Database System Principles
《数据库系统原理》

李文根/Wengen Li
Email: lwengen@tongji.edu.cn
先进数据与机器智能系统实验室
Advanced Data and Machine Intelligence Systems (ADMIS) Lab
https://admis-tongji.github.io
同济大学 计算机科学与技术学院
2026年03月

关于课程

- **课程学分**：3
- **理论课时**：48 (理论与课堂研讨)
- **实验课时**：16 (6-8次实验课)
- **实践课时**：2026年秋课程设计
- **课程教师**：**李文根**、关佶红、张毅超

- **助教（关老师）**
  - 刘凌诚：18807331552，2512055@tongji.edu.cn
  - 陈凌锐：17663003539，lingchan@tongji.edu.cn
  - 韩嘉睿：15815673200 2534007@tongji.edu.cn
- **助教（张老师）**
  - 彭俊熙：17871940933，2534121@tongji.edu.cn
  - 程瑞真：18755932687，2534005@tongji.edu.cn
  - 杨皓钦：18264688411 2512075@tongji.edu.cn
- **助教（李老师）**
  - 彭兆祥：19946039880，2533987@tongji.edu.cn
  - 张铭锐：13615669776，2534017@tongji.edu.cn
  - 王腾博：15690353715 tbwang@tongji.edu.cn

课程教师

李文根

课程时间: Mon.10:00~11:35(G101), Wed.10:00~11:35 (单周，G101)

答疑时间: Tue. 13:30-16:30, Thu.13:30-16:30 (智信馆410)

Email: lwengen@tongji.edu.cn

Online: QQ, Email

研究方向:

多模态AI（面向工业生产和社会治理）

时空智能（面向海洋）

视觉智能（卫星遥感图像处理）

课程教师（续）

Prof. Guan Jihong (关佶红)

ADMIS Lab: Room 429B, Zhixin Building, Research: Data Management, Data Mining, Big Data, Machine Learning, AI, Bioinformatics, et al

Lecture hours: Mon.10:00~11:35, Wed.10:00~11:35(Odd)

Office: Room 429B/458, Zhixin Building, Jiading campus

Office Hour: Mon.12:30-16:00, Wed.13:00-15:00

Online: Tencent Meeting, ML/AI/BD/Bioinfomatics

Tel: 186-1610-2875; Email: jhguan@tongji.edu.cn

课程教师（续）

Associate Prof. Yichao Zhang (张毅超)

Lecture Hours: Mon.10:00~11:35, Wed.10:00~11:35 (单周)

Office Hours: Tue. 13:30-16:30, Thu.9:30-11:30 (智信馆410)

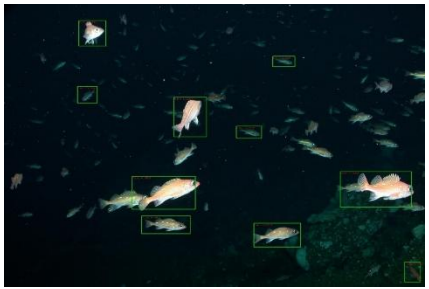
Email: yichaozhang@tongji.edu.cn

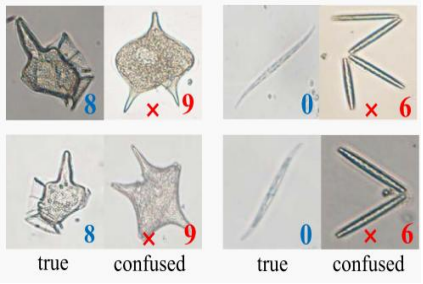
Online: Email, Wechat

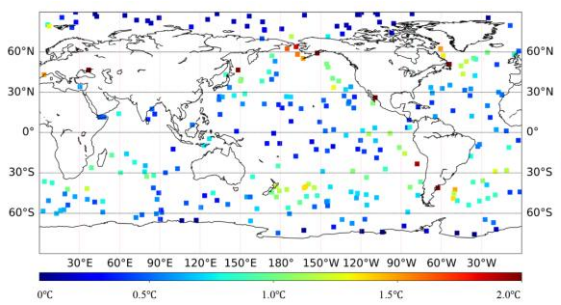
Research interests: Graph neural networks, link prediction, weighted network modeling, random diffusion, network games, financial data analysis, and urban disaster prevention.

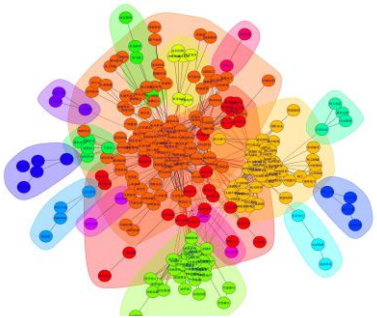
ADMIS课题组研究方向

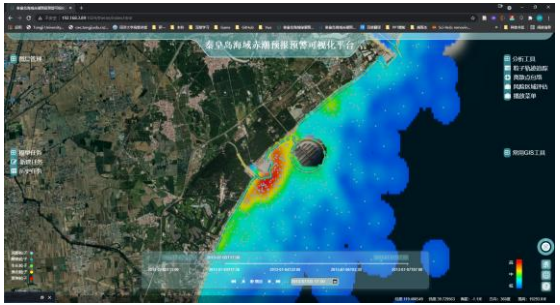
- **Spatio-temporal intelligence**: ocean computing, urban computing
- **Multi-modal AI**: sentiment analysis, knowledge graph, cross-modal retrieval
- **Bioinformatics**: drug discovery, protein data analytics, cell data analytics











海洋时空智能（OSTI）

应用场景

海洋环境监测与保护 | 海洋气候监测与预测 | 海洋渔业管理 | 海洋灾害预防与处置 | 航运与海事安全 | 海洋能源与工程 | 海洋探索

时空分析任务与方法

时空预测	目标检测与识别	异常/事件检测	时空模式挖掘
ESWA26、TGRS25、TIST25 TGRS24、TIST23、JSTARS23	TGRS25, IET CV23 IJCNN21、JEI22	ES26、Ultrasonics25 OE24、RS23、CIKM22	TKDE25、TKDD25 TAI24、T-Cyber22
Temperature, Chlorophyll-a, Ice, Current, DO, Salinity, Tide	Ship, Submarine, UUV, Marine Animal, Plankton	Eddy, Trajectory, El Nino, Typhoon, Pollution, OMZ	Clustering, Periodicity, Correlation, Causality

SVM | DRL | AE | CNN | RNN | Transformer | Attention | GCN | GAN | ··· | Diffusion

时空数据治理

TGRS25、CVPR25
TGRS24、RS23

- 数据质量控制
- 数据同化
- 卫星数据反演

时空数据补全

- 数值模型
- 传统机器学习模型
- 深度学习模型

RS26、TIST25、TAI25
TAFFC25、TAFFC24

时空数据融合

- 同构融合
- 异构融合

多源海洋时空数据

原始数据：卫星数据、原位数据、船舶数据、再分析数据

特征

- 大空间尺度
- 高稀疏性
- 数据多源异构性
- 复杂的时空依赖

可视化

- 环境要素
- 海洋事件
- 海洋模式

时空基础模型

海洋知识图谱

多模态AI：工业产线多模态带参知识图谱

构建融合多源异构数据与连续工艺参数的多模态带参知识图谱，支撑创成式设计及协同创新

工业工艺多源异构数据

人—机—料—法—环

- **操作人员技能数据**（人：人员管理）
- **设备布局图**（机：设备实体）
- **物流路径轨迹**（料：物料流动）
- **工艺参数文档**（法：工艺规划）
- **环境传感器数据**（环：温湿度/光照）
- ⋮

多源异构数据融合

- **“设备-工序-物料-参数”多维本体模型构建**
- **语义增强的多模态**（文本:工艺作业指导书/图像:车间设备布局图/表格:节拍平衡表）**特征对齐方法**
- **“布局拓扑-工艺语义-参数区间”深度融合的统一语义空间映射**

多模态带参知识联合抽取

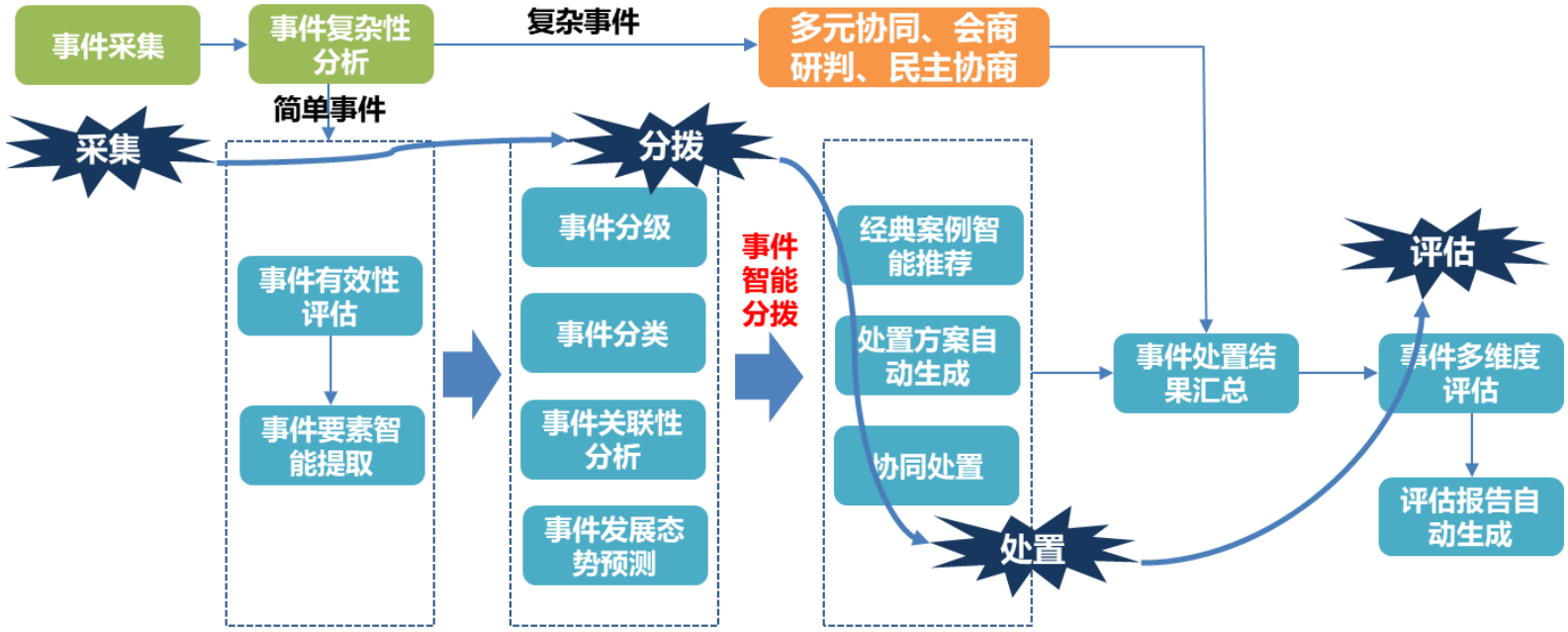
- **基于融合数据的实体-关系-参数联合抽取**
- **面向物流节拍的“轨迹-时序”时空模式挖掘算法**
- **支持区间约束与趋势描述的连续参数建模**

知识图谱检索与分析

- **面向混线产线配置场景的结构关系与连续工艺参数约束查询**
- **支撑创成式方案生成的结构化图拓扑与多模态向量语义混合检索**
- **适配仿真验证需求的连续参数索引与计算表达**

多模态AI：基层社会治理事件全周期治理

构建事件智能分拨与全周期管理方法体系，有效解决事件有效性判断、事件要素自动提取、事件自动分级分类、事件关联性因果性分析、事件发展态势研判、事件处置支持、事件综合评估等技术难题



数据（Data）

- **什么是数据?**
  - 在现实生活，数据是可识别的抽象符号，即描述事物的符号记录
  - 在计算机中，数据是所有能被计算机处理的符号的总称
  - 在数据库中，数据是存储的基本对象
- **数据的种类:**
  - 数字，如 1、2、3
  - 文本，如张三、李四
  - 时间，如2022年2月22日
  - 向量，如图片或文本的表征
  - 图片，如同济logo
  - 音频，如电视语音
  - …

数据的类型

- **结构化数据**：关系数据（表）
- **半结构化数据**
  - 键值对Key-Value
  - Markdown、XML、JSON
  - 图
  - 向量
- **非结构化数据**
  - 文本文档、电子邮件、图像、音频、视频

关系数据（Relational Data）

同济大学 TONGJI UNIVERSITY

Sno（学号）	Sname（姓名）	Sgender（性别）	Sage（年龄）	Sdept（系别）
2021310721	李博	男	17	CS
2021310722	赵宇	男	19	CS
2021310723	张敏	女	18	CS
2021310724	王勇	男	18	MA
2021310725	刘佳	女	17	MA

Sno（学号）	Cno（课程号）	Grade（成绩）
2021310721	5	98
2021310722	1	87
2021310723	1	92
2021310723	5	76
2021310724	7	84
2021310725	4	95

Cno（课程号）	Cname（课程名）	Cpno（先修课）	Ccredit（学分）
1	数据库	2	4
2	数据结构与算法	6	4
3	操作系统	2	3
4	高等数学		4
5	软件工程	6	2
6	程序设计		3
7	数值分析	4	2

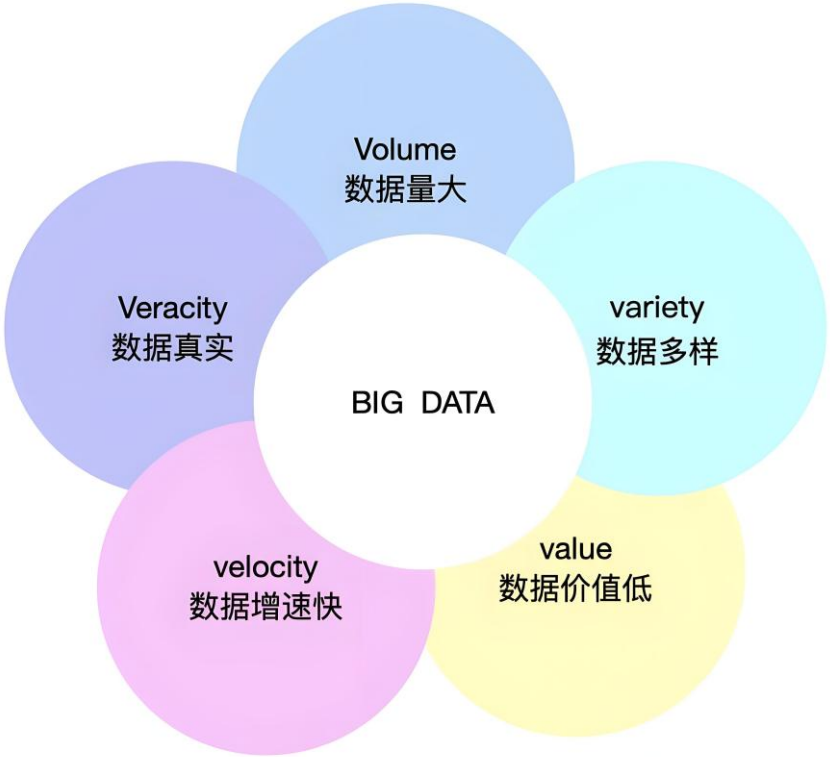
关系模型：二维表格（表）来表示数据

- 表中的每一行代表一个记录（实体或元组）
- 表中的每一列代表一个属性（字段）

查询赵宇数据库课程成绩

1) Student表查找赵宇学号
2) Course表查找数据库课程号
3) SC表查找成绩

大数据（Big Data）


Volume
数据量大
variety
数据多样
value
数据价值低
velocity
数据增速快
Veracity
数据真实
BIG DATA

Big Data:A Joke

某比萨店的电话铃响了，客服人员拿起电话。
客服：XXX比萨店。您好，请问有什么需要我为您服务？
顾客：你好，我想要一份……
客服：先生，烦请先把您的会员卡号告诉我。
顾客：16846146***。
客服：陈先生，您好！**您是住在泉州路一号12楼120x室**，请问您想要点什么？
顾客：我想要一个海鲜比萨……
客服：陈先生，海鲜比萨不适合您。
顾客：为什么？
客服：**根据您的医疗记录，你的血压和胆固醇都偏高**。
顾客：那你们有什么可以推荐的？
客服：您可以试试我们的低脂健康比萨。
顾客：你怎么知道我会喜欢吃这种的？
客服：**您上星期一在中央图书馆借了一本《低脂健康食谱》**。
顾客：好。那我要一个家庭特大号比萨，要付多少钱？
客服：**99元，这个足够您一家六口吃了。但您母亲应该少吃，她上个月刚刚做了心脏搭桥手术，还处在恢复期**。
顾客：那可以刷卡吗？
客服：陈先生，对不起。请您付现款，因为**您的信用卡已经刷爆了，您现在还欠银行4807元，而且还不包括房贷利息**
顾客：那我先去附近的提款机提款。
客服：陈先生，**根据您的记录，您已经超过今日提款限额**。
顾客：算了，你们直接把比萨送我家吧，家里有现金。你们多久会送到？
客服：大约30分钟。如果您不想等，可以自己骑车来。
顾客：为什么？
客服：**根据我们全球定位系统的车辆行驶自动跟踪系统记录。您登记有一辆车号为XX-748的摩托车，而目前您正在解放路东段华联商场右侧骑着这辆摩托车**。
顾客：……

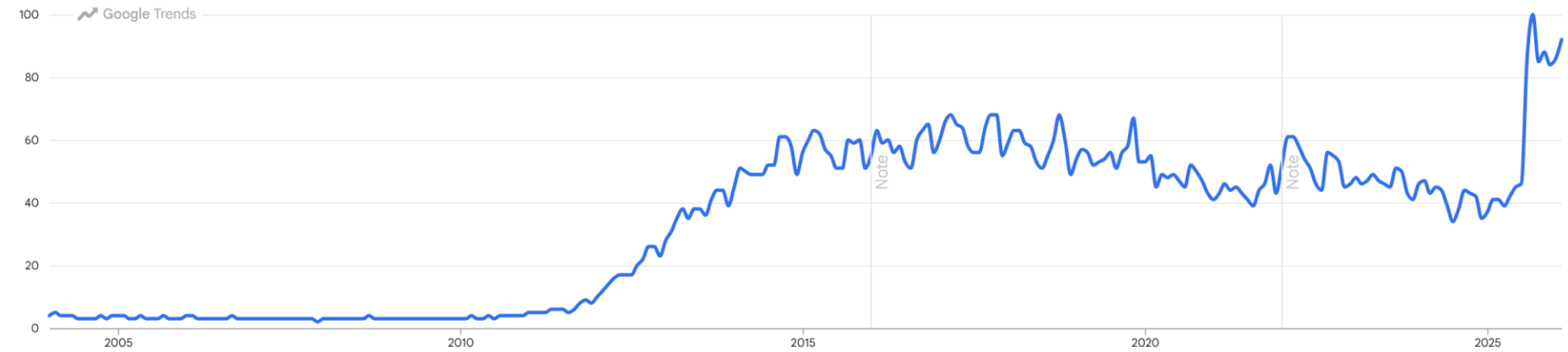
家庭住址
医疗记录
医疗记录
借阅记录
借贷记录
定位跟踪

全球大数据热度变化

Worldwide | 2004 – present | Web Search

Interest over time

Worldwide · 2004 – present



Big Data

From Google trends (https://trends.google.com/trends), 2026/3/1

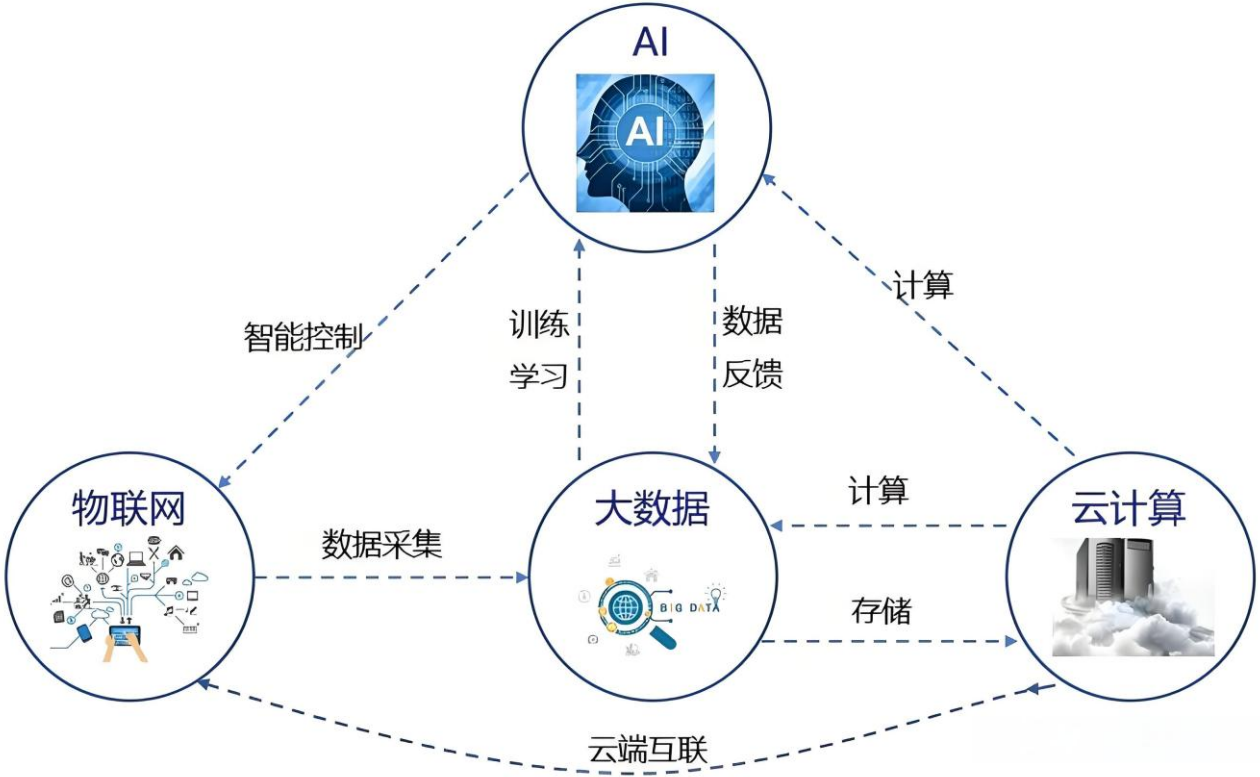
全球LLM热度变化

From Google trends (https://trends.google.com/trends), 2026/3/1

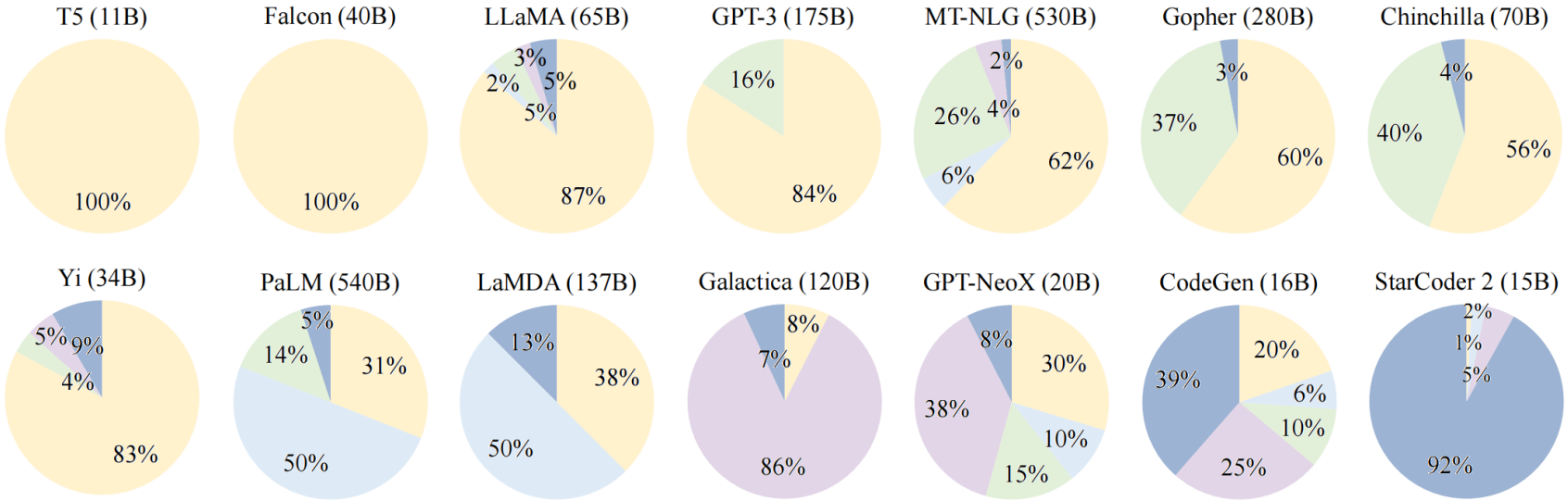
生成式大模型

- **大语言模型（Large Language Model，LLM)**
  - 一种基于深度学习和海量文本数据训练的人工智能模型，能够理解、生成和推理自然语言
  - GPT系列、Claude系列、LLaMA系列、DeepSeek、通义千问、文心一言
- **视觉大模型（Large Vision Model, LVM）**
  - 基于大规模视觉数据训练，具有通用视觉理解与生成能力的深度学习模型
  - Flamingo、InternVL、Imagen、Sora、VideoPoet
- **多模态大模型（Multimodal Large Model，MLM）**
  - 能够同时处理和理解多种模态数据（如文本、图像、视频、音频等）的通用人工智能模型，通过跨模态对齐学习，实现不同模态信息的统一表征与联合推理
  - GPT-4V、Flamingo、Qwen-VL
- **图大模型（Large Graph Model, LGM）**
  - 处理图结构数据的大规模深度学习模型，能够对节点、边和图全局进行表征学习、关系推理与生成，挖掘图数据中的拓扑结构、语义关联和动态演化规律
  - GraphGPT、GraphGym、GraphMAE

Big Data & AI



大模型训练数据



Webpages	C4 (800G, 2019), OpenWebText (38G, 2023), Wikipedia (21G, 2023)
Conversation Data	the Pile - StackExchange (41G, 2020)
Books & News	BookCorpus (5G, 2015), Gutenberg (-, 2021), CC-Stories-R (31G, 2019), CC-NEWES (78G, 2019), REALNEWs (120G, 2019)
Scientific Data	the Pile - ArXiv (72G, 2020), the Pile - PubMed Abstracts (25G, 2020)
Code	BigQuery (-, 2023), the Pile - GitHub (61G, 2020)

多源、异构、多模态大数据

基础设施
Infrastructure

车辆和人员
Vehicles and Employees

人口、土地、经济
Population, Land, Economics

气象、环境等
Weather, Activities

交通检测数据
Detection Devices

交通行为数据
Behavior Data

综合交通调查
Comprehensive Census

交通基础数据 Basic Data

动态运行数据 Dynamic Data

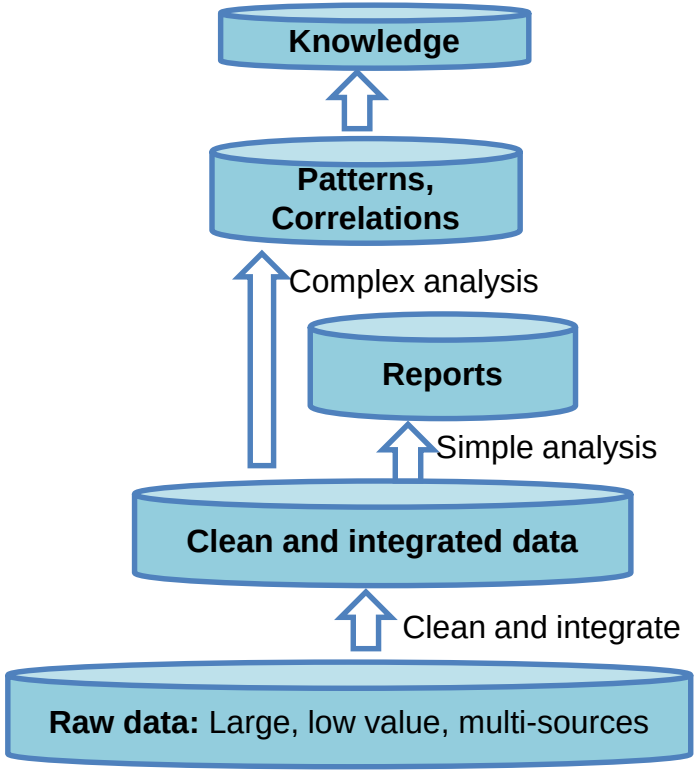
城市活动数据 Context Data

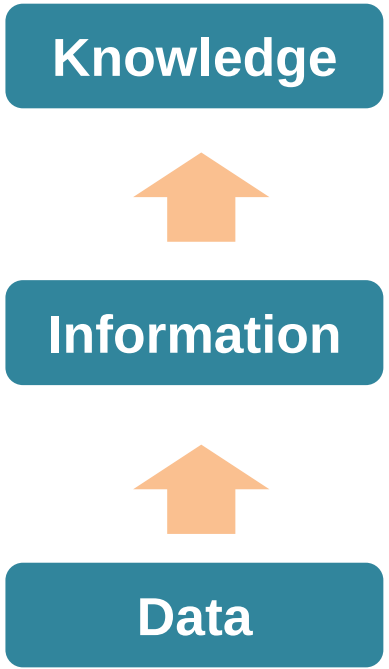
交通调查数据 Census Data

中国经济普查 CHINA ECONOMIC CENSUS

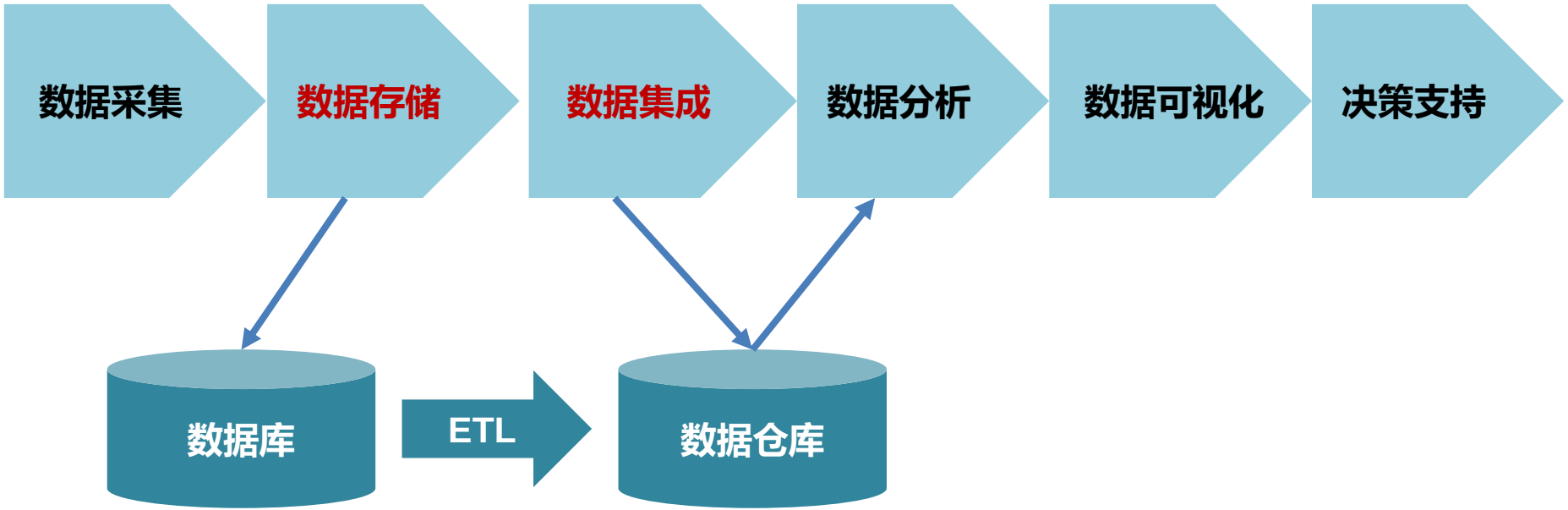
Transportation Economic Census

Data, Information & Knowledge





数据处理流程



ETL: 抽取(Extract)、**转换**(Transform)、**加载**(Load)

数据库（Database, DB）

- **什么是数据库?**
  - 一组相互有关联的数据集合
  - 长期储存在计算机中，有组织、可管理和可共享

- **数据库的基本特征**
  - 数据按一定的数据模型组织、描述和储存
  - 支持数据的增、删、改、查
  - 支持并发查询处理

数据库系统（Database System）

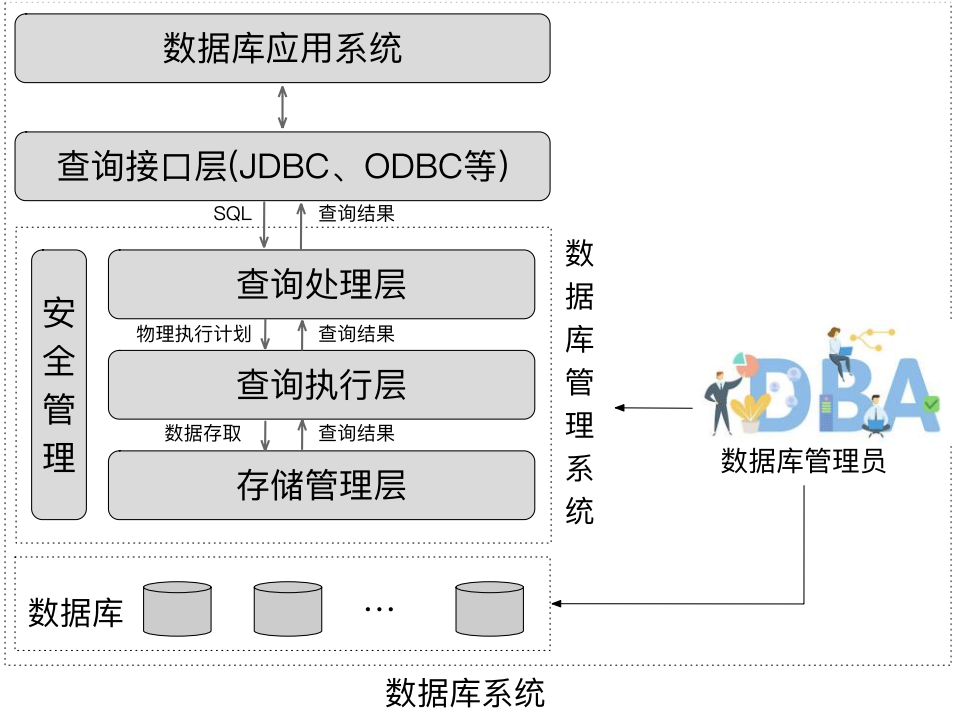
- **什么是数据库系统?**
  - 数据库系统是指由数据库管理系统和相关工具组成的软件系统，用于管理和操作大量数据
  - 一般包括
    - 数据库
    - 数据库管理系统
    - 开发工具、应用系统
    - 数据库管理员和终端用户

- **什么是数据库管理系统?**
  - 管理数据库的软件

数据库管理系统(DB Management System, DBMS)

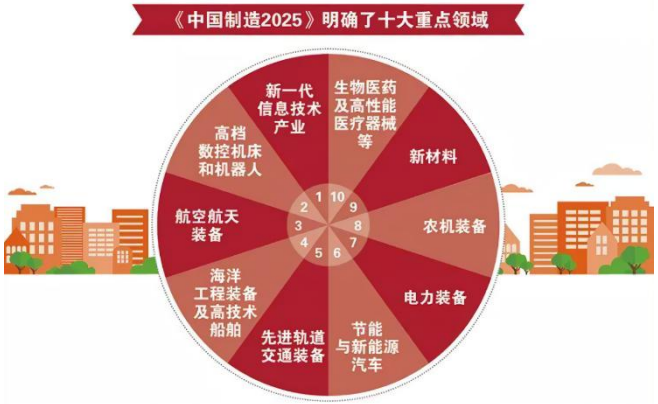
同济大学 TONGJI UNIVERSITY

- **什么是数据库管理系统?**
  - **定义1**：用户（应用程序）与操作系统之间的数据库管理软件
  - **定义2**：一个管理数据的大型复杂基础软件系统
- **数据库管理系统的用途**
  - 优雅查询和数据抽象
  - 高效组织和存储数据
  - 正确一致的并发更新
  - 低时延高吞吐的查询
  - 并行高效的有序执行
  - 可用性和高可靠保证
  - 安全可信的统一控制
  - 方便易用的用户接口



为什么要学习数据库?

- **Need for DB has exploded in the past years**
  - Customer Relationship Mgmt (CRM), Supply Chain Mgmt, Enterprise Resource Planning (ERP), Business intelligence(BI), etc.
  - Industry 4.0 (工业4.0), Made in China 2025(中国制造2025)



为什么要学习数据库?

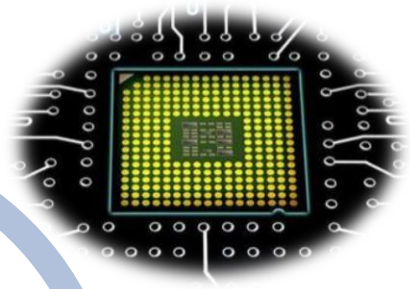
- **Need for DB has exploded in the past years**
  - Internet of things (IoT), Edge Computing, Smart City, Smart Ocean, Autopilot, etc.

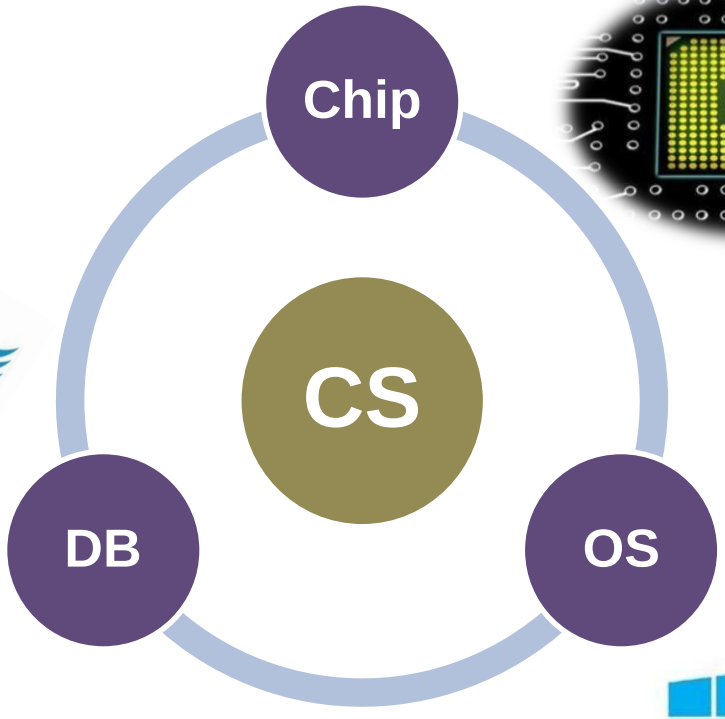
物联网

智慧城市

自动驾驶

为什么要学习数据库?





OceanBase
A Powerful Database System

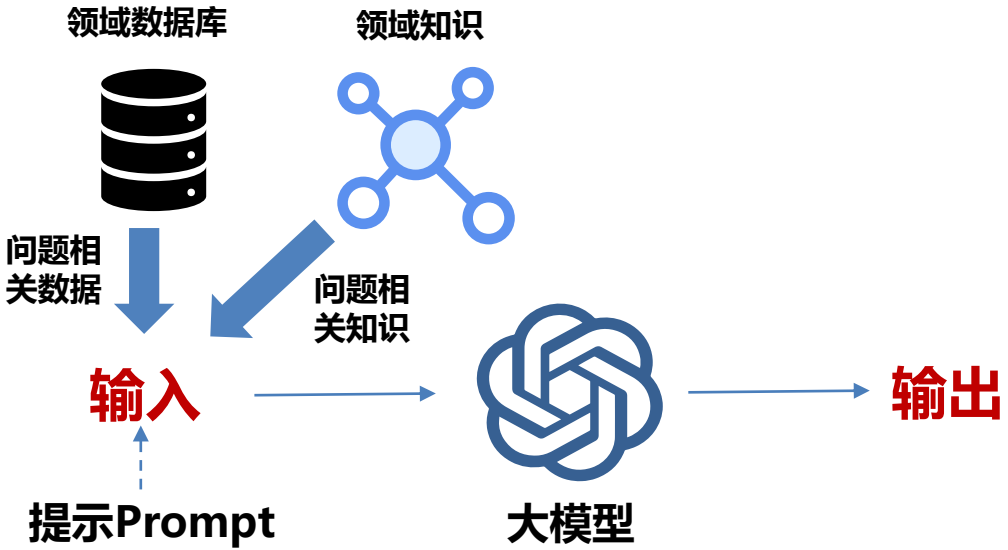
openGauss

Rank			DBMS
Feb 2026	Jan 2026	Feb 2025	
1.	1.	1.	Oracle
2.	2.	2.	MySQL
3.	3.	3.	Microsoft SQL Server
4.	4.	4.	PostgreSQL
5.	5.	5.	MongoDB
6.	6.	↑ 7.	Snowflake
7.	7.	↓ 6.	Redis
8.	8.	↑ 13.	Databricks
9.	9.	9.	IBM Db2
10.	10.	↓ 8.	Elasticsearch

Harmony OS

Windows Mac Linux UNIX Unix

为什么要学习数据库?



检索增强生成（RAG）：检索效率、检索质量

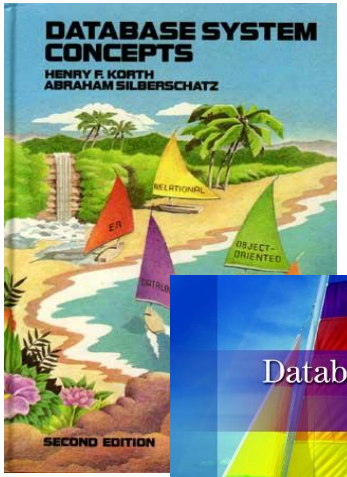
课程目标

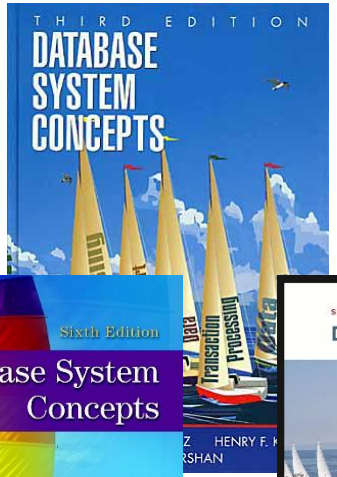
- **目标一**
  - 学习掌握数据库设计、存储管理、查询处理与优化、事务管理等基础知识
  - 针对特定数据管理需求，具备设计和开发数据库解决方案的能力
  - 了解数据库的前沿发展动态和相关先进技术

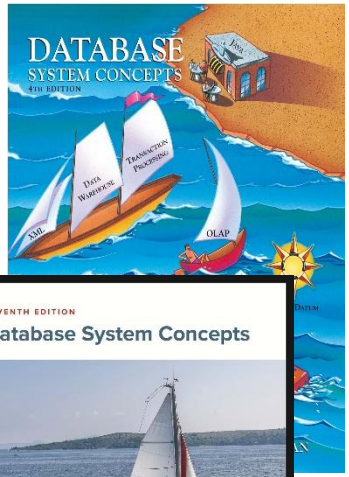
- **目标二**
  - 通过小组合作，提高团队协助能力
  - 通过课程实验和大作业，提高个人实践能力
  - 通过前沿探索，培养科研素养

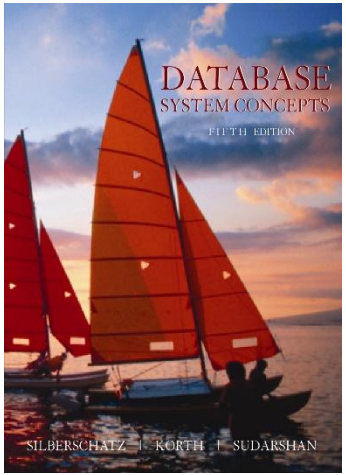
教材

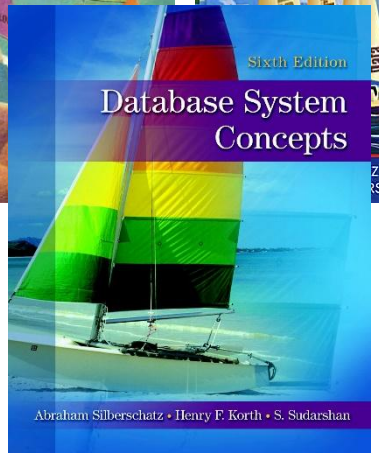
- Abraham Silberschatz(耶鲁), Henry F. Korth(理海大学), and S. Sudarshan (印度理工学院), **Database System Concepts**

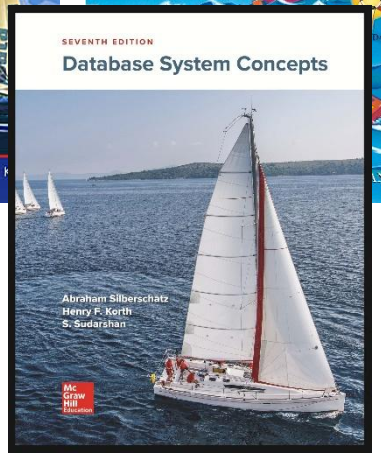








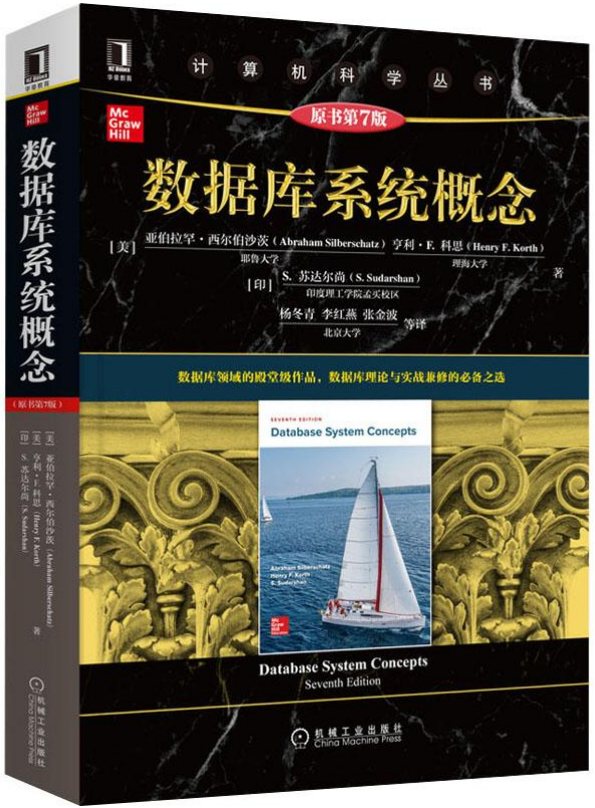




2010 **2019**

教材

- 中文版（**原书第7版**），北京大学杨冬青等译, 2021年



课程概要

- **Part 0: Overview**
  - Ch1: Introduction
- **Part 1 Relational Languages**
  - Ch2: Relational model
  - Ch3&4: SQL
  - Ch5: Advanced SQL
- **Part 2 Database Design**
  - Ch6: Database design via E-R model
  - Ch7: Relational database design
- **Part 3 Application Design & Development**
  - Ch8: Complex data types
  - Ch9: Application development
- **Part 4 Big Data Analytics**
  - Ch10: Big data
  - Ch11: Data analytics
- **Part 5 Storage Management & Indexing**
  - Ch12: Physical storage systems
  - Ch13: Data storage structures
  - Ch14: Indexing
- **Part 6 Query Processing & Optimization**
  - Ch15: Query processing
  - Ch16: Query optimization
- **Part 7 Transaction Management**
  - Ch17: Transactions
  - Ch18: Concurrency control
  - Ch19: Recovery system
- **Part 8 Parallel & Distributed Database**
  - Ch20: Database system architecture
  - Ch21-23: Parallel & distributed storage, query processing & transaction processing
- **Advanced topics**
  - OceanBase, MongoDB, Neo4J
  - RAG, Multimodal retrieval, …

课程考核

- **出勤和课堂练习**：10%（三次以上无故缺席按0分计算）
- **作业、实验和前沿调研**：20%
  - 课程作业
  - 课程实验
  - 数据库前沿调研报告与分享 (蚂蚁OceanBase, 华为OpenGauss等)
- **课程大作业**：20%
  - 数据库设计大作业报告
- **期末考试**：50%
- **课程相关科研、竞赛**：额外加分，上限10%

开展小组专题研究与讨论

同济大学 TONGJI UNIVERSITY

- **研究对象：蚂蚁科技OceanBase或华为OpenGauss**
  - **存储：** Traditional + Cloud
  - **索引：** Traditional + Spatial
  - **查询：** Traditional + Complex
  - **优化：** Different levels, strategies, techniques
  - **事务：** 事务处理、并发、恢复
  - **分布式DB:** 分布式存储、查询处理、事务

- **Advanced topics**
  - NoSQL数据库, HTAP数据库, 内存数据库，云原生数据库，新硬件数据库
  - BlockChain, AI4DB, DB4AI, 跨模态检索
  - …

课程设计（2026年秋）考核方式

同济大学 TONGJI UNIVERSITY

- **系统开发 (70%)**
  - System (50%): an application system, and a system report
  - Oral presentation (20%)：mid term(10%) + final(10%)
- **前沿报告 (20%)**
  - Case studies for 蚂蚁OceanBase, 华为OpenGauss，PostgreSQL, MySQL, NoSQL, etc.
- **考勤 (10%)**
- **科学相关科研、竞赛**
  - 额外加分，上限10%

其他参考资料

- **数据库国际学术会议**
  - SIGMOD/PODS, VLDB, ICDE
  - CIKM, ICDT, EDBT, ER, DASFAA, SSTD, etc.
- **数据库相关学术期刊**
  - ACM Trans. on Database Systems (TODS)
  - IEEE Trans. on Knowledge and Data Engineering (TKDE)
  - VLDB Journal
  - Data and Knowledge Engineering (DKE)
- **其他网上资源**
  - DBLP: http://dblp.uni-trier.de/
  - Google Scholar, Citeseer, etc.

- **Wechat (微信)关注**
  - 数据分析精选
  - 大数据
  - 大数据文摘
  - 互联网分析沙龙
  - 网络数据大全
  - 数据分析
  - 战略前沿技术
  - 大数据魔方
  - ……

课程群

课程QQ群：教辅答疑、发布通知、交流分享

课程要求

- 因故请假尽量提前一天联系助教或授课老师，突发情况除外
- 按时在Canvas上提交作业和相关报告，超期原则上不再接收
- 课程作业、实验报告和大作业报告等允许使用大模型，不过需注意内容的**正确性**和**逻辑连贯性**，以及可能的**重复情况**（两份文件重复超过30%视为抄袭，按0分计算）

课程问卷