大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

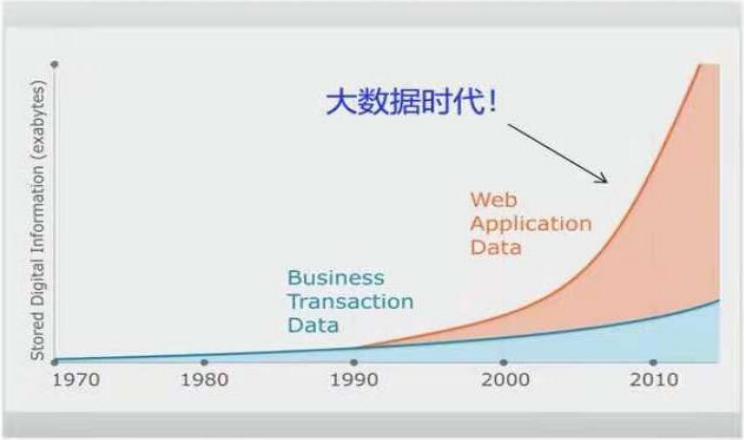
北京理工大学计算机学院　孙新

2019年1月

目　录

大数据时代，数据扮演着重要的角色

数据量增加



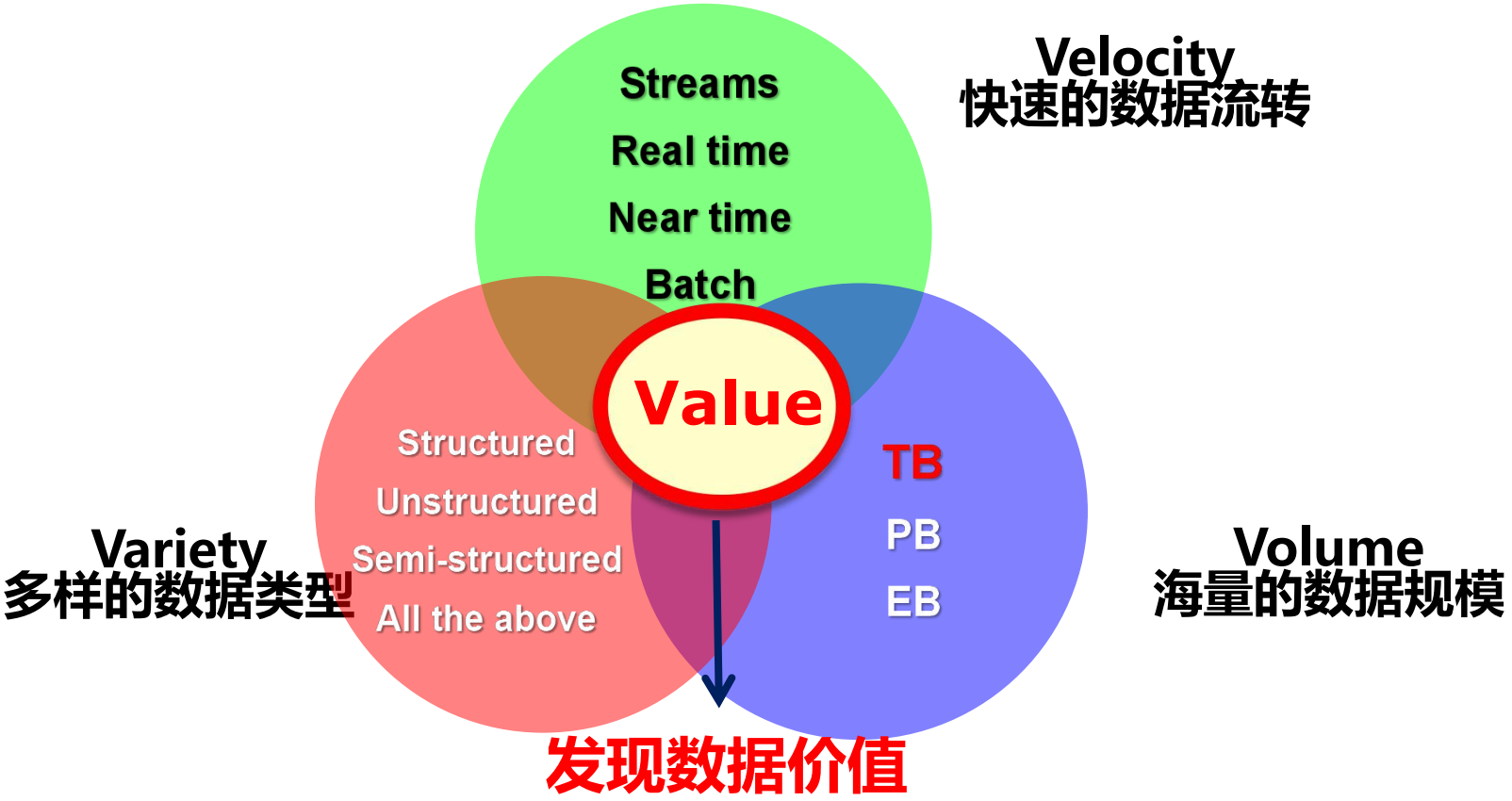
1 EB = 10^3 ZB = 10^6 PB = 10^9 TB = 10^{12} GB

数据结构日趋复杂

- 全球每秒钟发送 2.9 百万封电子邮件
- 每天会有 2.88 万个小时的视频上传到Youtube,
- 每天亚马逊上将产生 6.3 百万笔订单
- 每个月网民在Facebook 上要花费7 千亿分钟
- 被移动互联网使用者发送接收的数据达1.3EB
- Google 上每天需要处理24PB 的数**据...**

- **大数据技术飞速发展和人工智能的运用是这个时代的趋势**
- **如何利用大数据，使用数据驱动的方法解决智能问题和决策问题已经成为信息领域的共识**

数据分析需要解决的问题



数据分析需要解决的问题

然而，数据通常并不能直接被人们利用。如何从大量看似杂乱无章的数据中，发掘有用的知识、揭示其中隐含的内在规律，指导人们进行科学的推断与决策?

数据—信息—知识—价值

数据分析赋予我们洞察未来的能力

马云成功预测2008 年经济危机

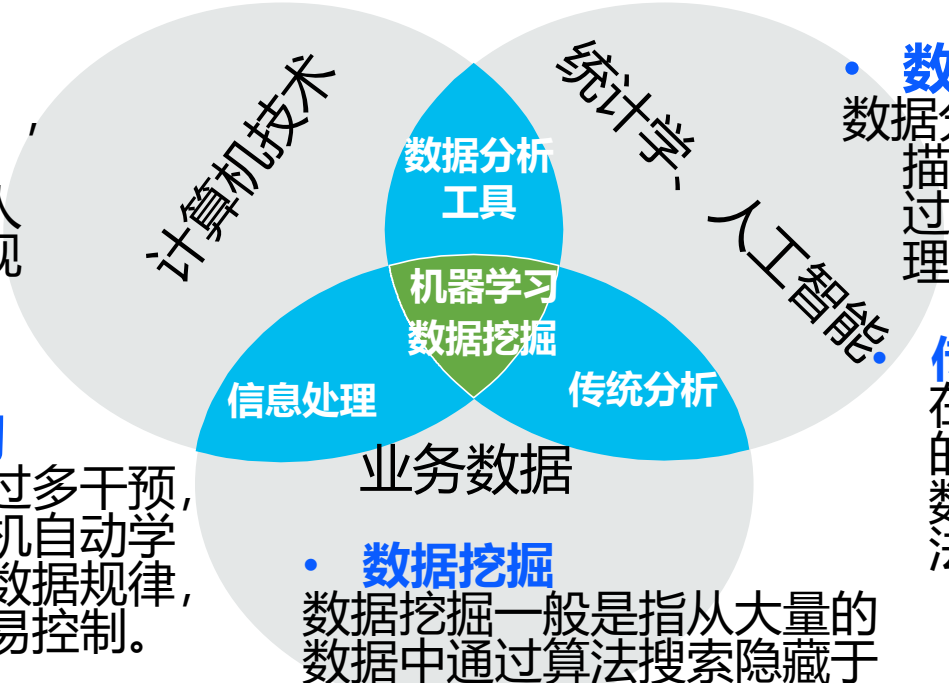
"2008 年初,阿里巴巴平台上整个买家询盘数急剧下滑，欧美对中国采购在下滑。海关是卖了货，出去以后再获得数据；我们提前半年从询盘上推断出世界贸易发生变化了"

- 通常而言，买家在采购商品前，会比较多家供应商的产品，反映到阿里巴巴网站统计数据中，就是查询点击的数量和购买点击的数量会保持一个相对的数值，综合各个维度的数据可建立用户行为模型。因为**数据样本巨大,保证用户行为模型的准确性**。因此在这个案例中，询盘数据的下降，自然导致买盘的下降。

人类从依靠自身判断做决定到依靠数据做决定的转变，也是大数据作出的最大贡献之一。——《大数据时代》

随着计算机技术发展和数据分析理论的更新，当前的数据分析逐步成为机器语言、统计知识两个学科的交集

- **数据分析工具**
  各种厂商开发了数据分析的工具、，将分析模型封装，使不了解技术的人也能够快捷的实现数学建模。

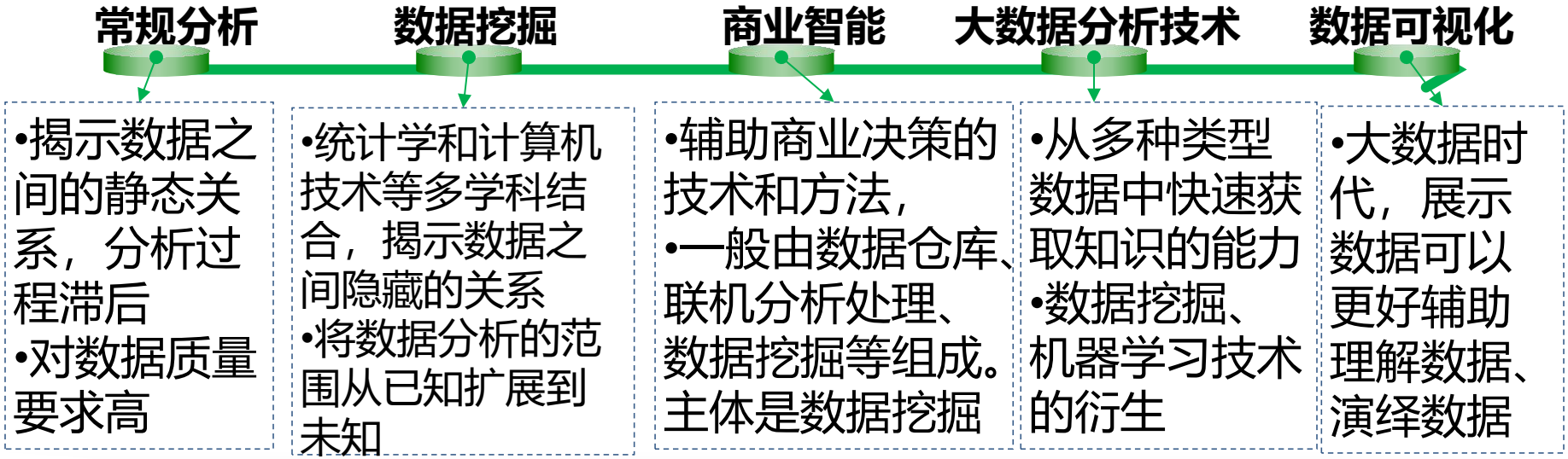


- **数学&统计学知识**
  数据分析的基础，将整理、描述、预测数据的手段、过程抽象为数学模型的理论知识

- **传统分析**
  在数据量较少时，传统的数据分析已能够发现数据中包含的知识，方法成熟，应用广泛，

- **机器学习**
  不需要人过多干预，通过计算机自动学习，发现数据规律，但结论不易控制。

- **数据挖掘**
  数据挖掘一般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。



数据分析

- 本节重点介绍通用的数据分析方法。
- 随着数据量的不断扩大，数据分析理论正处于飞速发展期，因此本文的方法侧重于基础原理介绍。

数据分析框架

业务理解

理解背景，评估需求

- **理解业务背景：**

数据分析的本质是服务于业务需求，如果没有业务理解，缺乏业务指导，会导致分析无法落地。

- **评估业务需求：**

判断分析需求是否可以转换为数据分析项目，

数据理解

数据收集 数据清洗

- **数据收集：**

抽取的数据必须能够正确反映业务需求。否则分析结论会对业务将造成误导

- **数据清洗：**

原始数据中存在数据缺失和坏数据，如果不处理会导致模型失效，因此需要对数据过滤“去噪”

数据准备

数据探索数据转换

- **探索数据：**

运用统计方法对数据进行探索，发现数据内部规律。

- **数据转换：**

为了达到模型的输入数据要求，需要对数据进行转换。

数据分析框架

建立模型 → 模型评估 → 应用

选择方法建立模型

- **建立模型：**

综合考虑业务需求精度、数据情况、花费成本等因素，选择最合适的模型。

在实践中对于一个分析目的，往往运用多个模型，然后通过模型评估，进行优化、调整，以寻求最合适的模型。

过程评估 结果评估

- **建模过程评估：**

对模型的精度、准确性、效率和通用性进行评估。，

- **模型结果评估：**

评估是否有遗漏的业务，模型结果是否回答了当初的业务问题，需要结合业务专家进行评估。

结果应用 模型改进

- **结果应用：**

将模型应用于业务实践，才能实现数据分析的真正价值：产生商业价值和解决业务问题。

- **模型改进：**

对模型应用效果的及时跟踪和反馈，以便后期的模型调整和优化。

谢 谢

Thank you for your attention!