《数据科学与工程的数学基础》课程设计——探索与思考

黄定江
djhuang@dase.ecnu.edu.cn

華東師範大學

2018/10/14 中国.大连

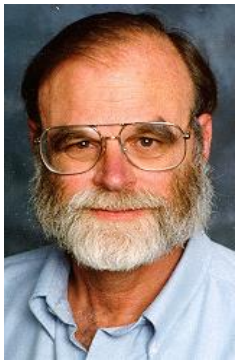
提纲

1. 课程设计的背景和动机
2. 国内外相似课程
3. 我们的课程规划

华东师范大学数据科学与工程学院
School of Data Science and Engineering at ECNU

1 课程设计的背景和动机

数据科学代表了未来科学发展的一种趋势



Jim Gray
图灵奖得主

T. H. Davenport
埃森哲战略变革研究院前主任

D.J. Patil
白宫首席数据官

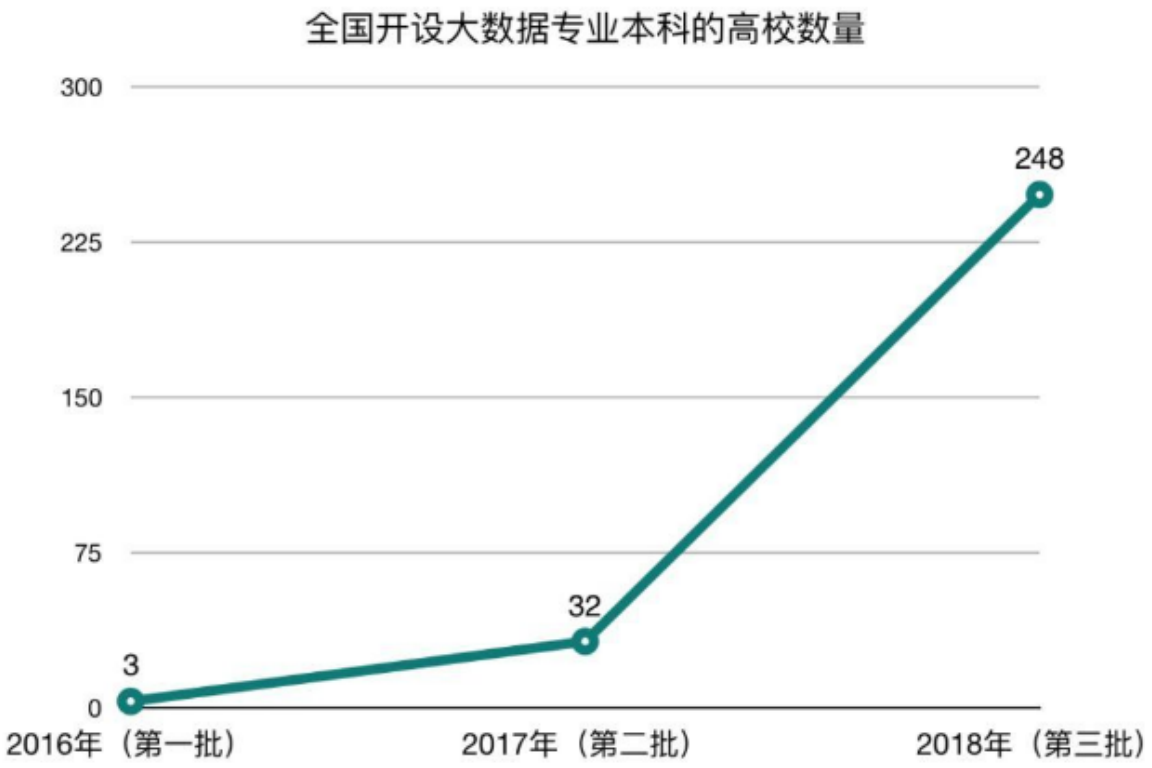
IEEE 计算机学会

第四范式：实验科学、理论科学、计算科学、**数据科学** 2007

21世纪最具吸引力的职业：**数据科学家**——《哈佛商业评论》，2012

未来九大科技趋势之一：**数据科学**——IEEE CS，2016

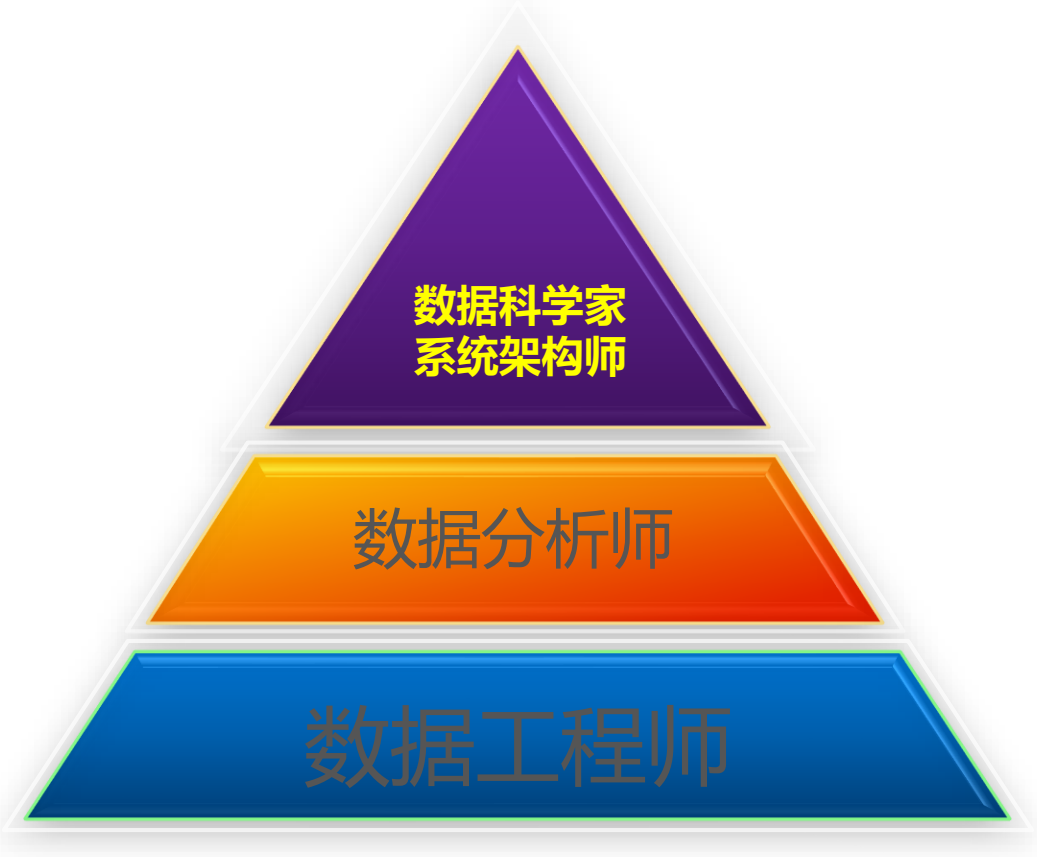
数据科学与大数据技术专业全国开设情况



数据来源大数据文摘

华东师范大学：2017年获批

专业培养的人才目标



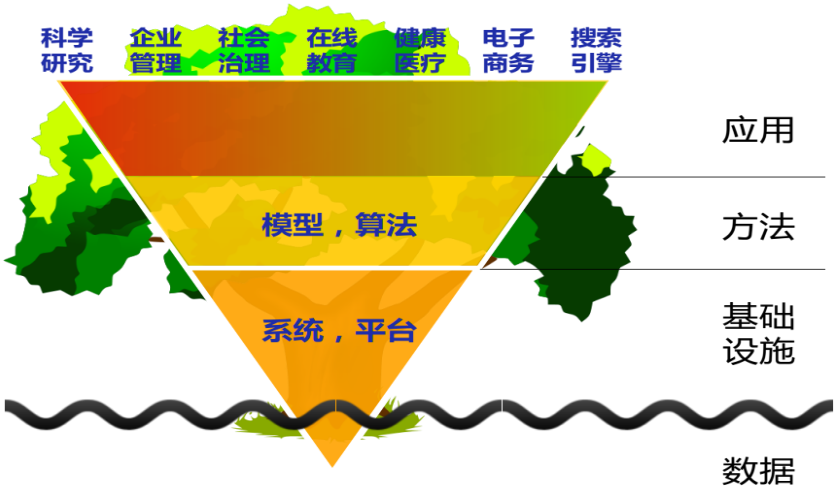
理论能力

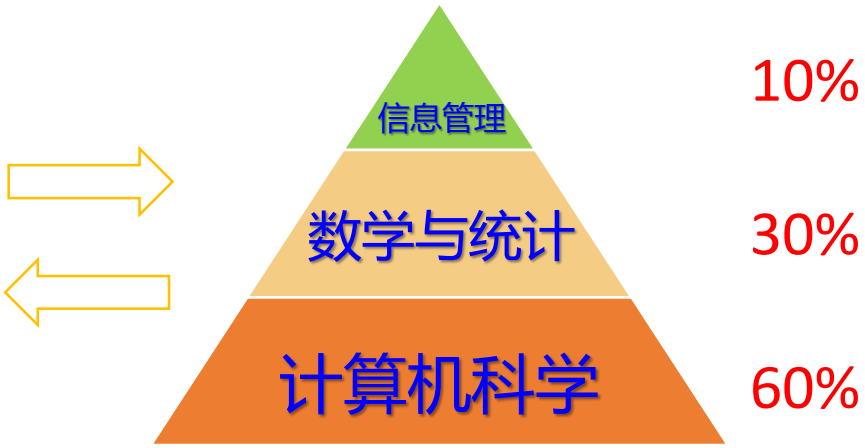
实践能力

应用能力

优秀的师资队伍和课程体系不容易建成

数据科学与工程是一个跨学科领域的交叉专业





应用驱动创新

协同创新

新的数据科学与工程课程体系和教材

思考：30%的数学与统计基础课程和知识占比其实是希望用来支撑数据科学与工程自身的理论规范的！！！

数据科学与工程专业需要什么样的数学课程体系？

什么是数据?

数据是信息的载体，是可以被计算机识别存储并加工处理的描述客观事物的信息符号的总称。

N	数据类型	
1	关系数据	结构化数据
2	时间序列	半结构化
3	图数据	半结构化
4	文本数据	非结构化
5	图片	非结构化
6	视频	非结构化
7	音频	非结构化

N	大数据特性	数量
1	高维	
2	海量	
3	多模	
4	高速	
5	噪声	
6	缺失	
7	非平衡	
8	稀疏	

大数据的结构是什么?

计算机科学处理问题的步骤

计算机科学是一门研究用数学和计算机进行数据表示和处理，并获得知识的学科。这里面涉及到两个问题：

- 数据的表示
- 数据的处理

计算机科学解决一个具体问题时，大致需要经过三个步骤：

- 从具体问题中抽象出一个适当的数学模型
- 设计一个解此数学模型的算法（Algorithm）
- 编出程序、进行测试、调整直至得到最终解答

数据结构

构建数学模型的实质是分析问题，从中提取操作的对象，并找出这些操作对象之间含有的关系，然后用数学的语言加以描述。

- 数值计算，所用的数学模型是用数学方程描述，所涉及的运算对象一般是简单的整形、实型和逻辑型数据，因此程序设计者的主要精力集中于程序设计技巧上，而不是数据的存储和组织上。

- 计算机科学应用的更多领域是“非数值型计算问题”，它们的数学模型无法用数学方程描述，而是用**数据结构**描述，解决此类问题的关键是设计出合适的**数据结构**

算法

- 与数据的结构密切相关，依附于具体的数据结构
- 运算是由计算机来完成，也就是说，数据结构还需要给出每种结构类型所定义的各种运算的算法

思考：这里的运算实际上是指任务！

经典的数据结构和计算机科学的数学体系

N	经典的数据结构	离散关系
1	逻辑结构	集合、线性、树形、图形（常用数据结构：数组、栈、队列、链表、树、图、堆）
2	物理结构	顺序、链接、索引、散列
3	运算结构（结构算法）	检索、插入、删除、更新和排序

数据结构：在同一类有限的数据集中，研究数据元素离散关系和数据运算

计算机科学是以“离散数学”为重点的数学体系

大数据结构和数据科学与工程的数学体系

N	数据特性	数据集
1	高维	无限
2	海量	无限
3	多模	多元
4	高速	快速增长
5	噪声、缺失、非平衡、稀疏	奇异性

大数据结构：在多元无限快速增长的数据集中，研究数据元素表示和数据运算

- 表示：相关关系表示（向量，矩阵，张量，拓扑空间、流形和李群）和随机表示
- 运算：分类、聚类、回归、降维和排序等

数据科学是以“矩阵、概率和优化”为重点的数学体系

庞大的内容体系

数据科学与工程涉及的数学基础知识，除了微积分、线性代数和概率论与数理统计这三大基础中的基础以及离散数学外，还涉及：

1. 矩阵分析：投影定理
2. 张量代数：张量分解
3. 优化理论：凸分析，凸优化模型和方法
4. 泛函分析：再生核希尔伯特空间，Mercer定理
5. 几何：解析几何、拓扑学、微分流形（嵌入定理）
6. 随机过程

《数学科学与工程的数学基础》

2 国内外相似课程

相似课程介绍 1

- **课程名称：** Topics in Mathematics of Data Science
- **学校：** Massachusettes Institute of Technology
- **面向对象：** Undergraduate
- **使用教材：** (讲义)
- Ten Lectures and Forty-Two Open Problems in the Mathematics of Data Science
- **先修课程：** 线性代数、概率论与统计、优化和算法的基础知识
- **课程主页：** https://ocw.mit.edu/courses/mathematics/18-s096-topics-in-mathematics-of-data-science-fall-2015/index.htm

Topics in Mathematics of Data Science

- **课程介绍：**

  - 以一系列未解决的问题为教材的切入点，引入对数学工具的讨论
  - 独立的研究型课程，专为本科学生设计，针对从数据中提取信息的各类算法理论的学习，讲义解决以下问题：
    - 主成分分析和一些随机矩阵理论、流形学习和扩散图
    - 谱聚类及其性能保证、数据集中度及尾界
    - 降维、压缩感知/稀疏恢复，矩阵补全等
    - 近似算法和Max-Cut问题
    - 随机图上的聚类、图中的同步
    - 反问题以及紧致群上成对比率的未知变量估计

- **课程特点：**

- 以实际问题为先导来介绍各个独立的数学工具，没有成体系化。

教学大纲

- **课程涵盖以下内容：**
- 主成分分析（PCA）和一些随机矩阵理论，用于通过spike模型了解高维PCA的性能。
- 流形学习和扩散图：非线性降维工具，替代PCA。半监督学习及其与Sobolev嵌入定理的关系。
- 光谱聚类及其性能保证：Cheeger的不等式。
- 标量变量和矩阵变量的测量浓度和概率尾界。
- 通过Johnson-Lindenstrauss Lemma和Gordon通过网格定理逃逸来减少维数。
- 压缩感知/稀疏恢复，矩阵完成等。如果时间允许，将介绍数论理论的测量矩阵构造。
- 小组测试。使用组合工具来确定测试程序的下限，如果有时间，可能会给出纠错码的速成课程，并在组测试中展示它们的使用。
- 理论计算机科学中的近似算法和Max-Cut问题。
- 随机图上的聚类：随机块模型。优化中的二元性基础。
- 图中的同步，反问题以及紧致群上成对比率的未知变量估计。

相似课程介绍 2

- **课程名称：** Mathematical foundations of data sciences
- **学校：** ENS Paris
- **面向对象：** 数学与应用硕士
- **使用教材：**

  Mathematical foundations of data sciences
- **先修课程：** 分析、代数、概率、优化的本科基础课程
- **课程主页：** http://math.ens-paris-saclay.fr/version-francaise/formations/master-mva/contenus-/mathematical-foundations-of-data-sciences-214256.kjsp?RH=1242430202531

Math. Foundations of Data Sci

- **课程介绍：**

- 本书概述了现代数据科学的重要数学和数值工具：
  - 信号和图像处理的基础知识（傅里叶、小波及其在去噪和压缩中的应用）
  - 成像科学（反问题、稀疏性、压缩感知）
  - 机器学习（线性回归、逻辑分类、深度学习）
- 重点是数学方法论的方法论工具，特别是：
  - 线性算子、非线性近似
  - 凸优化、最优传输，以及如何将它们映射到有效的计算算法

- **课程特点：**

- 建立在学生已有较为全面的本科数学基础之上，涵盖现代数据科学的主流工具，为研究信号、图像、机器学习的研究生提供指导。

相似课程介绍 3

- 课程名称：Mathematical foundations of data sciences
- 学校：Michigan State University
- 面向对象：Graduate
- 使用教材：(讲义)
- Math. Foundations of Data Sci
- 课程主页：https://matthewhirn.wordpress.com/teaching/spring-2017-cmse-820/

Mathematical foundations of data sciences

口课程介绍：

- 本课程将介绍数据科学中遇到的一些主要问题和挑战，介绍数据科学的基本数学原理，这些原理是算法，过程，方法和以数据为中心的思想的基础，内容涵盖PCA，核方法，正则化，统计，谱图理论，流形学习等。

Mathematical foundations of data sciences

Michigan State University	Math. Foundations of Data Sci	讲义介绍	
涉及知识点	相应的本科课程	难度	应用
内积	线性代数	简单	二分类
PCA	多元统计分析	中等	面部特征提取
Marchenko-Pastur distribution	一般无此内容介绍	难简化处理	
Hilbert空间，Banach空间	泛函分析	中等	核方法
正定	线性代数	中等	核方法
等价关系	离散数学	简单	核方法
正则化	一般无此内容介绍	中等	岭回归
表示理论	泛函分析	中等	
偏差方差分解	统计推断	简单	误差分析
图与矩阵	图论/离散数学	中等	谱图聚类
K-均值聚类	多元统计分析	简单	聚类
图拉普拉斯	一般无此内容介绍	困难	谱图聚类
流形	微分几何	困难	流形学习

相似课程介绍 4

- 教材名称： Foundations of Data Science
- 使用学校： Indiana University/ University of California, Santa Barbara/ Freie Universität Berlin/ University of Illinois - Chicago
- 面向对象：Undergraduate / Graduate
- 作者：
  - Avrim Blum　　ACM Fellow
  - Ravindran Kannan　　ACM Fellow
  - John Hopcroft　　ACM Fellow, Turing Award

Foundations of Data Science

内容介绍

- 本书目标涵盖预计在未来40年内有用的理论，强调概率统计与数值方法，包括高维空间线性代数与几何，奇异值分解，随机游走，机器学习，VC维，感知机，随机梯度下降，正则化，采样，聚类，矩阵分解，稀疏表示等。

教材特色

- 强调概率统计与数值方法，强调智能思想和数学基础，而不拘泥于特定应用。

相似课程介绍 5

- 教材名称： Mathematics for Machine Learning
- 面向对象：Undergraduate
- 作者：Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong
- 先讲数学基础，后讲应用

相似课程介绍 5

内容介绍：

- Part I: Mathematical Foundations
  - Introduction and Motivation
  - Linear Algebra
  - Analytic Geometry
  - Matrix Decompositions
  - Vector Calculus
  - Probability and Distribution
  - Continuous Optimization

- Part II: Central Machine Learning Problems
  - When Models Meet Data
  - Linear Regression
  - Dimensionality Reduction with Principal Component Analysis
  - Density Estimation with Gaussian Mixture Models
  - Classification with Support Vector Machines

3 我们的课程设计

课程设计理念一

- 围绕理解大数据结构这一目标：在多元无限快速增长的数据集中，研究数据元素的表示和数据运算
  - 表示：空间表示和随机表示
  - 运算：分类、聚类、回归、排序和降维等

课程设计理念二

- 依托现有的工科数学课程设置，不做大范围全新的设计
  - 高等数学
  - 线性代数
  - 概率论
  - 离散数学

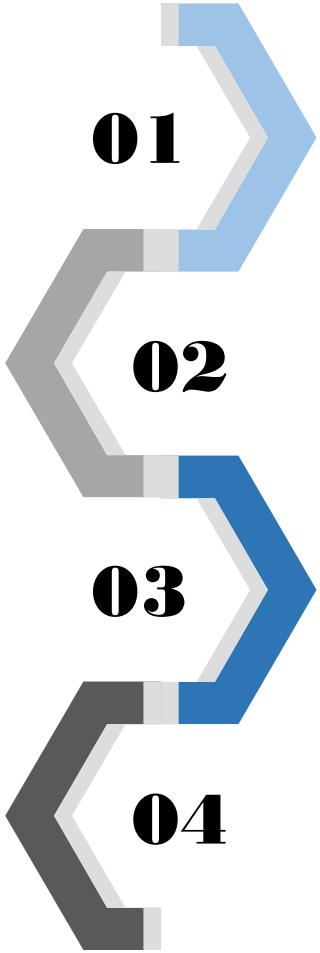
课程设计理念三

- 按照本科生和研究生两个阶段来设计

课程目标

夯实数学基础

- 连续数学
- 几何
- 代数



培养思维能力

- 抽象思维
- 连续方法

掌握应用数学的能力

- 发现问题
- 分析问题
- 建立模型

紧密对接后续课程

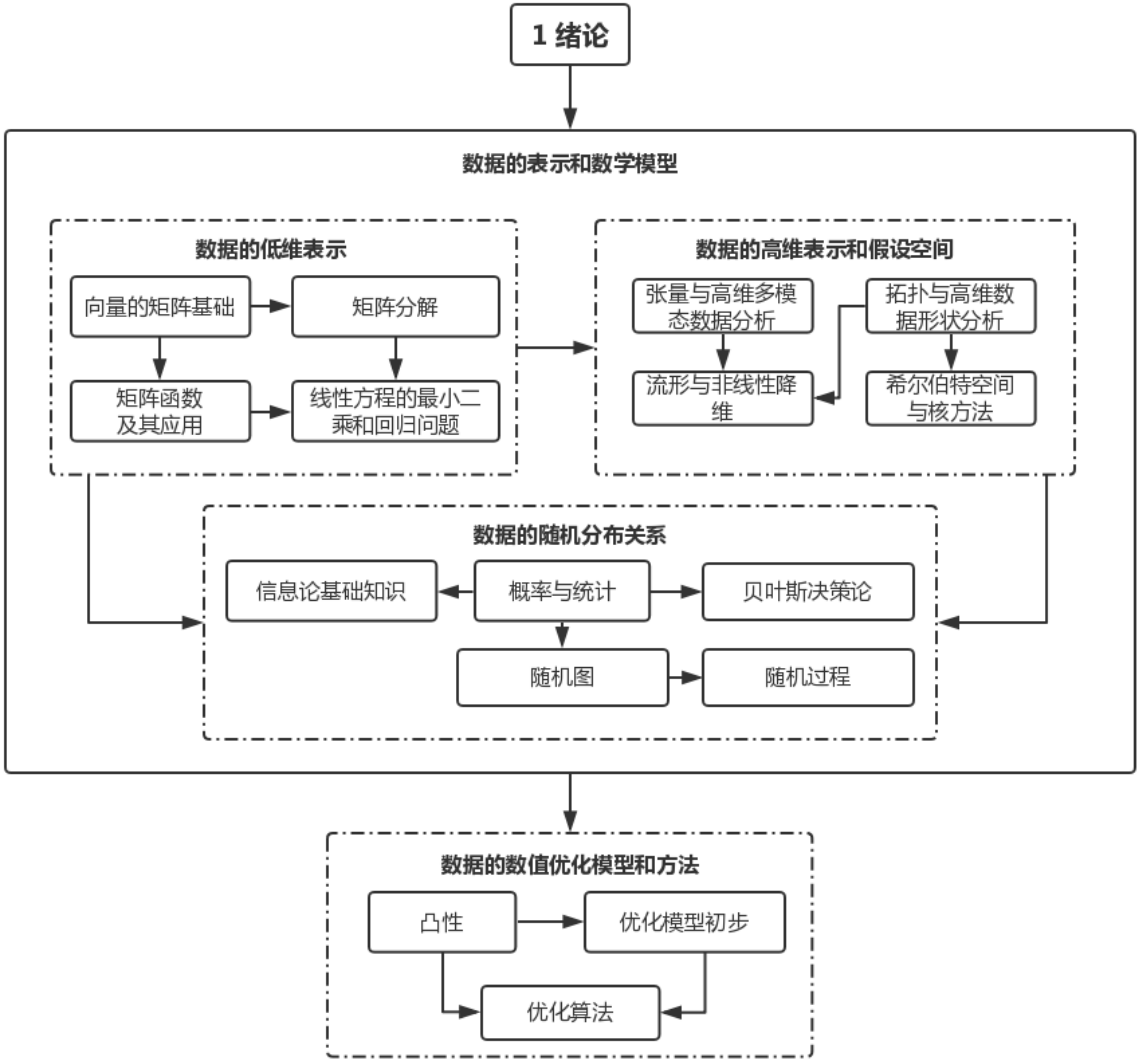
- 算法基础
- 机器学习

课程特点

- 内容：面广，强调数学基础的全面性和系统性，突出让学生了解数据科学所涉及的数学基础理论全貌，为学生后续系统学习各专业模块课程打下基础

- 强调数学基础的实用性
  - 数据科学实例驱动、贯穿全书

- 难度：不会太难，但又不会太浅显，内容循序渐进


1 绪论
数据的表示和数学模型
数据的低维表示
向量的矩阵基础
矩阵分解
矩阵函数及其应用
线性方程的最小二乘和回归问题
数据的高维表示和假设空间
张量与高维多模态数据分析
拓扑与高维数据形状分析
流形与非线性降维
希尔伯特空间与核方法
数据的随机分布关系
信息论基础知识
概率与统计
贝叶斯决策论
随机图
随机过程
数据的数值优化模型和方法
凸性
优化模型初步
优化算法

课程内容的总体框架

谢谢，请批评指正！