

（3）专业基础课：程序设计、计算机系统基础及组成原理、离散数学、计算机网络、算法与数据结构、数据库系统、操作系统、软件工程等；

（4）专业核心课：大数据导论、网络爬虫与数据采集、数据清洗、NoSQL数据库、数据可视化、分布式并行编程、机器学习等；

（5）专业课：云计算、数据安全、数据仓库、数据挖掘等。

1.9.5大数据专业的编程语言

1.C语言
2.C++
3.Java
4. Python
5.Scala
6.R语言

1.9.5大数据专业的编程语言

1. C语言

C语言是一门面向过程的计算机编程语言，与C++、Java等面向对象编程语言有所不同。C语言的设计目标是提供一种能以简易的方式编译、处理低级存储器、仅产生少量的机器码以及不需要任何运行环境支持便能运行的编程语言。C语言描述问题比汇编语言迅速、工作量小、可读性好、易于调试、修改和移植，而代码质量与汇编语言相当。C语言一般只比汇编语言代码生成的目标程序效率低10%～20%。因此，C语言可以编写系统软件。C语言在一些编程语言排行榜中长期排在第一的位置。

1.9.5大数据专业的编程语言

1. C语言

C语言具有很多优点，主要如下：
（1）它具有现代高级程序设计语言的基本语法特征，并且是编写操作系统的首选语言，与计算机硬件打交道时灵巧且高效，目前几乎所有的操作系统（如Windows、Unix和Linux等）均是由C语言编写的；
（2）常用的面向对象程序设计语言（例如C++和Java），其基本语法源于C语言。C语言甚至是其它编程语言的母语言，比如Java语言就是用C语言编写的。
（3）简洁紧凑，灵活方便。C语言一共只有32个关键字，9种控制语句，程序书写自由，主要用小写字母表示，它把高级语言的基本结构和语句与低级语言的实用性结合了起来。
C语言一般作为学习计算机程序设计语言的入门语言。

1.9.5大数据专业的编程语言

2.C++

C++是C语言的继承，是一门以C为基础发展而来的、面向对象的高级程序设计语言，它既可以进行C语言的过程化程序设计，又可以进行以继承和多态为特点的面向对象的程序设计。C++不仅拥有计算机高效运行的实用性特征，同时还致力于提高大规模程序的编程质量与程序设计语言的问题描述能力。

C++的优点主要包括：
（1）实现了面向对象程序设计，处理运行速度非常快，大部分的游戏软件都是由C++来编写的。
（2）语言非常灵活，功能非常强大。
（3）非常严谨、精确和数理化，标准定义很细致。
（4）语言的语法思路层次分明。

1.9.5大数据专业的编程语言

2.C++

大数据领域的不少产品都是使用C++开发的（即产品本身是由C++编写的），包括一些NoSQL数据库（ScyllaDB、MongoDB、Aerospike、Kudu、SequoiaDB）、数据仓库Impala、实时流计算框架Hurricane和Heron、资源调度框架Mesos等。

但是，谈到大数据开发语言，C++要明显逊色于Java，很多大数据应用程序（比如Hadoop程序等）都是使用Java开发的，而不是使用C++。

1.9.5大数据专业的编程语言

3.Java

Java是目前最热门的编程语言之一，在一些编程语言排行榜中长期排在前三名。虽然Java没有和 R、Python一样好的可视化功能，也不是统计建模的最佳工具，但是，如果需要建立一个庞大的应用系统，那么Java通常会是较为理想的选择。由于 Java具有简单、面向对象、分布式、鲁棒、安全、体系结构中立、可移植、高性能、多线程以及动态性等诸多优良特性，因此，被大量应用于企业大型系统开发中，企业对于Java人才的需求一直比较旺盛。

Java语言与大数据存在较为紧密的联系，Java在大数据领域有着广泛的应用，是大数据应用程序开发的常用语言。作为大数据领域热门的大数据处理框架Hadoop和Flink等，其框架本身都是采用Java语言开发的，编写Hadoop应用程序也首选Java语言。而目前热门的分布式计算框架Spark，也支持采用Java语言编写应用程序。

1.9.5大数据专业的编程语言

4. Python

Python是目前国内外很多大学里流行的入门语言，学习门槛低，简单易用，开发员可以使用Python来构建桌面应用程序和Web应用程序，此外，Python在学术界备受欢迎，常被用于科学计算、数据分析和生物信息学等领域。Python是最近几年发展最为迅速的编程语言，在一些编程语言排行榜当中甚至已经进入了前三名。

Python的主要优点如下：

（1）可以使用多种执行方式。可以直接在命令行执行相关命令，也可以用函数的方式执行相关命令，或者也可以用面向对象的方式执行相关命令。

（2）语法简洁，且强制缩格，程序具有很好的可读性。

（3）跨平台。支持多种开发平台，如Windows、Linux、Mac OS X、Solaris等。

（4）面向对象。Python既支持面向过程，又支持面向对象，这使得其编程更加灵活。

（5）丰富的第三方库。Python有丰富且强大的库，而且由于Python的开源特性，第三方库非常多，如Web开发、爬虫、科学计算等。

在数据分析领域，Python是广受欢迎的编程语言，网络数据采集（比如网络爬虫）、数据清洗、数据分析与挖掘、数据可视化等环节，通常都使用Python语言编写程序。

1.9.5大数据专业的编程语言

5.Scala

Scala是一门类似Java的多范式语言，它整合了面向对象编程和函数式编程的最佳特性，具有诸多优点，主要包括以下几个方面：
（1）具备强大的并发性，支持函数式编程，可以更好地支持分布式系统；
（2）Scala兼容Java，可以与Java互操作；
（3）Scala代码简洁优雅；
（4）Scala支持高效的交互式编程；
（5）Scala是Spark的开发语言。
Spark是当前热门的大数据处理技术，开发Spark应用程序时，首选编程语言是Scala，因为Spark框架自身就是使用Scala语言开发的，用Scala语言编写Spark应用程序，可以获得最高的性能。Spark的流行也迅速提升了Scala的影响力。流计算框架Flink的部分模块也是使用Scala语言开发的，也可以使用Scala语言编写Flink应用程序。

1.9.5大数据专业的编程语言

6.R语言

R是专门为统计和数据分析开发的语言，具有数据建模、统计分析和可视化等功能，简单易上手。R语言主要具有如下优点：

（1）免费开源。R的源代码可以自由下载使用，也有已编译的可执行文件版本可以下载。

（2）简单易学。虽然R与其他程序设计语言相比结构相对松散，使用变量前不需要明确定义变量类型等，但是，仍然保留了程序设计语言的基础逻辑与自然的语言风格。

（3）几乎兼容全部平台。除了支持OS X、Linux、Windows之外，甚至可以在iOS设备上编辑和运行R程序，还可以在iPhone等移动设备上安装R程序。

（4）多领域的统计资源。学者和数据分析师开发了很多R语言包，涉及到统计的各个方面，资源很丰富。

（5）出色的图形统计功能。除了基本统计直方图、折线图等，还可以绘制一些高级的图形，而这些是SPSS这类软件所不能匹敌的。

总体而言，R和Python都是比较流行的数据分析语言。相对而言，数学和统计领域的工作者更多使用R语言，而计算机领域的工作者更多使用Python。大数据处理框架Spark也提供了对R语言的支持。

1.10 本章小结

人类已经步入大数据时代，我们的生活被数据所“环绕”，并被数据深刻变革。作为大数据时代的公民，我们应该接近数据，了解数据，并利用好数据。因此，本章首先从数据入手，讲解了数据的概念、类型、组织形式、数据价值等内容，然后，把视角切入到大数据时代，介绍了大数据时代到来的背景及其发展历程。接下来，讨论了大数据的“4V”特性以及大数据对科学研究、社会发展、就业市场和人才培养的影响，并简要介绍了大数据在不同领域的应用和大数据产业。最后，对高校大数据专业的建设做了简要探讨。

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），全国高校知名大数据教师，现为厦门大学计算机科学系副教授，曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革局副局长。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、2017年度和2020年度厦门大学教学类奖教金获得者，荣获2019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项国家自然科学青年基金项目(No.61303004)、1项福建省自然科学青年基金项目(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1000万字高价值的研究和教学资料，累计网络访问量超过1000万次。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200万次，累计访问量超过1000万次。

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：林子雨大数据系列教材

林子雨大数据系列教材

用于导论课、专业课、实训课、公共课

了解全部教材信息：http://dblab.xmu.edu.cn/post/bigdatabook/

附录D：《大数据导论（通识课版）》教材

开设全校公共选修课的优质教材

本课程旨在实现以下几个培养目标：

- 引导学生步入大数据时代，积极投身大数据的变革浪潮之中
- 了解大数据概念，培养大数据思维，养成数据安全意识
- 认识大数据伦理，努力使自己的行为符合大数据伦理规范要求
- 熟悉大数据应用，探寻大数据与自己专业的应用结合点
- 激发学生基于大数据的创新创业热情

高等教育出版社 ISBN:978-7-04-053577-8 定价：32元

教材官网：http://dblab.xmu.edu.cn/post/bigdataintroduction/

附录E：《大数据技术原理与应用》教材

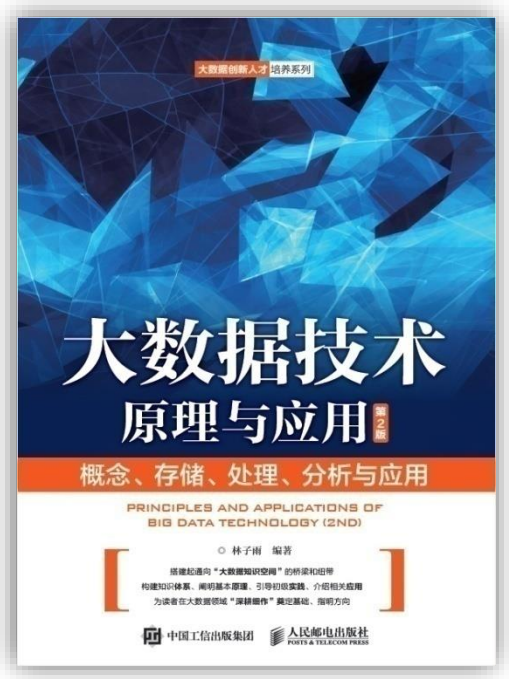
《大数据技术原理与应用——概念、存储、处理、分析与应用（第2版）》，由厦门大学计算机科学系林子雨博士编著，是国内高校第一本系统介绍大数据知识的专业教材。人民邮电出版社　ISBN:978-7-115-44330-4　定价：49.80元

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://dblab.xmu.edu.cn/post/bigdata

扫一扫访问教材官网



附录F：《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

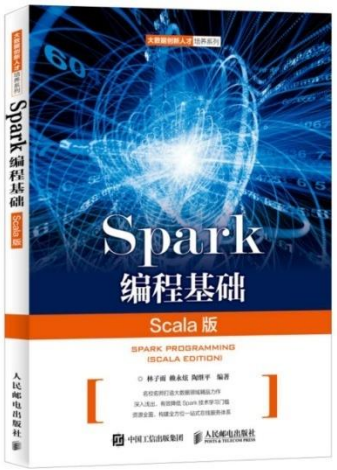
厦门大学林子雨编著

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

林子雨编著《大数据基础编程、实验和案例教程》
清华大学出版社 ISBN:978-7-302-47209-4 定价：59元

附录G：《Spark编程基础（Scala版）》



《Spark编程基础（Scala版）》

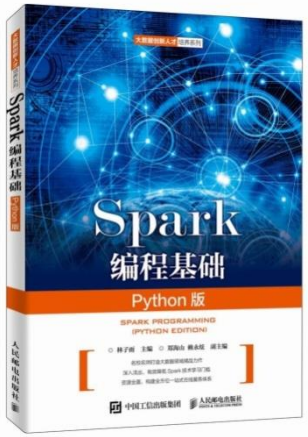
厦门大学 林子雨，赖永炫，陶继平 编著

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://dblab.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H：《Spark编程基础（Python版）》



《Spark编程基础（Python版）》

厦门大学 林子雨，郑海山，赖永炫 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:978-7-115-52439-3
教材官网： http://dblab.xmu.edu.cn/post/spark-python/

本书以Python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、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、Structured 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、上机实验指南等。

附录I：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/

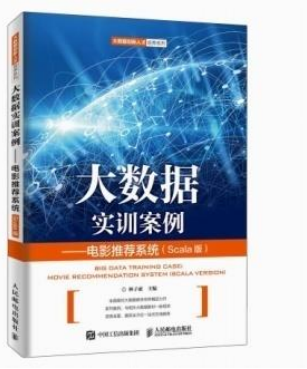
扫一扫访问平台主页

扫一扫观看3分钟FLASH动画宣传片

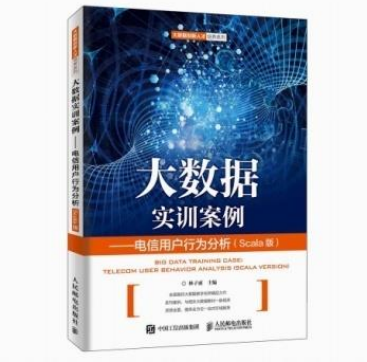
附录J：高校大数据实训课程系列案例教材

为了更好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：

《电影推荐系统》（已经于2019年5月出版）
《电信用户行为分析》（已经于2019年5月出版）
《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

Department of Computer Science, Xiamen University, 2020