数据科学与大数据技术专业人才培养

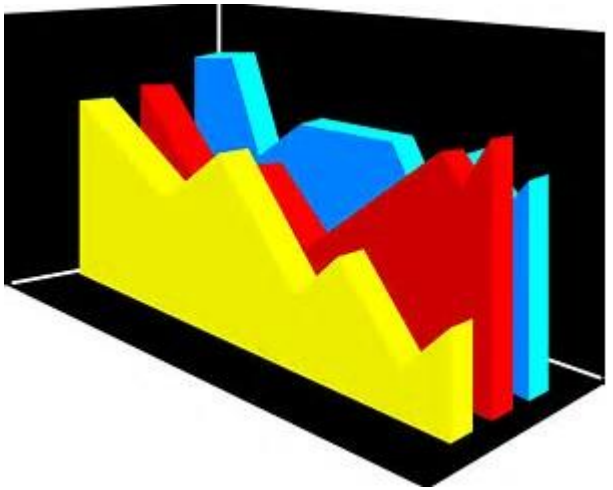
哈尔滨工业大学的实践

哈尔滨工业大学
王宏志
wangzh@hit.edu.cn
http://homepage.hit.edu.cn/pages/wang


BIG
DATA
USE
ONE
SINCE
SOFTWARE
UBIQUITOUS
REQUIRING
BUSINESS
MOVING
ANALYTICS
RADIO-FREQUENCY
COMPLEXITY
SAN
SOLID
WIRELESS
TOLERABLE
PARALLEL
SIZE
NEEDED
QUALITIES
PETABYTES
INTERNET
PROCESSING
BURIED
LOGS
TECHNOLOGIES
USED
DISTRIBUTED
CAPTURE
MAY
MANAGEMENT
MANAGE
GROW
SOCIAL
LARGE
STORE
DEFINING
CASE
EVERY
LARGER
STORAGE
RESEARCH
MPP
TERABYTES
CONTINUES
COST
CITATION
COMPLEX
DATABASES
INDEXING
DIFFICULTY
TARGET
ABILITY
SETS
EXAMPLES
APPLIED
AMOUNT
DESCRIBING
ELAPSED
CURRENT
THOUGHT
SENSOR
ARCHIVES
CONNECTOMICS
RECORDS
INFORMATION
RECONSIDER
OPPORTUNITIES
SEARCH
DEFINITION
BIOGEOCHEMICAL
NETWORKS
PROCESS
WITHIN
INCLUDE
SYSTEMS
COMBAT
SIGNIFICANT
TIME
RELATIONAL
TOOLS
SET
COMPUTING
SHARED
ZETTABYTES
GENOMICS
DISK
PERFORMANCE
ALSO
RELATED
HUNDREDS
CREATED
BIOLOGICAL
GARTNER
MASSIVELY
USING
TYPES
WORKING
DESKTOP
CURRENTLY
FC
WORLD'S
PRESENTATIONS
FORMS
CAPACITY
TENS
NOW
PRACTITIONERS
ORGANIZATIONS

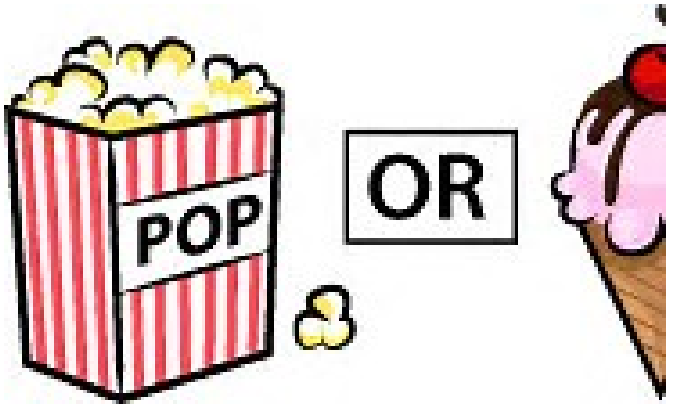
“数据科学与大数据技术”需要的新思路

系统



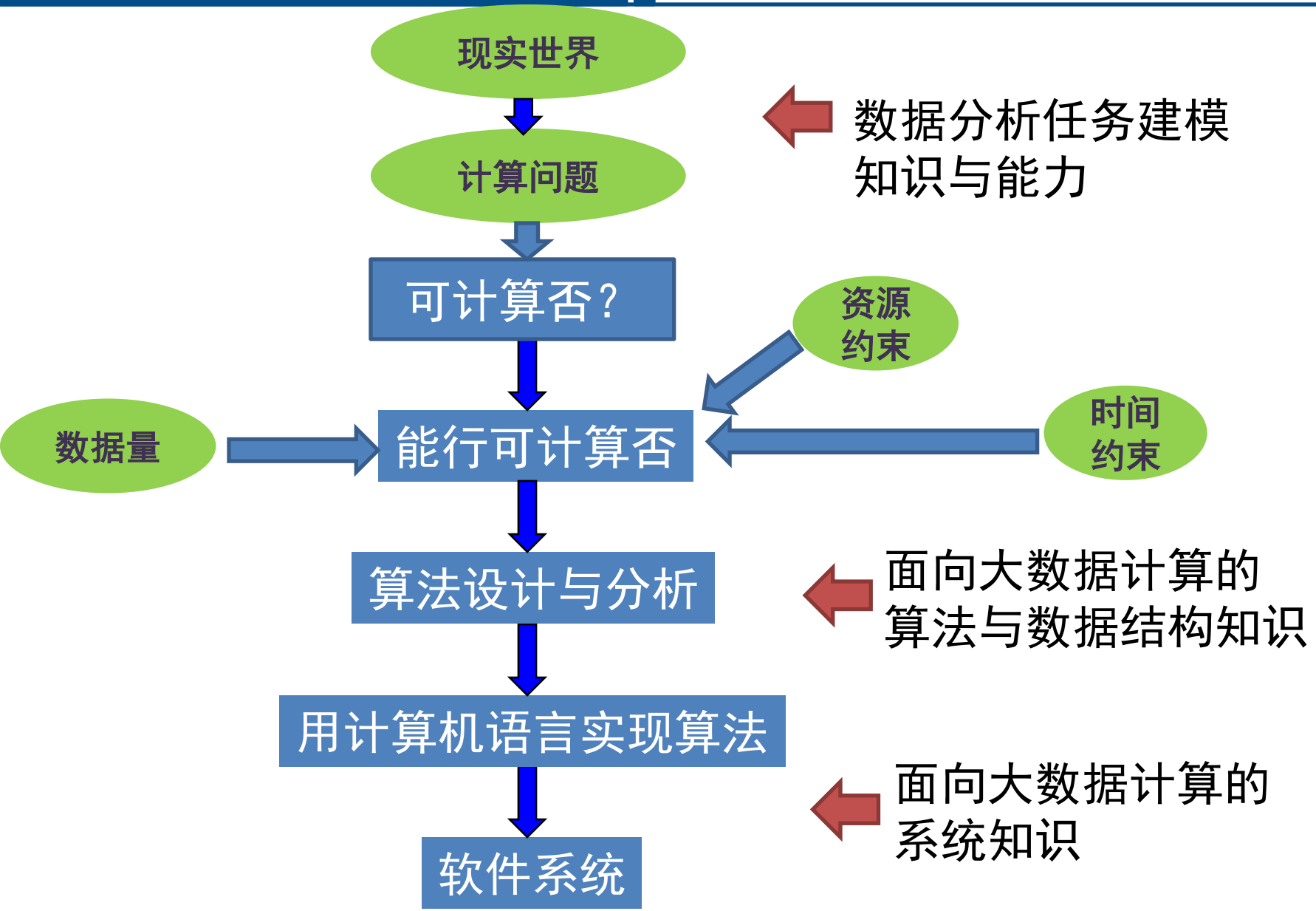
建模

实现

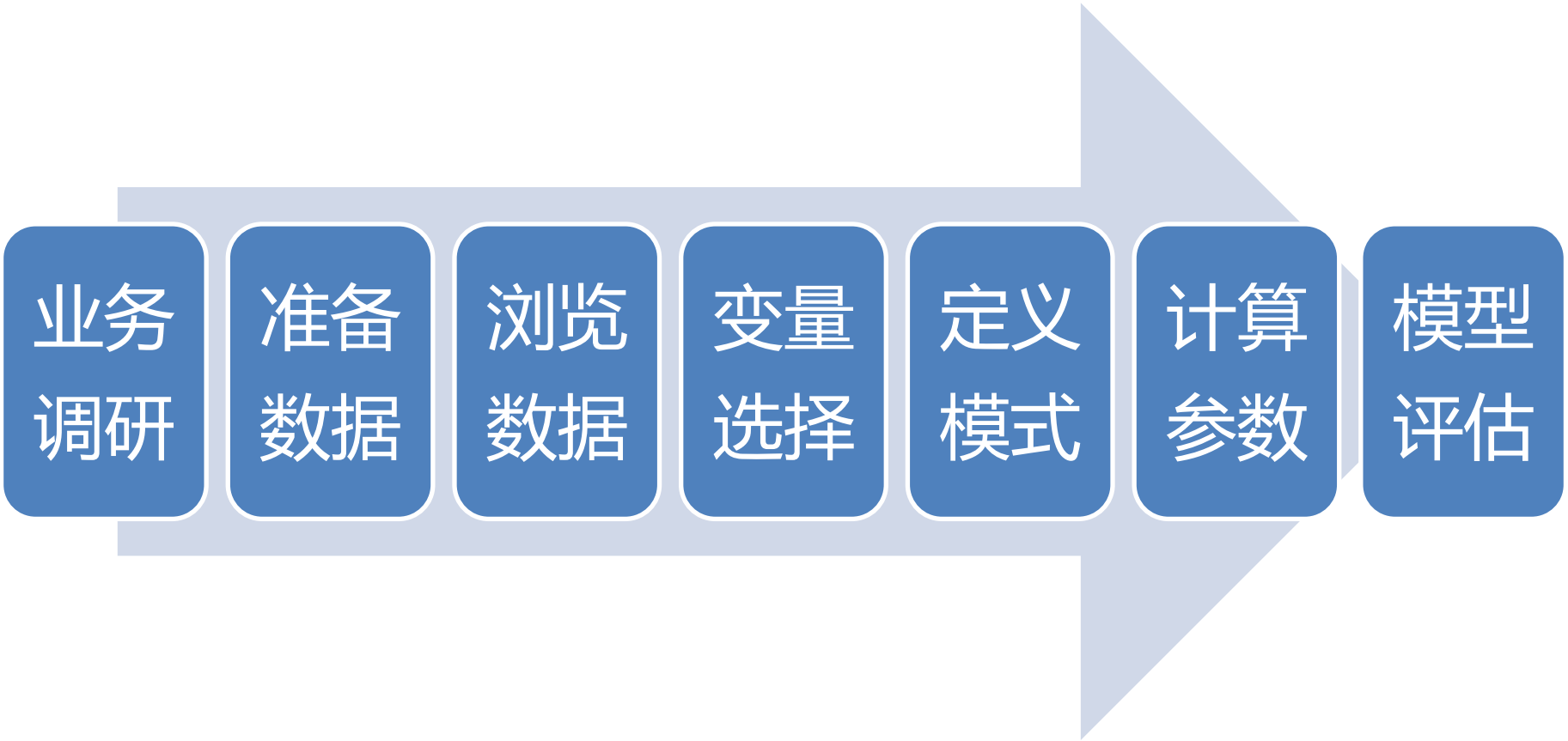


折衷

求解大数据计算问题的过程



“现实”到“模型”



数据驱动业务的逻辑思维能力

“模型”到“算法”

- 资源限制
  - 内存不足
    - 内存算法
    - 空间亚线性算法
  - 处理器计算能力不足
    - 并行算法
- 实时性要求
  - 问题计算复杂度下界难以满足要求
    - 时间亚线性算法

面向大数据计算的算法知识

“好算法”与“好系统”

- 适合的大数据计算软硬件平台
- 设计高效的大数据存取结构
  - 数据存储结构
  - 数据分布策略
  - 数据索引方法
- 编写适用于大数据的“好程序”
  - 避免使用系统垃圾回收机制
  - 减少内存拷贝
  - 减少数据重分布次数
  - 减小重分布数据量

✓ 面向大数据计算的分布式系统知识
✓ 面向大数据计算的数据管理知识
✓ 面向大数据计算的程序设计知识

专业区分度

- 独特的学科基础和内涵
  - 大数据表达理论、大数据计算理论与技术、大数据应用基础理论
  - 不同于计算机科学与技术、软件工程等学科

    不仅需要掌握计算方法和工具，还需要认知数据本身的现象和规律、数据管理和处理的基础理论、全生命周期的数据管理方法和系统等专门知识
  - 不同于商业智能和统计学

    不仅包含统计和商业智能的方法和模型，还包括算法设计与分析和计算系统的设计、研发、运维、评测、优化、应用等
- 独特的专业课程设置
  - 突出数据科学基础课程教学
  - 裁剪传统计算机统类课程
  - 强调大数据管理与处理的全生命周期
  - 充分结合行业，突出实用性
- 独特的能力要求
  - 数据密集型计算系统的设计、构建、运维及应用的能力
  - 数据密集型计算平台的开发及应用能力
  - 大数据理论、系统及应用的创新能力
  - 面向数据密集型问题，将现实问题抽象为数据计算模型的能力
  - 建立由多源异构数据到全面智能应用的建模及求解算法能力

培养目标

- 力求培养在教育/研究/工业/政务/社会服务等领域，能够引领社会发展的未来领军型“数据科学家”与“大数据计算系统架构师”
- 要求
  - 具有正确的世界观、人生观与价值观
  - 深刻理解数据的获取、建模、管理、利用的全生命周期，熟知相关技术、系统和应用的前沿动态
  - 具有创新精神和国际视野
  - 掌握数据处理和管理的基础理论，具备深度数据分析和数据挖掘技能
  - 能够综合运用所学知识，独立解决与大数据计算中相关的科学研究和复杂工程技术问题
  - 具有跨学科能力、团队合作能力和有效的交流能力

培养要求

- 基本素质

(1) 社会素质 (2) 人文素质 (3) 身心素质 (4) 研究素质 (5) 工程素质 (6) 个性素质 (7) 领袖素质

- 基本能力

(1) 数据科学思维能力 (2) 数据密集型算法设计与分析能力 (3) 程序设计与实现能力 (4) 系统应用能力 (5) 数据密集型计算系统设计与实现能力 (6) 系统分析与评价能力 (7) 组织、协调与项目管理能力 (8) 表达与沟通能力 (9) 英语理解与交流能力 (10) 自学、独立思考与创新能力

- 必要知识：

(1) 数学与自然科学基础 (2) 人文社会科学类知识 (3) 数据科学与大数据技术专业基础知识涵盖计算机科学的基础知识 (4) 专业核心知识覆盖大数据计算系统、大数据算法设计与分析、大数据分析、机器学习与数据挖掘、大数据获取与清洗、大数据管理 (5) 工程实践和毕业设计。专业核心课程的学习需包含较强工程实践内容。完成毕业设计

数据科学与大数据技术专业课程设置思路

- 依托现有计算机科学与技术等优势专业，建设具有特色的数据科学与大数据技术专业
  - 采用与现有专业一致的核心课程和专业课程，充分发挥现有专业的教学资源和师资优势。
  - 教学内容与培养方式向数据科学与大数据技术方向延伸扩展
  - 重点培养学生的数据建模、数据管理、数据分析与数据挖掘等能力，形成专业优势与特色
- 设计三门大学分课程(72[48+16]+72[48+16]+48[32+16])
  - 大数据计算基础、大数据分析、数据挖掘
- 引入计算机学科交叉课程
  - 自然语言处理、信息检索、随机算法…
- 建设一系列选修课
  - Web技术、工业大数据、时空数据库.

数据科学与大数据技术辅修专业课程设置

数据科学与大数据技术辅修专业（学位）第一学年教学进程表

开课学期	课程编号	课程名称	学分	学时分配						考核方式	备注
				学时	讲课	实验	上机	习题	课外辅导		
秋季	CS33271M	计算机数学基础	3.0	48	40	8				考试	
	合计		3.0	48	40	8					
春季	CS33272M	计算机系统基础	3.0	48	40	8				考试	
	合计		3.0	48	40	8					

数据科学与大数据技术辅修专业（学位）第二学年教学进程表

开课学期	课程编号	课程名称	学分	学时分配						考核方式	备注
				学时	讲课	实验	上机	习题	课外辅导		
秋季	CS33273M	计算机算法基础	3.0	48	40	8				考试	
	合计		3.0	48	40	8					
春季	CS33274M	大数据算法	3.0	48	40	8				考试	
	CS33275M	大数据系统	3.0	48	40	8				考试	
	合计		6.0	96	80	16					

数据科学与大数据技术辅修专业（学位）第三学年教学进程表

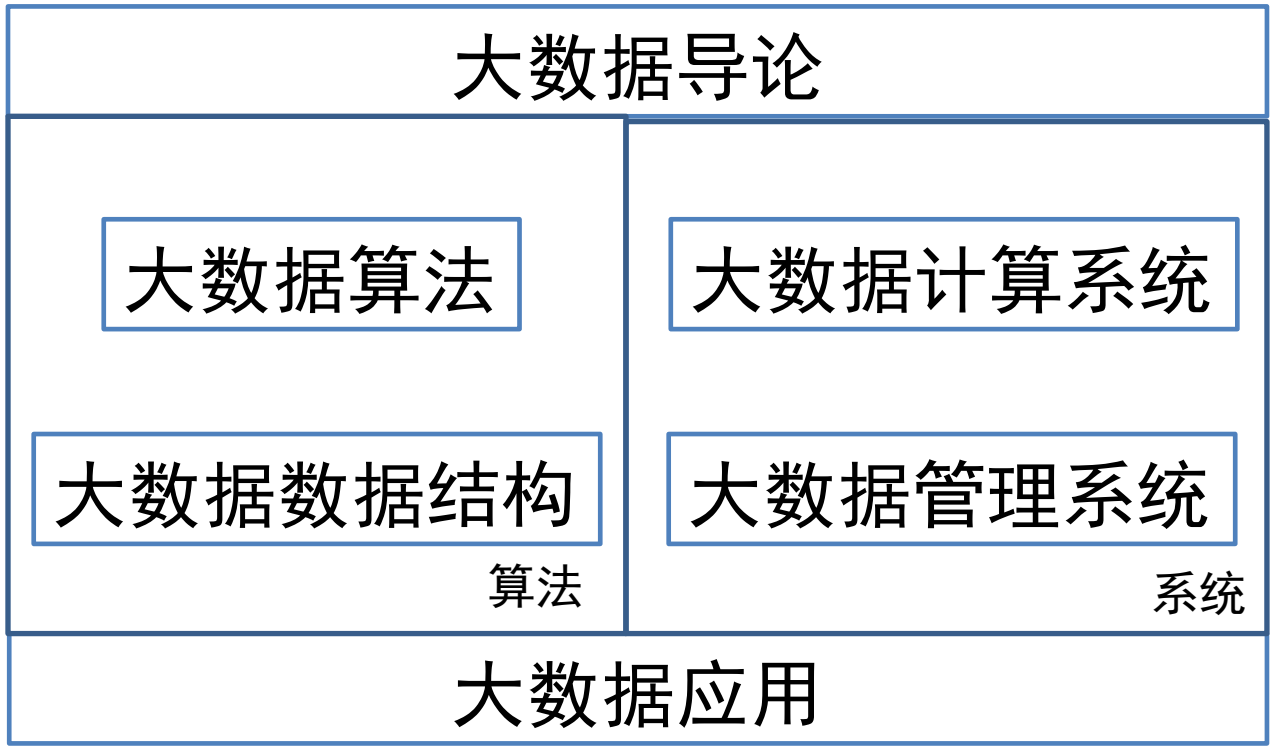
开课学期	课程编号	课程名称	学分	学时分配						考核方式	备注
				学时	讲课	实验	上机	习题	课外辅导		
秋季	CS33276M	大数据智能	2.0	32	32	0				考试	
	CS33277M	大数据分析课程设计	3.0	48	0	48				口试	
	小计		5.0	80	32	48					
春季	CS34999M	毕业设计（论文）	10	10 周							
			10.0	10 周							

前置课程要求

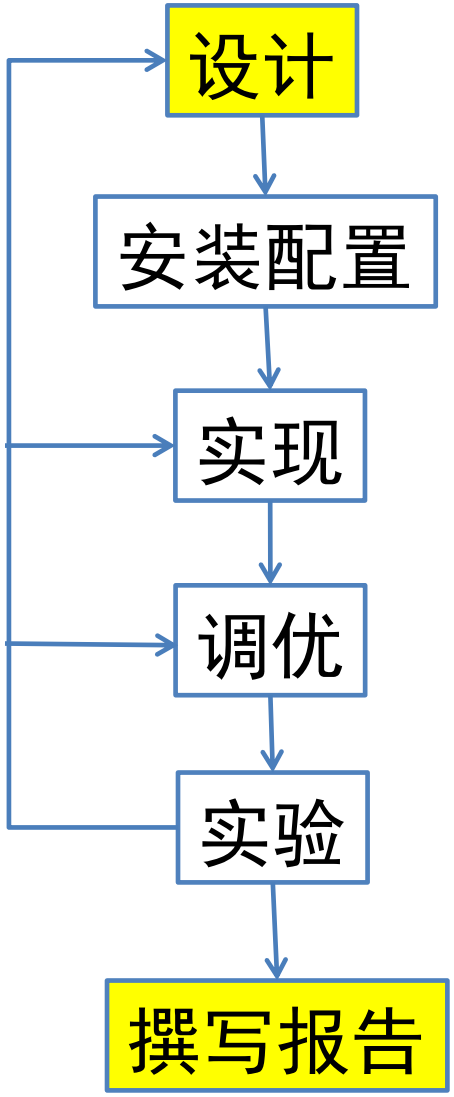
- “大学计算机-计算思维导论”、“高级语言程序设计”课程在辅修之前需完成学习并取得相应学分（该学分不含在辅修学分要求内）。
- 在修读“大数据算法”、“大数据智能”、“大数据分析课程设计”课程前，学生需先完成 “概率论与数理统计”、“线性代数”、“高等数学”课程的学习并获得学分。

“大数据计算基础”脉络图

课堂授课



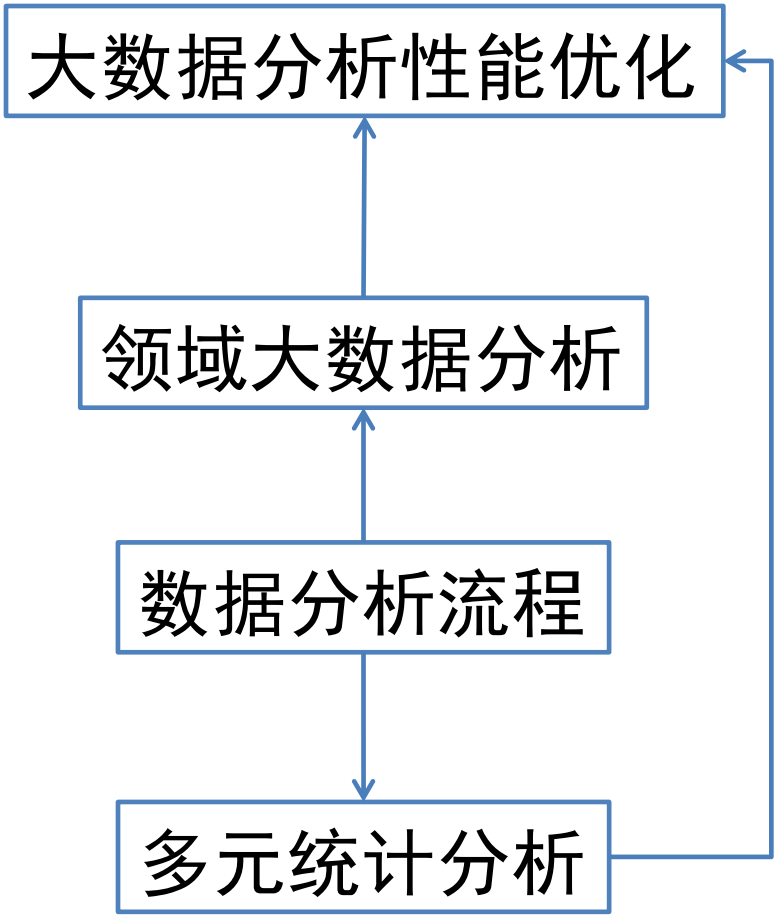
实验与作业



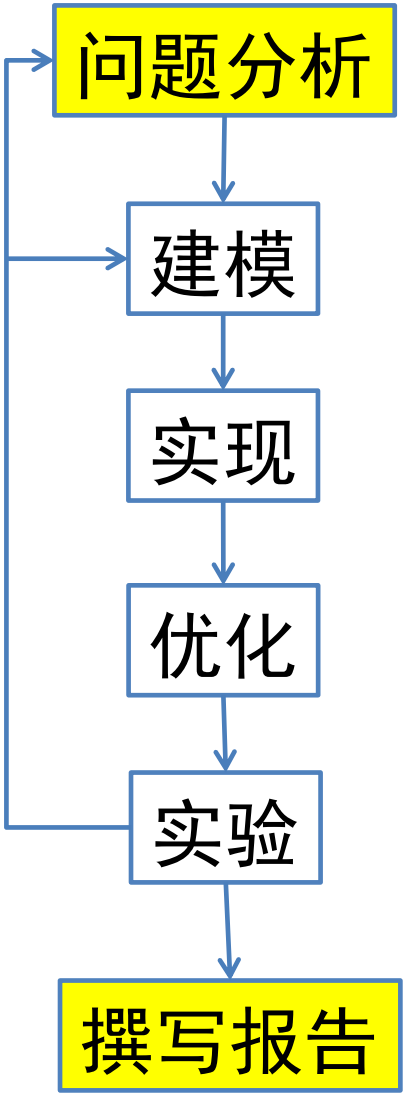
“大数据分析”脉络图

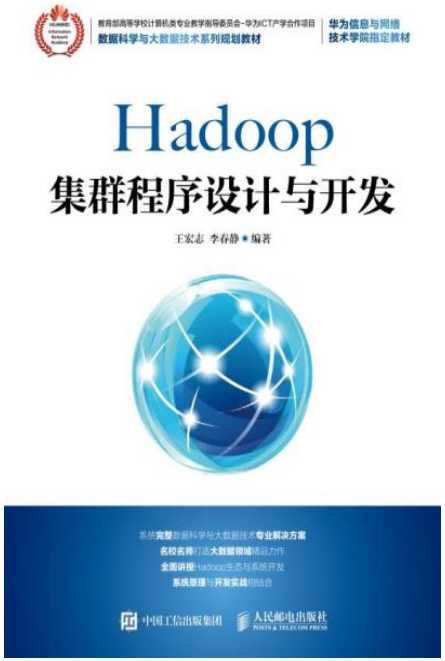
大数据分析

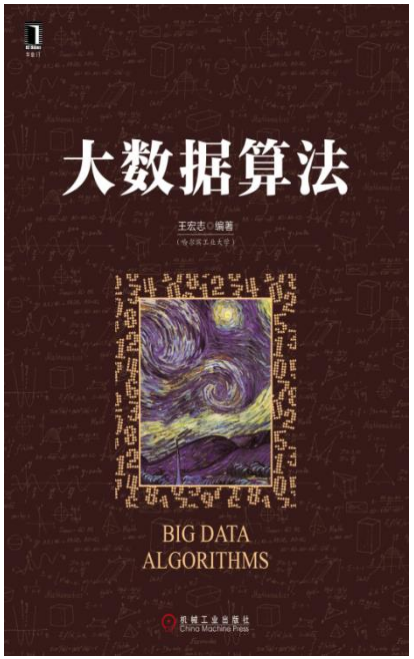
课堂授课

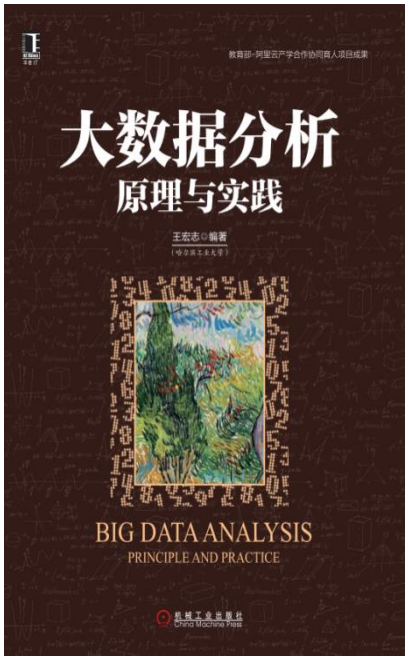


实验与作业









谢谢！

Thanks for your attention!

报告人：王宏志
wangzh@hit.edu.cn
http://homepage.hit.edu.cn/wang