



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

1

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/DS2026.html>

助教: 罗胤玻, 李宁

ds_intro2026@163.com



数据科学基础

61

- 数据蕴含着巨大的价值—文化创作
 - 机器作诗



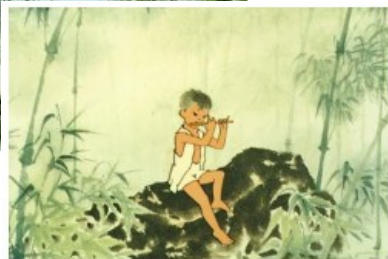
用户写作意图



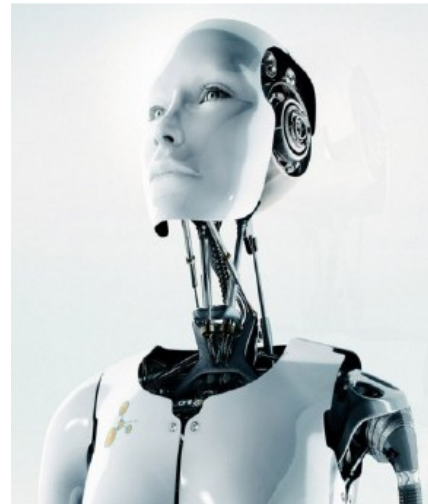
江、船、秋风



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。

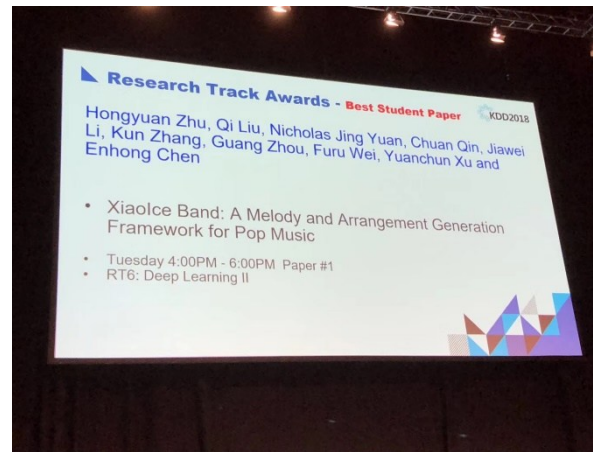
乔松古木两三间，
松竹阴中一径斜。
白鸟不知山路远，
牧童踏过野人家。



□ 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>

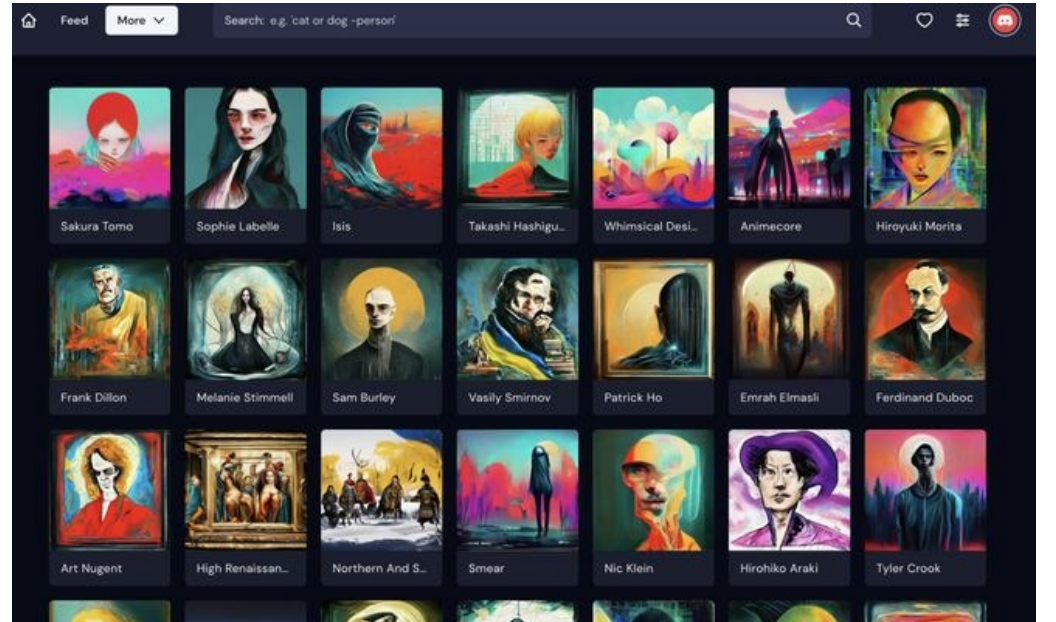




数据科学基础

63

- 数据蕴含着巨大的价值—文化娱乐
 - 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

- 数据蕴含着巨大的价值—文化娱乐
- AI短剧生成：海量视频数据帮助模型学习人物、场景、运镜和声画关系，输入文本、图片等内容生成具有角色动作、镜头切换和音效的短视频。
- 音频创作：AI学习歌手音色进行歌曲翻唱，也可以编曲作词，重新设计新的音乐



Seedance 2.5

具备 30 秒长叙事、精准参考与编辑能力的新一代视频生成模型



AI孙燕姿（不断更新中，建议收藏）

AI音乐 >

1666.2万 播放

更新至第48集





数据科学基础

65

■ 2007年，Jim Gray总结出了四个科学范式





数据科学基础

66

■ 2007年，Jim Gray总结出了四个科学范式





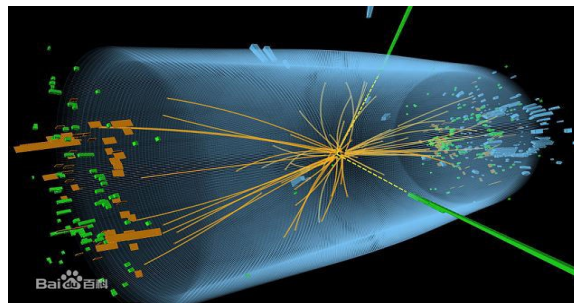
数据科学基础

67

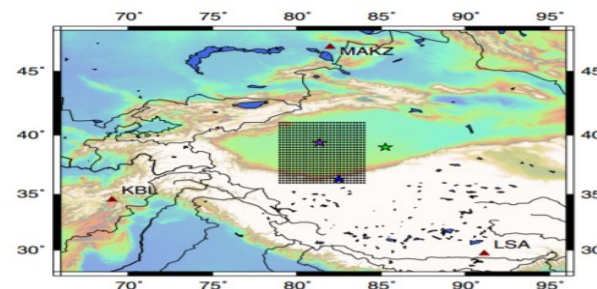
大数据推动科学新技术发现



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



数据挖掘保护知识产权



数据科学基础

68

- **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)





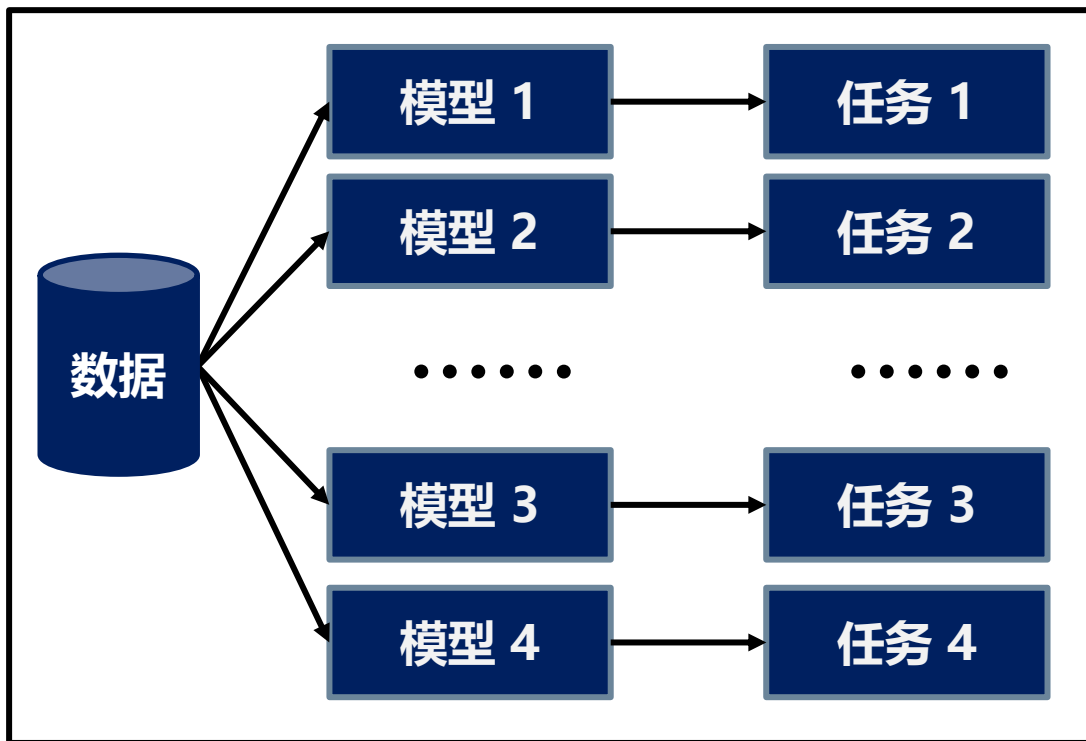
数据科学基础

69

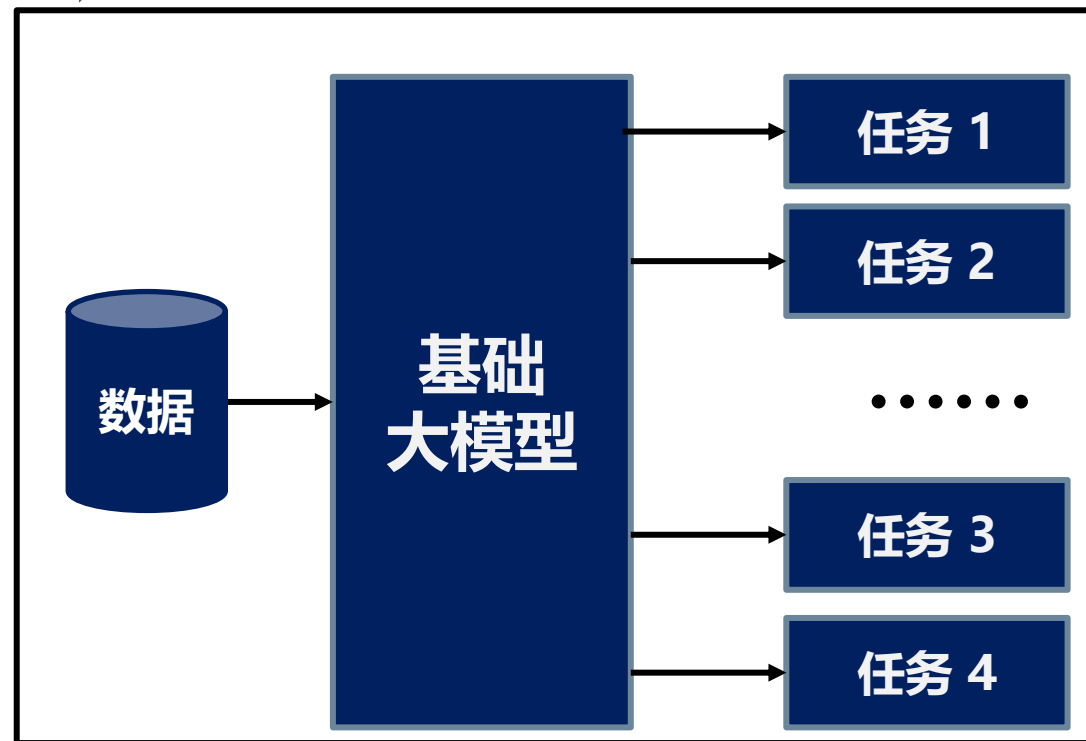
一任一模 “**小作坊**” 范式

具备任务泛化能力

千任一模 “**基座平台**” 范式



- ✓ 需要为**每个任务维护一个模型**
- ✓ 任务间参数**难以共享迁移**
- ✓ 稀疏标签样本场景**性能不佳**



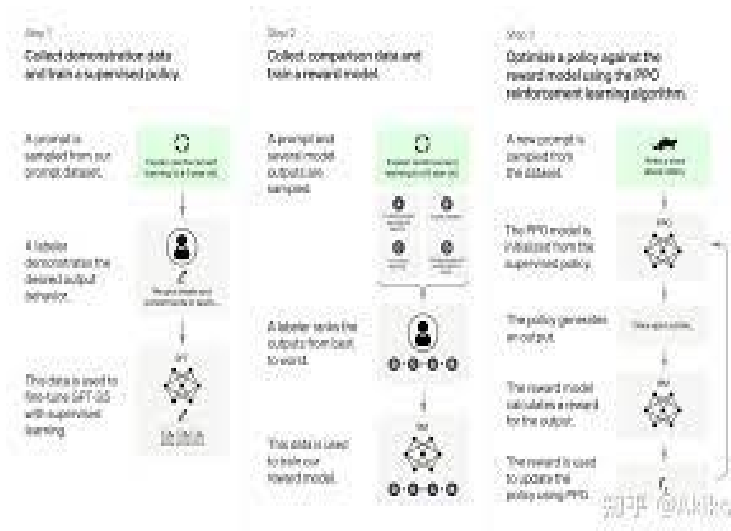
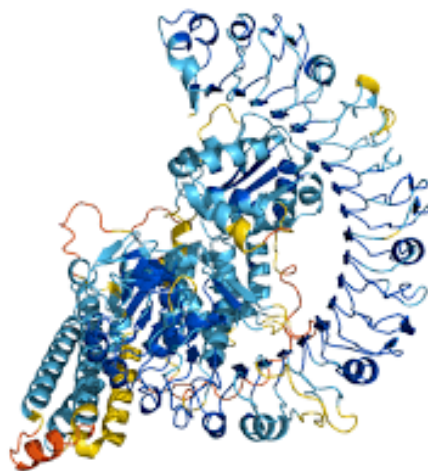
- ✓ 在基础模型之上**完成目标子任务**
- ✓ 任务间能够**共享基础模型参数**
- ✓ 基础模型能够**缓解稀疏样本窘境**



大数据推动科学新技术发现—强化学习

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
- 2025年，Deepseek创新GRPO算法

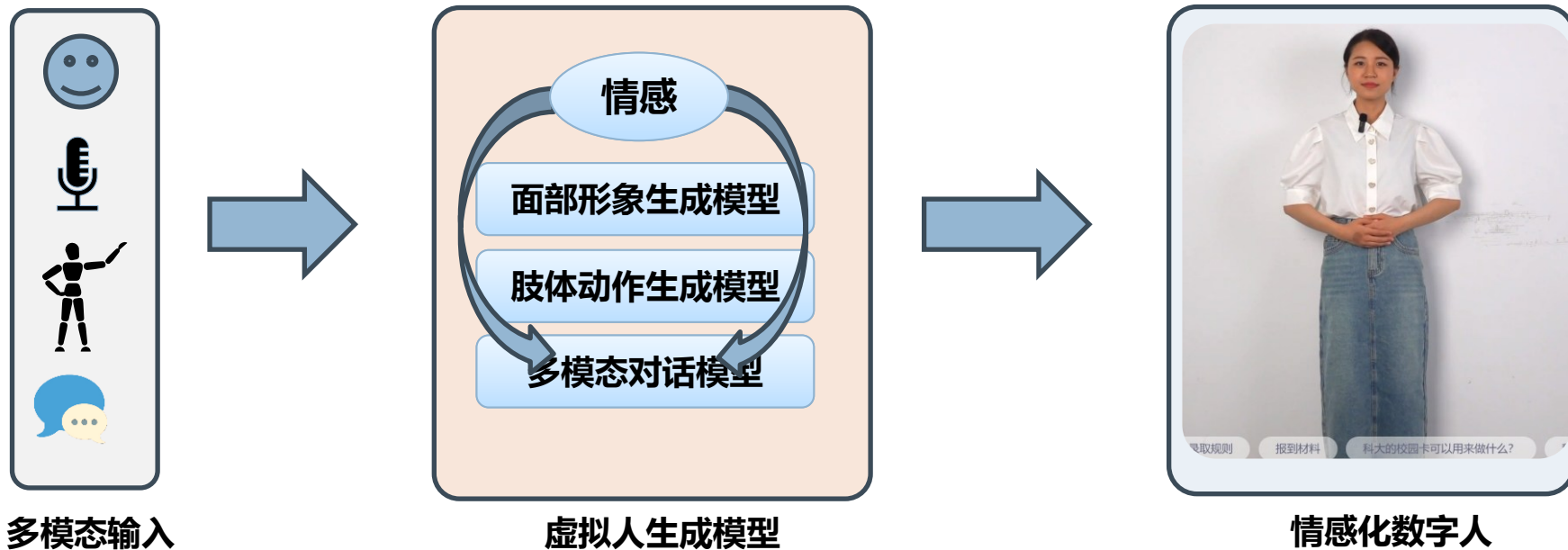




数据科学基础

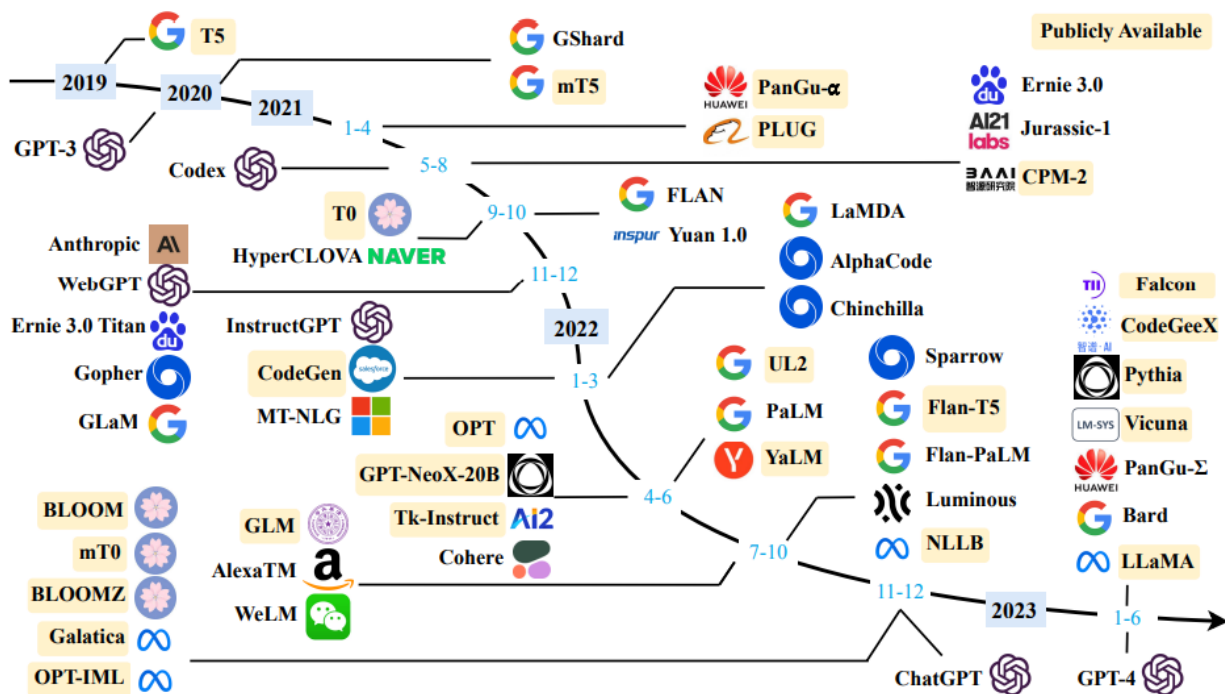
71

- 大数据推动科学新技术发现—数字人
 - 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





- 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

73

- 大数据推动科学新技术发现—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. ✅

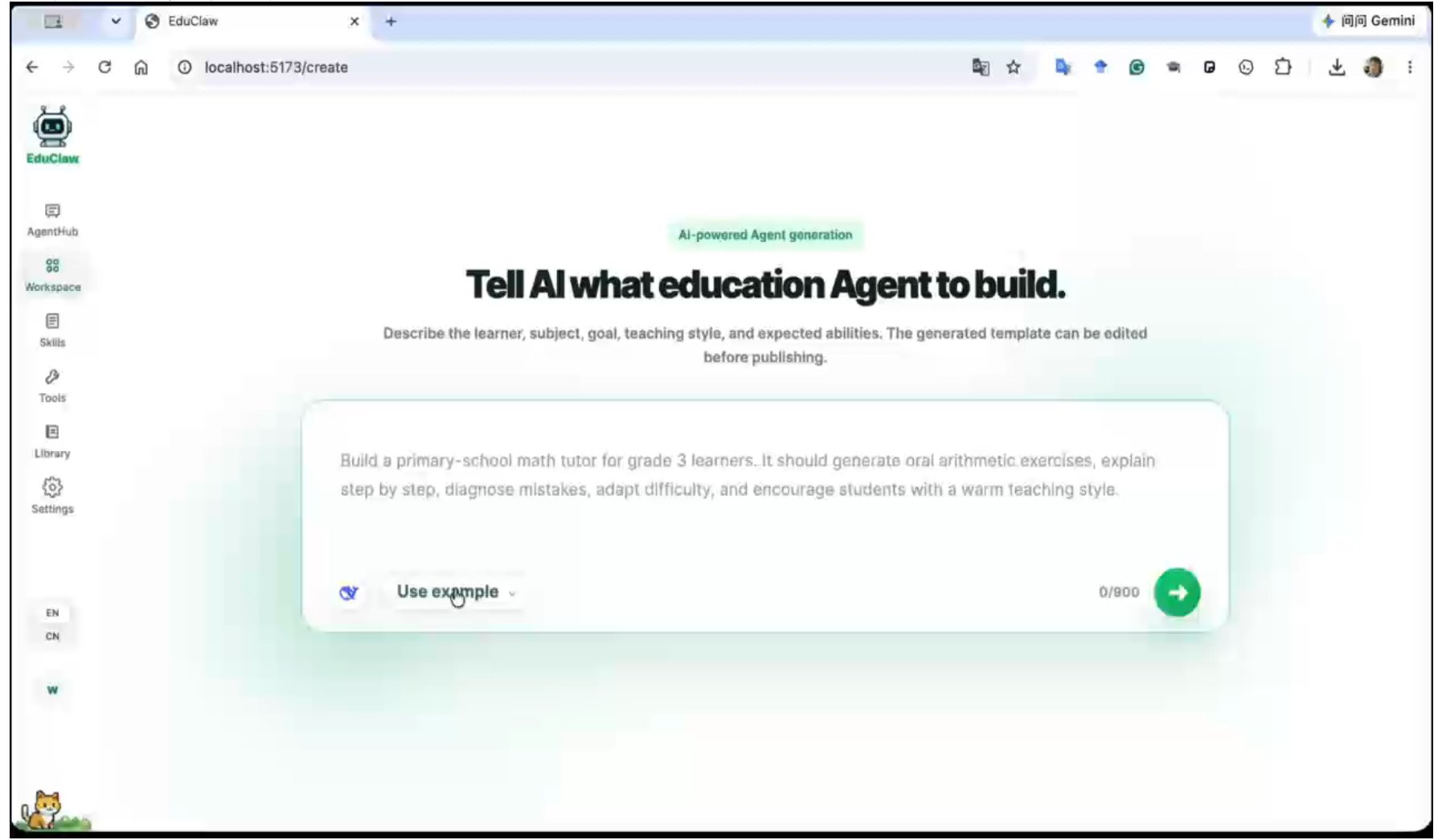
向大语言模型展示少量样本并解释推理过程，模型在回答时也会显示推理过程从而引出正确回答



数据科学基础

□ 大数据推动科学新技术发现—智能体

■ 复杂工具调用





数据科学基础

75

□ 数据蕴含着巨大的价值—AI4Science

中国科大机器化学家平台：采用机器智能去查找和阅读文献，从海量研究数据中汲取专家经验，调度2台移动机器人和15个自主开发的智能化学工作站，完成高通量合成、表征、测试的化学实验全流程



传统方法

传统的“炒菜式”遍历搜索方法效率低下，需要检查每一个可能的组合...

1400年

机器化学家

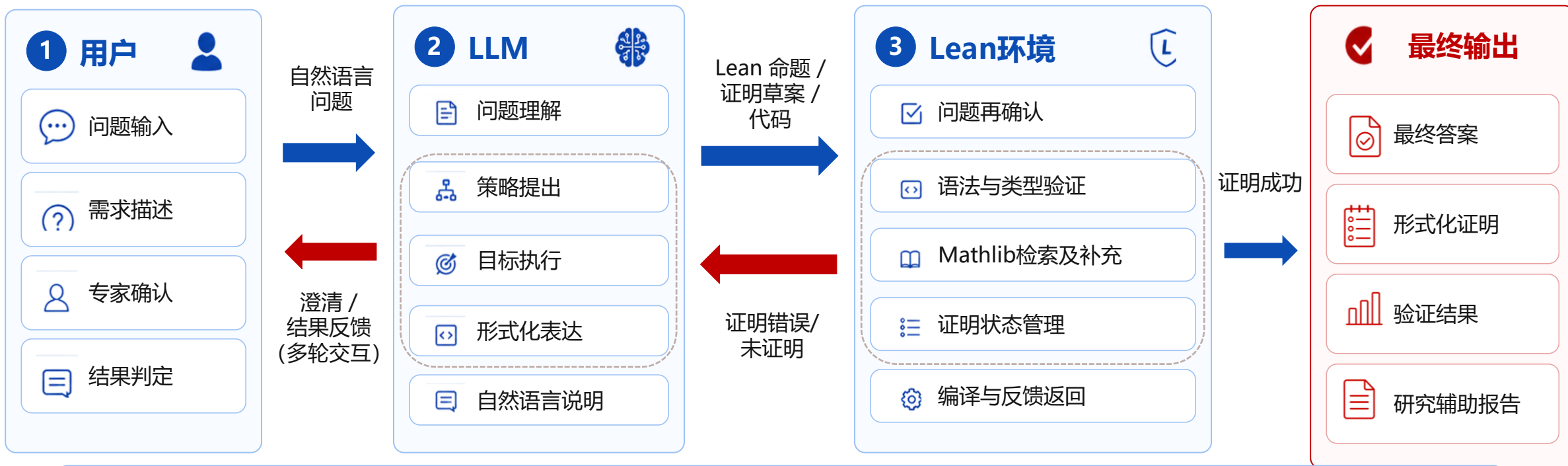
发挥其数据驱动和智能优化的优势。

5周

智能阅读16000篇论文并自主遴选出5种非贵金属元素，融合2万组理论计算数据和207组全流程机器实验数据，建立了理实交融的智能模型，指导贝叶斯优化程序从55万种可能的金属配比中找出最优的高熵催化剂



数据蕴含着巨大的价值—AI4Science



近期已完成的数学推理与形式化验证方面工作:

1. 非均匀网格上二维线性输运方程 P^k 元 upwind DG方法的最优误差估计定理
2. 非均匀网格上一维非线性守恒律 CDG方法最优误差估计定理
3. 非均匀网格上二维非线性守恒律 P^k 元和 Q^k 元 CDG 最优误差估计定理

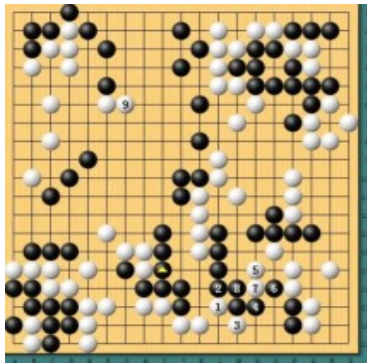


数据科学基础

- 大数据研究—数据与知识融合，让人工智能更“聪明”

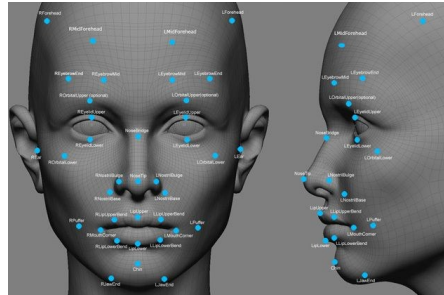
计算智能

- 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



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—AGI

➤ 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。

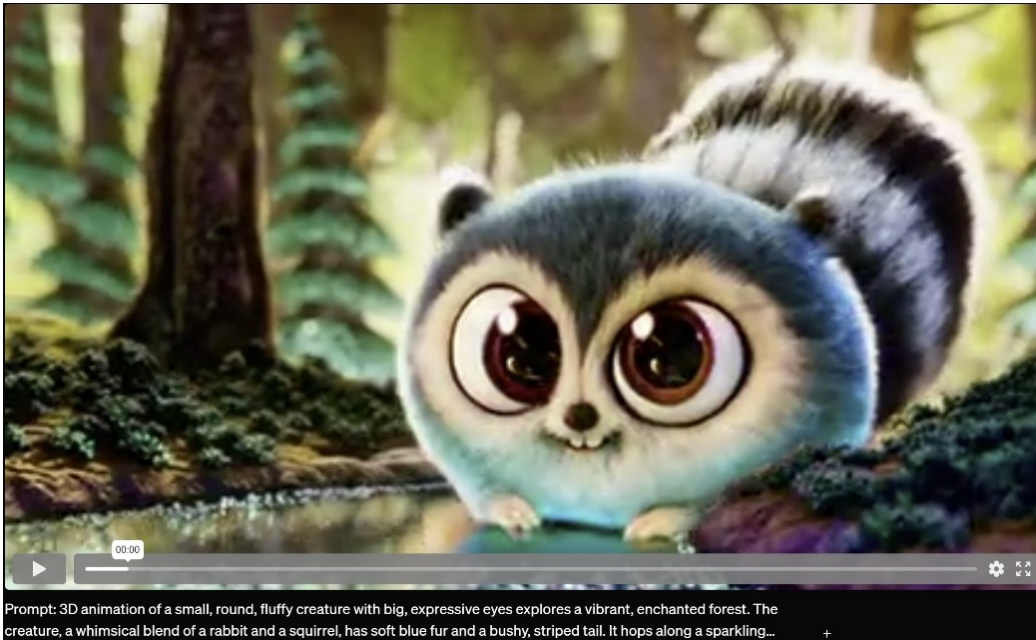
OpenAI's 5 Step to AGI

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





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

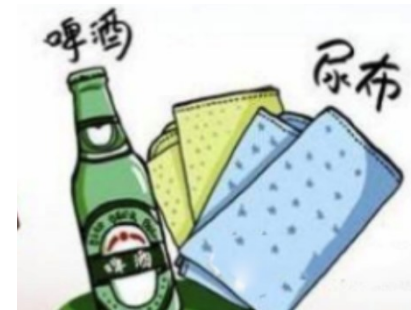
- 归纳法、数理逻辑、布尔代数系统

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

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



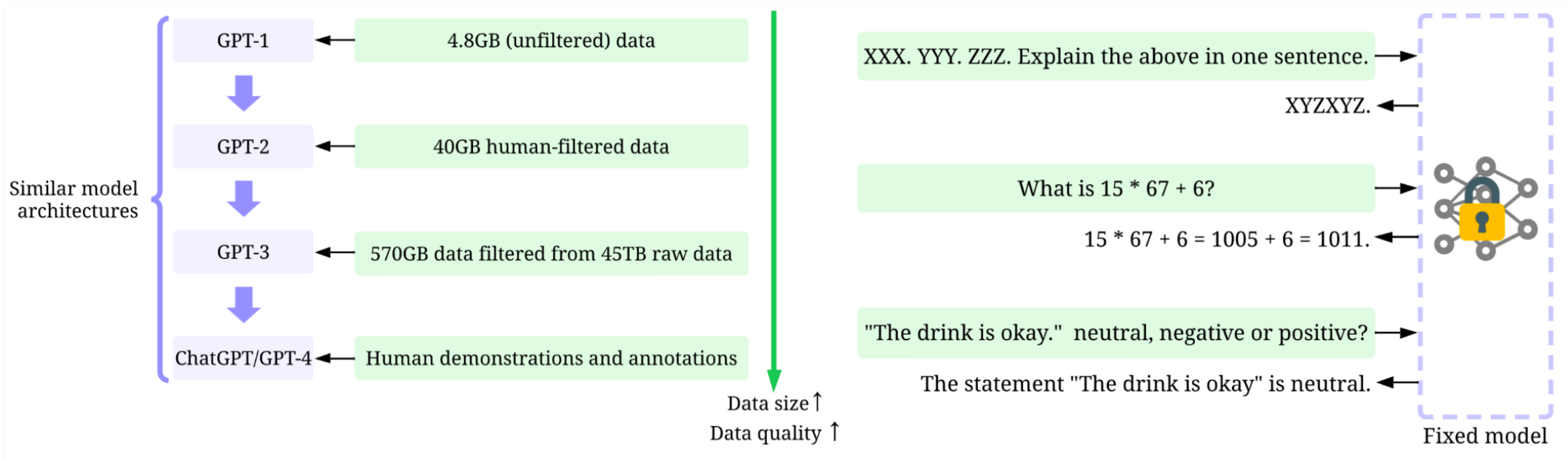
数据分析+
逻辑推理



数据科学基础

81

- 人工智能：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



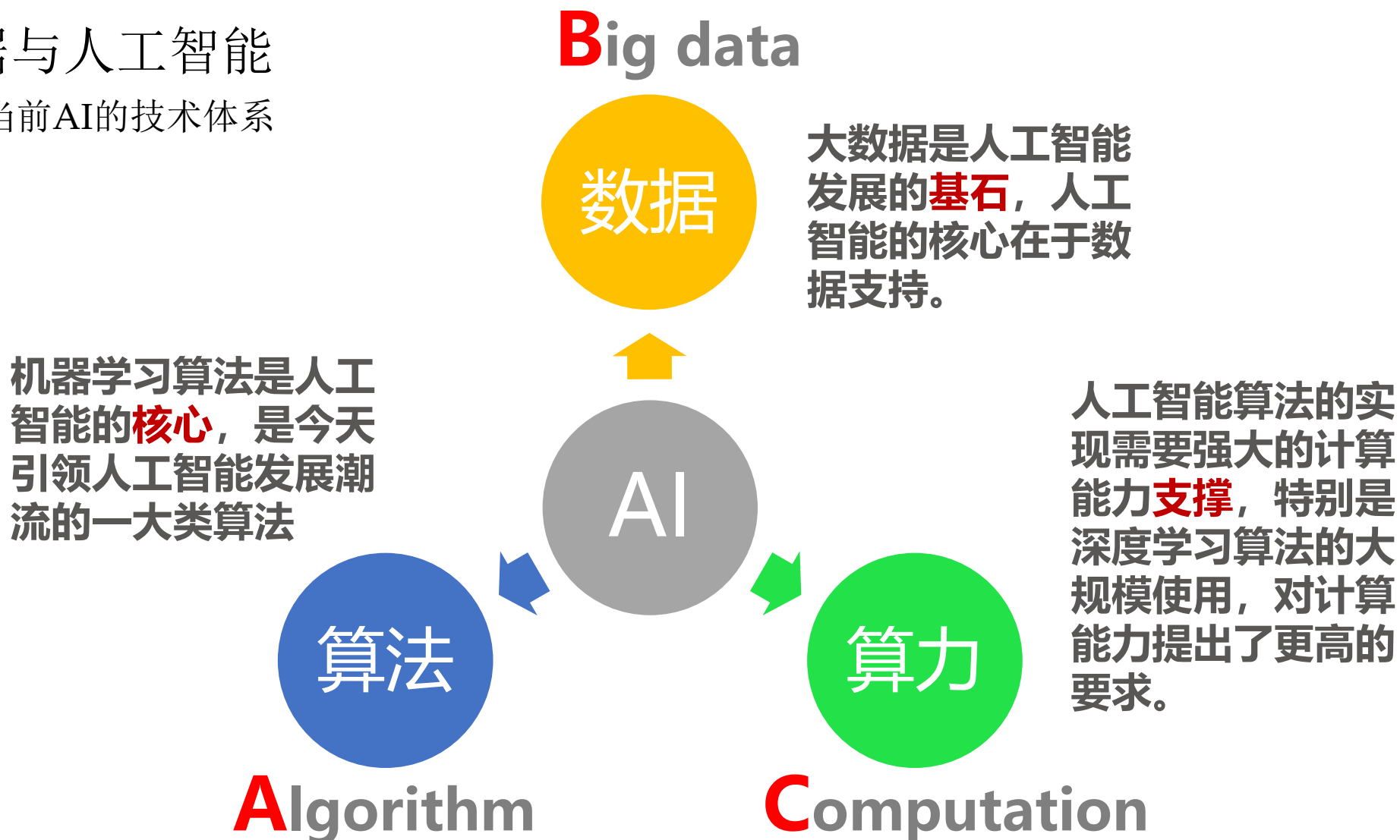


数据科学基础

82

□ 大数据与人工智能

□ ABC当前AI的技术体系



□ 大数据与人工智能的关系

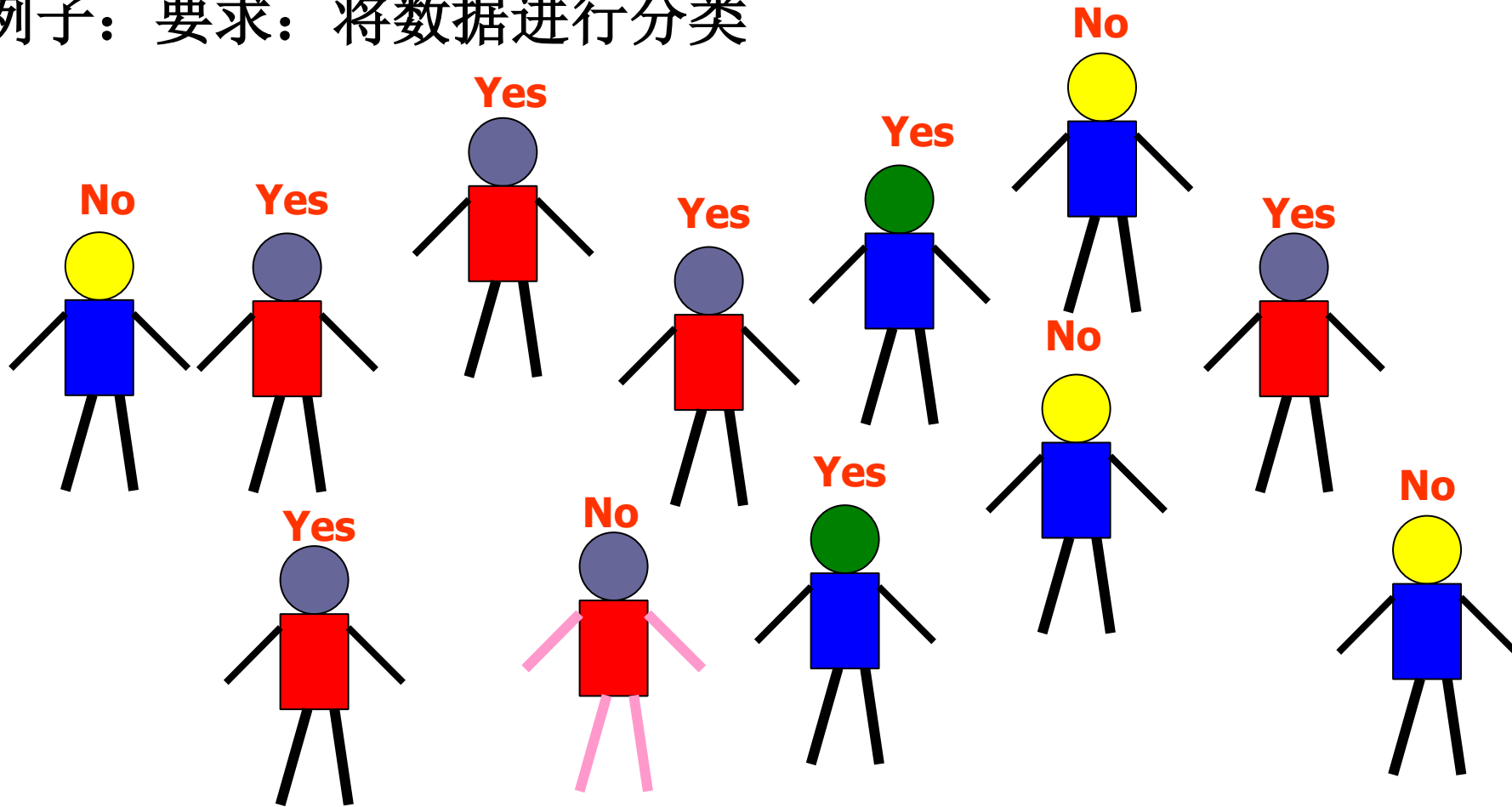
- 现阶段，人工智能的核心是对大数据进行的**特征抽取**与**机器学习算法**





数据+分类学习的方法

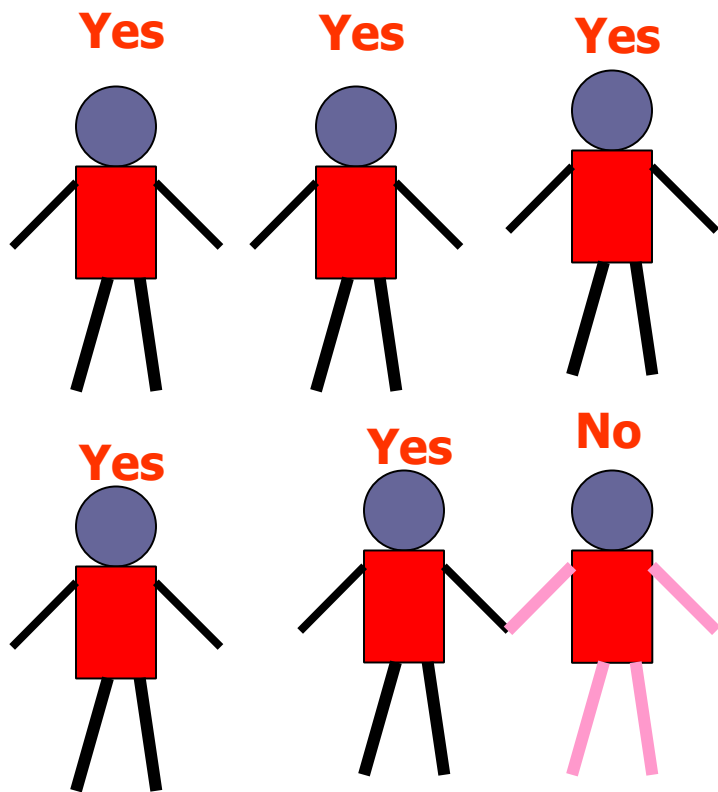
□ 一个例子：要求：将数据进行分类



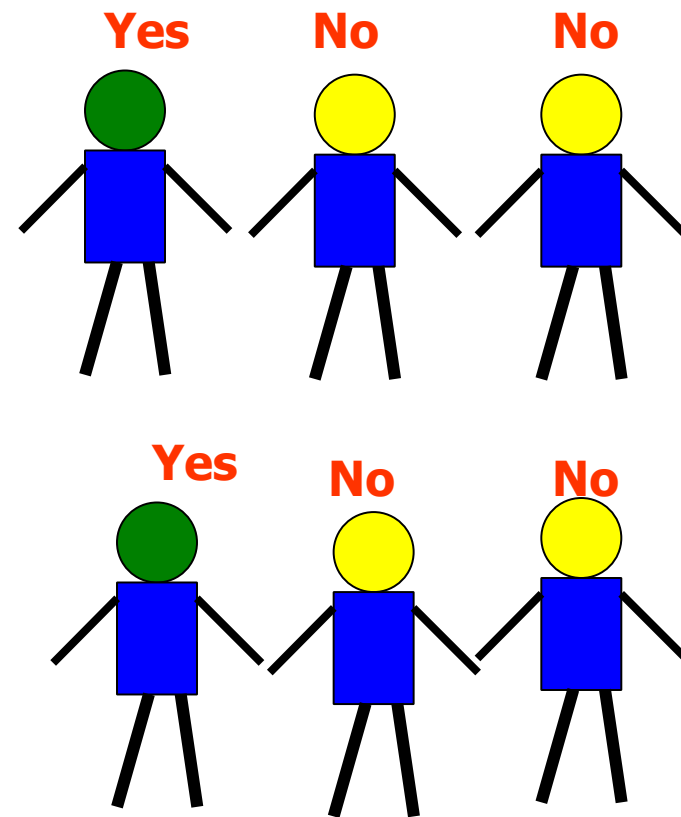


数据+分类学习的方法

躯干：红色



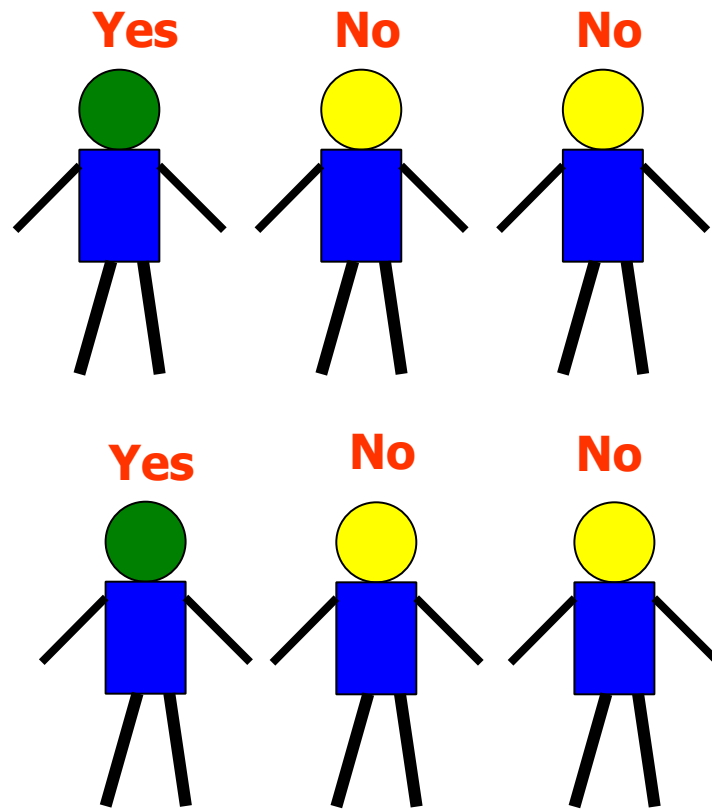
躯干：蓝色





数据+分类学习的方法

躯干：蓝色



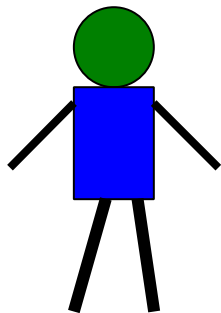


数据+分类学习的方法

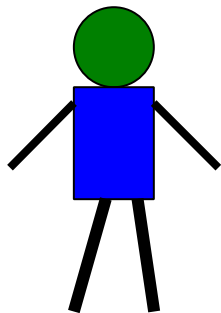
躯干：蓝色

头：绿色

Yes

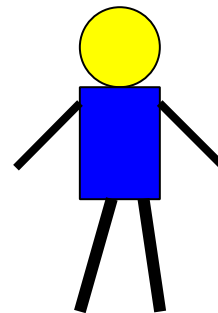


Yes

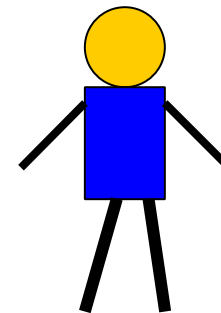


头：黄色

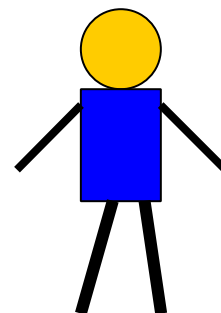
No



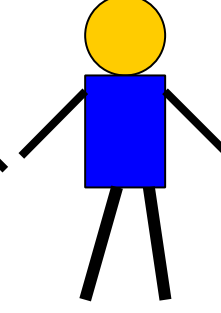
No



No



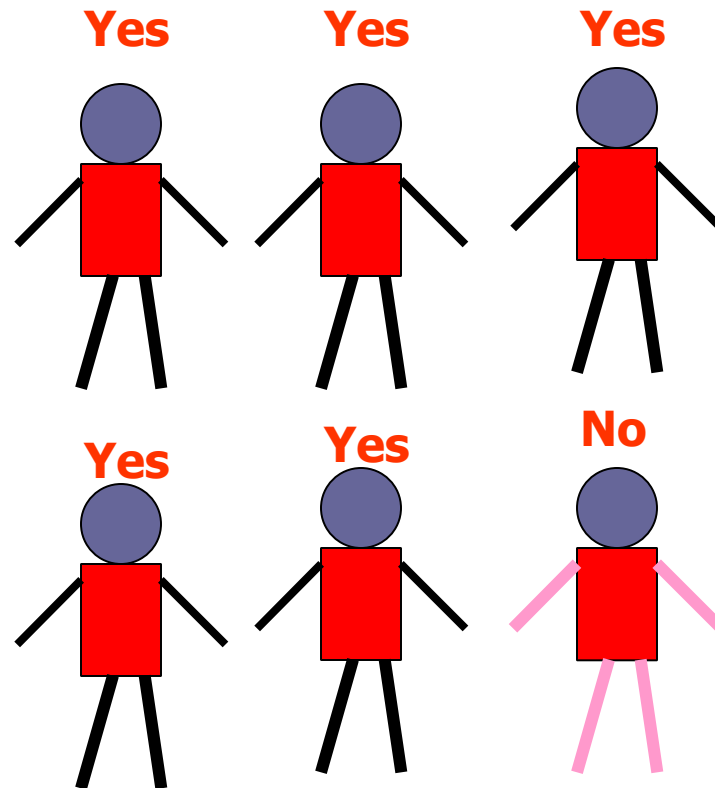
No





数据+分类学习的方法

躯干：红色

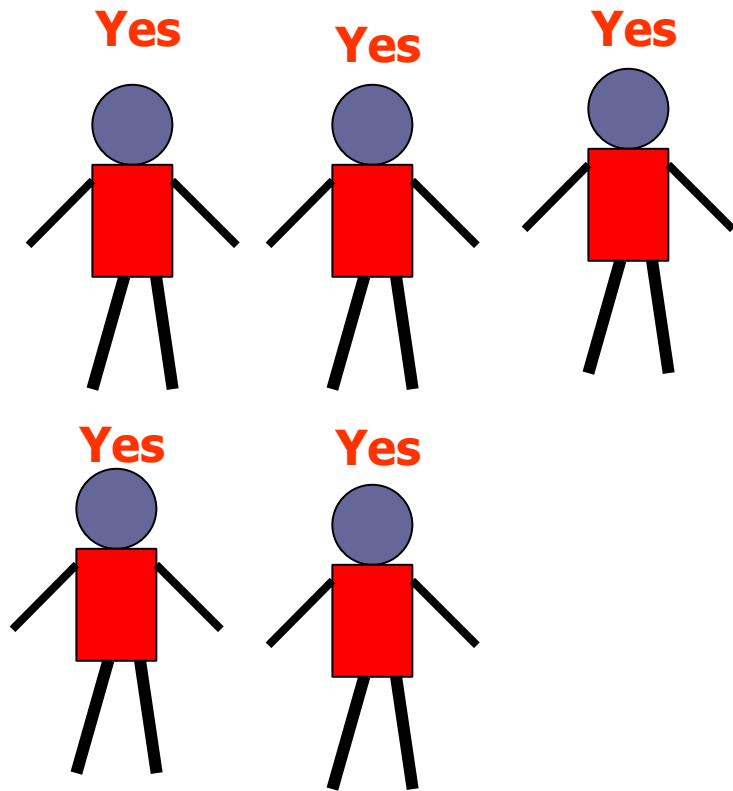




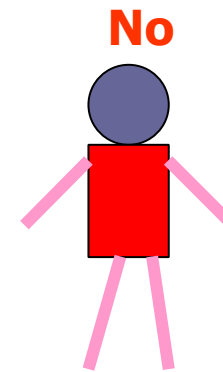
数据+分类学习的方法

躯干：红色

四肢：黑色

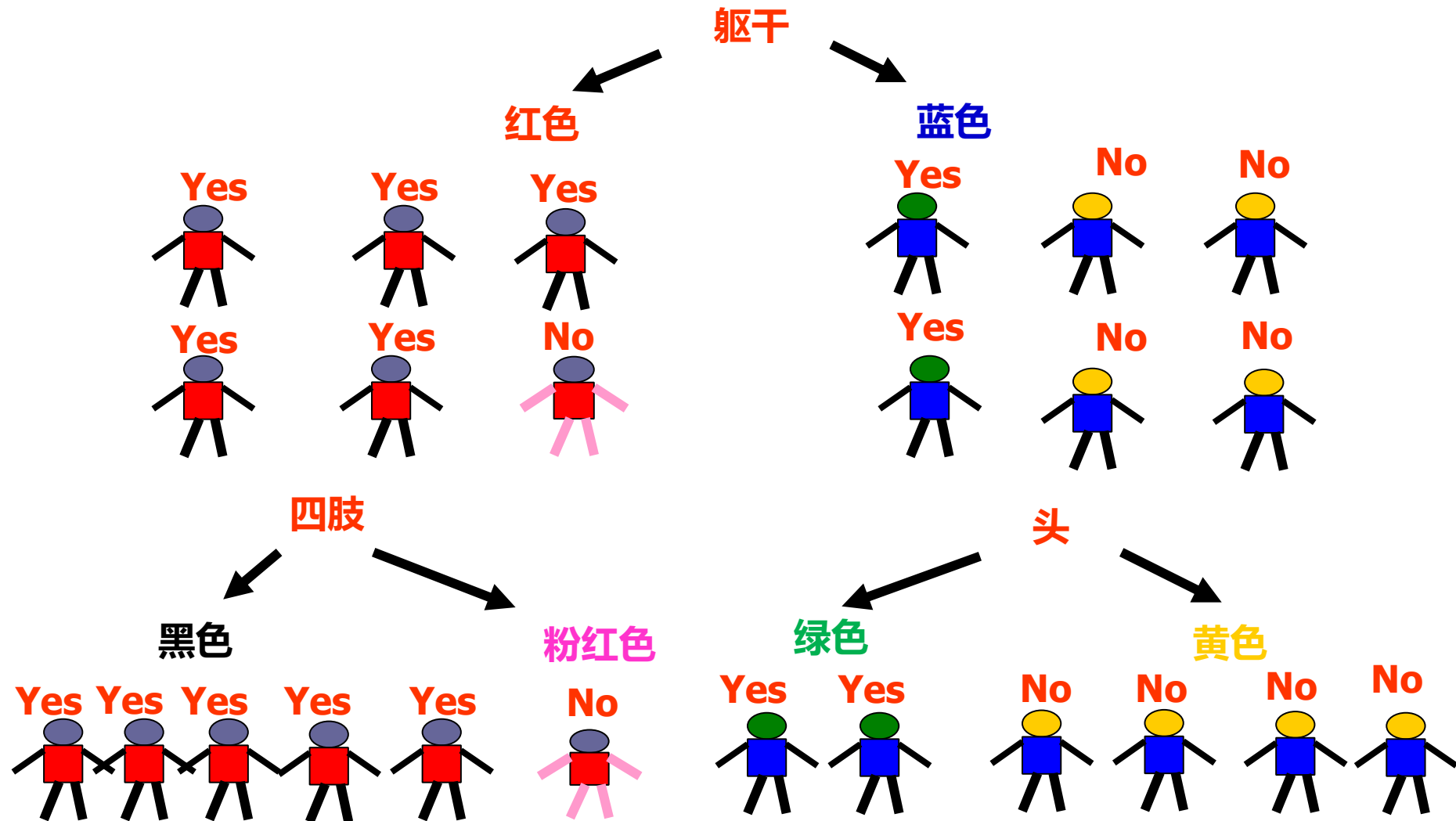


四肢：粉红色





数据+分类学习的方法



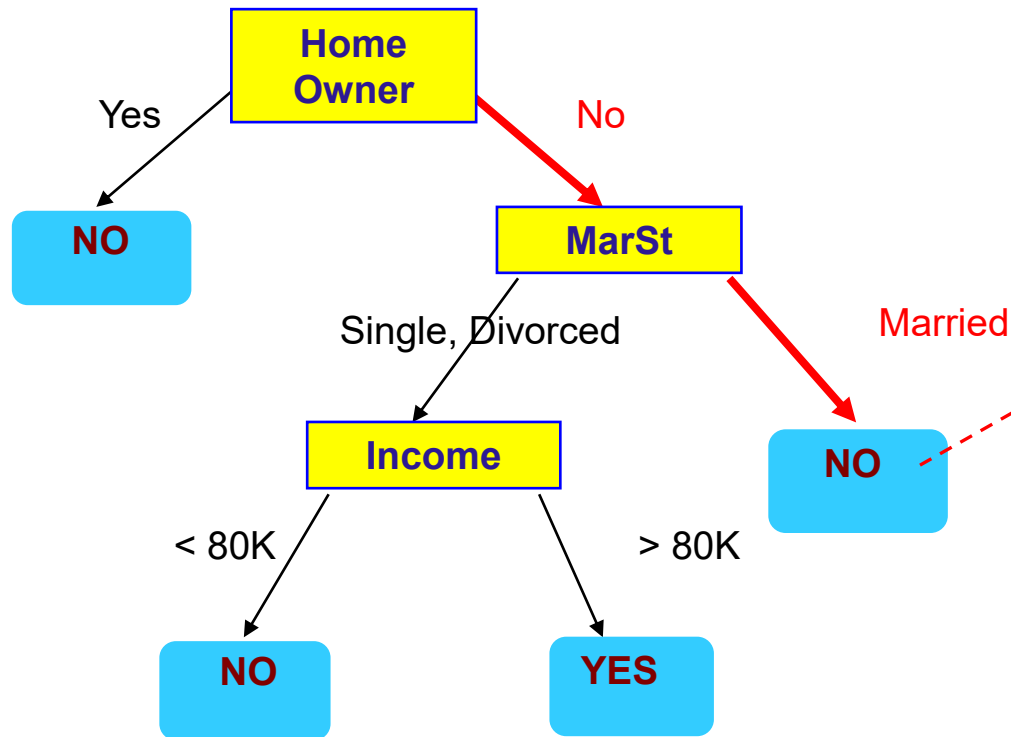


数据+分类学习的方法

□ 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

92

- 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

93

- 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: TeslaV100 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.MovidiusMyriad 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

94

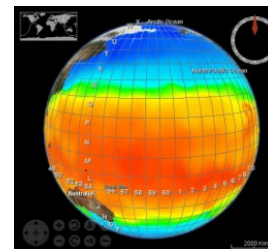


存储 (如硬盘 数据库)

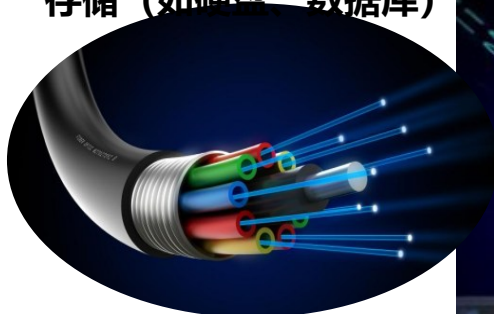


$$\min_X f(X) + \lambda \cdot \text{rank}(X)$$

分析、挖掘和学习



可视化



收集、传输



数据安全和个人隐私



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）

95

- 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
 - 人工智能技术与应用
 - 自然语言处理与大模型技术前沿
 - AI4Science
 - etc
 - 课程汇报与课程回顾，1周



课程要求与考核方式

96

- 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

97

- 授课教师：陈恩红，黄振亚
- 课程主页：<http://staff.ustc.edu.cn/~huangzhy/Course/DS2026.html>
- QQ群：1050506077
- 联系方式
 - 助教：罗胤玻，李宁
 - ds_intro2026@163.com
 - 教师：黄振亚，陈恩红
 - huangzhy@ustc.edu.cn

