

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	商务大数据分析 Business Analytics with Big Data				
2.	授课院系 Originating Department	金融系 Department of Finance				
3.	课程编号 Course Code	FET306				
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	春季 Spring				
7.	授课语言 Teaching Language	中英双语 English & Chinese				
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)	胡大宁, 金融系, hudn@sustech.edu.cn HU Daning, Department of Finance, hudn@sustech.edu.cn				
9.	实验员/助教、所属学系、联系方式 Tutor/TA(s), Contact	滕琪 (实验员), 金融系, tengq@mail.sustech.edu.cn TENG Qi (Tutor), Department of Finance, tengq@mail.sustech.edu.cn				
10.	选课人数限额(可不填) Maximum Enrolment (Optional)	50				
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	32		32		64

12. 先修课程、其它学习要求
Pre-requisites or Other Academic Requirements

13. 后续课程、其它学习规划
Courses for which this course is a pre-requisite

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**

现代商业通过交易系统，内外部的档案系统，以及线上活动记录可以获得越来越多的商务数据。如何运用先进的商务数据分析方法和模型，配合数据科学与信息系统技术来从这些数据中发现价值以支持商业决策和应用，是现代商业的重要挑战。

本课程结合计量经济学分析等商务分析方法和机器学习等数据挖掘方法来分析大规模的商务数据。这门课主要涵盖以下课题：数据的获取，处理和挖掘，社交网络分析，计量经济因果推断方法，以及对计量经济学方法和机器学习的结合等。该课程的考核包括期末考试和平时作业。

Firms now have more data than ever generated from transaction processing systems, internal and external documents, and online activities such as web sites, shopping experiences, and social networks. The challenge is to apply novel analytical and computing methods to derive insights from such business data and develop effective strategies or applications accordingly.

This course aims to utilize and integrate analytical models and methods from business domains such as econometrics and technology domains such as machine learning (data mining), to analyze large-scale business data sets. The main topics include Data Mining, Social Network Analysis, Causal Inference, and Integration of Econometrics and Machine Learning. The assessment concludes with a final exam and assignments.

16. **预达学习成果 Learning Outcomes**

1. 对商务大数据进行计量经济因果识别的建模和分析。
2. 对商务大数据进行结合实际的介绍，学习数据清洗预处理。
3. 学习如何进行社交网络的商业智能分析。
4. 学习如何进行各种类型的商务数据的挖掘和解析，如何验证和解释数据结果。

1. Students will gain knowledge of causal inference and econometric analysis.
2. Students will learn tools and techniques to extract and clean business data. Students will learn how to analyze data mining unstructured business data such as text data including documents, web pages, emails, etc.
3. Students will learn how to model and analyze large-scale social network data.
4. Students will learn how to mine and analyze large-scale business data and interpret the results.

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.)

理论（32 学时）

第一部分 商务大数据分析基础（4 学时）

1.1 商务大数据分析介绍（2 学时）

本部分介绍商务大数据分析的基本理论。

1.2 描述性数据分析基础（2 学时）

本部分介绍商务大数据分析的主要统计方法和相关计算机科学基础知识。

第二部分 数据库管理（4 学时）

2.1 数据库管理系统介绍（2 学时）

本部分介绍 DBMS（数据库管理系统）如何处理商务大数据分析所需数据的储存、操纵和预处理。

2.2 结构化查询语言基础（2 学时）

本部分介绍 SQL（结构化查询语言）以及在商业分析中的应用。

第三部分 社交网络数据分析（8 学时）

3.1 节点级分析：各类中心度分析（2 学时）

本部分介绍网络度中心性的概念的计算方法。

3.2 链接级分析（2 学时）

本部分介绍社交网络中的链接。

3.3 群组级分析：社区侦测（2 学时）

本部分介绍如何在社交网络中侦测社区，即连通性非常密集的图的子图。

3.4 网络级分析：网络拓扑结构分析（2 学时）

本部分介绍社交网络数据分析中的方法应用：随机、小世界、无标度网络拓扑模型。

第四部分 因果推断基础（6 学时）

4.1 线性回归和因果性（2 学时）

本部分介绍线性回归、假设检验以及因果推断的可能偏差和解决方法。

4.2 工具变量，自然实验和断点回归（2 学时）

本部分介绍计量因果识别的经典方法：工具变量和 2SLS，自然实验和双重差分，断点回归。

4.3 面板数据和匹配（2 学时）

本部分介绍面板数据的固定效应和随机效应，介绍匹配方法以及如何使用。

第五部分 商务分析的先进话题（8 学时）

5.1 数据挖掘基础（2 学时）

本部分介绍数据挖掘的基础，以及如何使用数据挖掘分析商务大数据。

5.2 金融科技案例分析（2 学时）

本部分介绍金融科技的经典案例。

5.3 市场案例分析（2 学时）

本部分介绍市场的经典案例。

5.4 随机森林（2 学时）

本部分介绍决策树与随机森林的构建及优点。

第六部分 期末复习（2 学时）

本部分回顾和复习本课程所有重点内容。

实验（32 学时）

第一部分 Excel 数据分析（4 学时）

1.1 描述统计（2 学时）

本部分主要是辅导学生使用 Excel 中的“数据分析”功能对数据进行描述统计。

1.2 数据透视表（2 学时）

本部分主要是辅导学生使用 Excel 中的“数据透视表”功能对数据进行可视化分析。

第二部分 ER 建模（2 学时）

本部分主要是辅导学生使用 VISIO 对现实世界数据集进行 ER 建模。通过本章的学习，学生会设计并绘制 E-R 图。

第三部分 数据库管理（4 学时）

3.1 select 查询语句和聚合函数（2 学时）

本部分主要讲解 SQL 语句中的 select 查询语句以及聚合函数的使用。

3.2 数据库、基本表的操作（2 学时）

本部分主要讲解在 SQL Server 中使用 SQL 语句创建数据库、表以及对表中数据进行插入、修改和删除等处理语句。

第四部分 网络可视化——NetDraw（4 学时）

4.1 SQL Server 数据处理（2 学时）

本部分主要讲解使用 SQL 从原始数据中获得所需的信息，以及网络可视化工具 NetDraw 的介绍。

4.2 NetDraw 网络分析（2 学时）

本部分主要介绍 NetDraw 的功能，通过本部分的学习，学生会使用 NetDraw 进行社会网络分析。

第五部分 网络可视化——R（10 学时）

5.1 R 语言的基础语法（4 学时）

本部分主要讲解 R 语言包的使用，从最基础的语法开始讲解，让学生对 R 语言有初步的认识。

5.1.1 vector、factor、matrix、list（2 学时）

本部分主要讲解 R 基础语法中的 vector、factor、matrix、list。

5.1.2 Data Frames、Flow Control、R plot（2 学时）

本部分主要讲解 R 基础语法中的 Data Frames、Flow Control 和 R plot。

5.2 R 语言的 igraph 包的使用（6 学时）

本部分重点介绍 R 语言中 igraph 包的使用，通过本部分的学习，学生会使用 R 语言绘制相关图谱，对社会网络进行分析。

5.2.1 R igraph（2 学时）

本部分重点介绍 R 语言中 igraph 包，以及如何读取使用 R 语言读取社会网络文件。

5.2.2 Plotting networks with igraph（2 学时）

本部分主要辅导学生如何使用 igraph 包画出社会网络图。

5.2.3 Network descriptives（2 学时）

本部分主要辅导学生对社交网络数据进行基本测度。

第六部分 Python 基础（2 学时）

本章主要讲解 Python 的基础语法。通过本章的学习，学生对 Python 语言有初步的认识。

第七部分 Python 爬虫（4 学时）

7.1 Requests 库（2 学时）

本部分主要介绍 HTML 标签，以及 Python 中 Requests 库的使用

7.2 BeautifulSoup 库（2 学时）

本部分主要介绍 Python 中 BeautifulSoup 库，并使用其来抓取网络上的图片和文本。

第八部分 Final Project (2 学时)

本部分主要是学生利用 SQL Server 或 Mysql、NetDraw、R 或 Python 完成社会网络分析项目。

Part 1: Basics of Business Analytics (BA) (4 hours)

1.1 Introduction to BA (2 hours)

This part introduces the basic concepts in business analytics.

1.2 Basic Descriptive Data Analysis (2 hours)

This part introduces main statistical techniques and computer sciences basic in business analytics.

Part 2: Database Management (4 hours)

2.1 Introduction to Database Management Systems (2 hours)

This part introduces how DBMS handles the storage, manipulation and preprocessing of the data needed in business analytics.

2.2 Basics of Structural Query Language (2 hours)

This part introduces SQL and its applications in business analytics.

Part 3: Social Network Analysis (8 hours)

3.1 Node Level Analysis: Centrality (2 hours)

This part introduces the definition and calculations of degree centrality.

3.2 Link Level Analysis (2 hours)

This part introduces links in a social network.

3.3 Group Level Analysis: Community Detection (2 hours)

This part introduces how to detect a community in a social network, that is, a subgraph of a graph with dense connectivity.

3.4 Network Level Analysis: Network Topologies (2 hours)

This part introduces the methodologies in social network analysis: random, small-world or scale free topological models.

Part 4: Causal Inference Basics (6 hours)

4.1 Linear Regression and Causality (2 hours)

This part introduces linear regression, hypothesis testing as well as the possible biases and solutions to causality.

4.2 Instrument Variables, Natural Experiment and Regression Discontinuity (2 hours)

This part introduces the traditional econometric methods to do causal inference: instrument variables and 2SLS, natural experiment and Difference-in-differences, as well as regression discontinuity.

4.3 Panel Data and Matching (2 hours)

This part introduces the fixed/random effect of panel data as well as the definition and application of matching.

Part 5: Advanced Topic in Business Analytics (8 hours)

5.1 Data Mining Basics (2 hours)

This part introduces data mining basics and how to use it in business data analytics.

5.2 Case Study in FinTech (2 hours)

This part introduces the classical studies in FinTech.

5.3 Case Study in Marketing (2 hours)

This part introduces the classical studies in Marketing.

5.4 Random Forest (2 hours)

This part introduces how to grow decision tree and random forest as well as their strength.

Part 6: Final Review (2 hours)

This part reviews all the important content in this course.

LAB (32 hours)

Part 1 Descriptive Data Analysis with Excel (4 hours)

1.1 Data Analysis (2 hours)

This section mainly shows how to use the “data analysis” function in Excel to analyze data.

1.2 Pivot Table (2 hours)

This section mainly shows how to use the “Pivot Table” to visually analyze the data.

Part 2 ER Modeling (2 hours)

This part introduces the ER modeling of real world datasets. In this part, students will learn how to draw E-R diagrams by VISIO software.

Part 3 Database Management (4 hours)

3.1 Select Query Statement and Aggregate Function (2 hours)

This section mainly explains the use of select queries and aggregate functions in SQL statements.

3.2 Database and Table Operation (2 hours)

This section mainly explains how to create databases, tables in SQL Server, and insert, modify, and delete data in the table.

Part 4 Network Visualization—NetDraw (4 hours)

4.1 SQL Server Data Processing (2 hours)

This section mainly explains how to use SQL to search relevant information from such raw data, and introduces the network and data visualization tools—NetDraw.

4.2 NetDraw for Network (2 hours)

This section mainly explains the functions of the network visualization tool—NetDraw. In this part, students will learn how to use NetDraw for social network analysis.

Part 5 Network Visualization—R (10 hours)

5.1 R Language (4 hours)

This part mainly explains the use of the R language package, starting with the most basic grammar, and giving students a preliminary understanding of the R language.

5.1.1 Vector, Factor, Matrix, List (2 hours)

This section mainly explains vector, factor, matrix, and list in R.

5.1.2 Data Frames, Flow Control, R Plot (2 hours)

This section mainly explains Data Frames, Flow Control, R plot in R.

5.2 R igraph Basics (6 hours)

This part focuses on the use of the igraph package in R. In this part, students will learn how to use the R to draw maps and analyze the social network.

5.2.1 R igraph (2 hours)

This section focuses on the use of the igraph package in R, and shows how to read network data from files.

5.2.2 Plotting Networks with igraph (2 hours)

This section mainly tutors students how to use the igraph package to draw a social network map.

5.2.3 Network Descriptive (2 hours)

This section mainly tutors students to make measures of social network data.

Part 6 Python Basics (2 hours)

This part focuses on the basics of Python. In this part, students will have a preliminary understanding of Python.

Part 7 Web Crawler (4 hours)
 7.1 Requests Library (2 hours)
 This section explains the HTML tags and the use of Requests library in Python.
 7.2 BeautifulSoup Library (2 hours)
 This section explains the use of BeautifulSoup library in Python, and shows how to program web crawler in Python to grab images and text from the internet.

Part 8 Final Project (2 hours)
 This section is mainly for students to complete social network analysis projects using SQL Server/MySQL, NetDraw, R/Python.

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

乔舒亚·安格里斯特, 约翰-斯特芬·皮施克. 基本无害的计量经济学[M]. 格致出版社, 2012.

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				
课堂表现 Class Performance				
小测验 Quiz				
课程项目 Projects				
平时作业 Assignments		30		
期中考试 Mid-Term Test				
期末考试 Final Exam		70		
期末报告 Final Presentation				
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