新媒体大数据分析

New Media Big Data Analysis

第一章 数据科学基础

黄振亚，朱孟潇，张凯
Email: huangzhy@ustc.edu.cn，mxzhu@ustc.edu.cn

课程主页：
http://staff.ustc.edu.cn/~huangzhy/Course/NM2023.html

助教：陆文灏， 沙云浩
bigdata_2023@163.com

9/16/2023

课程目标

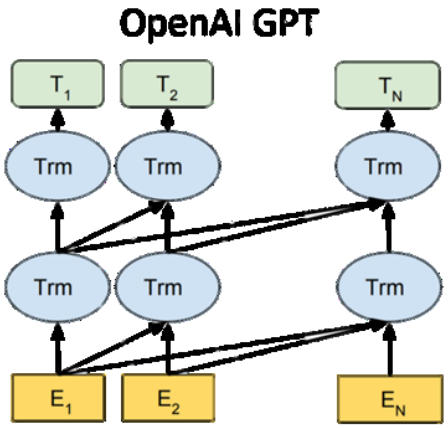
- 全面了解数据科学的基础知识
  - 包括数据分析的常用技术、发展前沿和应用案例
  - 了解数据的“能”与“不能”
- 树立数据科学的基本思路
- 初步掌握使用数据分析手段解决实际应用问题的能力

用科学的方法研究和应用数据

选修新媒体大数据分析课程的同学将来可能从事不同行业的科学研究、技术开发、产品管理等，希望这门课程带给你们的是终身受用的数据思维和创新能力。

数据科学基础

- ChatGPT：大数据催生人工智能新浪潮
  - 参数量从1.17亿增加到1750亿
  - 数据量从5GB增加到45TB
    - 96%以上是英文，其它20个语种不到4%



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

- 数据
  - 从计算机科学的角度，所有能够输入到计算机并被计算机程序处理的符号的总称
  - 新时代的生产要素（十四五）
- “人-机-物”三元融合，世界已经成为数据化的世界

Google 谷歌

Baidu 百度

当文字成为数据

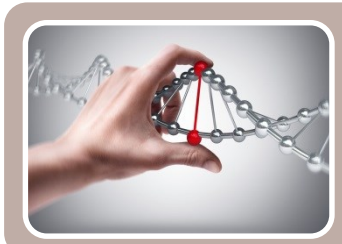
当方位成为数据

一切事物的数据化

当沟通成为数据

数据科学基础

- 科技传播系统致力于培养的“科技媒体和科技传播”英才应具备以下素质



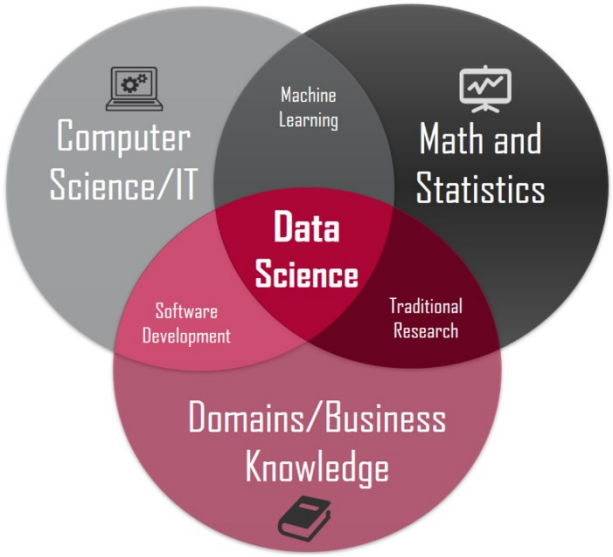
专业基础扎实，有卓越的数理基础、创意设计、管理能力

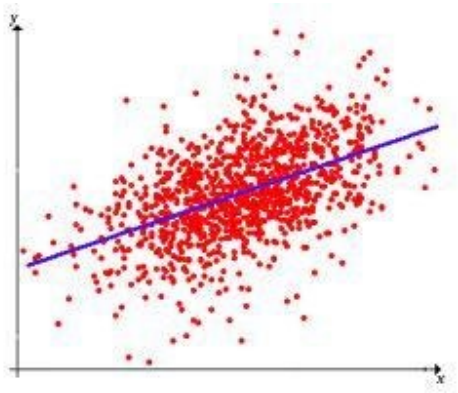
实践能力强，具有处理多媒体大数据的能力

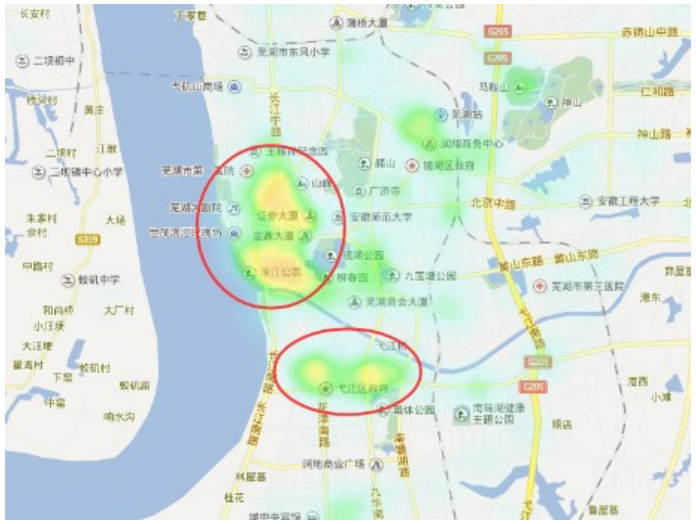
跨界能力强，能够解决新媒体行业的大数据应用问题

数据科学基础

- **大数据新工科人才需要具备以下素质**
  - 学习理论知识：数学（基础）+ 计算机科学 + 交叉学科知识
  - 锻炼实践能力：编程、数据分析、数据可视化等
  - 培养跨界能力：应用场景、领域知识







数据科学基础

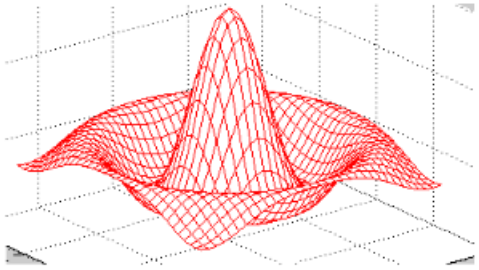
- **1. 理论基础扎实，能理解运用数据科学中的理论模型**
- 数学是学习数据科学的基础
  - 数学与优化：数学分析的应用
    - 梯度下降
    - 搜索方向：负梯度方向、牛顿方向
    - 算法收敛性
  - 数学与聚类：线性代数的应用
    - 社交网络聚类的问题形式化
    - 线性代数知识求解
  - 数学与图卷积网络：傅里叶变换的应用
    - 图表征学习
    - 图上的傅里叶变换与卷积
  - 。 。 。

数据科学基础

- 1. 理论基础扎实，能理解运用数据科学中的理论模型
  - 数学是学习数据科学的基础
    - 数学与优化：数学分析的应用

模型学习（机器学习）

找到合适的 w ，使 f(w, x) 最接近 D



例如，线性回归损失函数

$$L(w)=\sum_{d_i \in D} f(w, x_i)-y_i$$

$$w=argmin_w L(w)$$

优化方法

常见问题： $\min_{x \in R^n} f(x)$

- 梯度下降
- 牛顿法/拟牛顿法
- 。。。

数据科学基础

- **1. 理论基础扎实，能理解运用数据科学中的理论模型**
  - **数学是学习数据科学的基础**
    - **数学与聚类：线性代数的应用**

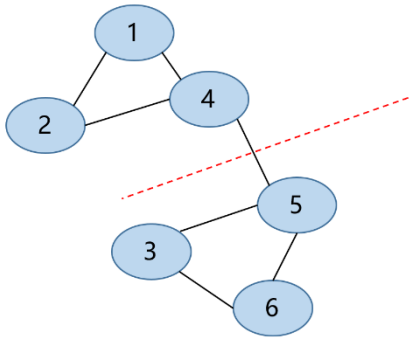
社交网络划分：物以类聚，人以群分

- ✓ 将全校学生划分为不同班级?
- ✓ 将员工划分为不同公司?
- ✓ 将用户划分为不同追星圈?

关键点：利用不同个体之间的联系

问题转化

无向图分割问题



$$W = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

$$D = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

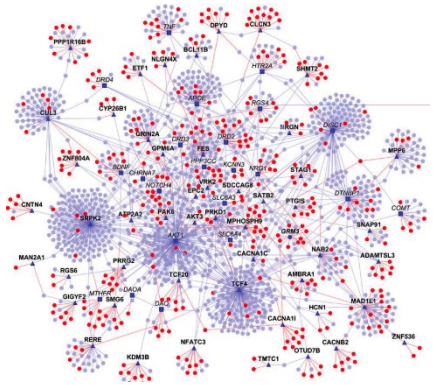
常用知识:

- 特征值分解
- 奇异值分解
- QR分解
- 矩阵求逆相关定理

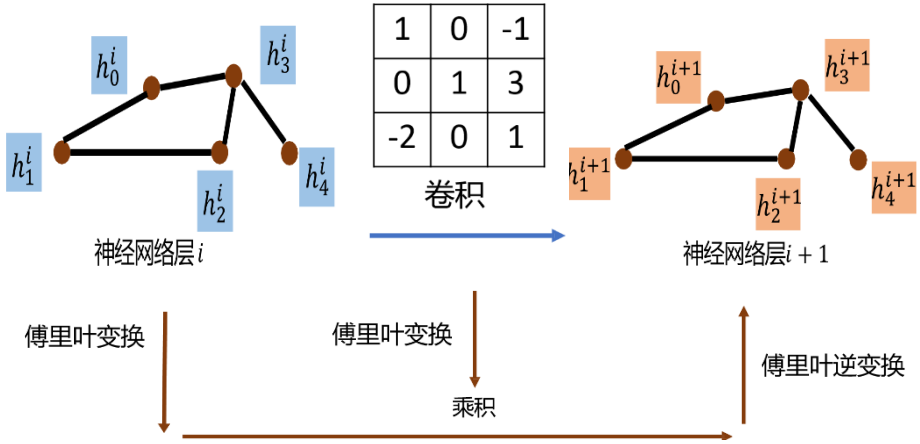
数据科学基础

- 1. 理论基础扎实，能理解运用数据科学中的理论模型
  - 数学是学习数据科学的基础
    - 数学与图卷积网络：傅里叶变换的应用

图数据：分子图、社交网络等



图卷积网络：一类典型方法



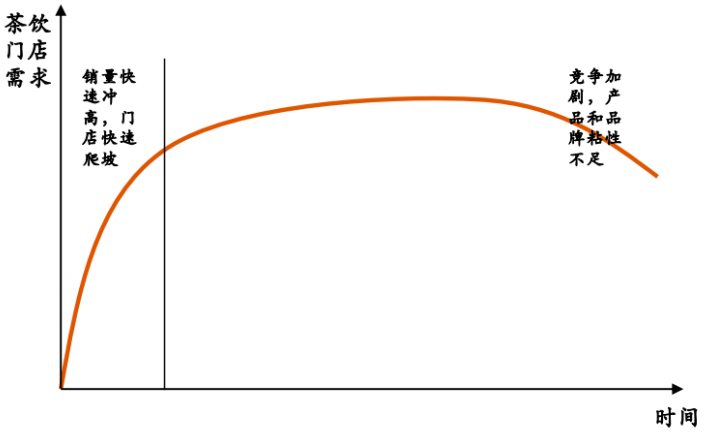
- ✓ 典型任务
  - ✓ 节点分类，关系（边）预测等
  - ✓ 图分类，图属性预测，图生成

- ✓ **Idea：卷积定理：函数卷积的傅里叶变换是函数傅立叶变换的乘积**
- ✓ **一般傅里叶变换　至　图上傅里叶变换**

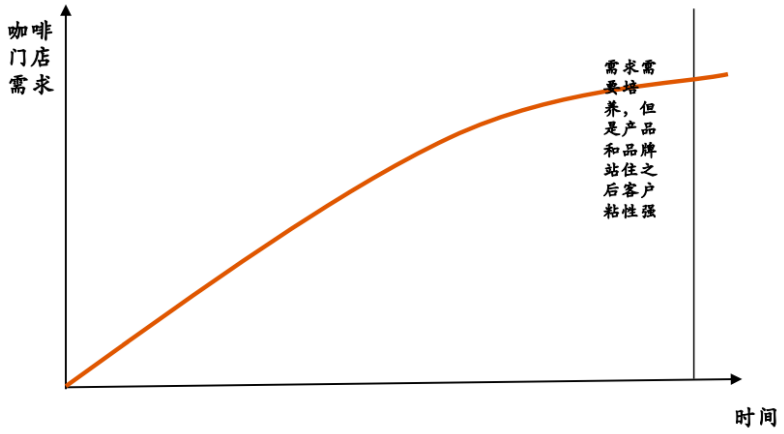
数据科学基础

- 1. 创意设计与管理能力
  - 运用数据分析结论帮助产品设计与营销管理
  - 例：瑞幸咖啡
    - 竞品分析，价格分析，用户分析，地理分析等
    - 新媒体营销等

茶饮需求相对成熟，门店能够快速形成赚钱效应



咖啡需求需要培养，单店爬坡周期更长



数据科学基础

- **2. 实践能力强，具有处理大数据的能力**
  - Python等编程技术，Web技术、数据库技术、可视化技术等
  - 常用工具使用：如大模型，可视化工具

```
def SumOfKsubArray(arr, n, k): SumOfKsubArray(arr, n, k):
  Sum = 0Sum = 0
  S = deque()= deque()
  G = deque()= deque()
  for i in range(k):for i in range(k):
    while (len(S) > 0 and arr[S[-1]] >= arr[i]):while (len(S) > 0 and arr[S[-1]] >= arr[i]):
      S.pop().pop()
    while (len(G) > 0 and arr[G[-1]] <= arr[i]):while (len(G) > 0 and arr[G[-1]] <= arr[i]):
      G.pop().pop()
    G.append(i).append(i)
    S.append(i).append(i)
  for i in range(k, n):for i in range(k, n):
    Sum += arr[S[0]] + arr[G[0]]Sum += arr[S[0]] + arr[G[0]]
    while (len(S) > 0 and S[0] <= i - k):while (len(S) > 0 and S[0] <= i - k):
      S.popleft().popleft()
    while (len(G) > 0 and G[0] <= i - k):while (len(G) > 0 and G[0] <= i - k):
      G.popleft().popleft()
    while (len(S) > 0 and arr[S[-1]] >= arr[i]):while (len(S) > 0 and arr[S[-1]] >= arr[i]):
      S.pop().pop()
    while (len(G) > 0 and arr[G[-1]] <= arr[i]):while (len(G) > 0 and arr[G[-1]] <= arr[i]):
      G.pop().pop()
    G.append(i).append(i)
    S.append(i).append(i)
  Sum += arr[S[0]] + arr[G[0]]Sum += arr[S[0]] + arr[G[0]]
  return Sumreturn Sum
```

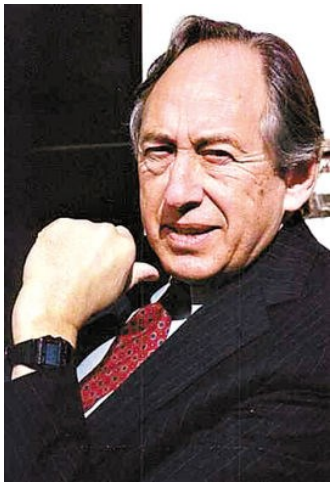
数据科学基础

- 3. 跨界能力强，能够解决特定行业的大数据应用问题

改变这个世界的四种力量

暴力

知识



世界著名未来学家托夫勒
《第三次浪潮》作者

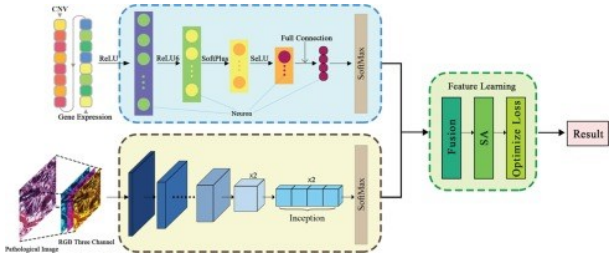
大数据

金钱

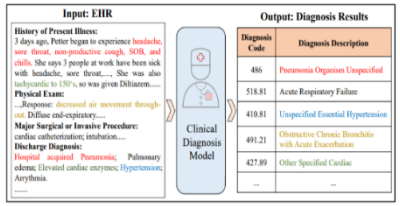
数据科学基础

- 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

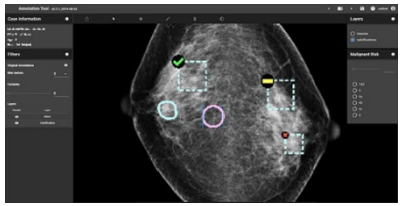
AI电子病历



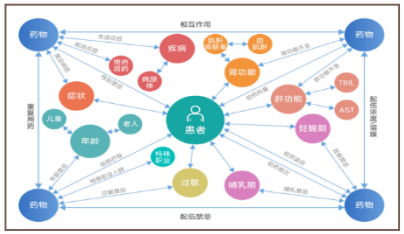
多模态医疗数据挖掘模型



疾病诊断



病灶识别



药物推荐

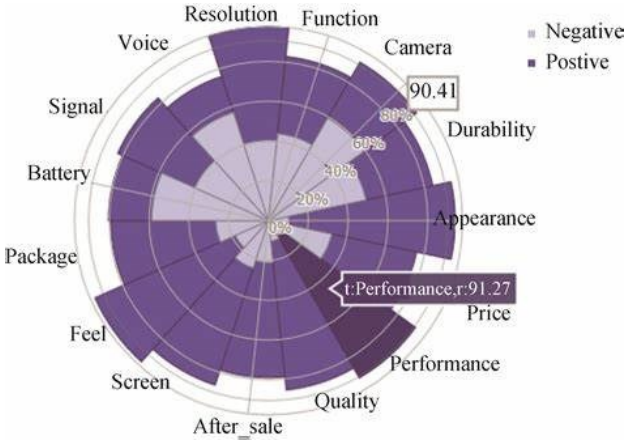
- 数据蕴含着巨大的价值—安防领域
  - 公安监控智能分析

区间超速判定

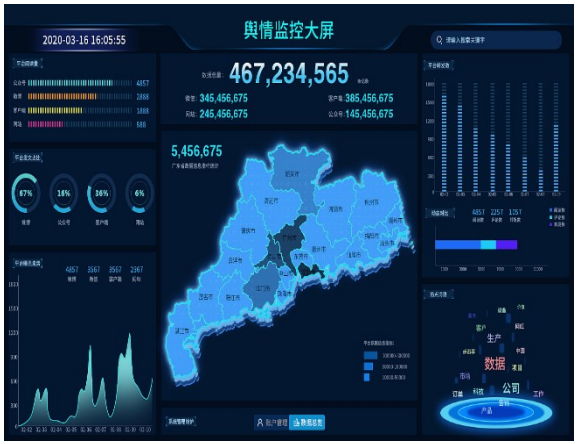
“天眼”追凶

数据科学基础

- 数据蕴含着巨大的价值—安防领域
  - 舆情监测



舆情情感分析

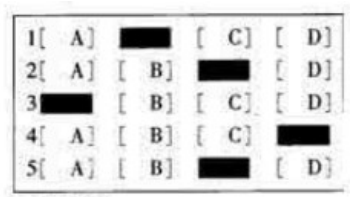


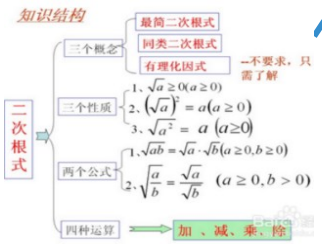
传播途径监测

数据科学基础

- 数据蕴含着巨大的价值—智慧教育
  - 因材施教

学生的**学习行为数据**





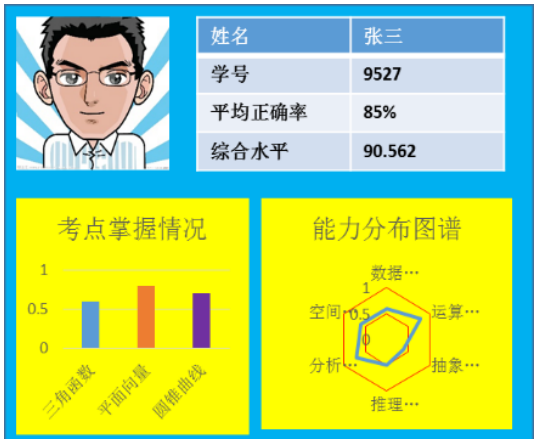
试题-知识点

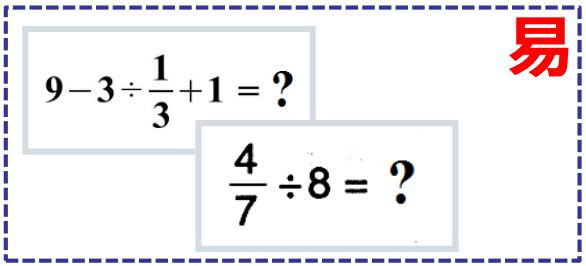
大数据分析

学生认知水平画像

试题难度等特征的预测

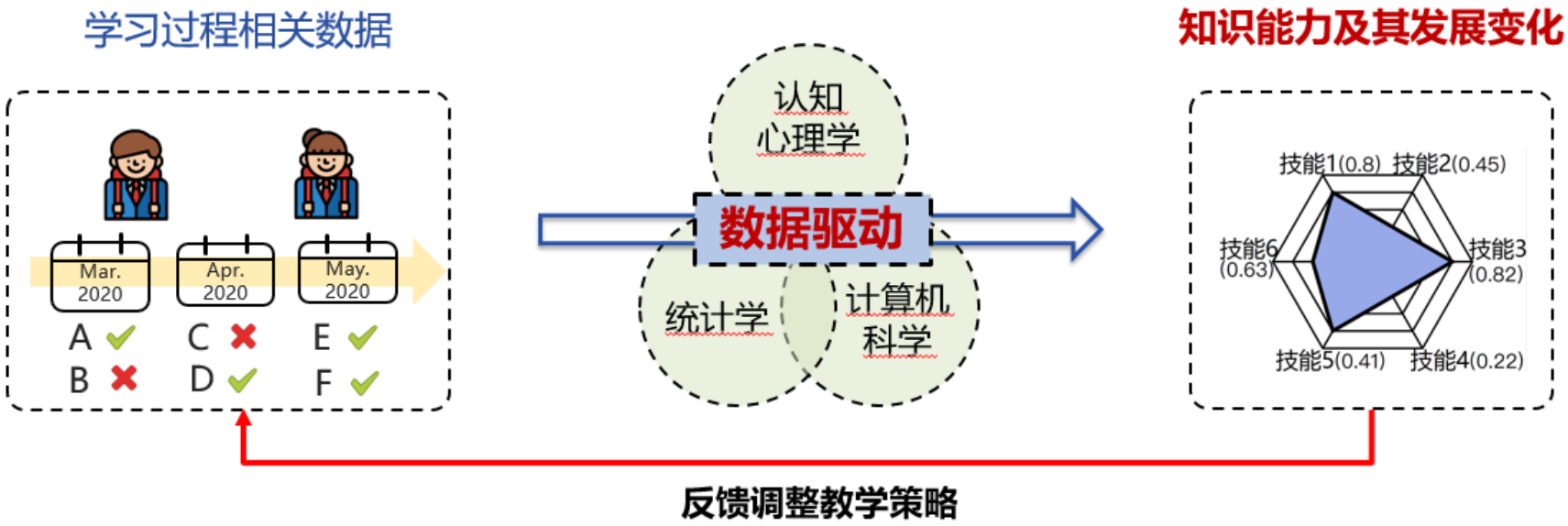
个性化学习推荐





数据科学基础

- 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



数据科学基础

- 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

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]

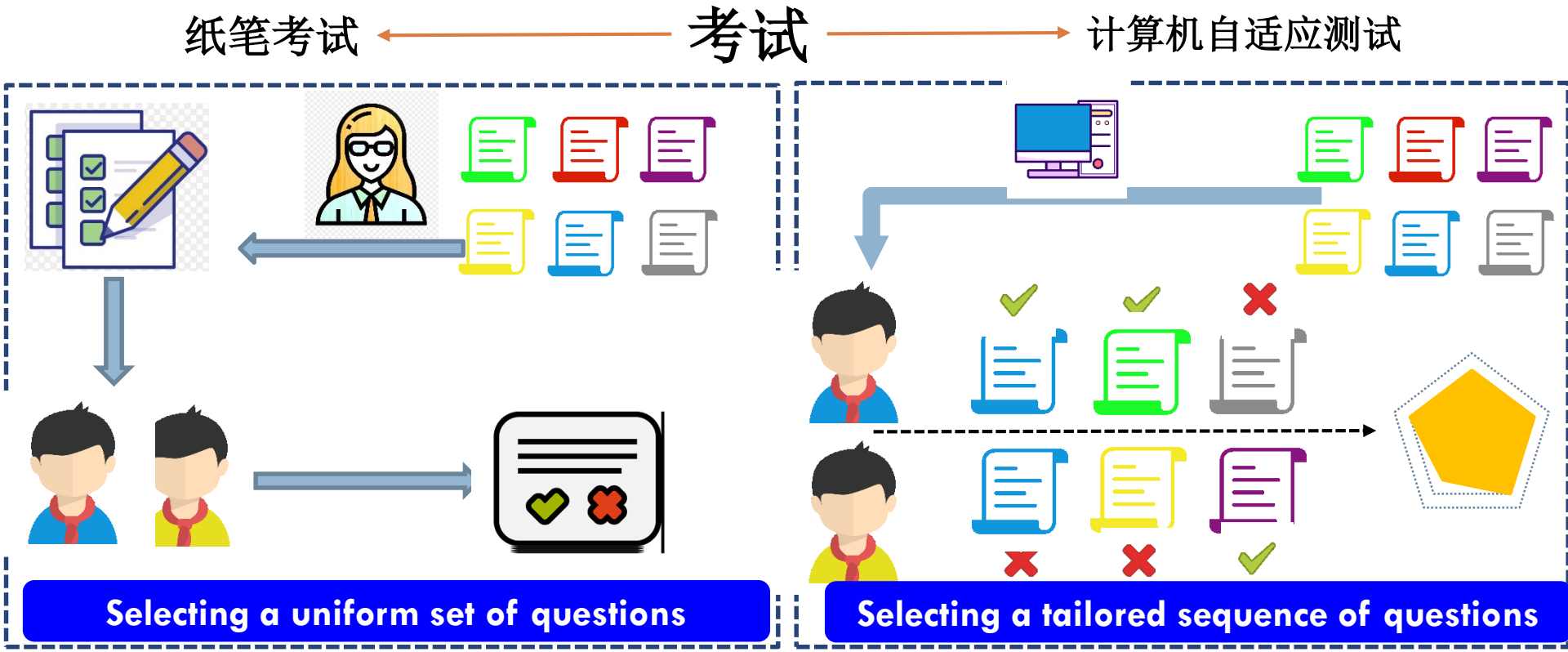
浙江省高考英语科目一年安排2次考试，考生可报考2次，选用其中较高1次的成绩。在2018年11月刚结束的英语科目考试中，根据答卷试评情况，发现部分试题与去年同期相比难度较大。为保证不同次考试之间的试题难度大体相当，浙江省招委组织专家研究论证，在制订评分细则时，决定面向所有考生，对难度较大的第二部分（阅读理解）、第三部分（语言运用）的部分试题进行难度系数调整，实施加权赋分。其他试题未作调整。

浙江省教育考试院
2018年11月27日

数据科学基础

- 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

- 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

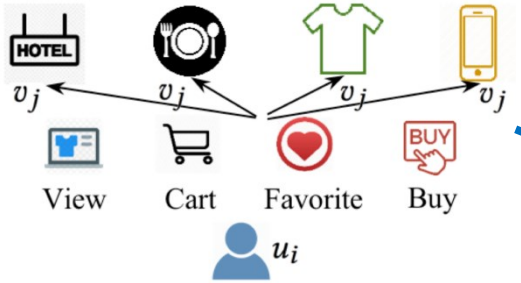
15万名奥巴马支持者在Facebook安装了“奥巴马2012”应用，而通过这个程序，总统竞选团队可以间接得到这些支持者数百万的Facebook好友信息。

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

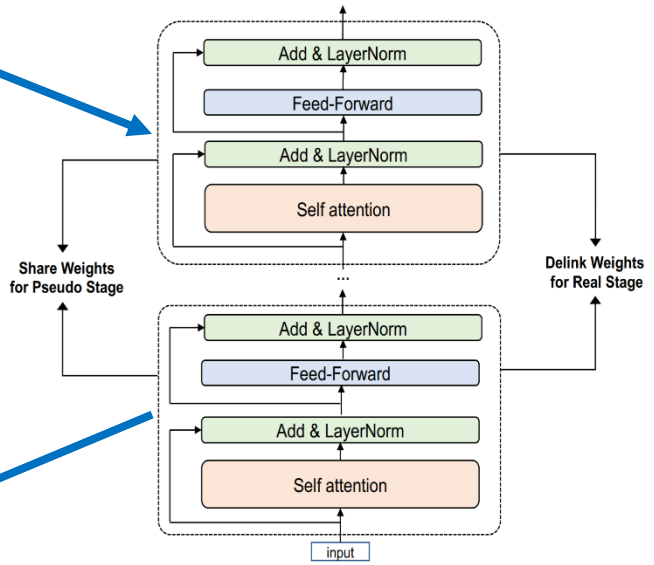
- 数据蕴含着巨大的价值—电子商务
  - 计算广告

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

- 数据蕴含着巨大的价值—电子商务
  - 精准搜索、个性化消费推荐

天猫双11：破亿交易时间

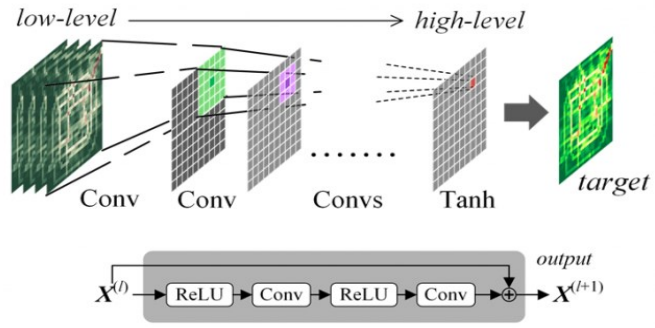
年份	10亿	100亿	500亿
2014	3分	38分28秒	21时12分
2015	1分12秒	12分28秒	9时52分22秒
2016	52秒	6分58秒	2时30分20秒
2017	28秒	3分01秒	40分12秒
2018	21秒	2分05秒	26分03秒
2019	14秒	1分36秒	12分49秒

数据科学基础

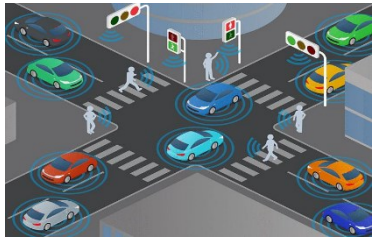
- 数据蕴含着巨大的价值 — 智慧城市
  - 基于运营商基站数据、交通路口和车辆移动轨迹数据等
  - 提升城市**交通管控**、**交通服务**和**规划水平**，实现用户未来行程规划、交通路口信号灯调控、合理规划道路建设等

车辆移动数据

交通流预测模型



行程规划

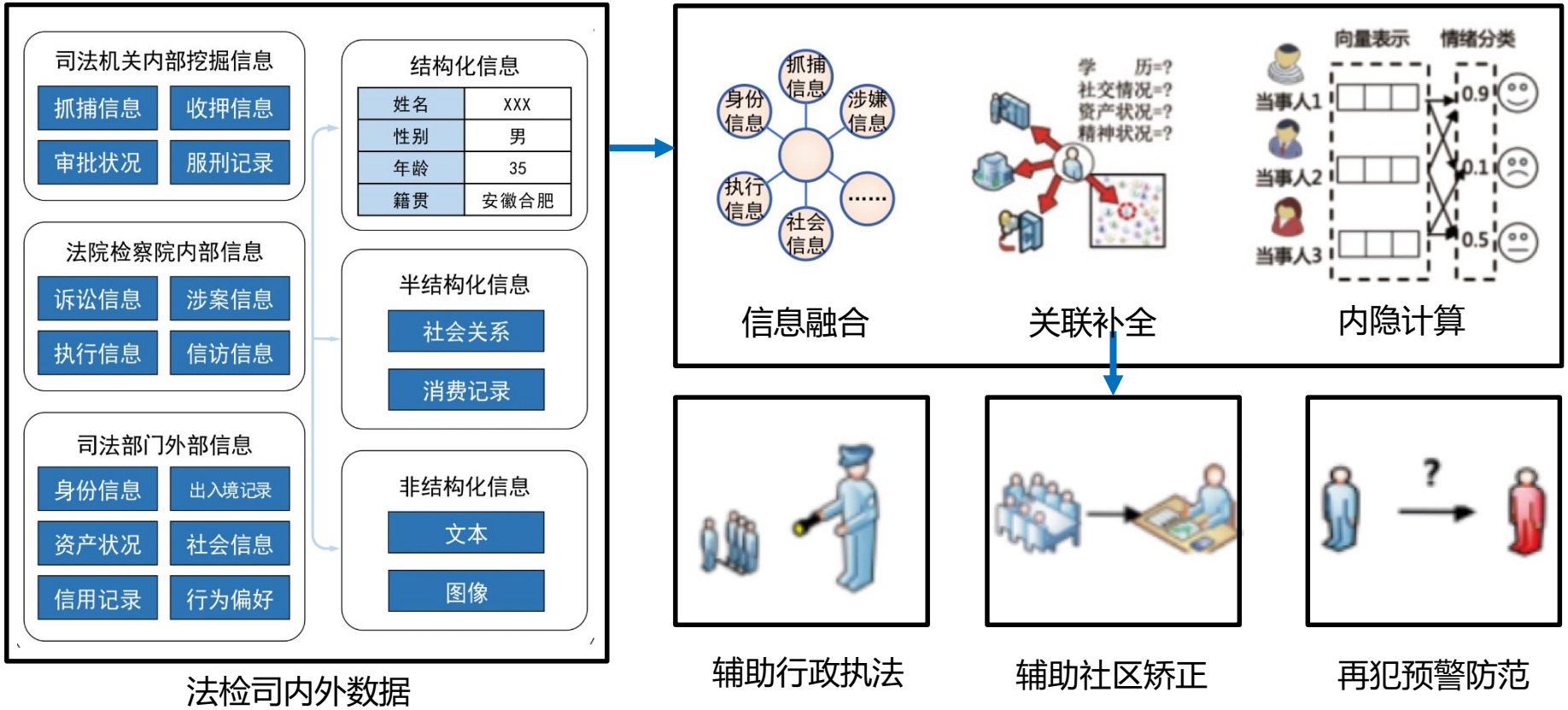


信号灯调控

道路规划

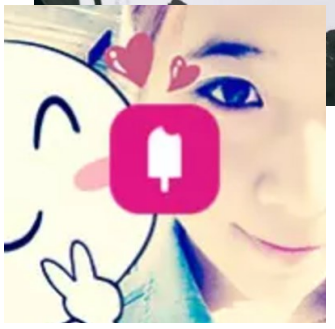
数据科学基础

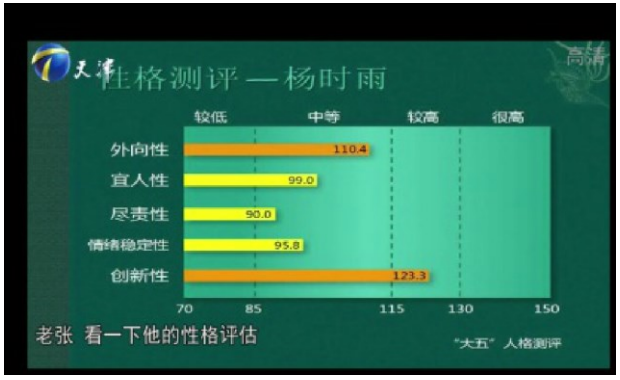
- 数据蕴含着巨大的价值——智慧司法
  - 基于法、检、司等部门关于涉案当事人的内部数据与外部数据等
  - 构建涉案当事人画像，辅助行政执法、社区矫正、进行再犯预警防范等，**提升司法管理服务的质量**和**效率**



数据科学基础

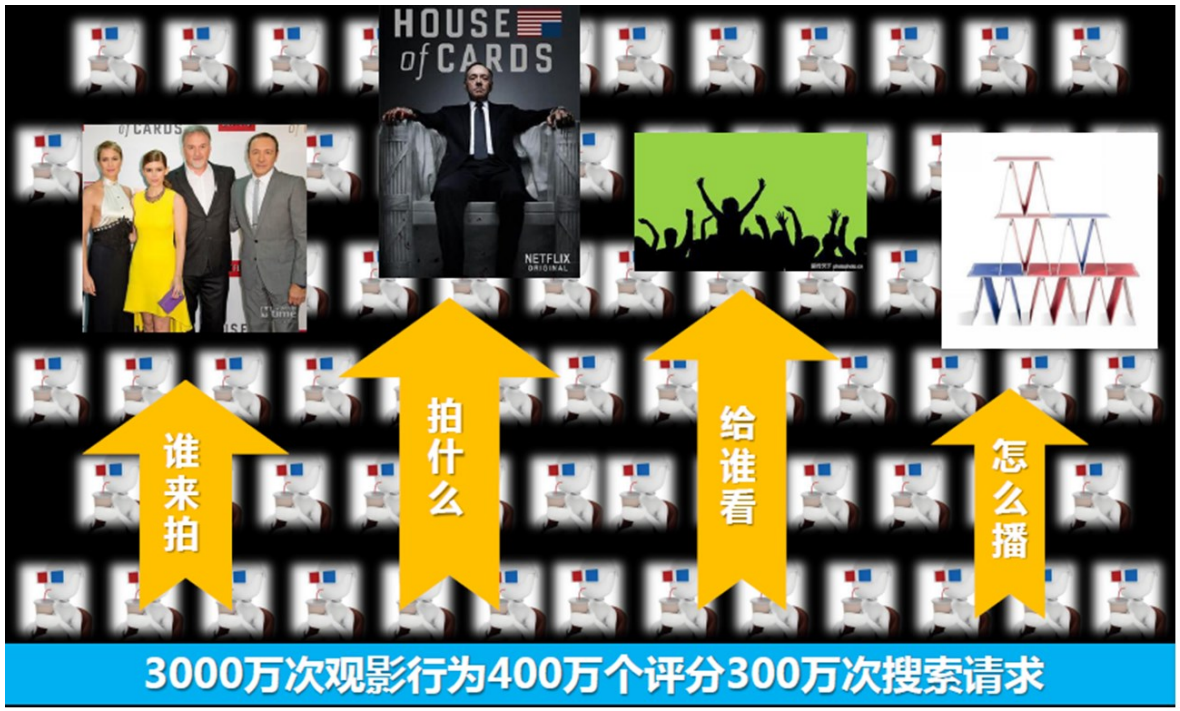
- 数据蕴含着巨大的价值—智能助手





数据科学基础

- 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



数据科学基础

- 数据蕴含着巨大的价值—文化娱乐
  - 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。

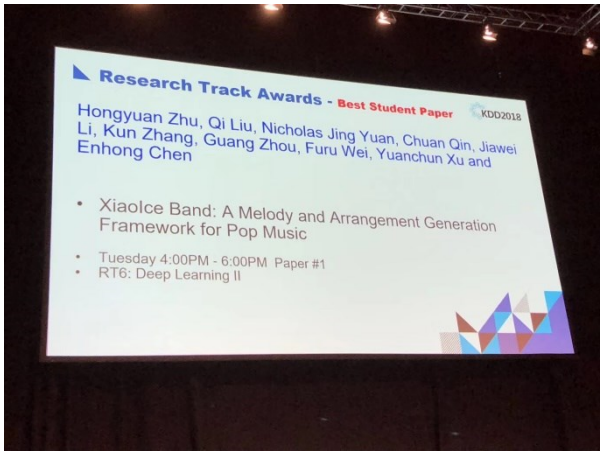
- 机器作诗 PK 古代诗人

哪首诗是人写的?

秋夕湖上	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。	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

- 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/VIDEo7JWp0u0oWRmPbM4uCBt171124.shtml



【KDD18最佳论文揭晓】中科大等斩获最佳学生论文，刘兵获创新奖，清华大学唐杰任副主席

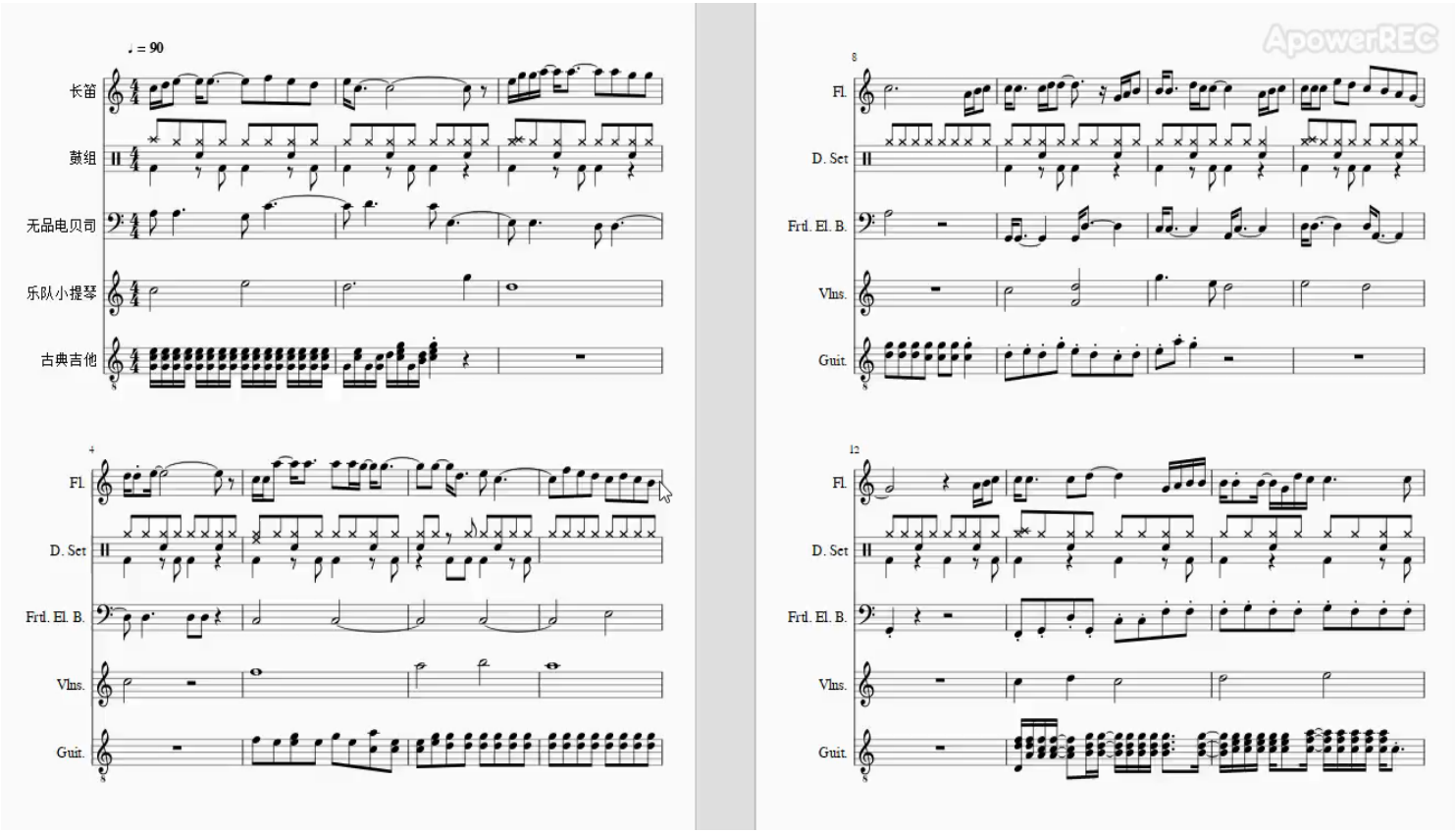
新智元 2018-08-21 06:24:19 阅读(0) 评论(0)

首页 新闻博览

我校获数据挖掘领域顶级国际会议KDD 2018最佳学生论文奖

数据科学基础

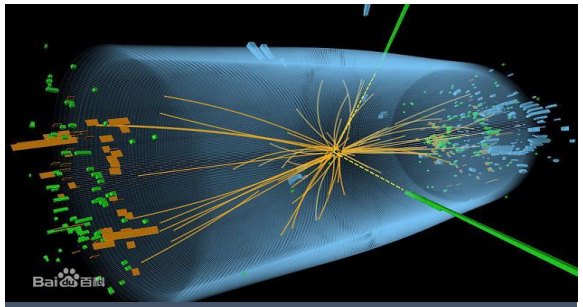
- 数据蕴含着巨大的价值—文化娱乐
  - 流行音乐的旋律与编曲生成



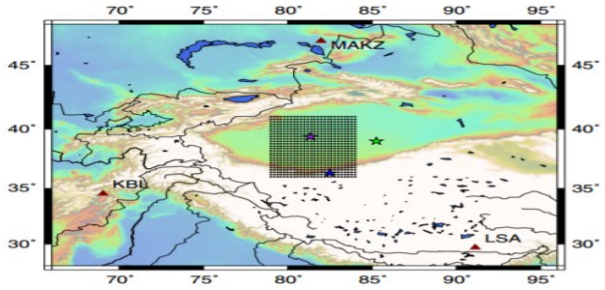
数据科学基础

- 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

专利数据挖掘保护知识产权

数据科学基础

- **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

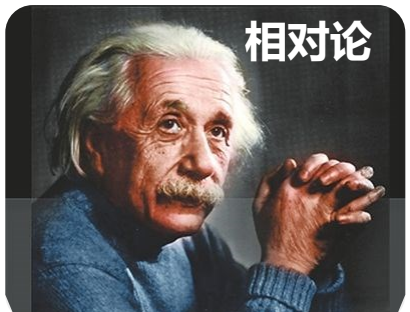
单一学科 → 数据驱动

多学科交叉

关系型数据库的鼻祖Jim Gray（右）

数据科学基础

- **2007年，Jim Gray总结出了四个科学范式**

几千年前	几百年前	几十年前	今天
经验科学	**理论科学**	**计算科学**	**数据密集型科学**
• **第一范式**	• **第二范式**	• **第三范式**	• **第四范式**
• 以**归纳法**为主，带有盲目性的观测和实验	• 以**演绎法**为主，关注理论总结和理性概括	• 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	• 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
• **科学实验**	• **数学模型**	• **科学计算**	• **机器学习**
 自由落体实验	 相对论	 冯诺依曼计算机	 天文大数据

数据科学基础

- 把握大数据带来的机遇
  - 零售业
    - Winners: Amazon, Ebay
    - Traditional: 传统书店、电子产品零售店
  - 旅游业
    - Winners: Expedia, Ctrip
    - Traditional: 旅行中介商
  - 金融服务业
    - Winners: E*trade, TD Ameritrade
    - Traditional: 股票中介商公司

数据科学基础

视频数据

- 把握大数据带来的机遇
- 影像租赁业
  - Winners: 视频流媒体公司(Netflix, Amazon, Hulu)
  - Traditional : DVD租赁公司
- 软件应用业
  - Winners: 软件数据服务公司(Saleforce.com)
  - Traditional : 软件产品公司
- 新闻报纸业
  - Winners: Google, Twitter, Facebook, Bloomberg
  - Traditional : 传统报纸业, Washington Post, WSJ
- 出租车行业
  - Winners: Uber, DiDi

数据科学基础

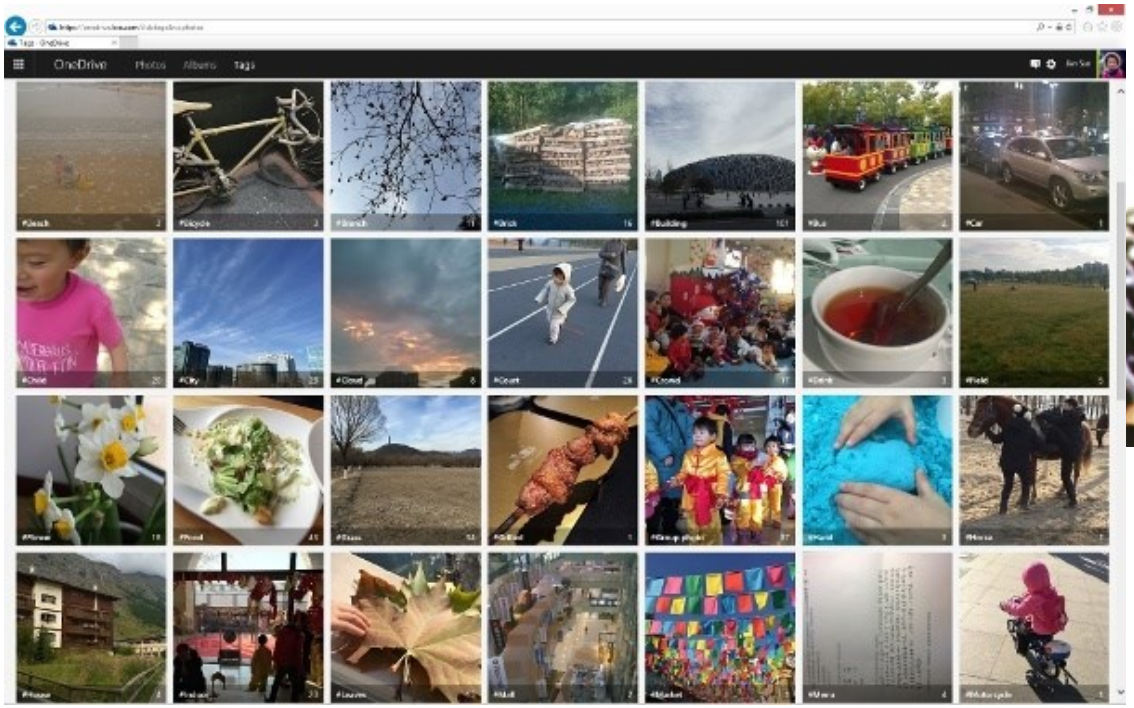
- 新媒体
  - 利用数字技术、网络技术和移动通信技术，通过互联网、宽带局域网、无线通信网和卫星等渠道，以电视、电脑和手机为主要输出终端，向用户提供视频、音频、语音数据服务、连线游戏、远程教育等集成信息和娱乐服务的所有新的传播手段或传播形式的总称，包括“新兴媒体”，也包括“新型媒体”
- 新媒体大数据
  - 新媒体服务场景中收集获取的数据
  - 按数据的模态类型：语音，图片，视频，网络，文本

数据科学基础

- 多媒体大数据研究热点—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

数据科学基础

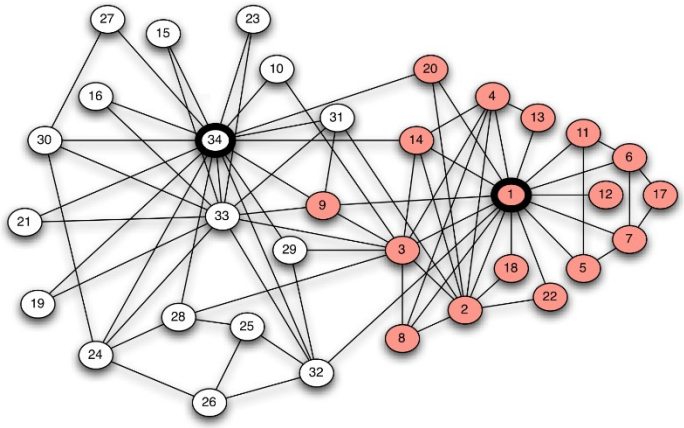
- 多媒体大数据研究热点—图片大数据
  - 图像识别
    - ImageNet图像数据库上，人工智能已达到2.99%的错误率（公安部三所），低于人类5.1%的错误率

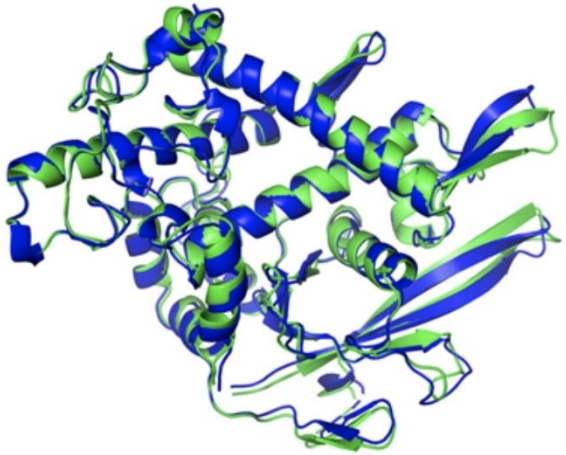


李飞飞
斯坦福大学、谷歌Ai前任首席科学家

数据科学基础

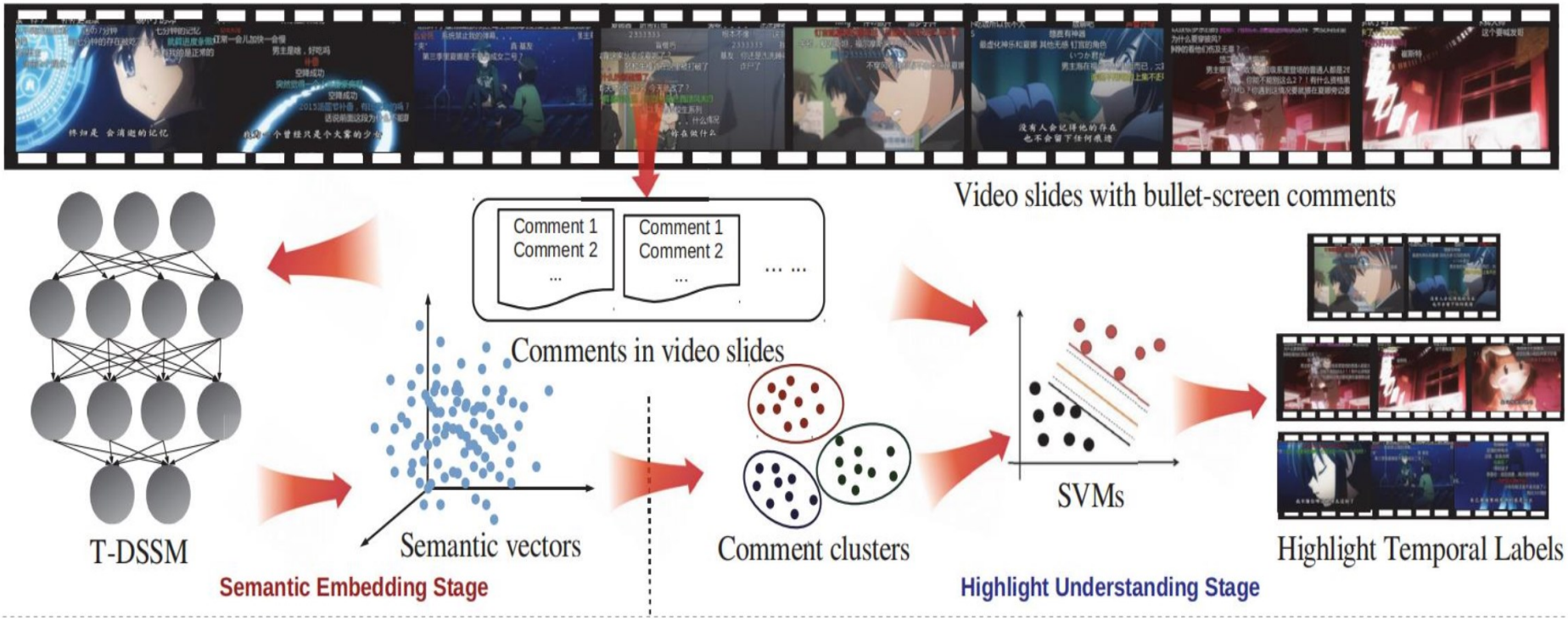
- 多媒体大数据研究热点—网络大数据
  - 社会网络分析
  - 蛋白质结构图
  - 知识图谱





数据科学基础

- 多媒体大数据研究热点—视频大数据
  - 社会网络分析



数据科学基础

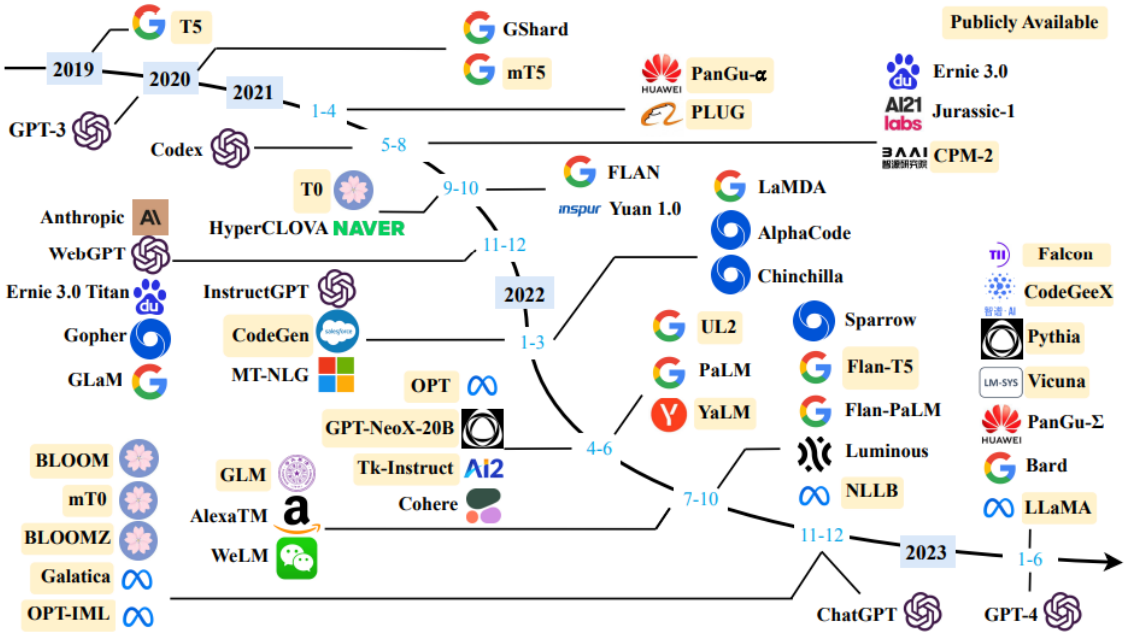
- 多媒体大数据研究热点—文本大数据
  - 自然语言处理
    - 机器阅读，机器翻译，文本推理，知识图谱等

通用语言理解评估（GLUE）基准

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B
1	Alibaba DAMO NLP	StructBERT		90.3	75.3	97.1	93.9/91.9	93.0/92.5
2	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8
3	ERNIE Team - Baidu	ERNIE		90.1	72.8	97.5	93.2/91.0	92.9/92.5
4	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5
5	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5
6	Huawei Noah's Ark Lab	NEZHA-Large		88.7	67.4	97.2	93.2/91.0	92.2/91.6
7	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1
8	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0
9	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9
10	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7
11	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6

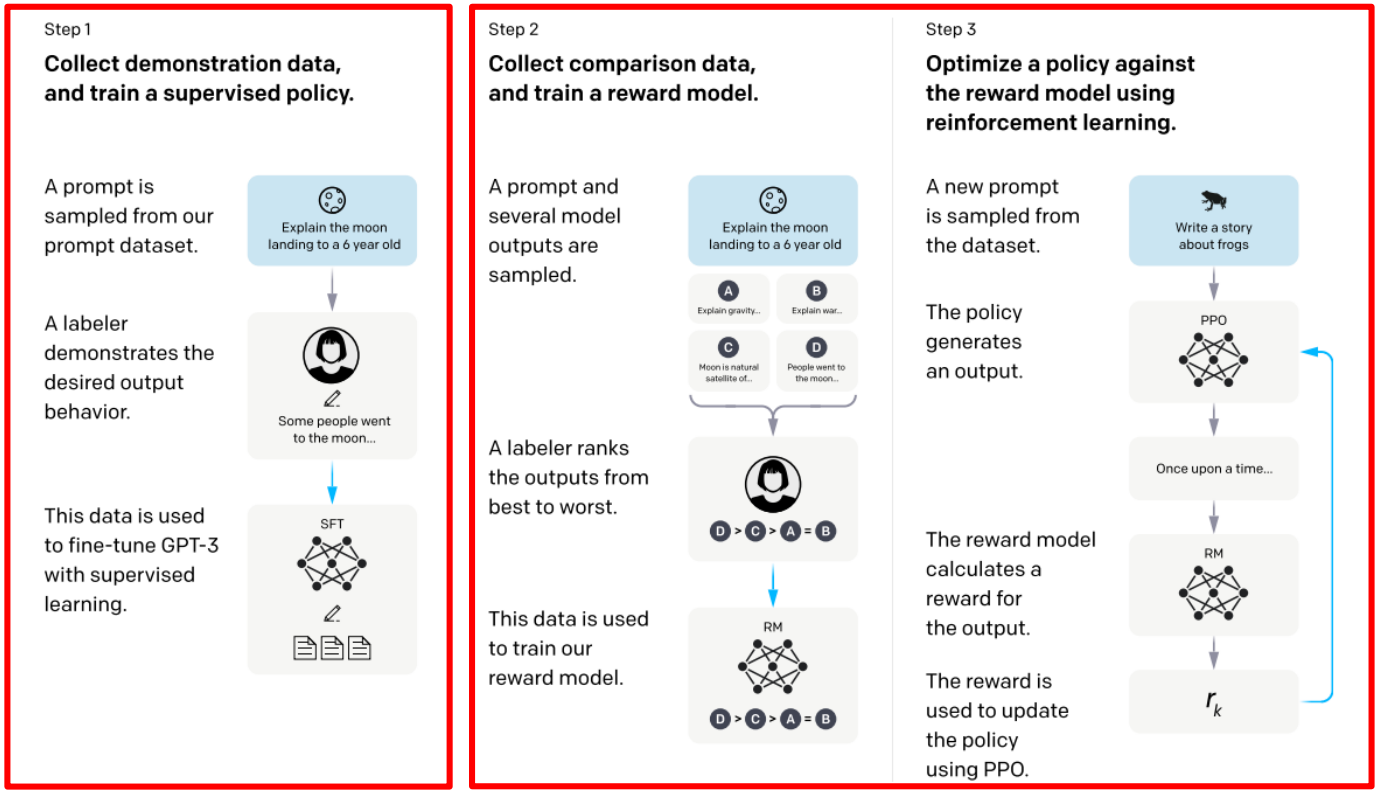
数据科学基础

- 多媒体大数据研究热点—文本大数据
  - 大语言模型——多项技术融合
    - 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。



- 多媒体大数据研究热点—文本大数据
  - 大语言模型

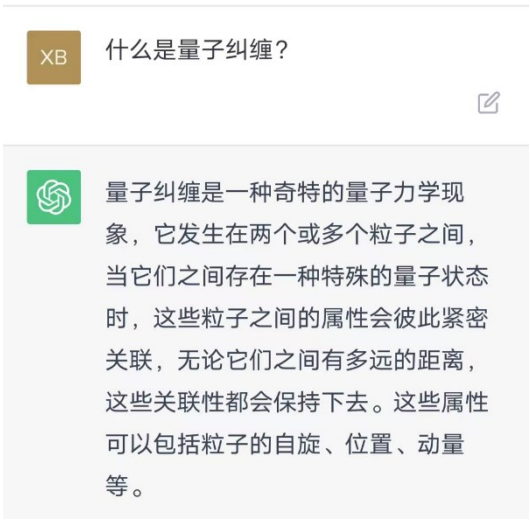
指令微调，使大模型能够泛化到多种任务上



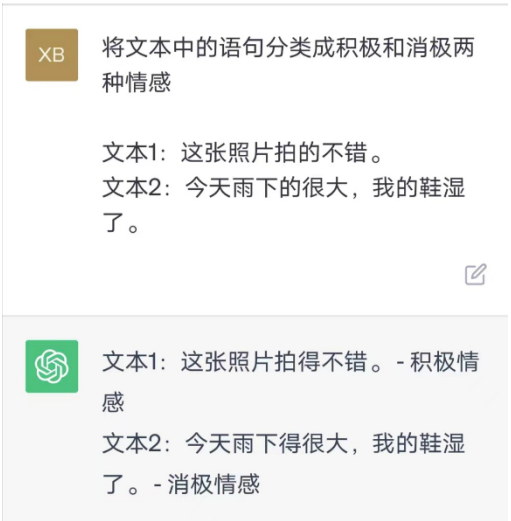
人工反馈的强化学习，使大模型与人类价值观对齐

数据科学基础

- 多媒体大数据研究热点—文本大数据
  - 大语言模型
    - 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



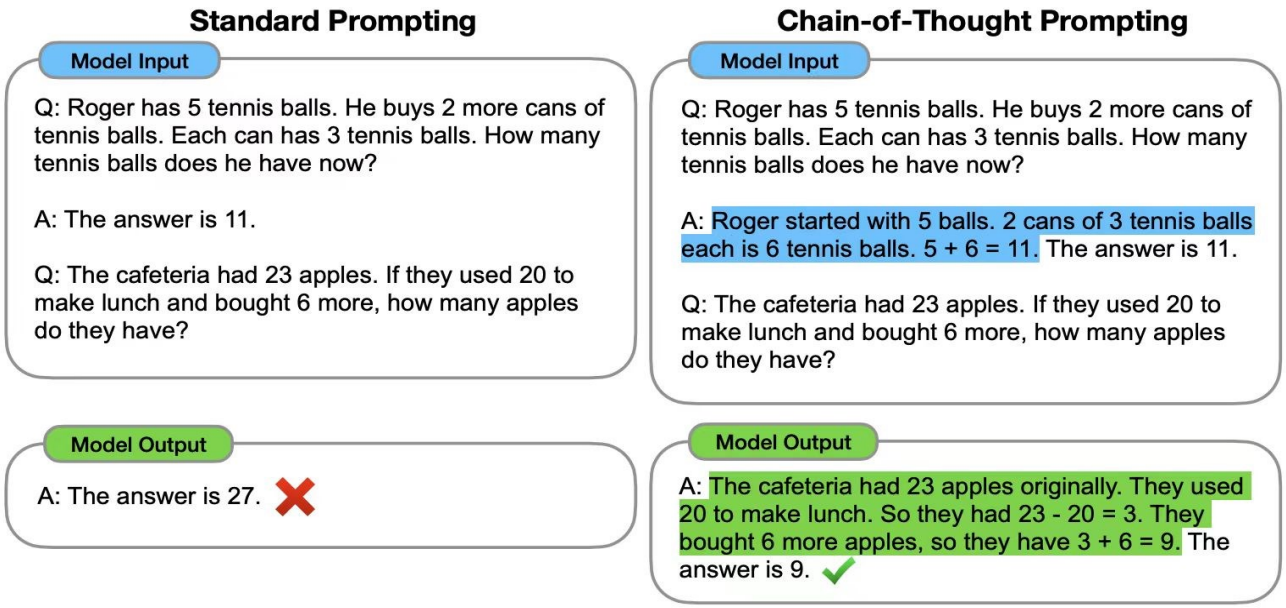
ChatGPT作为知识库回答问题



ChatGPT仅通过任务描述完成文本情感分类

数据科学基础

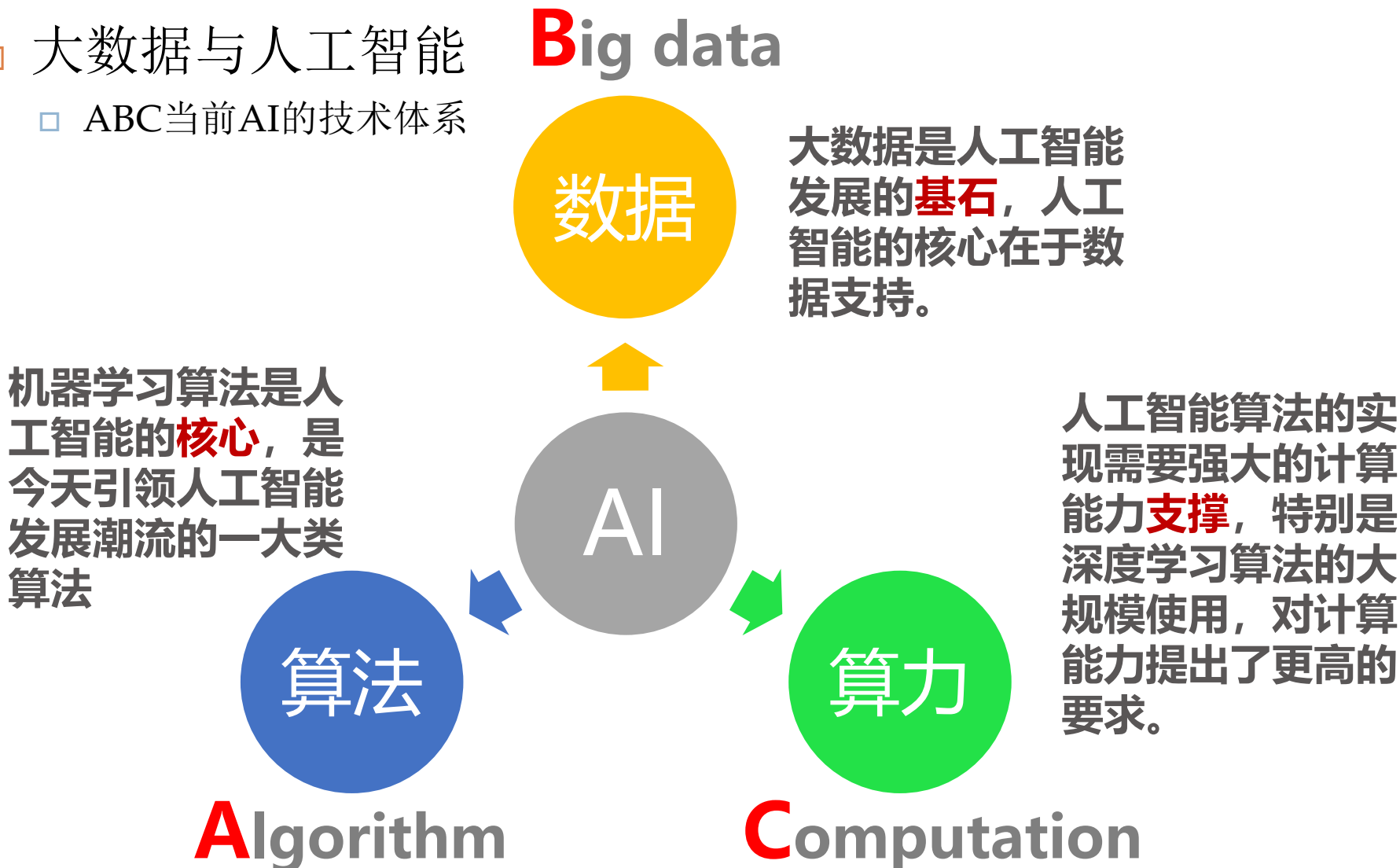
- 多媒体大数据研究热点—文本大数据
  - 大语言模型
    - 通过思维链（CoT）解决复杂问题



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

- 大数据与人工智能
  - ABC当前AI的技术体系

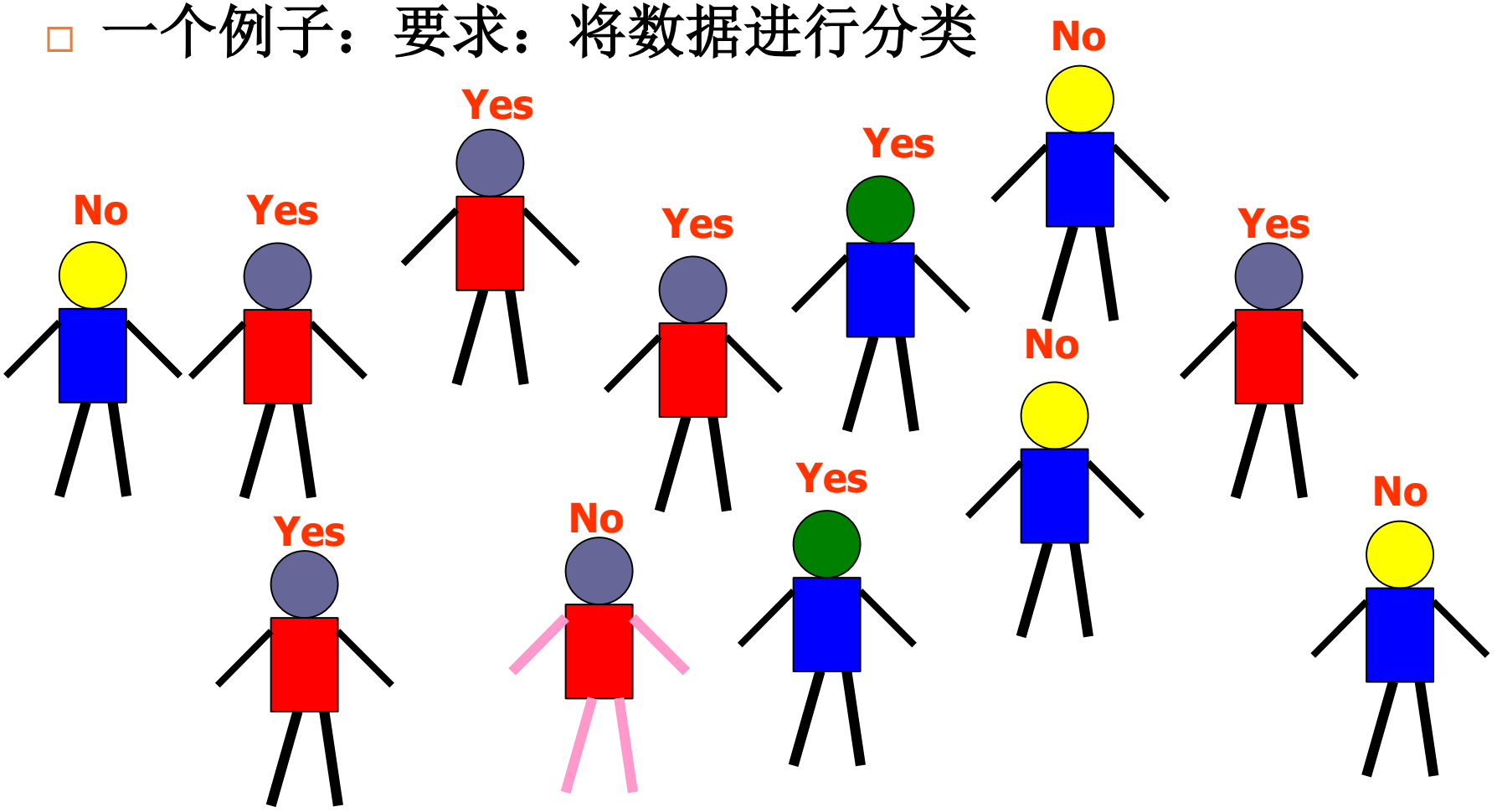


数据科学基础

- 大数据与人工智能
  - 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

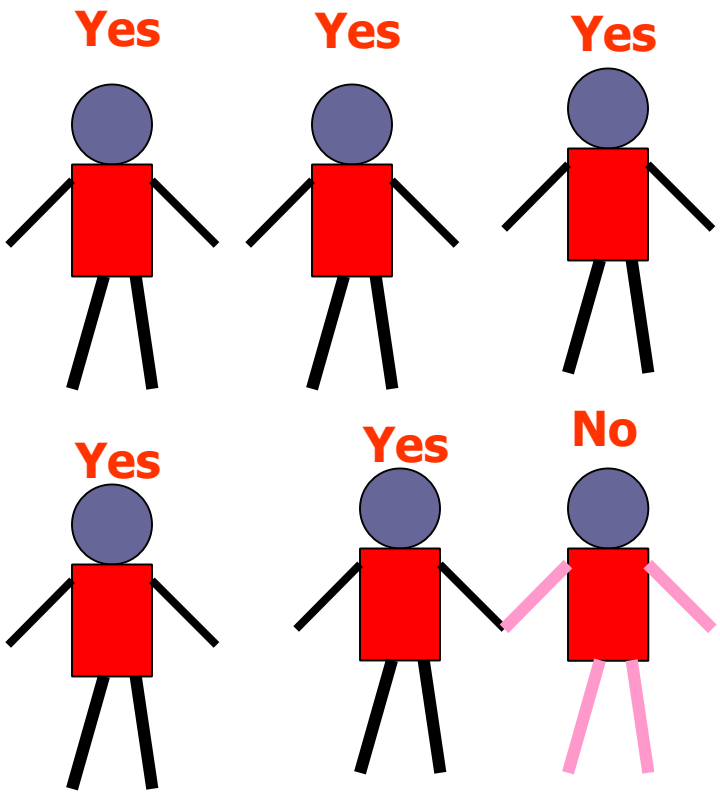
数据+分类学习的方法

- 一个例子：要求：将数据进行分类

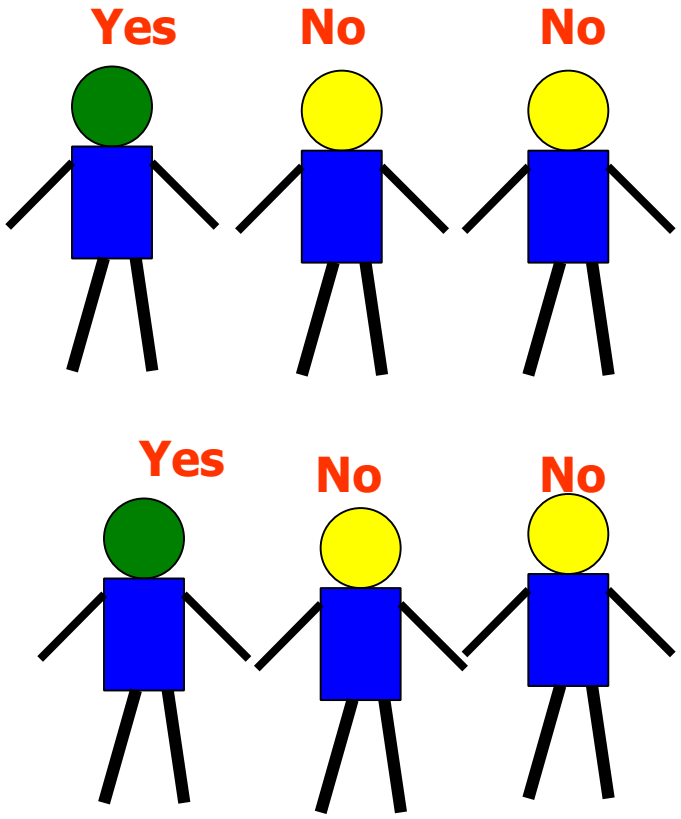


数据+分类学习的方法

躯干：红色

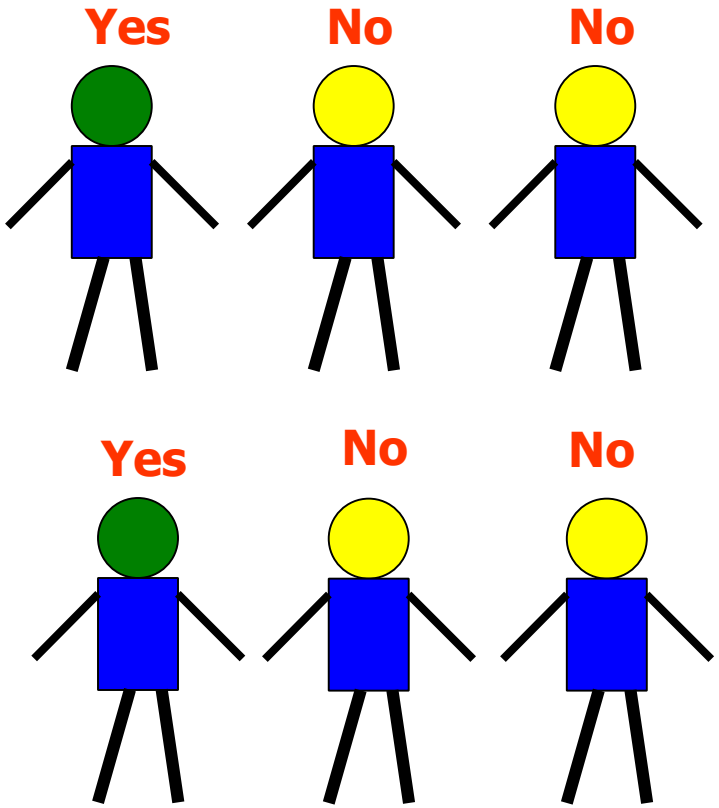


躯干：蓝色



数据+分类学习的方法

躯干：蓝色

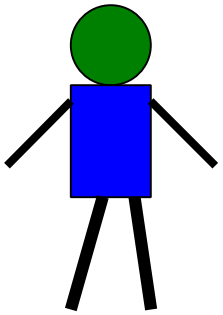


数据+分类学习的方法

躯干：蓝色

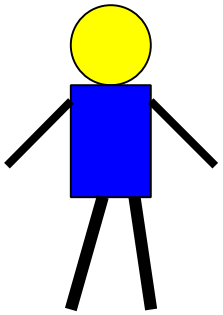
头：绿色

Yes

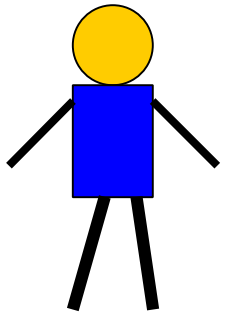


头：黄色

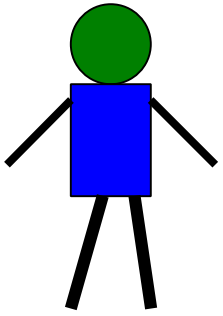
No



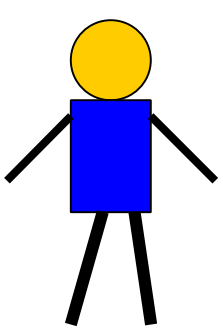
No



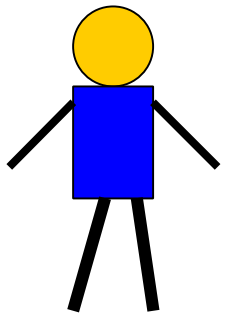
Yes



No

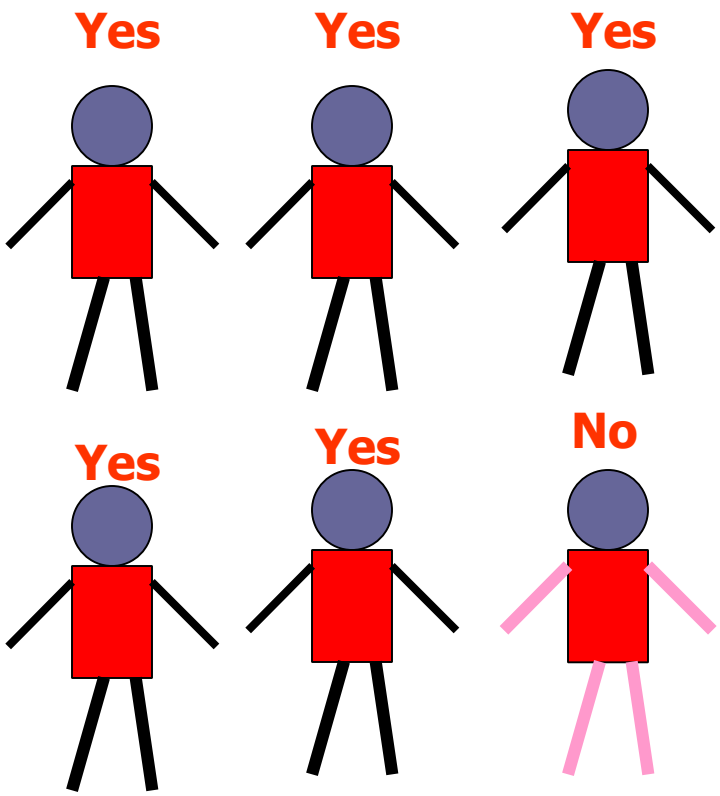


No



数据+分类学习的方法

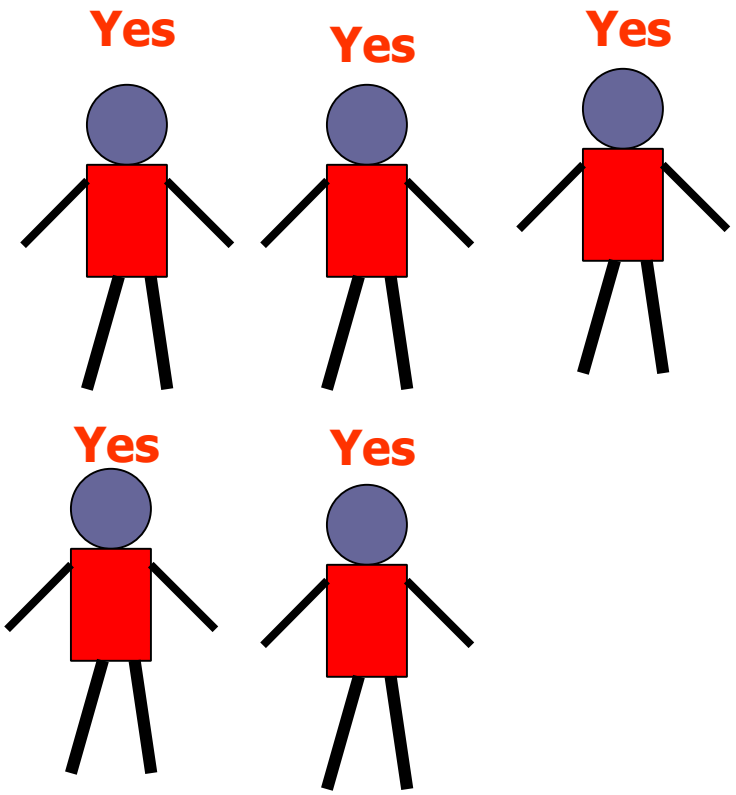
躯干：红色



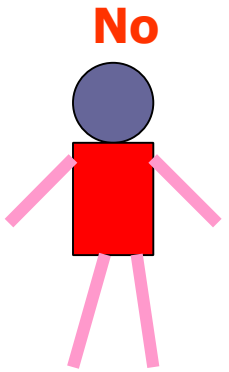
数据+分类学习的方法

躯干：红色

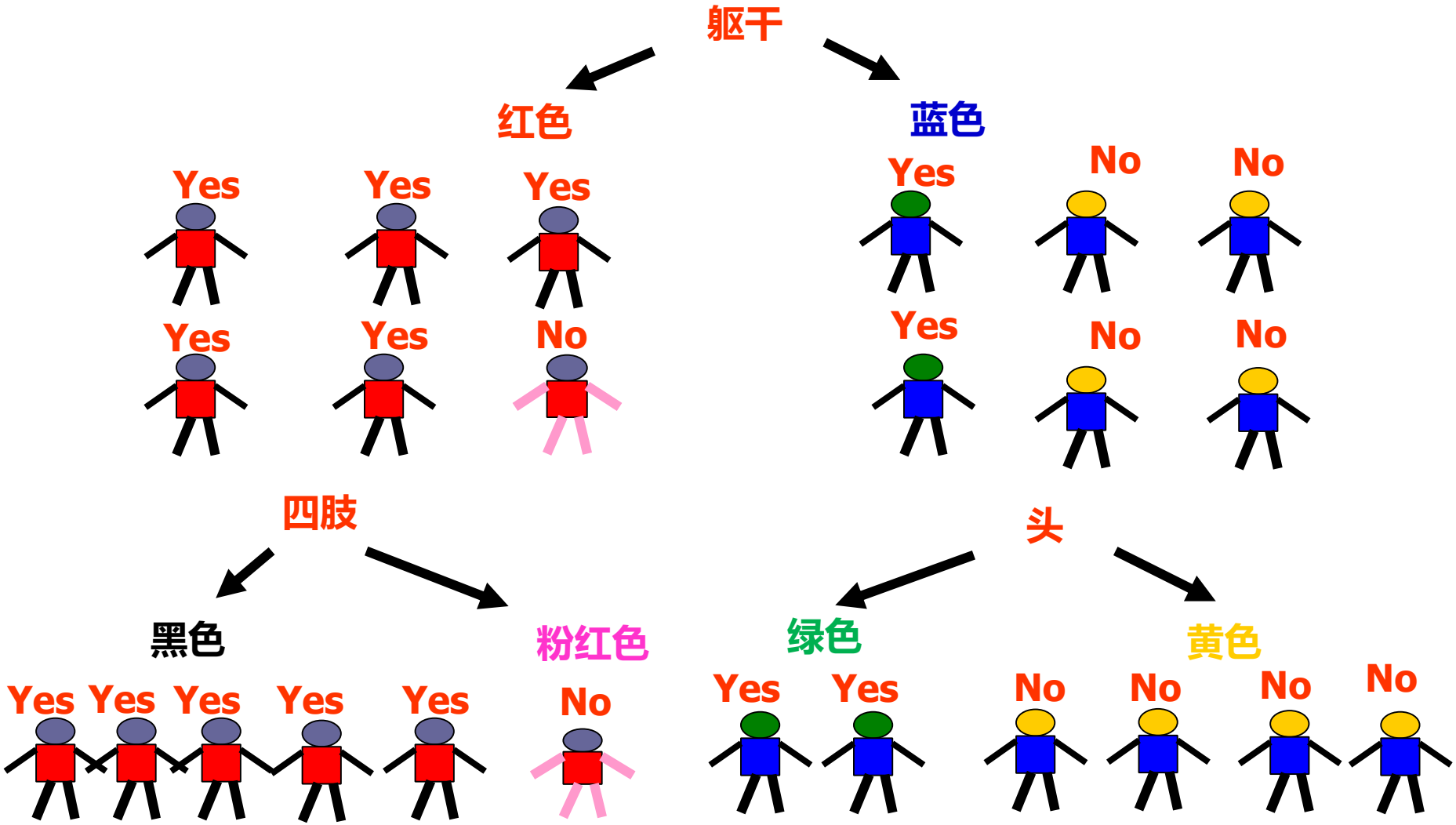
四肢：黑色



四肢：粉红色

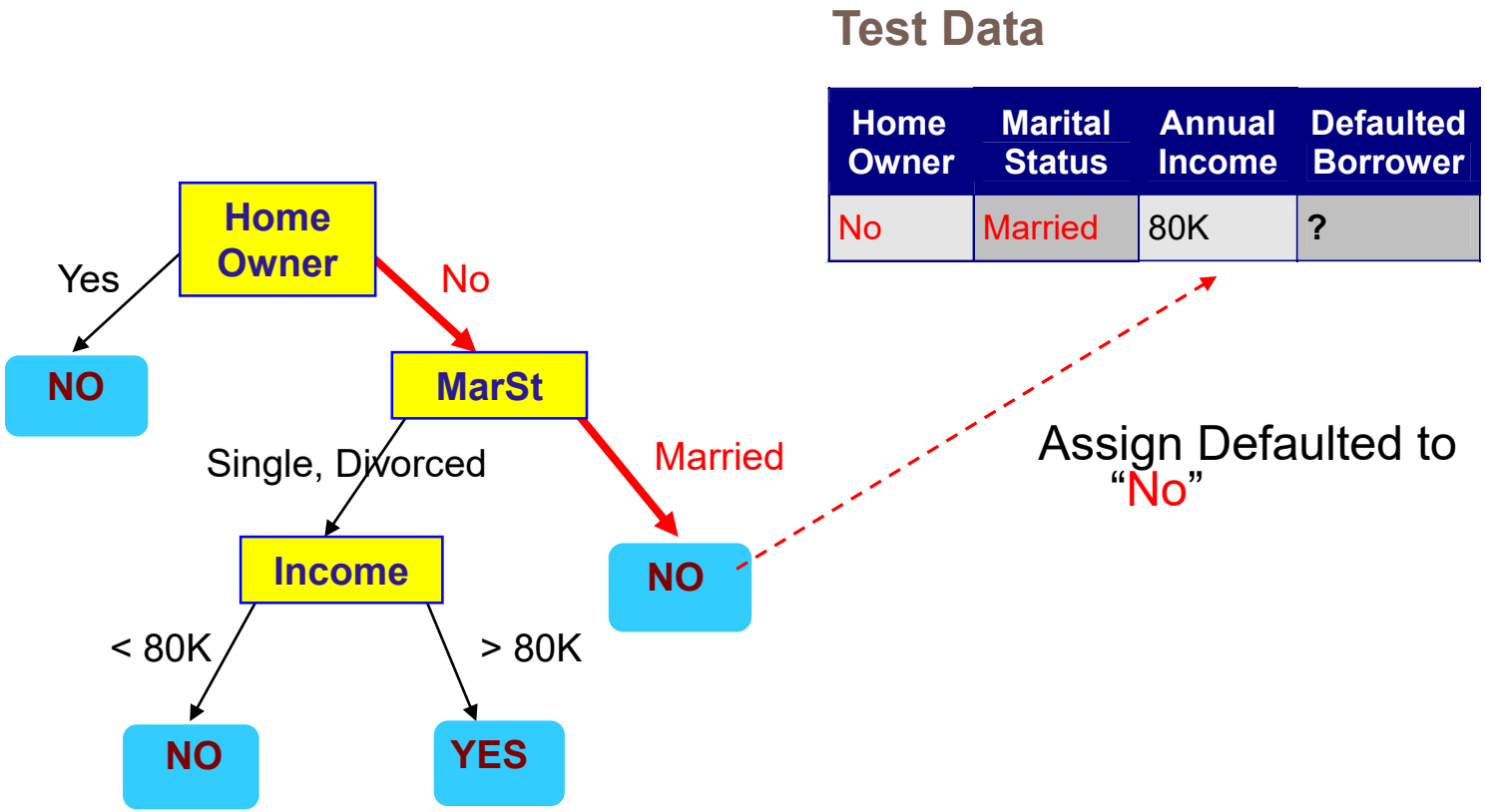


数据+分类学习的方法


躯干
红色
蓝色
Yes
Yes
Yes
Yes
Yes
No
Yes
No
No
Yes
No
No
四肢
头
黑色
粉红色
绿色
黄色
Yes
Yes
Yes
Yes
Yes
No
Yes
Yes
No
No
No
No

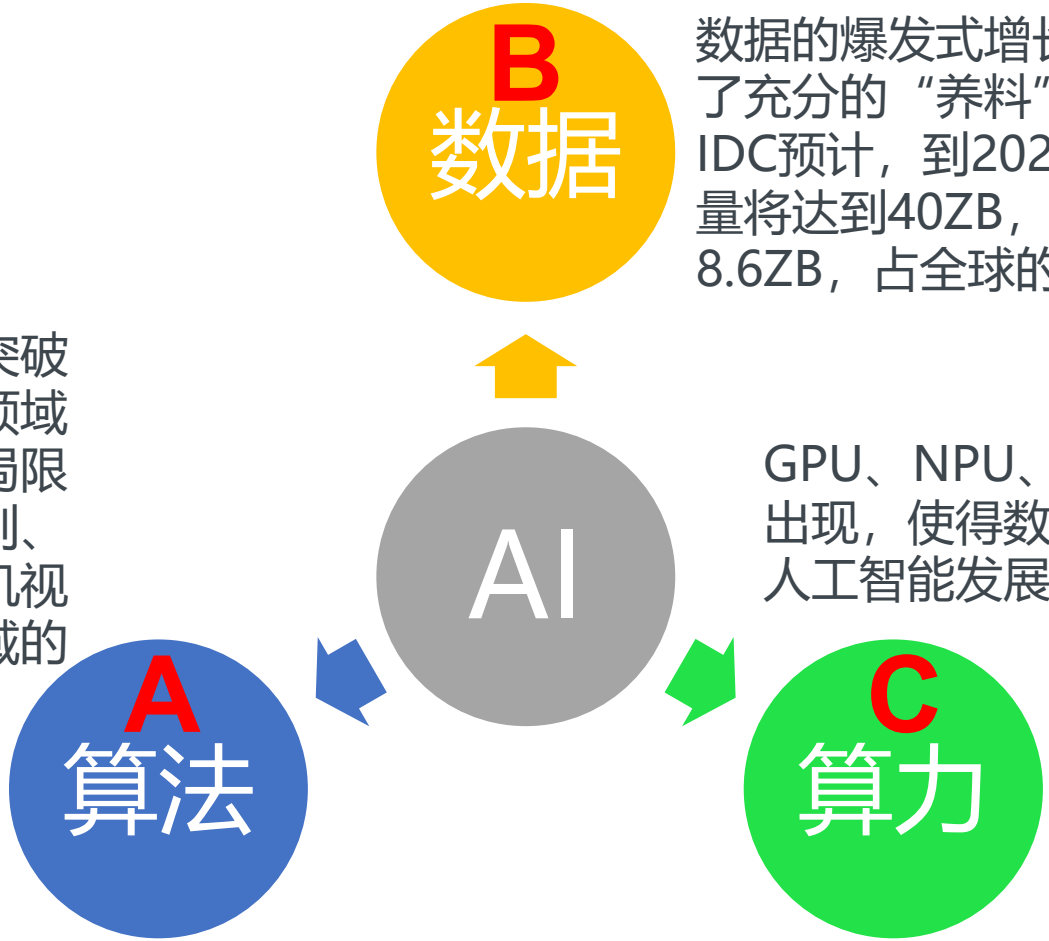
数据+分类学习的方法

- 决策树（第四章）——使用模型对测试数据分类



数据科学基础

- **大数据的未来**—数据驱动人工智能成熟与商业化



数据的爆发式增长为人工智能提供了充分的“养料”，市场调研机构IDC预计，到2020年，全球数据总量将达到40ZB，我国数据量将达到8.6ZB，占全球的21%左右。

深度学习的出现突破了过去机器学习领域浅层学习算法的局限，颠覆了语音识别、语义理解、计算机视觉等基础应用领域的算法设计思路

GPU、NPU、FPGA等专用芯片的出现，使得数据处理速度不再成为人工智能发展的瓶颈

数据科学基础

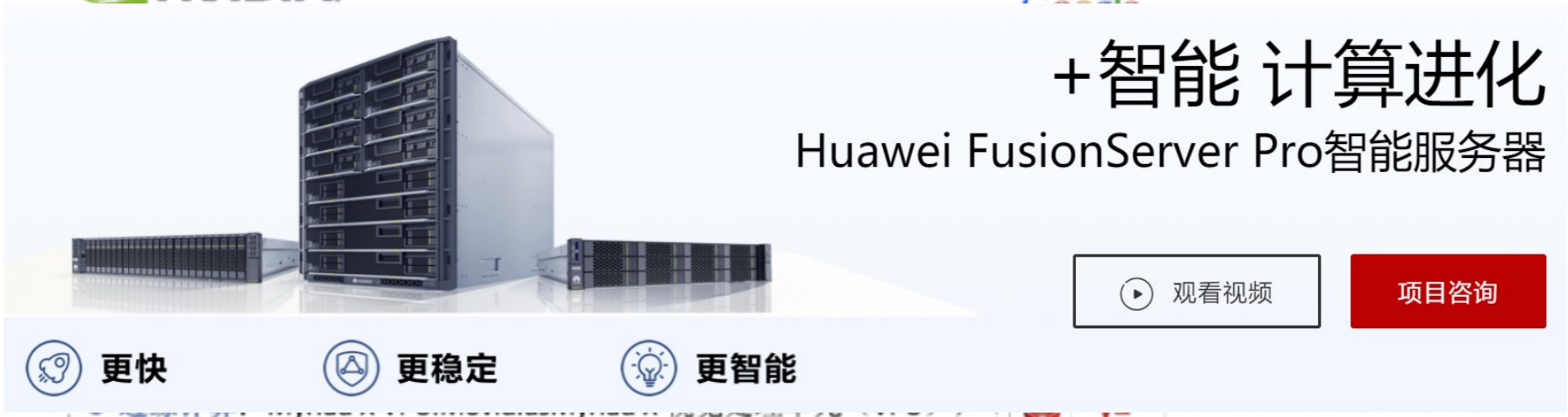
- 包括高效的CPU/GPU、云计算、 AI芯片、多机集群并行化处理等技术手段

AMD
- 云计算：EPYC（霄龙）处理器；Project 47服务器

NVIDIA
- 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



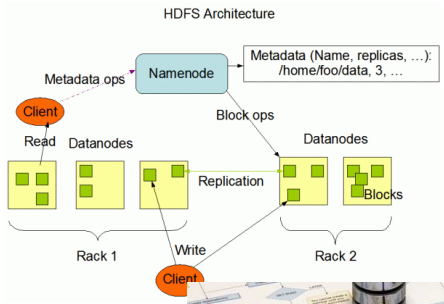
- 移动端：麒麟980芯片

NXP
- 跨界处理器：i.MX RT1060

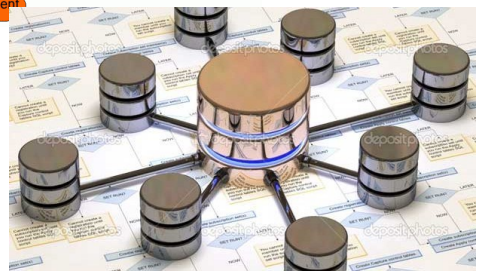
数据科学基础

- 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

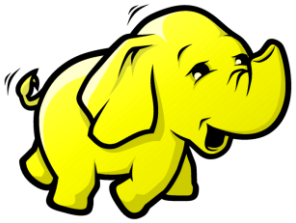
分布式文件系统



关系型数据库



分布式数据库



数据科学基础

- **大数据的未来**—数据驱动人工智能成熟与商业化
  - **向垂直行业渗透**已成为大势所趋
  - 把相关技术赋能给**具体的垂直行业**，比发掘一个适用于所有行业的通用问题好很多
- 从应用成效来看，在电商等领域有较好发展，**一些领域（如农业）没有充分发**

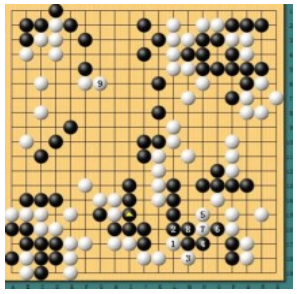
数据科学基础

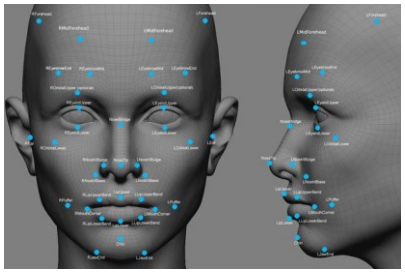
- **大数据的未来**—数据与知识融合，让人工智能更“聪明”

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

- **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

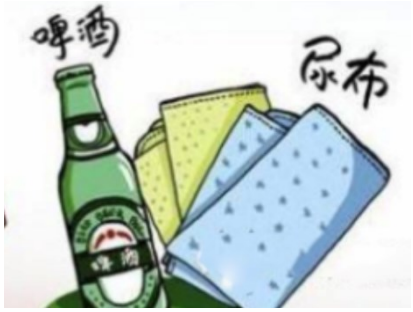
- **归纳法、数理逻辑、布尔代数系统**

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

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

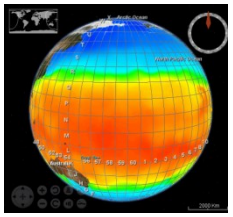
数据分析+逻辑推理

数据科学基础

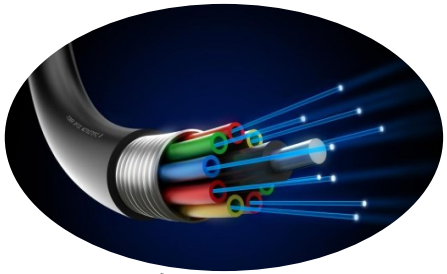

存储（如硬盘、数据库）

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

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）