《数据科学导论》 课程大纲

# 课程概述

主讲教师

××××

电话/微信

××××

电子邮件

××××

办公地点

××××

办公时间

××××

* 本课程主要探讨大数据时代及其背后的科学问题——数据科学的理念、理论、方法、技术、工具、应用及最佳实践，为学生更好地掌握专业知识以及提升自己的核心竞争力打下坚实基础。
* 本课课程对学生专业背景和知识基础不做严格限制，鼓励跨学科选修；
* 本课程主讲老师——××××

# 教材

* 朝乐门. 数据科学理论与实践[M].北京：清华大学出版社，2017.

# 教学形式

* 讲解：培养数据科学家的理论水平（数据科学的基础理论）
* 编程：培养数据科学家的实战能力（数据科学的代表性实践及R/Python编程）
* 讨论：培养数据科学家的精神；

# 考核方式

* 平时：10分
* 作业（含编程）：50分
* 课程报告：40分

# 课程安排

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 周 | 主要内容 | 教学目的 | 教学重点 | 教学难点 | 教学方法 |
|  | 第1章 基础理论  1.1.学习本章知识的方法、建议及注意事项  1.2.术语定义  1.3.研究目的  1.4.数据科学的发展简史 | 1.掌握：数据科学中的基础理论，尤其是核心术语、研究目的、理论体系与基本原则；  2.理解：数据科学家的主要职责与能力要求；  3.了解：数据科学的发展简史 | 1.大数据挑战的本质  2.数据科学的研究目的  3.数据科学的理论体系  4.数据科学的基本原则 | 1.大数据挑战的本质  2.数据科学的基本原则 | 1.理论讲解  2.随堂讨论  3.课后调研 |
|  | 1.5. 数据科学的理论体系  1.6. 数据科学的基本原则  1.7.数据科学家  1.8.继续学习本章知识的建议及注意事项 |
|  | 第2章 基础理论  2.1.学习本章知识的方法、建议及注意事项  2.2.数据科学的学科地位  2.3.统计学 | 1.掌握：数据科学的学科地位；  2.理解：机器学习、统计学、可视化理论对数据科学的主要影响；  3.了解：数据科学的理论基础——机器学习、统计学、可视化理论——的知识体系及代表性方法。 | 1.数据科学的学科地位  2.数据科学视角下的机器学习知识体系  3.数据科学视角下的统计学知识体系  4.数据科学视角下可视化理论知识体系 | 1.机器学习、统计学、数据可视化之间的区别及它们在数据科学中的不可替代作用；  2.统计学、机器学习和数据可视化之间的融合。 | 1.理论讲解  2.随堂讨论  3.课后调研 |
|  | 2.4.机器学习  2.5.数据可视化  2.6.继续学习本章知识的建议及注意事项 |
|  | 第3章 流程与技术  3.1.学习本章知识的方法、建议及注意事项  3.2.基本流程  3.3.数据加工  3.4.数据审计 | 1.掌握：数据科学的基本流程及常用方法；  2.理解：结合数据科学的基本流程和常用方法，理解数据科学与数据工程的区别；  3.了解：数据科学中的项目管理。 | 1.数据加工方法  2.探索性数据分析方法  3.数据洞见方法  4.数据故事化方法  5.数据可视化方法 | 1.数据加工方法  2.数据故事化方法  3.数据可视化方法 | 1.理论讲解  2.随堂讨论  3.课后调研 |
|  | 3.5.数据分析  3.6.数据可视化  3.7.数据故事化  3.8.项目管理  3.9继续学习本章知识的建议及注意事项 |
|  | 第4章 技术与工具  4.1学习本章知识的方法、建议及注意事项  4.2技术体系 | 1.掌握：MapReduce和Spark的核心技术与主要特征；  2.理解：NoSQL和NewSQL的核心技术与主要特征；  3.了解：数据科学的技术体系；R与Python的区别；Hadoop生态系统；数据计算技术的发展趋势；数据管理技术的发展趋势。 | 1.MapReduce及其开源实现  2.Spark主要特点及核心技术  3.NoSQL和NewSQL关键技术  4.Hadoop生态系统 | 1.大数据计算技术与传统数据计算技术的区别——以MapReduce和Spark为例  2.大数据管理技术与传统数据管理技术的区别——以NoSQL、NewSQL和关系云为例 | 1.理论讲解  2.随堂讨论  3.课后调研 |
|  | 4.3 MapReduce  4.4 Hadoop  4.5 Spark |
|  | 4.6 NoSQL与NewSQL |
|  | 4.7 R与Python  4.8 发展趋势  4.9 继续学习本章知识的建议及注意事项 |
|  | 第5章数据产品开发  5.1 学习本章知识的方法、建议及注意事项  5.2 定义  5.3 主要特征  5.4 关键活动 | 1.掌握：数据产品的类型、特征及开发方法；  2.理解：数据能力评估方法、数据治理的主要内容、数据柔术的基本思想；  3.了解：数据战略的制定要求。 | 1.数据产品的开发方法；  2.数据能力的评估方法；  3.数据治理的重要内容；  4.数据柔术的基本思想。 | 1.数据产品的设计；  2.数据柔术的基本思想；  3.DMM模型的应用。 | 1.理论讲解  2.随堂讨论  3.课后调研  4.小组展示 |
|  | 5.5 数据柔术  5.6 数据能力  5.7 数据战略  5.8 数据治理  5.9 继续学习本章知识的建议及注意事项 |
|  | **第6章 典型案例及实践**  6.1 学习本章知识的方法、建议及注意事项  6.2 2012美国总统大选  6.3 统计分析实践 | 1.掌握：基于R的数据科学实践；SparkR（或pyspark）编程；  2.理解：结合2012美国总统大选理解数据科学的典型应用；  3.了解：基于R（或Python）的统计分析、机器学习和数据可视化。 | 1.2012美国总统大选；  2.基于R（或Python）的数据科学实践；  3.SparkR（或pyspark）编程。 | SparkR 与sparklyr编程的区别与联系。  （或pyspark的编程方法） | 1.现场演示  2.个人联系  3.课后作业  4.小组展示 |
|  | 6.4 机器学习实践  6.5 数据可视化实践 |
|  | 6.6 SparkR（或pyspark）编程实践  6.7 继续学习本章知识的建议及注意事项 |
|  | **机动** | | | | |

# 附加信息

参考阅读书目见课程PPT.

|  |
| --- |
| **特别声明**  1.本文件为中国人民大学朝乐门老师的课程《数据科学导论》（对应教材为《数据科学理论与实践》（清华大学出版社，2017））的配套教学资源。  2.非版权所有人的使用范围仅限于教师的课堂教学。版权归朝乐门所有，未经许可，不得以任何形式用于商业目的。  3.不要擅自放在包括百度网盘/文库、CSDN、豆丁网等第三方平台，否则版权所有人有权追究上传者和第三方平台的法律责任。  4.本文件将以【开源模式】更新和维护，欢迎您的补充与修改。通过chaolemen@ruc.edu.cn联系总负责人朝乐门老师，总负责人将根据您的贡献动态更新作者列表及排名。  朝乐门  2017-11-18 |