清华大学数据库技术与应用

课程介绍

授课教师：计算机系王健楠
授课学期：2026年（春季）

目录

关于我 - About Me

清华大学 Tsinghua University

🎓教育背景与任职经历

学历背景： 清华大学博士，加州大学伯克利分校 (UC Berkeley) 博士后

任职经历： 现任清华大学计算机系长聘副教授；曾任加拿大西蒙菲沙大学 (SFU) 副教授

教学经验： 近十年海内外教学经验，长期讲授数据库系统、大数据编程、数据科学课程

CMPT 354：数据库系统

Topics

- Database History
- Relational Model
- SQL
- Relational Algebra
- Storage and Indexing
- Database Design
- Query Processing
- Transaction Processing
- NoSQL & Big Data & Cloud Databases

CMPT 732: 大数据编程

1. Course Introduction slides
2. Hadoop Concepts ["Hadoop Concepts" slides]
3. Python Preliminaries ["Python Preliminaries" slides]
4. Spark Concepts ["Spark Concepts" slides]
5. Spark DataFrames Concepts ["Spark DataFrames Concepts" slides]
6. Cloud & Data Management
7. Why MapReduce? ["Why MapReduce?" slides]
8. 21st Century Databases
9. NoSQL & Cassandra ["NoSQL & Cassandra" slides]
10. Spark Machine Learning ["Spark Machine Learning" slides]
11. Spark Streaming ["Spark Streaming" slides]
12. Data Warehouse/Lake Explainer
13. Other Big Data Tools ["Other Big Data Tools" slides]
14. Small Data ["Small Data" slides]
15. NumPy/Pandas Speed ["NumPy/Pandas Speed" slides]
16. Dask & Ray ["Dask & Ray" slides]

CMPT 733: 大数据科学

Topics

- Introduction to Data Science
- Data Preparation
- Visualization
- Statistics
- Deep Learning
- Practical Machine Learning (AutoML, Explainable AI,
- Anomaly Detection
- Cloud Computing
- Responsible Data Science
- Communication

你与“数据”的距离

基于数据的决策

你最难忘的“数据驱动”决策是什么？

选专业　选课　人生大事

核心概念辨析

你理解这些术语以及区别吗？

- **Database** vs **DataFrame**
- **Data Warehouse** vs **Data Lake**

面临的数据痛点

在你所在的专业（计算机/化学/社科等）中，最头疼的数据问题是什么？

格式混乱　数据量大　模态丰富　工具难用

技能树快速自测

现场投票：你的工具箱里有什么？

- Python
- SQL
- Excel
- ChatGPT

目录

课程定位：培养 AI 时代的数据基础能力

AI 时代背景

在人工智能迅速发展的今天，数据已成为驱动 AI 突破的核心要素。如何高效管理复杂多源数据，是 AI 时代学生不可或缺的底层能力。

大模型训练　AI4Science　数据驱动决策

数据科学

学习如何运用核心技术手段：

- 数据处理与清洗
- 可视化分析
- 统计建模

从数据中发现规律
形成可信结论

数据工程

构建稳健的数据系统基础设施：

- SQL 查询
- 性能优化
- 数据建模

从实验环境 Demo
推进到生产级系统

Data + AI

探索数据库与 AI 的双向赋能：

- **AI4Data:** AI 赋能数据库
- **Data4AI:** 数据库支撑 AI

掌握 AI 时代
数据底层核心能力

AI 时代的数据挑战

大模型训练：大规模数据量

从传统小样本学习向海量数据预训练范式转变，数据规模呈指数级飞跃。

45 TB	PB+
GPT-3 文本	多模态模型

AI4Science：高质量科学数据

高质量科学数据成为科研新基建，AI 正在加速基础科学的发现过程。

AlphaFold　材料科学数据库　AI 辅助新药研发

AI Agent：数据分析不可信

AI Agent 数据分析结果不可盲信，必须打破“黑盒”，建立完全可控的调试机制。

- 可观测行为
- 可调试过程
- 可控制结果

数据模态：多模态数据处理

突破单一 Table 限制，掌握全链路异构数据处理，构建从非结构化到结构化的价值管道。

结构化 (Table)　半结构化 (JSON)　非结构化（文本）

完成本课你将获得什么

硬核能力

数据科学维度：Pandas数据清洗、高维可视化与统计分析，从数据中发现规律。

数据工程维度：SQL查询及优化、索引设计与数据建模，半结构化数据处理。

Data+AI维度：向量数据库、Text-to-SQL、AI数据湖、数据智能体等

01

实战产出

全栈实验集：8次循序渐进的实验(A1-A8)，覆盖全链路技能。

AI+Data项目：结合真实领域数据与AI技术的端到端解决方案。

工程化交付：符合工业界规范、可复现的代码仓库 (GitHub/GitLab)。

02

前沿素养

Data+AI融合思维：深度掌握 AI4Data 与 Data4AI 的双向赋能逻辑。

可信AI能力：能看懂生成式AI处理分析数据的路程，并进行修改，保证结果的可信

跨学科与AI4Science：运用AI与数据技术解决跨领域科学问题的能力。

03

目录

课程内容全景图

清华大学 Tsinghua University

阶段 1: 数据科学

- W1 课程介绍与数据库历史
- W2 Pandas I：基础数据操作
- W3 Pandas II：高级数据操作
- W4 数据准备 I：结构化数据
- W5 数据准备 II：非结构化数据
- W6 数据可视化
- W7 数据统计
- W8 数据驱动的机器学习

阶段 2: 数据工程

- W9 SQL I：基础查询
- W10 *五一假期*
- W11 SQL II：高级查询
- W12 查询性能优化
- W13 数据建模
- W14 半结构化数据

阶段 3: Data + AI & 项目展示

- W15 **LLM4Data**
  LLM赋能数据分析
- W16 **Data4LLM**
  面向LLM的数据系统
- W17 **课程总结与项目展示**
  Final Presentation

第1周：为什么学习数据库历史

清華大學 Tsinghua University

1. 理解技术演进逻辑

技术变革不是无缘无故的，每一次改变都是为了解决当时最棘手的痛点。理解背景才能在面对新问题时选对工具。

2. 避免重复造轮子

学会站在巨人的肩膀上，前人已经解决过的难题，不需要我们从零开始探索。理解历史，避免闭门造车。

第2-3周：为什么学习Pandas

极高的社区热度

4.5 亿+

PyPI 月下载量

45k+

GitHub Stars

数据科学领域的通用语言。无论是学术界还是**工**业界，Pandas **都是** Python **数据分析生**态的基**石**。

强大且灵活的API

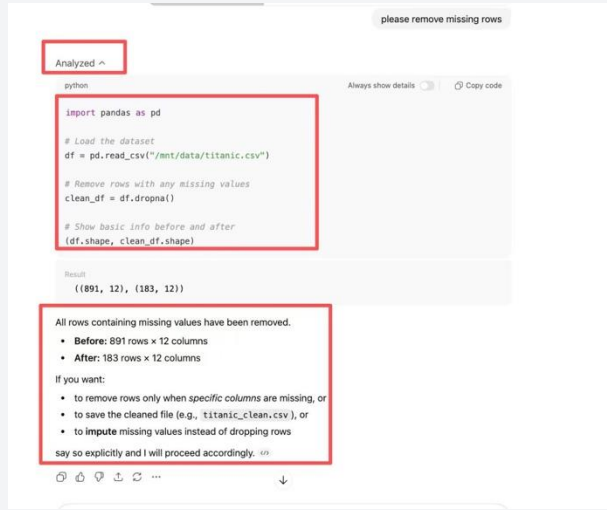
1000+

API 方法函数

清洗　转换　聚合　可视化

提供丰富且一致的 API **接口**，**一行代**码即可完成**复**杂的数据变换、时间序列处理和多表关联。

大模型很懂Pandas



作为最流行的库，Pandas **是大模型**训练语料中**的"一等公民"**。LLM **默**认使用pandas**来**处理分析**数据**。

第4-5周：为什么学习数据准备

真实数据极其“脏乱”

真实世界的数据充斥着缺失值、重复项和异常格式。遵循 **"Garbage In, Garbage Out"** 原则：如果输入数据质量差，再好的算法也无法产生有价值的结论。

数据质量问题示例

ID	Age	Date	Score
01	25	2024-01-01	98
02	NaN	Jan 2, 24	85
03	199	2024/01/03	-1
01	25	NULL	98 (重复)

数据分析的“瓶颈”

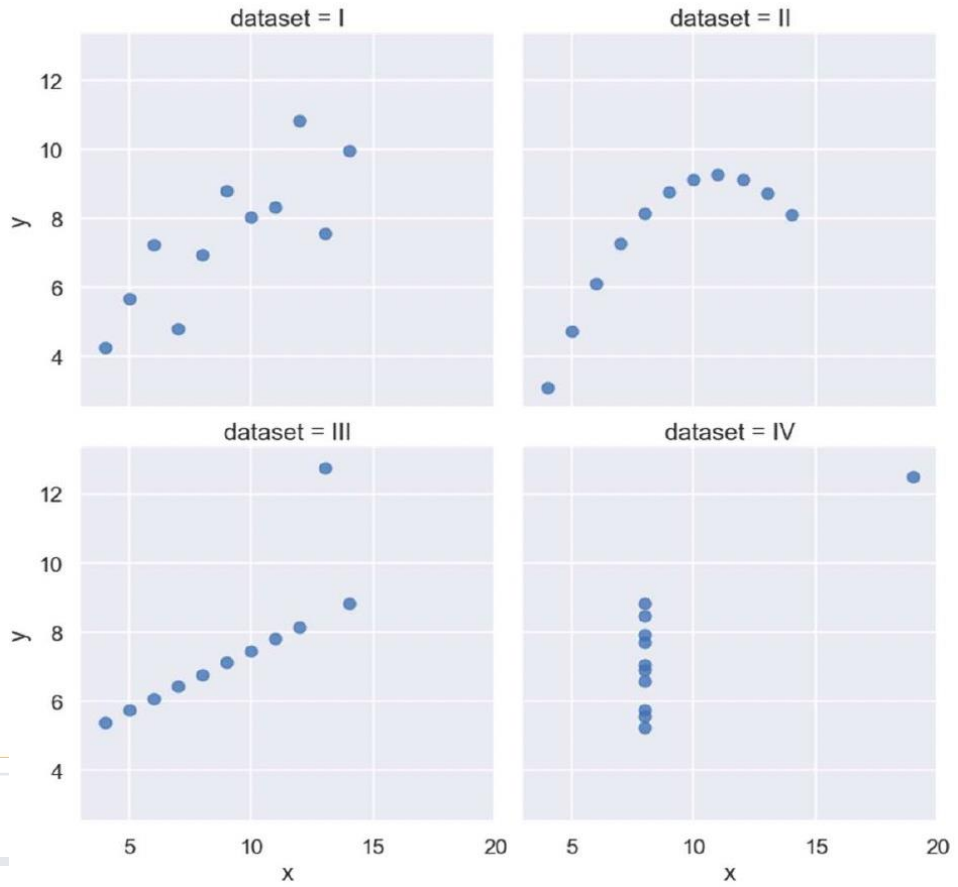
数据准备通常占用数据科学家 **50%-80% 的工作**时间。这是最耗时、**最枯燥，但也是最关**键的环节，直接决定项目的成败。

典型数据科学项目时间分配

第6周：为什么学习数据可视化

可视化帮助理解数据

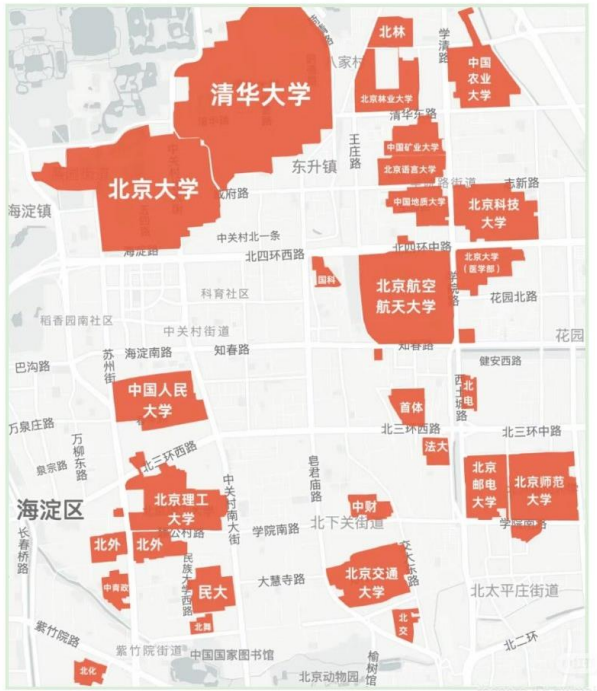
有些规律仅凭统计指标无法发现。只有通过可视化，数据的真实形态（异常、非线性、聚类）才能显现。



可视化帮助展示洞察

"A picture is worth a thousand words." —— 用直观图形降低沟通成本

北京市海淀区高校分布图



第7周：为什么学习数据统计

统计学会说谎

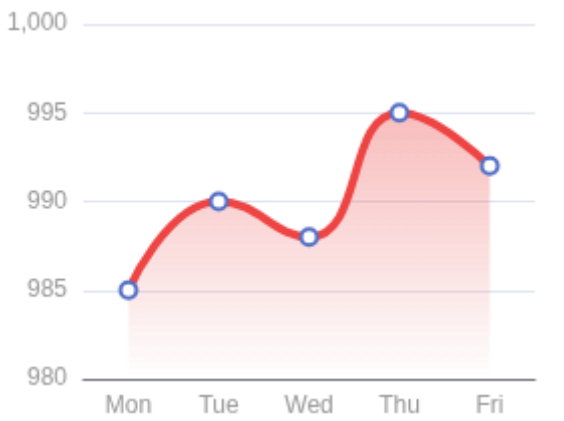
How Statistics Lie

数据本身不会说谎，但呈现数据的方式会。**通**过改变坐标轴、选择性展示或不当分组，**同一**组数据可以被包装成完全相反的结论。

> *“**学**习统计学不仅是为了用数据说话，更是为了识**破数据**谎言，避免被误导。”*

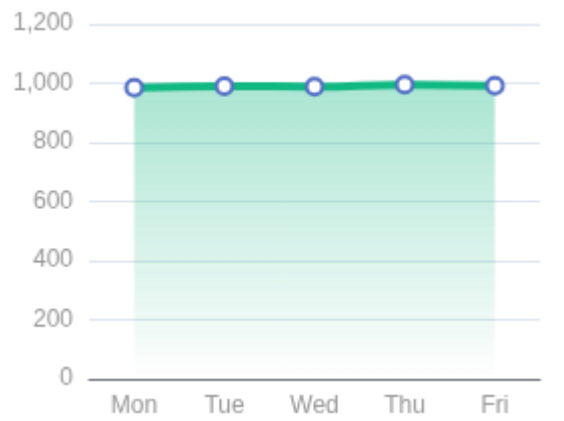
经典案例：Y轴截断的视觉魔术

误导性展示 (截断Y轴)



结论：业绩似乎在“爆发式增长”

真实展示 (完整Y轴)



结论：业绩实际上基本持平

特征工程 > 算法选择

> “数据和特征决定了机器学习的上限，
> 而模型和算法只是逼近这个上限。”

80%

数据科学家认为
特征工程是影响性能的首要因素

Source: Kaggle 2023 Survey

简单模型 + 好特征 **>** **复杂模型** + 差特征

数据质量胜过算法复杂度

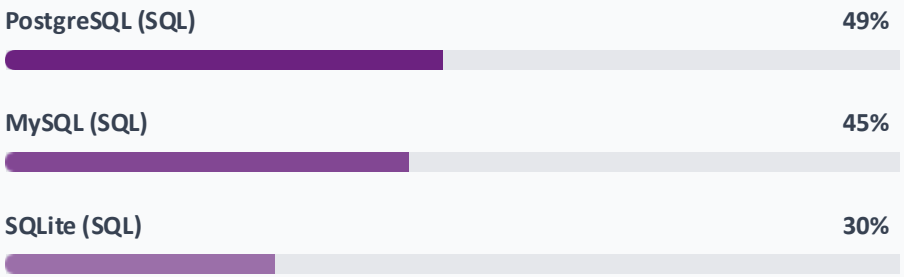
第9-11周：为什么学习SQL

数据查询的
通用语言

无论后端使用 MySQL、PostgreSQL 还是 Oracle，SQL **都是沟通的**标准协议。它是数据从业者的**“普通”**话，**也是**连接数据与价值的桥梁。

Stack Overflow 开发者调查

最受欢迎的数据库技能：



* 数据来源: Stack Overflow Survey 2023

声明式编程
降低门槛

只需告诉数据库**“要什么”** (What)，**而无需关心“怎么做”** (How)。底层执行计划、索引选择和数据检索由数据库引擎自动优化完成。

SQL 示例

需求：**找出所有薪**资高于 20k **的**员工

query.sql

```sql
SELECT name, department, salary
FROM employees
WHERE salary > 20000
ORDER BY salary DESC;
-- 数据库自动选择最优索引路径
```

无需编写循环、文件IO或内存管理逻辑

第12周：为什么学习查询性能优化

解决 SQL 运行慢的问题

避免用户在屏幕前苦苦等待，掌握从“慢”到“快”的核心技术。

优化前 Before

10 s

漫长等待

学习优化

优化后 After

0.01 s

秒级响应

第13周：为什么学习数据建模

- Ted Codd 因提出关系模型而获得图灵奖。
- 排名前五的数据库引擎中，有四个是关系型数据库。

Rank			DBMS	Database Model
n 2	Dec 2021	Jan 2021		
..	1.	1.	Oracle	Relational, Multi-model
2.	2.	2.	MySQL	Relational, Multi-model
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model
4.	4.	4.	PostgreSQL	Relational, Multi-model
5.	5.	5.	MongoDB	Document, Multi-model

第14周：为什么学习半结构化数据

response.json

```
{
  "user_id": 1024,
  "username": "data_master",
  "tags": ["admin", "contributor"],
  "metadata": {
    "last_login": "2024-05-20T10:30:00Z",
    "device": "iPhone 15",
    "preferences": null
  },
  "logs": [
    { "id": "A1", "action": "click" },
    { "id": "B2", "error": true }
  ]
}
```

真实世界不是完美的表格

Real World Data != Perfect Tables

现代 Web 和移动应用每秒都在产生海量数据，它们不再是整齐划一的二维表，而是包含嵌套、层级和稀疏字段的**半结构化数据**。传统关系型数据库（Schema-on-Write）难以应对这种灵活多变的结构。

- REST API 响应
- 系统配置文件
- 应用运行日志
- NoSQL 文档存储

掌握 JSON/XML 处理能力，是适应现代数据生态（Data Ecosystem）的必备技能。

第15-16周：为什么学习Data+AI

AI4Data

AI 赋能数据库系统

LLM 生成 SQL (Text-to-SQL)

大模型理解自然语言，自动生成复杂的查询语句，降低使用门槛。

自动化数据清洗

AI 自动识别异常值、补全缺失数据，智能提升数据质量。

智能查询优化

基于学习的优化器 (Learned Optimizer) 预测查询成本，从根本上提升性能。

相互赋能

Data4AI

数据系统支撑 AI 应用

向量数据库支持 RAG

存储和检索高维向量，为大模型提供私有知识库，解决幻觉问题。

高质量数据训练模型

数据工程保障数据血缘与质量，决定了 AI 模型的上限。

数据管理保障 AI 可信

通过权限控制、隐私计算与合规审计，确保企业级 AI 的安全性。

目录

教学日历 - Syllabus

清華大學 Tsinghua University

周次	日期	教学主题 (Topic)	作业发布 (Released)	作业截止 (Due)
W1	02-26	**课程介绍与数据库历史**	A1	—
W2	03-05	Pandas I：基础数据操作	A2	—
W3	03-12	Pandas II：高级数据操作	—	A1
W4	03-19	数据准备 I：结构化数据	A3	A2
W5	03-26	数据准备 II：非结构化数据	—	—
W6	04-02	数据可视化	A4	A3
W7	04-09	数据统计	—	—
W8	04-16	数据驱动的机器学习	A5	A4
W9	04-23	SQL I：基础查询	—	选题报告
W10	*04-30*	*五一假期 (Labor Day Holiday)*	*A6*	*A5*
W11	05-07	SQL II：高级查询	—	—
W12	05-14	查询性能优化	A7	A6
W13	05-21	数据建模	—	—
W14	05-28	半结构化数据	—	—
W15	06-04	LLM4Data (LLM × Data Analytics)	A8	A7
W16	06-11	Data4LLM (Data Systems for LLMs)	—	—
Final	—	**课程总结与项目展示**	—	A8 项目终审

评分标准

考核项	占比	说明
平时实验	32%	共8次个人实验 • 每次占比 4%
课程项目	68%	团队合作综合大作业 • 选题汇报 (18%) • 项目汇报 (25%) • 项目成果 (25%, 含报告、代码与视频)

课程项目

项目形式与选题

团队规模

2-3人/组，鼓励跨专业组队（如CS + 化学/金融），发挥多学科背景优势，**共同解决复**杂问题。

数据来源

必须使用**真实场景数据，拒**绝"**玩具数据集**"。**数据量需达到一定**规模，体现数据处理的真实挑战。

项目类型

- **AI4Science**：**数据**驱动的科学发现
- **Domain DS**：**特定**领域的数据科学项目
- **LLM App**：**基于大模型的数据**应用
- **Data Eng**：**端到端数据工程系**统

核心实施要求

完整流程

项目需覆盖数据科学全生命周期，用真实的数据解决真实的问题

1 问题定义 → 2 数据获取 → 3 **整合清洗** → 4 建模分析 → 5 结论展示

技术栈要求

必须包含 **SQL 查询** 与 **Python 数据处理** 环节。鼓励结合使用 Pandas、Scikit-learn、PyTorch **等工具**库。

可复现性

提交的代码库需包含完整的 **README 文档**、环境配置文件 (requirements.txt) **与运行**说明，可一键复现。

课程政策

学术诚信与作业规范

不要迟交

每个人有 **2天** 的作业延期预算

一旦预算用完，每迟交一天扣除 **20%** 分数

不要作弊

我们将对所有代码和报告进行严格的 **查重检测**

一旦发现抄袭，最终总成绩将 **扣除 30%**

AI 使用态度与规范

平时实验

禁止直接使用 AI 完成代码。
实验旨在训练基础编码能力与底层原理理解，过度依赖 AI 将导致核心能力缺失。

课程项目

鼓励使用 AI 作为辅助工具。
在解决复杂问题时，可以使用 AI 进行思路启发、代码优化或辅助 debug，但需在报告中声明 AI 的使用部分。

参考的课程资源

UC Berkeley

Department of Electrical Engineering and Computer Sciences

Data 100

数据科学原理与技术。涵盖数据生命周期：数据收集、清洗、可视化和建模。

Data 101

数据工程。专注于可扩展数据系统、SQL 和大数据处理框架。

Simon Fraser University

School of Computing Science

CMPT 354

数据库系统 I。介绍数据库设计、SQL 编程和数据库应用开发。

CMPT 733

大数据实验室。涵盖大数据分析、分布式系统和机器学习流水线的高级主题。

联系方式与期待

联系方式

邮箱: jnwang@tsinghua.edu.cn

办公室: 清华大学 自强科技楼1号楼-813

课程交流: http://learn.tsinghua.edu.cn (网络学堂)

寄语与期待

- 保持好奇
- 数据思维
- 动手实践
- 学术诚信

祝大家在本学期学有所获，学习顺利！