

《数据科学导论》课程教学大纲

课程英文名	Introduction to Data Science				
课程代码	B0501570	课程类别	专业课	课程性质	专业选修
学 分	2		总学时数	32	
开课学院	计算机学院		开课基层教学组织	智能信息处理课程组	
面向专业	计算机科学与技术、智能计算与数据科学（计算机科学与技术）、计算机科学与技术(第二学士学位)、经济学		开课学期	第4/5学期	

注：课程类别是指学科基础课/专业课/实践环节/通识公共课/公共基础课/其他；

课程性质是指通识必修/通识选修/学科必修/专业必修/专业选修/实践必修/实践选修。

一、课程目标

数据科学家是大数据时代最急需的人才，他们应具有宽广的视野，同时具有扎实的理论和技术功底。“数据科学模块”的整个课程体系，涵盖一系列基础课程包括数据科学导论、大数据基础、大数据开发实践、数据挖掘等一系列专业课程，为最终培养能够解决复杂业务问题的合格的数据科学家服务。本课程作为“数据科学模块”的先导课和认知类课程，致力于以形象生动的教学模式为学生普及数据挖掘、大数据相关的基础知识、核心概念和思维模式，从工程技术、法律规范、应用实践等不同角度描绘数据科学的美好蓝图，把学生引进数据科学的大门。

通过此课程学习，使得学生对数据分析的基本原理和关键技术有一个初步了解，培养以数据为中心的问题求解的思想方法和初步能力，为他们今后从事计算机相关领域的工作打下坚实的基础。通过课程的理论知识讲解，并开展课程相关实验，使得学生在了解数据科学技术理论及其应用的同时，结合国家建设和民族复兴的新时代背景，增强学生家国情怀与文化自信，激发学生使命感和责任心。

通过理论教学和实践活动，达到以下课程目标：

课程目标 1：能够了解数据分析的基本思路、基本理论与基本方法，包括数据抽取、数据预处理、数据建模以及模型评价、数据分析方法、数据可视化等。

课程目标 2：能够应用现代信息技术工具，通过文献检索及研究，理解数据分析常用的技术和工具。

课程目标 3：培养实践应用能力，具备将所掌握的数据分析技术运用到解决实际问题的能力。

课程目标 4：具备基本的科学素养，及时了解数据科学在国内外的实际应用案例，及时掌握国

家相关方面的科技战略需求，树立强烈的爱国主义使命感与责任心。

二、课程目标与毕业要求对应关系

课程的课程目标对计算机科学与技术专业毕业要求指标点的支撑情况如表 2 所示：

表 1 课程目标与毕业要求对应关系

毕业要求	指标点	课程目标及支撑权重
毕业要求 1：工程知识：掌握数学、自然科学、工程基础、计算机专业领域的知识，并能应用于计算机领域复杂工程问题的解决方案中。	1-3 能够运用计算机专业知识，对计算机领域复杂工程问题解决方案进行分析与优化。	目标 1：0.5 目标 2：0.5
	1-4 掌握计算机系统、物联网、人工智能、大数据、网络安全等某个专业领域的知识，并用于解决计算机领域的复杂工程问题。	目标 1：0.5 目标 2：0.5
毕业要求 2：问题分析：能够应用数学、自然科学和工程科学的基本原理，以及科学思维方法，对计算机领域的复杂工程问题进行识别、表达和分析，并通过文献查阅与研究获得有效结论。	2-1 能够应用数学、自然科学、工程科学和计算科学的基本原理识别、表达计算机领域的复杂工程问题。	目标 1：0.3 目标 2：0.4 目标 3：0.4
毕业要求 4：研究：能够基于包括计算学科在内的科学原理，采用科学方法研究计算机领域的复杂工程问题，包括设计实验、分析与解释数据、并通过信息综合得到合理有效的结论。	4-3 能够收集、分析与解释数据，并通过信息综合得到合理有效的结论。	目标 2：0.4 目标 3：0.4 目标 4：0.2
毕业要求 12：终身学习：具有自主学习和终身学习的意识，有不断学习和适应信息技术高速发展的能力。	12-1 能够在信息技术高速发展的大背景下，建立自主学习和终身学习的意识。	目标 1：0.3 目标 2：0.3 目标 3：0.2 目标 4：0.2

课程的课程目标对智能计算与数据科学（计算机科学与技术专业）毕业要求指标点的支撑情况如表 1 所示：

表 2 课程目标与毕业要求对应关系

毕业要求	指标点	课程目标及支撑权重
毕业要求 1：工程知识：能够掌握数学、自然科学、工程基础、计算机软硬件知识、人工智能、智能计算和数据科学的基础理论及专业知识，并应用在人工智能、智能计算和大数据专业领域及其他相关交叉领域的复杂工程问题的解决方案中。	1-3 能够运用计算机专业知识，对计算机领域复杂工程问题解决方案进行分析与优化。。	目标 1：0.5 目标 2：0.5
	1-4 掌握人工智能、智能计算、大数据等某个专业领域的知识，并用于解决计算机领域的复杂工程问题。	目标 1：0.5 目标 2：0.5
毕业要求 2：问题分析：能够应用数学、自然科学、工程科学、人工智能、智能计算和大数据的基本原理，对人工智能和大数据专	2-1 能够应用数学、自然科学、工程科学和计算科学的基本原理识别、表达计算机领域的复杂工程问题。	目标 1：0.3 目标 2：0.4 目标 3：0.4

业领域及其他相关交叉领域的复杂工程问题进行识别、表达、分析和抽象建模，并通过文献查阅与研究获得有效结论。	问题。	
毕业要求 4：研究：具有基本的科学素养和研究意识，能够采用科学方法研究人工智能、智能计算和大数据专业领域及其他相关交叉领域的复杂工程问题，包括设计实验、分析与解释数据、并通过信息综合得到合理有效的结论。	4-3 能够收集、分析与解释数据，并通过信息综合得到合理有效的结论。	目标 2：0.4 目标 3：0.4 目标 4：0.2
12. 终身学习：具有自主学习和终身学习的意识，有不断学习和适应信息技术高速发展的能力。	12-1 能够在信息技术高速发展的大背景下，建立自主学习和终身学习的意识。	目标 1：0.3 目标 2：0.3 目标 3：0.2 目标 4：0.2

三、课程目标与教学内容和方法的对应关系

《数据科学导论》课程目标与教学内容、建议教学方法的对应关系如表 3 所示。教师可根据实际情况和自身特长采用合适的教学方法。

表 3 课程目标与教学内容、教学方法的对应关系

教学内容	教学方法	课程目标
1. 数据科学概述	课堂讲授、课堂提问、课堂讨论	1,4
2. Python 与数据科学	课堂讲授、课堂提问、课后实践	2,3
3. OLTP 与数据服务	课堂讲授、案例分析、课后实践	2,3
4. OLAP 与结构化数据分析	课堂讲授、案例分析、课后实践	2,3
5. 数据清洗与数据集成	课堂讲授、案例分析、课堂讨论	1,2,3
6. 数据的深度分析	课堂讲授、案例分析、课后实践	1,2,3
7. 文本分析	课堂讲授、案例分析、课后实践	1,2,3
8. 社交网络分析	课堂讲授、案例分析、课堂讨论	2,3
9. 语义网与知识图谱	课堂讲授、案例分析、课堂讨论	1,4

该课程详细教学内容和方法如下所述。

1. 数据科学概述

(1) 教学内容：

- 数据科学的定义
- 数据科学的定位
- 数据科学家
- 数据科学的基本原则
- 数据处理流程：时间维度的纵向视角
- 数据处理系统的架构：系统维度的计算视角
- 数据的多样性：数据类型维度的横向视角

- 数据价值的挖掘：价值维度的价值提升视角

(2)教学重点：主要介绍数据科学的基本概念、大数据及其价值、数据的生命周期和数据处理的流程(包括数据的采集和获取、数据预处理/清洗和集成(Preprocessing)、数据管理(Management)、数据分析(Analysis)、可视化(Visualization)和解释等)、以及数据处理的不同模式。

(3)教学难点：数据科学的定位及如何进行其价值挖掘。

(4)教学要求：能够理解数据科学的基本定义和定位；了解数据科学的基本原则及处理流程。

思政融合点 1：引导学生查阅资料，分析并总结国内外数据科学相关的应用领域。特别调研我国在数据管理、数据分析、可视化等领域取得成就的相关文献，激发学生学习先进技术的热情。

2. Python 与数据科学

(1)教学内容：

- Python 概述
- Python 开发环境配置（Setup）
- 通过一系列实例学习 Python
- 第三方库和实例

(2)教学重点：Python 语言部分，包括 Python 语言基础，以及 Python 的几个重要的库(重点讲的介绍)。

(3)教学难点：Python 语言基础，以及 Python 的几个重要的库，包括 Pandas 入门、数组和 Numpy 等)。

(4)教学要求：通过本小节学习，掌握 Python 的基本语法并对了解数据分析的常见库。除此之外，完成对应的上机实验，即 Python 安装配置、python 入门、pandas 入门。

3. OLTP 与数据服务

(1)教学内容：

- 面向 OLTP 应用的 RDBMS 数据库技术
- 面向数据服务的 NoSQL 数据库技术
- NewSQL 数据库技术

(2)教学重点：OLTP 概念、RDBMS 数据库技术、NoSQL 数据库技术。

(3)教学难点：RDBMS 数据库与 NoSQL 数据库之间的区别及演化原因。

(4)教学要求：通过本小节学习，了解关系型数据库的原理、种类，非关系型数据库的原理、种类，两者之间的区别。除此之外，完成对应的上机实验，即 SQL 入门、OLTP 实验。

4. OLAP 与结构化数据分析

(1)教学内容：

- 联机分析处理（OLAP）与结构化数据分析

- 高性能 OLAP 系统的关键技术
- 结构化数据分析工具介绍

(2)教学重点：联机分析处理意义、OLAP 主要操作、关键技术及工具。

(3)教学难点：OLAP 系统关键技术，包括列存储技术、位图索引技术、内存数据库技术等。

(4)教学要求：通过本小节学习，掌握联机分析处理的概念、意义及关键技术，了解常见的结构化数据分析工具。对除此之外，完成对应的上机实验，即 OLAP 实验。

5. 数据清洗与数据集成

(1)教学内容：

- 数据抽取、转换与装载
- 数据清洗
- 数据集成

(2)教学重点：ETL 操作、数据清洗的意义/方法/步骤、数据集成模式。

(3)教学难点：数据清洗的常见做法、数据集成即多源数据整合的全流程。

(4)教学要求：通过本小节学习，掌握 ETL 操作的概念及意义，了解数据清洗的常见做法及数据集成的过程。

6. 数据的深度分析（数据挖掘、机器学习）

(1)教学内容：

- 机器学习与数据挖掘简介
- 主流机器学习与数据挖掘方法
- 主流数据深度分析工具

(2)教学重点：介绍数据模型和通用的数据分析方法入门、数据可视化等。

(3)教学难点：重点对分类（Decision Tree, SVM, Naïve Bayes）、聚类(K-Means、DBSCAN)、回归(线性回归和 SVR)、和降维(SVD、PCA)等类别的主流方法给予介绍。同时介绍 Python 的机器学习库 scikit learn 和可视化库 matplotlib。

(4)教学要求：通过本小节学习，了解数据模型和通用的数据分析方法的种类、各自的原理和作用。除此之外，完成对应的上机实验，即首先进行 python 中机器学习库 scikit learn 和可视化库 matplotlib 的实验，然后再利用上述两个库实现集中主流的数据分析方法。

思政融合点 2：数据挖掘和机器学习目前已成为计算机领域的热门研究课题，介绍我校在数据挖掘领域取得的国际领先成果，激发学生对于参与到前沿课题研究的自信感和热情。引导学生思考使用数据挖掘技术的可能性、可靠性、可行性，以及其可能引发的社会问题，探讨如何利用数据深度分析造福人类社会。

7. 文本分析

(1)教学内容:

- 文本分析的意义
- 文本分析的任务和方法
- 文本分析可视化
- 文本分析软件和工具

(2)教学重点: 依赖于分布式的大数据处理平台(Hadoop/Spark/Lucene)和各种工具, 包括 Scrapy 爬虫、Jieba 分词、NLTK、scikit learn 等, 使用各种文本分析方法, 对文本进行分词、索引与检索、实体识别、情感分析、话题发现、可视化等。

(3)教学难点: 重点讲解文本爬取、文本表示、文本分类、文本检索。

(4)教学要求: 通过本小节学习, 了解文本分析的原理、流程和使用工具, 特别掌握文本数据可视化的方法如词云等。除此之外, 完成对应的上机实验, 即使用 python 实现 Twitter 情感分析。

思政融合点 3: 文本分析是当前数据挖掘的热门技术之一, 指导学生学习文本技术流程, 并应用课堂上学习的文本分析技术对当下热门的话题进行实际分析挖掘, 引导学生思考通用文本分析技术在其他领域的可能性、可靠性、可行性。

8. 社交网络分析

(1)教学内容:

- 简介
- 社交网络分析的应用
- 社交网络分析方法
- 软件

(2)教学重点: 介绍社交网络的概念、应用及其分析方法。

(3)教学难点: 掌握社交网络的基本属性、复杂网络的拓扑特性及其可达性、最短路径、最小生成树等概念。

(4)教学要求: 通过本小节学习, 了解社交网络分析的基本概念、应用及其分析方法, 为理解人类各种社交关系的形成、行为特点分析以及信息传播的规律提供的基础。

9. 语义网与知识图谱(选学)

(1)教学内容:

- 语义网的基本概念
- 语义网体系结构
- 语义网的关键技术
- 知识库与知识图谱

(2)教学重点：掌握语义网的体系架构（每一层的作用）及其关键技术；了解知识图谱创建过程及挖掘原理。

(3)教学难点：知识图谱的创建及进一步挖掘。

(4)教学要求：通过本小节学习，了解语义网和知识图谱的基本概念、应用及其分析方法，掌握用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系的能力。

思政融合点 4：知识图谱是人工智能领域的一颗掌上明珠，也是目前的一个新兴研究热点，引导学生了解我国目前在知识图谱领域取得的成就以及领域应用，激发学生学习新技术、开展新应用的热情。

思政融合点 5：总结课堂内容，让学生深入了解成为一个合格的数据科学家需要具备的各项素质，增强学生家国情怀与文化自信，激发学生使命感和责任心。

四、 实践环节及要求

1. 实验项目和基本要求

表 4 实验项目和基本要求

序号	实验项目	学时	基本要求	实验性质	实验类别
1	搭建 python 环境、python 入门	2	搭建 Python 环境、熟悉 python 基本语法并熟悉 python 常见库如 pandas、numpy	必做	综合
2	数据预处理相关实验	2	练习使用数据采集软件	必做	综合
3	数据分析方法实现	2	熟悉 Python 中机器学习库 scikit-learn 和可视化库 matplotlib 的实验，利用上述两个库实现集中主流的数据分析方法如 knn、svm、kmeans、线性回归、关联分析等	必做	综合
4	文本分析实现	2	使用现成分类器或自行训练分类器，实现 r 情感分析	必做	综合

2. 实验报告基本要求

实验报告至少包含以下几个部分：(1)实验目的；(2)实验环境；(3)实验过程(含实验方案、步骤、程序等)；(4)实验结果及结果分析；(5)实验总结。

五、 与其它课程的联系

先修课程：程序设计基础。

后续课程：无

六、 学时分配

总学时 32 学时，其中讲课 24 学时，课内上机 8 学时。如表 5 所示。

表 5 《数据科学导论》学时分配表

教学内容	讲课时数	实验时数	实践学时	课内上机时数	课外上机时数	自学时数	习题课时	讨论时数
1. 数据科学概述	2							
2. Python 与数据科学	4			2		2		
3. OLTP 与数据服务	2				2			
4. OLAP 与结构化数据分析	2				2			
5. 数据清洗与数据集成	2			2		2		
6. 数据的深度分析	4			2				
7. 文本分析	4			2		2		
8. 社交网络分析	2				2			
9. 语义网与知识图谱	2				2	2		
合计	24			8	8	8		
总计						32		

七、 课程目标达成途径及学生成绩评定方法

1. 课程目标达成途径

课程目标的达成途径如表 6 所示。

表 6 课程目标与达成途径

课程目标	达成途径
课程目标 1： 能够了解数据分析的基本思路、基本理论与基本方法，包括数据抽取、数据预处理、数据建模以及模型评价、数据分析方法、数据可视化等。	以引导式、启发式和总结式教学方法为主，通过重点/难点内容讲解、课后作业、进行随堂提问、课堂程序演示等模式，由浅入深，再宽度展开，旨在培养学生宽广的视野，形成完整的知识体系。
课程目标 2： 能够应用现代信息技术工具，通过文献检索及研究，理解数据分析常用的技术。	以启发式、分析式和研讨式教学方法为主，结合典型应用案例，通过上机实践、作业、随堂提问、课堂讨论等模式，帮助学生具备数据分析的思维方式，为实际应用问题建立问题模型，培育数据分析的能力。
课程目标 3： 培养实践应用能力，具备将所掌握的数据分析技术运用到解决实际问题的能力。	通过重点/难点内容讲解、上机实践、实验讲解等模式，帮助学生掌握数据科学的实践应用，培养学生的动手能力和浓厚兴趣。
课程目标 4： 具备基本的科学素养，及时了解数据科学在国内外的实际应用案例，及时掌握国家数据	通过课堂讲授、课后自学、文献查阅、课堂讨论、分析对比、总结报告等各种方式，让学生建立终生学习的意识；同时，进一步了解目前国内数据科学的发展水平和应用现状，从而建立

科学方面的科技战略需求，树立强烈的爱国主义使命感与责任心。	强烈的民族自豪感与爱国主义使命感。
-------------------------------	-------------------

2.学生成绩评定方法

该课程为考查课程，考查方式为期末提交学生作品。该课程采用形成性评价与终结性评价相结合的评价方法，学期总评成绩由二部分构成：平时成绩，占比 50%，包括课程思政实践、课堂成绩、上机实验报告等，教师可按照课堂实际情况进行调整；期末考查成绩，占比 50%。各部分的具体评价环节、关联课程目标、评价依据及方法和在总成绩中的占比，如表 7 所示。

表 7 课程考核与成绩评定方法

成绩构成	考核项目	考核关联的课程目标	考核依据与方法	占总评成绩的比重
平时成绩	课程思政实践	4	基于数据科学主题，通过课外文献查阅、课堂展示、课堂小组讨论、阅读报告等多种形式，考查学生对数据科学及其应用的了解情况以及核心价值观状况。	5%
	课堂成绩	1,2,3,4	考察考勤、课堂讨论、课堂互动、课堂练习等。	15%
	上机实验报告	1,2,3	考查上机实验完成情况。	30%
期末考试	期末考查	1,2,3,4	考查学生提交的期末作品的成绩。	50%
总评成绩		1,2,3,4	=平时成绩*50%+期末作品成绩*50%	100%

表 8. 考核内容详细评分标准

考核内容	评分标准			
	90-100	75-90	60-75	<60
课程思政实践	报告条理清晰，文字流畅，字数 \geq 4000，参考文献数量 \geq 8 且相关性强；内容完整且材料丰富，体现强烈的使命感、责任心与民族自豪感	报告条理清楚，字数 \geq 3000，参考文献数量 \geq 5 且相关性较好；内容完整，材料不够丰富，能体现学生的使命感、责任心与民族自豪感	报告有一定条理，字数 \geq 1000，参考文献数量 \geq 2 且基本相关；内容基本完整但材料较少，能体现学生的使命感与民族自豪感	报告字数 $<$ 1000，参考文献数量 $<$ 2；内容少，或有抄袭现象，体现不出学生的使命感与民族自豪感
课堂成绩	雨课堂测试、课堂练习、回答问题等据实评价；或参与回答次数在教学班前 15%	雨课堂测试、课堂练习、回答问题等据实评价；或参与回答次数在教学班前 50%	雨课堂测试、课堂练习、回答问题等据实评价；或参与回答次数在教学班前 85%	雨课堂测试、课堂练习、回答问题等据实评价；或参与回答次数在教学班后 15%
上机实验报告	程序运行流畅，功能完善，性能好；代码独创性好；算法与数据结构设计或代码实现等有 2 处及以上创	程序运行正常，功能基本实现，但不够完善、优化等，代码自创率应高于 50%；算法与数据结构设计或代	程序运行基本正常，有少量 bug；功能实现至少达到 80%，代码自创率应高于 30%；基本没有创新	程序运行 bug 多，功能实现低于要求的 80%

	新点	码实现等至少有一处创新点	性	
期末考查	报告条理清晰，文字流畅内容完整，字数 ≥ 3500 ，数据结构及算法设计合理且效率高，有2个以上创新点或改进，参考文献 ≥ 5 篇，查重率 $\leq 20\%$ ，汇报PPT图表清晰，设计美观，答辩过程脱稿讲解，分析条理清晰，问题回答准确，小组协作好	报告条理清楚，内容较完整，字数 ≥ 2500 ，数据结构及算法设计合理且性能较好，有1个以上创新点或改进，参考文献 ≥ 3 篇，查重率 $\leq 35\%$ ，汇报PPT设计较美观，答辩过程脱稿讲解，条理较清楚，问题回答基本正确，分工较合理	报告内容基本完整，字数 ≥ 1500 ，数据结构及算法设计基本合理，参考文献 ≥ 2 篇，查重率 $\leq 50\%$ ，汇报PPT美观性及内容一般，答辩过程大部分内容脱稿讲解，大部分问题能基本正确回答，分工基本合理	报告内容不完整，字数少于1500，数据结构及算法设计不能满足题目基本要求，查重率 $> 60\%$ ，有抄袭现象，汇报PPT界面文字多，内容不完整，答辩过程基本念PPT，只能回答少量问题，分工不够合理

八、教学资源

表9 课程的基本教学资源

资源类型	资源
教材	覃雄派等:《数据科学概论》，中国人民大学出版社，2018。
参考书籍或文献	1. Alberto Boschetti等:《数据科学导论-Python语言实现》，机械工业出版社，2016。 2. 石川等:《数据科学导论》，清华大学出版社，2021。 3. 俞东进等:《大数据基础、技术与应用》，科学出版社，2022.
教学文档	无

九、课程目标、毕业要求指标点达成度定量评价

在课程结束后，需要对每一个课程目标(含思政课程目标)进行达成度的定量评价，用以实现课程的持续改进。

课程目标达成度的定量评价算法：

- 使用教学活动(如课程思政实践、课后作业、课堂表现、文献阅读报告、期末作品等等)成绩或期末考试部分题目得分率作为评价项目，来对某个课程目标进行达成度的定量评价；
- 为保证考核的全面性和可靠性，要求对每一个课程目标的评价项目选择超过两种；
- 根据施教情况，评价项目可以由教师自行扩展，权重比例可以由教师自行设计；
- 对某一个课程目标有支撑的各评价项目权重之和为1；
- 使用所有学生(含不及格)的平均成绩计算。

表10. 课程目标达成度定量评价方法

课程目标	课程目标达成度评价方式
课程目标 1： 能够了解数据分析的基本思路、基本理论与基本方法，包括数据抽取、数据预处理、数据建模以及模型评价、数据分析方法、数据可视化等。	课程思政实践：0.2 课堂成绩：0.4 期末考查：0.4
课程目标 2： 能够应用现代信息技术工具，通过文献检索及研究，理解数据分析常用的技术。	课堂成绩：0.3 上机实验报告：0.3 期末考查：0.4

课程目标	课程目标达成度评价方式
课程目标 3： 培养实践应用能力，具备将所掌握的数据分析技术运用到解决实际问题的能力。	课堂成绩: 0.2 上机实验报告: 0.4 期末考查: 0.4
课程目标 4： 具备基本的科学素养，及时了解数据科学在国内外的实际应用案例，及时掌握国家数据科学方面的科技战略需求，树立强烈的爱国主义使命感与责任心。	课程思政实践: 0.2 课堂成绩: 0.2 上机实验报告: 0.3 期末考查: 0.3

十、说明

本课程大纲主要用于计算机科学与技术专业、智能计算与数据科学（计算机科学与技术专业）的《数据科学导论》课程的教学目标、教学内容、教学方法、教学要求以及考核评价方法等，承担该课程的教师必须遵照本大纲安排授课计划、实施教学过程，完成学生各个阶段与各方面的学习成果考核与评价；在学期末，需对课程目标和课程支撑的毕业要求指标点进行达成度评价。

本课程大纲于 2021 级开始执行，生效之日原先版本均不再使用。

十一、编制与审核

表 11 大纲编制与审核信息

工作内容	责任部门或机构	负责人	完成时间
编制	智能信息处理组	俞东进、孙笑笑	2022.03
审核	智能信息处理组	彭勇	2022.03
审定	计算机学院教学工作委员会	黄孝喜	2022.05