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/DS2022.html

助教：覃龙虎，刘嘉聿
ds_intro2022@163.com

9/6/2022

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

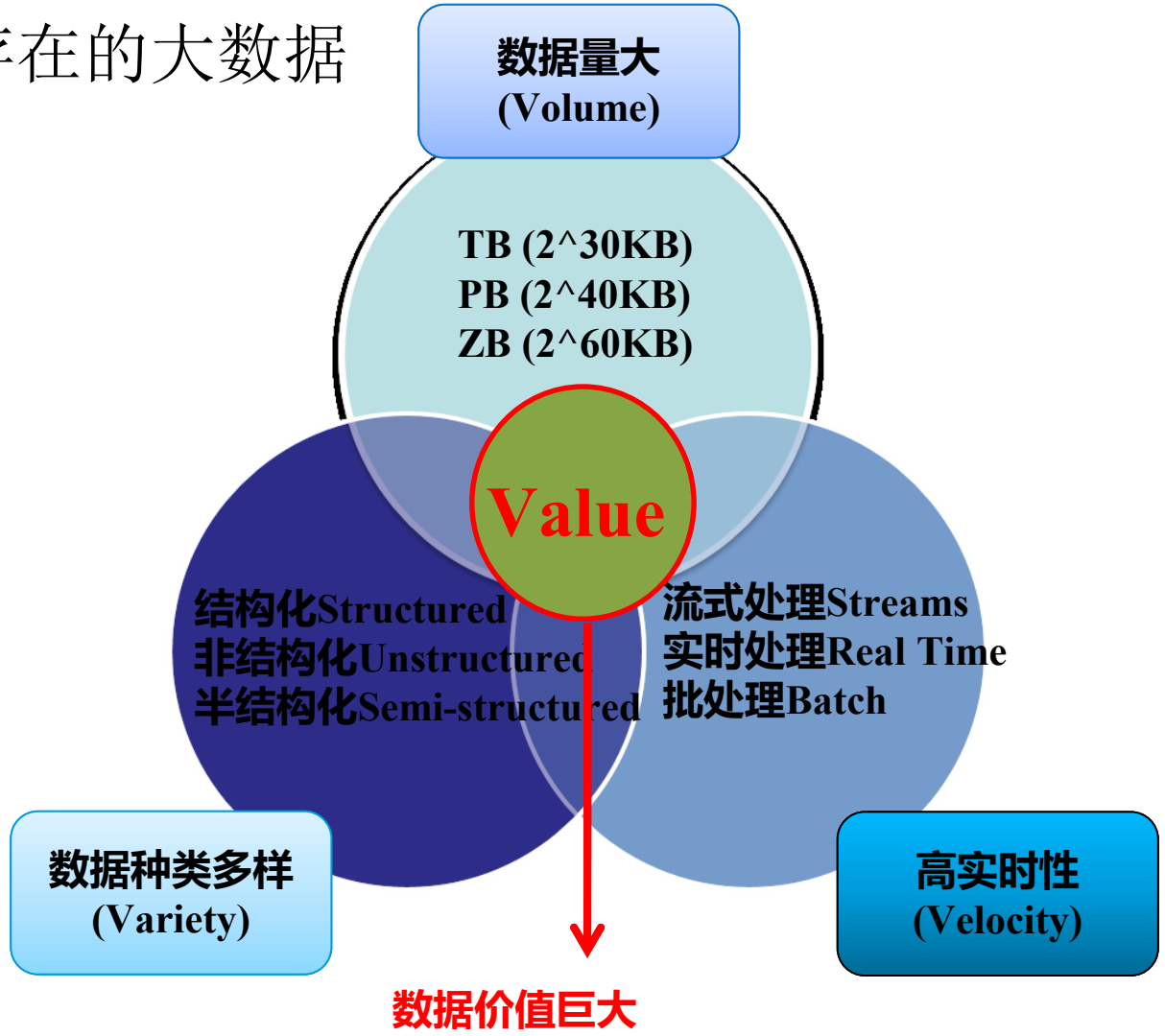
用科学的方法研究和应用数据

选修数据科学与导论课程的同学将来可能从事不同领域的科学研究或者技术开发，

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

- 客观存在的大数据



数据科学基础

- 大数据新工科人才需要具备以下素质

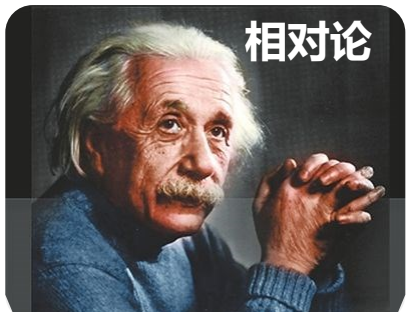
理论基础扎实，能理解运用数据科学中的理论模型

实践能力强，具有处理大数据的能力

跨界能力强，能够解决特定行业的大数据应用问题

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

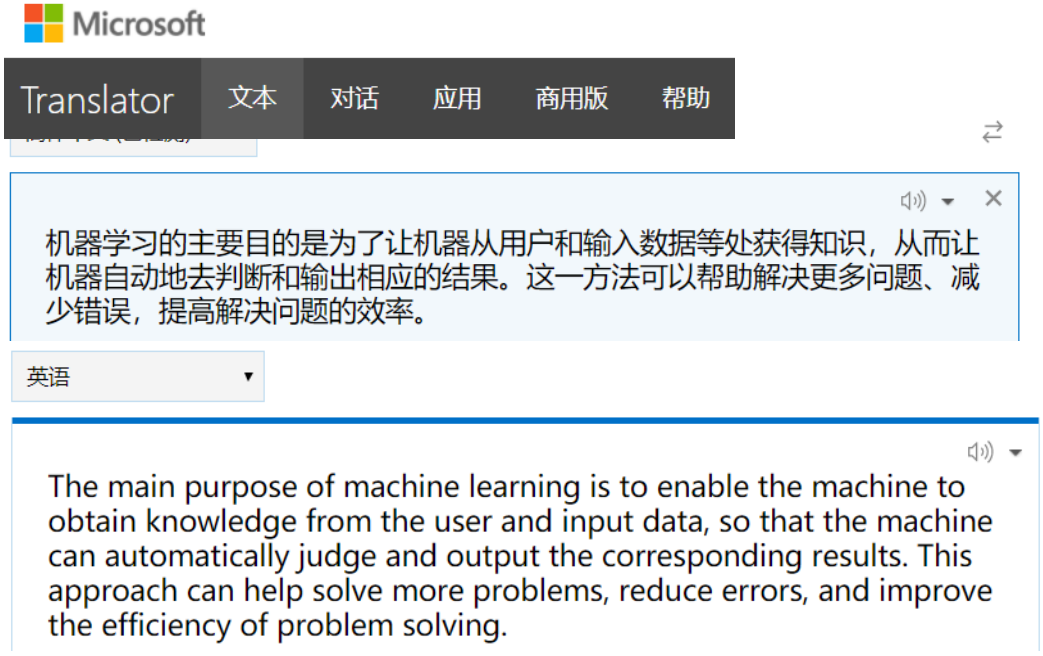
- 把握大数据带来的机遇
- 影像租赁业
  - Winners: 视频流媒体公司(Netflix, Amazon, Hulu)
  - Traditional : DVD租赁公司
- 软件应用业
  - Winners: 软件数据服务公司(Saleforce.com)
  - Traditional : 软件产品公司
- 新闻报纸业
  - Winners: Google, Twitter, Facebook, Bloomberg
  - Traditional : 传统报纸业, Washington Post, WSJ
- 出租车行业
  - Winners: Uber, DiDi

数据科学基础

- 大数据带来的技术创新-当前进展
  - 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。近来，科大讯飞等的语音识别词错率仅有3%左右

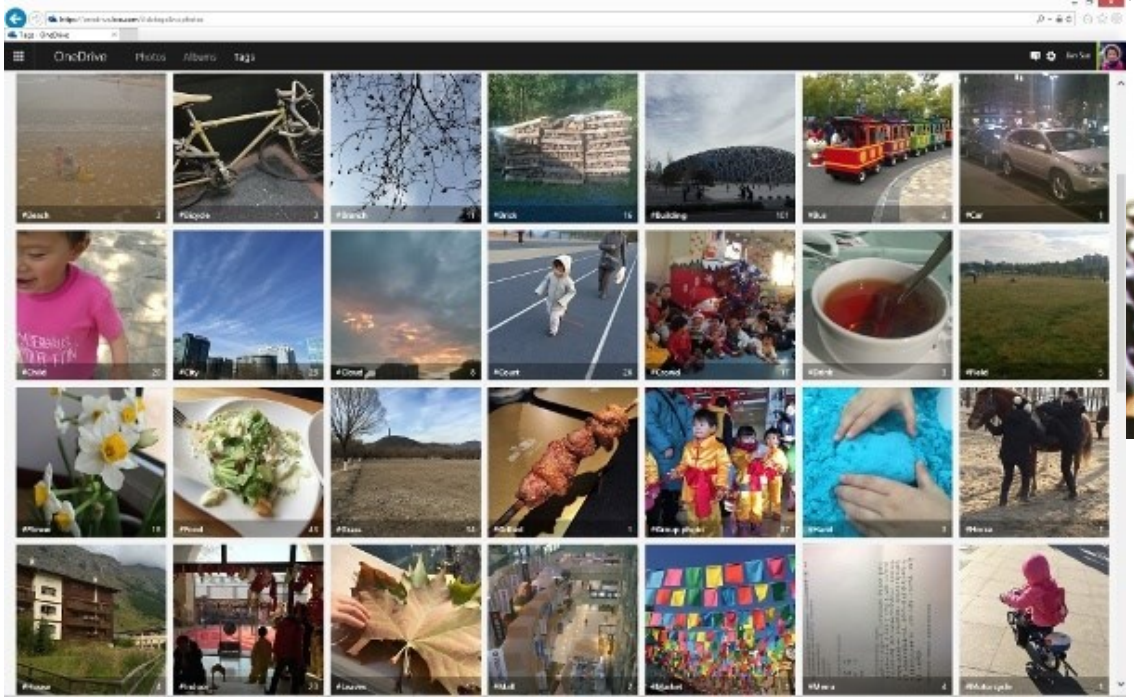
数据科学基础

- 大数据带来的技术创新-当前进展
- 机器翻译
  - 2018年3月，微软亚洲研究院与雷德蒙研究院宣布，其共同研发的机器翻译系统在通用新闻报道测试集newstest2017的中-英测试集上，**达到了可与人工翻译媲美的水平**



数据科学基础

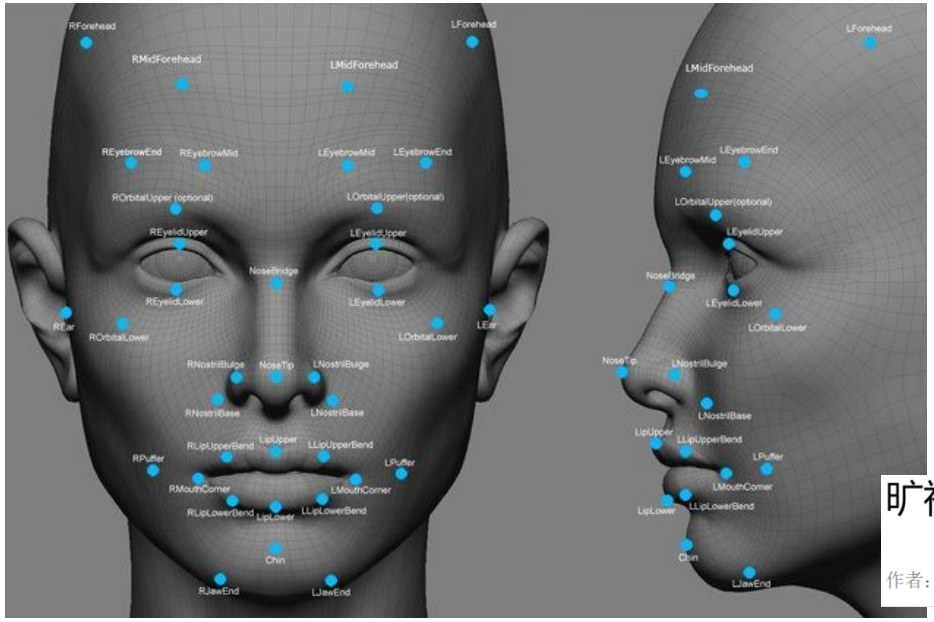
- 大数据带来的技术创新-当前进展
- 图像识别
  - ImageNet图像数据库上，人工智能已达到2.99%的错误率（公安部三所），低于人类5.1%的错误率



李飞飞
斯坦福大学、谷歌Ai前任首席科学家

数据科学基础

- 大数据带来的技术创新-当前进展
- 人脸识别
  - 2014年，Facebook人脸识别系统DeepFace达到97.53%准确率，达到人类水平
  - 不少科研人员从研究机构离职，或自己创办公司，或加入创业公司



SEETA

中科视拓

Face++ 旷视

旷视科技融资4.6亿美元布局城市大脑 创AI融资记录

作者：　　来源：中国证券报・中证网　　2017-11-01 10:55

数据科学基础

- 大数据带来的技术创新-当前进展
- 自然语言处理

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

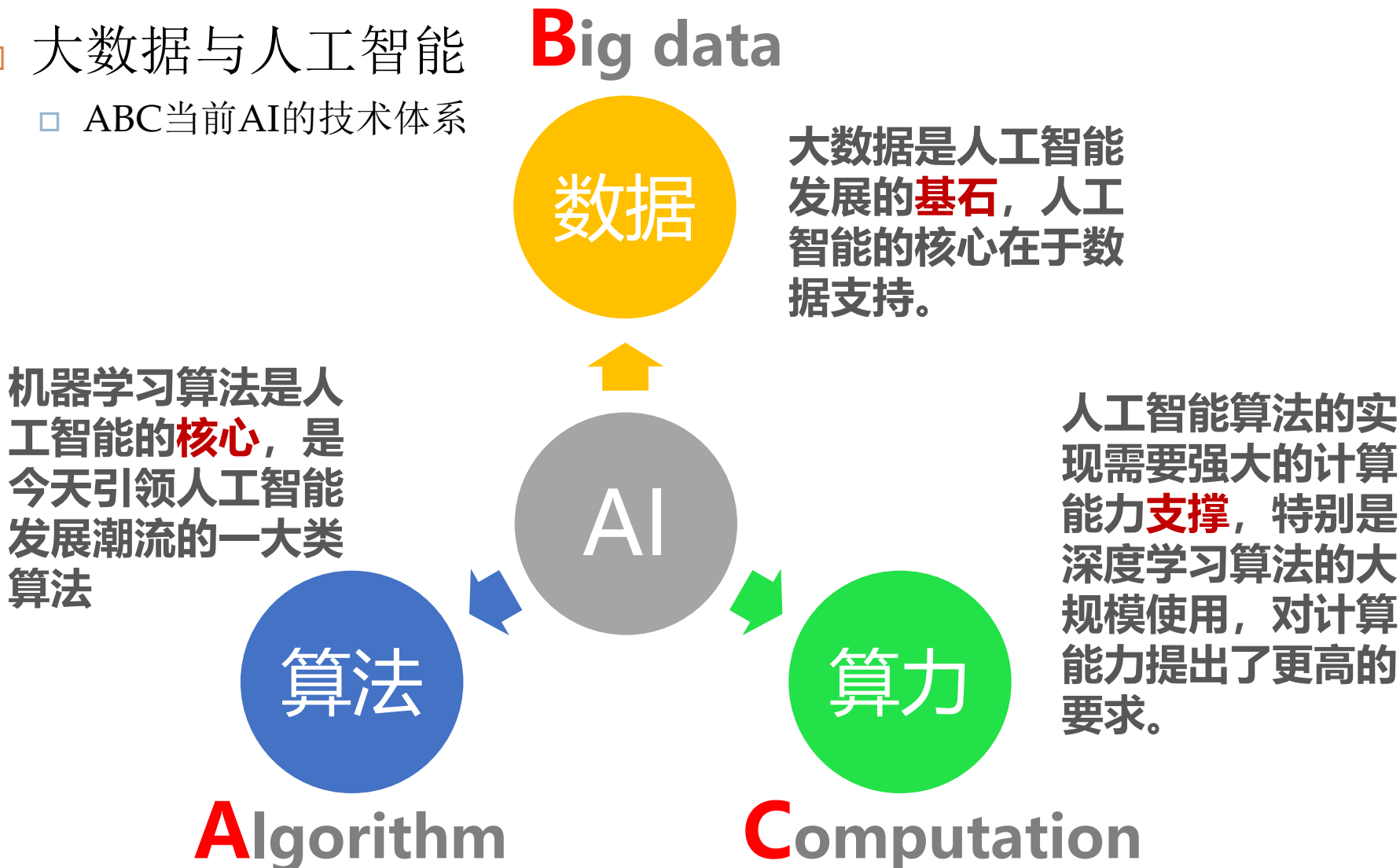
- 大数据带来的技术创新-当前进展

人机对弈

- 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
- 2017年5月，AlphaGo的升级版Master在围棋快棋上击败柯杰，聂卫平等高手，取得60胜0负的战绩
- 2017年10月，AlphaGo Zero从0学起，在不到3 天的时间内以100:0完虐AlphaGo

数据科学基础

- 大数据与人工智能
  - ABC当前AI的技术体系

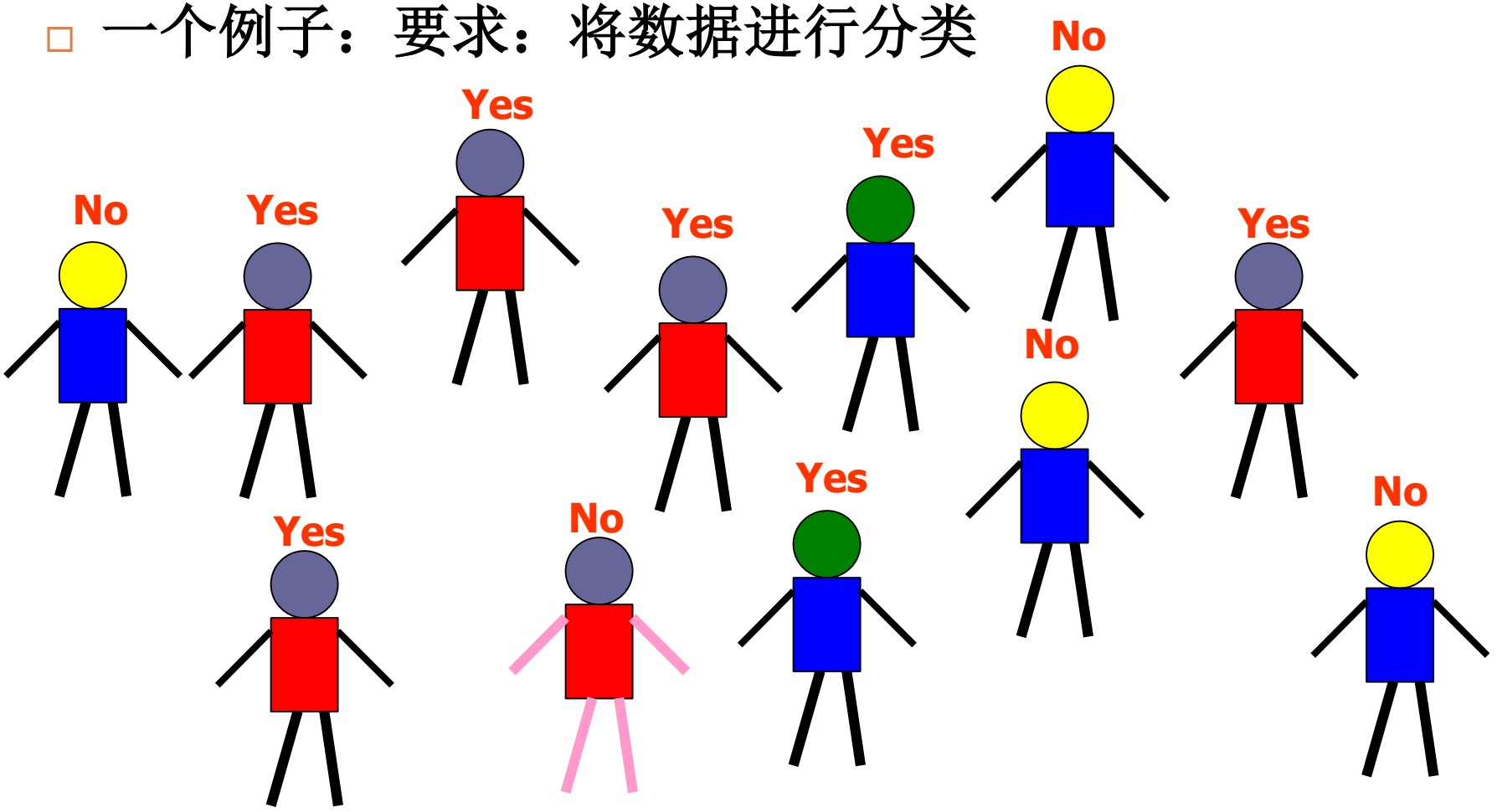


数据科学基础

- 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

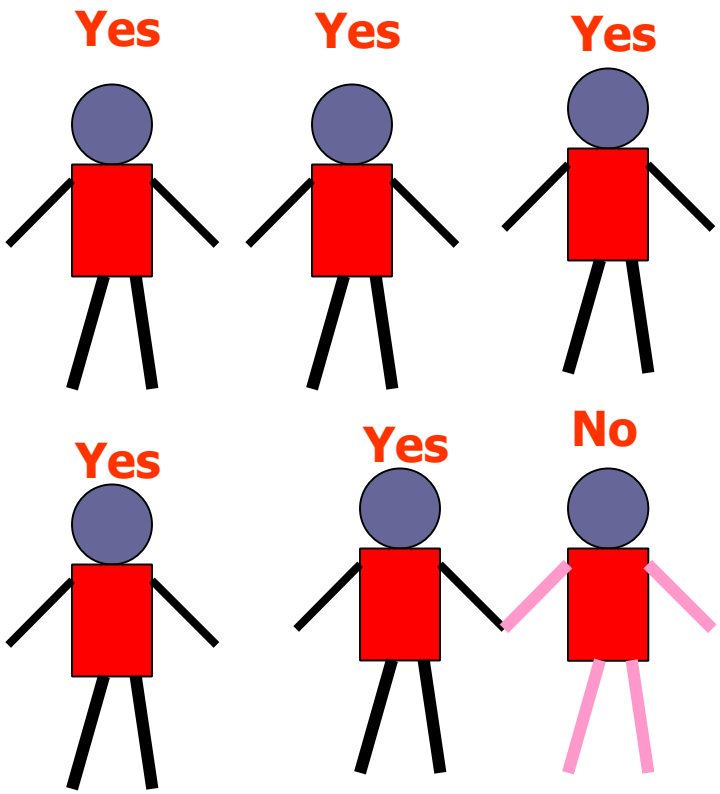
数据+分类学习的方法

- 一个例子：要求：将数据进行分类

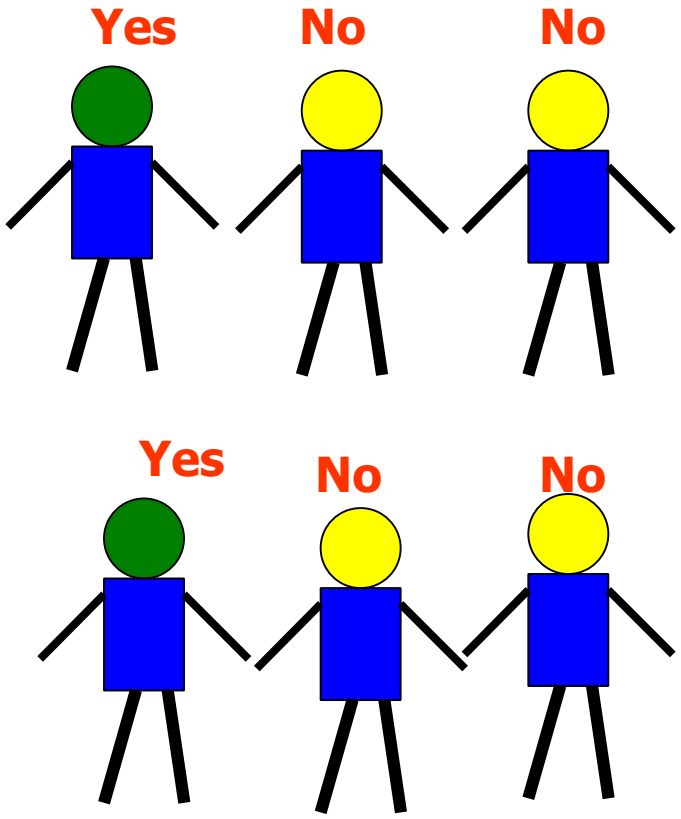


数据+分类学习的方法

躯干：红色

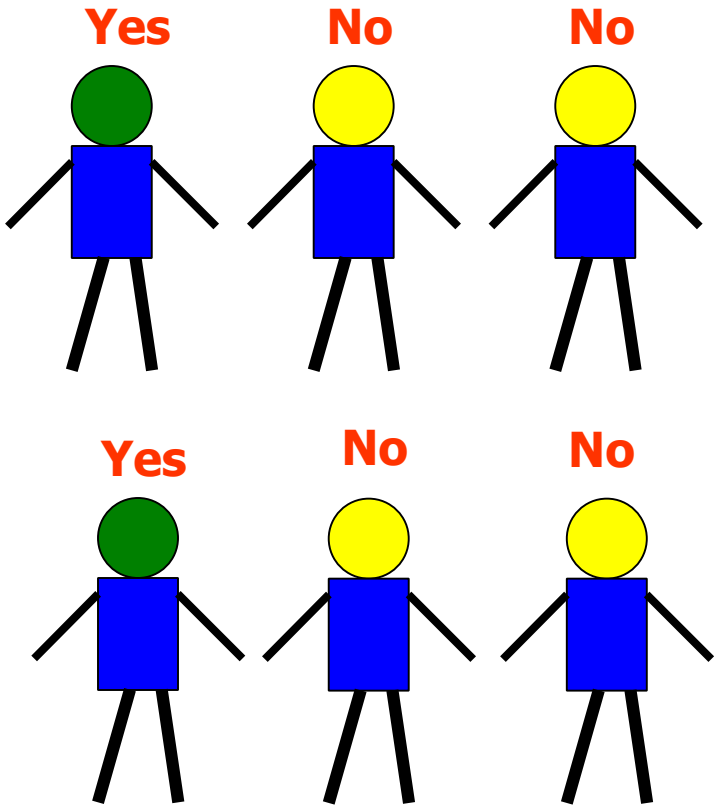


躯干：蓝色



数据+分类学习的方法

躯干：蓝色

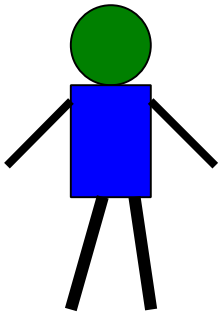


数据+分类学习的方法

躯干：蓝色

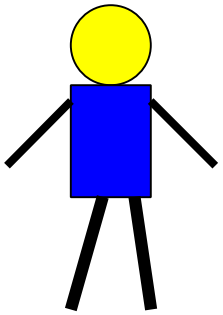
头：绿色

Yes

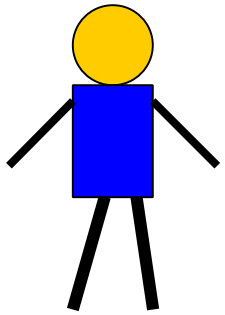


头：黄色

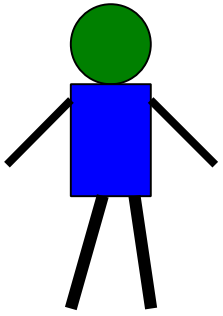
No



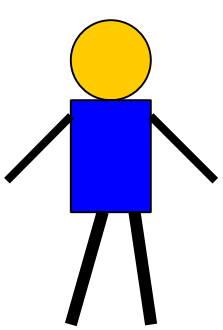
No



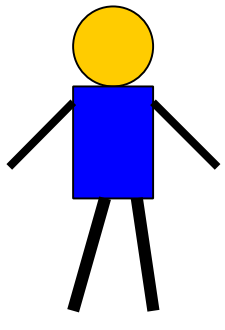
Yes



No

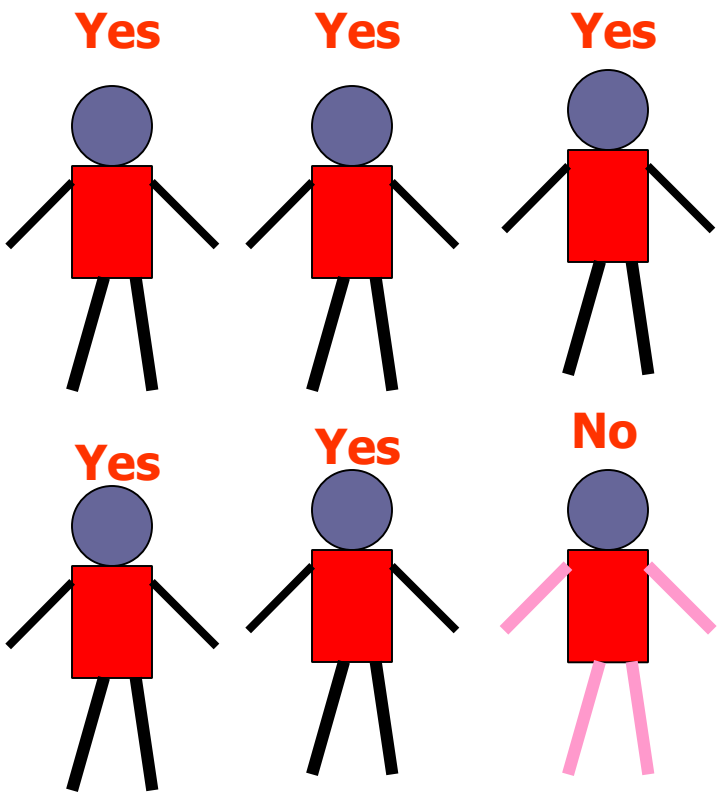


No



数据+分类学习的方法

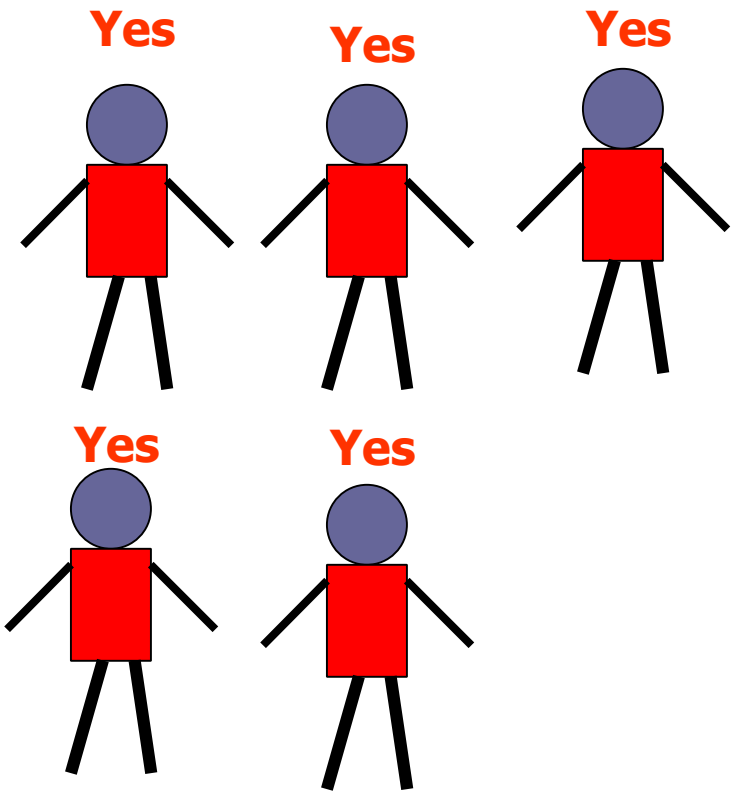
躯干：红色



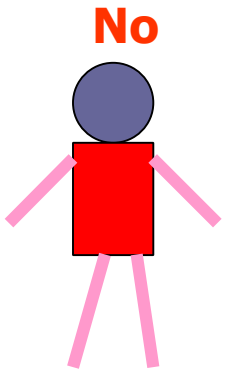
数据+分类学习的方法

躯干：红色

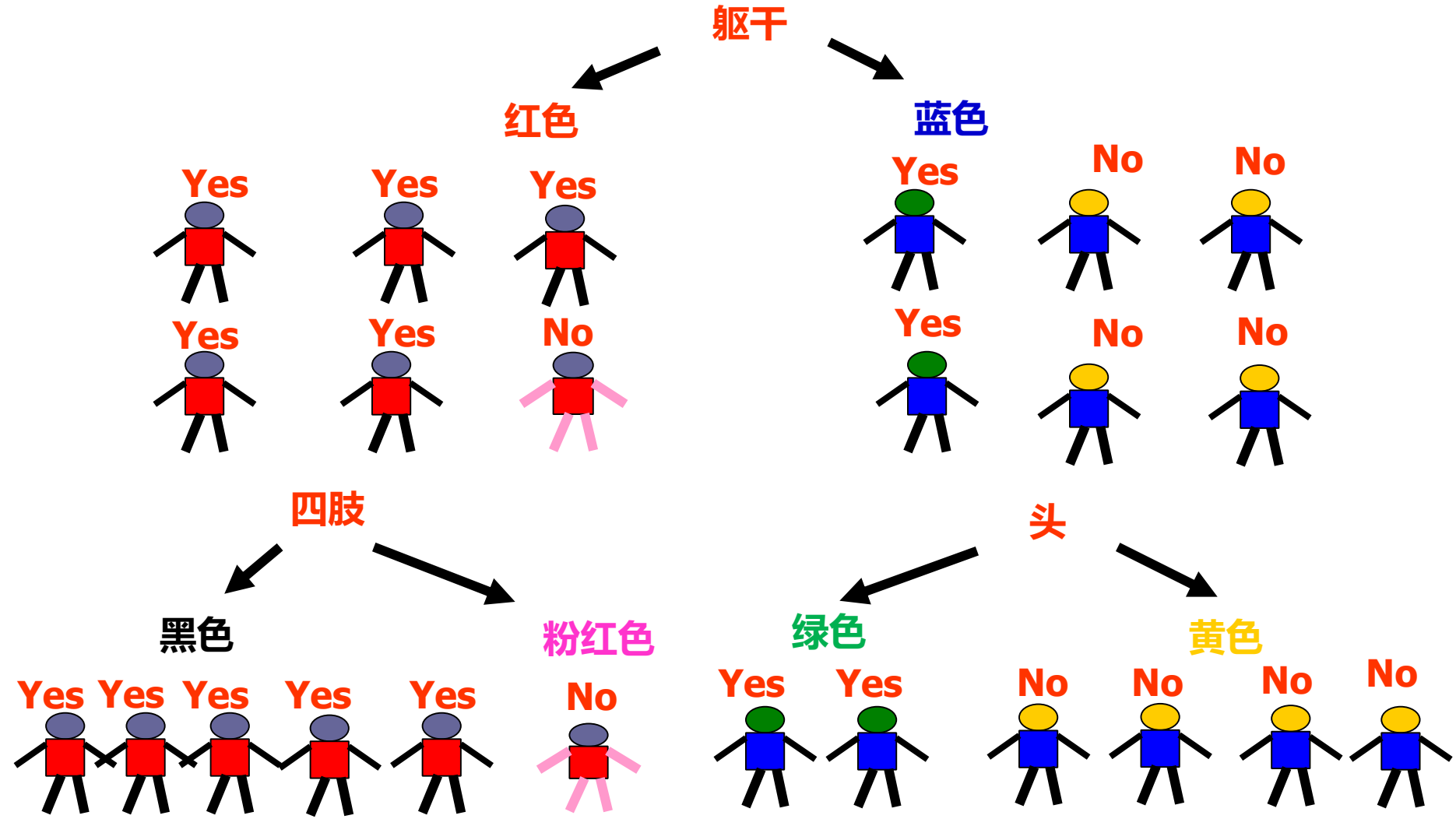
四肢：黑色



四肢：粉红色

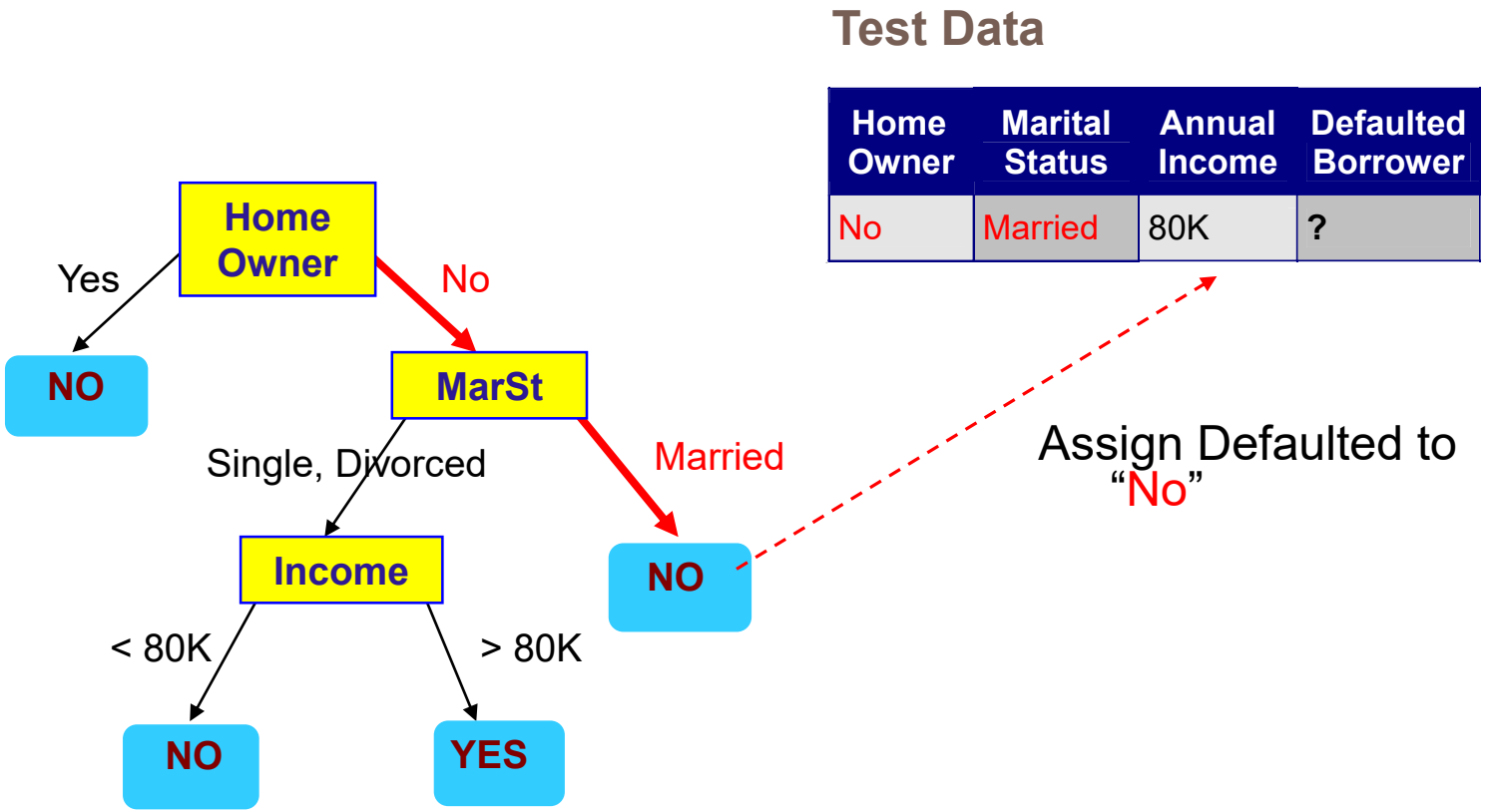


数据+分类学习的方法


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

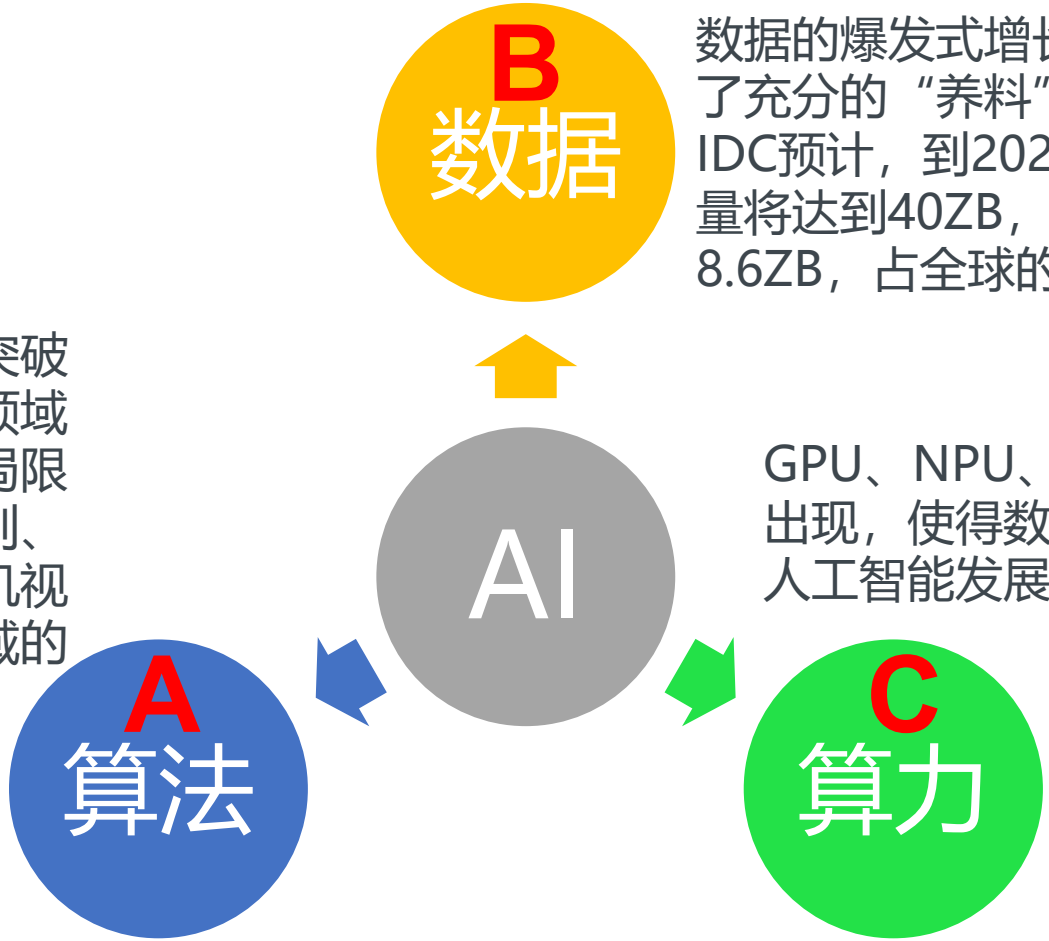
数据+分类学习的方法

- 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

arm
- 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

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

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

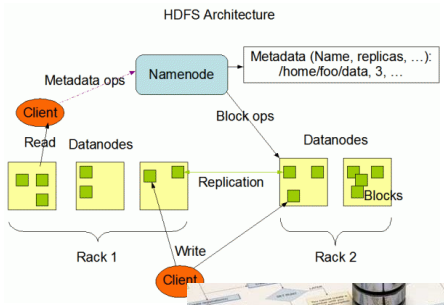
- 移动端：麒麟980芯片

NXP
- 跨界处理器：i.MX RT1060

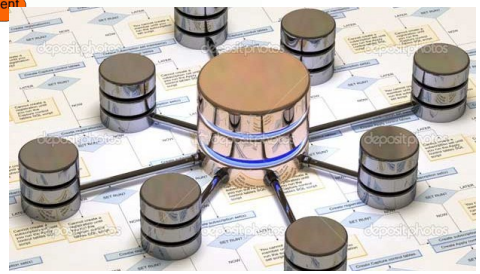
数据科学基础

- 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

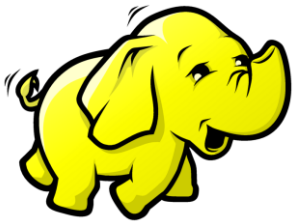
分布式文件系统



关系型数据库



分布式数据库



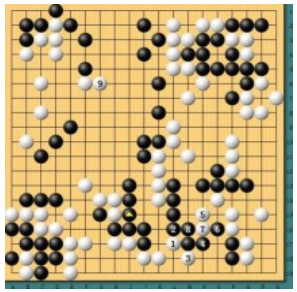
数据科学基础

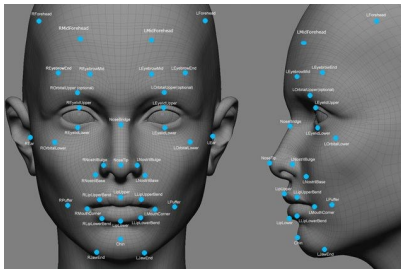
- **大数据的未来**—数据驱动人工智能成熟与商业化
  - **向垂直行业渗透**已成为大势所趋
  - 把相关技术赋能给**具体的垂直行业**，比发掘一个适用于所有行业的通用问题好很多
- 从应用成效来看，在电商等领域有较好发展，**一些领域（如农业）没有充分发**

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	感知智能	认知智能
• 规则明确、特定领域	• 语音、图像、视频	• 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

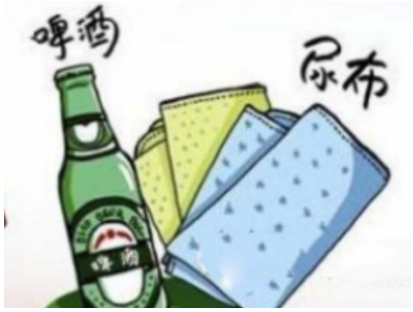
- **归纳法、数理逻辑、布尔代数系统**

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

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

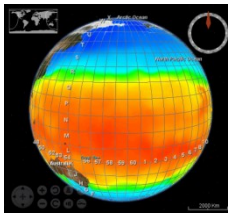
数据分析+逻辑推理

数据科学基础

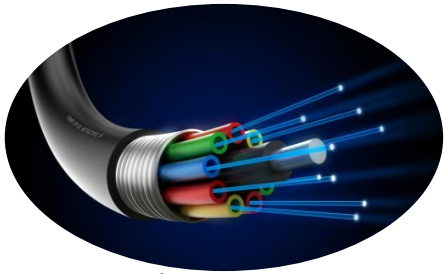
存储（如硬盘、数据库）

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

分析、挖掘和学习



可视化



收集、传输

数据安全与个人隐私

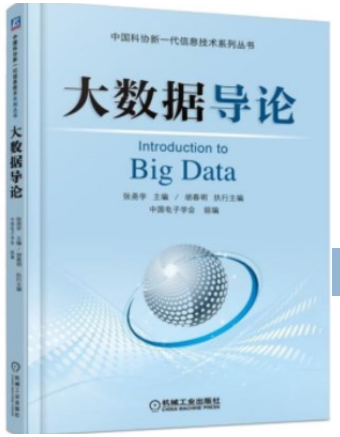
生产、记录

基本程序与算法

……

计算（平台与架构等）

数据科学基础



- **什么是数据科学？**
  - 基于传统的数学、统计学的理论和方法，运用计算机技术进行大规模数据计算、分析和应用的一门学科（**摘录自《大数据导论》**）
  - 用数据的方法研究科学，用科学的方法研究数据（**摘录自鄂维南院士“数据科学的基本内容”**）
- 数据科学的提出
  - 1974年，丹麦计算机科学家Peter Naur(图灵奖得主)提出了“数据学”概念，建议计算机学界不仅要关注科学计算，**也要关注数据处理**
  - 1996年，数据科学(Data Science)这一名称（主要是从**统计学**领域）开始出现
  - 2002年，数据科学杂志创办：主要以科学数据为对象
  - 2007年，Jim Gray提出**数据密集型科学发现**（科学研究的第四范式）
  - 2013年，有些媒体称为中国的大数据元年

课程章节及学时分配（计划）

- 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15-16次课
    - 信息检索与推荐系统
    - 特定类型数据分析：文本挖掘
    - 大数据典型应用：教育大数据分析
    - etc
  - 课程汇报，第17次课
  - 课程回顾，第18次课

教学内容

- 数据分析：数据采集、数据预处理、特征工程
- 数据统计：数据分布、参数估计、假设检验、抽样
- 数据处理工具入门和数据处理实践（Python）
- 数据挖掘：聚类、分类、关联规则、异常检测，etc
- 机器学习：神经网络与深度学习，etc
- 大数据处理平台：分布式存储和处理框架、常用工具
- 文本挖掘：文本处理、主题模型、自然语言处理
- 推荐系统：兴趣建模、协同过滤、深度推荐
- 大数据典型应用：智慧教育、智慧城市等

课程要求与考核方式

- **课程目标：用科学的方法研究和应用数据**
- **课程要求**
  - **文献调研报告 1份**
    - 每人一份
    - 时间节点：第10周上课前
  - **实验报告 1份（需要编程），以下为初步计划**
    - 以小组为单位提交，每小组一份，包含每个人的工作介绍
    - 时间节点：第15周上课前
- **考核方式**
  - 课堂出勤（30%）+调研报告（30%）+实验报告（40%）

任务一：文献调研

- 结合本学期上课的内容，调研相关的文献并撰写报告
  - 文献可以从模型、应用或大数据平台等任一角度进行调研，调研的内容尽可能广，并将学到的内容和心得以报告形式提交
  - 建议：大数据（数据分析）+本人学科
  - 语言要求：中英文不限（包括调研的内容）
  - 格式要求：专业、美观
- 报告内容：
  - 题目
  - 调研结果综述
  - 学习心得与思考
  - 参考文献

任务一：文献调研

- 可以通过以下途径了解更多更前沿的研究领域
  - 期刊及其子刊：
    - Nature, Science, IEEE/ACM Transactions等
  - 国际/国内会议
    - 《中国计算机学会推荐国际学术会议和期刊目录》
    - 《清华大学计算机学科推荐学术会议和期刊列表》
    - 《中国科大计算机学院学位分委会认定的顶级会议与期刊》
    - 《信息与智能学部学位分委员会认定的A档/B档会议与期刊》
    - 例如，KDD, WSDM, ICDM, AAAI, IJCAI, SIGIR, CIKM, NIPS, ICML, VLDB, SIGMOD, ICDE, ACL,EMNLP, WWW
  - 各种科技论坛、公众号
    - 机器之心，新智元，专知
  - 在线课程

任务一：文献调研

- 例如，从模型角度出发：
  - 采样方法
    - MCMC - Metropolis-Hastings
    - MCMC - SLICE SAMPLING
  - 分类
    - 决策树
    - 支持向量机(SVM)
    - 贝叶斯分类器
  - 聚类
    - K均值算法(K-means)
    - 基于密度的聚类(DBScan)
    - 层次聚类
  - 概率主题模型
    - 隐马尔科夫模型(HMM)
    - 隐狄利克雷分配模型(LDA)
  - 深度学习
    - 卷积神经网络(CNN)
    - 循环神经网络(RNN)
  - 强化学习
    - 马尔科夫决策过程(MDP)
    - 深度Q网络(DQN)

任务一：文献调研

- 例如，从应用角度出发（推荐几篇）
  - 实验室主页上可下载http://staff.ustc.edu.cn/~cheneh/
  - **教育**
    - EKT: Exercise-aware Knowledge Tracing for Student Performance Prediction
    - DisenQNet: Disentangled Representation Learning for Educational Questions
  - **金融**
    - Product Supply Optimization for Crowdfunding Campaigns
    - P2P Lending Survey: Platforms, Recent Advances and Prospects
  - **推荐系统**
    - Relevance meets Coverage: A Unified Framework to Generate Diversified Recommendations
    - Personalized Travel Package Recommendation
  - **社交网络**
    - An Influence Propagation View of PageRank
    - From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring

文献调研-评分标准

- **多文献**调研报告
  - 独立写作程度如何、是否抄袭
  - 是否有明确主题、主题是否符合课程内容、表述是否合适
  - 是否有背景和现状综述
  - **对现状综述的总结是否有框架、有逻辑**
  - 对各文献的单独表述是否清晰具体
  - **是否对各文献涉及模型和方法做有意义的比较分析**
  - 逻辑表达是否通顺
  - 参考文献是否完整
  - 提交是否及时

文献调研-评分标准

- **单文献**调研报告
  - 独立写作程度如何、是否抄袭
  - 模型或方法部分表述是否清晰
  - 实验部分表述如何
  - **是否切中论文核心贡献**
  - **对论文是否有自己的思考**
  - 逻辑表达是否通顺
  - 参考文献是否完整
  - 提交是否及时

任务二：实践与实验

- 参与实验，撰写实践报告
- 实践部分重点：大家在实践中熟悉数据科学知识，锻炼团队合作能力，报告中叙述清楚、内容合理即可
- 完成一个实际的大数据分析任务
  - 下载数据，完成相应任务
  - 参加比赛
  - 需要coding
- 编程语言：Python
- 工具包：numpy，sklearn，matploit，pytorch

实践与实验

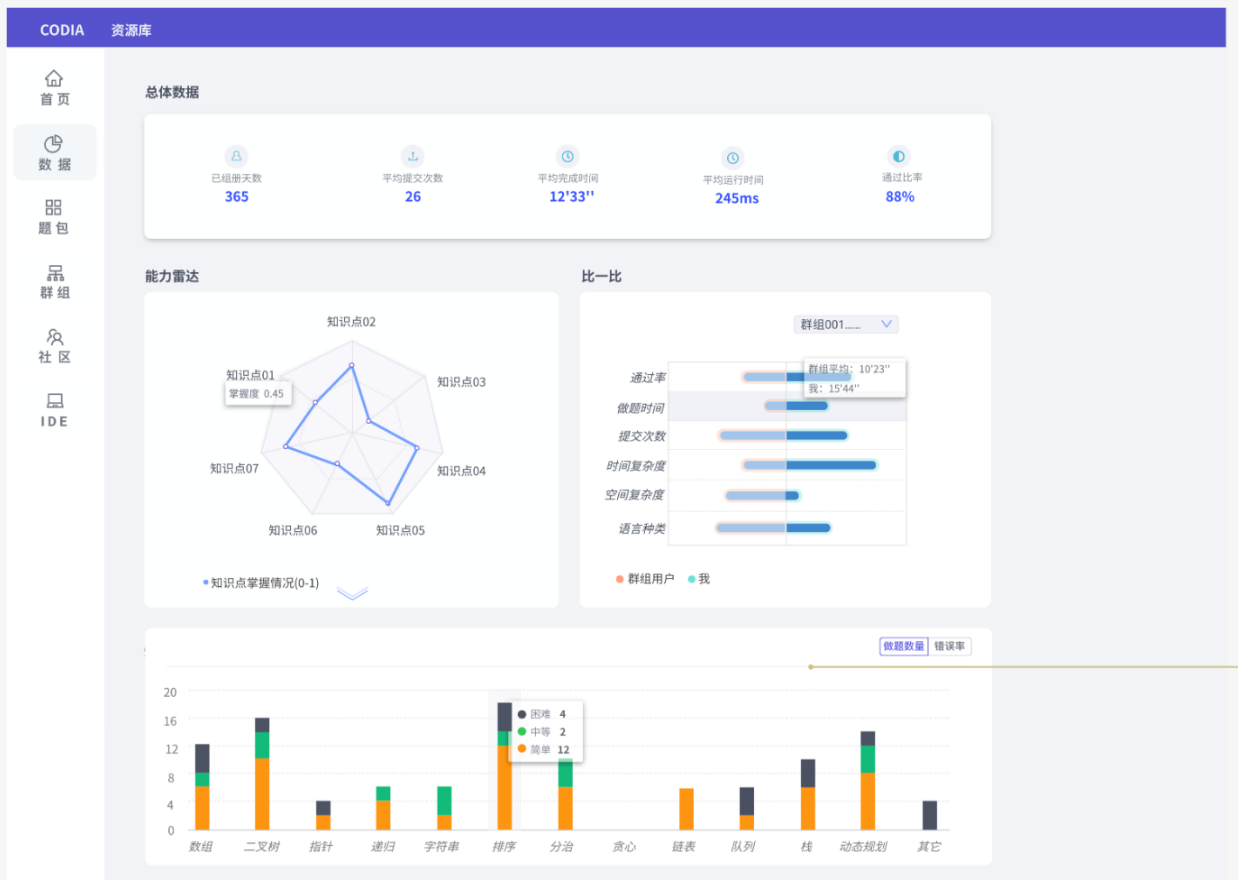
- 推荐大家使用实验室平台练习编程
  - CODIA：https://code.bdaa.pro

实践与实验

- 推荐大家使用实验室平台练习编程
  - CODIA：https://code.bdaa.pro

实践与实验

- 推荐大家使用实验室平台练习编程
  - CODIA：https://code.bdaa.pro



联系方式

- 授课教师：黄振亚，陈恩红
- 课程主页：
  http://staff.ustc.edu.cn/~huangzhy/Course/DS2022.html
- QQ群：766486017
- 联系方式
  - 教师：
    - 黄振亚，huangzhy@ustc.edu.cn
  - 助教：
    - 覃龙虎，刘嘉聿
    - ds_intro2022@163.com