STA217 数据科学导论课程大纲

1、 2021　秋季学期

2、 2022　秋季学期起

课程详述

COURSE SPECIFICATION

以下课程信息可能根据实际授课需要或在课程检讨之后产生变动。如对课程有任何疑问，请联系授课教师。

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	数据科学导论 Introduction to Data Science
2.	**授课院系 Originating Department**	统计与数据科学系 Department of Statistics and Data Science
3.	**课程编号 Course Code**	STA217
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业选修课 Major Elective Courses
6.	**授课学期 Semester**	秋季 Fall 【2021秋季学期】
7.	**授课语言 Teaching Language**	英文 English
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）**	马一方 助理教授 Assistant Professor Yifang Ma 统计与数据科学系 Department of Statistics and Data Science
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	待公布 To be announced
10.	**选课人数限额(可不填) Maximum Enrolment （Optional）**	

11. 授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
学时数	48				48

Credit Hours					

12.	先修课程、其它学习要求 **Pre-requisites or Other Academic Requirements**	MA102a 数学分析 II / MA102B 高等数学（下）A MA102a Mathematical Analysis II / MA102B Calculus II A
13.	后续课程、其它学习规划 **Courses for which this course is a pre-requisite**	数据分析、社交网络分析、数据可视化及应用 Data Analysis, Social Network Analysis, Data Visualization and Application
14.	其它要求修读本课程的学系 **Cross-listing Dept.**	

教学大纲及教学日历 SYLLABUS

15. **教学目标 Course Objectives**

本课程通过理论与实践相结合的形式帮助学生充分掌握数据科学中基本工具、理论和方法，其中包括：数据科学中的数学理论和常用方法；不同数据类型的分析和可视化；复杂数据的清理、降维和建模介绍等。

This course uses a combination of theory and practice to help students fully understand the basic tools, theories and methods in data science, including: mathematical theories and common methods in data science; analysis and visualization of different data types; complex data cleaning, dimensionality reduction and modeling, etc.

16. **预达学习成果 Learning Outcomes**

通过本课程的学习，学生预期可达到：

- 使用 Python 和其他工具来获取，清理和处理数据
- 使用统计方法进行快速探索、可视化、描述复杂数据结构
- 使用数据科学理论对数据进行分析、建模和预测

On successful completion of the course, students should be able to:

- Use Python and other tools to collect, clean, and process data.
- Use statistical methods to quickly explore, visualize, and describe complex data structures.
- Use data science theory to analyse, model, and predict real data.

17. **课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）**
Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

前言：简介 （2 学时）
数据科学导论
Part 0: Course Overview (2 hours)
Introduction to data science

第一部分：Python 程序设计（12 学时）
Python 基本数据类型；流程控制；输入输出；函数与模块等；标准库，内置函数；科学计算包 Numpy & Scipy；正则表达式；网页爬虫；表格数据处理；数据清理；数据可视化等
Part 1 Python Programming (12 hours)
Basics of python: data types; flow control; IO; function & modularity
Python standard library; built-in functions
Numerical computing using numpy & scipy
Regular expressions
Scraping data from web
Analyzing tabular data using pandas
Data Wrangling
Data visualization with matplotlib

第二部分： 数据科学中的理论（12 学时）
概率，变量，分布，数据关系，模拟实验设计，面向对象，程序优化，Skewed data，图算法和网络科学理论，邻接矩阵等
Part 2 Foundational Mathematics for Programming and Data Science (12 hours)
Basic probabilities
Single variable analysis
Normal distributions
Data relationships
Simulation and top-down design
Object-oriented programming
Code optimization
Skewed data
Basic graph/network theory/matrix

第三部分： 数据分析和可视化（12 学时）
数据分析与可视化方法（pandas/matplotlib/seaborn）;时间序列数据；文本数据；图像数据；数据降维；网络数据分析与可视化；交互可视化简介等。
Part 3 Data Analysis and Visualization (12 hours)
Exploratory Data Analysis and effective visualization: pandas/matplotlib/seaborn
Time series
Text analysis
Image data
Dimensionality Reduction/PCA/MDS/LLE
Network analysis and visualization: Gephi
Interactive Visualization: Plotly/Bokeh/ /D3

第四部分：机器学习简介（10 学时）
回归分析，贝叶斯分析， Scikit-Learn 简介；决策树与随机森林简介等。
Part 4 Introduction to Machine Learning (10 hours)
Regression
Bayes
Introduction to scikit-learn, learning a model
Decision trees and Random forests

18. 教材及其它参考资料 Textbook and Supplementary Readings

参考资料：

McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.". ISBN-10: 1491957662

Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie，The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) ISBN-10: 0387848576

课程评估 ASSESSMENT

19.

评估形式 Type of Assessment	评估时间 Time	占考试总成绩百分比 % of final score	违纪处罚 Penalty	备注 Notes
出勤 Attendance		10		
课堂表现 Class Performance				
小测验 Quiz				
课程项目 Projects		20		
平时作业 Assignments		50		
期中考试 Mid-Term Test				
期末考试 Final Exam				
期末报告 Final Presentation		20		
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)				

20. 记分方式 GRADING SYSTEM

√ A. 十三级等级制 Letter Grading
☐ B. 二级记分制（通过/不通过） Pass/Fail Grading

课程审批 REVIEW AND APPROVAL

21. 本课程设置已经过以下责任人/委员会审议通过
This Course has been approved by the following person or committee of authority

马一方

南方科技大学 SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

COURSE SPECIFICATION

The course information as follows may be subject to change, either during the session because of unforeseen circumstances, or following review of the course at the end of the session. Queries about the course should be directed to the course instructor.

1.	**课程名称 Course Title**	数据科学导论 Introduction to Data Science
2.	**授课院系 Originating Department**	统计与数据科学系 Department of Statistics and Data Science
3.	**课程编号 Course Code**	STA217
4.	**课程学分 Credit Value**	3
5.	**课程类别 Course Type**	专业选修课 Major Elective Courses
6.	**授课学期 Semester**	秋季 Fall【2022 秋季学期起】
7.	**授课语言 Teaching Language**	英文 English
8.	**授课教师、所属学系、联系方式（如属团队授课，请列明其他授课教师） Instructor(s), Affiliation& Contact （For team teaching, please list all instructors）**	马一方 助理教授 Assistant Professor Yifang Ma 统计与数据科学系 Department of Statistics and Data Science
9.	**实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact**	待公布 To be announced
10.	**选课人数限额(可不填) Maximum Enrolment （Optional）**	

11. 授课方式 Delivery Method	讲授 Lectures	习题/辅导/讨论 Tutorials	实验/实习 Lab/Practical	其它(请具体注明) Other（Please specify）	总学时 Total
学时数 Credit Hours	48				48

12. 先修课程、其它学习要求 Pre-requisites or Other Academic Requirements	MA102a 数学分析 II / MA102B 高等数学（下）A MA102a Mathematical Analysis II / MA102B Calculus II A
13. 后续课程、其它学习规划 Courses for which this course is a pre-requisite	数据分析、社交网络分析、数据可视化及应用 Data Analysis, Social Network Analysis, Data Visualization and Application
14. 其它要求修读本课程的学系 Cross-listing Dept.	

教学大纲及教学日历 SYLLABUS

15. 教学目标 Course Objectives

本课程通过理论与实践相结合的形式帮助学生充分掌握数据科学中基本工具、理论和方法，其中包括：数据科学中的数学理论和常用方法；不同数据类型的分析和可视化；复杂数据的清理、降维和建模介绍等。

This course uses a combination of theory and practice to help students fully understand the basic tools, theories and methods in data science, including: mathematical theories and common methods in data science; analysis and visualization of different data types; complex data cleaning, dimensionality reduction and modeling, etc.

16. 预达学习成果 Learning Outcomes

通过本课程的学习，学生预期可达到：

- 使用 Python 和其他工具来获取，清理和处理数据
- 使用统计方法进行快速探索、可视化、描述复杂数据结构
- 使用数据科学理论对数据进行分析、建模和预测

On successful completion of the course, students should be able to:

- Use Python and other tools to collect, clean, and process data.
- Use statistical methods to quickly explore, visualize, and describe complex data structures.
- Use data science theory to analyse, model, and predict real data.

17. 课程内容及教学日历 （如授课语言以英文为主，则课程内容介绍可以用英文；如团队教学或模块教学，教学日历须注明主讲人）

Course Contents (in Parts/Chapters/Sections/Weeks. Please notify name of instructor for course section(s), if this is a team teaching or module course.)

前言：简介 （2 学时）
数据科学导论
Part 0: Course Overview (2 hours)
Introduction to data science

第一部分：Python 数据分析基础（12 学时）
Python 基本数据类型；流程控制；输入输出；函数与模块等；标准库，内置函数；科学计算包 Numpy & Scipy；表格数据处理；数据清理；非结构数据分析等
Part 1 Python Programming (12 hours)
Basics of python: data types; flow control; IO; function & modularity
Python standard library; built-in functions
Numerical computing using numpy & scipy
Analyzing tabular data using pandas
Data Wrangling
Unstructured data analysis

第二部分： 数据科学中的理论（12 学时）
概率，变量，分布，数据关系， Skewed data，图算法和理论等
Part 2 Foundational Mathematics for Programming and Data Science (12 hours)
Basic probabilities
Single variable analysis
Normal distributions
Data relationships
Skewed data analysis
Basic graph/network theory

第三部分： 数据分析和可视化（12 学时）
数据分析与可视化方法;网络数据分析与可视化；交互可视化简介等。
Part 3 Data Analysis and Visualization (12 hours)
Exploratory Data Analysis and effective visualization
Trends, Category, Uncertainty visualization
Network visualization
Interactive Visualization

第四部分：实践（10 学时）
数据建模，模型优化，模拟实验设计，网络数据分析与可视化。
Part 4 Practice (10 hours)
Modeling
Code optimization
Simulation
Network analysis and visualization

18. 教材及其它参考资料 Textbook and Supplementary Readings

参考资料：

McKinney, W. (2012). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. " O'Reilly Media, Inc.". ISBN-10: 1491957662

Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie，The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) ISBN-10: 0387848576

课程评估 ASSESSMENT

19.	评估形式	评估时间	占考试总成绩百分比	违纪处罚	备注

Type of Assessment	Time	% of final score	Penalty	Notes
出勤 Attendance		10		
课堂表现 Class Performance				
小测验 Quiz				
课程项目 Projects		15		
平时作业 Assignments		30		
期中考试 Mid-Term Test		30		
期末考试 Final Exam				
期末报告 Final Presentation		15		
其它（可根据需要改写以上评估方式） Others (The above may be modified as necessary)				

20. 记分方式 **GRADING SYSTEM**

√ A. 十三级等级制 **Letter Grading** ☐ B. 二级记分制（通过/不通过） **Pass/Fail Grading**

课程审批 REVIEW AND APPROVAL

21. 本课程设置已经过以下责任人/委员会审议通过
This Course has been approved by the following person or committee of authority

马一方