第四节　数据科学与大数据

一、数据科学

定义	一门通过系统性研究获取与数据相关的知识体系的学科，最早由丹麦的计算机科学领域先驱彼得•诺尔提出。
研究对象	数据，即从“数据”整合成“信息”进而组织成“知识”的整个过程，包含对数据进行采集、储存、处理、分析、表现等活动。

第四节　数据科学与大数据

二、大数据

定义	无法在一定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，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
特性	数据量大、数据多样性、价值密度低、数据的产生和处理速度快

第四节　数据科学与大数据

三、数据挖掘

含义	1. 数据源必须是真实的、大量的、含噪声的 2. 发现的是用户感兴趣的知识 3. 发现的知识是可接受、可理解、可运用的 4. 不要求发现放之四海而皆准的知识，仅支持特定的发现问题
核心	以解决实际问题为出发点；核心任务是对数据关系和特征进行探索。

第四节　数据科学与大数据

三、数据挖掘

常见方法	监督学习	每个观测单位既有自变量（特征）又有因变量（标签）
	无监督学习	每个观测单位只有自变量（特征），没有因变量（标签）
	半监督学习	是监督学习与无监督学习相结合的一种学习方法。数据集中，一部分观测单位既有自变量又有因变量，另一部分观测单位只有自变量，没有因变量，而且没有标签的观测单位数量远大于有标签的观测单位数量。

典型真题

【真题·2022多选】关于数据科学说法正确的（　）。

A. 只研究数据本身的特点和变化规律

B. 从数据整合成信息进而组织成知识

C. 可视化

D. 人工智能

E. 数据挖掘只包括监督学习和无监督学习

典型真题

答案：BCD

解析：数据科学是一门通过系统性研究获取与数据相关的知识体系的学科，研究对象是数据。

本章内容总结

第23章内容总结

- **1.两大分支**
  - 1.描述统计：展示
  - 2.推断统计：参数估计；假设检验
- **2.变量（数据）**
  - 数量变量（数值型数据）、分类变量（分类数据）、顺序变量（顺序数据）
- **3.数据的来源**
  - 1.按收集方式：观测和实验数据(2新)
  - 2.按来源：一手（观测、实验）和二手数据

本章内容总结

- 第23章内容总结
  - 4.统计调查
    - 分类
      - 按调查对象的范围不同
        - 全面调查（全面统计报表和普查）、非全面调查（非全面统计报表、抽样调查、重点调查和典型调查等）
      - 按调查登记的时间是否连续
        - 连续调查（总量）；不连续调查（时点）
    - 普查
      - 一次性或者周期性；规定统一的标准调查时间；数据准确；使用范围窄
      - ①经济普查逢 3、8 年份（第二 、 三产业）②人口普查逢0年份③农业普查逢6年份
    - 抽样调查
      - 经济性(最显著的优点)；时效性强；适应面广；准确性高
    - 重点调查
      - 标志值来说在总体中占绝大比重
    - 典型调查
      - 代表性或典型意义，对普查进行补充性调查
    - 质量标准
      - 真实性、准确性、完整性、及时性、适用性、经济性、可比性、协调性、可获得性
  - 5.数据科学与大数据
    - 最早由丹麦的计算机科学领域先驱彼得•诺尔提出
    - 数据挖掘：监督学习、无监督学习、半监督学习
    - 大数据“4V”特性
      - 数据量大、数据多样性 、价值密度、数据的产生和处理速度快

谢谢 观看

THANK YOU