



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

黄振亚，陈恩红，刘淇

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

课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html>



课程目标

2

- 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，

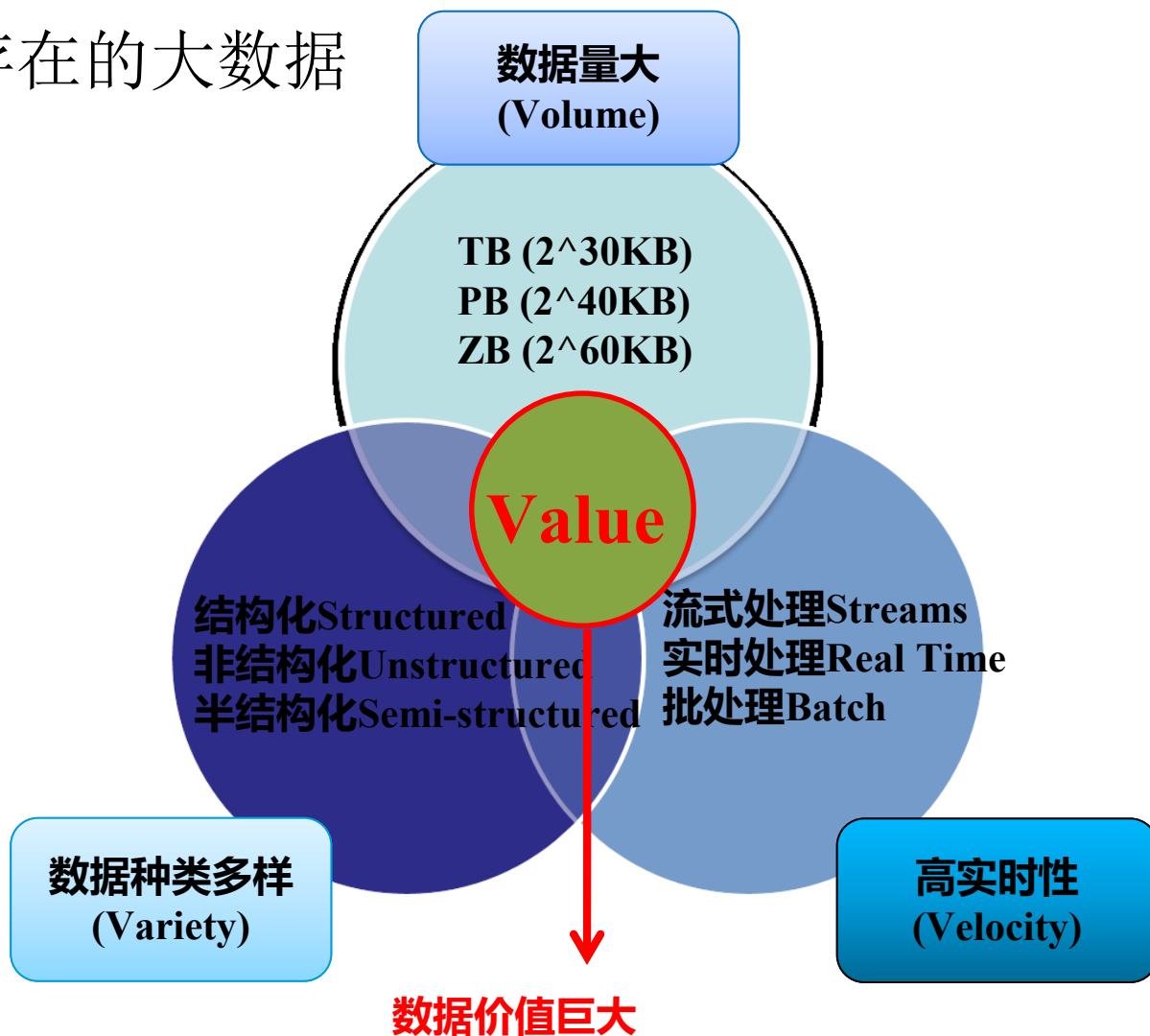
希望这门课程带给你们的是终身受用的数据思维和创新能力。



回顾：数据科学基础

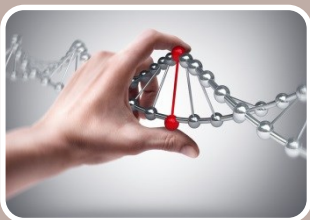
3

□ 客观存在的大数据





4





数据科学基础

5

□ 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。

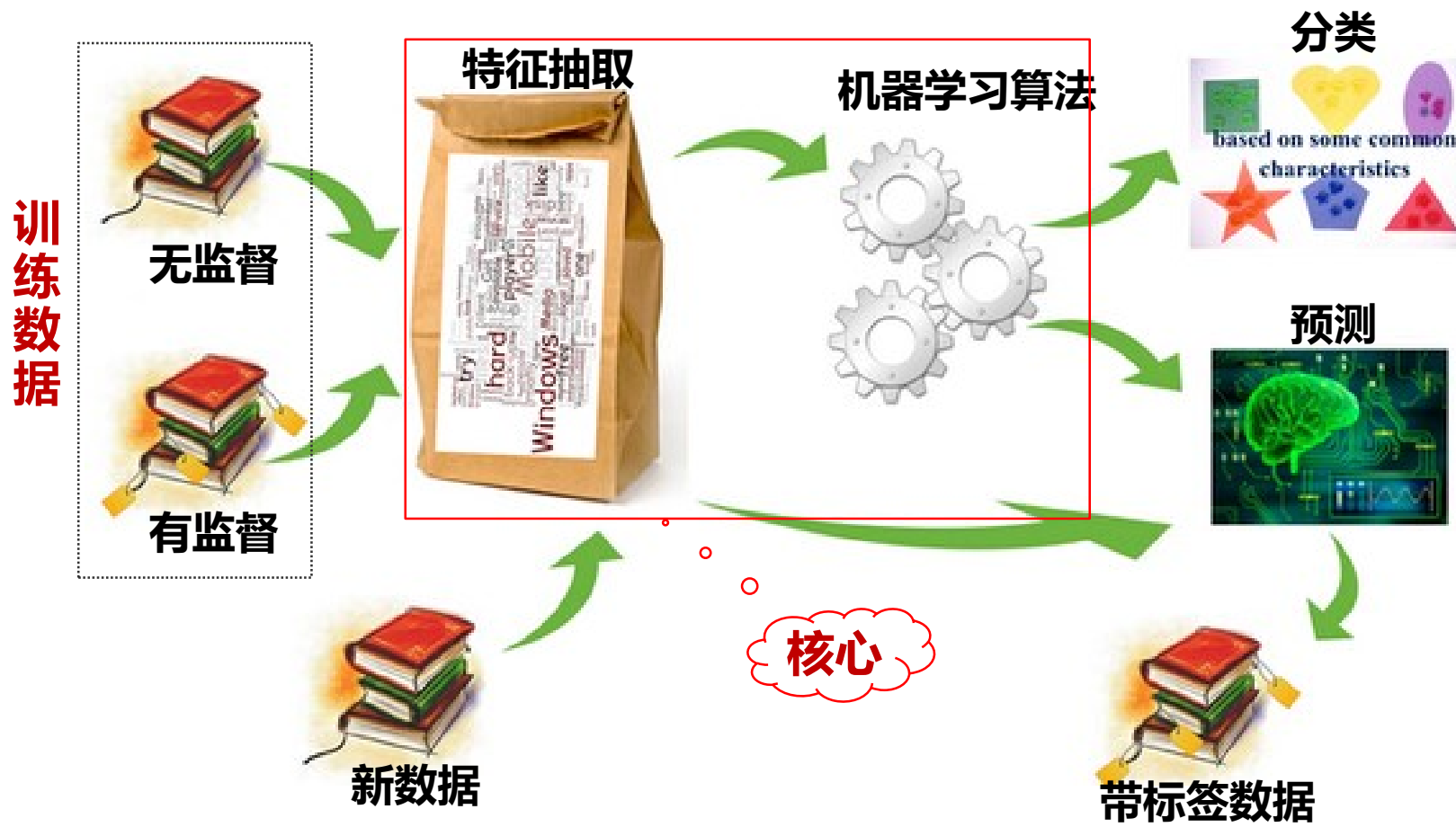


数据科学基础

6

大数据与人工智能

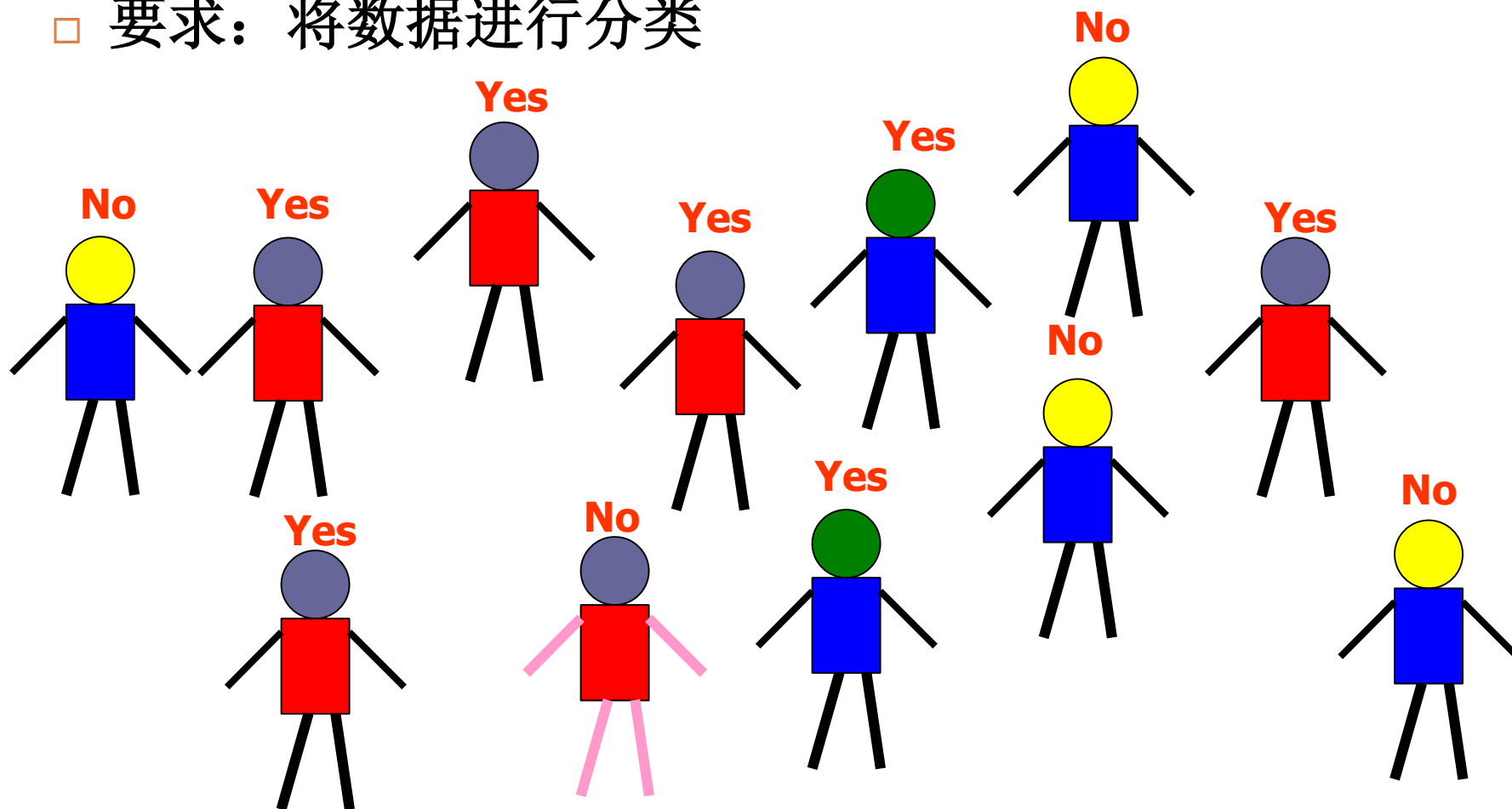
现阶段，人工智能的核心是对大数据进行的**特征抽取**与**机器学习算法**





数据+分类学习的方法

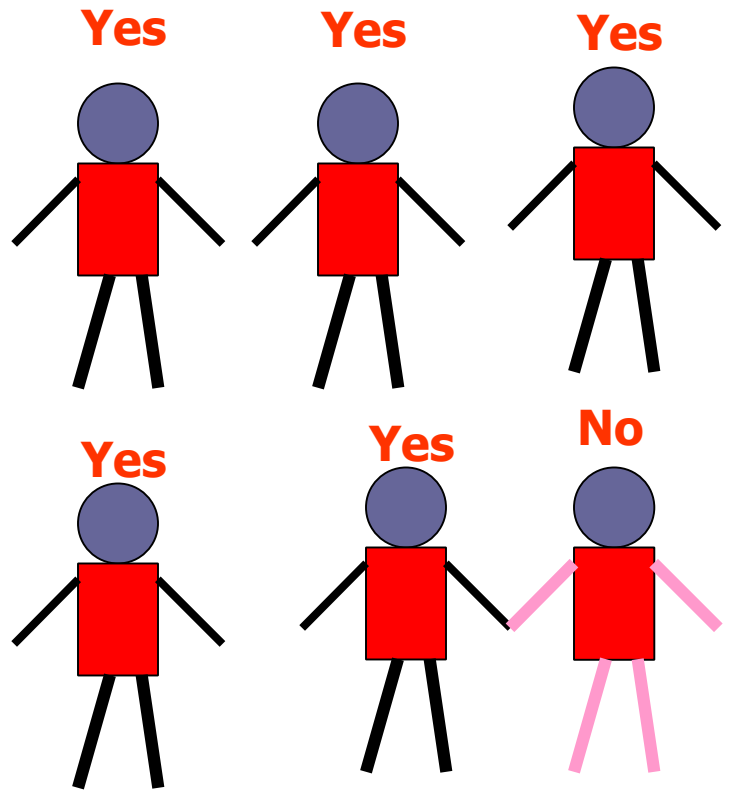
□ 要求：将数据进行分类



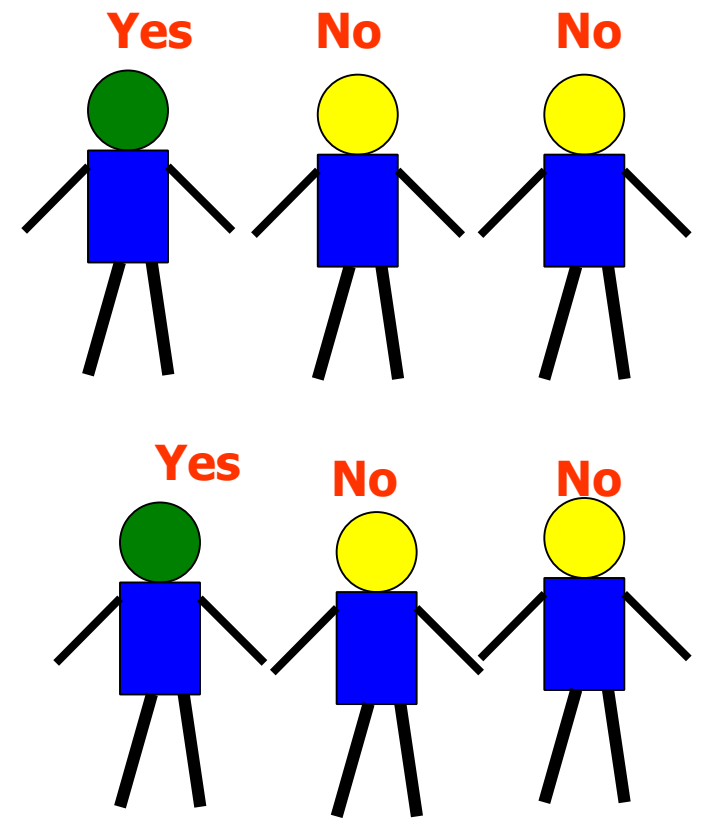


数据+分类学习的方法

躯干：红色



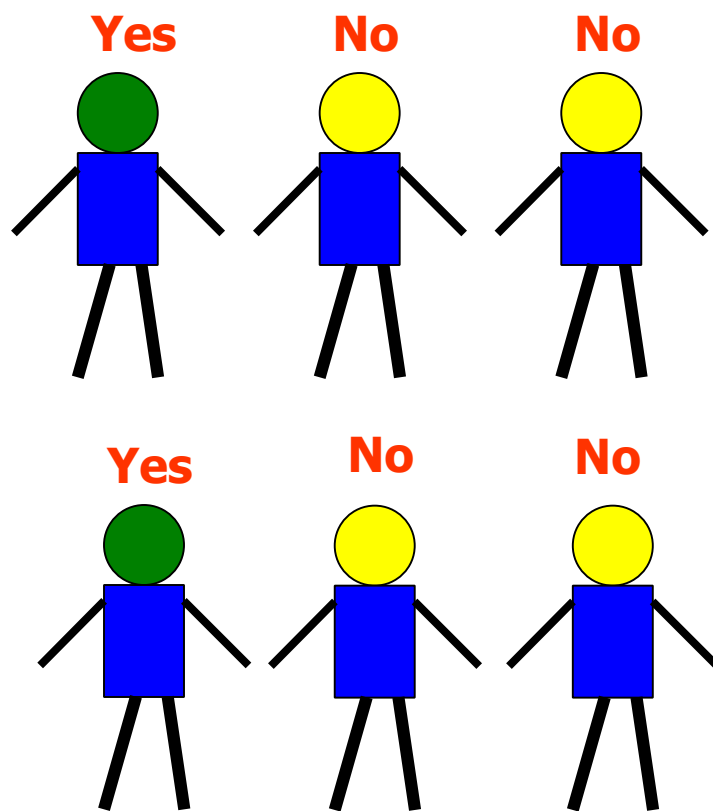
躯干：蓝色





数据+分类学习的方法

躯干：蓝色



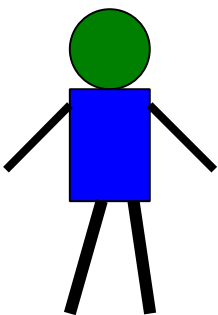


数据+分类学习的方法

躯干：蓝色

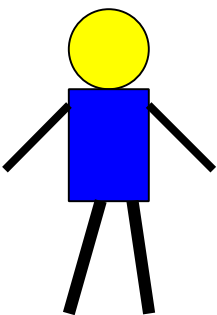
头：绿色

Yes

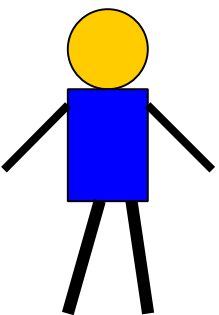


头：黄色

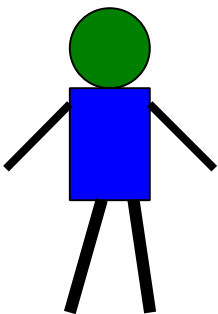
No



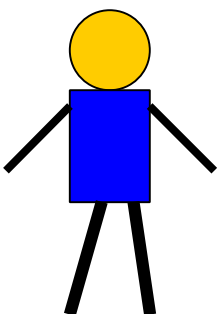
No



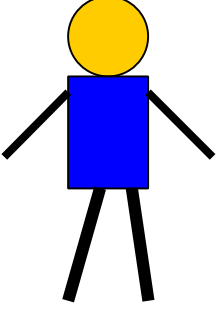
Yes



No



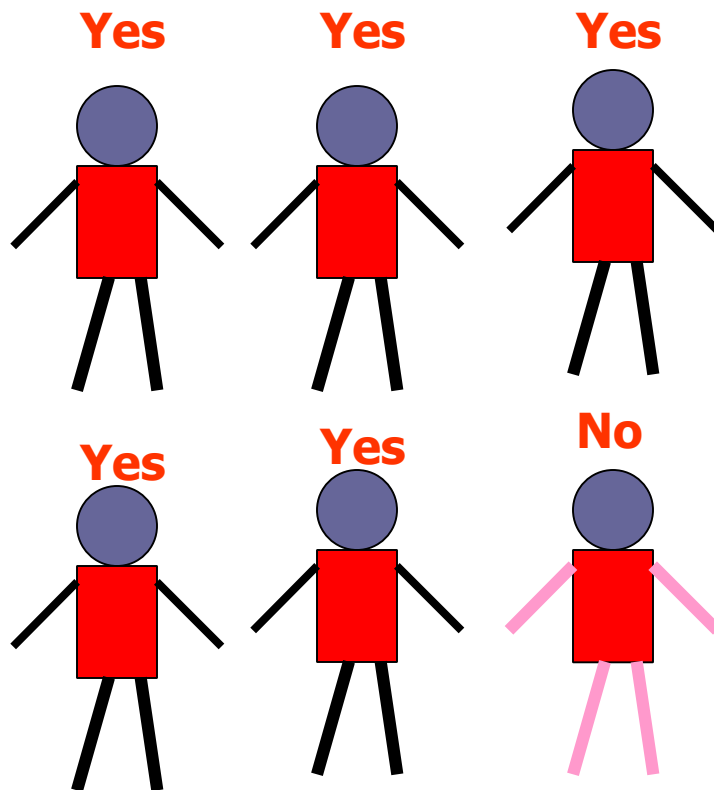
No





数据+分类学习的方法

躯干：红色



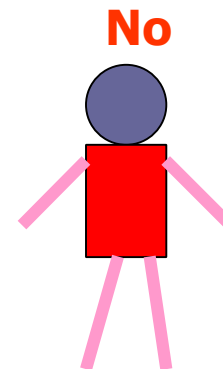
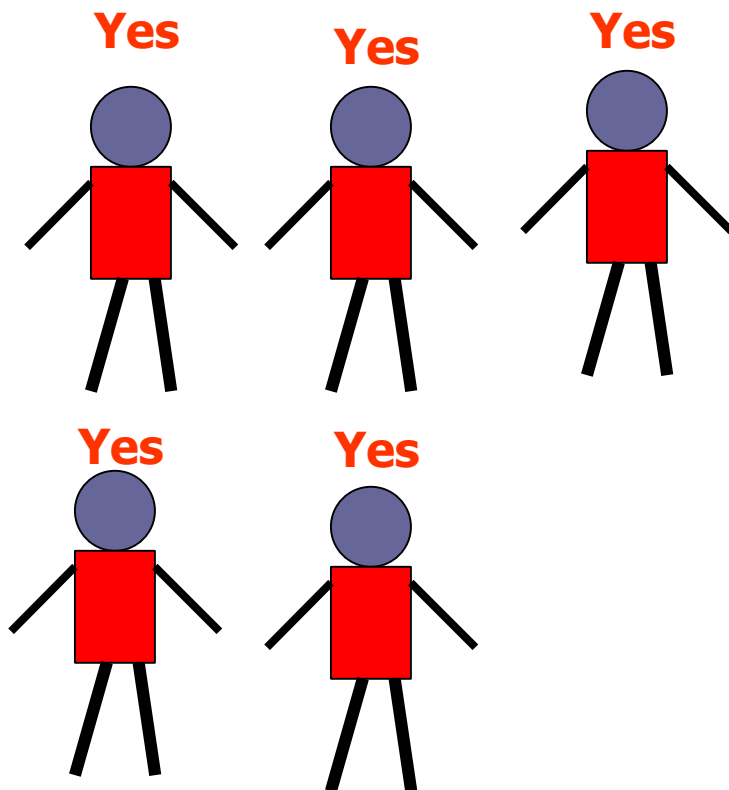


数据+分类学习的方法

躯干：红色

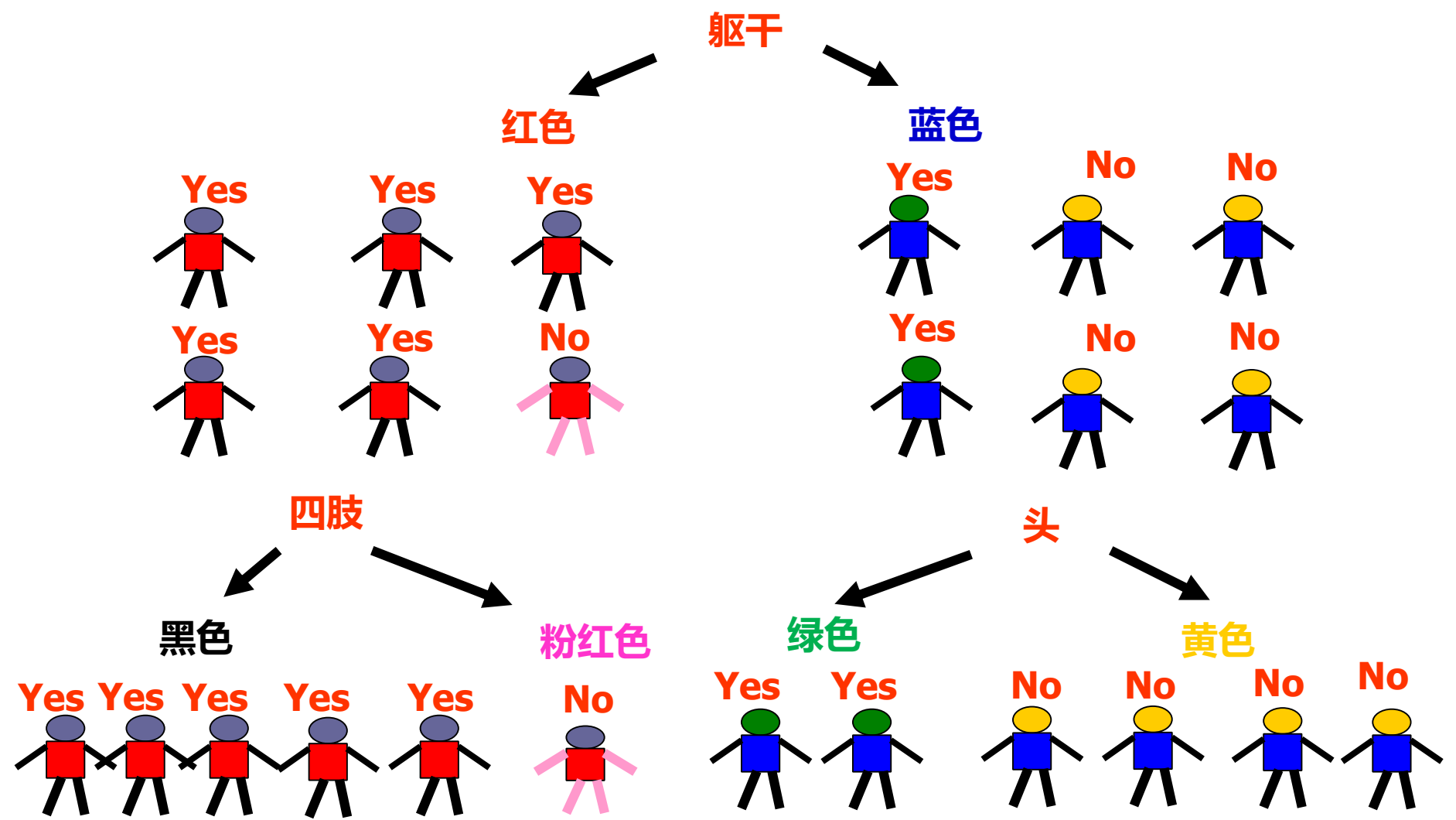
四肢：黑色

四肢：粉红色





数据+分类学习的方法



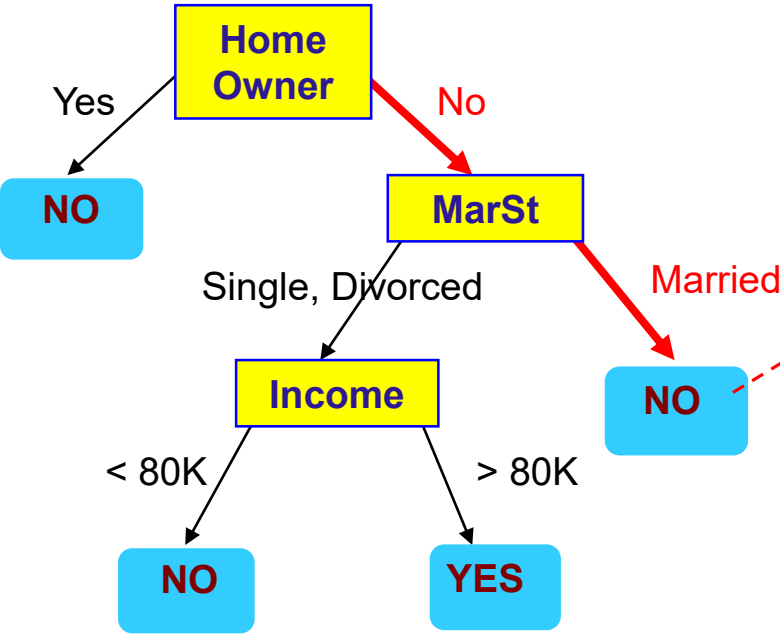


数据+分类学习的方法

决策树（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	?

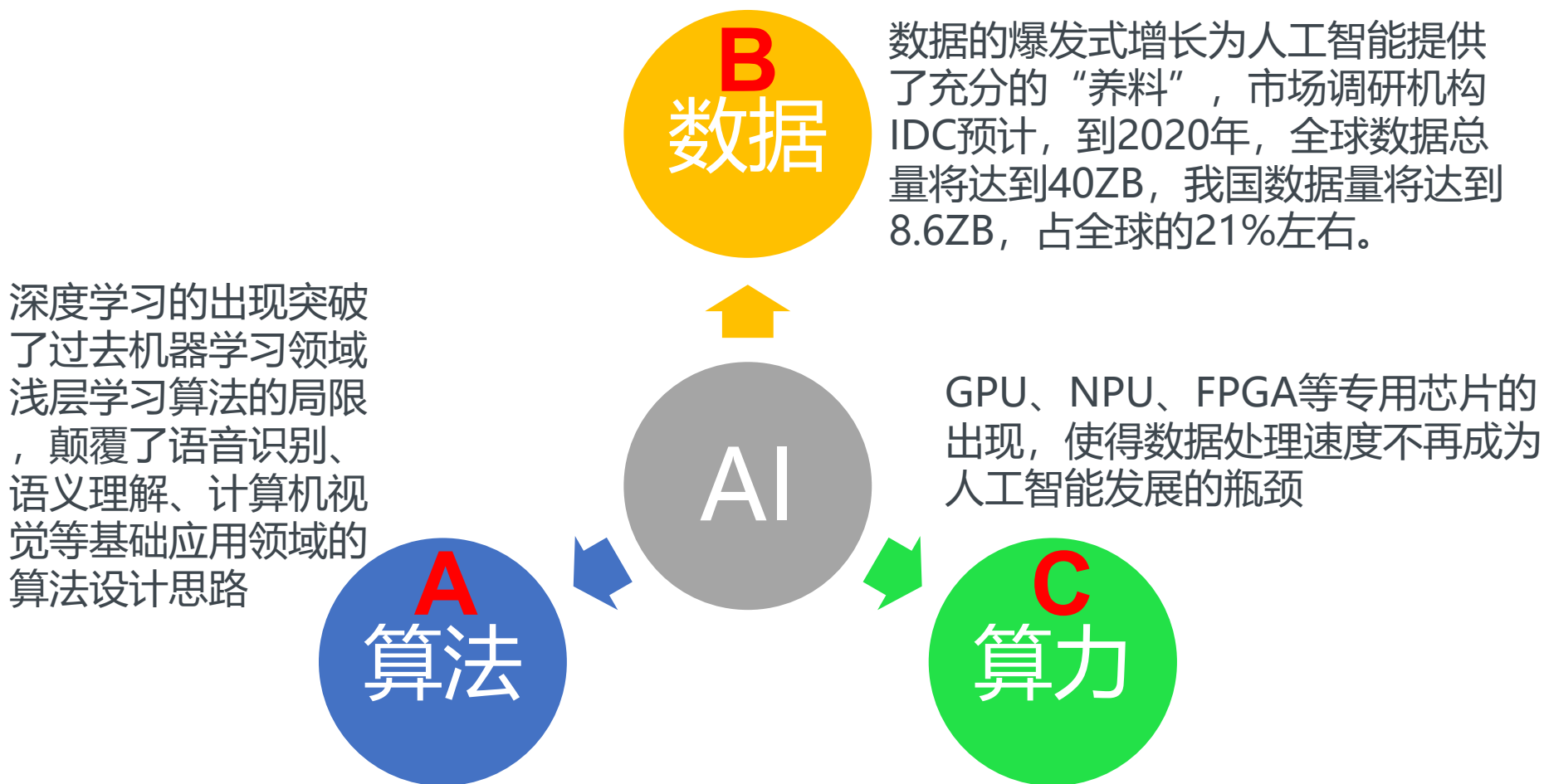


Assign Defaulted to
“No”



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





数据科学基础

16

- 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

HUAWEI HISENCON

- 移动端: 麒麟980芯片

NXP

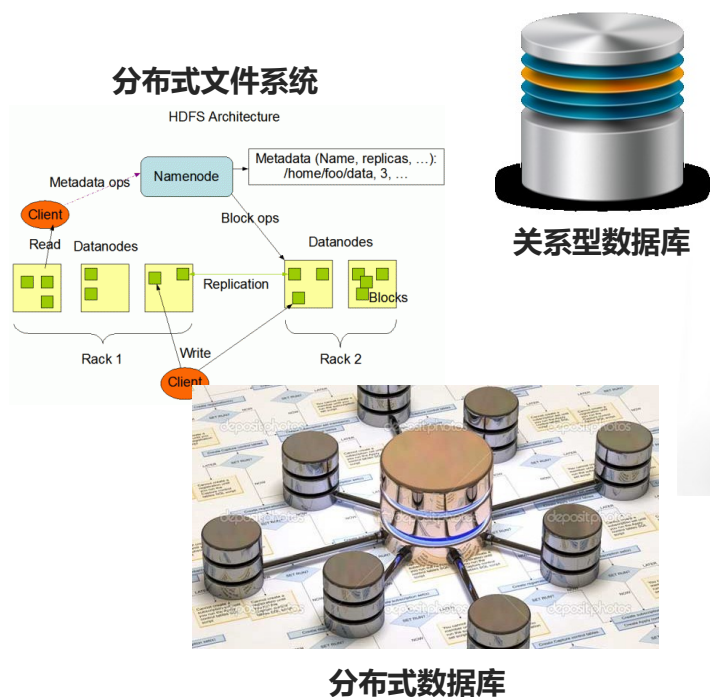
- 跨界处理器: i.MX RT1060



数据科学基础

17

- 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



11/10/2021



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
- 从产值来看，在电商等领域有较好发展，**一些领域（如农业）没有充分发展**



改造方式



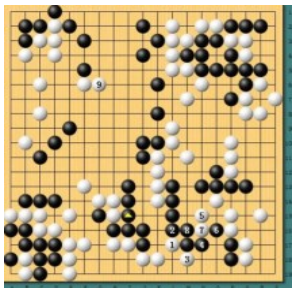


数据科学基础

□ 大数据的未来—数据与知识融合，让人工智能更“聪明”

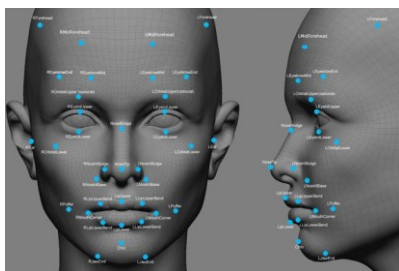
计算智能

- 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

21

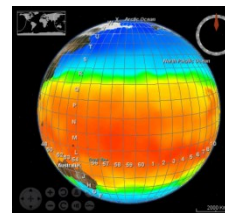


存储（如硬盘、数据库）

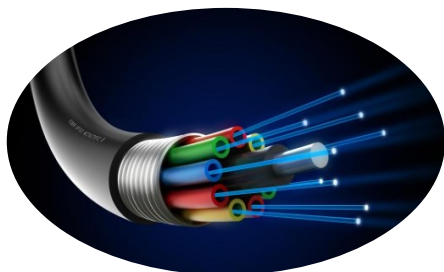


$$\min_X f(X) + \lambda \cdot \text{rank}(X)$$

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

.....

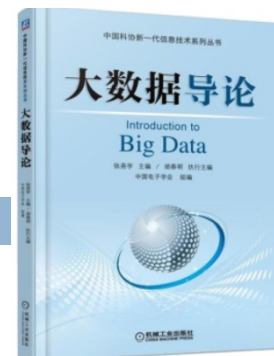


计算（平台与架构等）



数据科学基础

22



□ 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）

23

- 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
 - 数据竞赛课程汇报，第17次课
 - 课程回顾，第18次课



教学内容

24

- 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



联系方式

25

- 授课教师：黄振亚，陈恩红，刘淇
- 课程主页：
<http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html>
- QQ群：697196774
- 联系方式
- 教师：黄振亚，huangzhy@ustc.edu.cn
- 助教：刘嘉聿，ds_intro2021@163.com





课程要求与考核方式

26

- 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（2021年11月9日）
 - 实践报告 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（2021年12月14日）
- 考核方式
 - 课堂出勤（30%）+调研报告（30%）+实践报告（40%）



任务一：文献调研（2021年11月9日）

27

- 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



文献调研（2021年11月9日）

28

- 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》
 - 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



文献调研（2021年11月9日）

29

□ 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)



文献调研（2021年11月9日）

30

□ 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
- Question Difficulty Prediction for READING Problems in Standard Tests

□ 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

31

□ 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

32

- 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

33

- 参与实践与实验，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



实践与实验

34

□ 推荐大家使用实验室平台练习编程

□ CODIA: <https://code.bdaa.pro>

{CODIA}

编程到来，智

编程已成为智能时代的必备
纸上谈兵。在 CODIA 在线
学习、持续进步。

现在加入

登录

注册

请选择您使用CODIA的目的？

@username



大学课程同步练习



准备考研上机

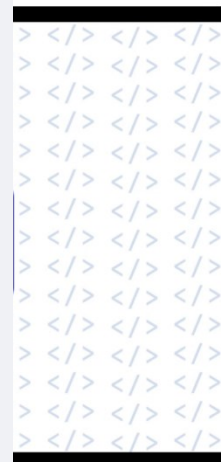


日常编程技能学习



准备工作机试

跳过



11/10/2021



实践与实验

35

□ 推荐大家使用实验室平台练习编程

□ CODIA: <https://code.bdaa.pro>

The screenshot displays the CODIA platform interface. At the top, there's a header bar with the CODIA logo and the text '资源库'. Below this, a navigation sidebar on the left contains icons for '首页' (Home), '数据' (Data), '题库' (Question Bank), '群组' (Groups), '社区' (Community), and 'IDE'. The main content area features several widgets: a top row of statistics (注册天数 213, 通过率 0.00%, 加入群组 4, 通过题数 0/240, 未读消息 0), a '最近轨迹' (Recent Trajectory) section showing a list of practice problems (e.g., '2021中科大计院复试'), a '动态' (Dynamic) feed showing user activity (e.g., 'JustWu 完成了 爬楼梯'), and a '今日任务' (Today's Tasks) section for '中国科学技术大学' (USTC) with multiple task cards for different computer science exams (e.g., '20-中科大计算机学院机试').



实践与实验

36

□ 推荐大家使用实验室平台练习编程

□ CODIA: <https://code.bdaa.pro>

