



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/DS2025.html>

助教：于峻浩，程程

data_science_2025@163.com

9/22/2025

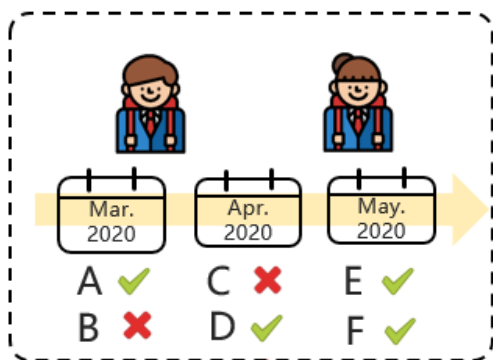


数据科学基础

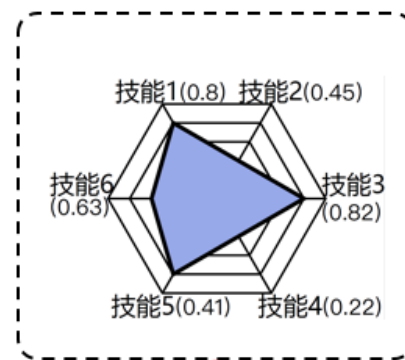
44

- 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 数据驱动的模式
 - 规模化因材施教：千人千面的自适应学习

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

45

- 数据蕴含着巨大的价值—智慧教育
 - 考试试题质量与公平性



浙江省教育考试院
ZHEJIANG EDUCATION EXAMINATIONS AUTHORITY

组织机构 信息公开 政策法规 政策解读 2018年11月27日 星期二 11:11:39 请输入关键字

普通高考 | 学考选考 | 研究生考试 | 成人高考 | 自学考试 | 社会考试 | 教师资格考试 | 海外考试

关于英语科目考试成绩的说明

[发布时间:2018-11-27 阅读量:1570]

浙江省高考英语科目一年安排2次考试，考生可报考2次，选用其中较高1次的成绩。在2018年11月刚结束的英语科目考试中，根据答卷试评情况，发现部分试题与去年同期相比难度较大。为保证不同次考试之间的试题难度大体相当，浙江省招委组织专家研究论证，在制订评分细则时，决定面向所有考生，对难度较大的第二部分（阅读理解）、第三部分（语言运用）的部分试题进行难度系数调整，实施加权赋分。其他试题未作调整。

浙江省教育考试院

2018年11月27日



数据科学基础

46

- 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

大模型学习助手

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 $[X*Y]$ 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 $[X*Y]$ 的补码：

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

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： $[X*Y]_{补} = 1.01111100$

因此，根据布斯算法， $[X*Y]$ 的补码为1.01111100。



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

47

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
 - 2021-2025年，实验室保研/考研机试





数据科学基础

48

- 数据蕴含着巨大的价值——社会科学
 - ◆ 社交媒体 比 问卷调查 提供了更有代表性的结果
 - ◆ 智能引导社会成员的行为



15万名奥巴马支持者在Facebook安装了“奥巴马2012”应用，而通过这个程序，总统竞选团队可以间接得到这些支持者数百万的Facebook好友信息。



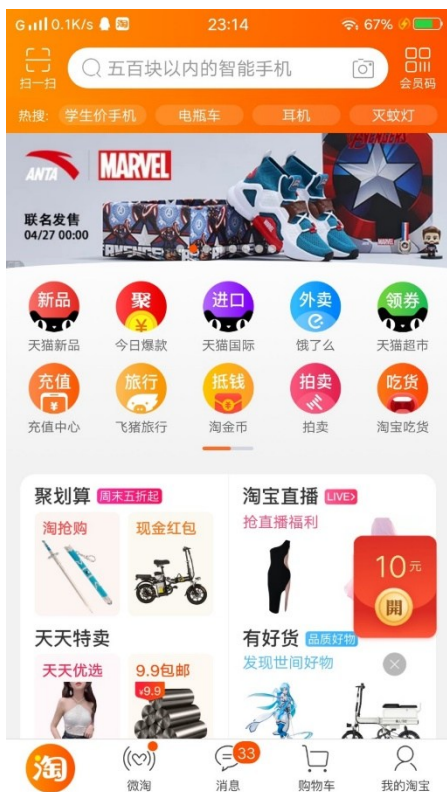
有一种说法称，特朗普的团队聘用数据分析公司，做了精准的广告投放，影响了那些徘徊不定的选民，拿下了决定性的关键州选举人票



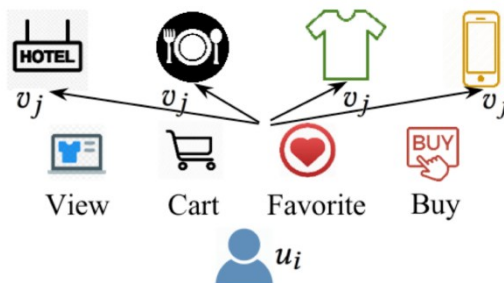


数据科学基础

数据蕴含着巨大的价值—电子商务 计算广告



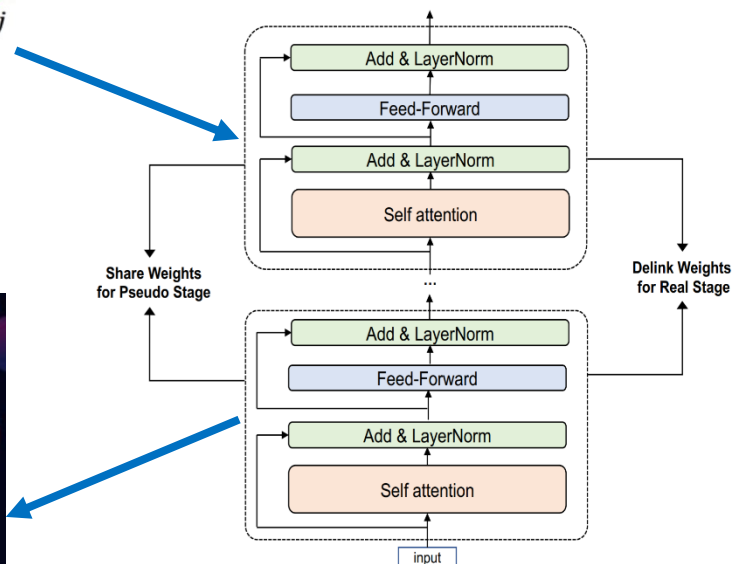
电商平台



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



数据科学基础

50

- 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

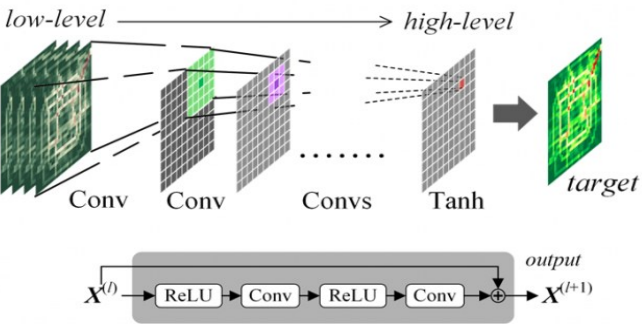
车辆移动数据



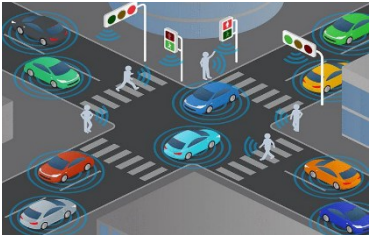
行程规划



交通流预测模型



信号灯调控



道路规划



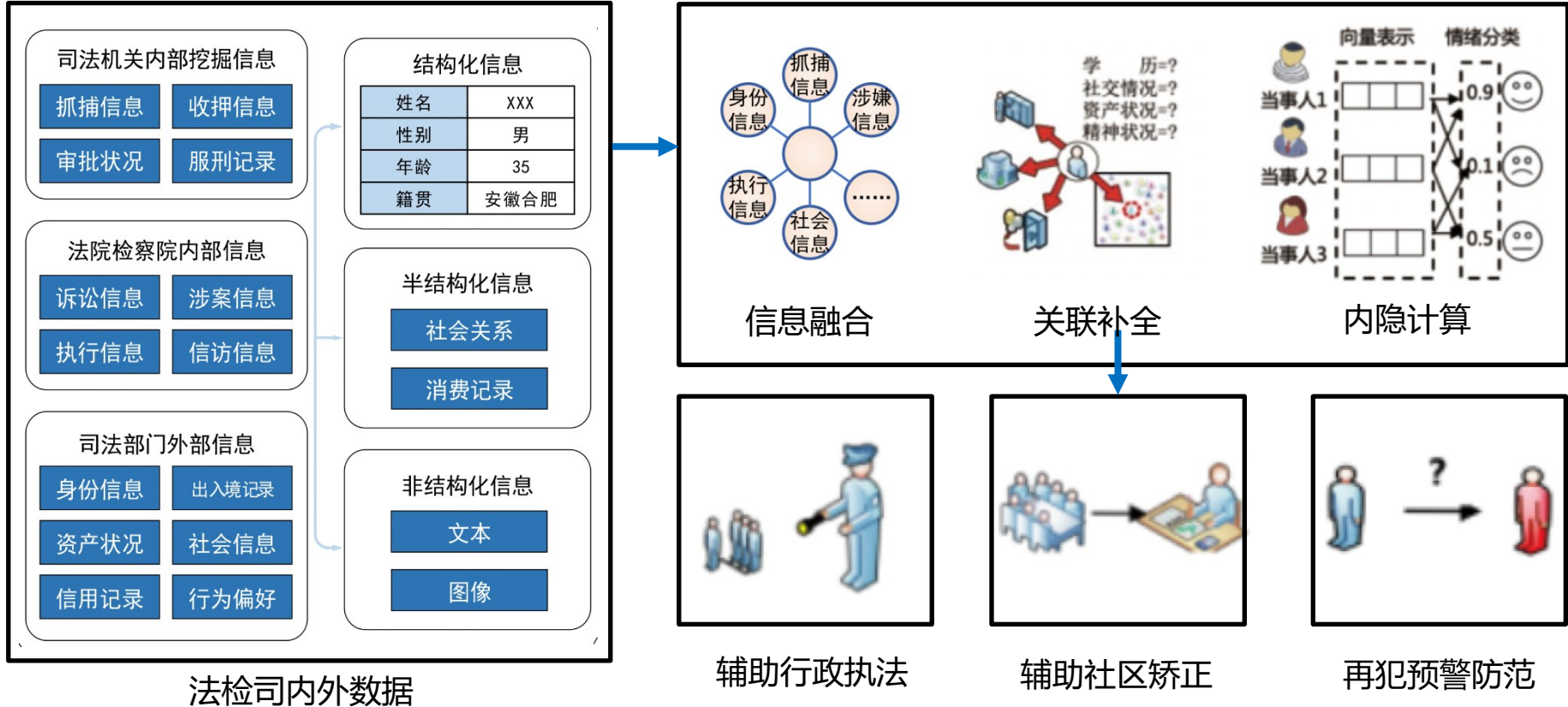


数据科学基础

51

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

52

- 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

53

- 数据蕴含着巨大的价值—文化创作
 - 机器作诗



用户写作意图



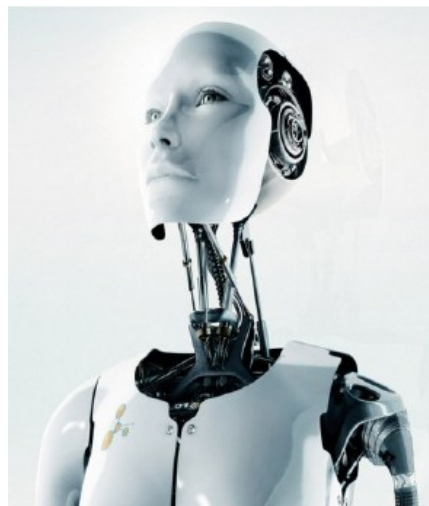
江、船、秋风



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

54

□ 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。

机器

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。

宋代诗人葛绍体



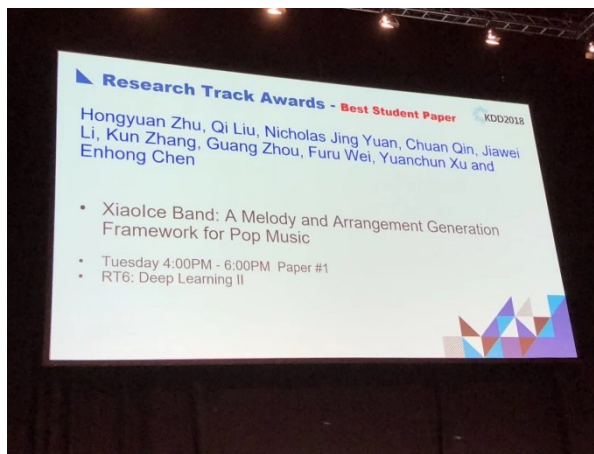
数据科学基础

55

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>





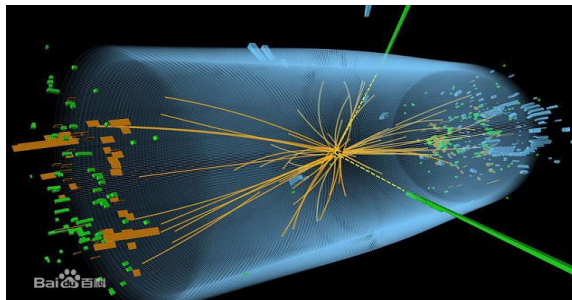
数据科学基础

56

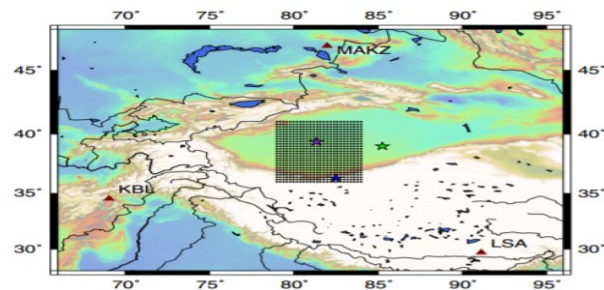
- 数据蕴含着巨大的价值—科学技术
 - ◆ 大数据推动科学新技术发现



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

57

- **科技大数据来自于物理世界**
 - 科学实验数据或传感数据
 - 技术描述型数据—专利、论文
- **集多种特点于一身**
 - 采集的高代价性
 - 复杂性
 - 超高维度
 - 高度计算复杂性
 - 高度的不确定性
 - 学科知识壁垒
 - 信息与通信技术高度集成性

单一学科



数据驱动

多学科交叉



关系型数据库的鼻祖Jim Gray (右)





数据科学基础

58

■ 2007年，Jim Gray总结出了四个科学范式

几千年前

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

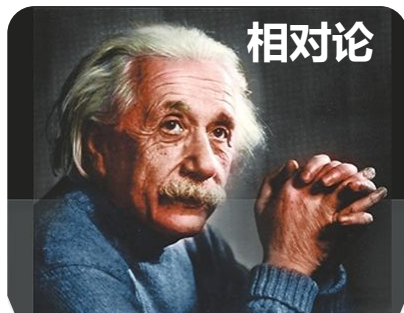
几百年前

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



相对论

几十年前

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

今天

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



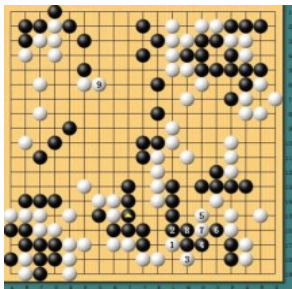


数据科学基础

□ 大数据研究—数据与知识融合，让人工智能更“聪明”

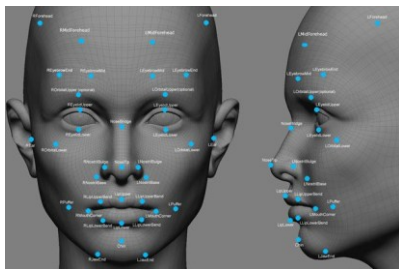
计算智能

- 规则明确、特定领域



感知智能

- 语音、图像、视频



认知智能

- 理解、推理、解释





数据科学基础

60

大数据研究—语音大数据

语音识别

- 微软英语语音识别实现词错率5.9%的突破，第一次超越人类
- 科大讯飞语音识别词错率3%左右。中文领域突出





□ 大数据研究—图片大数据

- ImageNet图像数据库上，人工智能已达到2.99%的错误率（公安部三所），低于人类5.1%的错误率



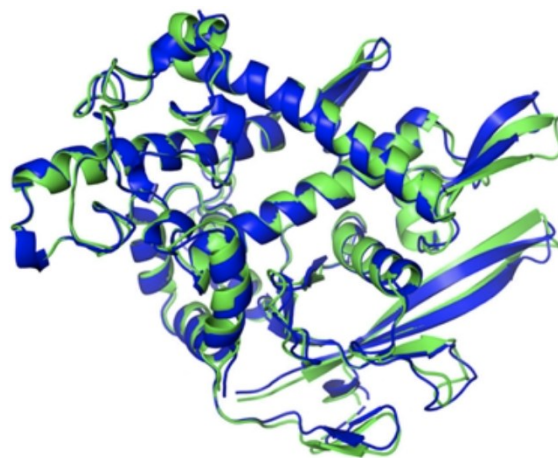
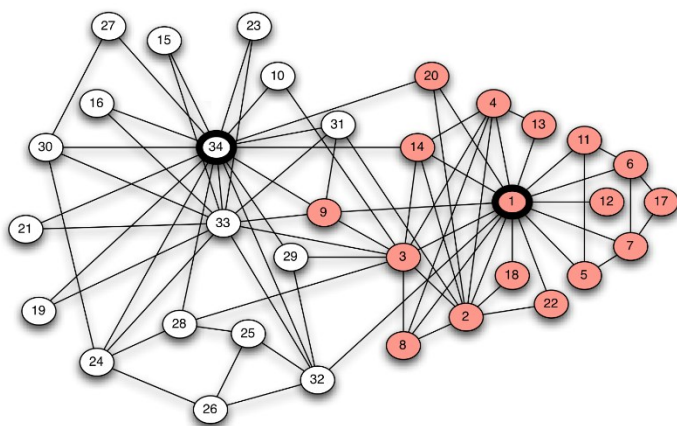
李飞飞
斯坦福大学、谷歌AI前任首席科学家



数据科学基础

62

- 大数据研究—网络大数据
 - 社会网络分析
 - 蛋白质结构图
 - 知识图谱





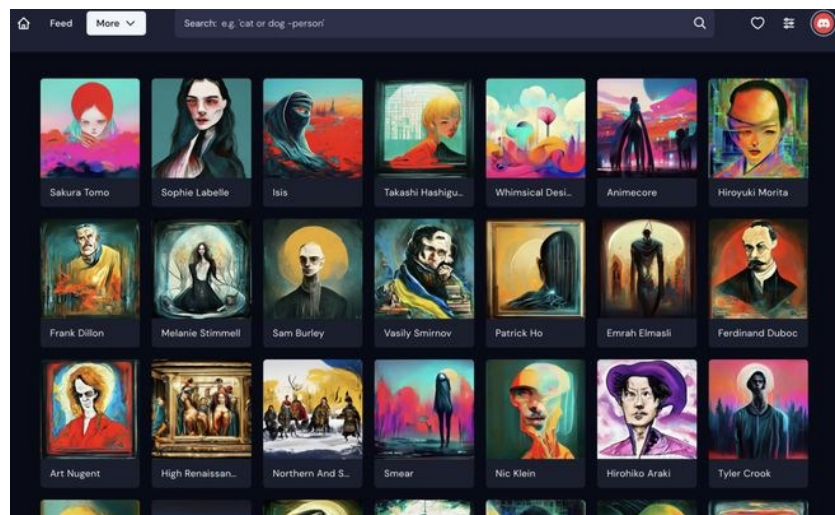
数据科学基础

63

□ 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

64

大数据带来的技术创新-当前热点

微表情生成与识别

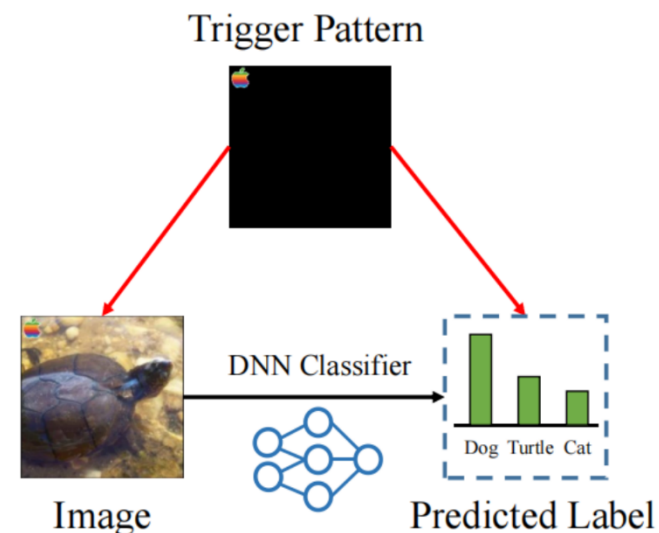




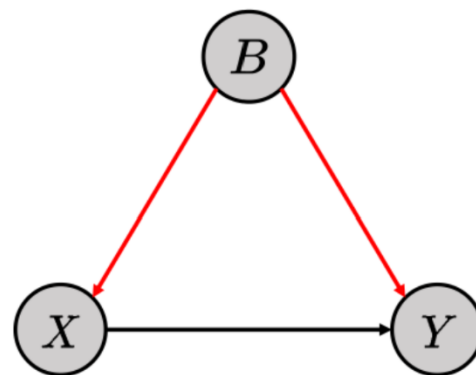
数据科学基础

65

- 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



(a) Illustration of Backdoor Attack



(b) Causal Graph $\mathcal{G}$



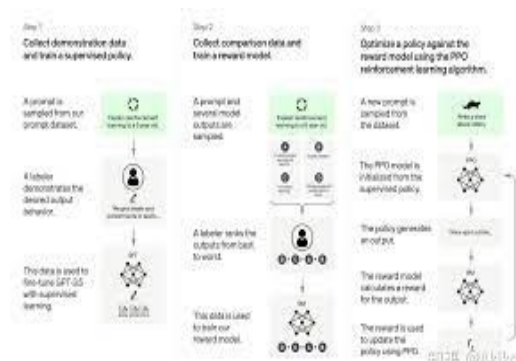
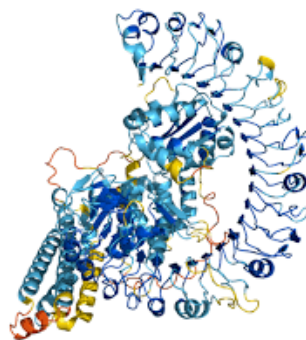
数据科学基础

66

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

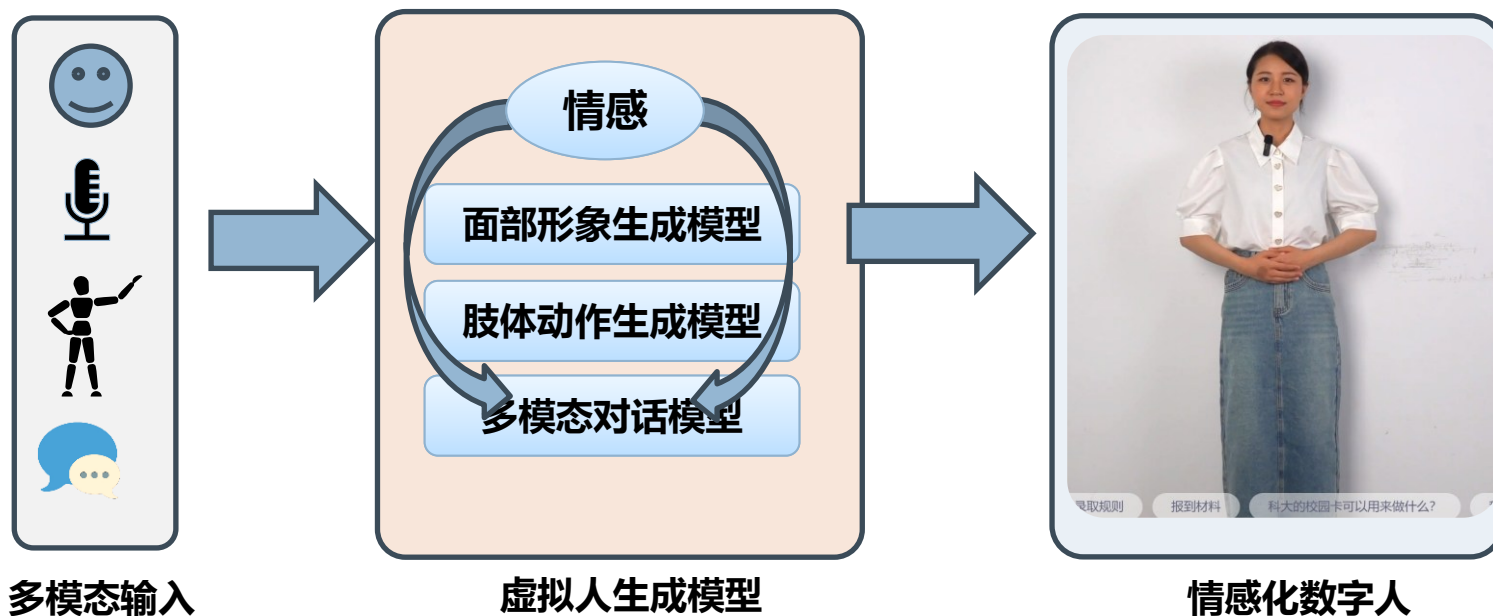




数据科学基础

67

- 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





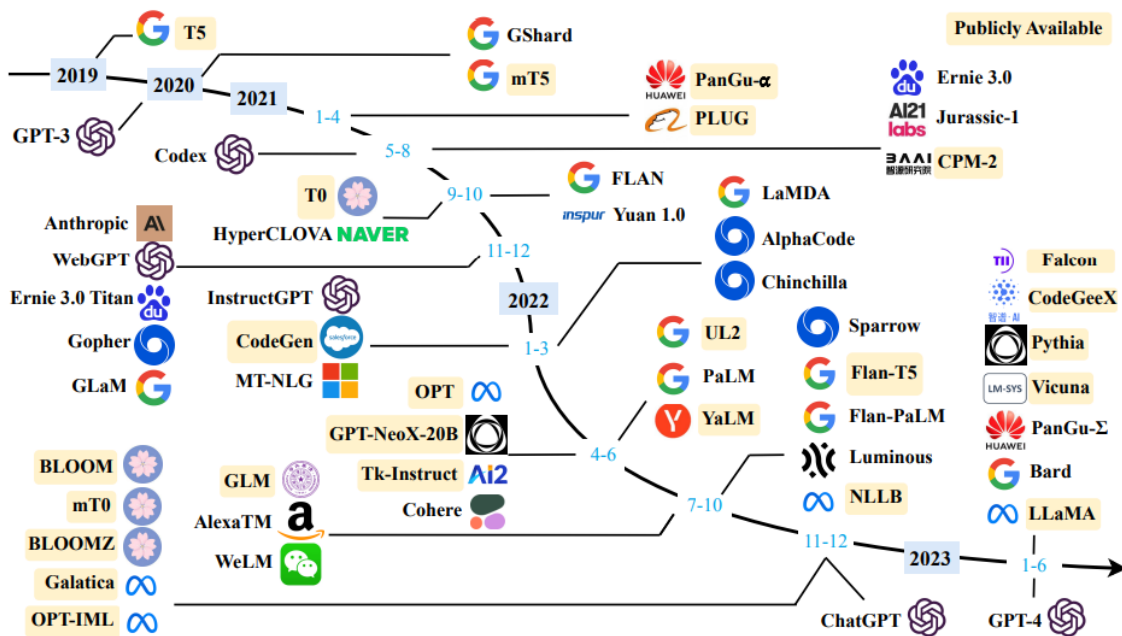
数据科学基础

68

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

69

- 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。

ChatGPT作为知识库回答问题

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仅通过任务描述完成文本情感分类



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——逻辑推理
 - 通过思维链（CoT）解决复杂问题

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

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

➤ L1: 基础对话者

- 对话与交互能力: 当前(如GPT-4),能流畅的对话
- 特点: 以对话为核心, 提供信息、解答问题、辅助创作等

➤ L2: 推理者

- 推理与问题解决能力: OpenAI-O1接近L2。 PhD-Level。
- 特点: 像人类一样分析问题、制定方案, 甚至在某些领域超越人类

➤ L3: 智能行动者

- 自主决策能力: AI能独立思考并根据复杂情境采取行动(接近人类)
- 特点: 不仅能够思考, 还能在真实世界中执行决策, 实现人机协同

➤ L4: 创新者

- 创造能力: AI能进行创造性思维和协助人类进行发明和创造, 推动科技进步
- 特点: 具备创新思维和创造力, 能够提出新的想法和解决方案

➤ L5: 组织者

- AGI: AI能执行和组织人类所有工作, 标志着真正的人工通用智能的实现
- 特点: 具备人类智能和组织能力, 能管理和优化整个社会的运行。

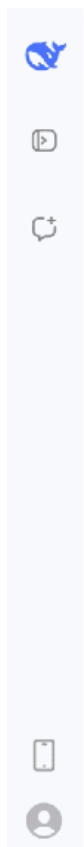
OpenAI's 5 Step to AGI

- | | |
|---------|---|
| Level 1 | Chatbots, AI with conversational language |
| Level 2 | Reasoners, human-level problem solving |
| Level 3 | Agents, systems that can take actions |
| Level 4 | Innovators, AI that can aid in invention |
| Level 5 | Organizations, AI that can do the work of an organization |



数据科学基础

□ 大数据的未来—多模态大模型



我是 DeepSeek, 很高兴见到你!

我可以帮你写代码、读文件、写作各种创意内容, 请把你的任务交给我吧~

给 DeepSeek 发送消息



DeepSeek



联网搜索

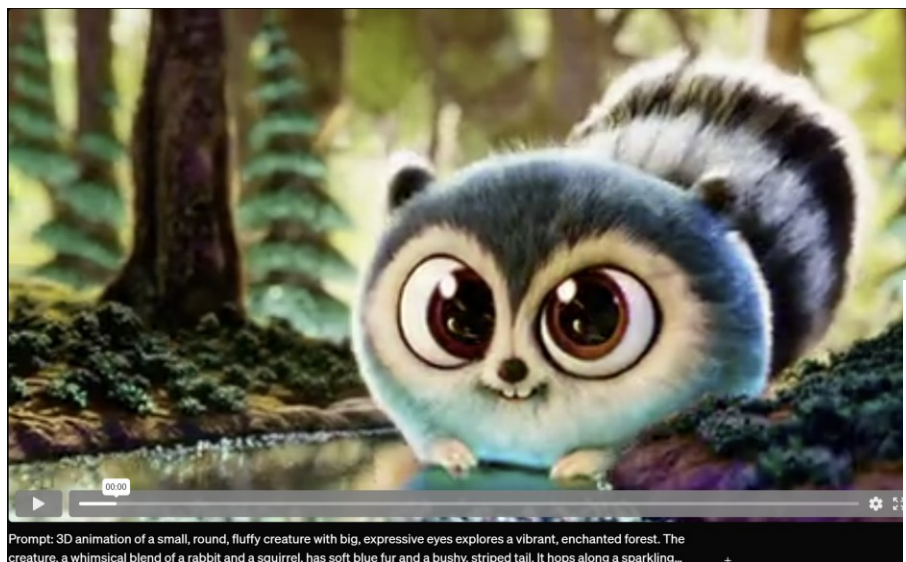




数据科学基础

□ 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>





数据科学基础

□ 大数据的研究—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

$$(a \vee b) \vee c = a \vee (b \vee c) \quad \text{重推理}$$
$$(a \wedge b) \wedge c = a \wedge (b \wedge c)$$

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

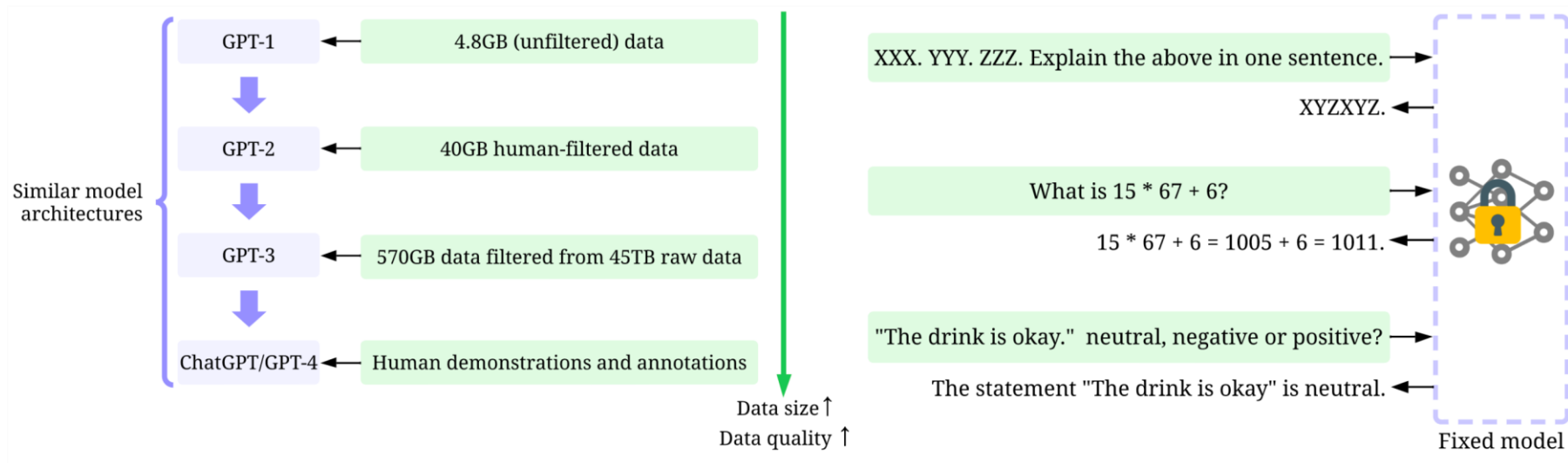
数据分析+
逻辑推理



数据科学基础

75

- 人工智能逐渐从以模型为中心过渡到以数据为中心
 - **GPT成功依赖数据基石：** 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

76

□ 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

Algorithm

AI

算力

Computation

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。



数据科学基础

77

□ 大数据与人工智能的关系

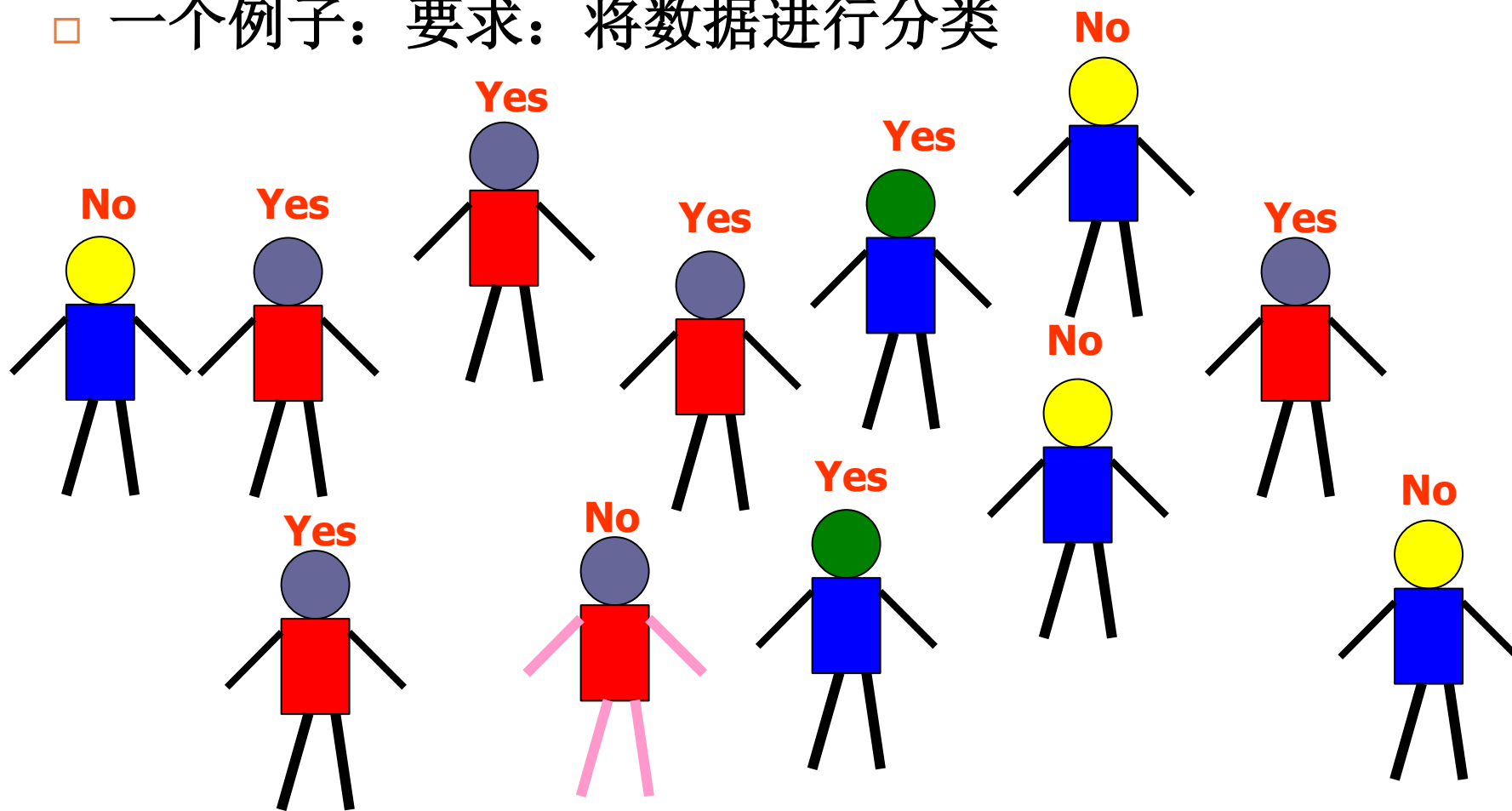
- 现阶段，人工智能的核心是对大数据进行的**特征抽取**与**机器学习算法**





数据+分类学习的方法

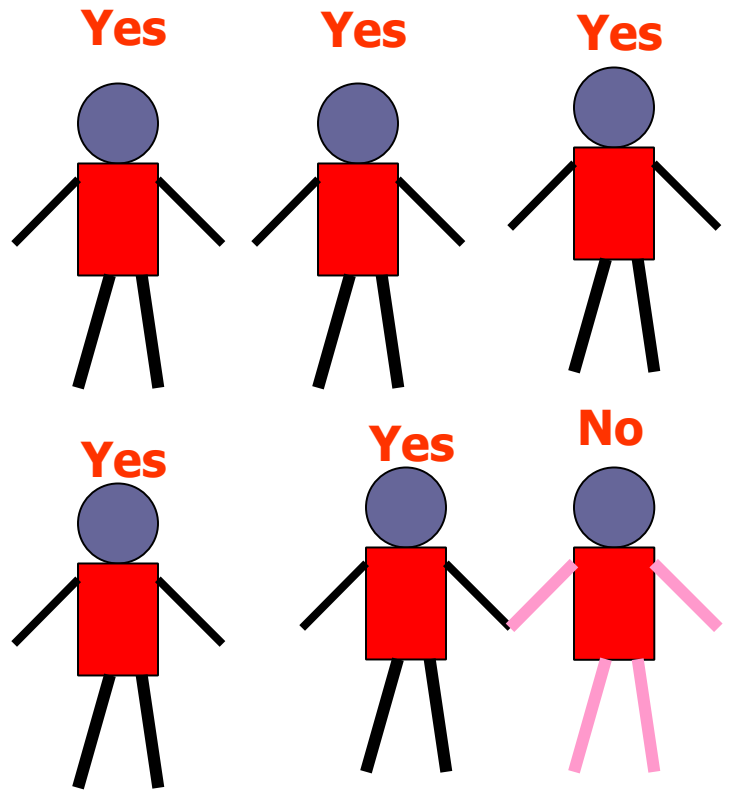
□ 一个例子：要求：将数据进行分类



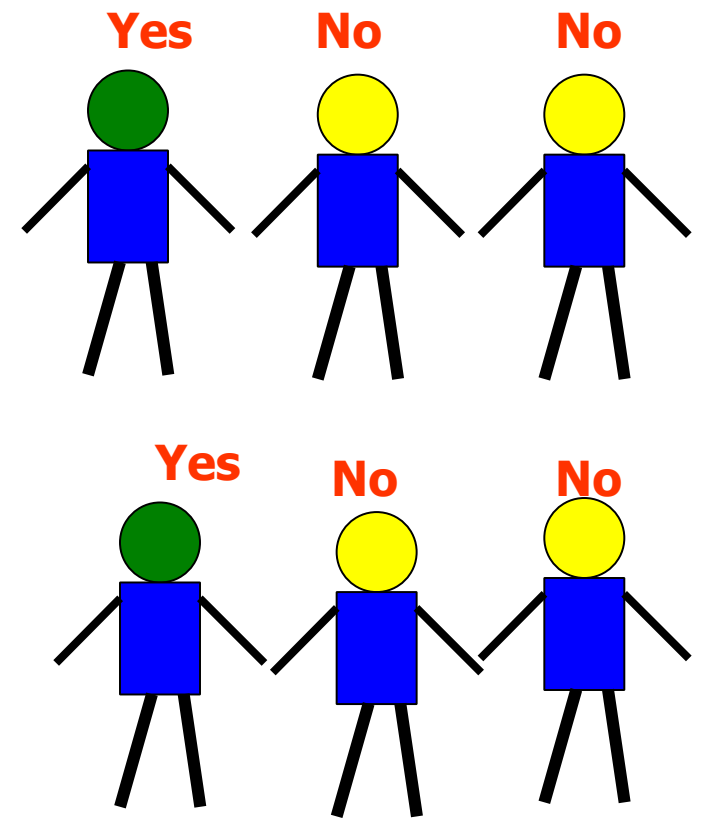


数据+分类学习的方法

躯干：红色



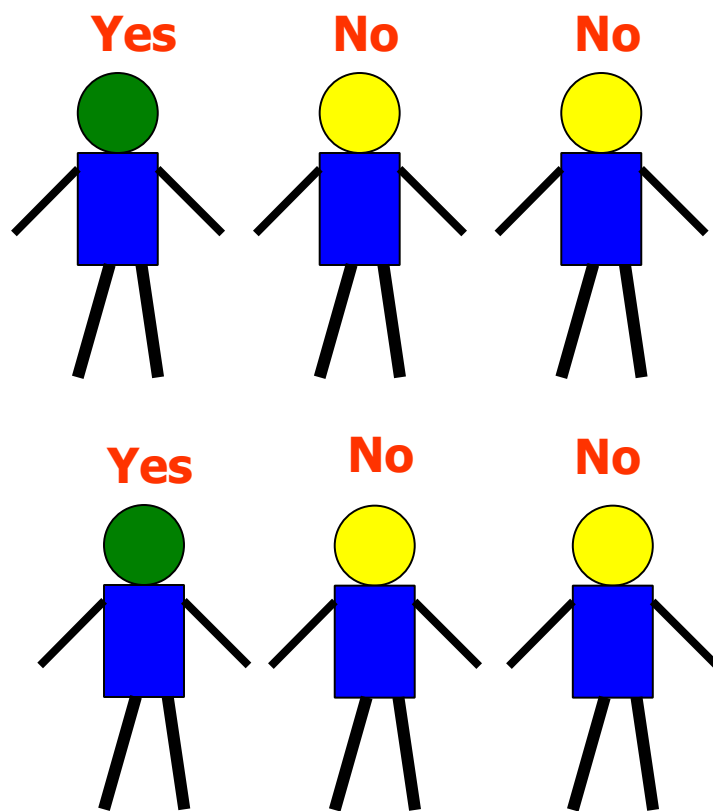
躯干：蓝色





数据+分类学习的方法

躯干：蓝色



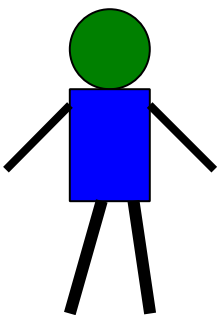


数据+分类学习的方法

躯干：蓝色

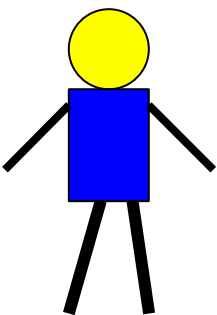
头：绿色

Yes

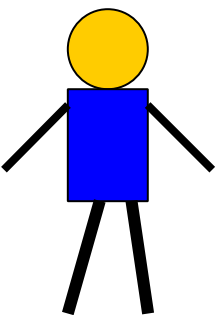


头：黄色

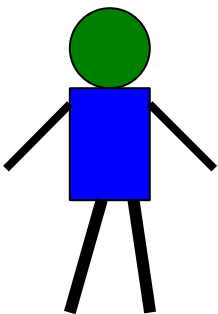
No



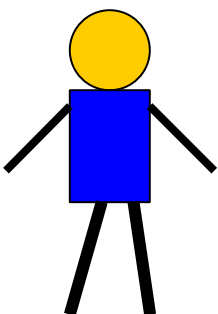
No



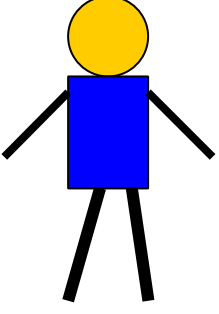
Yes



No



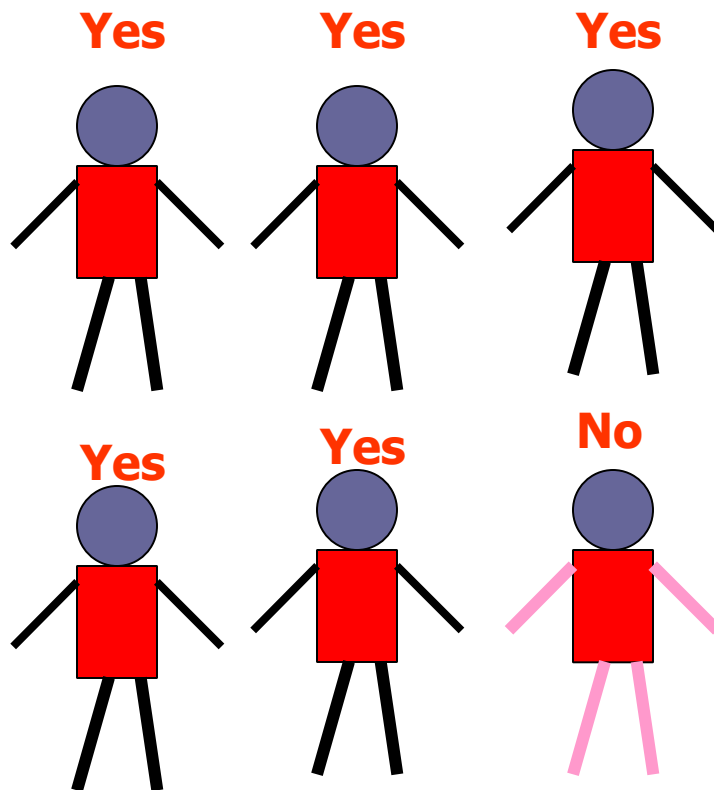
No





数据+分类学习的方法

躯干：红色



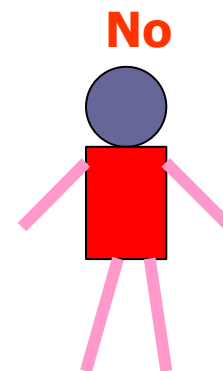
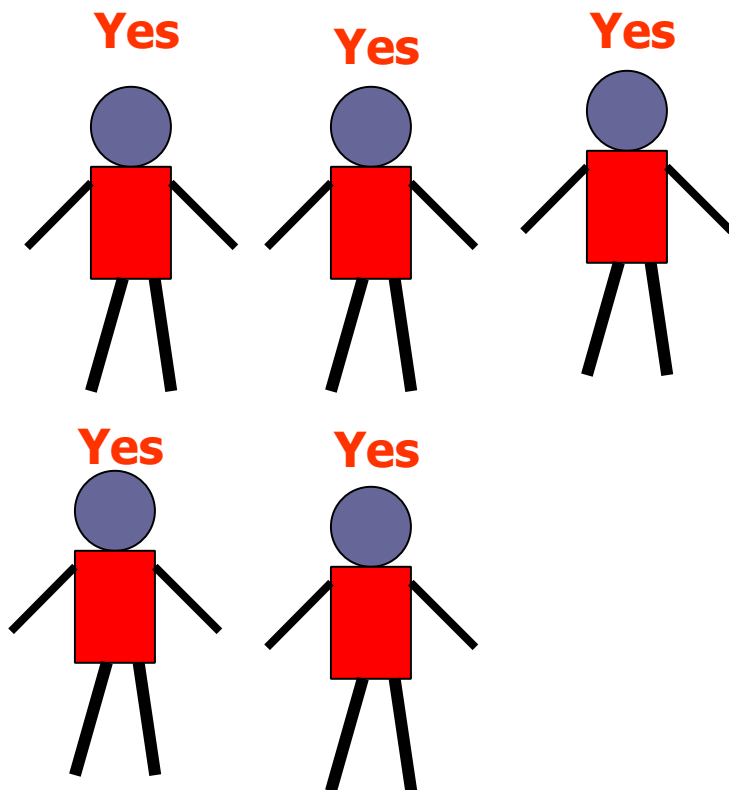


数据+分类学习的方法

躯干：红色

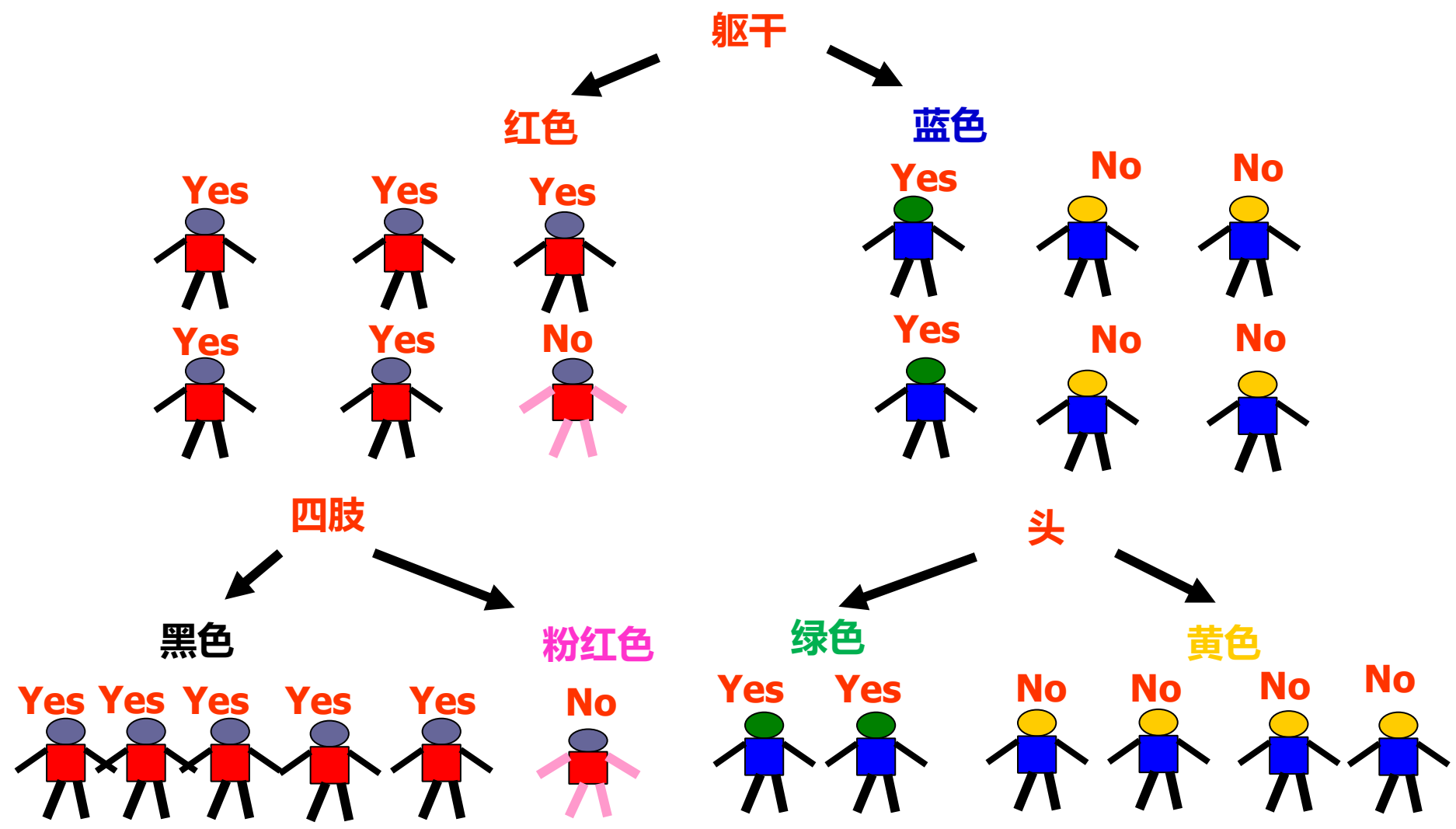
四肢：黑色

四肢：粉红色





数据+分类学习的方法



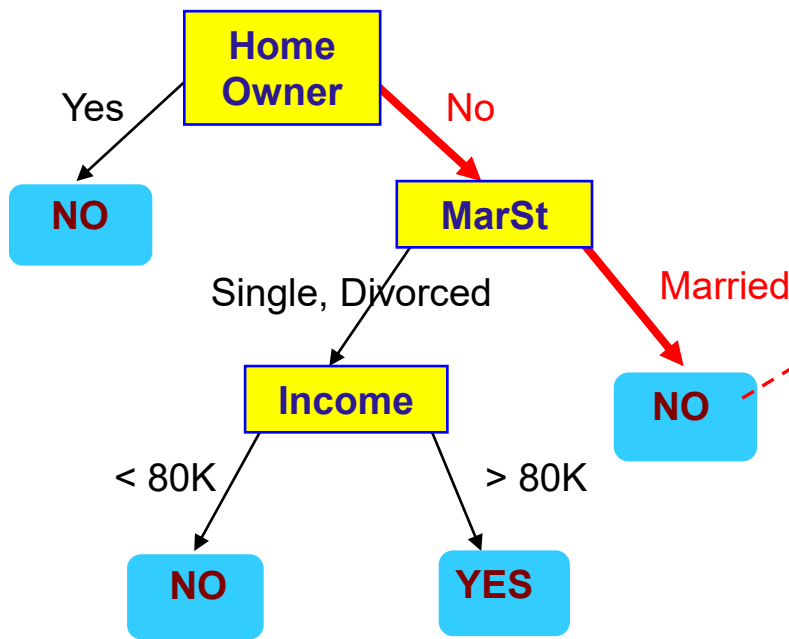


数据+分类学习的方法

□ 决策树——使用模型对测试数据分类

Test Data

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



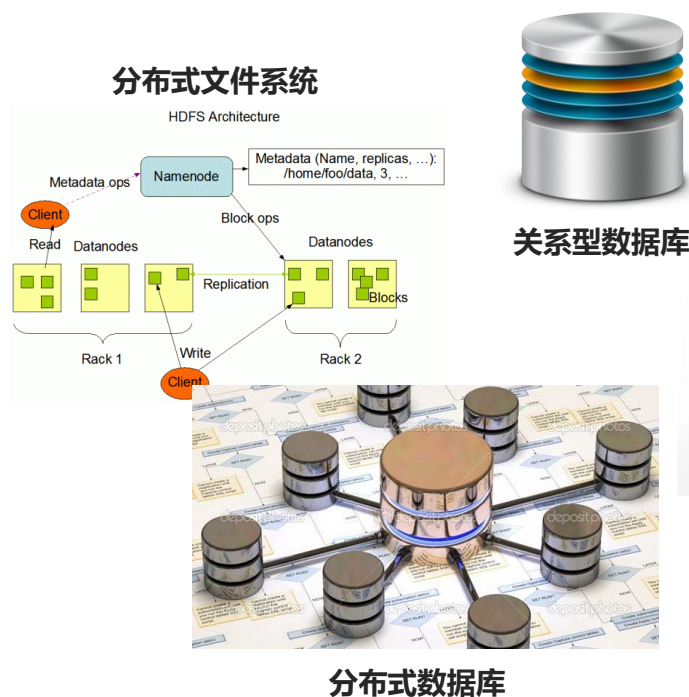
Assign Defaulted to
"No"



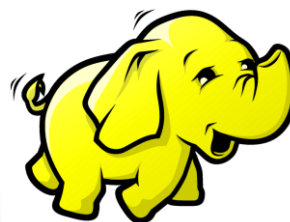
数据科学基础

86

- 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



APACHE
STORM™



9/22/2025



数据科学基础

87

- 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

AMD

- 云计算: EPYC (霄龙) 处理器; Project 47服务器

NVIDIA

- 新一代处理器架构**VOLTA**: 新一代NVIDIA NVLink高速互联技术
- 云计算: Tesla V100 GPU 加速器; DGX-2 全球最大GPU; GPU云平台
- 机器人: Jetson Xavier 机器人专用AI芯片
- GPU工作站: 基于Volta架构的GV100
- 自动驾驶: Drive Xavier 首个自动驾驶处理器

intel

- 云计算: 至强可扩展处理Purley; Xeon+FPGA (云端/设备端低功耗性能计算); Xeon Phi+Nervana (云端高性能计算)
- 无人驾驶: EyeQ 4/EyeQ 5 SoC
- 边缘计算: Myriad X VPU. Movidius Myriad X 视觉处理单元 (VPU), 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

XILINX

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

arm

- CPU架构: Cortex-A76
- GPU架构: Mali G76

Google

- 云计算: TPU 3.0; Cloud TPU
- 移动端: Pixel VisualCore

Qualcomm

- 移动端: 骁龙855处理器
- 智能驾驶: C-V2X芯片组



- 移动端: A12 芯片



- 移动端: 麒麟980芯片



- 跨界处理器: i.MX RT1060



数据科学基础

88

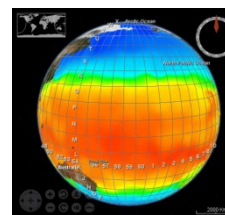


存储（如硬盘、数据库）

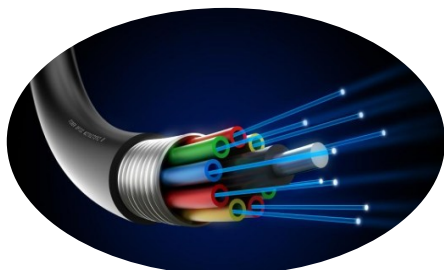


$$\min_X f(X) + \lambda \cdot \text{rank}(X)$$

分析、挖掘和学习



可视化



收集、传输



数据安全与个人隐私



生产、记录



基本程序与算法

.....



计算（平台与架构等）



课程章节及学时分配（计划）

89

- 课程共计18周（1-18周，约18次课）
 - 数据科学基础，2周
 - 数据分析，4周
 - 数据统计，3周
 - 数据挖掘与机器学习，4周
 - 数据挖掘前沿专题，4周
 - 信息检索与推荐系统
 - 人工智能技术与应用
 - 自然语言处理与大模型技术前沿
 - etc
 - 课程汇报与课程回顾，1周



课程要求与考核方式

90

- 课程目标：用科学的方法研究和应用数据
- 课程要求，具体要求之后布置
 - 课堂出勤与作业
 - 文献调研报告
 - 实验与报告（需要编程）
 - 课程交流与课程汇报
- 考核方式
 - 课堂（30%）+调研（30%）+实验（40%）



联系方式

91

- 授课教师：陈恩红，黄振亚
- 课程主页：
<http://staff.ustc.edu.cn/~huangzhy/Course/DS2025.html>
- QQ群：931835645
- 联系方式
 - 助教：于峻浩，程程
 - data_science_2025@163.com
 - 教师：黄振亚，陈恩红
 - huangzhy@ustc.edu.cn



数据科学导论2025

群号: 931835645

