深圳技术大学 SHENZHEN TECHNOLOGY UNIVERSITY
大数据与互联网学院 COLLEGE OF BIG DATA AND INTERNET

2023-2024秋季课程

数据科学与大数据导论

Introducation 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

2023年9月

关于课程 About this course

- 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 About this course

- 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

- 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!

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**。

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）
  - Introuction，3次课
    - 1）大数据概述Introducation 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
  - 国际周、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

课程网页（这周末开通）

- www.caojz.cn/courses/idsbd2023/
- 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！！！

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
  - 作业（10%）+课堂出勤（3%）+实验报告（27%）+期末考试（60%）
  - 出勤得分：每次主动回答问题，课后来讲台登记学号姓名，可获得考勤加分。
  - 实验报告：交电子版，具体上交方式见课程网页。每迟到1天上交，总分扣10分，超过10天未交0分。

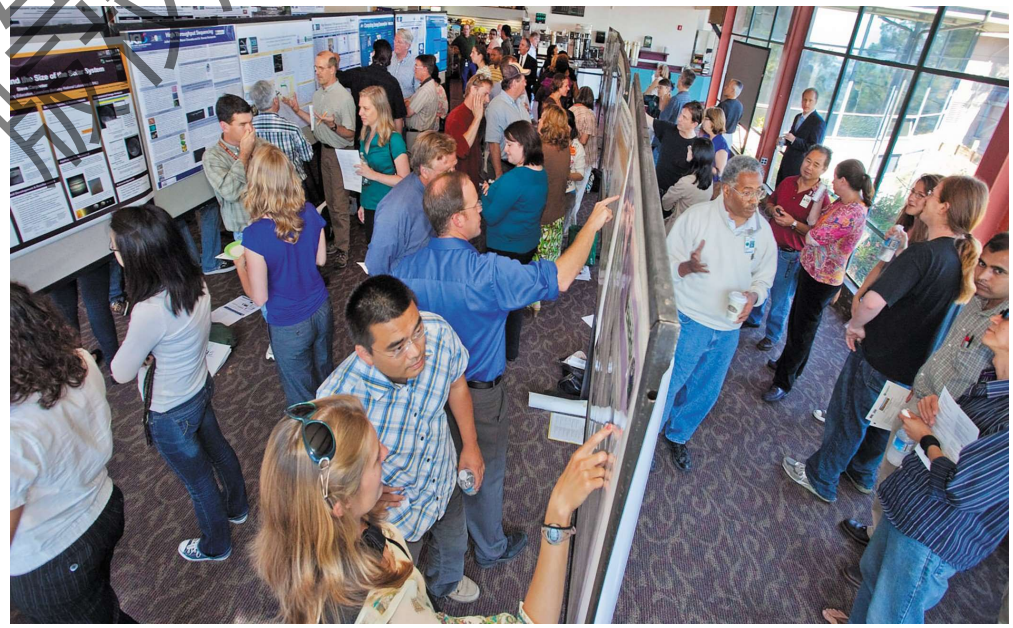
关于实验

- 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；1.2 Python 基础知识；1.3 Python 数据分析库（Numpy、Pandas、Matplotlib）	验证性	必做	6 学时	1	202309	实验室机房授课
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	

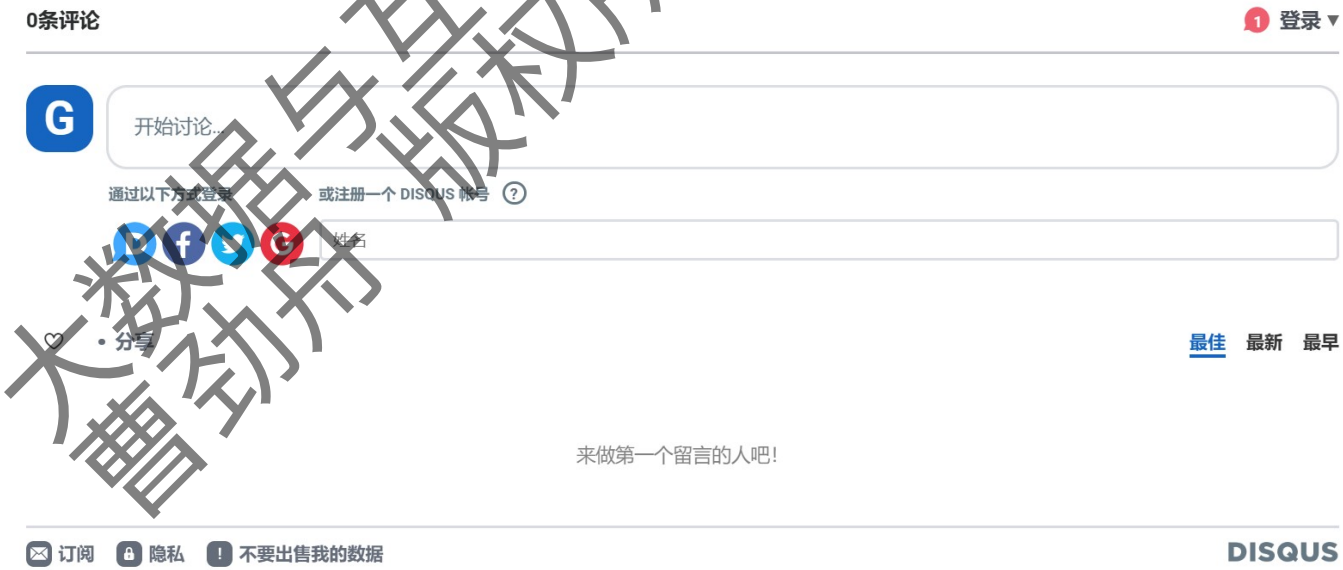
关于作业Final Project

- 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。
- 第2周完成小组成员组队，小组成员不超过5人。
- 任选题目，课程网页会公布一系列建议选题，供大家参考。
- 例如：1）地铁16号线开通后房价预测；2）微博数据的社交关系研究；
- 第一阶段：提交项目介绍书（第7周截止），须包含以下内容：
  - 文献调研
  - 问题陈述
  - 拟使用的数据介绍
  - 实现计划+拟运用的工具、方法、模型等
  - 研究计划（学期里程碑）
  - 小组成员分工
- 第二阶段：进度展示
  - 课堂PPT汇报，5页PPT，每小组5分钟
- 第三阶段：期末展示
  - 展示方式：海报展示
  - 实践报告1份



联系方式

- www.caojz.cn/courses/idsbd2023/ 讨论区
- 授课教师：曹劲舟
  - Email：caojinzhou@sztu.edu.cn
  - 办公室：C1-1419
- 微信群（不接受添加个人微信，有问题请在群内交流）



Questions?