

2024-2025秋季课程

# 数据科学与大数据导论

Introduction to Data Science and Big data

曹劲舟 博士 助理教授

深圳技术大学 大数据与互联网学院

caojinzhou@sztu.edu.cn

2024年8月

# 自我介绍

大数据与互联网学院  
曹劲舟 版权所有





# 城市与空间人工智能实验室

曹劲舟 博士 助理教授

caojinzhou@sztu.edu.cn

深圳技术大学 大数据与互联网学院

C1-305/604

# 导师背景

## 联合培养

### 本科

2009.09-2013.06  
武汉大学遥感信息工程学院

### 硕博连读

2013.09-2019.06  
武汉大学测绘遥感信息工程国家重点实验室

2017.09-2018.09  
美国华盛顿大学  
土木与环境工程学院

### 博士后

2019.09-2021.08  
深圳大学建筑与城市规划学院&广东省城市空间信息工程重点实验室  
&深圳市空间信息智能感知与服务重点实验室

### 副研究员

2021.09-2022.03  
深圳大学

### 助理教授

2022.04-至今  
深圳技术大学大数据与互联网学院

## 导师背景

---

曹劲舟博士，助理教授，深圳市C类高层次人才，美国华盛顿大学访问学者，硕士生导师。2019年获武汉大学测绘遥感信息工程国家重点实验室博士学位，导师为中国工程院院士、深圳大学党委书记李清泉院士。曾在深圳大学广东省城市信息重点实验室从事博士后、副研究员工作。研究方向为城市大数据挖掘，Geo-AI和地理/社会计算。主持国家自然科学基金青年项目，中国博士后科学基金，深圳市基础研究面上项目，自然资源部重点实验室开放基金等纵向科研项目7余项，参与国家自然科学基金中欧国际合作项目、面上项目、国家重点研发计划子课题、广东省自然科学基金、深圳市基础研究重点项目等多项。发表论文30余篇，其中1篇论文入选ESI高被引论文；Google Scholar被引750+次。授权发明专利10项，获准软件著作权3项。获得测绘科学技术奖一等奖（2021）等奖励。

# 研究方向



智慧城市应用



大数据分析

数据

计算

人工智能方法



新浪微博 weibo.com

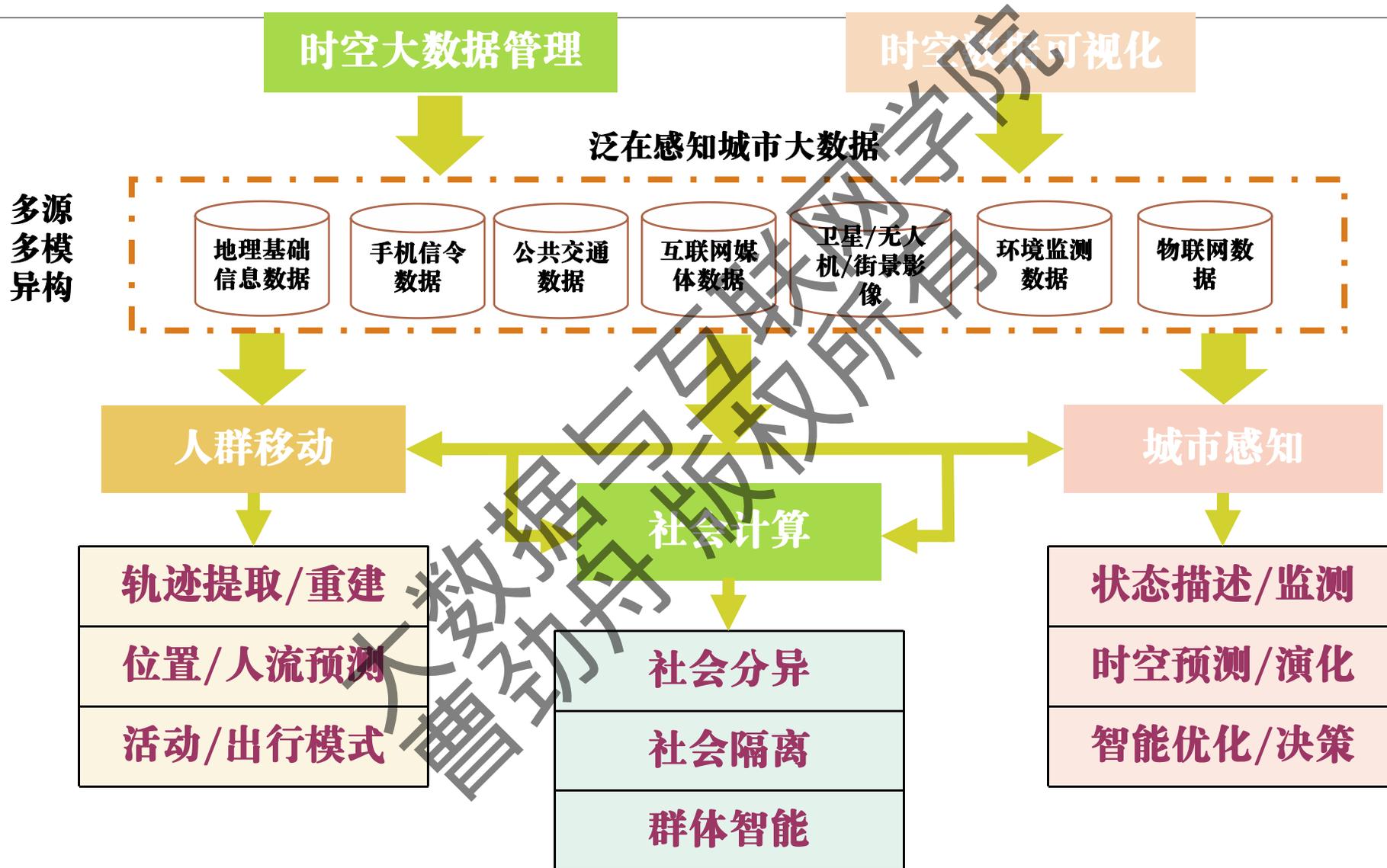


数据

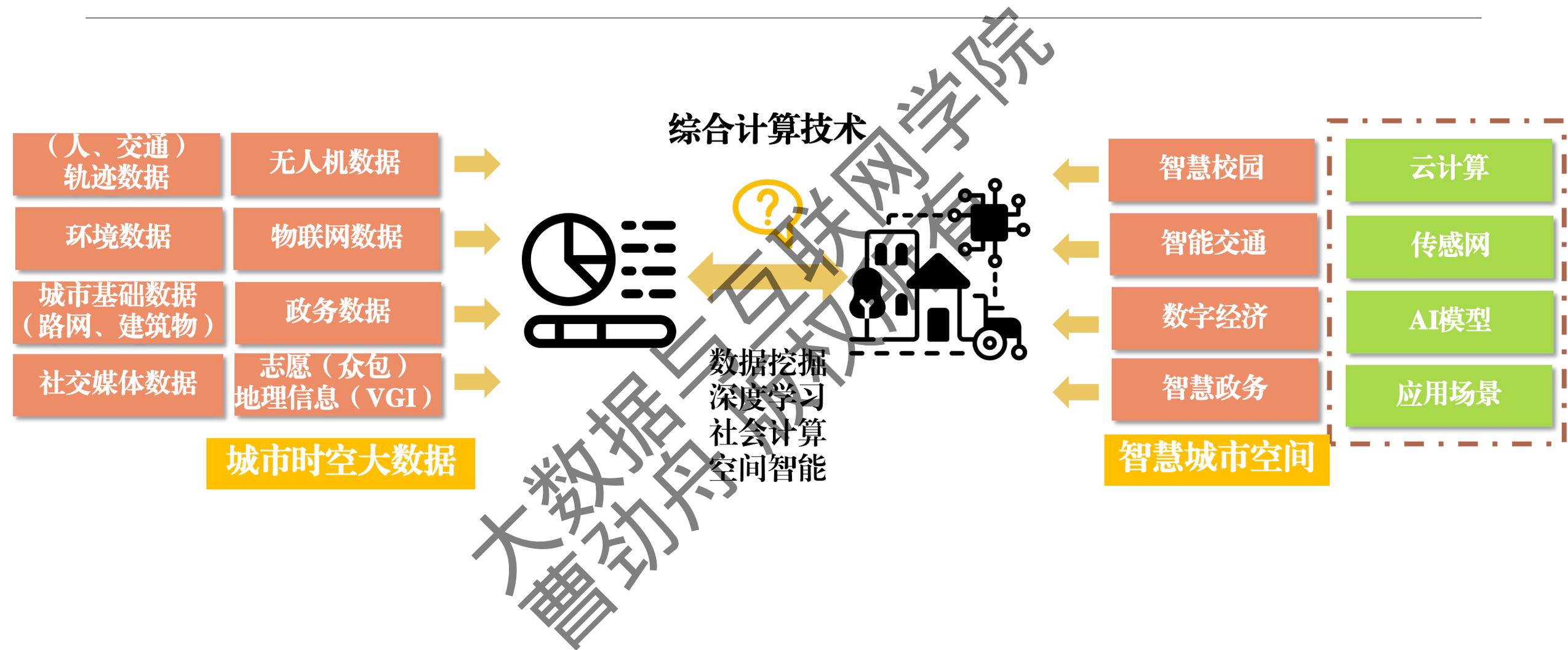
计算

如何结合新型数据源（城市时空大数据）和新模型方法（深度学习），更好地解决智慧城市建设中面临的挑战性问题？

# 研究框架



# 研究应用



# 课程内容

大数据与互联网学院  
大数据舟 版权所有



# | 关于课程

- 课程定位：数据科学与大数据系列课程的基础课和先导课
- 后续课程：大数据原理与技术、人工智能导论、大数据融合技术、大数据编程与可视化技术、高级统计学、深度学习方法与应用等
- 课程目标：
  - 让同学们对数据科学与大数据有一个整体的认识
  - 针对不同类型的数据进行深入讲解
  - 了解数据处理与分析的基本工具与常用技术、发展前沿和应用案例
  - 树立数据科学的基本思路，了解数据的“能”与“不能”
  - 利用实验课，初步掌握使用数据分析手段解决实际应用问题的能力，独立或小组的形式完成实验内容和大项目

# | 关于课程

- 这不是Python课
  - 对于Python基础知识靠同学们在实验过程中进行掌握
- 这不是数学课
  - 算法推导不是课程内容，感兴趣同学可以自学
- 这不是算法课
  - 更多的是让大家掌握数据科学的流程和应用方向

# | 关于课程 About this course



- 用科学的方法研究和应用数据  
希望这门课程带给你们的是终身受用的数据思维  
和创新能力。

数据科学与大数据导论课能让你成为数据科学家吗？

- 不能……
- 但我们希望这是一个好的开端！



# 课程介绍

## □课程覆盖的内容

- 处理和分析各种类型的数据
- 文本、图、空间、时间、关系、Web、时间序列、流数据……
- 解决数据科学的两个核心任务
  - 从数据中洞见真知：raw data → Insights
  - 数据驱动的决策支持：城市大数据分析、文本挖掘、图数据分析……
- 掌握数据分析的技能与工具
  - Python及其数据分析工具
  - 机器学习初步
  - 数据统计基础、深度学习、数据库系统、最优化……
- 了解大数据处理的工具
  - 初步介绍一些分布式数据处理工具、数据存储平台、数据可视化工具等

# 课程章节及学时分配（初步，可能会调整）

- 课程共计18周（1-18周，18次课）
  - 概论，3次课
    - 1) 大数据概述 Introduction to big data
    - 2) 数据科学基础 Data Science Fundamentals
    - 3) 大数据处理基础 Big Data Analytics Fundamentals
  - 大数据分析算法，4次课 Big Data Analytics Algorithms: 机器学习相关
    - 1) 聚类、分类 Clustering and Classification
    - 2) 回归、关联分析 Regression and Association Analysis
  - 大数据处理工具，3次课
    - 1) 大数据可视化 Big Data Visualization
    - 2) 大数据处理平台与数据存储 Big Data Platforms and Tools and Data Storage

# 课程章节及学时分配（初步，可能会调整）

- 课程共计18周（1-18周，18次课）
- 数据科学前沿专题，7-8次课
  - 城市大数据科学 Urban data science
    - 图数据计算 Graph data computing
      - 图的基本概念、图的构建与可视化、图的中心度分析、图的社区检测、影响力分析
    - 文本挖掘 Text mining
      - 文本的预处理(如中文分词)、文本的分类、文本的检索…
- 课程回顾与复习，1次课
- 法定节假日等会冲掉1-2次课，进度根据实际情况调整

# 课程介绍

课程不会深入的内容

- 数据库系统与技术（大二下）
  - 数据科学家需要非常熟练的掌握数据库技术
  - 留给后续数据库相关课程
- Python程序设计与数据分析编程实践【自学】（大二下）
  - 对成为一个数据科学家来讲非常重要
  - 认为能够通过自学+实验课掌握基本的技能
- （复杂的）机器学习与深度学习（大三上）
  - 讲解机器学习的基本思想与最简单模型，把更复杂的知识留给后续的课程

# 课程组织

大数据与互联网学院  
曹劲舟 版权所有



# 课程网页

- [www.caojz.cn/courses/idsbd2024/](http://www.caojz.cn/courses/idsbd2024/)
- 授课PPT将会每周课程结束后上传到课程网页
- 作业/项目安排/自学教程/阅读材料等资料将会不定时上传到课程网页
- 请同学们收藏网页，不定时check!!!

# 理论+实验

## ● 实验将于第六周开始

- 23 级大数据 2 班：周一 8-10 节
- 23 级大数据 3 班：周一 11-13 节

	周一	周二	周三	周四	周五
1					
2					
3		理论课 大数据 3 班 C-5-310=4			
4					
5					
6	理论课 大数据 2 班 C-5-310				
7					
8	实验课 大数据 2 班 第 6 周开始 C-5-360 机房				
9					
10					
11	实验课 大数据 3 班 第 6 周开始 C-5-360 机房				
12					
13					

# 关于实验

## ● 9个实验

- 第1个：Python基础
- 2-9个：根据课程进度，逐步开展
- 实验课上还有代码随堂考试和课后作业

实验项目编号	实验项目名称		实验类型	实验性质	实验学时	每组人数	首次开出年月	备注
1	Python基础	1.1 Python 开发环境搭建	验证性	必做	6 学时	1	202309	实验室机房授课
		1.2 Python 基础知识						
		1.3 Python 数据分析库 (Numpy, Pandas, Matplotlib)						
2	数据预处理与探索性分析实验		验证性	必做	2 学时	1	202309	
3	数据可视化实验		验证性	必做	4 学时	1	202309	
4	聚类算法实验		验证性	必做	4 学时	1	202309	
5	分类算法实验		验证性	必做	4 学时	1	202309	
6	回归算法实验		验证性	必做	4 学时	1	202309	
7	城市大数据分析与实践		验证性	必做	4 学时	1	202309	
8	图数据计算实验		验证性	必做	4 学时	1	202309	
9	文本挖掘实验		验证性	必做	4 学时	1	202309	

# 关于实验

- 所有代码均不得使用 ChatGPT 完成!
- 如果有使用，必须在作业最后进行声明：“本实验使用 ChatGPT 进行了代码调试/代码优化/...”

大数据与互联网学院



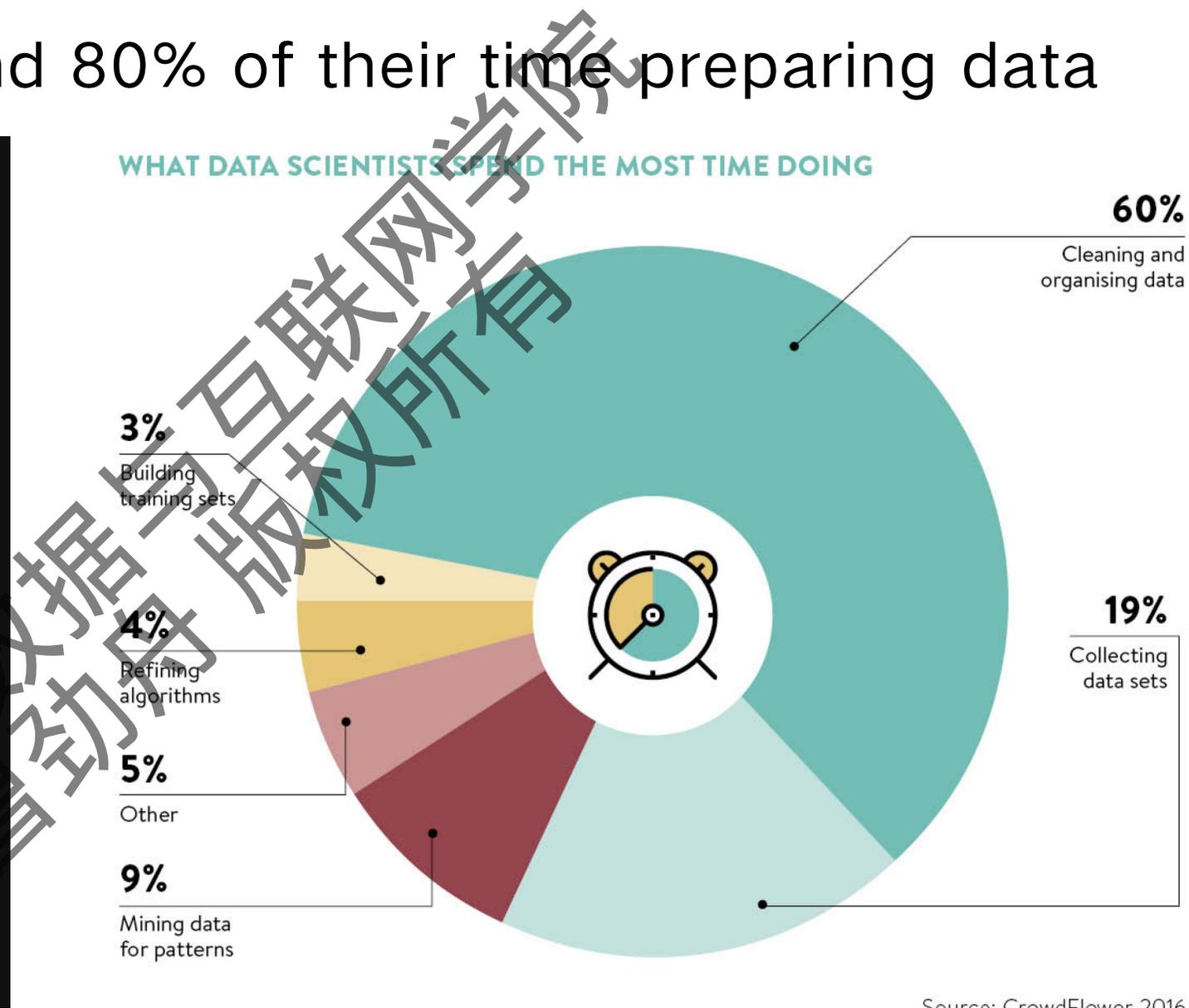
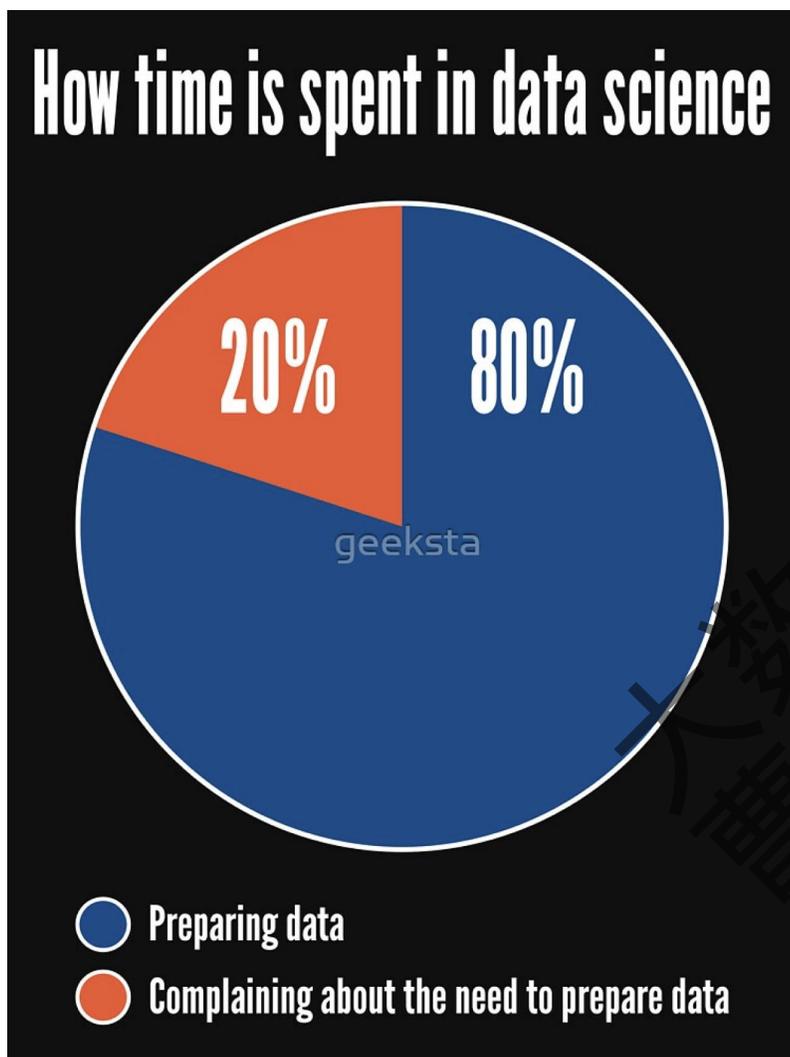
# 使用 Chatgpt 将摧毁你的代码能力

- 在这门课程中，解决练习、作业或编程测试时，请不要使用像 ChatGPT 这样的大型语言模型（LLMs），因为：
  1. 这门课程的目标是手动学习编程，这在你未来在 ITU 的学习和职业生涯中是绝对必需的，
  2. 课程的所有部分都设计为手动解决，
  3. 期末考试将是笔试，因此 ChatGPT 在那里也无济于事，
  4. 大型语言模型（LLMs）是不可靠的。



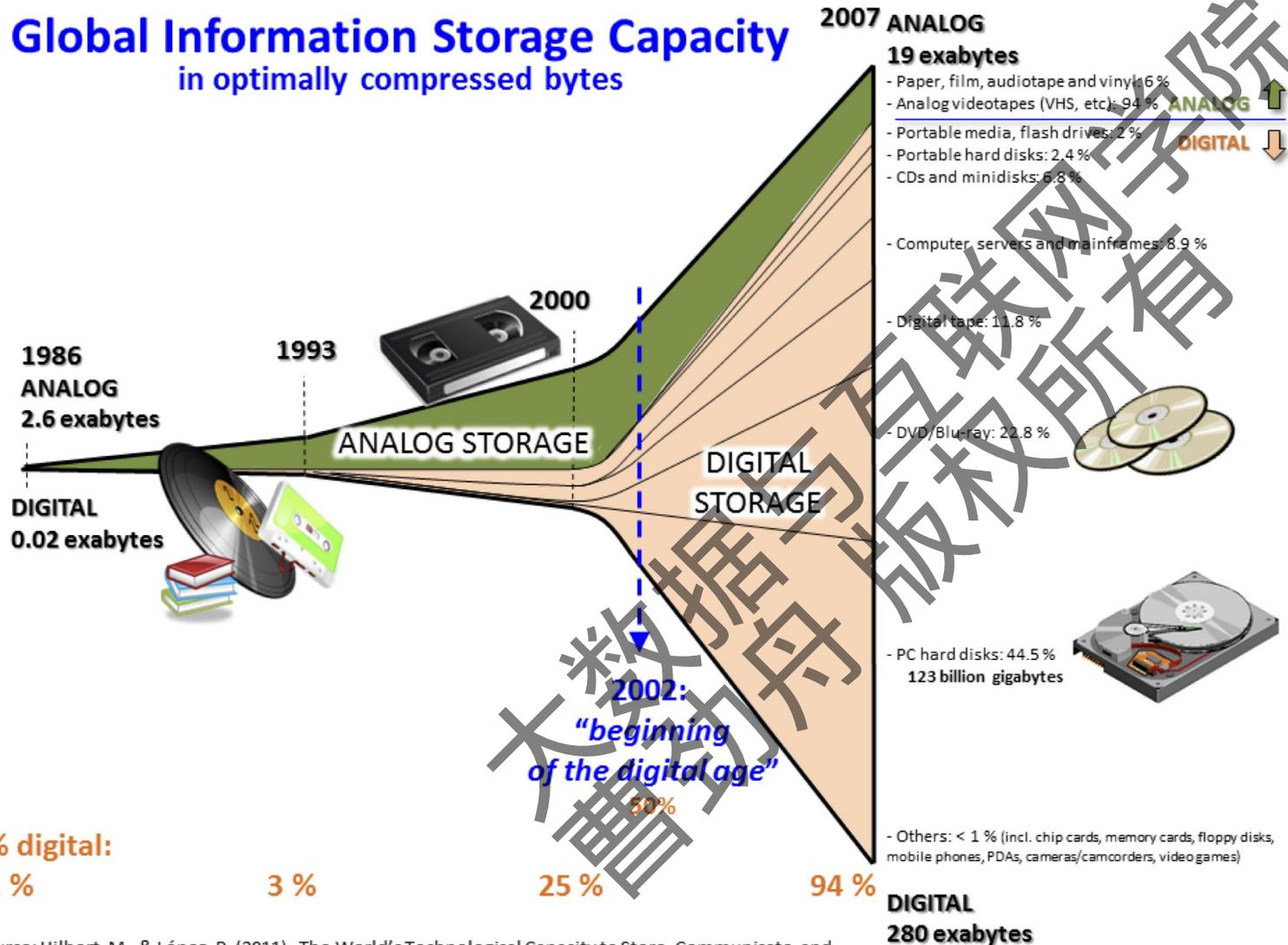
# 为什么一定要写代码

- Data Scientists spend 80% of their time preparing data



# 数据科学的历史

## Global Information Storage Capacity in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

# 数据科学家需要的四个技能：

## ● 编程与计算机科学

- 以高效处理大数据集

## ● 数学与统计学

- 以正确提出问题并分析数据

## ● 沟通与可视化

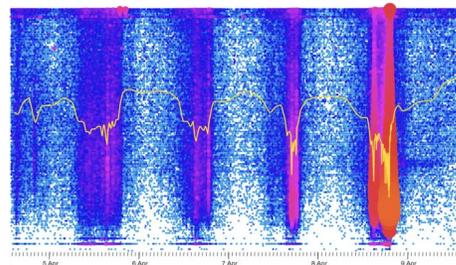
- 以交流洞见，特别是与对数据不太熟悉的人

## ● 领域知识

- 以提出正确的问题

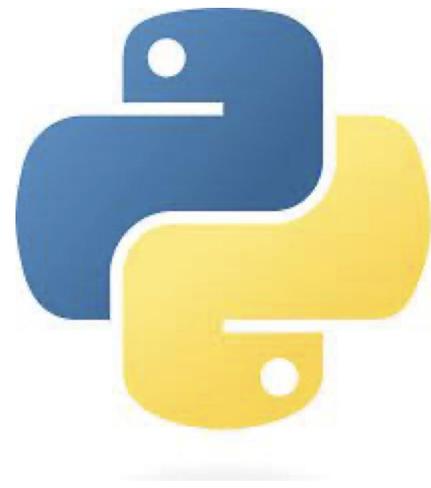


$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$



# 为什么是 Python 作为编程语言？

- 2023 级 Python 程序设计被列为选修课，请务必当成**必修课**对待！
- 简单，你可以马上开始编程。
- 如果你之前从未编程过，许多人推荐从 Python 开始。
- 一旦你具备了“编程思维”，学习新的编程语言就会容易得多。
- 广泛使用，尤其是在数据科学/人工智能领域。



# Python 编程环境

Jupyter notebook

交互式编程



运行脚本  
从命令行调用 .py 文件

lecture11\_numpy Last Checkpoint: 16 hours ago (autosaved)

### Creating numpy arrays

There are a number of ways to initialize new numpy arrays, for example from

- a Python list or tuples
- using functions that are dedicated to generating numpy arrays, such as `arange`, `linspace`, etc.
- reading data from files

### From lists

For example, to create new vector and matrix arrays from Python lists we can use the `numpy.array` function.

```
In [ ]: 1 # a vector: the argument to the array function is a Python list
        2 v = np.array([1,2,3,4])
        3 v
```

```
In [ ]: 1 type(v)
```

```
In [ ]: 1 # a matrix: the argument to the array function is a nested Python list
        2 M = np.array([[1, 2], [3, 4]])
        3 M
```

The vector has 1 dimension, the matrix has 2. We learn this with `numpy.ndim`.

```
In [ ]: 1 np.ndim(v), np.ndim(M)
```

The `v` and `M` objects are both of the type `ndarray` that the `numpy` module provides.

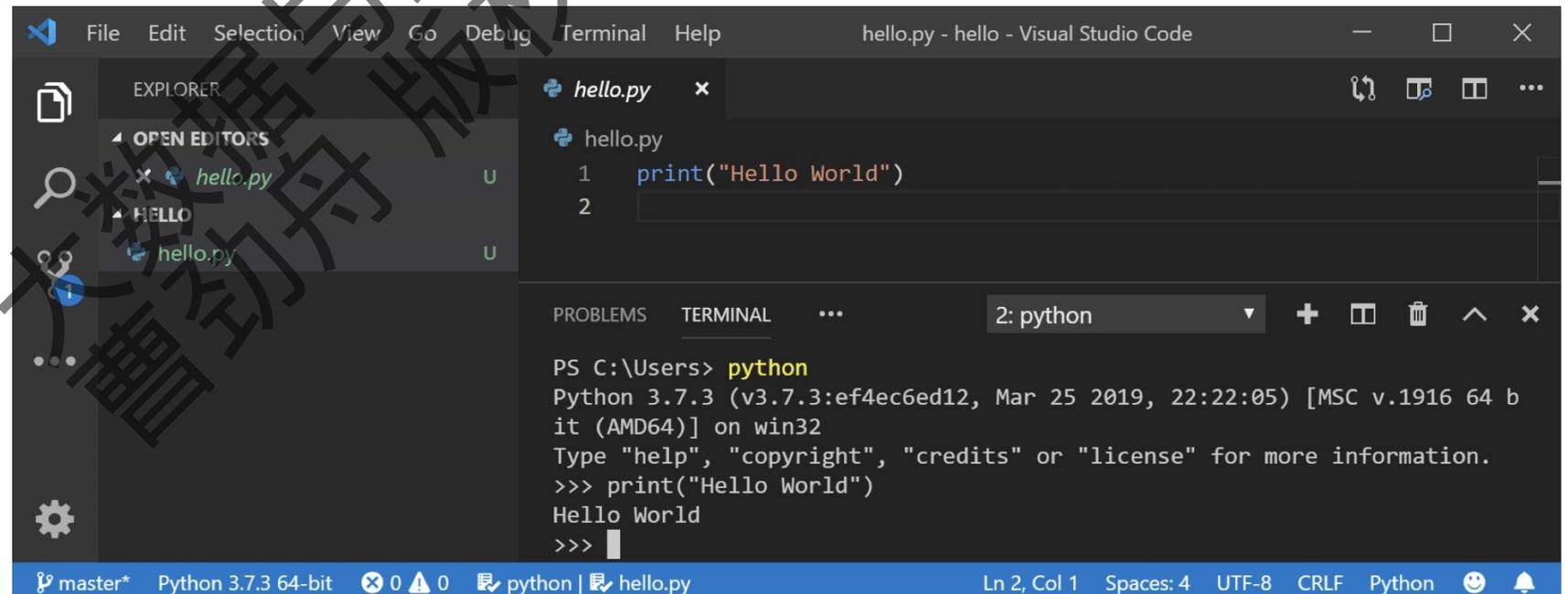
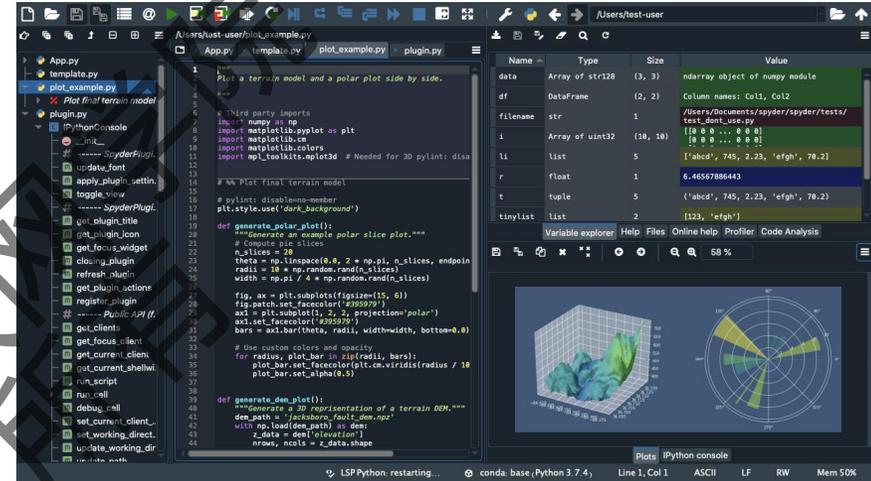
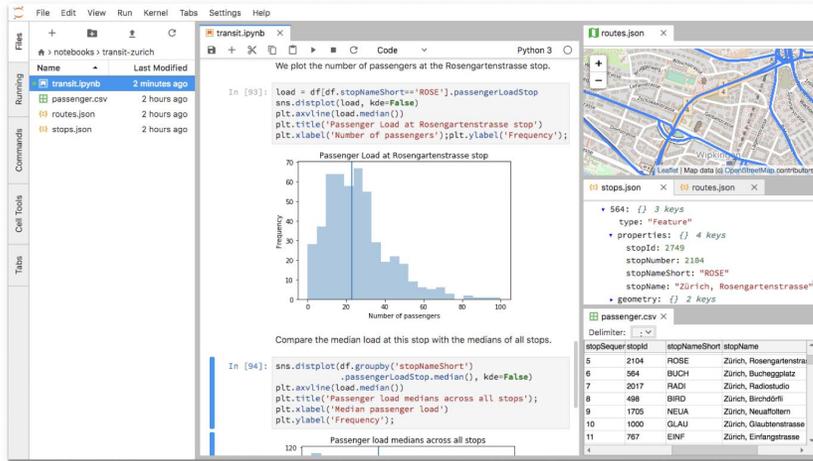
```
(base) anvy@mac622265 ~ % pwd
/Users/anvy
(base) anvy@mac622265 ~ % cd "OneDrive - ITU/teaching/ids/idsp-python/lectures/lecture07"
(base) anvy@mac622265 lecture07 % python myscript.py
The sum of 5 and 3 is 8
(base) anvy@mac622265 lecture07 %
```

```
myscript.py
1 x = 5
2 y = 3
3 mysum = x + y
4 print(f"The sum of {x} and {y} is {mysum}")
```

Line 1, Column 1 | 2023 | Tab Size: 4 | Python

文本编辑器: Sublime Text, VScode

# Python 编程 IDE (integrated development environments)



# 课程要求与考核方式

- 课程目标：用科学的方法研究和应用数据
- 考核方式
  - 课堂出勤（3%）+实验（27%）+期末项目大作业（20%）+期末考试（50%）
  - 出勤得分：每次主动回答问题，课堂表现积极，可获得考勤加分。
  - 实验作业：交电子版，具体上交方式见课程网页。**截止日期之后不接受任何补交。**
  - 期末考试：**笔试**

# 关于期末项目大作业 Final Project



# 关于期末项目大作业 Final Project



## 基于多元线性回归的奶粉评价量因素分析

### MLR-Based Milk Powder Evaluation Factors Analysis

黄梓扬 王秀文 凌荣 程哲翔 田野  
指导老师：曹劲舟

数据科学与大数据学院 | 课程作业—21号

以小组形式，提出一个有意思的研究假设或洞见，并用数据分析与大数据方法方法进行实现，并用可视化方法进行成果展示。

选题：会公布一系列建议选题，大家在建议选题中任选题目。

时间安排：

- 第2周完成小组成员组队，小组成员不超过5人。

第一阶段：提交项目介绍书（第7周截止）（篇幅至少5页，有模板），须包含以下内容：

- 文献调研——总结已有研究
- 问题陈述
- 拟使用的数据介绍，数据来源在哪里
- 拟运用的工具、方法、模型等
- 后续计划
- 小组成员分工

第二阶段：期末展示，第17或第18周

- 展示方式：海报展示
- 实践报告1份

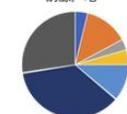
#### 研究背景

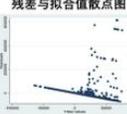
奶粉海量评价信息下的选择困境  
随着电商平台的迅猛发展，消费者在购买奶粉时越来越依赖于网络评价。面对海量的评价信息，消费者往往感到迷茫，无法准确判断奶粉的质量。了解奶粉评价量的各种因素，有助于做出更好地购买决策。

企业和商家提升市场竞争力的需求  
对于企业和商家而言，了解消费者对奶粉的评价，以及各因素对评价量的影响，有助于生产出让消费者更满意的产品、制定更具针对性的营销策略。

#### 研究过程

① 建立多元线性回归模型  
 $y = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + u_i$   
其中  $\beta_m$  是未知待估参数， $u_i$  是无法观测且满足一定条件的误差项。

② 部分数据描述性统计：  
奶源产地  


③ 异方差分析  
残差与拟合值散点图  


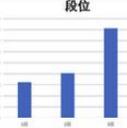
#### ④ OLS+稳健标准误回归

由怀特检验结果可知：该模型存在异方差，因此采用OLS+稳健标准误回归。

变量	评价量
团购价格	-28.5714**
商品毛重 kg	1557.143*
A1	6000*
A2	2327.345**
A3	2.189*
A4	0.0221506*
A5	6.335**
A6	2.007.084***

#### 研究数据

数据来源  
研究来自某电商平台864条关于婴幼儿奶粉的销售信息，每条信息由11个指标组成，其中评价量可以从一个侧面反映顾客对于产品的关注度

11个指标的总体情况如图：  


变量类型	变量名称	说明
评价量	评价量	间接反映顾客对产品的关注度
定量指标	商品毛重(kg)	数据位于0.12-8.64之间
	团购价(元)	数据位于9.9-25.98之间
	商品名称	共有84种不同品牌
	奶源产地	共有9个不同产地
	国产或进口	共有两个类别：国产和进口
定性指标	适用年龄岁	共有5种类别
	配方	共有3种不同配方
	分类	有2个类别：牛奶粉和羊奶粉
	段位	共有四种段位
	包装单位	共有4种包装单位

#### 研究思路

数据预处理：数据获取、检查，进行变量描述统计

异方差检验：制作拟合值与残差散点图、进行怀特检验

多重共线性检验：计算方差膨胀因子VIF，检验多重共线性

进行初步回归：建立多元线性回归模型，使用OLS+稳健标准误回归

向后逐步回归：使用向后逐步回归剔除部分变量，进行结果分析

标准化回归：使用标准化回归，探究最具影响力的因素

#### 结论

##### Conclusions

① 评价量与各因素间的关系  
通过多元线性回归分析，我们发现团购价格是影响奶粉评价量的显著因素。此外，常规配方的牛奶粉较受消费者青睐。表明消费者在选择奶粉时偏好传统而有效的配方。在奶源产地方面，中国的奶源得到了消费者的认可，同时，来自澳洲和荷兰的奶粉同样受到消费者的喜爱。在国产与进口奶粉的选择上，消费者倾向于进口产品，这可能与对品质的期待以及对品牌的信赖度有关。

② 最具影响力的因素  
由标准化回归以及评价量各因素贡献表可以发现，团购价格是最具影响力的因素。这突显了价格优惠在激发消费者购买欲望中的关键作用。团购价格的吸引力不仅能够增加消费者的购买意愿，还能够提升他们对产品的整体评价。

③ 现实意义  
对于消费者而言，本研究的结果提供了一定的购买决策参考。消费者在选购奶粉时，可以根据团购价格作为参考，寻找性价比高的产品。其次，常规配方的牛奶粉可能是一个安全的选择。此外，消费者在选择国产或进口奶粉时，可以根据自己的需求和预算进行权衡。同时，消费者也应关注产品的整体营养价值和成分安全性。对于企业和商家来说，这些发现有助于他们更精准地定位市场和制定销售策略。了解团购价格对消费者购买决策的影响，可以帮助商家制定更具针对性的营销策略。

# 联系方式

---

---

---

或注册一个 DISQUS 帐号 

 姓名

来做第一个留言的人吧!

要出售我的数据

- [www.caojz.cn/courses/idsbd2024/](http://www.caojz.cn/courses/idsbd2024/)  
讨论区
- 授课教师：曹劲舟
  - Email: caojinzhou@sztu.edu.cn
  - 办公室：C1-1402
- 微信群

大数据与互联网学院  
曹劲舟

Questions?

大数据与互联网学院  
大数据与互联网学院

