大数据分析

课程介绍与考核安排

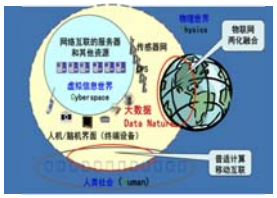
程学旗
靳小龙
刘盛华

授课团队

- 教师
  - 程学旗
    - 中科院计算所研究员，博导，副所长
    - 邮　箱：cxq@ict.ac.cn
    - 办公室：计算所 1138
  - 靳小龙
    - 中科院计算所研究员，博导
    - 邮　箱：jinxiaolong@ict.ac.cn
    - 办公室：计算所 932
  - 刘盛华
    - 中科院计算所副研究员，硕导
    - 邮　箱：liushenghua@ict.ac.cn
    - 办公室：计算所 937A
- 教师助教
  - 赵凯琳 zhaokailin17z@ict.ac.cn, 13287621599
- 学生助教
  - 李明龙 liminglong18@mails.ucas.ac.cn
  - 韩　宇 hanyu19@mails.ucas.ac.cn



大数据分析是大数据价值落地的关键环节



信息技术革命与人机物三元世界的交融 → 大数据
（数量巨大、种类繁多、增长极快、价值稀疏的复杂数据）

大数据 —分析处理能力→ 大价值

- 美国政府大数据计划和Google 等大公司目前最重视的都是数据价值，着力于大数据分析技术和系统的研究与应用；

我国当前痛点：大数据→小价值

据IDC统计数据显示，中国目前拥有的数据量占全球的14%，但数据利用率不到0. 4%

突破大数据分析技术瓶颈，推动大数据价值落地成为大数据领域的当务之急

大数据分析课程应有的定位

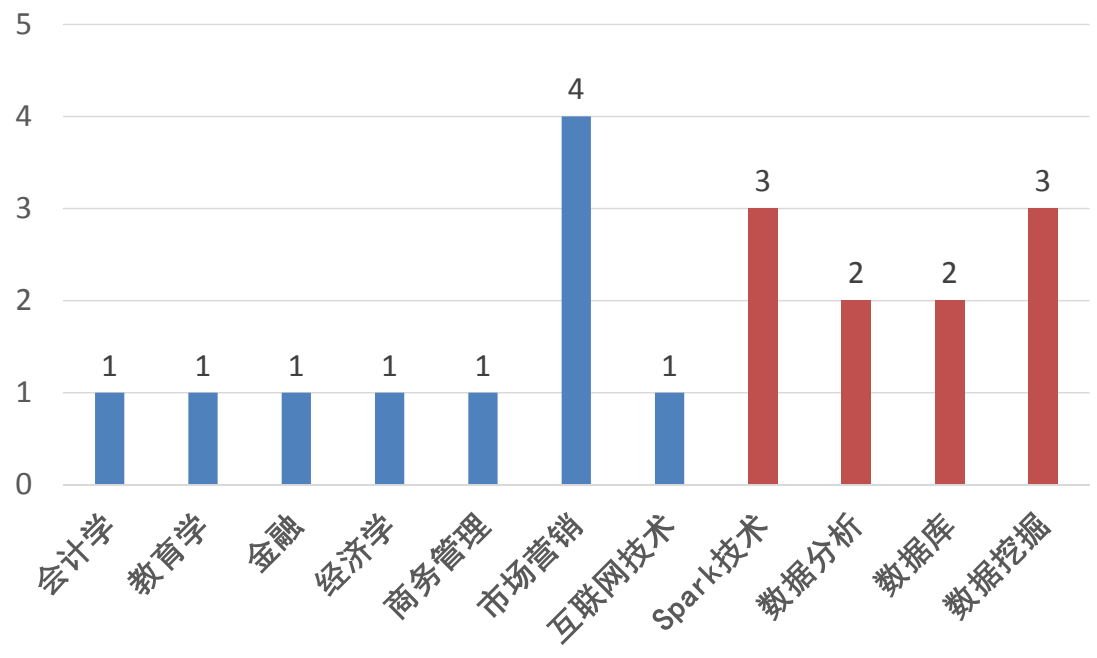
- 专业必修课/普及课，培养合格的大数据分析工程师和大数据分析科学家
  - 构建知识体系，阐明基本原理；
  - 引导初级实践，了解相关应用；
  - 为学生在大数据领域“深耕细作”奠定基础，指明方向；
- 搭起通向“大数据分析”系统之门的桥梁和纽带

大数据分析书籍概览

书名	著作者	出版社	时间
大数据分析	王星等编著	清华大学出版社	2013年09月
大数据分析	张重生编著	机械工业出版社	2016年12月
[illegible]	[illegible]	[illegible]	2017年11月
大数据分析	[illegible]	人民邮电出版社	2014年08月
大数据分析	[美]Thomas H.Davenport编	机械工业出版社	2015年03月
[illegible]	[illegible]	[illegible]	[illegible]
大数据分析的道与术	华然编著	电子工业出版社	2016年04月
大数据分析方法	[美] Michele Chambers, Thomas W.Dinsmore	机械工业出版社	2016年08月
[illegible]	[illegible]	[illegible]	[illegible]
大数据分析计算机基础	张延松 王成章 徐天成	中国人民大学出版社	2016年07月
大数据分析原理与实践	王宏志	机械工业出版社	2017年07月
[illegible]	[illegible]	[illegible]	2016年07月
大数据预测分析	[美]Lawrence S. Maisel, Gary Cokins等著	人民邮电出版社	2014年11月
数据科学与大数据分析	[美]EMC教育服务团队	人民邮电出版社	2016年07月
Spark大数据分析	经管之家	电子工业出版社	2017年07月
Spark大数据分析	[美]Mohammed Guller	机械工业出版社	2017年05月
Spark大数据分析实战	高彦杰 倪亚宇	机械工业出版社	2016年01月
Python金融大数据分析	[德]Yves Hilpsch著	人民邮电出版社	2015年12月
群体智能与大数据分析技术	陶乾等	暨南大学出版社	2018年04月
社交媒体大数据分析	[美] Lutz Finger, Soumitra Dutta	人民邮电出版社	2016年10月

- 从当当与Amazon上获取，总计48本，8本编，40本著
- 39本通用型，9本面向特定领域：金融、社交媒体与群体智能
- 12本面向特定系统/语言：Spark系统，Python语言、R语言
- 13本主要读者对象为本科生/研究生，适合做教材

著作者专业背景



大数据分析书籍的副标题

书名	副标题
大数据分析	方法与应用
大数据分析	决胜互联网金融时代
大数据预测分析	决策优化与绩效提升
大数据分析	数据驱动的企业绩效优化、过程管理
大数据分析的道与术	
大数据探索性分析	
数据科学与大数据分析	数据的发现、分析、可视化与表示
大数据分析计算机基础	
大数据分析方法	用分析驱动商业价值
大数据分析	数据挖掘必备算法示例详解
大数据分析方法	
大数据分析原理与实践	
大数据分析	R语言实现
大数据分析	创造价值 做聪明的市场决策
Python金融大数据分析	
Spark大数据分析实战	
社交媒体大数据分析	理解并影响消费者行为
Spark大数据分析	核心概念、技术及实践
Spark大数据分析	技术与实战
……	

16本利用副标题对内容作了进一步的阐释和限定，说明对大数据分析的内涵、内容设置、定位等有多种不同的理解

大数据分析教材

书名	著作者	出版社	内容简介	读者对象
大数据分析原理与实践	王宏志	机械工业出版社（2017）	介绍了大数据预处理，数据仓库，以及分类、聚类、关联分析、结构分析和文本分析模型；大数据的并行、流式与图分析平台等	计算机科学专业的本科生或者研究生，也可以作为从事大数据相关工作的工程技术人员的参考用书
大数据分析：方法与应用	王星等编著	清华大学出版社（2013）	主要介绍数据挖掘、统计学习和模式识别中与大数据分析相关的理论、方法及工具	统计学、管理学、计算机科学等专业的高年级本科生或研究生
大数据探索性分析	吴翌琳，房祥忠	中国人民大学出版社（2016）	介绍大数据的预处理、采样、大数据探索性分析案例、数据可视化等	有统计学基础的硕士研究生，统计专业高年级本科生

本门课程简介

大数据分析的思维方法以及基础知识框架，如：大数据分析方法的基本概念、基本知识等

基本的数据分析方法，如：数据采样方法，分类，聚类，排序等

常用的大数据分析应用，如：文本大数据分析，知识计算，网络数据挖掘等

大数据分析架构，如：数据并行分析架构、模型并行分析架构、流式分析框架等。

了解大数据分析的经典应用案例，包括互联网搜索与广告、社交网络分析、互联网舆情分析等。

课程内容安排

大数据分析应用案例与生态

数据驱动的自然语言处理

文本大数据分析

知识获取与计算

大图数据分析

社会媒体大数据分析

跨媒体大数据分析

大数据分析技术与系统

大数据机器学习

大数据统计分析

大数据与大数据分析简介

课程时间与内容安排

40个课时；13次课；2个学分

	时间	内容
1	09. 15	大数据与大数据分析概论 课程设置
2	09. 22	大数据分析技术与系统
3	09. 29	大数据统计分析
	国庆假期	
4	10. 13	大数据机器学习
5	10. 20	数据驱动的自然语言处理
6	10. 27	文本大数据分析
7	11. 03	知识计算
8	11. 10	大图挖掘与分析

课程时间与内容安排

	时间	内容
9	11. 17	社会媒体分析
10	11. 24	跨媒体分析
11	12. 01	大数据分析应用案例与生态
12	12. 08	大作业：分组报告
13	12. 15	闭卷考试

注意：每项内容的课时与时间安排可能会有微调

教材

高等教育出版社
程学旗 主编

- 没有强制指定教材
- 建议的课本和材料
  - 《大数据分析》高等教育出版社
    - 程学旗 主编
  - 《Mining of Massive Datasets》
    - Jure Leskovec, Anand Rajaraman, Jeff Ullman
  - 《Graph Mining: Laws, Tools and Case Studies》
    - Deepayan Chakrabarti and Christos Faloutsos
  - 《Individual and Collective Graph Mining: Principles, Algorithms, and Applications》
    - Danai Koutra, Christos Faloutsos. Synthesis Lectures on Data Mining and Knowledge Discovery, October 2017, 206 pages. Morgan & Claypool publishers.
  - 《Scalable algorithms for data and network analysis》
    - Teng, Shang-Hua， Foundations and Trends® in Theoretical Computer Science 12
  - F. R. K. Chung, Spectral graph theory, CBMS Regional Conference Series in Mathematics, Publication Year 1997: Volume 92

课程基础知识需求

- 基本统计知识
- 概率论基础
- 线性代数、矩阵论
- 算法分析
- 编程语言 Python，… …

建议提前回顾/自学一下上述基本知识

课程考核安排与成绩占比设置

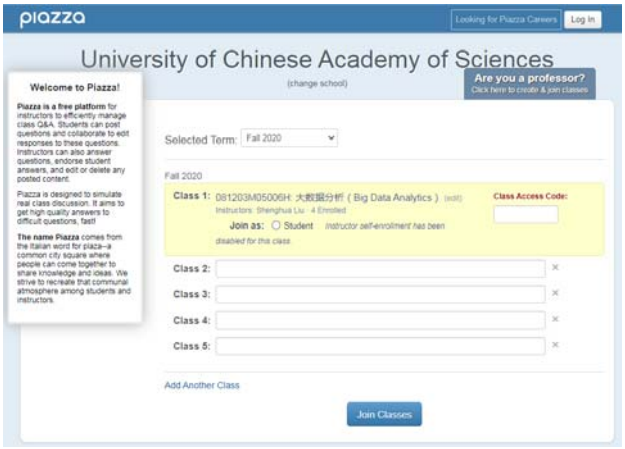
- 课后小作业： 2次（2x5%=10%）
  - 独立完成，不许互相抄袭（编程类作业简单的变量重命名也被认为是抄袭）；
- 调研报告 （15%）
  - 针对大数据分析的某个方向/领域，形成一份调研报告；
  - 有自己独立的观点与见解；
  - 逻辑清晰、结构合理、格式规范，有参考文献并逐个引用；
  - 严禁拷贝、抄袭、剽窃；
- 大作业 （25%）
  - 申请书 2%；
  - Word版终期报告 13%；
  - PPT汇报及答辩 10%；
- 闭卷考试 （50%）
- 加分项 （10%）
  - 积极参与Piazza/微信群的课程讨论与交流；

课后作业发布

- 发布在SEP上
  - 预计发布时间：
    - 大数据机器学习/大图分析；
    - 知识计算；
  - 作业提交形式
    - 电子版提交；
    - SEP系统；
  - Piazza上讨论作业

课程与作业讨论

- 注册Piazza账号
- 课程网址
  - http://piazza.com/university_of_chinese_academy_of_sciences/fall2020/081203m05006h
  - 班级访问码: bigdata2020
- 课程问答和讨论
  - 仅关于课程内容、作业、大作业等
- 学生积极参与



大作业

- 按小组进行，每个小组3-5人，自由组合；
- 选题：
  - CCF大数据大赛题目
    - https://www.datafountain.cn/competitions；
    - 将大数据分析方法应用到自己研究领域；
- 需提交的材料：
  - 申请书：包括但不限于题目、选题内容、研究思路，以及小组成员名单等；
  - Word版终期报告：包括但不限于题目、研究内容、技术路线、实验结果及成员名单等；
  - 答辩PPT；
- 重要日期：
  - 申请书提交日期：10月12日，23:59；
  - 终期报告提交日期：12月15日，23:59；
  - PPT答辩日期：第11-12周；

关于作业的问题

- 为了方便交流，作业的问题首先反馈给TA；
- 也可以到Piazza上讨论和提问；
- 针对共性问题，TA与教课老师反馈和讨论；

交流渠道

- 通知发布在Piazza和每次课的最后；
- Email；
- 利用Piazza、微信群与其他学生和TA讨论；
- 集中答疑时间：待定；