



2026年春季学期

1

数据库系统概论

AN INTRODUCTION TO DATABASE SYSTEMS

中国科学技术大学
人工智能与数据科学学院

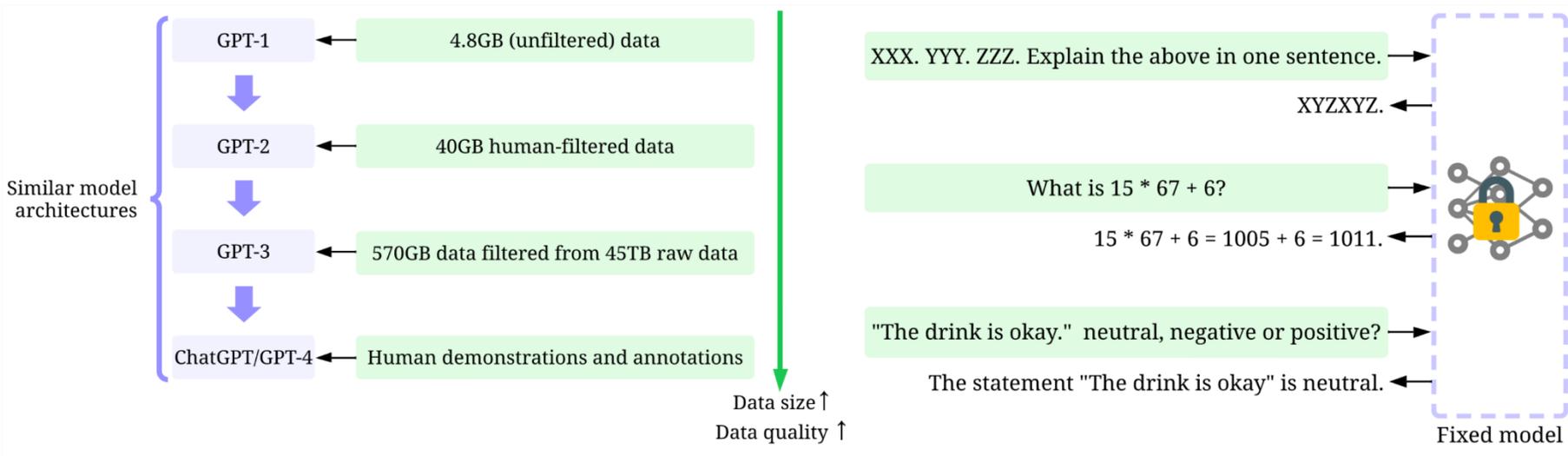
黄振亚, huangzhy@ustc.edu.cn



数据科学基础

2

- 人工智能逐渐从以模型为中心过渡到以数据为中心
 - GPT成功的数据基石**：GPT进化中，模型结构保持相似，训练数据的规模、质量得到极大提升
 - 数据导向的模型应用**：当模型足够强大，仅仅需要修改推理数据（提示工程）便可完成目标任务

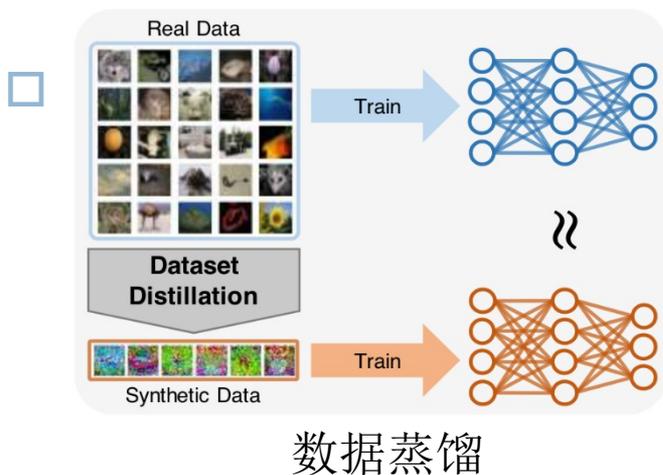
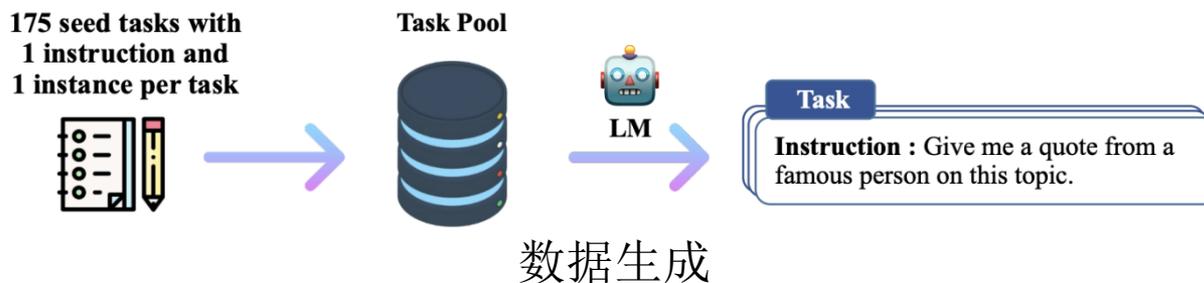




数据科学基础

3

- 以数据为中心(data-centric)的数据科学技术
 - 增加数据数量



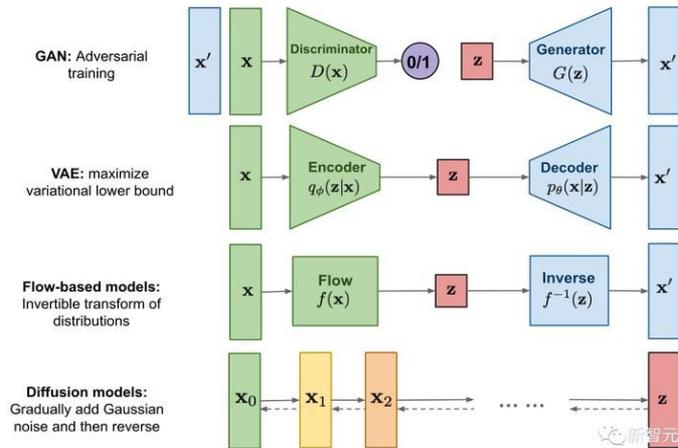


数据科学基础

4

大数据催生人工智能新浪潮—扩散模型-2022

- 任务：AI图像生成
- 应用数据集：LAION-5B
 - 80TB量级
 - 58.5亿个图像-文本对



- 图像数据集规模变化：
 - Cifar-10: 6万张
 - ImageNet: 1400万张
 - LAION-5B: 58.5亿张



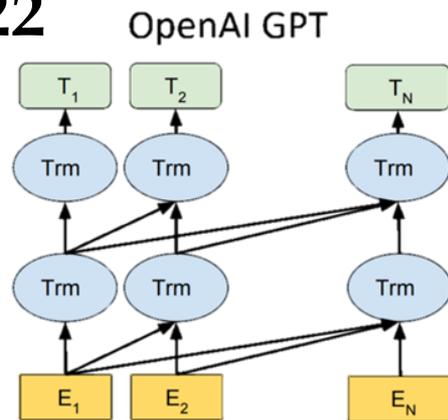


数据科学基础

5

大数据催生人工智能新浪潮- ChatGPT-2022

- 任务：文本对话
- 数据量：5GB增加到45TB
 - 96%以上是英文，其它20个语种不到4%
- 参数量：1.17亿增加到1750亿
- 文本数据规模变化：



GPT

无监督预训练，有监督微调

5G文本数据 | 1.17亿模型参数

在9/12任务上最优，包括问答、语义相似度、文本分类

2018

GPT-2

多任务、零样本学习 (zero-shot)

40G文本数据 | 15亿模型参数

在7/8任务上最优，包括阅读理解、翻译、问答

2019

GPT-3

小样本学习 (few-shot)

45T文本数据 | 1750亿模型参数

在阅读理解任务上超越当时所有zero-shot模型

2020



数据科学基础

- 大数据催生人工智能新浪潮- Deepseek
 - 使用超60万条COT数据进行训练
- 使用COT数据集的优势：
 - 复杂任务处理：擅长解决需要多步逻辑推导的难题，推理能力大幅度提升。
 - 可解释性增强：将推理过程透明化，每一步决策有据可查，提升可信度。

共35头，94脚，求鸡兔数量？

已思考 (用时 3 秒)

• 我们被问到：“共35头，94脚，求鸡兔数量？”这是一个经典的鸡兔同笼问题。设鸡有 x 只，兔有 y 只。那么 $x+y=35$ ， $2x+4y=94$ 。解这个方程组。从第一个方程， $x=35-y$ ，代入第二个： $2(35-y)+4y=94 \Rightarrow 70-2y+4y=94 \Rightarrow 70+2y=94 \Rightarrow 2y=24 \Rightarrow y=12$ ，则 $x=23$ 。所以鸡23只，兔12只。答案：鸡23只，兔12只。

鸡有23只，兔有12只。





数据科学基础

□ 大数据催生人工智能新浪潮—openclaw-2026

□ 功能：多任务自动化，全能助手

□ 多模态的数据格式：

- 文本
- 图像
- 语音

□ 智能体的协同工作：

- 通过MCP增加tool
- 根据用户习惯归纳skill
- 不同智能体间相互配合



OpenClaw 常用指令

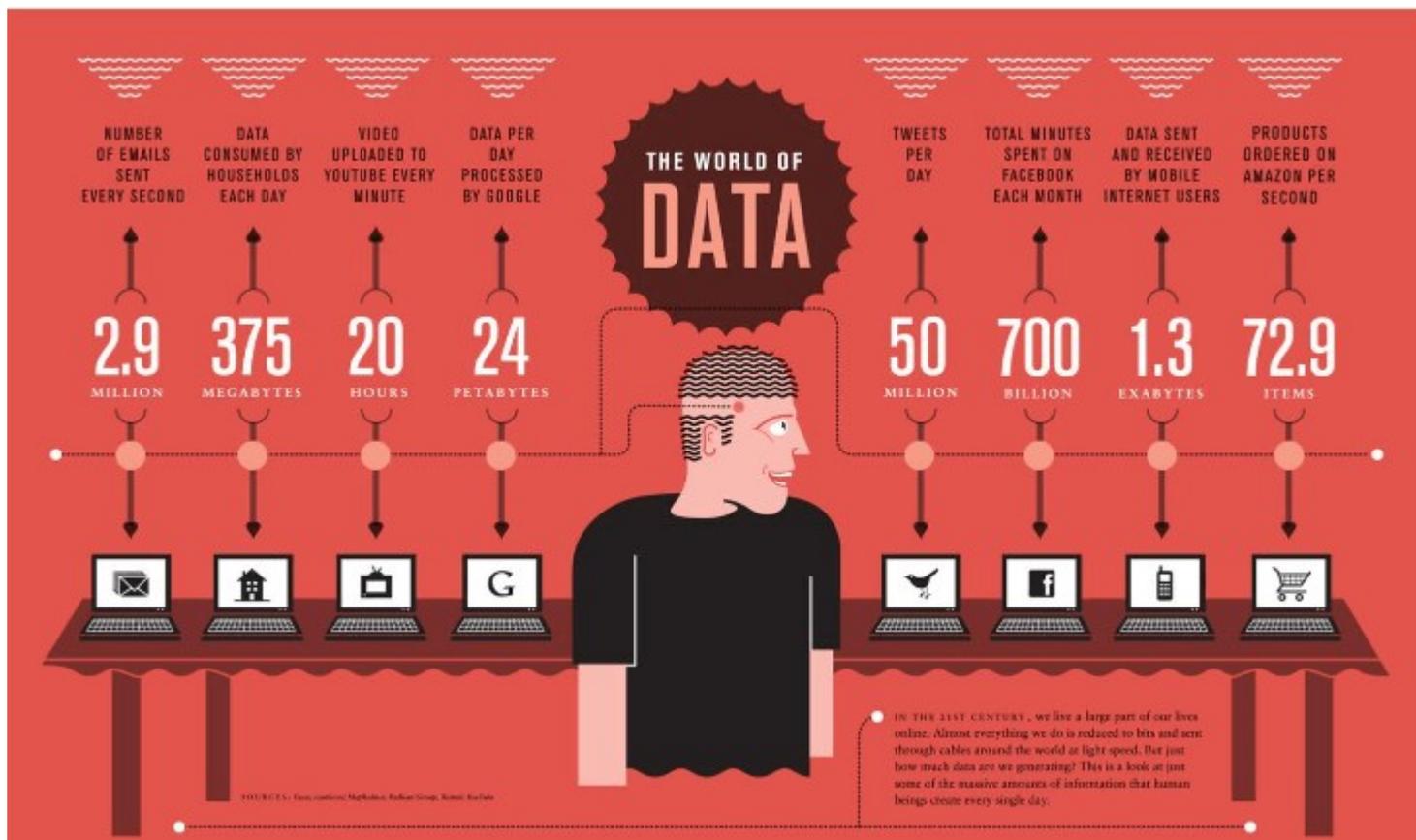
掌握这些命令,从新手到高手

/status 查看状态 ★★★★★ 监控消耗	/model 切换模型 ★★★★★ 成本优化	/compact 压缩上下文 ★★★★★ 长对话优化
/exec 执行控制 ★★★★ 安全设置	/approve 批准操作 ★★★★ 权限管理	/tts 语音控制 ★★★ 语音合成



大数据时代

- 我们生活在数据中，所有人都在制造和分享数据





数据科学基础

9

□ 大数据概念的提出



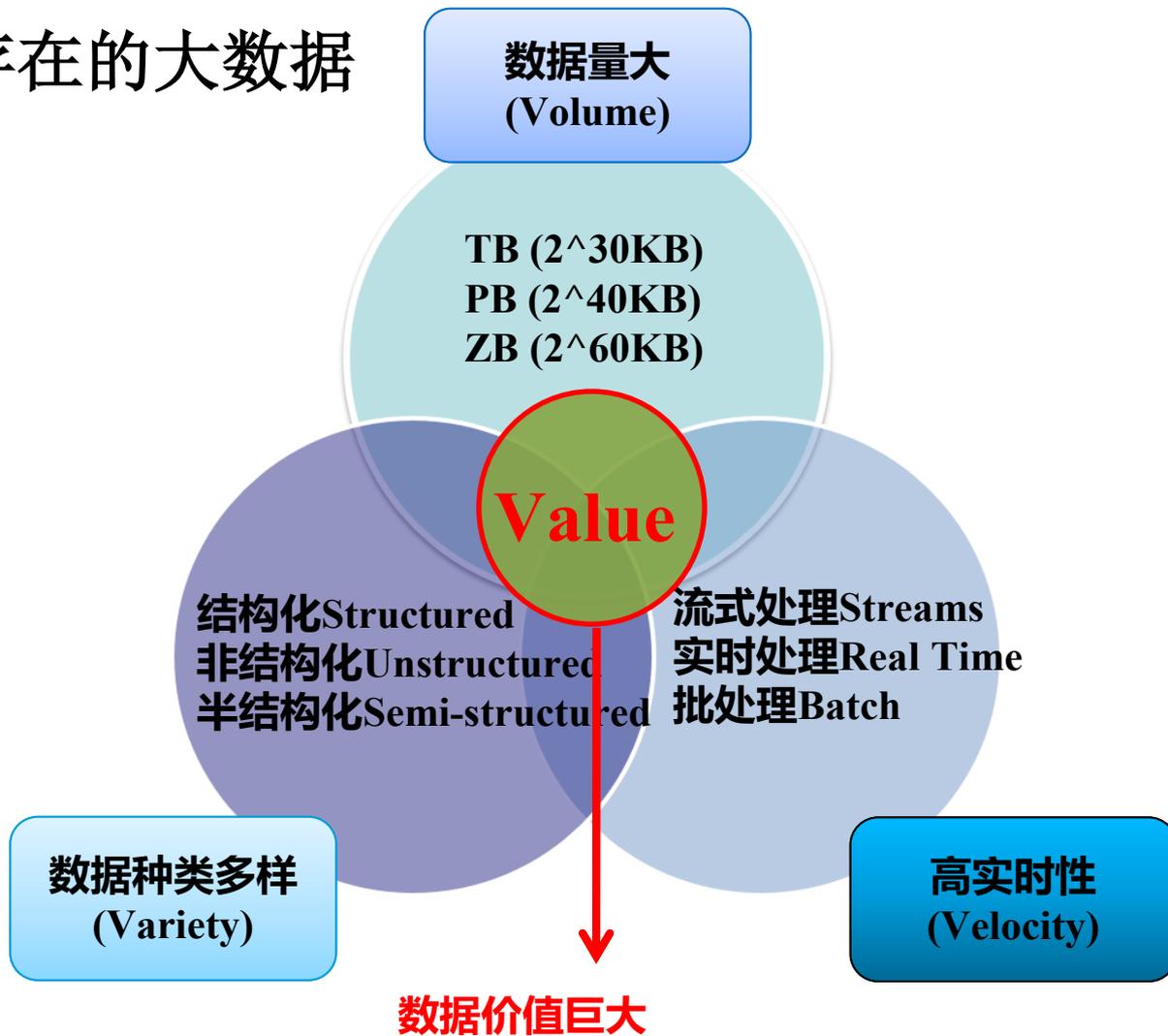
从2008年9月,《Nature》杂志首次出版一期大数据专刊,科学家们提出“大数据真正重要的是新用途和新见解,而非数据本身”



大数据时代

10

客观存在的大数据





大数据时代

11

□ 客观存在的大数据---Volume(数据量巨大)

阿里所保有的、经过清洗的历史数据已超过**100PB**。

——阿里数据仓库负责人七公（汪海）

百度现在的**数据规模**已经到了**EB级**，每天处理的数据量到了上百PB。

——百度大数据部总监薛正华

全球数据总量在2020年达到**60ZB**，预计2025年达到175ZB，我国数据占18%。

——IDC互联网数据中心

$$1 \text{ ZB} = 2^{10} \text{ EB} = 2^{20} \text{ PB} = 2^{30} \text{ TB} = 2^{40} \text{ GB}$$

- 1 ZB = 地球上沙粒的总量，1 EB = 4000个美国国会图书馆的藏书



大数据时代

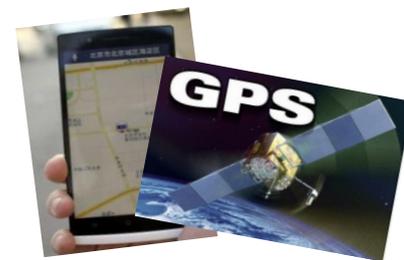
客观存在的大数据--- Variety(数据类型多)

数据形式的多样:

- 结构化数据, 半结构化数据, 非结构化数据
- 关系数据库数据、xml/JASON文档、音视频数据

数据来源的多样性:

- 不同的IT应用系统
- 各种设备 (手机、手环)
- 互联网、物联网
- 其它



时空数据



文本数据



图像数据



事务数据



视频数据



音频数据



大数据时代

客观存在的大数据--- Velocity(高实时性)

1秒定律: 对于大数据应用而言, 必须要在1秒钟内形成答案, 否则这些结果可能就是过时的、没有意义的

在百度输入关键字:
“汽车维修”、“挖掘机 学习”

某在线电影网站



某IT业界资讯网站



例如, 用户在合肥某台PC上, 打开百度输入关键字片刻之后, 再打开其它网站, 就会看到相关的广告, 并且所推荐的是地理位置信息相关的 (合肥、安徽)

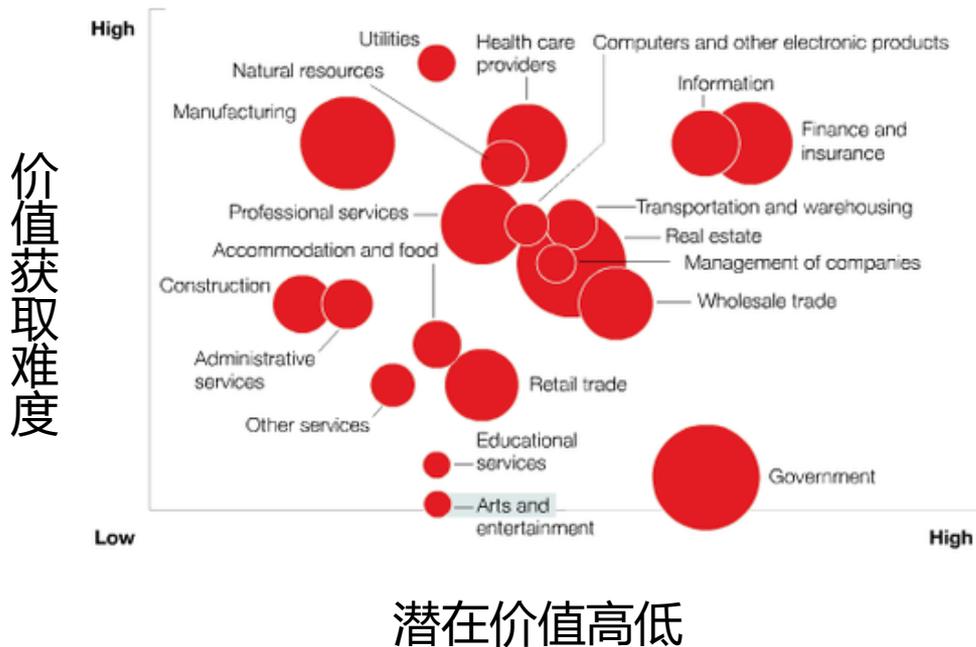


大数据时代

客观存在的大数据 --- Value (价值巨大但价值密度低)

挖掘大数据中的价值类似沙里淘金，需要从海量数据中挖掘稀疏但珍贵的信息

所有产业都可以应用大数据产生价值



● 各产业GDP占比 (以美国经济为例)

图：麦肯锡对各个行业从大数据中获得价值难易程度的分析 (2011年)



大数据蕴含巨大的价值

15

□ 大数据改变了人们的衣食住行

过去



现在



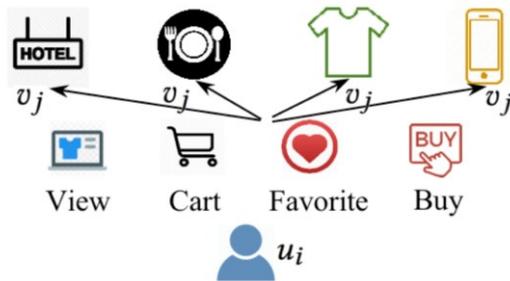


大数据蕴含巨大的价值

大数据促进了个性化服务



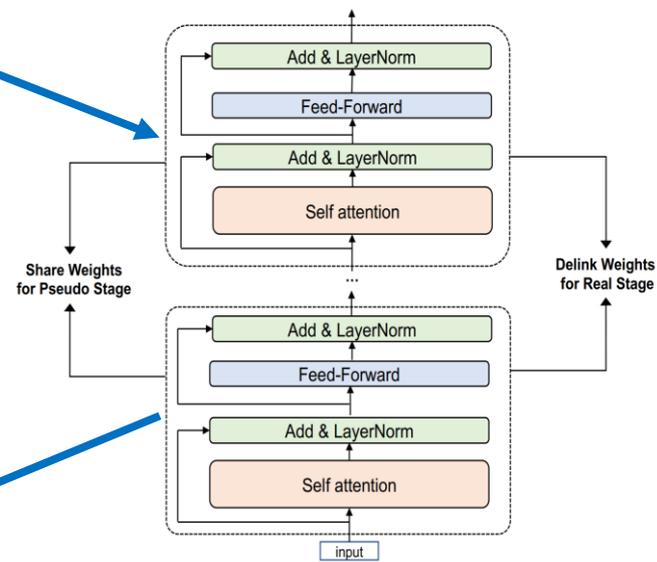
电商平台



海量用户多样交互行为



促进用户消费、提升平台收益



达摩院10万亿参数 M6-10T模型

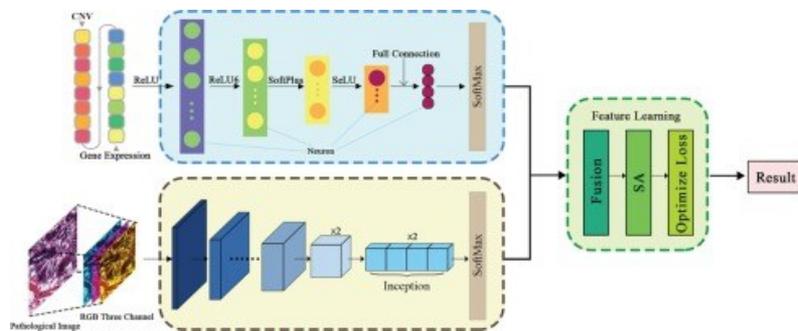
大数据蕴含巨大的价值

17

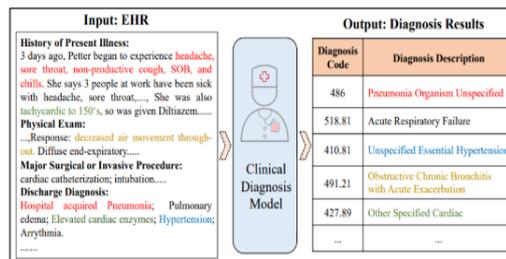
大数据辅助智慧医疗



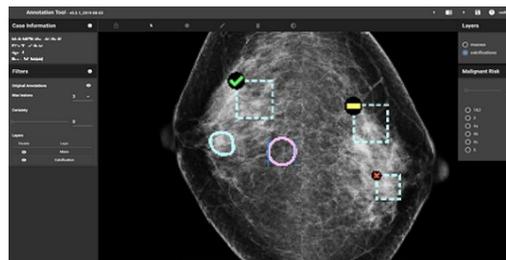
AI电子病历



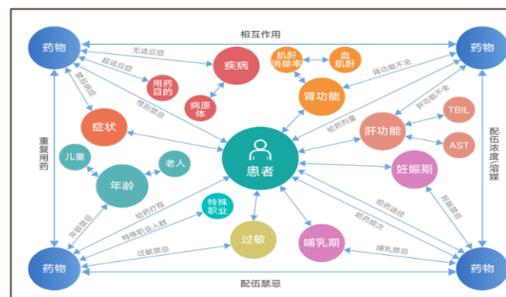
多模态医疗数据挖掘模型



疾病诊断



病灶识别



药物推荐



大数据蕴含巨大的价值

18

□ ChatGPT

- 参数量从1.17亿增加到1750亿
- 预训练数据量从5GB增加到45TB
 - 96%以上是英文，其它20个语种不到4%
 - 标注外包员工来自美国和东南亚

Introducing ChatGPT

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

FO

鸡蛋有120个，鸭蛋比鸡蛋多(1/6)，请问两种蛋共有几个？



鸭蛋一共有 $120 * (1/6) = 20$ 个。

所以鸡蛋和鸭蛋一共有 $120 + 20 = 140$ 个。



大数据蕴含巨大的价值

19

GPT

无监督预训练，有监督微调

5G文本数据 | 1.17亿模型参数

在9/12任务上最优，包括问答、语义相似度、文本分类

2018

GPT-2

多任务、零样本学习 (zero-shot)

40G文本数据 | 15亿模型参数

在7/8任务上最优，包括阅读理解、翻译、问答

2019

GPT-3

小样本学习 (few-shot)

45T文本数据 | 1750亿模型参数

在阅读理解任务上超越当时所有 zero-shot模型

2020

GPT-4o

多模态，可处理图像和文本输入

GPT-4的升级版模型，其中“O”是Omni的缩写，意为“全能”。其在响应速度、多模态能力、实时交互性方面较GPT-4能力有极大的提升

2024.5

GPT-4

多模态，可处理图像和文本输入

在大多数专业和学术考试中表现出人类水平，且能通过律师资格考试，排名考生中前10%，相较之下 GPT-3.5排名低于后10%

2023.3

ChatGPT (3.5)

基于InstructGPT进行优化

能生成更翔实的回复：标注数据质量更高
更擅长连续对话：源于标注人员标注的多轮对话数据

2022.11

捕获人类意图进一步优化

大规模预训练模型



Deepseek爆火出圈

评价一下“数据库系统概论”这门课



关于“数据库系统概论”这门课，网上有一个流传很广的评价：“这是大学四年里，最接近‘程序员’感觉的一门课。”

这门课通常被视为计算机专业由“语法学习”迈向“工程实践”的桥梁。它不教你如何写循环和判断，而是教你如何管理和组织核心的数据资产。

以下是从课程价值、学习挑战和后续影响三个维度的深度评价：

1. 核心价值：从“写代码”到“做系统”的认知跃迁

这门课最大的价值在于帮你建立“数据思维”。在学这门课之前，你可能更关注程序的控制流（if-else, for 循环），学完之后，你会开始关注数据流。

- **理论的高光点：**课程中讲的“范式理论”，是你第一次系统性地理解“为什么这么设计表才是好的”。当你用这个理论去审视师兄留下的混乱数据库，并成功将它拆分成几个清晰、低冗余的表时，那种茅塞顿开的感觉，是这门课独有的魅力。

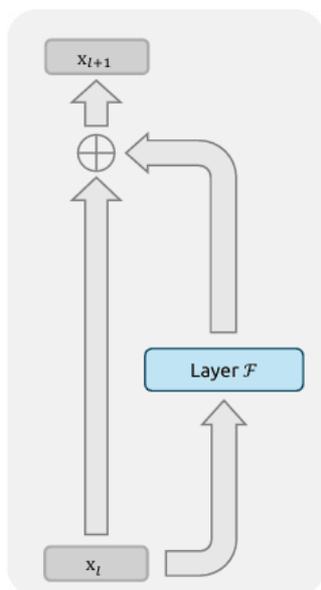




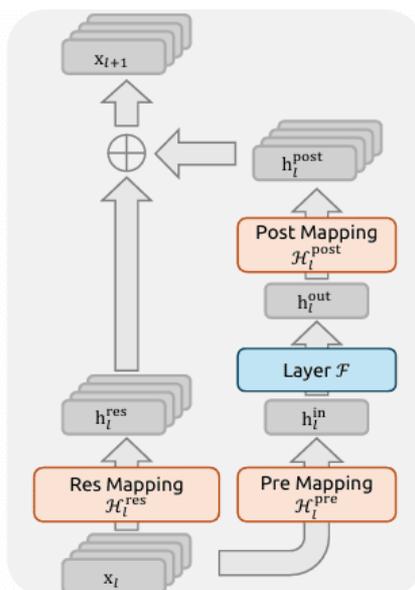
Deepseek-V4

21

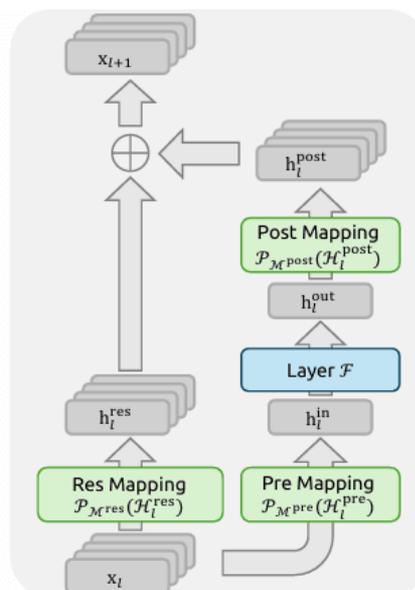
- DeepseekV4即将在一周后推出（2026. 3）
- DeepseekR1（文本模态） \longrightarrow DeepseekV4（多模态）
- 在底层实现了视觉、代码、文本的统一向量表征
- 使用mHC架构，增强长上下文能力



(a) Residual Connection



(b) Hyper-Connections (HC)



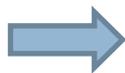
(c) Manifold-Constrained HC (mHC)



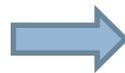
数据存储：大模型与数据库

大模型与数据库

数据库
(1970s)



大数据分析
(2010s)



智能大模型
(2020s)

数据库系统

```

1 Create Table DemoSQLTable (
2   id int,
3   myGETDATE smalldatetime default GETDATE(),
4   myCurrentTimeStamp datetime default CURRENT_TIMESTAMP,
5   mySYSDATETIME datetime2 default SYSDATETIME()
6 );
7 GO
8
9 insert into DemoSQLTable (ID) values (1);
10 GO
11
12 select * from DemoSQLTable;
13

```

id	myGETDATE	myCurrentTimeStamp	mySYSDATETIME
1	2021-12-25 03:25:00	2021-12-25 03:24:48.920	2021-12-25 03:24:48.9233333

SQL

互联网

搜索引擎

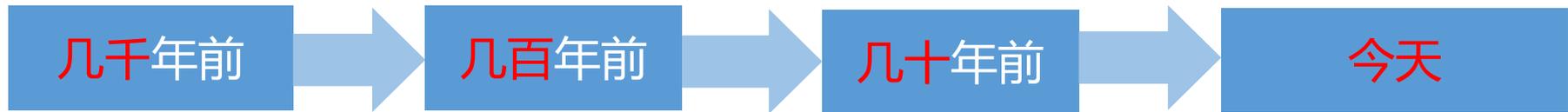
大模型

生成试问答



大数据蕴含巨大的价值

2007年，Jim Gray总结出了四个科学范式



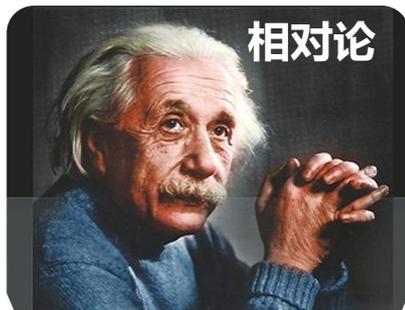
经验科学

- **第一范式**
- 以**归纳法**为主，带有盲目性的观测和实验
- **科学实验**



理论科学

- **第二范式**
- 以**演绎法**为主，关注理论总结和理性概括
- **数学模型**



计算科学

- **第三范式**
- 重视**数据模型构建、定量分析方法**，利用计算机来分析和解决
- **科学计算**



数据密集型科学

- **第四范式**
- 先有了**大量的已知数据**，然后通过计算得出之前未知的理论
- **机器学习**





大数据蕴含巨大的价值

数据分析挖掘技术是解决众多国家重大现实需求问题的共性基础 ---数据驱动的人工智能

社交媒体、人口流动、居住交通数据



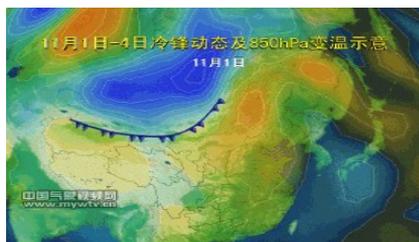
突发事件预测、关键人群监测

医疗、医保、健康、影像等大数据



医疗诊断方案

环境、气象、交通、社会发展等大数据



环境治理

交通流、医疗、商业、环境、劳动力等数据



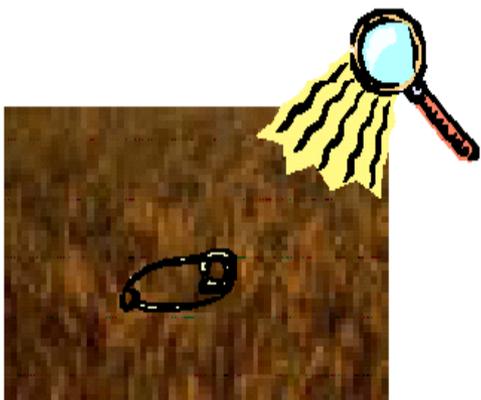
城市智慧管理

数据库是基础中的基础!



数据库是基础中的基础

25



- **数据库技术**是从大量数据里找某个具体数据，或是简单的数据统计信息。
- **数据库技术**做的事如同在草堆里去找别针。



- 数据挖掘技术找的不是一个已存在那里的信息。
- 它做的事如同是要设法搞清楚在草堆里有一根针，会造成什么样的后果。



数据库的位置和作用

数据/Data



结构化



半结构化



非结构化

分布式文件系统
数据库

2. 存储
管理

1. 采集、
预处理

5. 可
视化

4. 分析
挖掘

3. 计算
模式

6. 隐私
和安全

应用/App



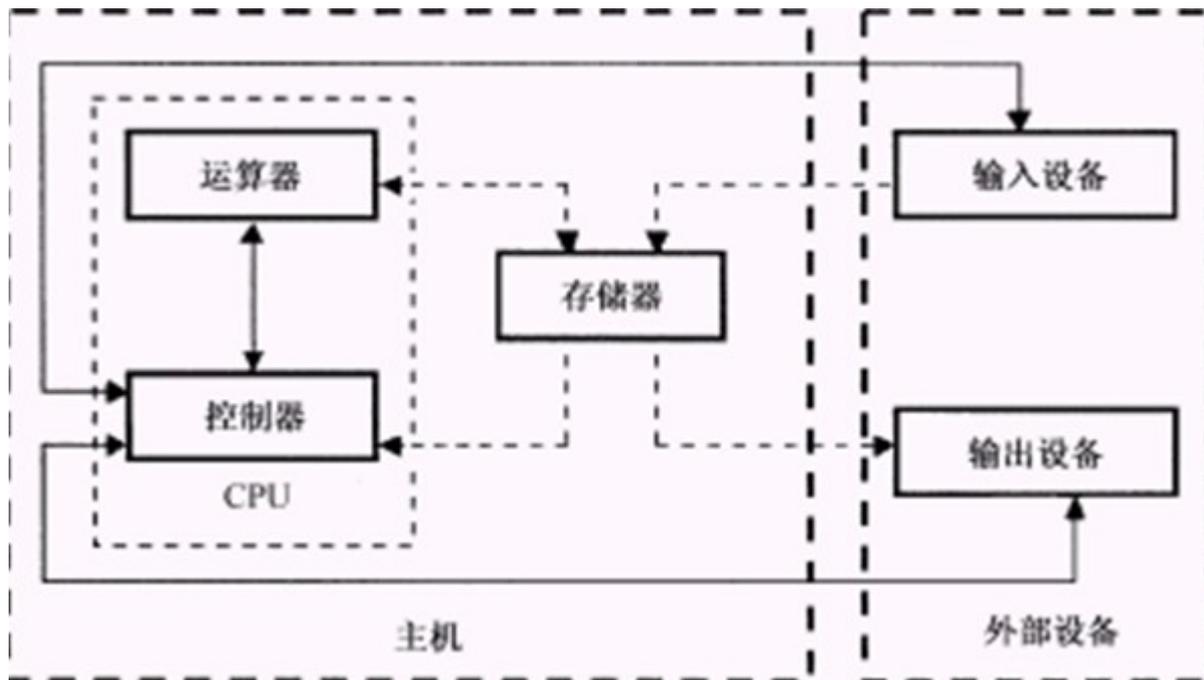
计算系统





数据库的位置和作用

• 计算机组成



• 存储层次

寄存器

缓存

内存

磁盘

• 存储功能

- 存储程序

- 存储数据

存储器





数据库的位置和作用

29





数据库的位置和作用

□ 结构化数据（本课程重点）

□ 可以使用关系型数据库表示和存储的数据，拥有固定结构

□ 半结构化数据（本课程关注）

□ 弱结构化，虽然不符合关系型数据模型的要求，但是含有相关的标记(自描述结构)，分割实体及期属性。如：XML，JSON

□ 非结构化数据

□ 没有固定数据结构，或很难发现统一数据结构的数据

□ 例如：文档，文本，图片，视频，音频等

姓 名	年	性 别
小明	12	男
小白	13	女
小奇	18	男

```
<province>  
  <name>黑龙江</name>  
  <cities>  
    <city>哈尔滨</city>  
    <city>大庆</city>  
  </cities>  
</province>
```





结构化数据与半结构化数据

学号	姓名	年龄	性别	专业
001	张三	20	男	计算机
002	李四	19	女	管理
003	王五	21	男	数学
...

```
{  
  "name": "中国",  
  "province": [{  
    "name": "黑龙江",  
    "cities": {  
      "city": ["哈尔滨", "大庆"]  
    }  
  }],  
  {  
    "name": "广东",  
    "cities": {  
      "city": ["广州", "深圳", "珠海"]  
    }  
  }  
},  
.....  
}]
```

对象，成员：键值对



数据库领域的顶级会议

□ 数据库顶级会议介绍：VLDB、SIGMOD、ICDE

中国计算机学会推荐国际学术会议 (数据库, 数据挖掘与内容检索)

一、A类

序号	会议简称	会议全称	出版社	网址	
1	SIGMOD	ACM Conference on Management of Data	ACM	http://www.sigmod.org	1974
2	SIGKDD	ACM Knowledge Discovery and Data Mining	ACM	http://www.acm.org/sigkdd/	1995
3	SIGIR	International Conference on Research an Development in Information Retrieval	ACM	http://www.acm.org/sigir/	1978
4	VLDB	International Conference on Very Large Data Bases	Morgan Kaufmann/ACM	http://www.vldb.org	1975
5	ICDE	IEEE International Conference on Data Engineering	IEEE	http://www.icde.org/	1984



数据库的图灵奖

四次图灵奖

2015年CNCC（合肥）

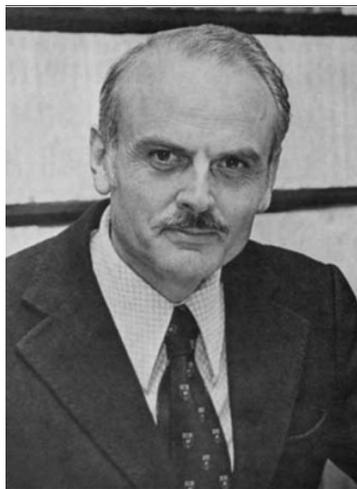
1973年



Charles W. Bachman
查尔斯·巴赫曼

- 网状数据库系统IDS
- 数据库标准的制定

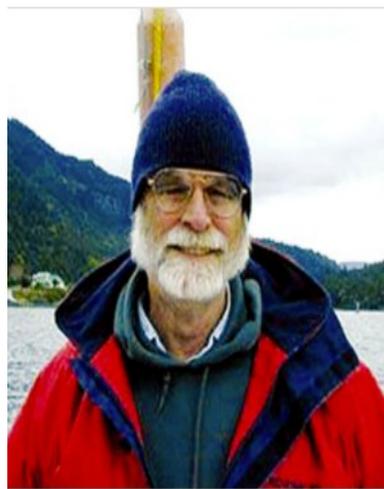
1981年



Edgar F. Codd
埃德加·科德

- 关系型数据模型原型理论

1998年



James Gray
詹姆斯·格雷

- 数据库事务处理原型系统向产业化

2014年



Michael Stonebraker
迈克尔·斯通布雷克

- 关系型数据库
- 新型数据库
- 大数据(库)系统



数据库

□ 数据库有哪些





我国的数据库

- 数据库有哪些
 - 人大金仓KingBase
 - 达梦数据库
 - 神通数据库
 - 阿里OceanBase, PolarDB
 - 腾讯的TDSQL
 - 华为的GaussDB



PRODUCT DISPLAY
神通大型通用数据库



神通数据库管理系统



神通数据库(MPP集群)



神通数据库 (openGauss 版)



神通高可用集群软件

OCEANBASE

用技术让海量数据的管理和使用更简单



PolarDB

TDSQL

腾讯分布式数据库

GaussDB

AI-Native分布式数据库



数据库

□ DB-Engines: <https://db-engines.com/en/ranking> (2026.03.02)

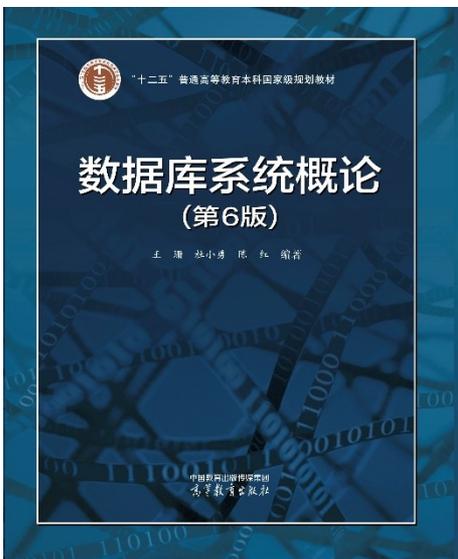
Rank			DBMS	Database Model	Score		
Mar 2026	Feb 2026	Mar 2025			Mar 2026	Feb 2026	Mar 2025
1.	1.	1.	Oracle	Relational, Multi-model ⓘ	1182.46	-21.05	-70.62
2.	2.	2.	MySQL	Relational, Multi-model ⓘ	858.34	-9.88	-129.79
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model ⓘ	711.47	+3.33	-76.67
4.	4.	4.	PostgreSQL	Relational, Multi-model ⓘ	680.08	+8.05	+16.66
5.	5.	5.	MongoDB	Document, Multi-model ⓘ	383.58	+4.85	-12.85
6.	6.	6.	Snowflake	Relational	211.24	+3.10	+49.46
7.	↑8.	↑13.	Databricks	Multi-model ⓘ	145.81	+1.29	+49.80
8.	↓7.	↓7.	Redis	Key-value, Multi-model ⓘ	145.19	-1.85	-10.17
9.	9.	9.	IBM Db2	Relational, Multi-model ⓘ	111.38	+0.16	-15.19
10.	10.	↓8.	Elasticsearch	Multi-model ⓘ	103.58	-2.88	-27.80
11.	11.	11.	Apache Cassandra	Wide column, Multi-model ⓘ	101.88	+0.28	-4.78
12.	12.	↓10.	SQLite	Relational	95.97	-3.22	-17.11
186.	↓182.	↑194.	TDSQL for MySQL	Relational, Multi-model ⓘ	1.07	-0.04	+0.14
253.	253.	↓242.	Kingbase	Relational, Multi-model ⓘ	0.46	-0.02	-0.05
208.	↓204.	↑209.	Alibaba Cloud AnalyticDB for MySQL	Relational, Multi-model ⓘ	0.79	-0.03	+0.03



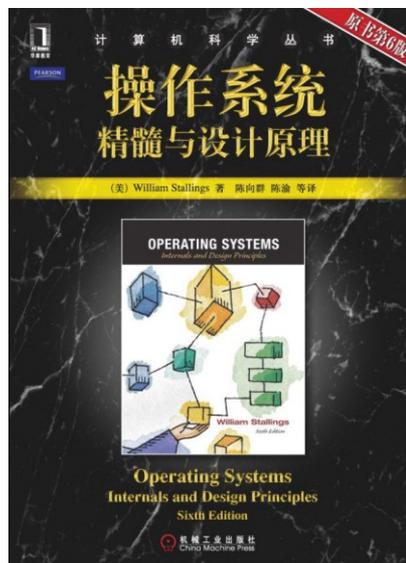
数据库是核心课程之一

37

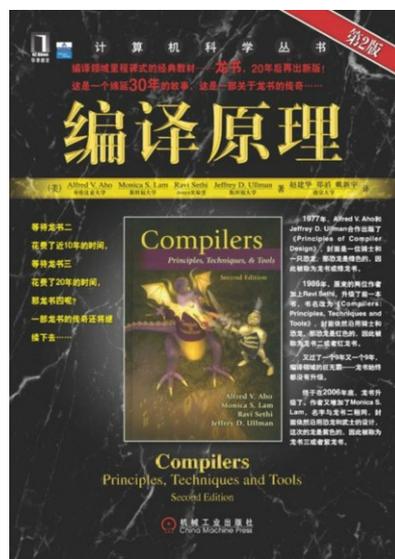
- 了解计算机系统的核心课程



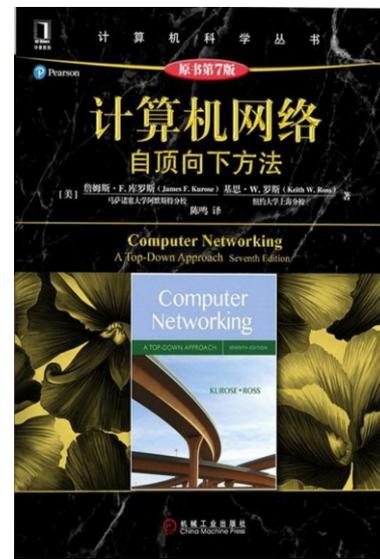
数据库系统



操作系统



编译系统



计算机网络



数据库是核心课程之一

38

• 了解计算机系统的核心课程

数据库系统概论课程评价与分析

最后，总结部分要给出建设性的意见，既指出问题，又给出解决方案，比如建议结合在线课程、实战项目来弥补课程的不足，或者自学新技术扩展视野。

“数据库系统概论”锐评：

——一门让你又爱又恨的“程序员必修课”

如果说计算机专业课程里有一门课能让人从入门到“入土”，从信心爆棚到怀疑人生，《数据库系统概论》绝对能进前三。它像一把双刃剑——学透了是屠龙技，学不会就是“删库跑路”的导火索。以下为无滤镜锐评：

槽点一：理论劝退，学完还是不会设计数据库

• “范式”是永恒的痛：

第一范式、第二范式、BCNF.....学的时候头头是道，一到课程设计就发现：“完全符合范式？这表拆得亲妈都不认识！”实际开发中为了性能往往要冗余字段，但考试时你敢不按范式答题？分分钟挂科警告。

• 锁机制和事务隔离级别：

讲道理，这些概念学完能当场背诵的学生都是勇士，但面对“幻读”“脏读”的灵魂拷问，多少人能真正用代码复现？“MVCC是啥？哦，面试八股文必备。”

槽点二：SQL从入门到“删库跑路”

• 实验课的真实体验：

“请用SQL查询所有成绩大于90的学生”——简单！

“请优化一个跨10张表、嵌套5层的关联查询”——告辞！

更致命的是，实验环境动不动就给你报个“ERROR 1064”，debug两小时发现是少了个逗号。

给 DeepSeek 发送消息

深度思考 (R1)

联网搜索





数据库课程

39

- 预备知识
 - 程序设计语言
 - 数据结构
 - 离散数学
 - 计算机导论
- 同步课程



课程安排

40

□ 课程+实验 (DS2002.01)

- 60学时，课堂学习
- 20学时，实验练习

□ 课堂学习

- 时间：周二 09:45-11:20/周五 15:55-17:30
- 地点：西区 3C104

□ 实验练习、答疑课 (暂定)

- 时间：XXX
- 地点：XXX



课程主页与课程群

课程主页

- <http://staff.ustc.edu.cn/~huangzhy/Course/DB2026.html>
- Linked in: <http://staff.ustc.edu.cn/~huangzhy/>

□ 课程联系方式

- 课程群： 1090416695（请所有选课同学加入）
- 课程邮箱：ustcdb_2026@163.com

□ 助教

- 纪天昀，王朝坤，邢祎航，陆文博
- ustcdb_2026@163.com



教材及参考书(1)

42

教材

□ 《数据库系统概论(第6版)》

王珊, 杜小勇, 陈红;

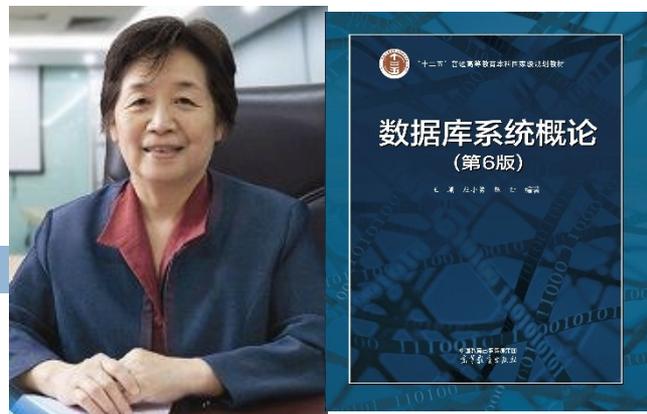
高等教育出版社, 2023.3

□ 《数据库系统基础教程》

岳丽华, 金培权, 万寿红等译

□ 《A First Course in Database Systems》

Jeffrey.D.Ullman, Jennifer Widom, Dept. Of Computer Science, Stanford University





教材及参考书(1)

中国人民大学杜小勇王珊团队成果获国家科学技术进步奖二等奖



1月8日上午
领导人出席会议
信息学院杜
心技术的创新与
科学技术奖项。

该成果由中国人
库管理系统内核研
突破了数据库管理
权19项，出版著作
务、电子党务、国
个重大信息化工程
领域的发展，全面

党和国家

理系统核
获得国家

在国产数据
主研究成果，
项，软件著作
在电子政
行业和六十多
理系统技术



上机软件

44

上机软件

□ MySQL

- MySQL workbench download:

<http://dev.mysql.com/downloads/workbench/>

□ *MongoDB

□ 前端开发

- Python, Java, PHP等, C/S或B/S架构不限

安装和使用方式可以参考课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/DB2026.html>



学习方式

45

听课

(启发式、讨论式)

读书

(预习、复习)

作业与实验

(课后作业, 实验练习, 课程设计)

报告

(实验练习, 课程设计)





课程成绩

46

□ 平时成绩 (30%)

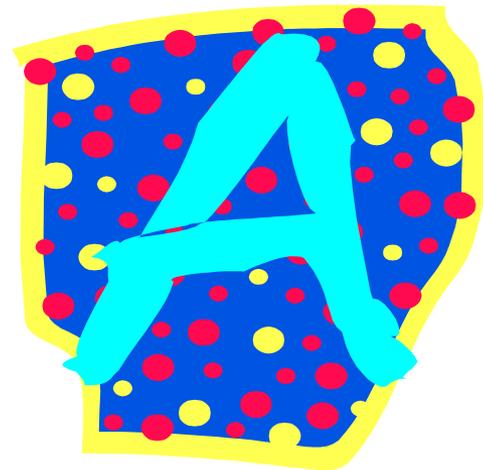
课后作业, 实验练习 + 课堂出勤 (三次不到, 取消考试成绩)

□ 实验成绩 (20%)

课程设计

□ 期末考试 (50%)

卷面成绩



3/6/2026



内容安排(1)

47

基础篇

- 第一章 绪论
- 第二章 关系模型
- 第三章 关系数据库标准语言SQL
- 第四章 数据库安全性
- 第五章 数据库完整性

设计与应用开发篇

- 第六章 关系数据理论
- 第七章 数据库设计
- 第八章 数据库编程



内容安排(2)

48

系统篇

- 第九章 关系数据库存储管理
- 第十章 关系查询处理和查询优化
- 第十一章 数据库恢复技术
- 第十二章 并发控制

新技术篇

- 第十三章 大数据管理
- 第十四章 NoSQL技术



课程目标

49

□ 基础篇

- 了解数据库系统的发展与组成
- 熟练掌握关系型数据库、SQL
- 基本掌握数据库安全性，完整性

□ 设计与应用开发

- 掌握关系型数据原理
- 熟练掌握并应用数据库设计和编程开发

□ 系统篇

- 基本掌握关系数据库优化、数据库恢复和并发

□ 新技术篇

- 基本掌握NoSQL



课程说明

50

- 专业基础课：只想听课堂课程的同学谨慎选择
- 几点说明
 - 平时30% + 课程设计20% + 期末考试50%
 - 作业，实验，大实验，笔试（闭卷）

 - 课堂与平时重要，大家保证出勤
 - 不许叠课

 - 不同专业中的学分认定，需跟相关专业教秘确认



Any Questions?

51

□ 课程QQ群: 1090416695

