



2024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 大数据学院

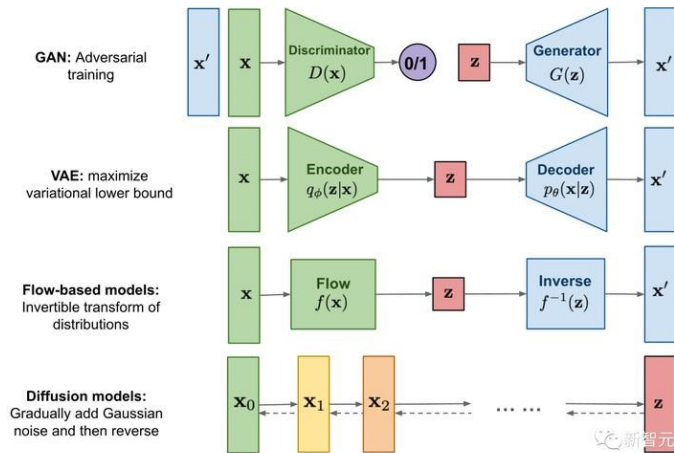
黄振亚, [huangzhy@ustc.edu.cn](mailto:huangzhy@ustc.edu.cn)



# 数据科学基础

## 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



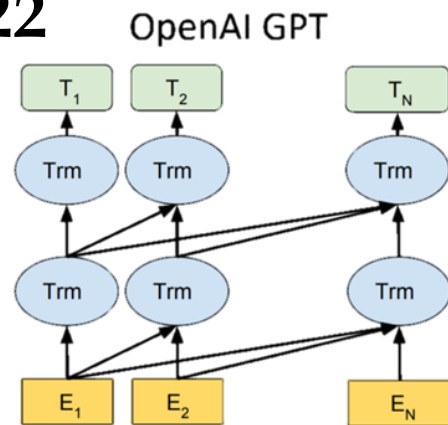


# 数据科学基础

3

## 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

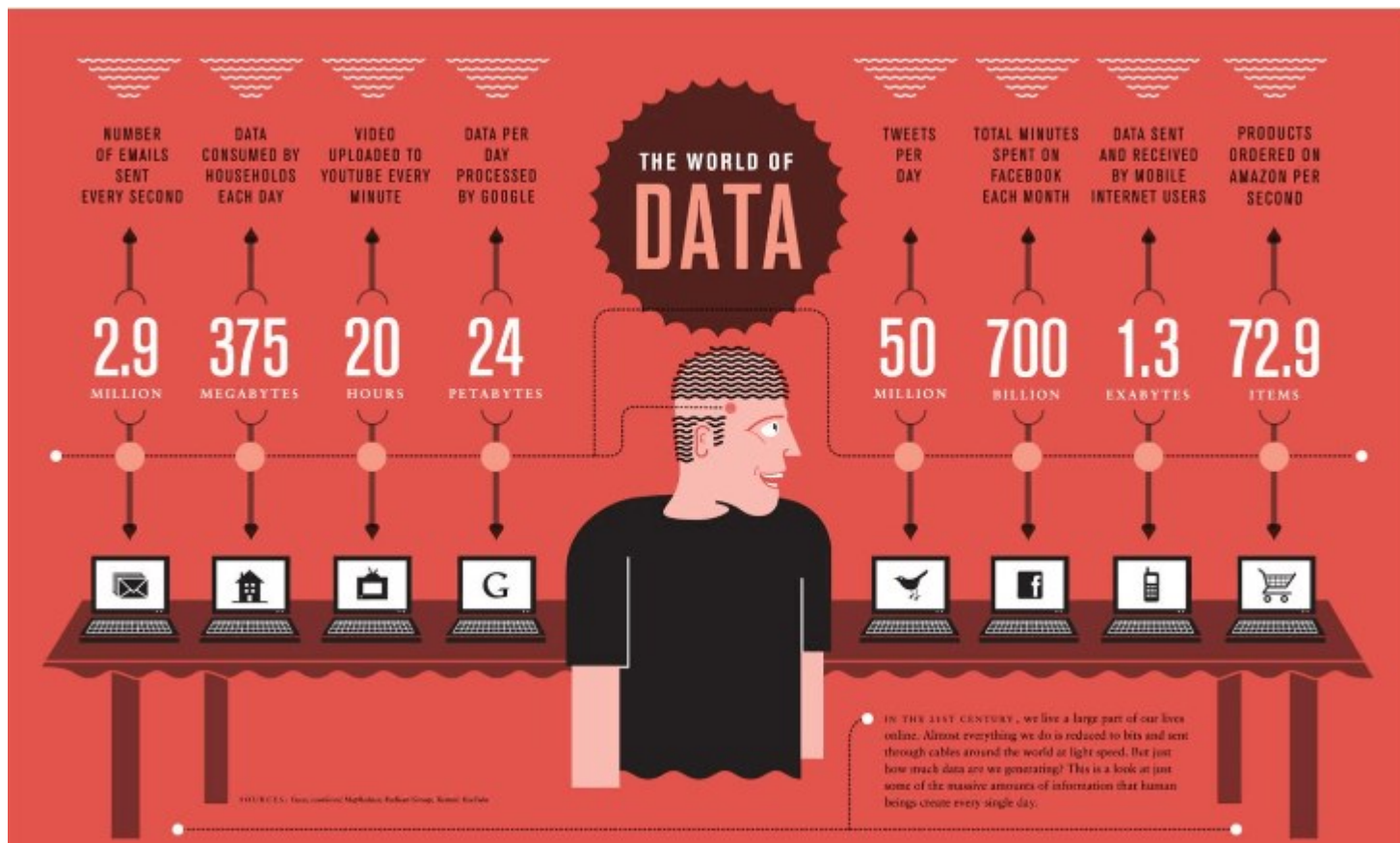
2020



# 大数据时代

4

- 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

5

## □ 大数据概念的提出



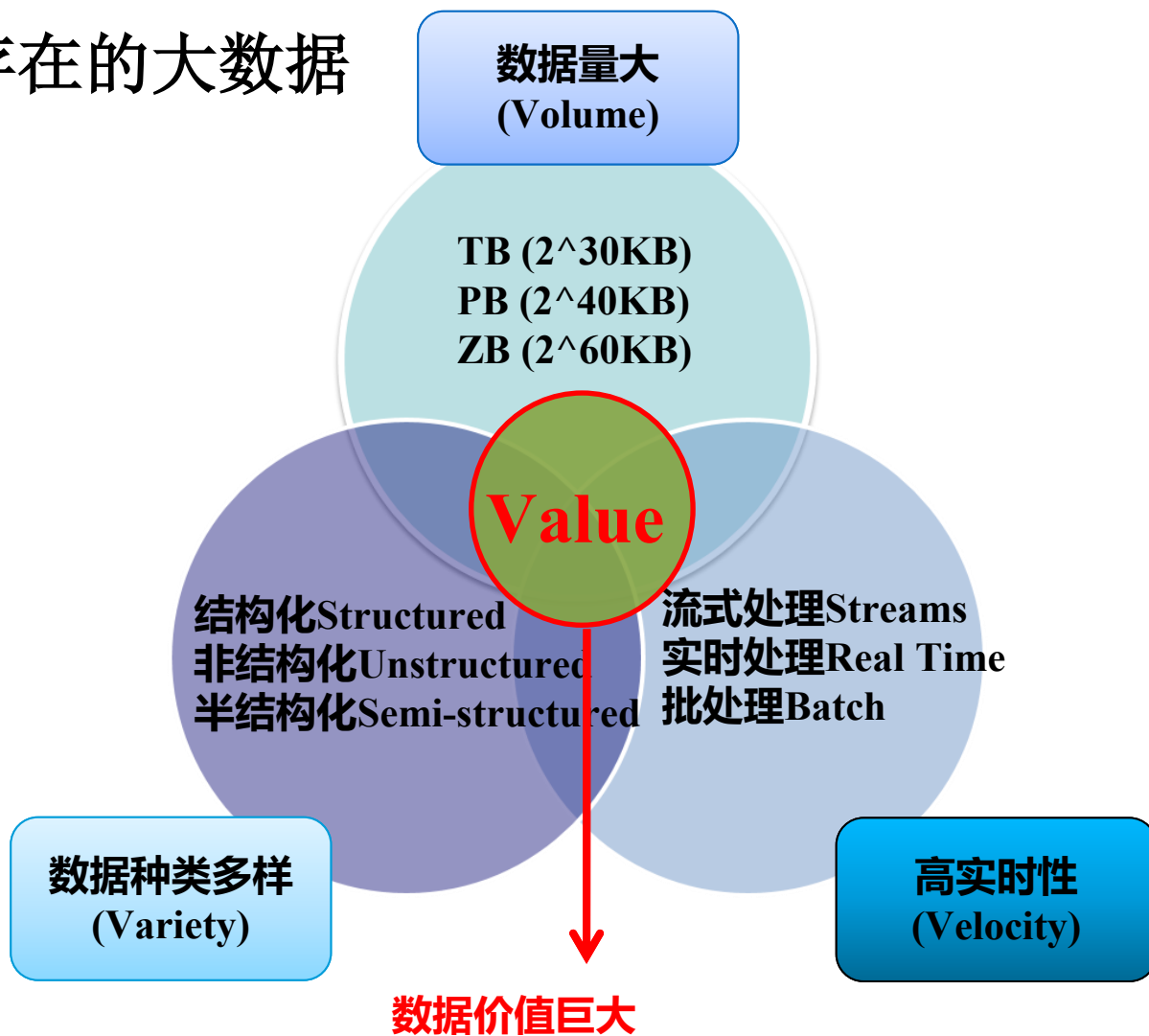
从2008年9月,《Nature》杂志首次出版一期大数据专刊,科学家们提出“大数据真正重要的是新用途和新见解,而非数据本身”



# 大数据时代

6

## 客观存在的大数据





# 大数据时代

7

## □ 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)

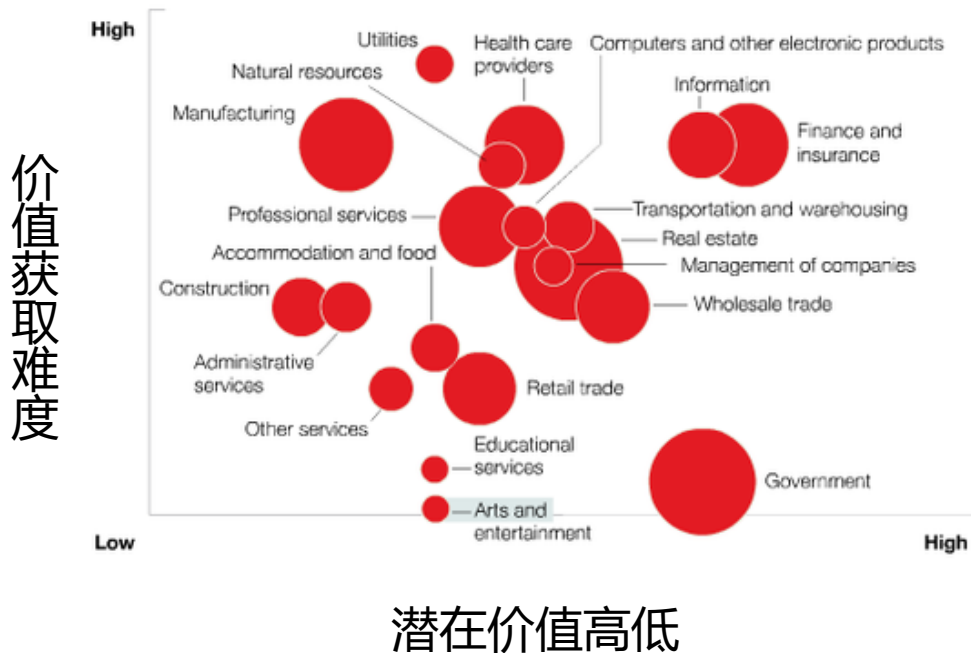


# 大数据时代

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

11

## □ 大数据改变了人们的衣食住行



过去

现在



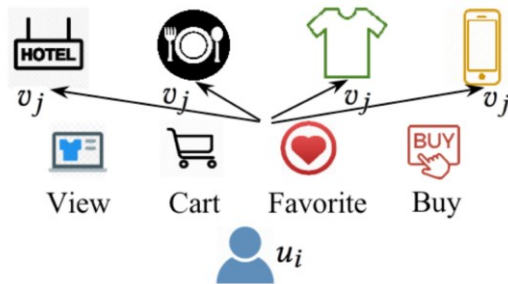


# 大数据蕴含巨大的价值

## 大数据促进了个性化服务



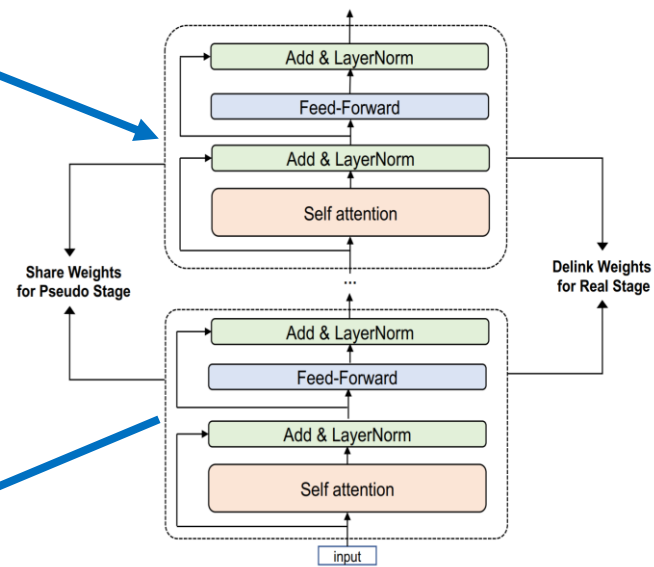
电商平台



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

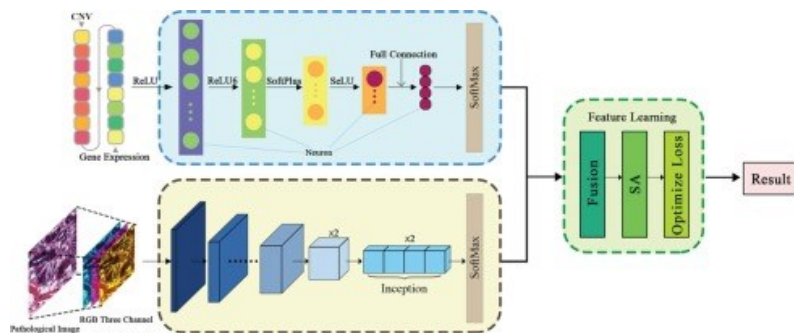
# 大数据蕴含巨大的价值

13

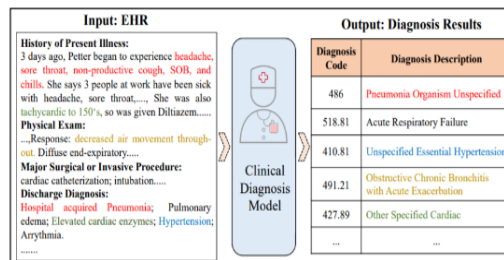
## 大数据辅助智慧医疗



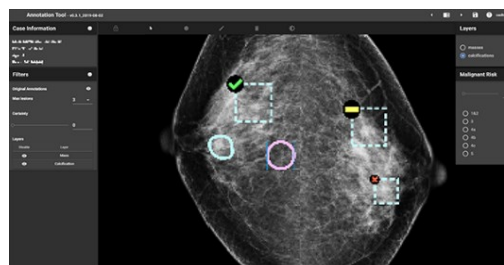
AI电子病历



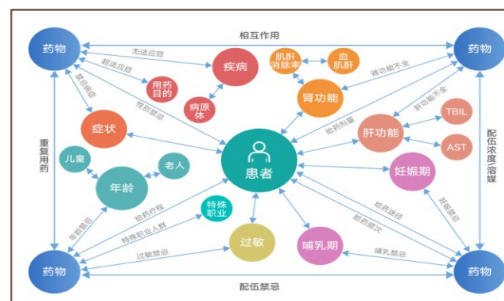
多模态医疗数据挖掘模型



疾病诊断



病灶识别



药物推荐



# 大数据蕴含巨大的价值

14

## □ 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。



# ChatGPT

15

+ New chat

五人买苹果

洛杉矶凌晨四点

Clear conversations

Dark mode

My account

Updates & FAQ

Log out

Model: GPT-4

有5个人去买苹果，他们买的苹果数分别是A, B, C, D, E, 已知A是B的3倍, C的4倍, D的5倍, E的6倍, 则A+B+C+D+E最小为多少?

已知A是B的3倍, C的4倍, D的5倍, E的6倍。为了使A+B+C+D+E的和最小, 我们可以假设B、C、D、E分别为A的倍数的最小公倍数。

首先我们找到3、4、5、6的最小公倍数。分解质因数, 得到:

$$3 = 3^1$$

$$4 = 2^2$$

$$5 = 5^1$$

$$6 = 2 * 3 = 2^1 * 3^1$$

对于最小公倍数, 需要取每个质因数的最高次幂:

$$\text{LCM} = 2^2 * 3^1 * 5^1 = 4 * 3 * 5 = 60$$

所以我们可以令B=60, C=60, D=60, E=60。根据已知条件, 我们可以求出A:

$$A = 3B = 3 * 60 = 180$$

$$A = 4C = 4 * 60 = 240$$

$$A = 5D = 5 * 60 = 300$$

$$A = 6E = 6 * 60 = 360$$

显然, 要使A+B+C+D+E的和最小, A应取最小值180。此时, A+B+C+D+E的和为:

$$180 + 60 + 60 + 60 + 60 = 420$$

所以, A+B+C+D+E最小为420。

Regenerate response



|



# 国内外ChatGPT类技术概览







# 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**

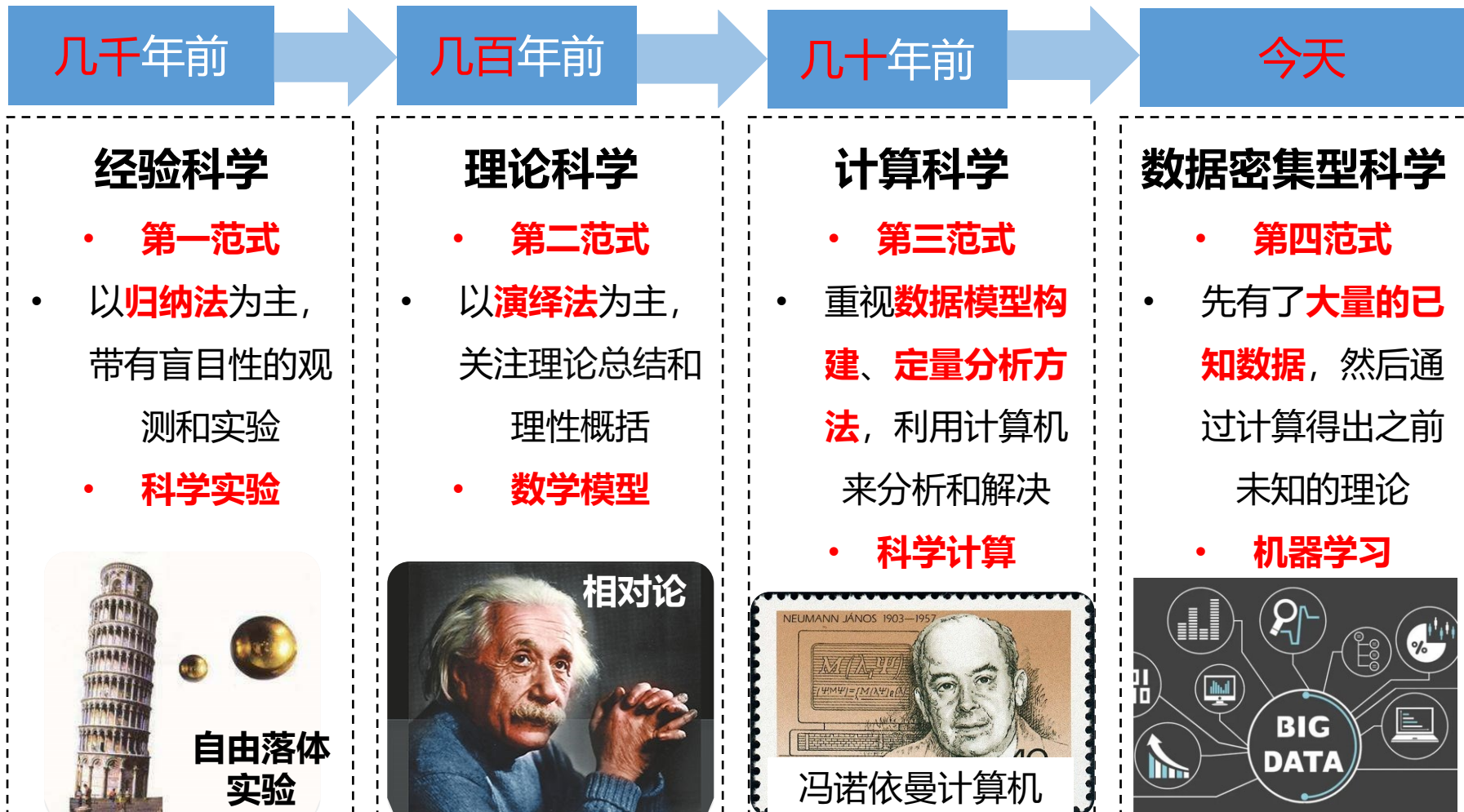
**大模型**

**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





# 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



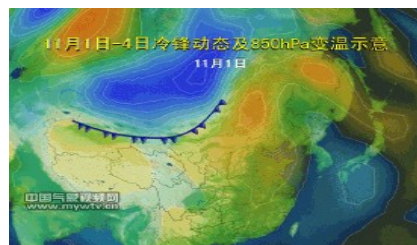
突发事件预测、关键人群监测

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



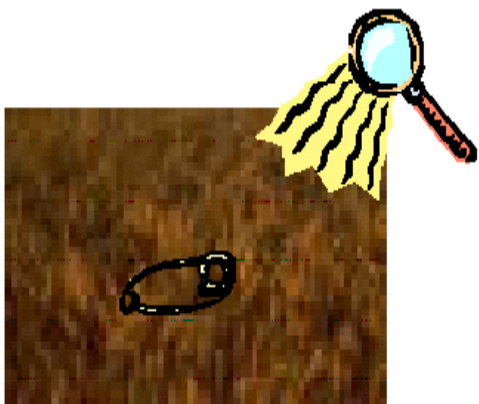
城市智慧管理

# 数据库是基础中的基础!



# 数据库是基础中的基础

20



- **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

21

数据/Data



结构化



半结构化



非结构化

1. 采集、预处理

5. 可视化

4. 分析挖掘

3. 计算模式

应用/App



分布式文件系统  
数据库

计算系统

2. 存储管理

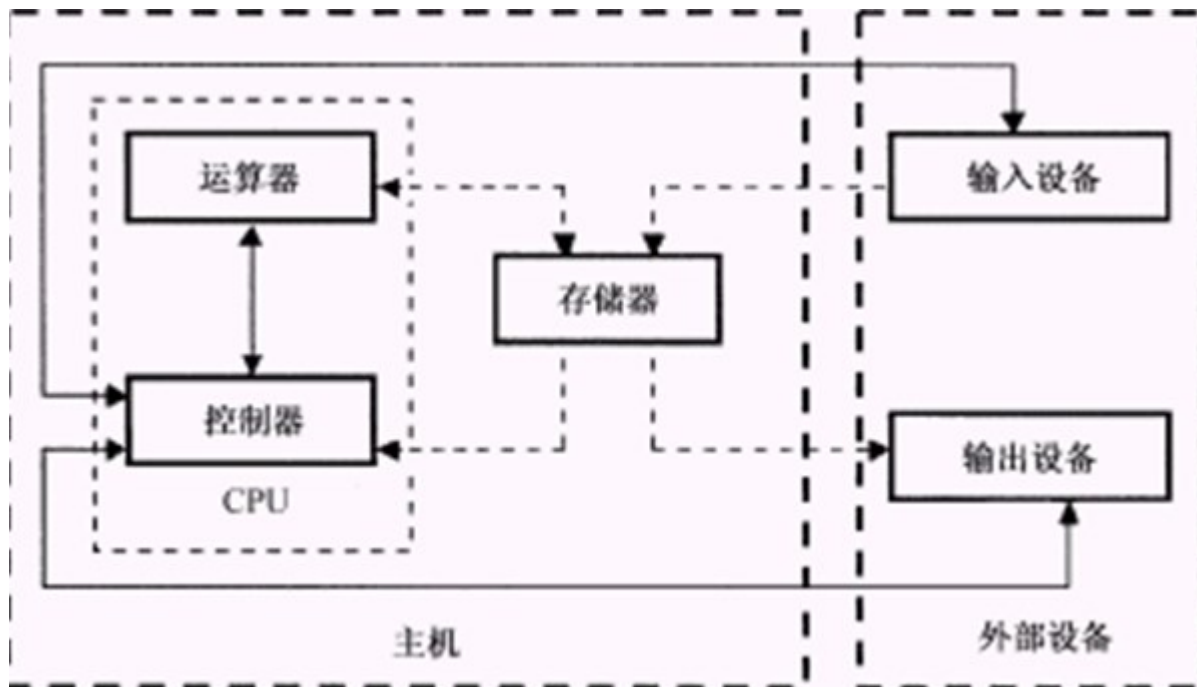
6. 隐私和安全





# 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



# 数据库的位置和作用

## 存储器





# 数据库的位置和作用

24







# 数据库的位置和作用

25

## □ 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

26

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 <b>1974</b>	ACM	<a href="http://www.sigmod.org">http://www.sigmod.org</a>
2	SIGKDD	ACM Knowledge Discovery and Data Mining <b>1995</b>	ACM	<a href="http://www.acm.org/sigkdd/">http://www.acm.org/sigkdd/</a>
3	SIGIR	International Conference on Research and Development in Information Retrieval <b>1978</b>	ACM	<a href="http://www.acm.org/sigir/">http://www.acm.org/sigir/</a>
4	VLDB	International Conference on Very Large Data Bases <b>1975</b>	Morgan Kaufmann/ACM	<a href="http://www.vldb.org">http://www.vldb.org</a>
5	ICDE	IEEE International Conference on Data Engineering <b>1984</b>	IEEE	<a href="http://www.icde.org/">http://www.icde.org/</a>



# 数据库的图灵奖

## 四次图灵奖

2015年CNCC（合肥）

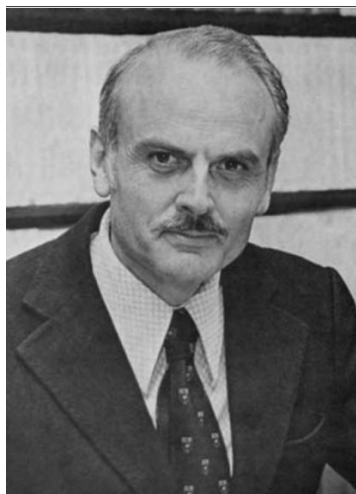
1973年



**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

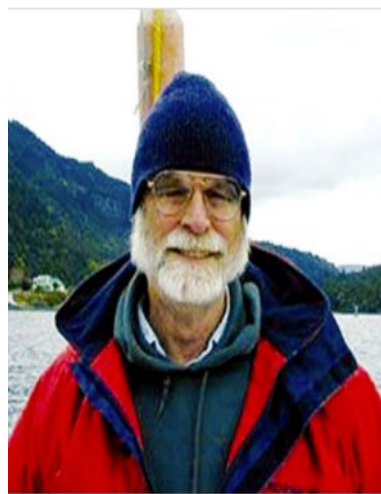
1981年



**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





# 数据库

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

417 systems in ranking, February 2024

Rank			DBMS	Database Model	Score		
Feb 2024	Jan 2024	Feb 2023			Feb 2024	Jan 2024	Feb 2023
1.	1.	1.	Oracle +	Relational, Multi-model ⓘ	1241.45	-6.05	-6.08
2.	2.	2.	MySQL +	Relational, Multi-model ⓘ	1106.67	-16.79	-88.78
3.	3.	3.	Microsoft SQL Server +	Relational, Multi-model ⓘ	853.57	-23.03	-75.52
4.	4.	4.	PostgreSQL +	Relational, Multi-model ⓘ	629.41	-19.55	+12.90
5.	5.	5.	MongoDB +	Document, Multi-model ⓘ	420.36	+2.88	-32.41
6.	6.	6.	Redis +	Key-value, Multi-model ⓘ	160.71	+1.33	-13.12
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model ⓘ	135.74	-0.33	-2.86
8.	8.	↓ 7.	IBM Db2	Relational, Multi-model ⓘ	132.23	-0.18	-10.74
9.	9.	↑ 12.	Snowflake +	Relational	127.45	+1.53	+11.80
10.	↑ 11.	↓ 9.	SQLite +	Relational	117.28	+2.08	-15.38
11.	↓ 10.	↓ 10.	Microsoft Access	Relational	113.17	-4.50	-17.86
12.	12.	↓ 11.	Cassandra +	Wide column, Multi-model ⓘ	109.27	-1.77	-6.95
219.	↓ 217.	↓ 214.	TDSQL for MySQL +	Relational, Multi-model ⓘ	0.90	-0.13	-0.12
259.	↑ 262.	↓ 241.	Kingbase	Relational, Multi-model ⓘ	0.57	-0.09	-0.19
233.	↑ 248.	↑ 283.	Alibaba Cloud AnalyticDB for MySQL +	Relational, Multi-model ⓘ	0.76	-0.02	+0.27
234.	↑ 245.		Vald	Vector	0.76	-0.04	
235.	↓ 230.	↓ 198.	Blazegraph	Multi-model ⓘ	0.75	-0.16	-0.43



# 数据库

## “中国知网”：2019年，2021年



### 行业知识服务与知识管理平台

农林牧渔、卫生、科学研究

农业 食品 医疗 药业 公共卫生 国土  
检验检疫 环保 水利 气象 海洋 地震

### 研究学习平台

研究型学习平台

研究生 本科生 高职学生  
中职学生 中学生

大数据研究平台

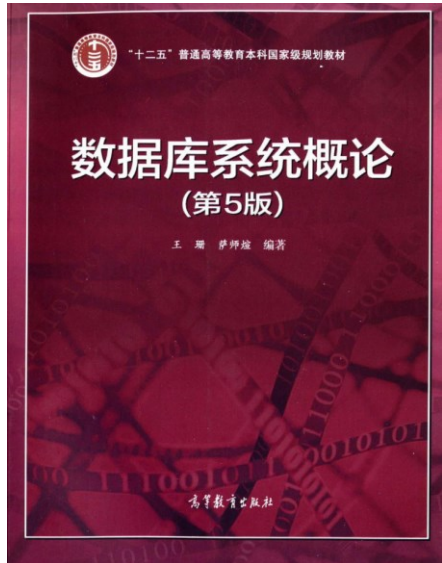
专利分析 学术图片 统计数据 学术热点  
学者库 统计分析



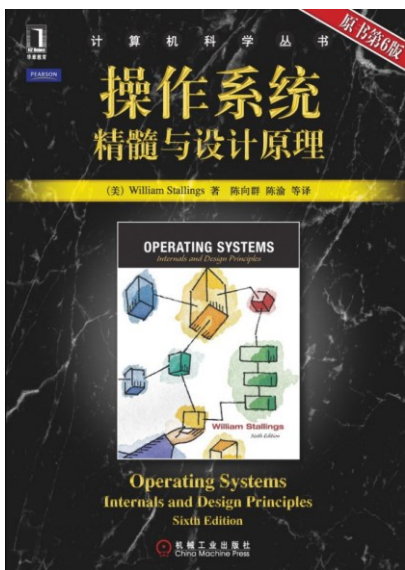


# 数据库是核心课程之一

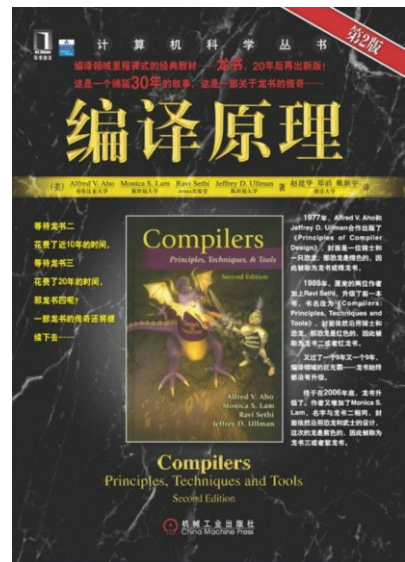
- 了解计算机系统的核心课程



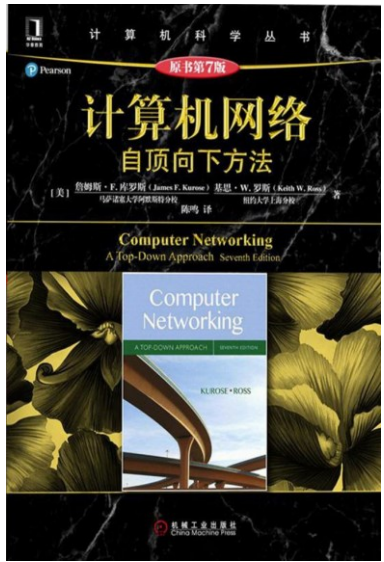
数据库系统



操作系统



编译系统



计算机网络



作为一门计算机科学领域中非常重要的课程，《数据库》是学习计算机科学的必修课之一。



以下是我对这门课程的评价：

优点：

1. 实用性强：在现代社会中，数据库系统已经成为企业和组织管理数据的主要方式。学习数据库可以让学生掌握如何设计、实现和维护数据库系统，对于学生未来的职业发展有很大的帮助。
2. 广泛应用：数据库技术不仅在企业中广泛应用，而且在各种领域都有着重要的地位。比如，在金融、医疗、教育等行业中，都需要使用数据库来管理和分析数据。
3. 丰富的知识内容：数据库课程不仅包括数据库的基本概念和原理，还包括数据库设计、SQL语言、数据库应用开发等多个方面的知识，能够全面地培养学生的数据库技能。
4. 实践性强：数据库课程通常包括实验环节，学生需要使用数据库管理系统来完成实验任务，这有助于学生更好地理解数据库技术的应用和实现。

缺点：

1. 抽象性较强：数据库技术本身较为抽象，学习过程中需要学生理解和掌握一定的理论知识。这对于一些学生而言可能较为困难。
2. 难度较大：数据库技术涉及到多个方面的知识，学习难度较大。同时，对于一些需要进行数据库设计和开发的实际应用来说，需要更多的实践和经验。
3. 需要较好的数学基础：数据库技术中涉及到一些数学概念和计算，需要学生具备较好的数学基础才能更好地理解和掌握。

总体而言，作为一门重要的计算机科学课程，《数据库》具有非常广泛的应用前景，对于学生未来的职业发展具有很大的帮助。同时，需要注意的是，学习难度较大，需要学生具备一定的抽象思维和数学基础，需要认真学习并进行实践。



# 数据库课程

35

- 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

36

## □ 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：周三 14:00-15:35/周五 15:55-17:30
- 地点：西区 3C104

## □ 实验练习、答疑课 (暂定)

- 时间：周五 晚上19:00-20:00
- 地点：西区 科技实验楼东楼 717



# 课程主页与课程群

## 课程主页

- <http://staff.ustc.edu.cn/~huangzhy/Course/DB2024.html>
- Linked in: <http://staff.ustc.edu.cn/~huangzhy/>

## □ 课程联系方式

- 课程群: 667690990, (请所有选课同学加入)
- 课程邮箱: ustcdb\_2024@163.com

## □ 助教

- 董艳民, ustcdb\_2023@163.com
- 黄弈骁, ustcdb\_2023@163.com
- 马志远, ustcdb\_2023@163.com



# 教材及参考书(1)

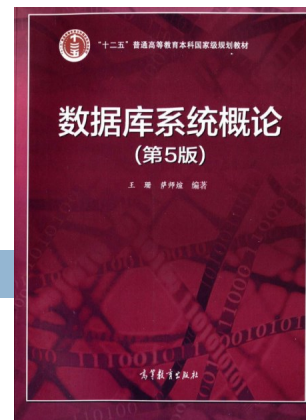
38

## 教材

- 《数据库系统概论(第五版)》  
王珊, 萨师焯; 高等教育出版社, 2014.9

- 《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  
库管理系统内核研制  
突破了数据库管理系  
权19项，出版著作6  
务、电子党务、国际  
个重大信息化工程  
领域的发展，全面推

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

40

## 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/DB2024.html>





# 学习方式

41

## 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

42

## □ 平时成绩 (30%)

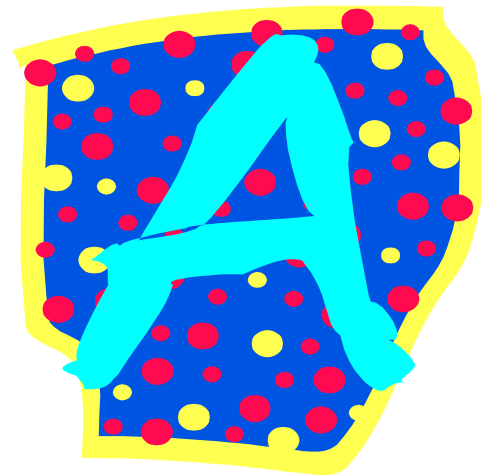
课后作业, 实验练习 + 课堂出勤(三次不到, 取消考试成绩)

## □ 实验成绩 (20%)

课程设计

## □ 期末考试 (50%)

卷面成绩



2/28/2024



# 内容安排(1)

43

## 基础篇

- 第一章 绪论
- 第二章 关系数据库
- 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)

44

## 系统篇

- 第九章 关系查询处理和查询优化
- 第十章 数据库恢复技术
- 第十一章 并发控制

## 新技术篇

- 第十二章 大数据管理
- 第十三章 NoSQL技术



# 课程目标

45

## □ 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

46

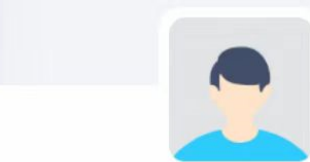
- 专业基础课：只想听课堂课程的同学谨慎选择
- 几点说明
  - 平时30% + 课程设计20% + 期末考试50%
    - 作业，实验，大实验（组队），笔试（闭卷）
  - 课堂与平时重要，大家保证出勤
    - 不许叠课
  - 不同专业中的学分认定，需跟相关专业教秘确认



# Any Questions?

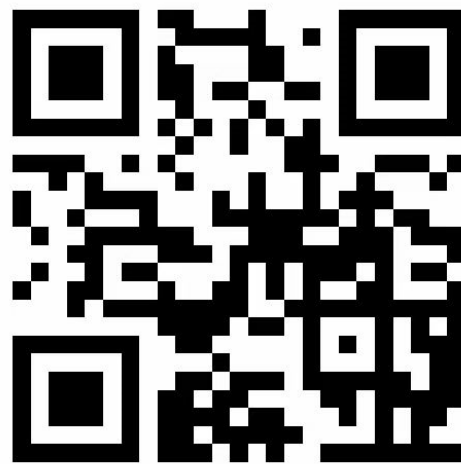
47

□ 课程QQ群： 667690990



数据库系统概论2024

群号：667690990



扫一扫二维码，加入群聊。

