



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

课程交流与总结

黄振亚，陈恩红

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。



教学内容

3

- 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



第二章 数据分析基础

4

□ 数据采集

Data Collection

- 信息检索
- 网络爬虫

□ 数据存储

Data Storage

□ 数据预处理

Data Preprocessing

- 数据清洗
- 数据集成
- 数据变换
- 数据规约

□ 特征工程

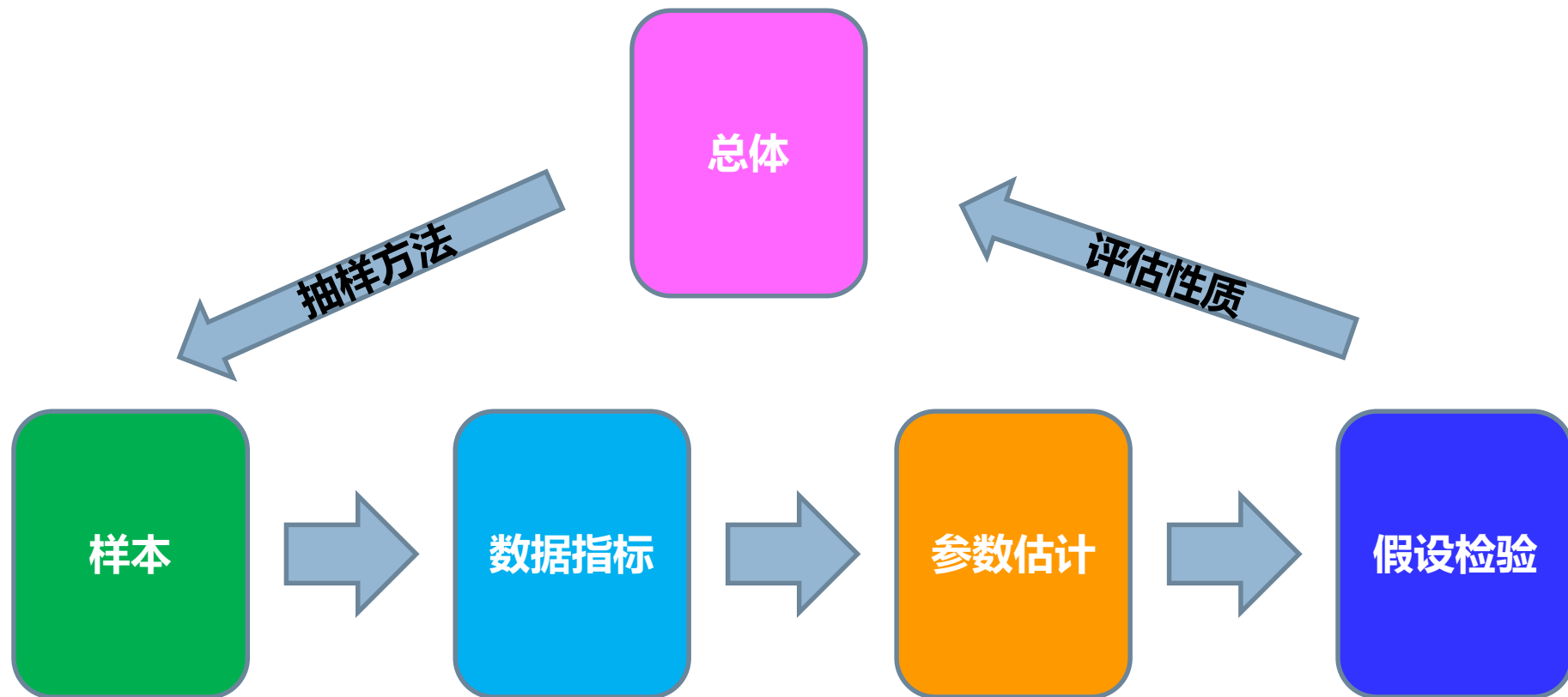
Feature Engineering

- 特征设计
- 特征选择



第三章 数据统计基础

5



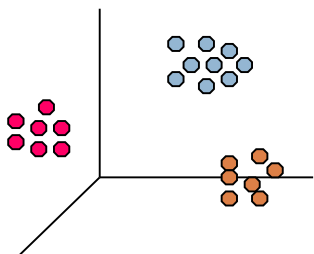


第四章 数据挖掘基础

6

数据挖掘——四个任务有哪些常用方法？

聚类



关联分析



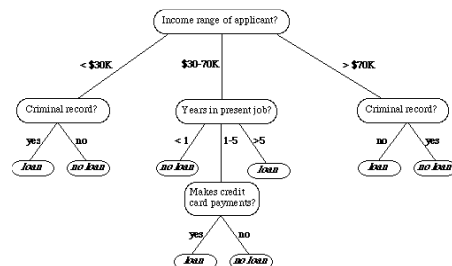
数据

	T		H		P	
	L	H	L	H	L	H
J	-6.0	8.8	60	100	986	1044
F	-2.8	10.9	48	100	973	1025
M	-5.6	17.7	34	100	976	1037
A	-1.2	22.2	27	100	996	1036
M	-0.8	27.8	25	100	1003	1034
J	5.2	29.1	26	100	998	1030
J	9.8	30.6	23	99	997	1027
A	5.6	26.1	31	100	992	1029
S	5.2	24.8	35	100	998	1028
O	-0.4	21.3	42	100	990	1031
N	-7.6	17.3	55	100	963	1023
D	-10.4	9.2	53	100	987	1039

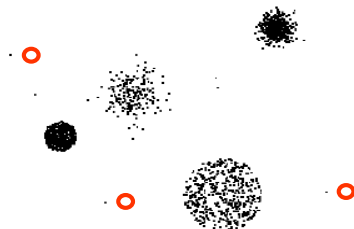
table 17a

2010 monthly weather variation, Cambridge (UK)

分类与预测



异常检测





专题一 数据分析编程

7

- 数据分析编程(Python, Numpy, Pandas)
- 机器学习工具包(Scikit-Learn)
- 深度学习工具包(TensorFlow, PyTorch)

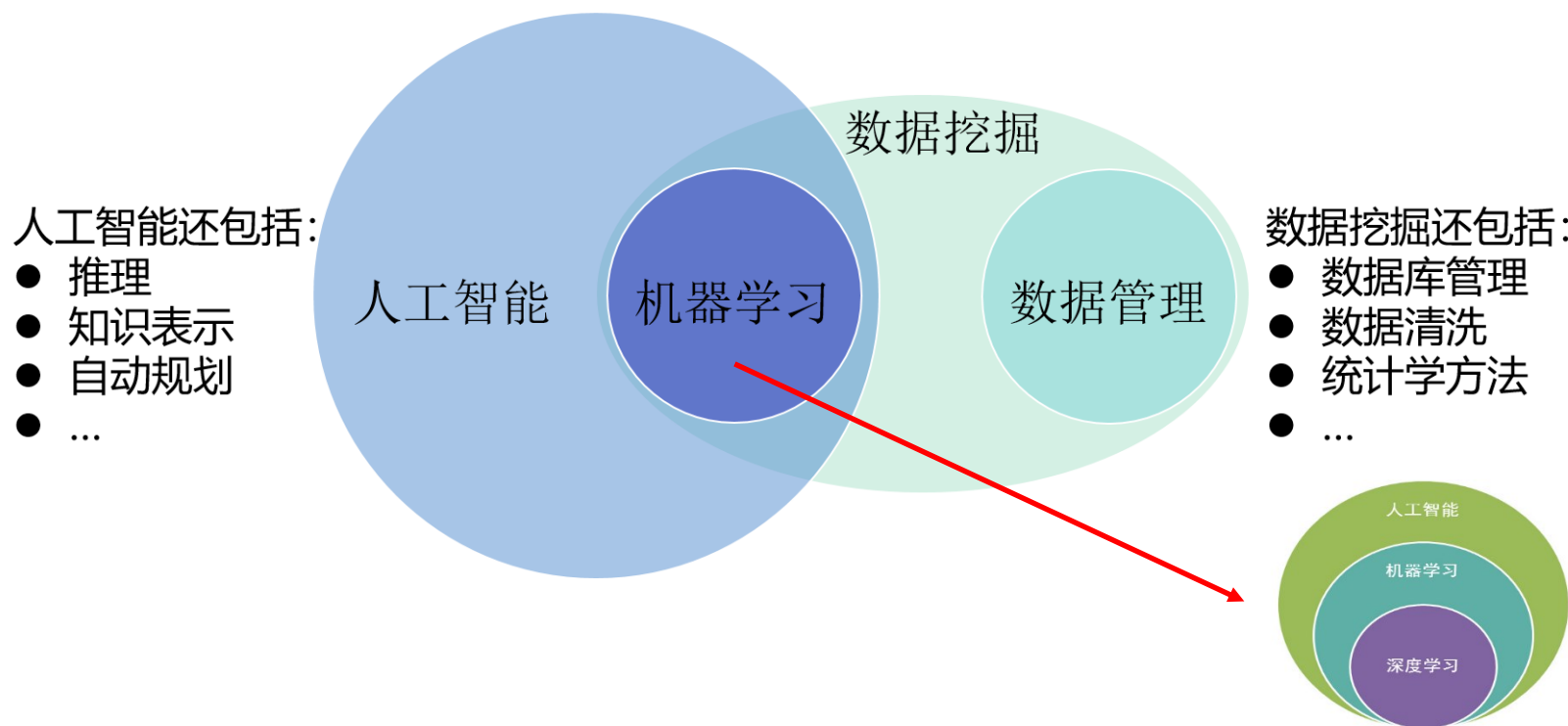




专题二 人工智能与深度学习

8

- 人工智能应用: 计算机视觉, 自然语言处理, 推荐系统
- 深度学习基础: 多层MLP, 前向传播, 反向传播
- 前沿方向: 图神经网络, 强化学习, 联邦学习, etc

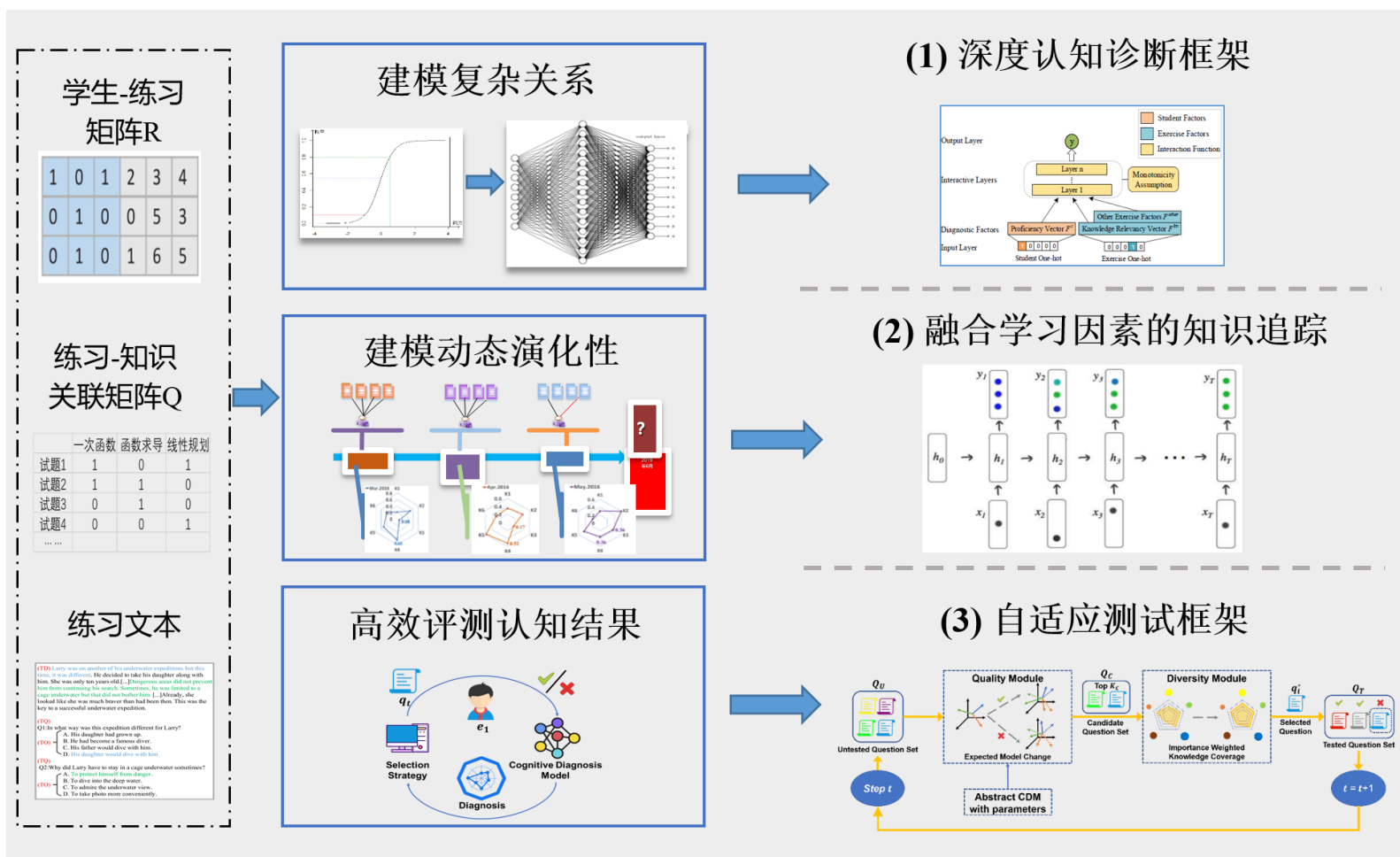




专题三 大数据分析应用

9

数据驱动的认知诊断研究与应用探索

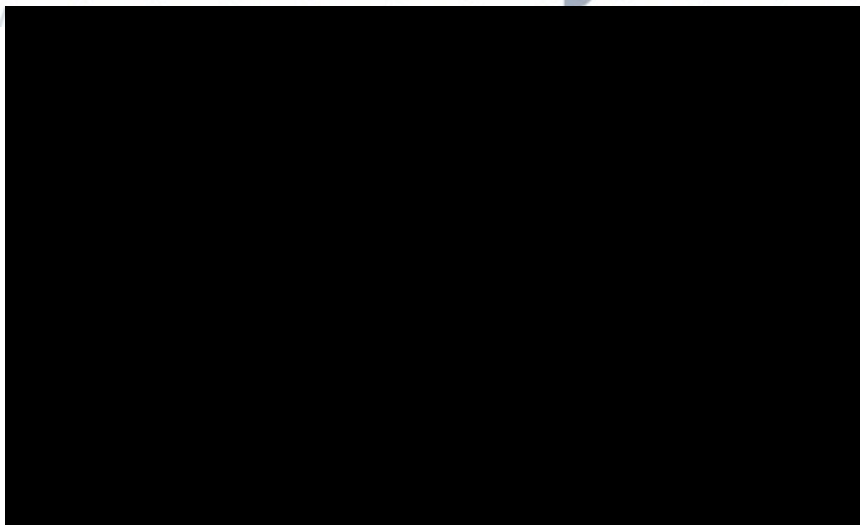




大数据分析与应用实验室

10

- **CoDia (Coder Diagnosis, 课达编程)**
 - <http://code.bdaa.pro>
- **LUNA (智慧教育知识图谱)**
 - <http://luna.bdaa.pro>



2500万+
道试题

9大
学科

33万+
套试卷

小学~成人
全学段覆盖

海量教学资源、智能处理

- 智能化、个性化的资源（试题、试卷）的录入、查询和管理服务
- 内置标准化、体系化的教学资源
- 使用高效准确的试题相似度评估算法，提升题库的资源质量

试题检索 → 试卷检索 →



《数据科学导论》

11

数据科学导论 Introduction to Data Science

谢谢大家