



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

5

数据科学导论

Introduction to Data Science

第一章 数据科学基础

陈恩红，黄振亚

Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2024.html>

助教：肖桐
ds_intro2024@163.com

9/10/2024



数据科学基础

43

改变这个世界的第四种力量——大数据的应用

暴力



金钱



世界著名未来学家托夫勒
《第三次浪潮》作者

知识



大数据



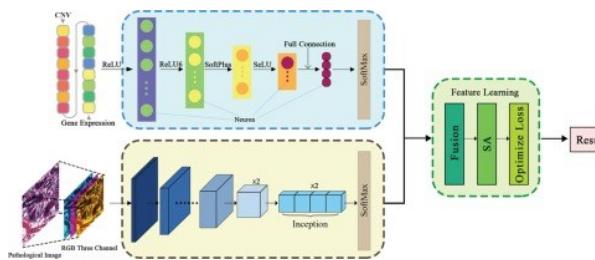
数据科学基础

44

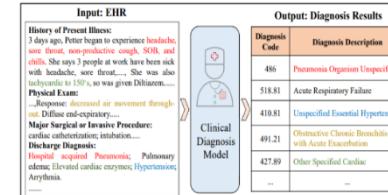
- 数据蕴含着巨大的价值—智慧医疗
 - 通过对患者建立AI电子病历
 - 整合患者的全时段、多模态的健康数据（病例文本、检查影像等）
 - 实现对患者的疾病诊断、病灶识别、药物推荐等



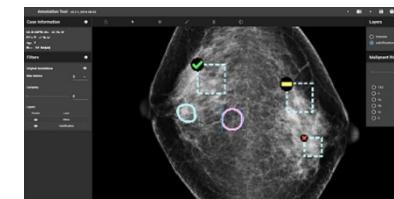
AI电子病历



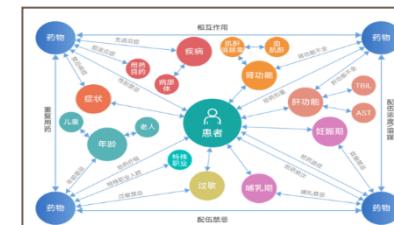
多模态医疗数据挖掘模型



疾病诊断



病灶识别



药物推荐



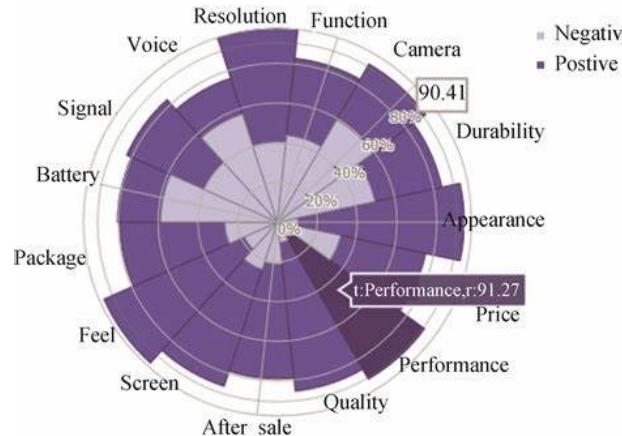
数据科学基础

46

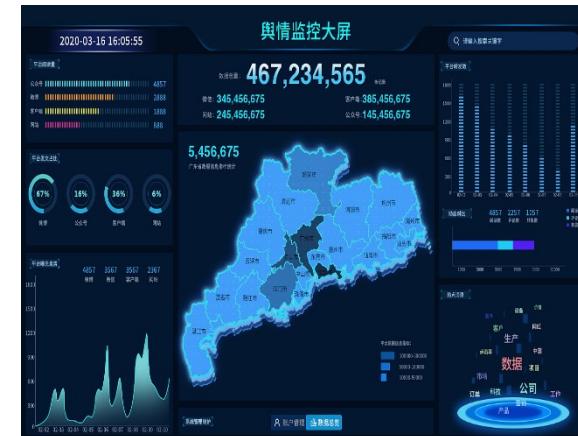
□ 数据蕴含着巨大的价值—安防领域

□ 舆情监测

- 通过对新闻网站、论坛、博客等的文章和评论进行文本挖掘和情感分析，安防系统能够实时捕捉公众对特定话题的反应。
- 通过大数据分析识别并跟踪网络上虚假信息的传播路径，并帮助相关机构及时干预，防止谣言扩散引发社会恐慌或动乱。



舆情情感分析



传播途径监测

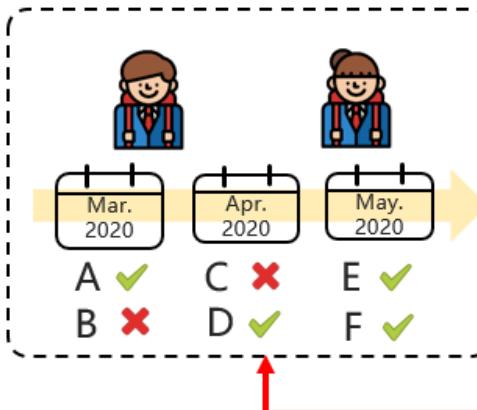


数据科学基础

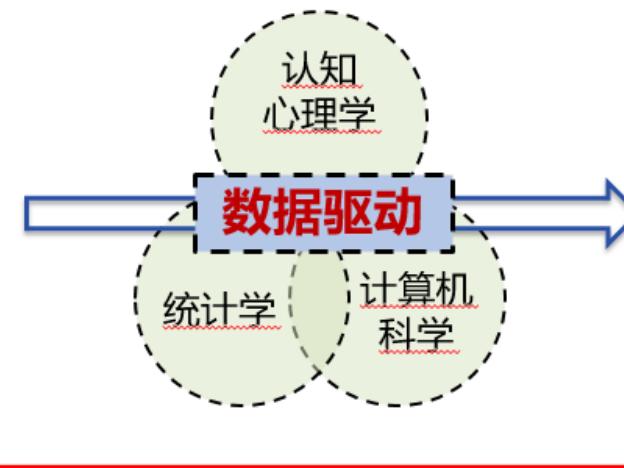
47

- 数据蕴含着巨大的价值—智慧教育
 - 学习者能力分析
 - 量表范式 到 数据驱动的范式

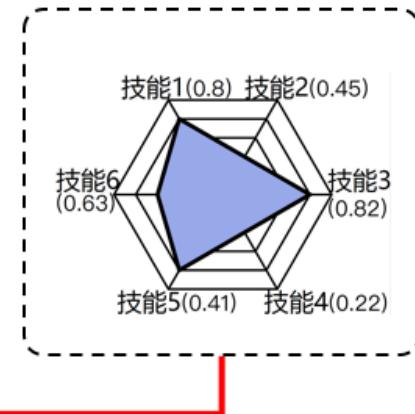
学习过程相关数据



知识能力及其发展变化



反馈调整教学策略

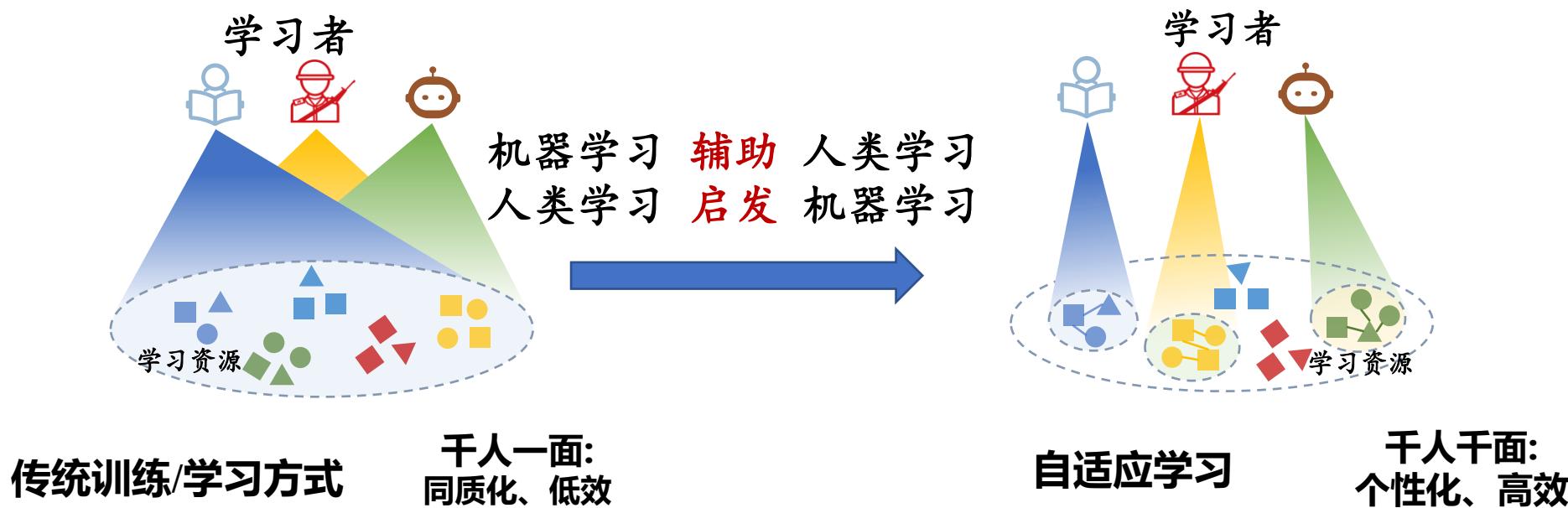




数据科学基础

48

- 数据蕴含着巨大的价值—智慧教育
 - 规模化因材施教：千人千面的自适应学习



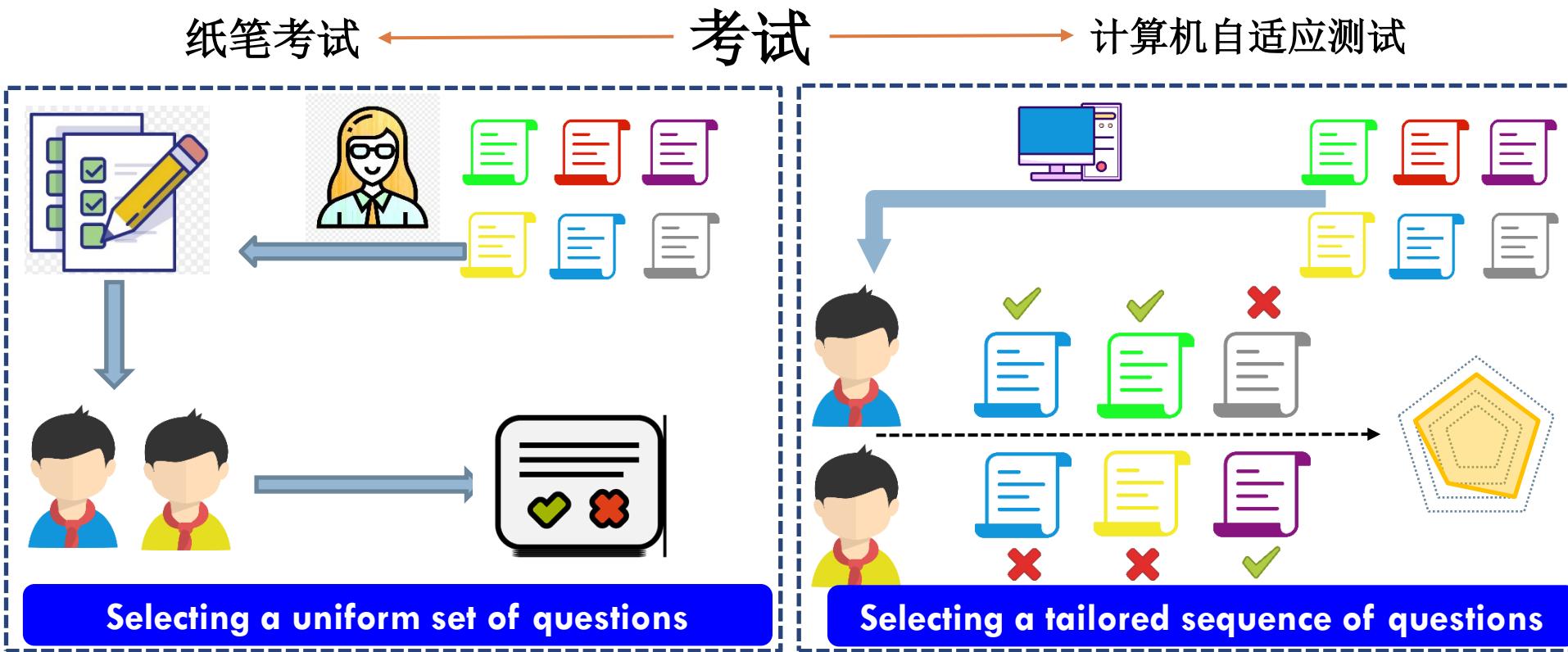


数据科学基础

49

- 数据蕴含着巨大的价值—智慧教育
 - 数据驱动的自适应测试 (CAT)

GRE®





数据科学基础

51

- 数据蕴含着巨大的价值—智慧教育
 - 大模型驱动的智慧教育

学：智慧助教 教：数字教师 研：科研助手 管：以智助管

大型学习助手

什么是布斯算法？已知二进制数 $X=0.1010, Y=-0.1101$ 。
利用布斯算法求 $[X \cdot Y]$ 的补码。

布斯算法是一种补码乘法算法，用于计算两个用补码表示的二进制数的乘积。在布斯算法中，被乘数和乘数均用补码表示，连同符号位一起参与运算。布斯算法的运算过程包括将两个数的补码相乘，并根据乘数相邻两位的情况进行特定操作，最终得到乘积的补码。

根据已知的二进制数 $X=0.1010, Y=-0.1101$ ，可以进行布斯算法求解 $[X \cdot Y]$ 的补码：

1. 将 X 和 Y 表示为补码： $[X]_{\text{补}} = 00.1010$
 $[Y]_{\text{补}} = 11.0011$ $[-X]_{\text{补}} = 11.0110$

2. 根据布斯算法的运算过程，进行乘法运算，得到乘积的补码： $[X \cdot Y]_{\text{补}} = 1.01111100$

因此，根据布斯算法， $[X \cdot Y]$ 的补码为
 1.01111100 。



数字化智能体系统

星火科研助手

星火科研助手致力于创建高效、可靠的科技文献服务平台，为学术研究人员、企业专业人士和其它相关用户提供服务，助力科技发展创新。

成果调研 论文研读 辅助写作



关注订阅号
及时了解科研服务前沿信息

关注服务号
随时随地了解科研助手动态





数据科学基础

52

□ 数据蕴含着巨大的价值—智慧教育

高校编程平台(CODIA): <https://code.bdaa.pro>

- 平台特色：利用人工智能技术辅助学生编程学习
 - 面向学生：提供**学生诊断、习题推荐、智能助教**等个性化服务
 - 面向老师：支撑教师**自主出题，测试情况跟踪**等
- 平台应用：已在中国科大本科生课程实践教学中使用
 - 2021-2024年，计算机学院《程序设计II》、《数据结构》课程实验教学
 - 2021-2024年，实验室**保研/考研机试**



最近更新

C语言初学者看过来，这里有属于你的一份惊喜
知识路径模块上线,迎来多重更新学以致练，能力进阶，知识路径C语言模块与配套习题重磅上线，整合智能三模块，打通学练全流程

NICE/WC
迎接新学员上线
立即查看 立即体验

14459人 用户数量 | 371259次 提交次数 | 1717天 安全运行

用户 ovu6f6kox 刚刚提交了题目 的结果 用时 29 ms

智慧助教 用户调用：
21825 times 28574341 tokens

编程到来，智能将来

掌握编程，是对智能时代的准备，亦是提升竞争力的重要途径。在这里，你可以获得：

- 历年考研及保研机试题库
- 数据分析及算法等课程相关题库
- 个性化能力评估和诊断

现在加入

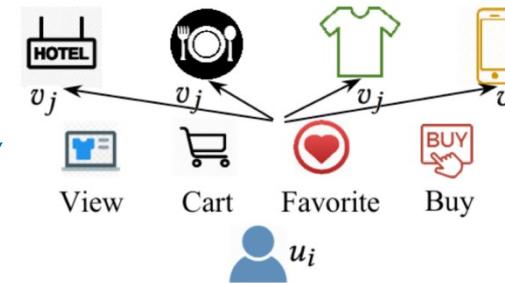
数据结构

- 算法基本操作
- 字符串处理
- 二叉树遍历
- 堆栈和队列
- 图的搜索算法



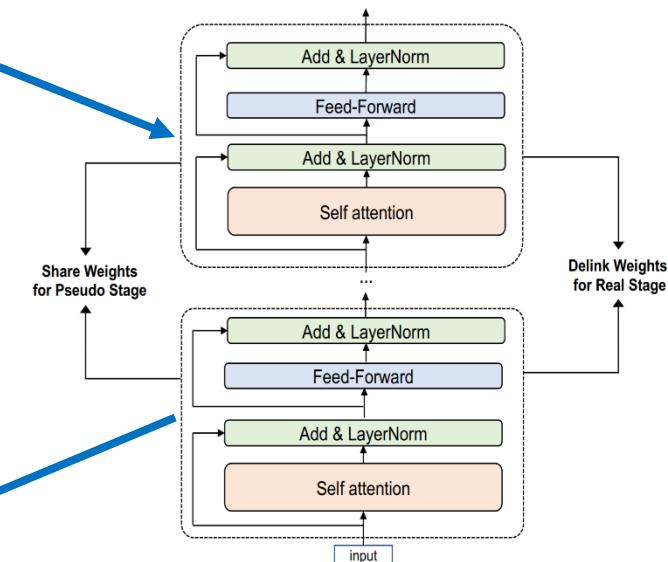
数据科学基础

- 数据蕴含着巨大的价值—电子商务
 - 计算广告



电商平台

促进用户消费、提升平台收益





数据科学基础

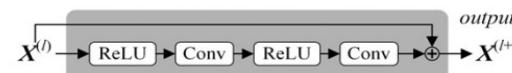
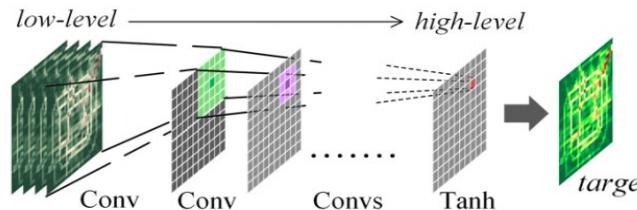
55

- 数据蕴含着巨大的价值 — 智慧城市
 - 基于运营商基站数据、交通路口和车辆移动轨迹数据等
 - 提升城市**交通管控、交通服务和规划水平**，实现**用户未来行程规划、交通路口信号灯调控、合理规划道路建设等**

车辆
移动
数据



交通
流预
测模
型



行程规划



信号灯调控



道路规划

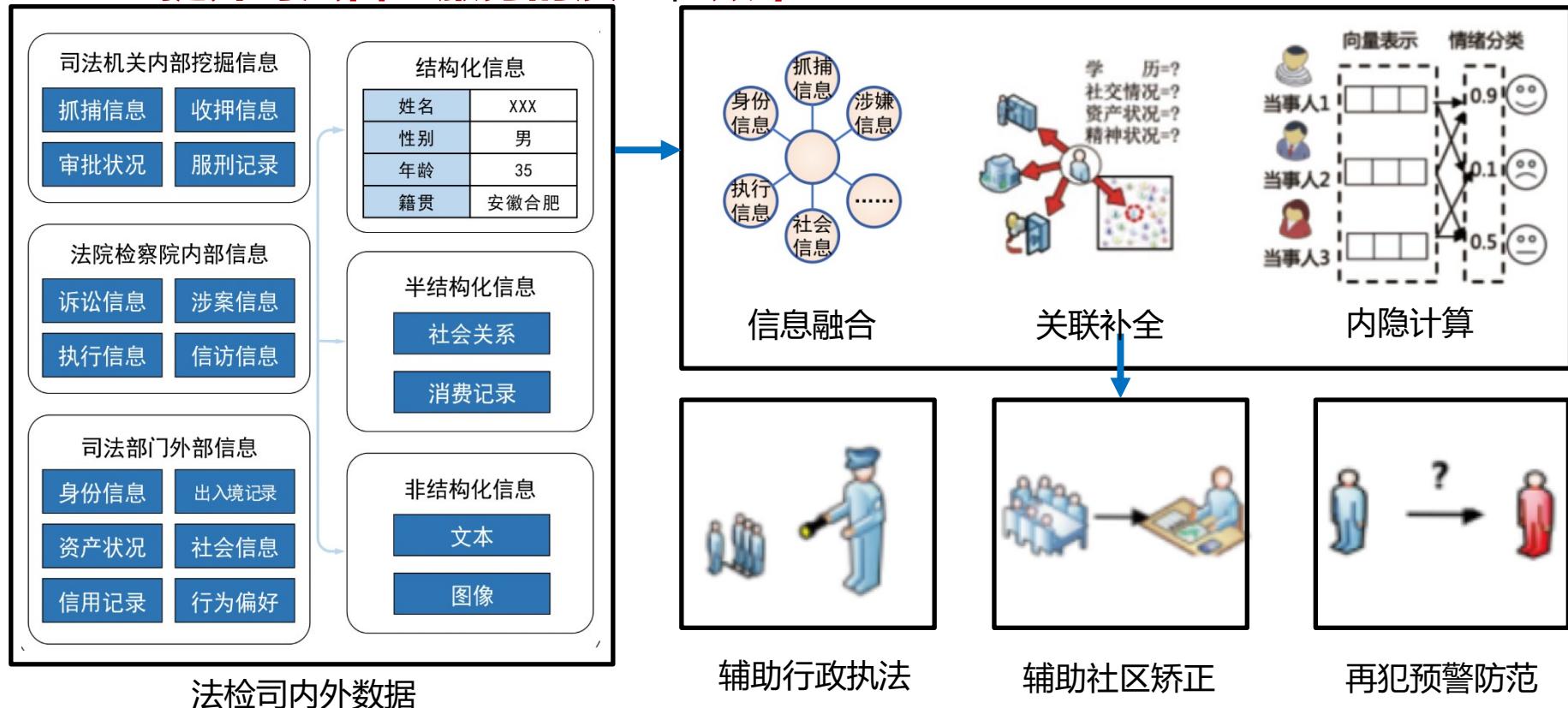


数据科学基础

56

□ 数据蕴含着巨大的价值——智慧司法

- 基于法、检、司等部门关于涉案当事人的内部数据与外部数据等
- 构建涉案当事人画像，辅助行政执法、社区矫正、进行再犯预警防范等，
提升司法管理服务的质量和效率

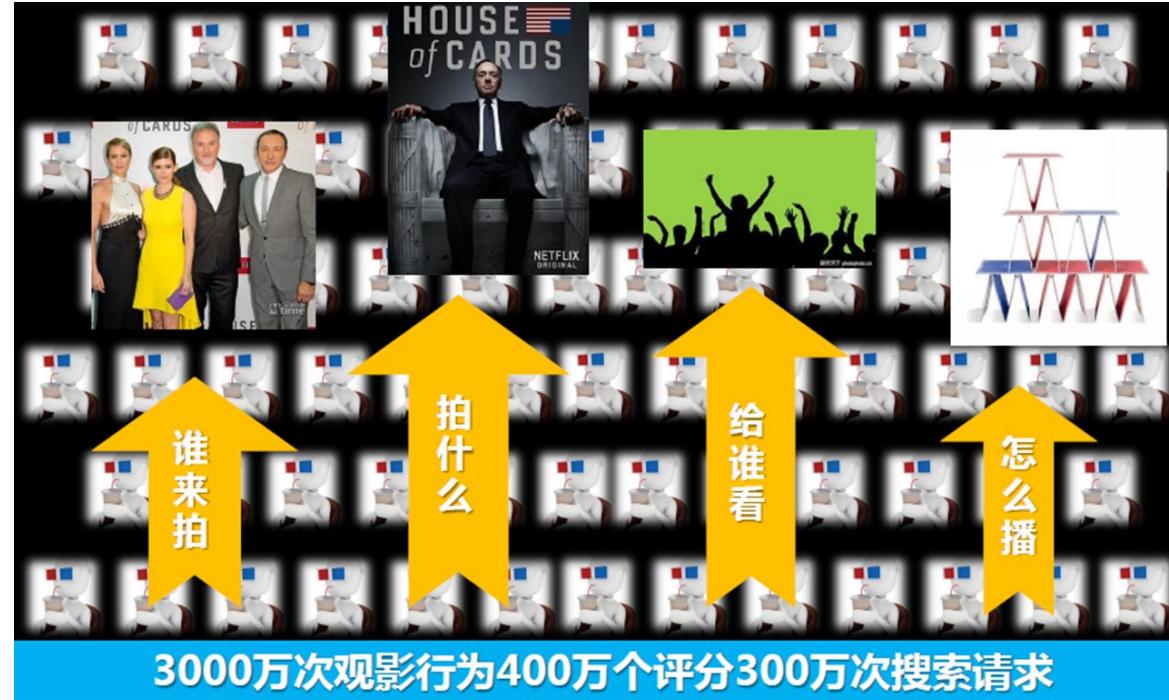




数据科学基础

57

- 数据蕴含着巨大的价值—文化娱乐
 - 纸牌屋效应：数据决定影视剧的内容



大卫·芬奇
凯文·史派西

老版《纸牌屋》

喜欢老版纸牌屋
及同类剧的用户

13集同时上线



数据科学基础

58

- 数据蕴含着巨大的价值—文化创作
 - 机器作诗

用户写作意图

江、船、秋风

春风、杨柳

松、竹、山、牧童



诗歌自动生成系统



对应诗词

江北江南万顷秋，
船头人去水悠悠。
一帆一棹秋风急，
又有离人万里愁。

杨柳千条拂地垂，
一川春水浸桃花。
游人不识湖中路，
游遍人间野水涯。

乔松古木两三间，
松竹阴中一径斜。
白鸟不知山路远，
牧童踏过野人家。



数据科学基础

59

□ 机器作诗 PK 古代诗人



哪首诗是人写的?

秋夕湖上
一夜秋凉雨湿衣，
西窗独坐对夕晖。
湖波荡漾千山色，
山鸟徘徊万籁微。

秋夕湖上
荻花风里桂花浮，
恨竹生云翠欲流。
谁拂半湖新镜面，
飞来烟雨暮天愁。

机器

宋代诗人葛绍体



数据科学基础

60

□ 数据蕴含着巨大的价值—文化娱乐

- 流行音乐的旋律与编曲生成
- 机智过人：

http://tv.cctv.com/2017/11/24/VIDEo7JWp0u0oWRmPbM4uC_Bt171124.shtml





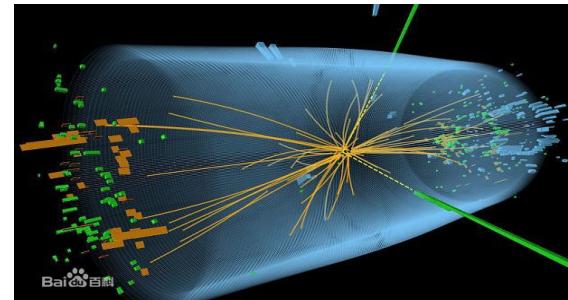
数据科学基础

61

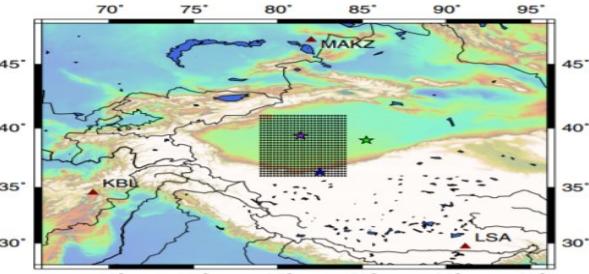
- 数据蕴含着巨大的价值—科学技术
 - ◆ 大数据推动科学新技术发现



天文大数据搜索新星



物理大数据预测分子属性



大数据地震速报、余震预测



生物大数据改良基因

科学
大数
据



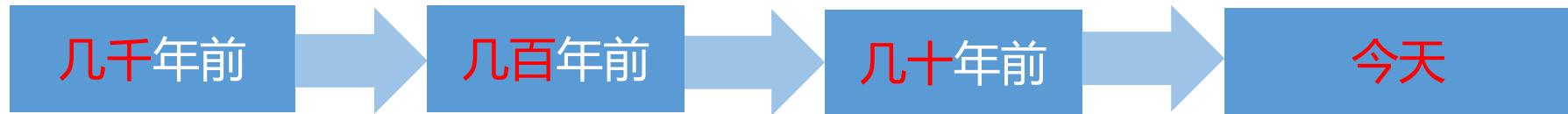
数据挖掘保护知识产权



数据科学基础

63

■ 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



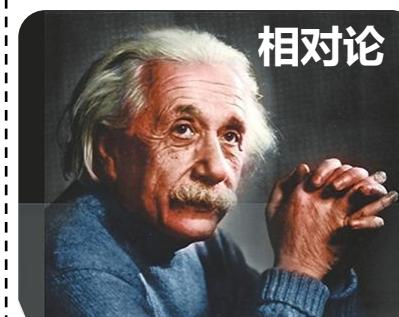
自由落体实验

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



冯诺依曼计算机

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

天文大数据





数据科学基础

64

□ 大数据带来的技术创新-当前热点

□ 音频生成

- 科大讯飞语音合成：中文领域突出、方言发言
- SO-VITS-SVC(2022)：让AI歌手翻唱任何歌曲

普通话发音人 方言发音人 多语言发音人

音色：全部 成年男声 成年女声 童年女声 童年男声 特性：全部 多风格 特色IP

场景推荐：全部 语音助手 通用场景 客服助理 新闻播报 智能客服 阅读听书 教育培训

此场景我们推荐以下发音人：

聆飞哲 成年男声 多风格 推荐：年轻时尚、亲切自然	聆小璐 成年女声 多风格 推荐：温柔活泼、亲切自然	聆小ワイ 成年女声 多风格 推荐：温柔时尚、亲切自然
聆小樱 成年女声 多风格 推荐：温柔放松、亲切自然	聆小瑶 成年女声 多风格 推荐：温柔甜美、亲切自然	聆飞晨 成年男声 推荐：年轻时尚、亲切自然
聆小璇 成年女声 多风格 推荐：温柔开朗、亲切自然	聆小珊 成年女声 推荐：温柔活泼、亲切自然	乐乐 童年女声 推荐：呆萌女童





数据科学基础

65

□ 大数据带来的技术创新-当前热点

□ AI图像生成

- Stable Diffusion-2022: 开源工具; 本地部署
- Midjourney-2022: 图像生成社区





数据科学基础

66

- 大数据带来的技术创新-当前热点
 - 微表情生成与识别



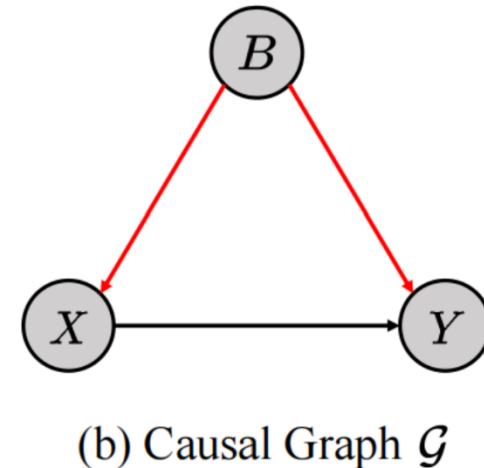
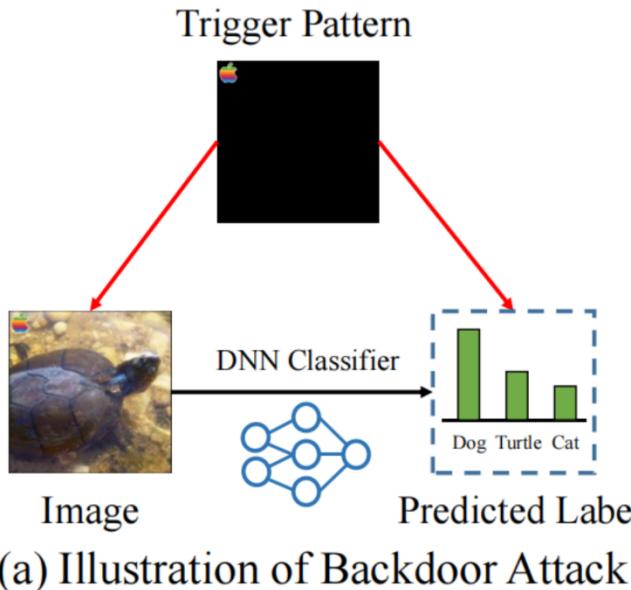
Wenhai Leng, Sirui Zhao, Yiming Zhang, Shiifeng Liu, Xinglong Mao, Hao Wang, Tong Xu*, Enhong Chen*, ABPN: Apex and Boundary Perception Network for Micro- and Macro-Expression Spotting , ACM MM 2022



数据科学基础

67

- 大数据带来的技术创新-当前热点
 - 可信AI与因果推断
 - 基于后门准则的去混表征学习-2023





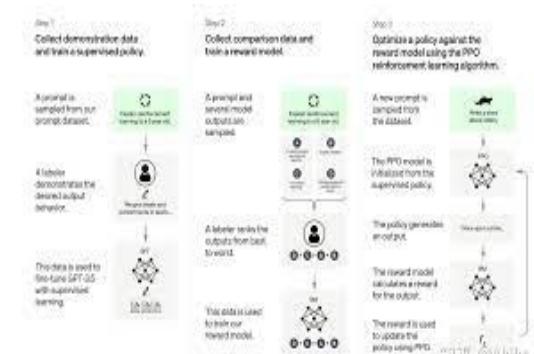
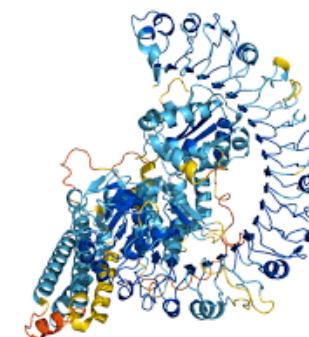
数据科学基础

68

□ 大数据带来的技术创新 – 当前热点

□ 强化学习

- 2016年，AlphaGo以4: 1的战绩击败李世石，机器第一次在围棋领域战胜人类顶尖高手
- 2020年，Alpha Fold在蛋白质折叠问题的预测平均准确率超过90%
- 2022年，ChatGPT应用RLHF进行微调，取得惊人效果





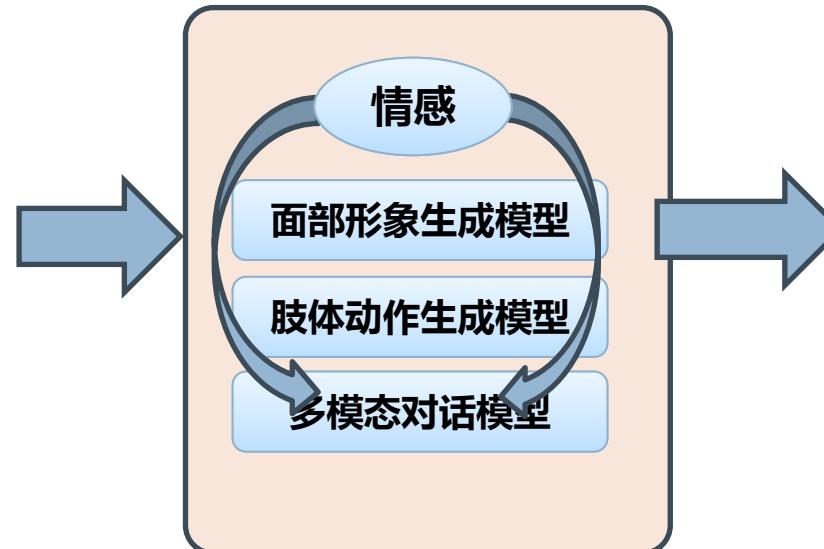
数据科学基础

69

- 大数据带来的技术创新-当前热点
 - 数字人技术；情感数字虚拟人生成
 - 中科大-华为云合作项目
 - “形”“声”“色”“态”多元角度生成高度拟人化、有温度的虚拟数字人



多模态输入



虚拟人生成模型



情感化数字人



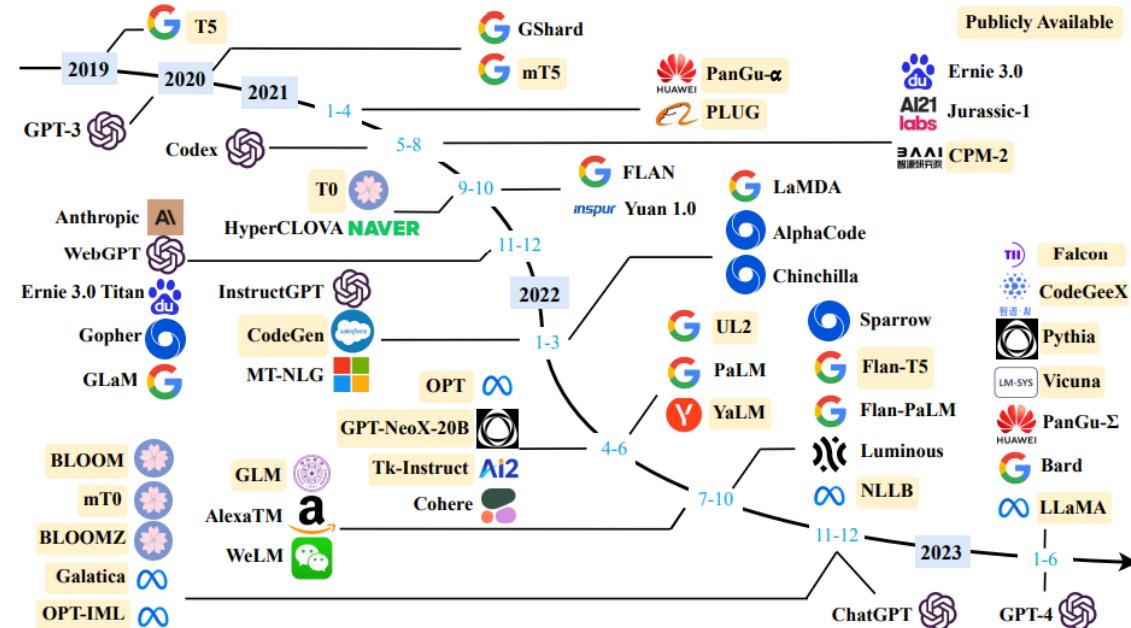
数据科学基础

70

□ 大数据带来的技术创新 – 当前热点

□ 大语言模型

- 如今大语言模型层出不穷，例如 ChatGPT，ChatGLM，LLaMa等，具备大量知识，能够以类人的方式和人类进行多轮对话，在多种自然语言理解相关任务上表现良好。





数据科学基础

71

- 大数据带来的技术创新-当前热点
 - 大语言模型——知识图谱，信息检索
 - 庞大的知识库
 - 通过指令实现在多种任务上的零样本/少样本学习

XB

什么是量子纠缠？



量子纠缠是一种奇特的量子力学现象，它发生在两个或多个粒子之间，当它们之间存在一种特殊的量子状态时，这些粒子之间的属性会彼此紧密关联，无论它们之间有多远的距离，这些关联性都会保持下去。这些属性可以包括粒子的自旋、位置、动量等。

XB

将文本中的语句分类成积极和消极两种情感

文本1：这张照片拍的不错。

文本2：今天雨下的很大，我的鞋湿了。



文本1：这张照片拍得不错。 - 积极情感

文本2：今天雨下得很大，我的鞋湿了。 - 消极情感

ChatGPT作为知识库回答问题

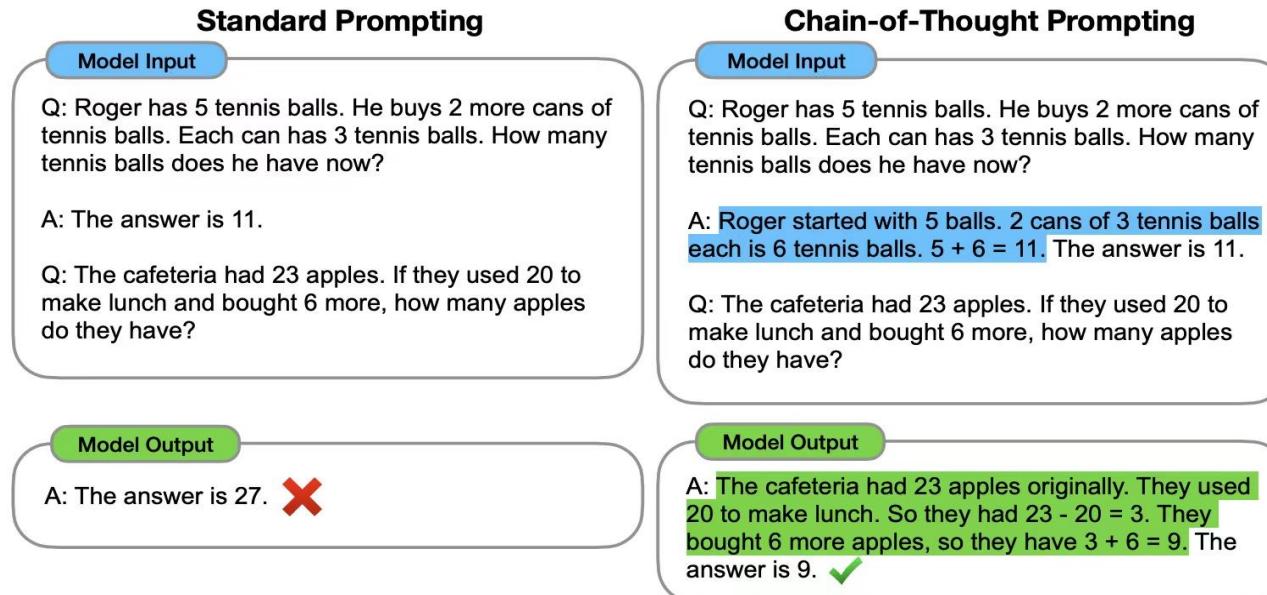
ChatGPT仅通过任务描述完成文本情感分类



数据科学基础

72

- 大数据带来的技术创新-当前热点
- 大语言模型——逻辑推理
 - 通过思维链 (CoT) 解决复杂问题



向大语言模型展示少量样本并解释推理过程，模型在回答时也会显示推理过程从而引出正确回答



数据科学基础

73

- 大数据带来的技术创新 – 当前热点
- AI录播课：以录播课的形式达到直播课的效果

The screenshot displays a virtual classroom interface. On the left, a slide titled "挑战讲解1:什么是BMI" (Challenge Lecture 1: What is BMI) shows the "BMI的标准范围 (中国)" (BMI Standard Range (China)) with four color-coded human silhouettes: blue (偏瘦), teal (正常), yellow (过重), and orange (肥胖). Below the silhouettes, the BMI ranges are listed: ≤ 18.4 (偏瘦), $18.4 < \text{bmi} \leq 23.9$ (正常), $23.9 < \text{bmi} \leq 28.0$ (过重), and $\text{bmi} > 28.0$ (肥胖). At the bottom of this slide, there is a "BMI体重健康管理" (BMI Weight Management) section with icons for Python, Java, C++, and others, along with a "学习人数" (Number of Learners) of 35 people. On the right, a video player window shows a male teacher named Meng Xiangyang. A text box from the AI teaching assistant says: "你好呀，我是你的专属AI编程小助手，你可以向我提问关于本节课需要知道的相关问题，我将会帮助你——解答哦~". Below the video player, there is a text input field asking "计算机如何进行四则运算？" (How does a computer perform arithmetic operations?).

支持师生实时交互，进而为
每个同学打造一堂专属虚拟
课堂，提高上课效果



数据科学基础

74

- 大数据与人工智能
 - ABC当前AI的技术体系

Big data

数据

大数据是人工智能发展的**基石**，人工智能的核心在于数据支持。

机器学习算法是人工智能的**核心**，是今天引领人工智能发展潮流的一大类算法

算法

AI

算力

人工智能算法的实现需要强大的计算能力**支撑**，特别是深度学习算法的大规模使用，对计算能力提出了更高的要求。

Algorithm

Computation



数据科学基础

75

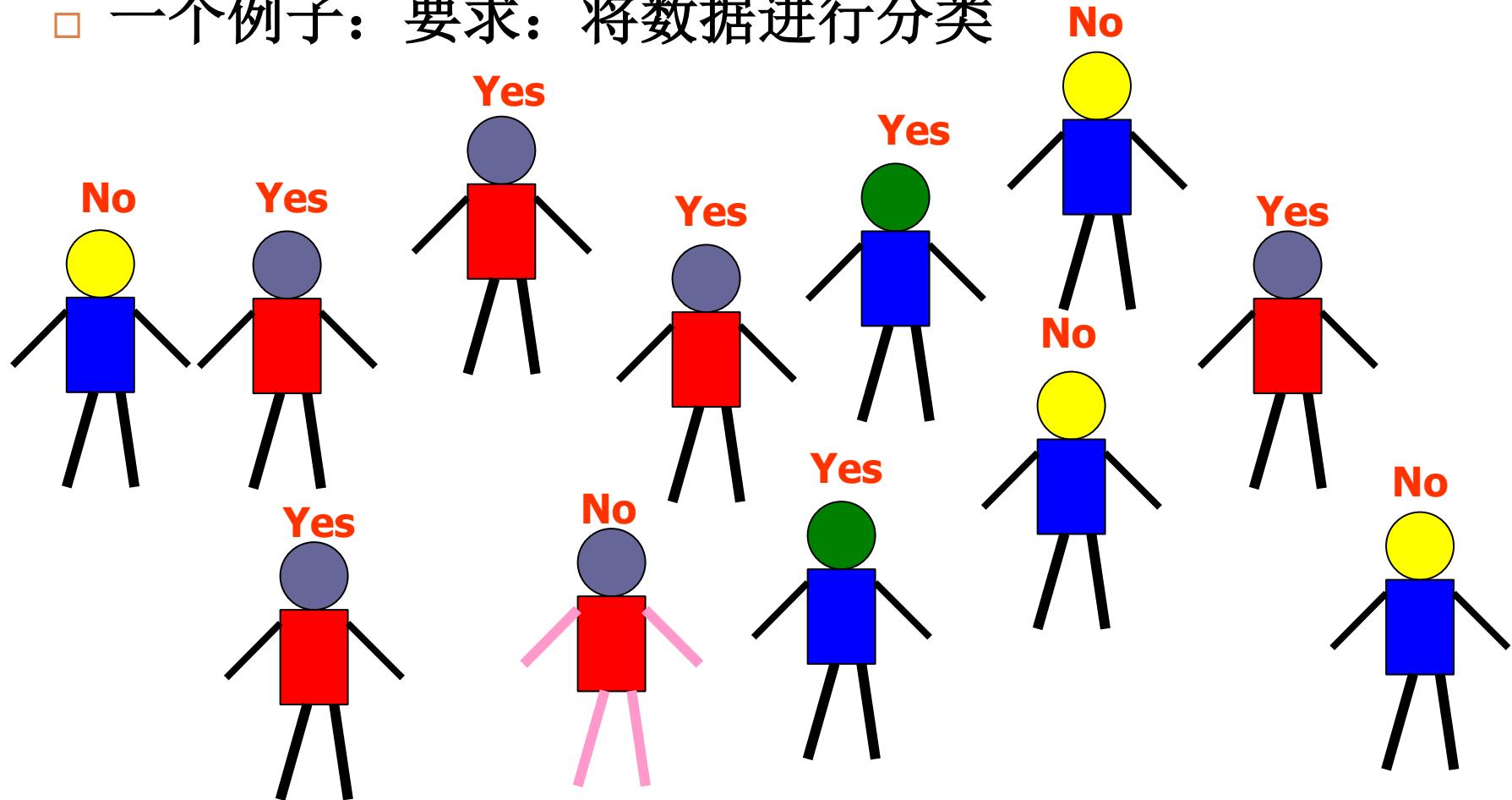
- 大数据与人工智能的关系
- 现阶段，人工智能的核心是对大数据进行的**特征抽取与机器学习算法**





数据+分类学习的方法

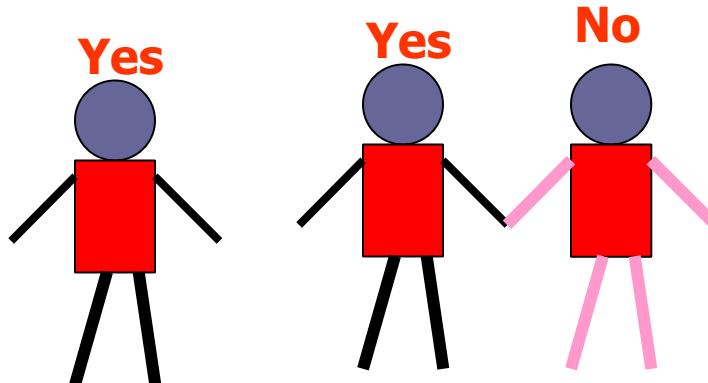
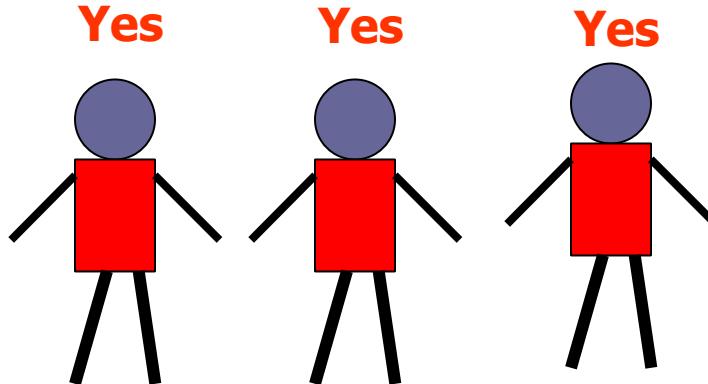
□ 一个例子：要求：将数据进行分类



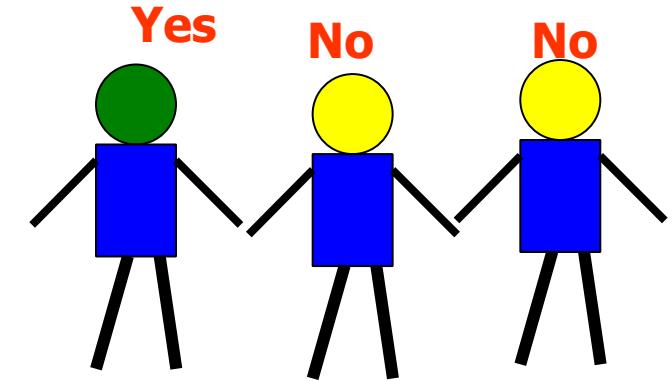
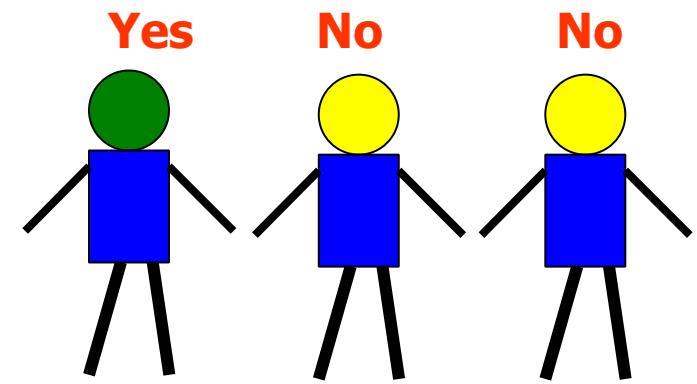


数据+分类学习的方法

躯干：红色



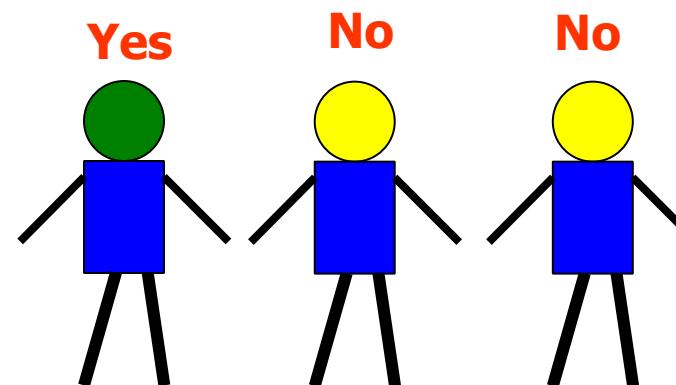
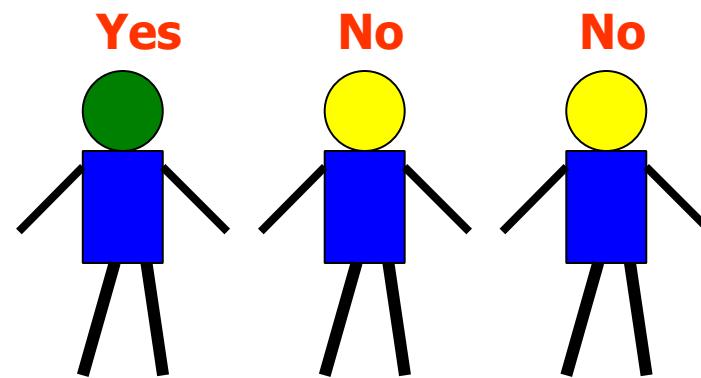
躯干：蓝色





数据+分类学习的方法

躯干：蓝色



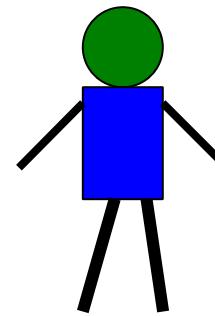


数据+分类学习的方法

躯干：蓝色

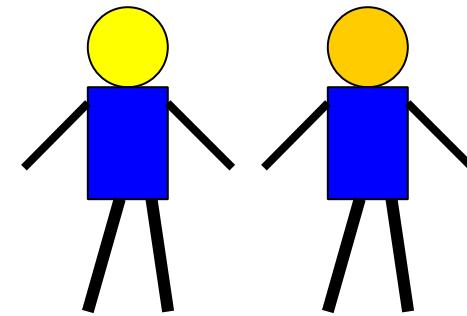
头：绿色

Yes

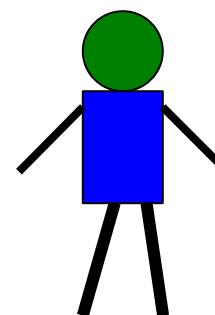


头：黄色

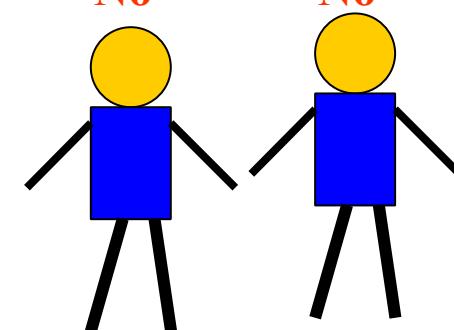
No



Yes



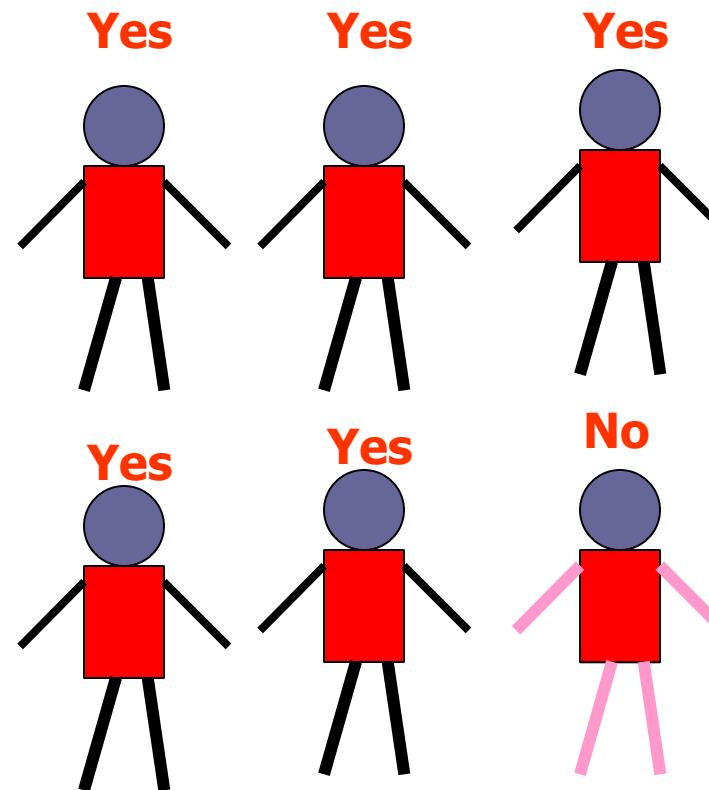
No





数据+分类学习的方法

躯干：红色





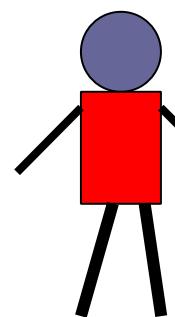
数据+分类学习的方法

躯干：红色

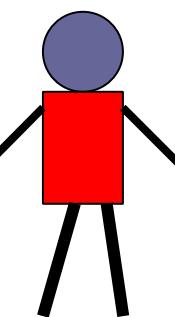
四肢：黑色

四肢：粉红色

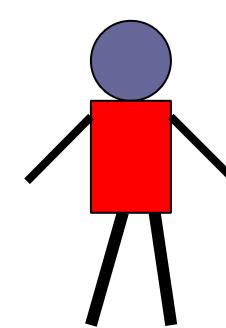
Yes



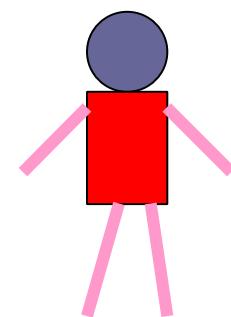
Yes



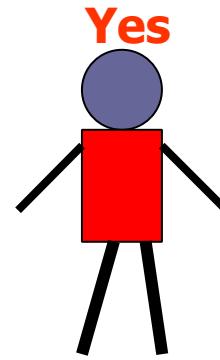
Yes



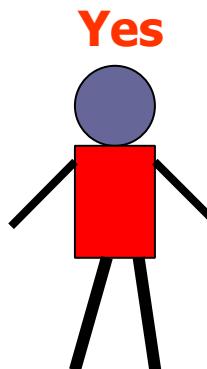
No



Yes

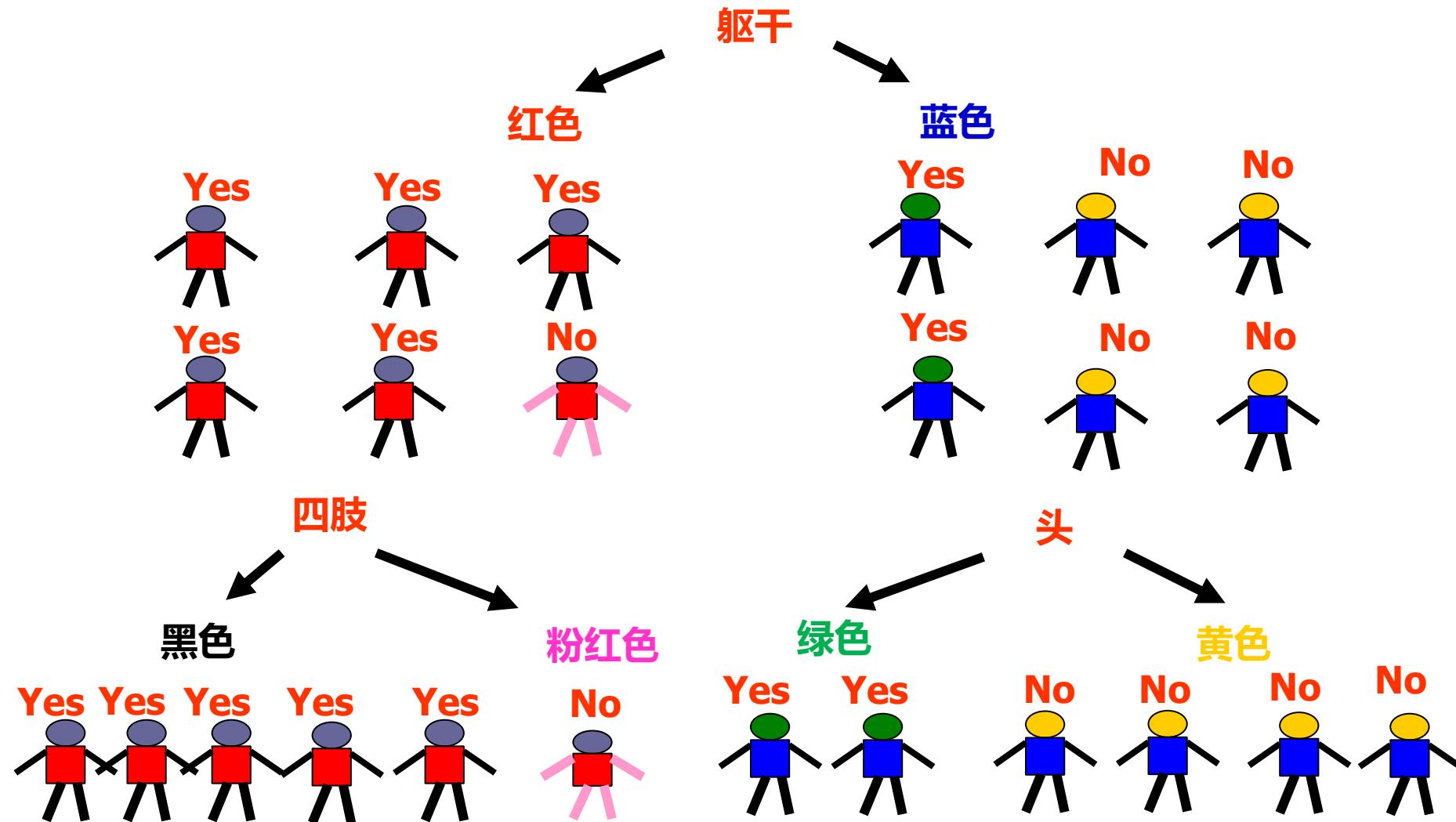


Yes





数据+分类学习的方法



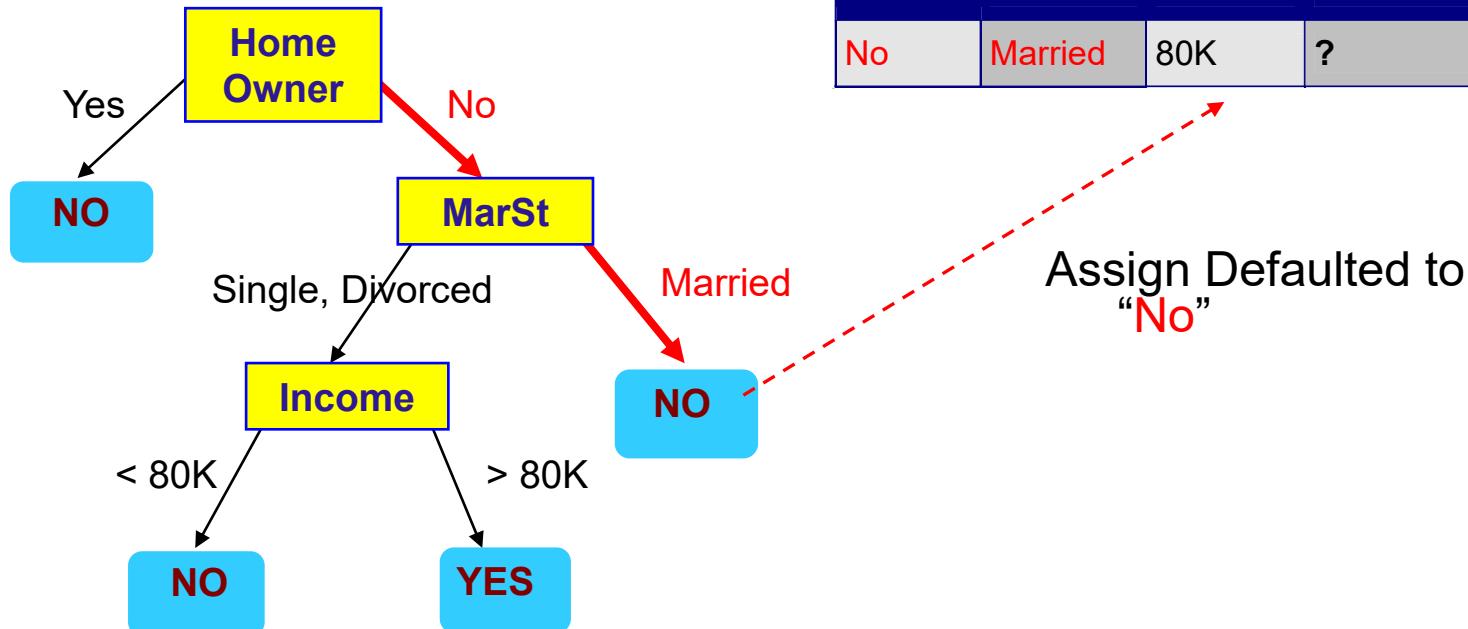


数据+分类学习的方法

- 决策树（第四章）——使用模型对测试数据分类

Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?





数据科学基础

84

- 包括高效的CPU/GPU、云计算、AI芯片、多机集群并行化处理等技术手段



- 云计算：EPYC（霄龙）处理器；Project 47服务器



更快



更稳定



更智能

是全球首个配备专用神经网络计算引擎的SoC

- 自学习神经元芯片：Loihi



- 云计算：可重配置加速堆栈（FPGA-Accelerator Stack）
- 设备端：reVISION加速堆栈



- CPU架构：Cortex-A76
- GPU架构：Mali G76

+智能 计算进化

Huawei FusionServer Pro智能服务器

观看视频

项目咨询



- 移动端：麒麟980芯片



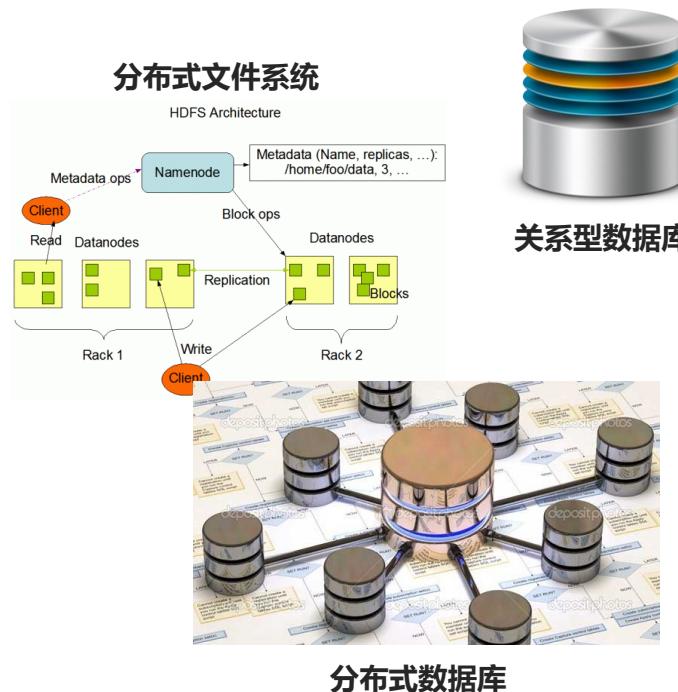
- 跨界处理器：i.MX RT1060



数据科学基础

85

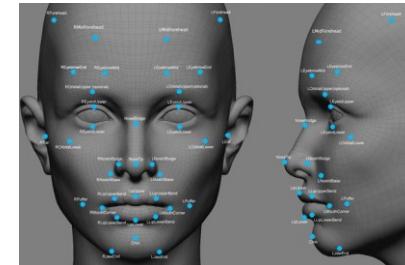
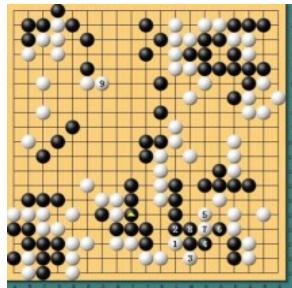
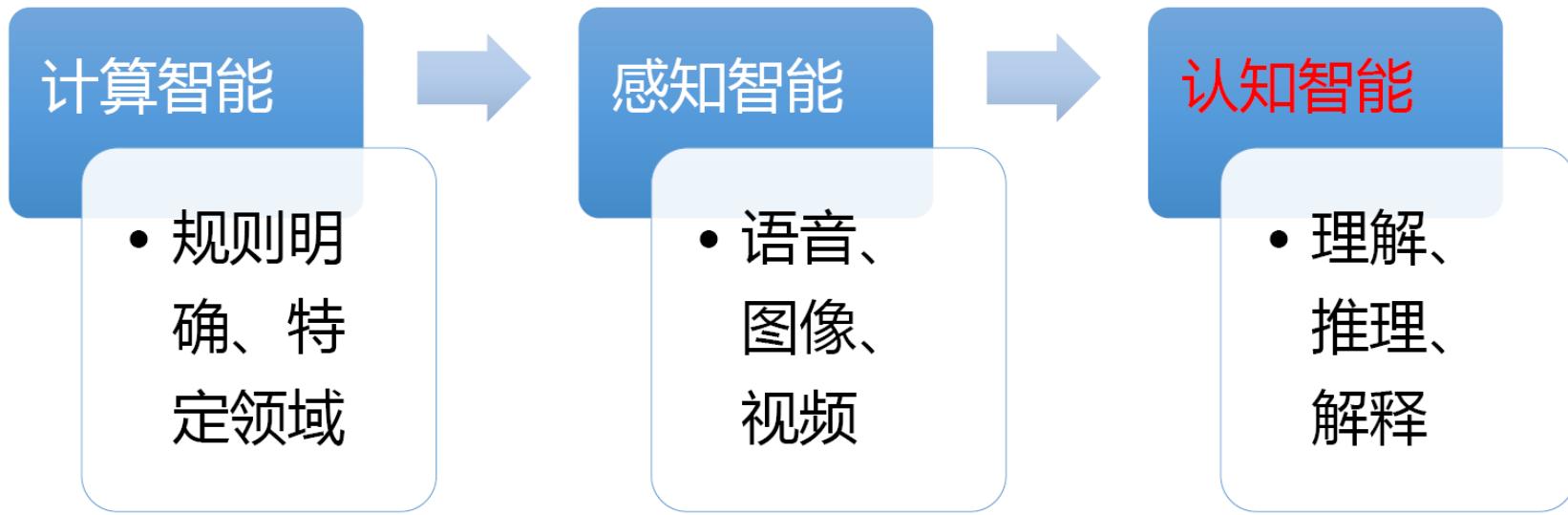
- 包括高效的CPU/GPU、云计算、AI芯片、多机集群并行化处理等技术手段
 - 数据处理和智能计算任务的多元化促使相关软件的多样化





数据科学基础

□ 大数据的未来—数据与知识融合，让人工智能更“聪明”



能穿多少穿多少



能穿多少穿多少



数据科学基础

□ 大数据的未来—从数据中的相关性到世界的因果推断

逻辑关系

- 归纳法、数理逻辑、布尔代数系统

$$(a \vee b) \vee c = a \vee (b \vee c)$$
$$(a \wedge b) \wedge c = a \wedge (b \wedge c)$$

重推理

相关关系

- 贝叶斯网络、机器学习、深度学习



重分析（
学习）

因果关系

- 因果关系是有方向的、存在时序先后性

万有引力



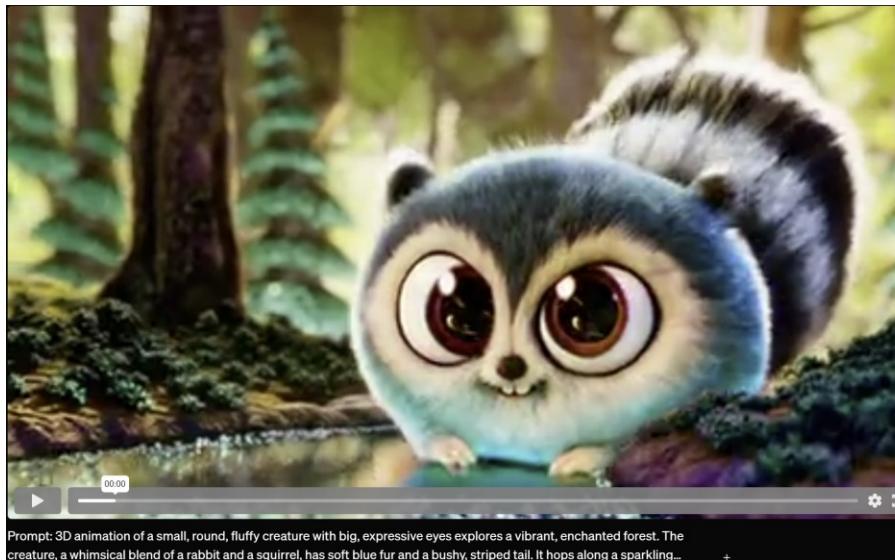
数据分析+
逻辑推理



数据科学基础

□ 大数据的未来—多模态大模型

- SORA: <https://openai.com/index/sora/>
- 智象未来: <https://hidreamai.com/home>



Prompt: 3D animation of a small, round, fluffy creature with big, expressive eyes explores a vibrant, enchanted forest. The creature, a whimsical blend of a rabbit and a squirrel, has soft blue fur and a bushy, striped tail. It hops along a sparkling...



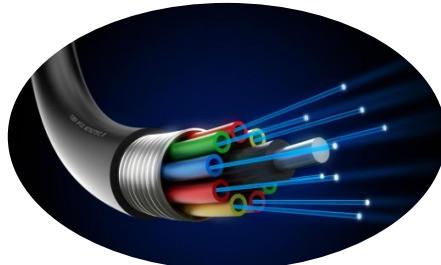


数据科学基础

89



存储（如硬盘、数据库）



收集、传输

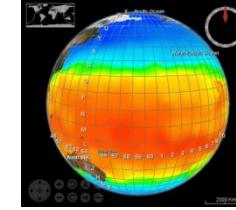


生产、记录



$$\min_X f(X) + \lambda \cdot \text{rank}(X)$$

分析、挖掘和学习



可视化



数据安全与个人隐私



计算（平台与架构等）



基本程序与算法

.....



课程章节及学时分配（计划）

90

- 课程共计18周（1-18周，约18次课）
 - 数据科学基础，第1-2次课
 - 数据分析入门，第3-6次课
 - 数据处理工具与实验基础（Python），第7次课
 - 数据统计基础，第8-10次课
 - 数据挖掘与机器学习基础，第11-14次课
 - 数据挖掘前沿专题，第15-17次课
 - 信息检索与推荐系统
 - 人工智能前沿专题
 - 大数据典型应用
 - etc
 - 课程汇报与课程回顾，第18次课