大数据分析

海量数据计算研究中心

杨东华

任课教师简介

杨东华 副教授，博士生导师

- 2008年获计算机软件与理论专业博士学位。目前在分析测试与计算中心任副主任，主要负责高性能计算相关的工作；同时在海量数据计算中心任副教授，博士生导师。主要从事大数据管理、大数据分析等方面的研究工作。
- 近年来，在国内外知名期刊和会议上发表论文30余篇，作为负责人主持国家、省部级等项目10余项，其中主持国家自然科学基金3项；参与“973”国家重点基础研究发展计划项目和国家自然科学基金重点项目各1项。
- 已经培养硕士研究生14名，毕业6名。目前在读博士生2名。

联系方式

电话：13766863397

邮箱：yang.dh@hit.edu.cn

QQ: 2899207

办公地址：科学园2H栋301房间；（计算中心）

科学园科创大厦K1417房间（实验室）

苗东菁 副教授，博士生导师

2018年获佐治亚州立大学计算机科学博士学位。现于海量数据计算中心从事大数据计算理论与算法的研究工作。

联系方式

电话：13091431693

邮箱：miaodongjing@hit.edu.cn

QQ: 251969680

办公地址：科学园科创大厦K1418房间

《大数据分析》课程介绍

课程概览

课程设置的目的和意义

- 为了使学生系统而全面地掌握与大数据分析相关的基础知识，了解和认识大数据分析领域的前沿成果，建立针对大数据分析问题的思维方式。
- 学校设置了大数据专业，海量数据计算中心开设了三门学科方向性课程：
  - ✓ 大数据计算基础（大三秋季学期）
  - ✓ 大数据分析（大三春级季学期）
  - ✓ 大数据挖掘（大四秋季学期）

课程基本信息介绍

课程编号：CS32272

课程名称：大数据分析

英文名称：Big Data Analytics

课程学时：72　讲课学时：48　实验学时：24

上机学时：0　习题学时：0

课程学分：4.5

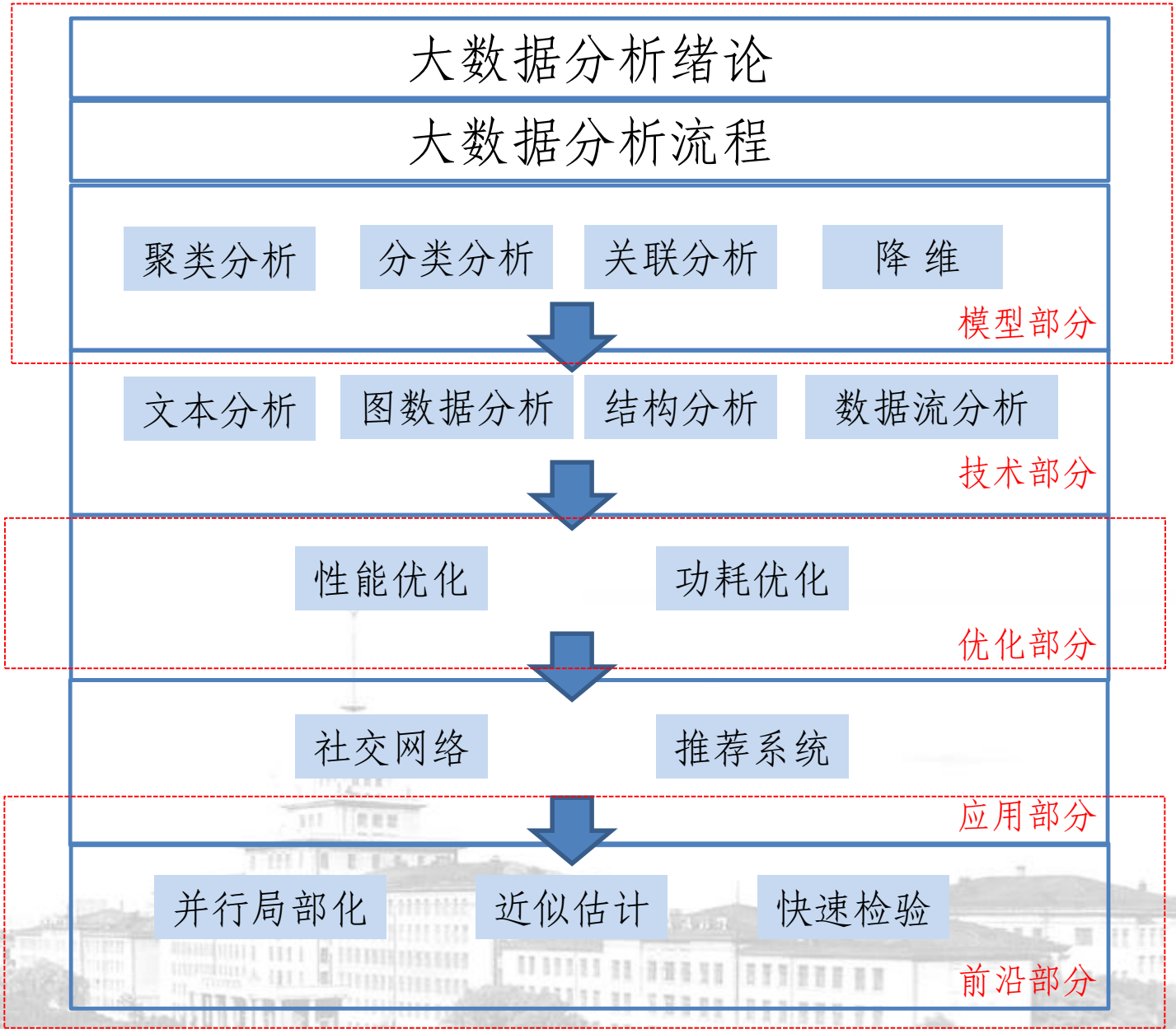
授课对象：计算机大类专业

开课学期：3春

先修课程：数据结构与算法、计算机系统、计算机网络、数据库系统、软件工程、程序语言设计、大数据计算基础

课程要求：学科方向性课程

课程脉络图


大数据分析绪论
大数据分析流程
聚类分析
分类分析
关联分析
降 维
模型部分
文本分析
图数据分析
结构分析
数据流分析
技术部分
性能优化
功耗优化
优化部分
社交网络
推荐系统
应用部分
并行局部化
近似估计
快速检验
前沿部分

课程脉络图

作业与实验

概念型作业；
设计型作业.

概念型作业；
计算型作业.

概念型作业；
计算型作业；
设计型作业；
综合型作业.

概念型作业；
设计型作业；
创新型作业.

系统搭建 → 准备数据 → 建立模型 → 分析优化 → 前沿把握 → 应用实践

实验一： 数据预处理

实验二：
聚类分析
分类分析

实验三：
分析性能优化

实验四：
近似估计
并行局部化
快速检验

课程的目标

1. 掌握和应用大数据基础知识和技术;

2. 能够对大数据分析平台和算法性能进行分析和评价;

3. 能够设计和搭建大数据分析系统;

4. 撰写大数据分析相关报告。

课程内容一：大数据分析绪论

- 了解什么是大数据分析

定义、意义、应用场景等

- 理解大数据分析涉及的关键技术(*)

数据采集、数据管理、基础架构、数据的理解与提取、统计分析、数据挖掘、数据可视化等

- 理解大数据分析的难点(*)

可扩展性、可用性、要与具体领域知识相结合、结果检验

- 了解大数据分析的前沿成果

课程内容二：大数据分析流程

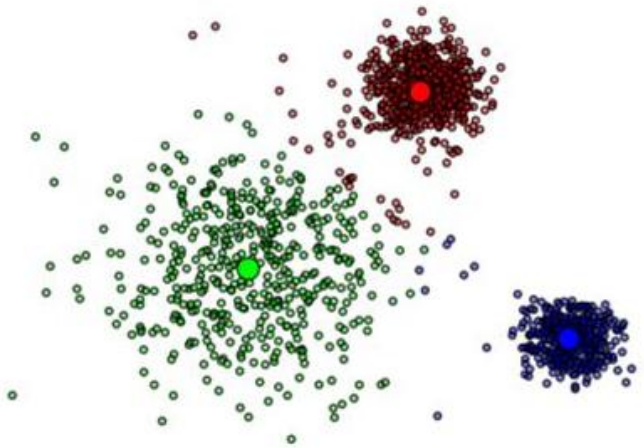
- 大数据的采集和存储
- 大数据预处理(*)
- 大数据分析建模(*)
- 大数据分析方法
- 大数据分析结果展示及评估

课程内容三：大数据统计分析

- 聚类分析
- 分类分析
- 关联分析
- 大数据降维

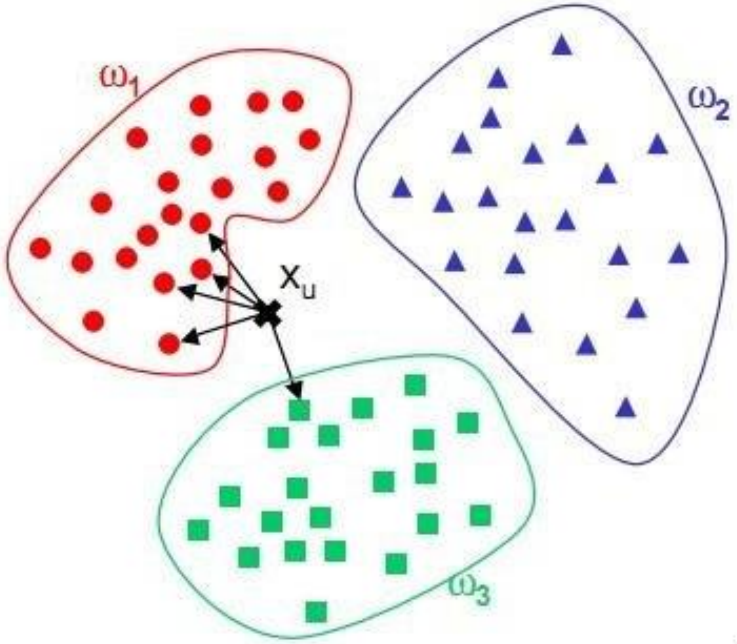
课程内容三：大数据统计分析

- 聚类分析
- 分类分析
- 关联分析
- 大数据降维



课程内容三：大数据统计分析

- 聚类分析
- 分类分析
- 关联分析
- 大数据降维



课程内容三：大数据统计分析

- 聚类分析
- 分类分析
- 关联分析
- 大数据降维

			类型						
			小说						
			小说						
			散文						
			散文						
			诗歌						
			诗歌						
			小说						
			小说						
			散文						
			诗歌						
			散文						
			小说						

课程内容三：大数据统计分析

- 聚类分析
- 分类分析
- 关联分析
- 大数据降维

			类型						
			小说						
			小说						
			散文						
			散文						
			诗歌						
			诗歌						
			小说						
			小说						
			散文						
			诗歌						
			散文						
			小说						

课程内容三：大数据统计分析

- 聚类分析
- 分类分析
- 关联分析
- 大数据降维

			类型						
			小说						
			小说						
			散文						
			散文						
			诗歌						
			诗歌						
			小说						
			小说						
			散文						
			诗歌						
			散文						
			小说						

课程内容四：大数据分析性能优化

- 分析查询性能优化策略
- 分析平台功耗优化策略

课程内容五：大数据分析算法前沿专题

在工程和应用视角之后，我们最后介绍理论计算机科学视角下的大数据分析算法研究前沿:

- 解近似估计：*k*-Median/Center
- 解代价近似估计：Vertex Cover
- 并行与局部化计算：Maximal Independent Set
- 快速检验：Property Testing

课程内容五：大数据分析算法前沿专题

1. 解近似估计：*k*-Median/Center

很多优化问题，其优化解通常只能在平方、立方等时间级别上给出近似解。但在大数据环境下，平方、立方等时间难以接受，因此需要研究如何在大数据环境下，快速给出近似解。我们将以聚类分析中经典问题“*k*-中位数”问题为例，介绍此类方法的最新研究结果。

课程内容五：大数据分析算法前沿专题

2. 解代价近似估计： Vertex Cover

当优化解规模仍然很大，很难快速输出的时候，通常会只求解其代价。我们将以图数据分析中经典问题“顶点覆盖”问题为例，介绍此类方法的最新研究结果。

课程内容五：大数据分析算法前沿专题

3. 并行与局部化计算：Maximal Independent Set

有时并不需要完整解，而只需要部分解。对于此类应用，通过并行算法局部化计算可以快速给出结果，我们将以社交数据分析中经典问题“极大独立集”问题为例，介绍此类方法的最新研究结果。

课程内容五：大数据分析算法前沿专题

4. 快速检验： Property Testing

最后将以统计分析中性质检验系列问题为例，介绍一大类统计性质快速检验方法的最新研究结果。典型的性质检验问题包括离散统计分布相似度检验等。

通过 系统、工程、理论、应用等四个角度给大家介绍大数据分析的前沿结果。

实验环节：覆盖大数据分析课程各个环节

实验内容	实验要求	学时
预处理 （实验一）	能够使用Hadoop/Spark平台，编程实现数据整合、数据预处理方法。	4
多元统计分析 （实验二）	能够使用Hadoop/Spark平台，使用高级编程语言编程实现聚类分析、分类分析等算法，并通过使用多元统计工具SSPS进行检验	8
大数据分析性能优化 （实验三）	能够编程实现对Hadoop/Spark系统中一个功能模块的优化；编程实现对现有工作中Hadoop/Spark算法的优化，并分析性能提升的效果和原因，完成项目开发。	8
大数据分析算法前沿 （实验四）	能够利用Hadoop/Spark系统，编程实现精确算法；编程实现前沿算法，实验对比性能提升的效果和求解精度的优劣。	4

考核方式

1. 开放式大作业（15%）

- 大作业题目：基于Hadoop/Spark的大数据分析及性能评估。
- 大作业内容：根据目前的大数据应用场景，如工业、医疗及科学领域，设计一款大数据分析的应用，要求覆盖从需求分析到系统设计、实现、评估的全过程，并撰写课程报告。

2. 实验（15%）

- 将项目开发分解为数据预处理、统计分析、大数据分析优化、大数据分析算法前沿等实验内容。共4次实验，24学时。

考核方式

3. 期末考试（60%)

期末考试覆盖所有教学内容，其比例为:

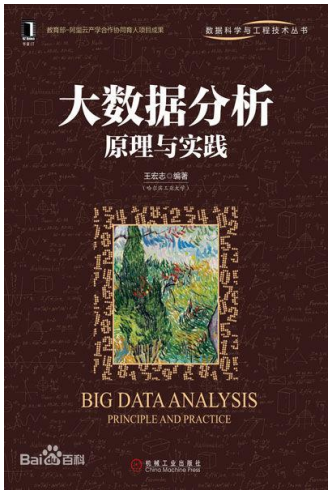
- 大数据分析基本概念（15%）;
- 大数据分析流程（10%）;
- 多元统计分析选讲（ 30% ）;
- 大数据分析性能优化（15%）;
- 大数据分析算法前沿（30%）

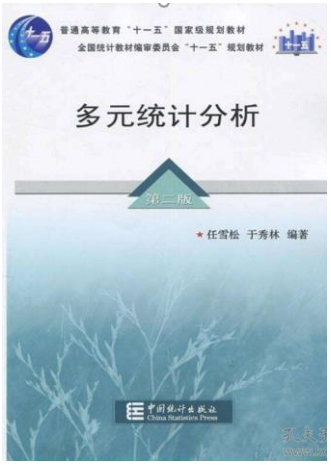
4. 课堂听课、随堂作业（10%)

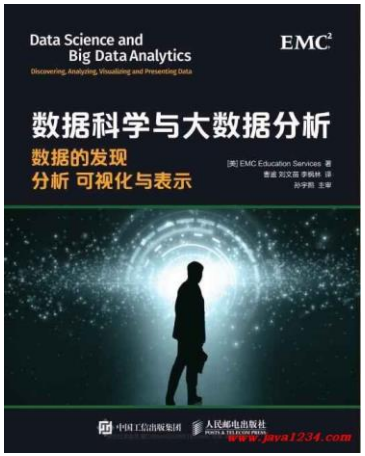
作业提交：hit_bda@163.com

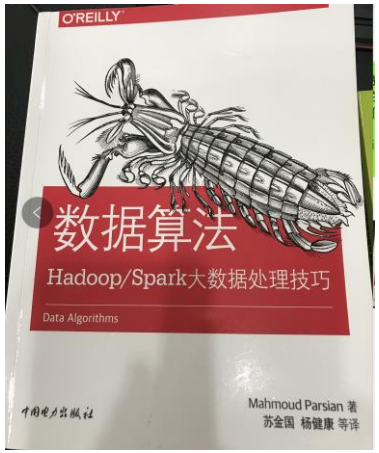
考核方式

参考书目









助教：

- 董洁琳（研二）：2657391549
- 夏维邑（研二）：407955243

实验部分

- 宋新彤（研一）：2039358225
- 鲁文博（研一）：805045718

课堂+作业

课程QQ群

命名方式：序号+学号+姓名

例如：

A01+1170300301+唐秋原；

B07+1180300201+赵翼

第 1 讲 绪论

目录

目录

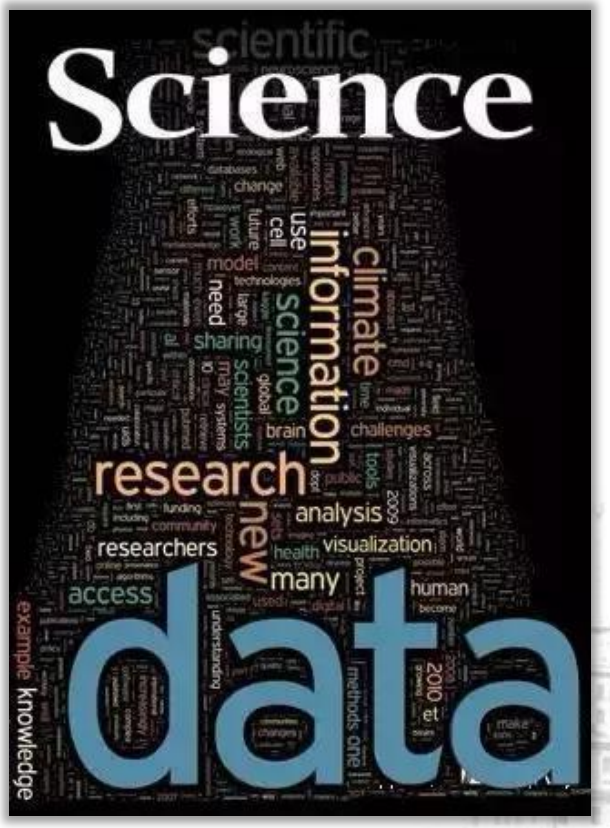
什么是大数据

时至今日，“数据”变身“大数据”，大数据开启了一次重大的时代转型。

“大数据”这一概念的形成，有三个标志性事件:

2008年9月，国际顶级科学期刊《自然》（Nature）杂志，出版专刊—The next google，第一次正式提出“大数据”概念。

什么是大数据



2011年2月1日，国际顶级科学期刊《科学》（Science）杂志出版专刊——Dealing with data。通过社会调查的方式，第一次综合分析了大数据对人们生活造成的影响，详细描述了人类面临的“数据困境”。

什么是大数据

2011年5月，麦肯锡研究院发布报告—Big data: The next frontier for innovation, competition, and productivity。第一次给大数据做出相对清晰的定义：“大数据是指其大小超出了常规数据库工具获取、储存、管理和分析能力的数据集。”

什么是大数据

1. 大数据的定义

但是，至今没有公认的定义。比较容易被接受的定义有如下三个:

定义1:

大数据是指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。

——维基百科

什么是大数据

1. 大数据的定义

定义2:

不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。

——《大数据时代》

（维克托·迈尔-舍恩伯格与肯尼斯·库克耶著）

什么是大数据

1. 大数据的定义

定义3:

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。

——研究机构 Gartner

什么是大数据

2. 大数据的背景

(1) 数据量的迅猛增长

- 近年来，人们能明显的感受到大数据的来势迅猛。
- 据有关资料显示，全球网民平均每月使用流量:
  - ✓ 1998年是1MB
  - ✓ 2003年是100MB
  - ✓ 2014年是10GB
- 全网流量累计达到1EB(即10亿GB)所用的时间:
  - ✓ 在2001年是一年
  - ✓ 在2004年是一月
  - ✓ 在2013年仅需要一天，即一天产生的信息量可刻满1.88亿张DVD光盘。

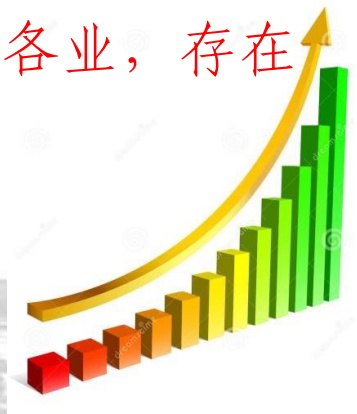
什么是大数据

2. 大数据的背景

(1) 数据量的迅猛增长

- 我国网民数量居世界首位，所产生的数据量位于世界前列。
  - ✓淘宝网站每天超数千万次的交易所产生的超50TB的数据
  - ✓百度搜索每天生成的几十PB数据

总之，大到学校，医院，银行，企业的系统行业信息，小到个人的一次百度搜索，一次地铁刷卡，大数据存在于各行各业，存在于民众生活的边边角角。



什么是大数据

2. 大数据的背景

(2) 大数据因自身可挖掘的高价值而受到重视

- ✓ 国家的宽带化战略的实施，云计算服务的起步，物联网的广泛应用和移动互联网崛起的同时，数据处理能力也迅速发展，数据积累到一定程度，其资料属性越明晰，显示出开发的价值。

- ✓ 同时，社会的节奏越来越快，要求快速反应和精细管理，急需借助对数据的分析做出科学的决策。

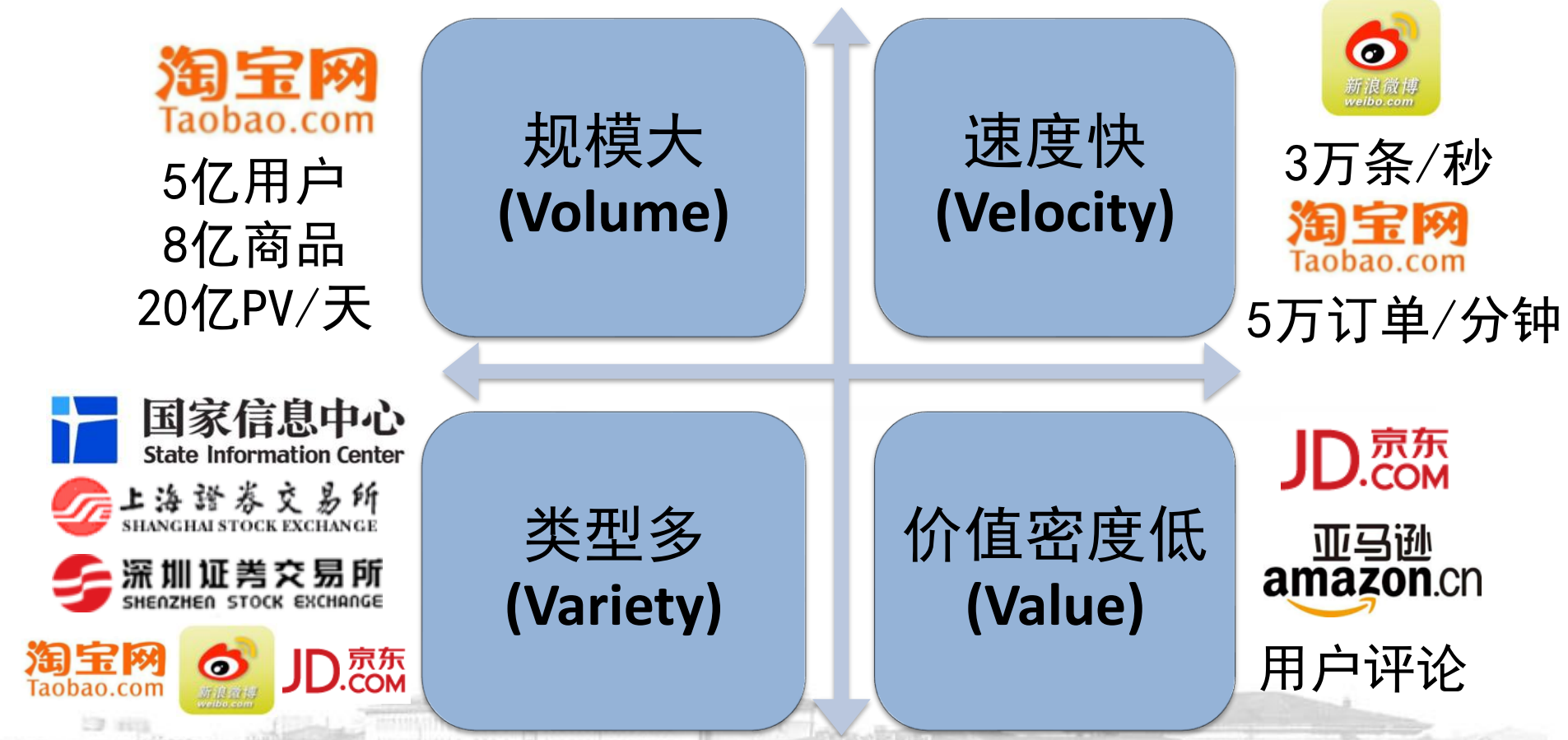
什么是大数据

2. 大数据的背景

大数据的时代到来了！

什么是大数据

3. 大数据的特点



什么是大数据

3. 大数据的特点

(1) Volume 规模性

- 规模大是大数据的最重要的标志之一。大数据之“大”，体现在数据的存储和计算均需要耗费海量规模的资源。
  - ✓ 从2013年至2020年，人类的数据规模扩大了50倍，每年产生的数据量增长到44万亿GB，相当于美国国家图书馆数据量的数百万倍，且每18个月翻一番。
- 数据的规模越大，通常对数据进行挖掘所得到的事物的演变规律越可信，数据的分析结果也越具有代表性。
  - ✓ 美国宇航局收集和处理的气候观察、模拟数据达到32PB;
  - ✓ 而FICO(费埃哲)的信用卡欺诈检测系统要监测全世界超过18亿个活跃信用卡账户。从而对个人消费信用进行评估。

什么是大数据

3. 大数据的特点

(2) Velocity 高速性

- 随着现代感测、互联网、计算机技术的发展，数据生成、储存、分析、处理的速度远远超出人们的想象力。
  - ✓ 大型强子对撞机实验设备中包含15亿个传感器，平均每秒钟收集超过4亿的实验数据;
  - ✓ 同样在一秒钟里，有超过3万次用户查询提交到谷歌，3万微博被用户撰写和更新。

什么是大数据

3. 大数据的特点

(3) Variety 多样性

- 与传统数据相比，大数据在来源和形式上的多样性愈加突出。
  - ✓ 除以结构化形式存在的文本数据，网络上也存在大量的位置、图片、音频、视频等非结构化信息。有数据表明，2016年，全部互联网流量中，视频数据达到55%，那么，有理由相信，大数据中90%都将是非结构化数据。
  - ✓ 大数据不仅仅在形式上表现出多元化，其信息来源也表现出多样性。大致可将其分为网络数据、企事业单位数据、政府数据、媒体数据等几种。

什么是大数据

3. 大数据的特点

(4) Value 高价值性

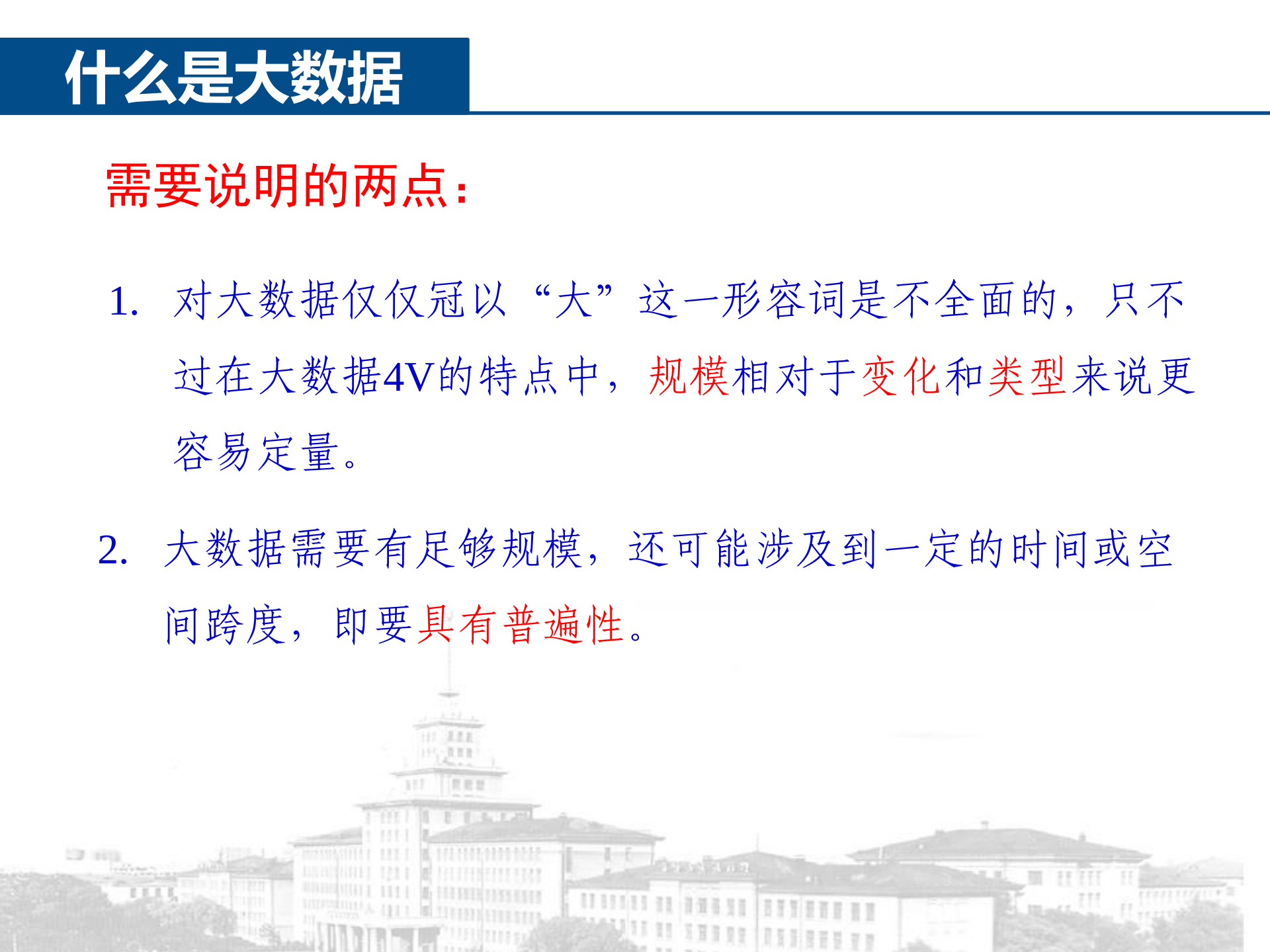
- 大数据有巨大的潜在价值，但同其呈几何指数爆发式增长相比，某一对象或模块数据的价值密度较低。IBM副总裁CTO Dietrich表示“可以利用Twitter数据获得用户对某个产品的评价，但是往往上百万记录中只有很小的一部分真正讨论这款产品”。
- 同一事件的不同数据集即便有相同的规模其价值也可以相差很多。例如，对同一观察对象收集的长时间稀疏数据和短时间密集数据，它们的“含金量”也可能是不同的。

什么是大数据

需要说明的两点：

1. 对大数据仅仅冠以“大”这一形容词是不全面的，只不过在大数据4V的特点中，规模相对于变化和类型来说更容易定量。

2. 大数据需要有足够规模，还可能涉及到一定的时间或空间跨度，即要具有普遍性。



目录

哪里有大数据 —处处皆是大数据

1. 社交网络/社会计算

互联网、移动网络和物联网产生大量数据。

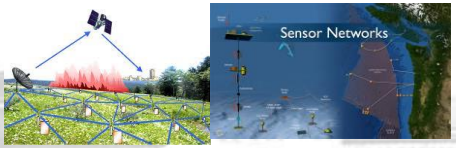
- Google公司通过大规模集群和MapReduce软件，每月处理的数据量超过400PB;
- Facebook注册用户超过10亿，每天需存储、访问和分析超过30PB的用户数据;
- 淘宝网在2019年6月拥有7.55亿用户，在线商品超过10亿件，每天交易超过数千万笔，单日数据产生量近100TB。

哪里有大数据 —处处皆是大数据

2. 电信、金融、电力等行业

- 电信行业，年度用户数据增长超过10%，比如，中国移动“大云”对7亿3千万以上用户的数据进行分析，将用户偏好和关注热点等进行归类，用于改善用户体验和辅助市场决策;
- 金融行业每年产生的数据超过数十PB;
- 国家电网采集获得的数据总量就达到了数十PB;
- 石油化工领域每年产生和保存下来的数据量也将近百PB级别。

移动设备



传感网

哪里有大数据 —处处皆是大数据

3. 科学研究

- 欧洲核子研究中心的大型强子对撞机(Large Hadron Collider, LHC)每年产生大约15PB的数据，这些数据足以刻满超过170万张的双面DVD光盘。
- 大型综合巡天望远镜（Large SynopticSurvey Telescope，LSST）将数十年持续观测整个天空的相关数据记录下来以供分析研究。在项目已经持续的十年中，大约会将产生的60PB数据存储到服务器中供分布在世界各地的科学家们协作分析数据，以便对整个宇宙做出新的发现。

哪里有大数据 —处处皆是大数据

4. 商业数据

商业活动产生大量的数据。

- 沃尔玛服务器每小时需要处理一百多万条顾客交易，这些信息被存储到超过2.5PB大小的数据库中；
- 劳斯莱斯公司对全世界数以万计的飞机引擎进行实时监控，每年传送PB数量级的数据。

哪里有大数据 —处处皆是大数据

5. 医疗卫生

- 医疗数据比如医药记录，收费记录，病例记录，处方药记录，X光片记录，磁共振成像记录，CT影像记录等都在不同程度上向数字化转化。

- 有报告显示，2011年，仅美国的医疗健康系统数据量就达到了150EB。医疗大数据分析应用将在提高医疗质量，强化患者安全，降低风险，降低医疗成本等方面发挥着巨大作用。

6. 制造业

制造业的大数据类型以产品设计数据、企业生产环节的业务数据和生产监控数据为主。

- 产品设计数据以文件为主，非结构化，共享要求较高，保存时间较长；
- 企业生产环节的业务数据主要是数据库结构化数据；
- 生产监控数据的数据量非常大。

目录

大数据的技术支撑

1. 存储：存储成本的下降

云计算出现之前	云计算出现之后
在云计算出现之前，数据存储的成本是非常高的。 例如，公司要建设网站，需要购置和部署服务器，安排技术人员维护服务器，保证数据存储的安全性和数据传输的畅通性，还会定期清理数据，腾出空间以便存储新的数据，机房整体的人力和管理成本都很高。	云计算出现后，数据存储服务衍生出了新的商业模式，数据中心的出现降低了公司的计算和存储成本。 例如，公司现在要建设网站，不需要去购买服务器，不需要去雇用技术人员维护服务器，可以通过租用硬件设备的方式解决问题。

存储成本的下降，使大家愿意把更久远的历史数据保存下来，有了历史数据的沉淀，才可以通过对比，发现数据之间的关联和价值。正是由于存储成本的下降，才能为大数据搭建最好的基础设施。

大数据的技术支撑

2. 计算：计算速度越来越快

海量数据从原始数据源到产生价值，期间会经过存储、清洗、挖掘、分析等多个环节，如果计算速度不够快，很多事情是无法实现的。所以，在大数据的发展过程中，计算速度是非常关键的因素。

- 分布式系统基础架构Hadoop的出现，为大数据带来了新的曙光；
- HDFS为海量的数据的存储提供了便利；
- MapReduce则为海量的数据提供了并行计算，从而大大提高了计算效率；
- Spark、Storm、Impala等各种各样的技术进入人们的视野。

大数据的技术支撑

3. 人工智能：机器拥有理解数据的能力

大数据带来的最大价值就是“智慧”，大数据让机器变得有智慧，同时人工智能进一步提升了处理和理解数据的能力。例如：

1	谷歌AlphaGo大胜世界围棋冠军李世石
2	iPhone上智能化语音机器人Siri
3	微信上与大家聊天的微软小冰

目录

什么是大数据分析

1. 大数据分析的定义

数据分析指的是用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。

什么是大数据分析

2. 大数据分析的三个层次

数据分析可以分为三个层次：描述分析、预测分析和规范分析。

描述分析

- 探索历史数据并描述发生了什么
- 包括聚类、相关规则挖掘、模式发现和可视化分析

- 描述性的数据分析，是数据分析中最简单的一个类型。
- 描述性分析的目的是探索历史数据并总结发生了什么事。

  超过80%的业务分析，特别是社会分析信息是描述性的。
- 包括聚类、相关规则挖掘、模式发现和可视化分析。

什么是大数据分析

2. 大数据分析的三个层次

预测分析

- 预测未来的概率和趋势
- 如基于逻辑回归的预测、基于分类器的预测等

- 预测性的数据分析，是利用各种统计、建模、数据挖掘工具对最近的数据和历史数据进行研究，从而对未来进行预测。
- 其目的并不是要准确告诉你将来会发生什么，只能预测未来会发生什么。因为所有的预测分析在本质上都只是一个概率。
- 比如分析情绪是一个常见的预测分析，这是一个纯文本的输入模型，该模型的输出结果是一个情绪的得分，包括积极的、消极的，或者中立的。在这种情况下，模型计算的得分，不一定预测到未来，并且预测的结果可能刚好相反。

什么是大数据分析

2. 大数据分析的三个层次

规范分析

- 对未来的决策给出建议
- 如基于模拟的复杂系统分析和基于给定约束的优化解生成

- 规范性数据分析可以通过一个或者多个动态指标显示每一个决策结果。
- 规范性数据分析每个环节、每个步骤、每个流程、每个岗位，都有一定的规矩和标准，信息更具准确性，业务决策者可以直接使用。
- 但规范性分析需要一个预测模型、两个附加条件，包括可操作的数据和一个反馈系统，并且这个系统要具备跟踪能力，通过行为预测产生结果。

什么是大数据分析

3. 大数据分析的意义

数据分析是整个大数据处理流程的核心。大数据分析蕴含着巨大的社会价值和产业空间，对社会、经济、科学研究等各个方面都具有重要的战略意义。

(1) 大数据分析可以带动经济发展。

(2) 大数据对社会进步具有重要作用

(3) 加强社会安全和稳定

什么是大数据分析

3. 大数据分析的意义

(1) 大数据分析可以带动经济发展

- 国际著名咨询机构Gartner在报告中说明，2016年全球大数据相关产业规模已达到2320亿美元。
- 麦肯锡全球研究所预测，仅医疗行业，大数据将带来每年3000亿美元的经济价值。
- 大数据的有效利用可以使得欧洲发达国家政府节省至少1000亿欧元的运作成本;
- 使美国医疗保健行业降低8%的成本(约每年3000多亿美元)；并使得大多数零售商的营业利润率提高60%以上。

什么是大数据分析

3. 大数据分析的意义

(1) 大数据分析可以带动经济发展

- 在制造业方面:
  - ✓ 华尔街对冲基金依据购物网站的顾客评论，分析企业的销售状况;
  - ✓ 一些企业利用大数据分析实现对采购和库存进行合理管理；通过分析网上数据了解客户需求，掌握市场动向。

什么是大数据分析

3. 大数据分析的意义

(1) 大数据分析可以带动经济发展

- 在商业领域，沃尔玛将每月4500万的网络购物数据，与社交网络上产品的大众评分结合，开发出“北极星”搜索引擎，方便顾客购物，在线购物的人数增加10%~15%。

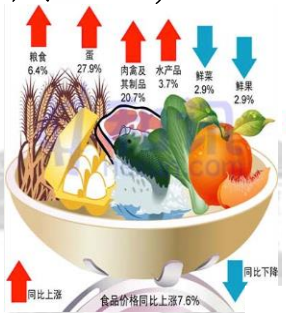
- 再如，有的电商平台将消费者在其平台上的消费记录卖给其他商家，商家得到这个消费记录对应的顾客IP地址后，就会留意其上网踪迹和消费行为，并适时弹出本公司商品的广告，这样就很容易做成交易，最终的结果是顾客，电商平台，商家，甚至相关网站都各有收益。

什么是大数据分析

3. 大数据分析的意义

(2) 对社会进步具有重要作用

- 在宏观经济领域方面，可以用于居民消费价格指数(CPI)的预测。
  - ✓ 国家统计局的预测依据则主要是刚性物品，如食品，百姓都要买，差别不大。
  - ✓ 淘宝利用化妆品、电子产品等网上成交额较高、购买量受经济影响较明显的商品价格来预测CPI;
  - ✓ 淘宝预测的CPI更能反映价格趋势。



什么是大数据分析

3. 大数据分析的意义

(2) 对社会进步具有重要作用

- 在金融领域：
  - ✓ 美国印第安纳大学利用谷歌公司提供的心情分析工具，从近千万的短信和网民留言中归纳出六种心情，进而预测道琼斯工业指数，准确率高达87%。
  - ✓ 华尔街“德温特资本市场”公司CEO利用计算机程序通过分析3.4亿留言判断民众情绪，以决定公司股票的买入和卖出。

什么是大数据分析

3. 大数据分析的意义

(2) 对社会进步具有重要作用

- 在医疗卫生领域，相关部门可以根据搜索引擎上民众对相关关键词的搜索数据建立数学模型进行分析，得出相应的预测进行预防。
  - ✓ 2009年，谷歌公司在甲型H1N1爆发前几周，就预测出流感形式，与随后的官方数据相关性高达97%;
  - ✓ 百度公司得出的中国艾滋病感染人群的分布情况，与后期的卫生部公布结果基本一致。

什么是大数据分析

3. 大数据分析的意义

(2) 对社会进步具有重要作用

- 在农业领域，可以预测产量
  - ✓ 硅谷的Climate公司，利用气候和产量的历史数据，以及气候和土壤的观察数据，建立模型。可以预测下一年的农产品产量、市场价格等信息。

什么是大数据分析

3. 大数据分析的意义

(2) 对社会进步具有重要作用

- 财政金融等大数据，可提高政府的管理决策水平和反腐败成效；
- 利用能源、交通、环境、地理等大数据，可提高城镇管理水平，促进能源节省、交通智能化、环境改善等，实现智慧城市；
- 利用教育大数据，可以提高全社会教育水平和教育效率。



什么是大数据分析

3. 大数据分析的意义

(3) 加强社会安全和稳定

- 通过对网络大数据的分析挖掘，能够及时发现社会动态与情绪，分析舆情，预警敏感、突发和重大事件，提高政府的应对能力，维护社会安全和稳定。
- 比如，通过对微博等网络大数据的挖掘分析能够发现社会动态，预警重大和突发性事件。

目录

大数据分析的过程

业务理解
理解需求，并制定初步计划

数据理解
熟悉数据，识别数据的质量问题

数据准备
将未处理的数据转化为模型工具的输入值

建模
选择和应用不同的模型技术，并对模型参数进行调整

评估
评估模型，检查构造模型的步骤

部署
产生简单的报告，或是实现一个比较可重复的数据挖掘过程

理解项目目标和从业务的角度理解需求，同时转化为数据分析问题的定义和制定初步计划。

大数据分析的过程

业务理解
理解需求，并指定初步计划

数据理解
熟悉数据，识别数据的质量问题

数据准备
将未处理的数据转化为模型工具的输入值

建模
选择和应用不同的模型技术，并对模型参数进行调整

评估
评估模型，检查构造模型的步骤

部署
产生简单的报告，或是实现一个比较可重复的数据挖掘过程

熟悉数据，识别数据的质量问题，发现数据的内部属性等。

大数据分析的过程

业务理解
理解需求，并指定初步计划

数据理解
熟悉数据，识别数据的质量问题

数据准备
将未处理的数据转化为模型工具的输入值

建模
选择和应用不同的模型技术，并对模型参数进行调整

评估
评估模型，检查构造模型的步骤

部署
产生简单的报告，或是实现一个比较可重复的数据挖掘过程

从未处理数据中构造最终数据集。
这些数据将是模型工具的输入值。

大数据分析的过程

业务理解
理解需求，并指定初步计划

数据理解
熟悉数据，识别数据的质量问题

数据准备
将未处理的数据转化为模型工具的输入值

建模
选择和应用不同的模型技术，并对模型参数进行调整

评估
评估模型，检查构造模型的步骤

部署
产生简单的报告，或是实现一个比较可重复的数据挖掘过程

选择和应用不同的模型技术，模型参数被调整到最佳的数值。

大数据分析的过程

业务理解
理解需求，并指定初步计划

数据理解
熟悉数据，识别数据的质量问题

数据准备
将未处理的数据转化为模型工具的输入值

建模
选择和应用不同的模型技术，并对模型参数进行调整

评估
评估模型，检查构造模型的步骤

部署
产生简单的报告，或是实现一个比较可重复的数据挖掘过程

从数据分析的角度建立了一个高质量显示的模型。检查构造模型的步骤，确保模型可以完成业务目标。

大数据分析的过程

业务理解
理解需求，并指定初步计划

数据理解
熟悉数据，识别数据的质量问题

数据准备
将未处理的数据转化为模型工具的输入值

建模
选择和应用不同的模型技术，并对模型参数进行调整

评估
评估模型，检查构造模型的步骤

部署
产生简单的报告，或是实现一个比较可重复的数据挖掘过程

产生简单的报告，或是实现一个比较复杂的、可重复的数据挖掘过程。在很多案例中，这个阶段是由客户而不是数据分析人员承担部署的工作。

目录

大数据分析涉及的技术

(1) 数据采集

- 数据采集

  利用多个数据库来接收客户端的数据，并且用户可以通过这些数据库来进行简单的查询和处理。

- 主要工具

  传统的MySQL与Oracle数据库，以及新兴的NoSQL数据库。

大数据分析涉及的技术

(1) 数据采集

- 挑战
  - ✓ 并发数高。如火车票售票网站12306、网上购物。因为同时有可能会有成千上万的用户来进行访问和操作，比如火车票售票网站和淘宝，它们并发的访问量在峰值时达到上百万，所以需要在采集端部署大量数据库才能支撑。
  - ✓ 如何在多个数据库间进行负载均衡和分片。如何在这些数据库之间进行负载均衡和分片的确是需要深入的思考和设计。

大数据分析涉及的技术

(2) 数据管理

➢ 数据管理

- ✓ 数据管理是大数据分析的基础。
- ✓ 使得大数据“存得下、查得出”，并为大数据的高效分析提供基本数据操作（如连接和聚集）。
- ✓ 实现数据有效管理的关键是数据组织。

➢ 大数据带来的挑战

- ✓ 应用场景的多样化、数据规模的不断增加，使得传统的关系数据库在很多情况下难以满足要求。
- ✓ 因此，学术界和工业界提出了NoSQL和NewSQL数据库。

大数据分析涉及的技术

(2) 数据管理

- NoSQL
  - ✓ NoSQL是指那些非关系型的、分布式的、不保证遵循ACID原则的数据存储系统。分为key-value存储、文档数据库和图数据库这3类。
  - ✓ 适用于处理大量数据的高访问负载以及日志系统的键值数据库(如Tokyo Cabinet/Tyrant, Redis, Voldemort, Oracle BDB)

大数据分析涉及的技术

(2) 数据管理

- NoSQL
  - ✓ 适用于分布式大数据管理的列存储数据(如Cassandra, HBase, Riak)
  - ✓ 适用于Web应用的文档型数据库(如 CouchDB, MongoDb, SequoiaDB)
  - ✓ 适用于社交网络、知识管理等的图数据库(如Neo4J, InfoGrid, Infinite Graph)，这些数据库统称为NoSQL。

大数据分析涉及的技术

(2) 数据管理

- NewSQL
  - ✓ NewSQL是对各种新的可扩展/高性能数据库的简称。
  - ✓ 这类数据库不仅具有NoSQL对海量数据的存储管理能力，还保持了传统数据库支持ACID和SQL等特性。
  - ✓ 典型的NewSQL包括VoltDB，ScaleBase，dbShards, Scalearc等。例如，阿里云分析型数据库可实现对数据的实时多维分析，百亿量级多维查询只需100毫秒。

大数据分析涉及的技术

(3) 基础架构

- 基础架构
  - ✓从底层来看，对大数据进行分析需要高性能的计算架构和存储系统。
- 举例
  - ✓分布式计算的MapReduce框架、Spark计算框架;
  - ✓用于大规模数据协同工作的分布式文件存储系统HDFS。

大数据分析涉及的技术

(4) 数据的理解与提取

大数据的多样性体现在多个方面

- 结构方面：大数据分析需要处理的数据，在很多情况下，并非传统的结构化数据。也包括多模态的半结构和非结构化数据；

- 语义方面：大数据的语义也有着多样性，同一含义有着多样的表达，同样的表达在不同的语境下也有着不同的含义。

大数据分析涉及的技术

(4) 数据的理解与提取

- 数据理解和提取要对具有多样性的大数据进行有效分析，需要对数据进行深入的理解，并从结构多样、语义多样的数据中提取出可以直接进行分析的数据。这方面的技术包括自然语言处理、数据抽取等。
  - ✓自然语言处理：研究人与计算机交互的语言问题的一门学科。它是人工智能(AI, Artificial Intelligence)的核心课题之一。
  - ✓数据抽取：把非结构化数据中包含的信息进行结构化处理，变成统一的形式。

大数据分析涉及的技术

(5) 统计分析

- 统计分析是指运用统计方法以及与分析对象有关的知识，从定量与定性的结合上进行的研究活动。

- 统计分析是在统计设计、统计调查、统计整理的基础上，通过分析从而达到对研究对象更为深刻的认识。

大数据分析涉及的技术

(5) 统计分析

- 统计分析又是在一定的选题下，集分析方案的设计、资料的搜集和整理而展开的研究活动。系统、完善的资料是统计分析的必要条件。

- 主要包括
  - ✓ 假设检验、显著性检验、差异分析、相关分析、回归分析、主成分分析、聚类分析、判别分析等。

大数据分析涉及的技术

(6) 数据挖掘

- 数据挖掘是指从大量的数据中通过算法搜索隐藏于其中信息的过程。

- 包括分类、估计、预测、关联规则、聚类、描述和可视化、复杂数据类型挖掘等。

大数据分析涉及的技术

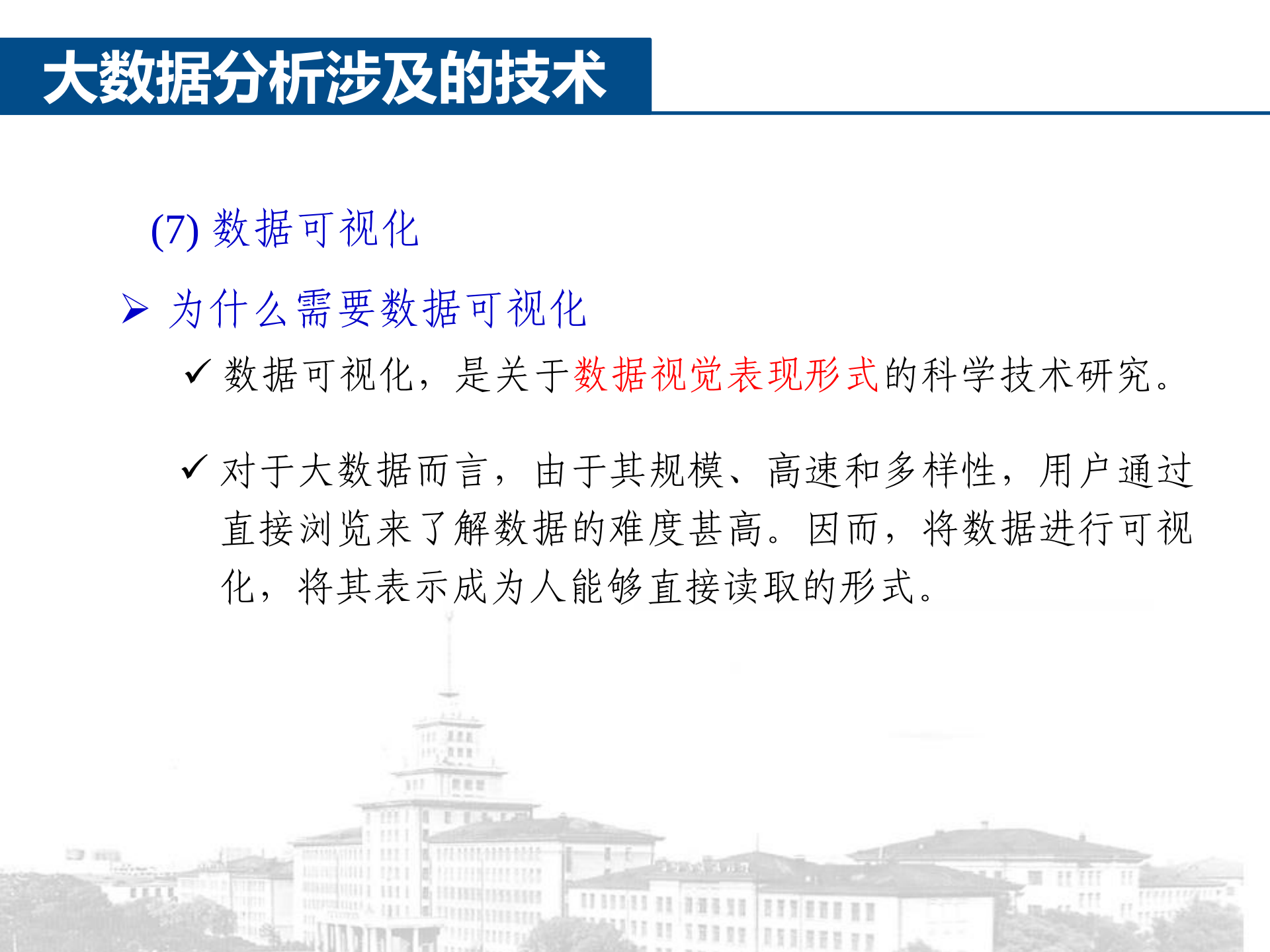
(6) 数据挖掘

- 与统计分析的区别
  - ✓ 数据挖掘一般没有什么预先设定的主题，主要是在现有数据上进行基于各种算法的计算，从而起到预测的效果，实现一些高级别数据分析的需求。
  - ✓ 例如阿里云品拥有一系列机器学习工具，可基于海量数据实现对用户行为、行业走势、天气、交通的预测，产品还集成了阿里巴巴核心算法库，包括特征工程、大规模机器学习、深度学习等。

大数据分析涉及的技术

(7) 数据可视化

- 为什么需要数据可视化
  - ✓ 数据可视化，是关于数据视觉表现形式的科学技术研究。
  - ✓ 对于大数据而言，由于其规模、高速和多样性，用户通过直接浏览来了解数据的难度甚高。因而，将数据进行可视化，将其表示成为人能够直接读取的形式。



大数据分析涉及的技术

(7) 数据可视化

- 数据可视化的种类
  - ✓ 按照原理分类:
    - ✓基于几何的技术
    - ✓面向像素的技术
    - ✓基于图标的技术
    - ✓基于层次的技术
    - ✓基于图像的技术和分布式技术等;
  - ✓ 按照数据类型分类:
    - ✓文本可视化
    - ✓网络（图）可视化
    - ✓时空数据可视化
    - ✓多维数据可视化等。

目录

大数据分析的难点

大数据分析不是简单的数据分析的延伸，由于大数据规模大、更新速度快、来源多样等性质，为大数据分析带来了一系列挑战。

(1) 可扩展性

大数据分析的首要任务是使得分析算法能够支持大规模数据；并在所要求的时间内得到结果。

大数据分析的难点

(2) 可用性

- 将大数据分析应用到实际中的前提是分析结果的可用性。

- 这里的“可用”包含两个方面：
  - ✓需要结果具有高质量，如数据完整、符合现实的语义约束等；
  - ✓需要结果的形式适用于实际的应用。

大数据分析的难点

(3) 要与具体领域知识相结合

- 一方面，领域知识具有多样性导致大数据分析方法的多样性，算法需要进行相应的调整;
- 另一方面，往往需要将领域知识的内容，进行合理的表示，用于大数据分析。

大数据分析的难点

(4) 结果的检验

- 由于有一些应用需要高可靠性的分析结果，否则会带来灾难性的后果。因而，大数据分析结果需要经过一定检验才可以真正应用。

本节课程小结

1. 什么是大数据
   1. 定义
   2. 背景
   3. 特点 ※
   4. 技术支撑
2. 什么是大数据分析
   1. 定义
   2. 三个层次
   3. 意义
3. 大数据分析的过程
4. 大数据分析涉及的技术 ※
5. 大数据分析的难点技术 ※

习题和讨论

一、在我们身边有哪些大数据？在这些大数据上有哪些分析任务？

- 体育赛事预测
  - ✓ 世界杯期间，谷歌、百度、微软和高盛等公司都推出了比赛结果预测平台。百度预测结果最为亮眼，预测全程64场比赛，准确率为67%，进入淘汰赛后准确率为94%。
  - ✓ 现在互联网公司取代章鱼保罗进行赛事预测，也意味着未来的体育赛事会被大数据预测所掌控。

习题和讨论

- 用户行为预测
  - ✓基于用户搜索行为、浏览行为、评论历史和个人资料等数据，互联网业务可以洞察消费者的整体需求，进而进行针对性的产品生产、改进和营销。
  - ✓百度基于用户喜好进行精准广告营销、阿里根据天猫用户特征包下生产线定制产品、亚马逊预测用户点击行为提前发货均是受益于互联网用户行为预测。

习题和讨论

- 灾害灾难预测
  - ✓气象预测是最典型的灾难灾害预测。地震、洪涝、高温、暴雨这些自然灾害如果可以利用大数据能力进行更加提前的预测和告知，便有助于减灾防灾救灾赈。
  - ✓与过往不同的是，过去的数据收集方式存在着死角、成本高等问题，物联网时代可以借助廉价的传感器摄像头和无线通信网络，进行实时的数据监控收集，再利用大数据预测分析，做到更精准的自然灾害预测。

习题和讨论

二*、比较数据分析和数据挖掘的区别。

1. 定义

- 数据挖掘用于识别和发现大型数据集中隐藏的模式，也被称为数据库中的知识发现；
- 数据分析是指根据分析目的，用适当的统计分析方法及工具，对收集来的数据进行处理与分析，提取有价值的信息，形成结论。

习题和讨论

2. 假设

- 数据挖掘不需要任何先入为主的假设即可识别数据的模式或趋势，也就是说，数据挖掘是在没有任何先入为主的假设的情况下进行的，因此来自数据的信息不能回答特定的问题；
- 数据分析是从一个假设出发，需要自行建立方程或模型来与假设吻合。也就是说，数据分析需要检验给定的假设，从数据中发现事实以回答特定的问题。

习题和讨论

3. 数据类型

- 数据挖掘研究主要针对结构化数据;
- 数据分析可以对结构化，半结构化或非结构化数据进行处理。

习题和讨论

4. 可视化

- 数据挖掘通常不涉及可视化工具;
- 数据分析总是伴随着结果可视化。

5. 输出

- 数据挖掘通常输出数据模型或规则，并且可相应得到模型得分或标签;
- 数据分析输出是对数据的经过验证的假设。

习题和讨论

- 数据分析的目标往往比较明确，分析条件也比较清楚，基本上就是采用统计方法，对数据进行多维度地描述；
- 数据挖掘的目标却不是很清晰，要依靠挖掘算法来找出隐藏在大量数据中的规律和模式，也就是从数据中提取出隐含的、未知的有价值的信息。

习题和讨论

三、比较电子商务和工业大数据分析任务的异同。

电子商务

- 大数据分析的任务是根据数以万计的交易产生大量交易数据和个人特征信息，准时地进行各类店铺排名和个性化智能推荐，进行用户行为数据分析，得到电商用户所需的个性化信息与产品，便于开展精确营销、客户管理；
- 商家根据购物历史信息进行生产、研究进货存货计划；买家也可获得更符合个性化需求的商品信息，从而提高客户满意度。

习题和讨论

三、比较电子商务和工业大数据分析任务的异同。

工业大数据

- 以工业数据为核心，围绕典型智能制造模式，从客户需求到产品设计、研发、工艺、制造、供应、销售、库存等整个产品生命周期的各个过程产生的数据以及相关技术和应用的总称。

习题和讨论

三、比较电子商务和工业大数据分析任务的异同。

工业大数据

- 相比于电子商务大数据更注重数据的量，工业大数据更加注重数据的连续性，一是同一工业流程时间上的连续，二是各个工业流程的连续，即从产品设计到产品销售的整个工业流程的连续。

习题和讨论

三、比较电子商务和工业大数据分析任务的异同。

工业大数据

- 工业大数据分析任务借助传感器、通信感知、过程工业实时数据库等技术，对数据实时性要求更高，更加注重数据处理后的数据质量，要求数据有真实性、完整性、可靠性，且由于数据之间关联性强，存储复杂，对数据的分析，建模更为复杂，不同工业领域涉及到的分析方法差别很大，精度和可靠度要求相对高。

习题和讨论

四*、大数据分析对技术提出了何种挑战？根据你的经验论述这些挑战应当如何应对。

- 数据采集：数据采集端需要部署大量的数据库，数据采集中如何在这些数据库之间进行负载均衡和分片需要深入思考和设计。
- 数据管理：传统关系数据库难以满足要求，需要可扩展/高性能数据库。
- 基础架构：需要高性能的计算架构和存储系统。

习题和讨论

四、大数据分析对技术提出了何种挑战？根据你的经验论述这些挑战应当如何应对。

- 数据理解与提取：需要用到自然语言处理，数据抽取等技术
- 统计分析：采取正确的统计分析技术
- 数据挖掘：从大量数据中通过算法搜索隐藏于其中的信息
- 数据可视化：将数据进行可视化表示成为人能够直接读取的形式。

习题和讨论

五*、试论述可视化在大数据分析过程中可能起到的作用。

- 数据可视化是指将大数据分析与预测结果以计算机图形或图像的直观方式显示给用户的过程，并可与用户进行交互式处理。
- 数据可视化技术有利于发现大量业务数据中隐含的规律性信息，以支持管理决策。
- 数据可视化环节可大大提高大数据分析结果的直观性，便于用户理解与使用，故数据可视化是影响大数据可用性和易于理解性质量的关键因素。