算法设计与分析

第十章 算法杂谈

丁小欧
dingxiaoou@hit.edu.cn

本章内容

10.1近似算法

10.2图论算法

10.3大数据算法

10.4总结

近似算法

- 应用：通信网络、工业工程、生物信息等领域
- 处理难解的组合优化问题的一个重要且有效的方法
- 对于难解问题：能得到的最好方案就是设计接近最优解的近似算法
- 在多项式时间内求得一个解，并使目标函数与最优解的目标函数值之间的比不超过一个常数

近似算法

- 近似算法内在理论发展+外部实际应用驱动——近似算法设计
- 对近似算法的评价：
  - 主要标准是运行时间
  - 一个同样重要的尺度——“性能比”度量得到的解与最优解的接近程度
  - 关键问题：
    - 在采取某种算法策略时，一个具体的近似算法性能比是多少？
    - 求解该优化问题的任意一个多项式时间近似算法最好性能比是多少？
    - 是否存在求解该问题的多项式时间近似方案？

本章内容

10.1近似算法

10.2图论算法

10.3大数据算法

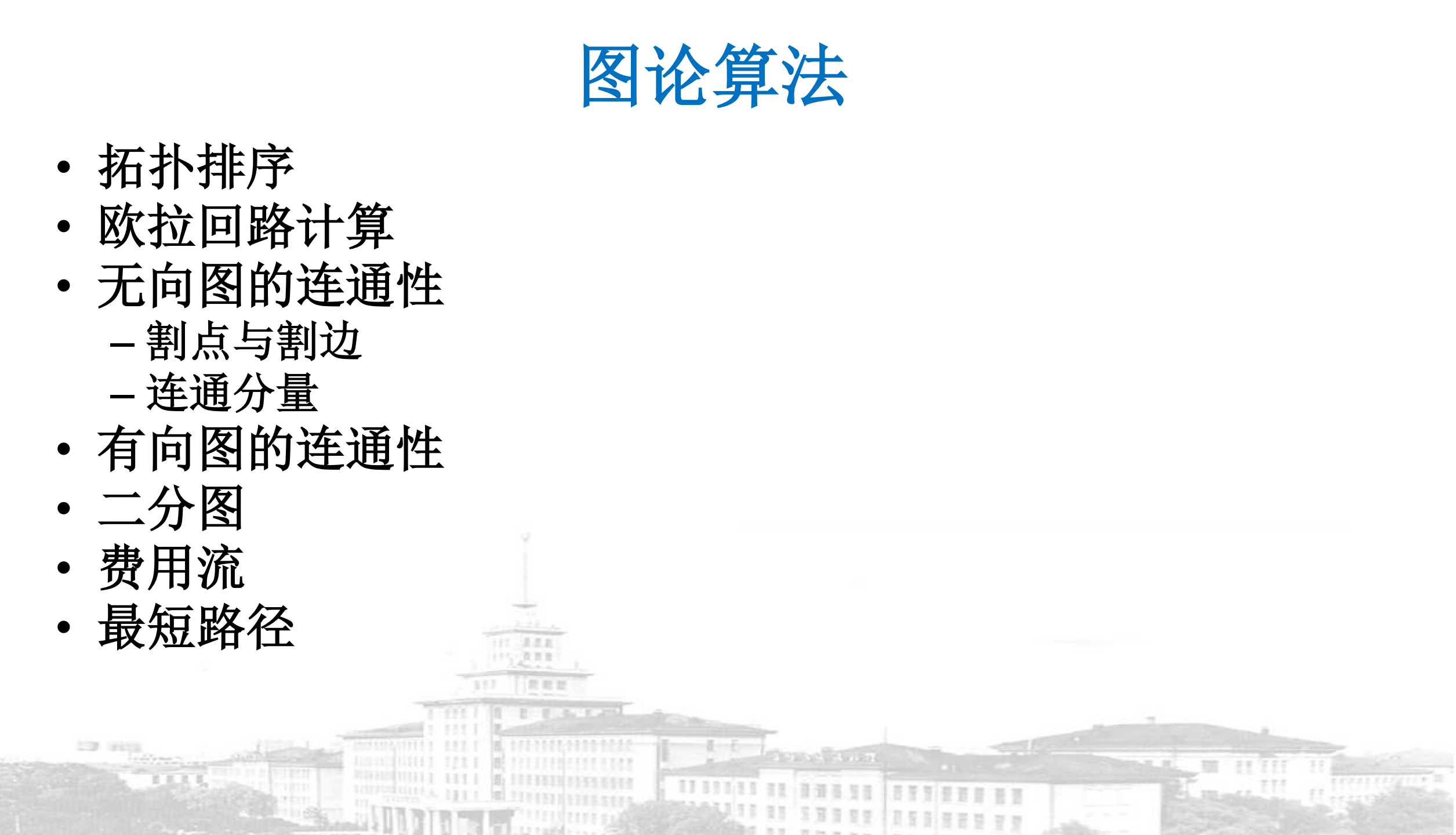
10.4总结

图论算法

- 图的存储
  - 邻接矩阵
  - 邻接表
  - …

图论算法

- 拓扑排序
- 欧拉回路计算
- 无向图的连通性
  - 割点与割边
  - 连通分量
- 有向图的连通性
- 二分图
- 费用流
- 最短路径



本章内容

10.1近似算法

10.2图论算法

10.3大数据算法

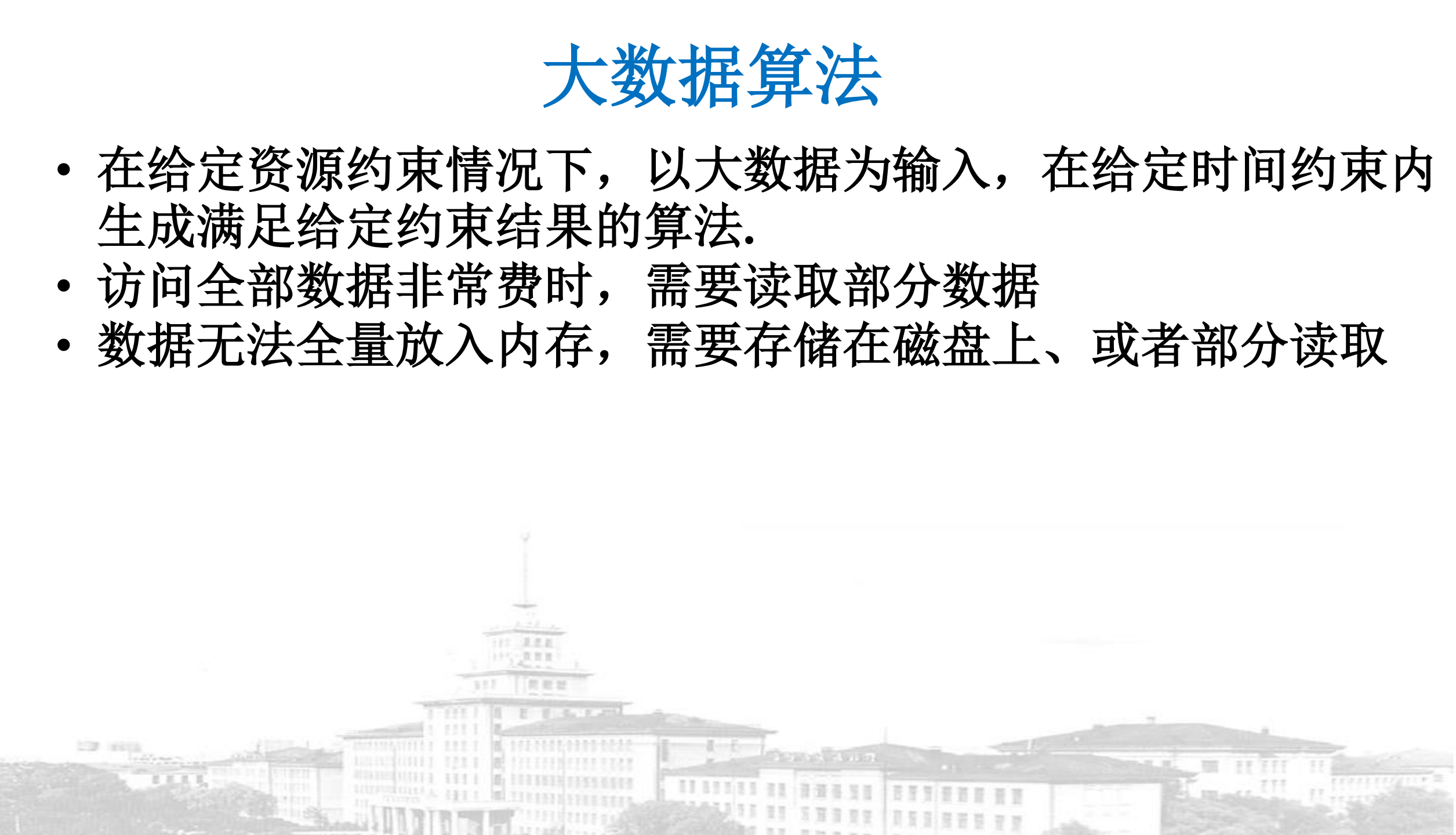
10.4总结

大数据算法

- 社交网大数据
- 天文大数据
- 工业大数据
- 能源大数据
- ……

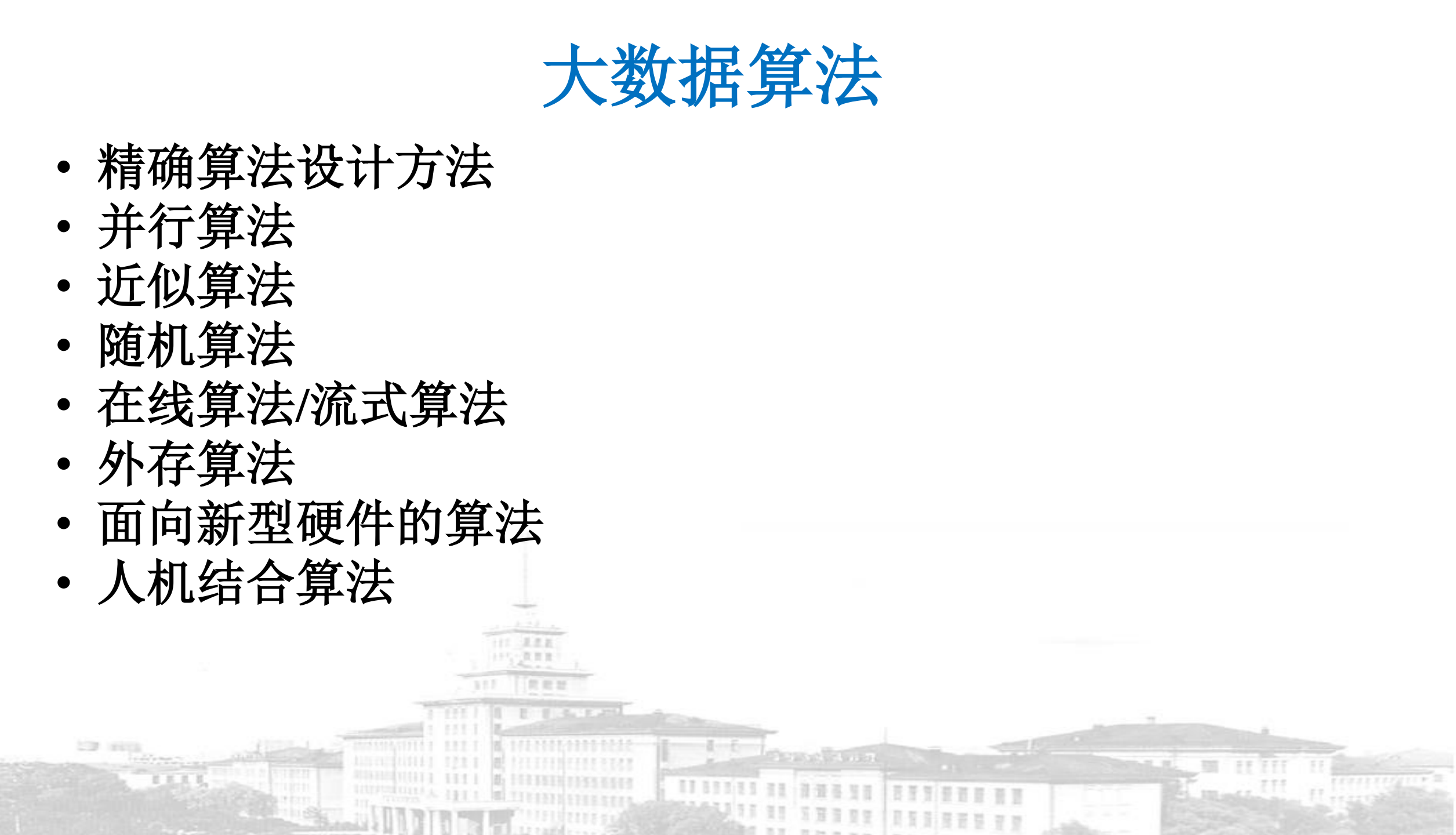
大数据算法

- 在给定资源约束情况下，以大数据为输入，在给定时间约束内生成满足给定约束结果的算法.
- 访问全部数据非常费时，需要读取部分数据
- 数据无法全量放入内存，需要存储在磁盘上、或者部分读取



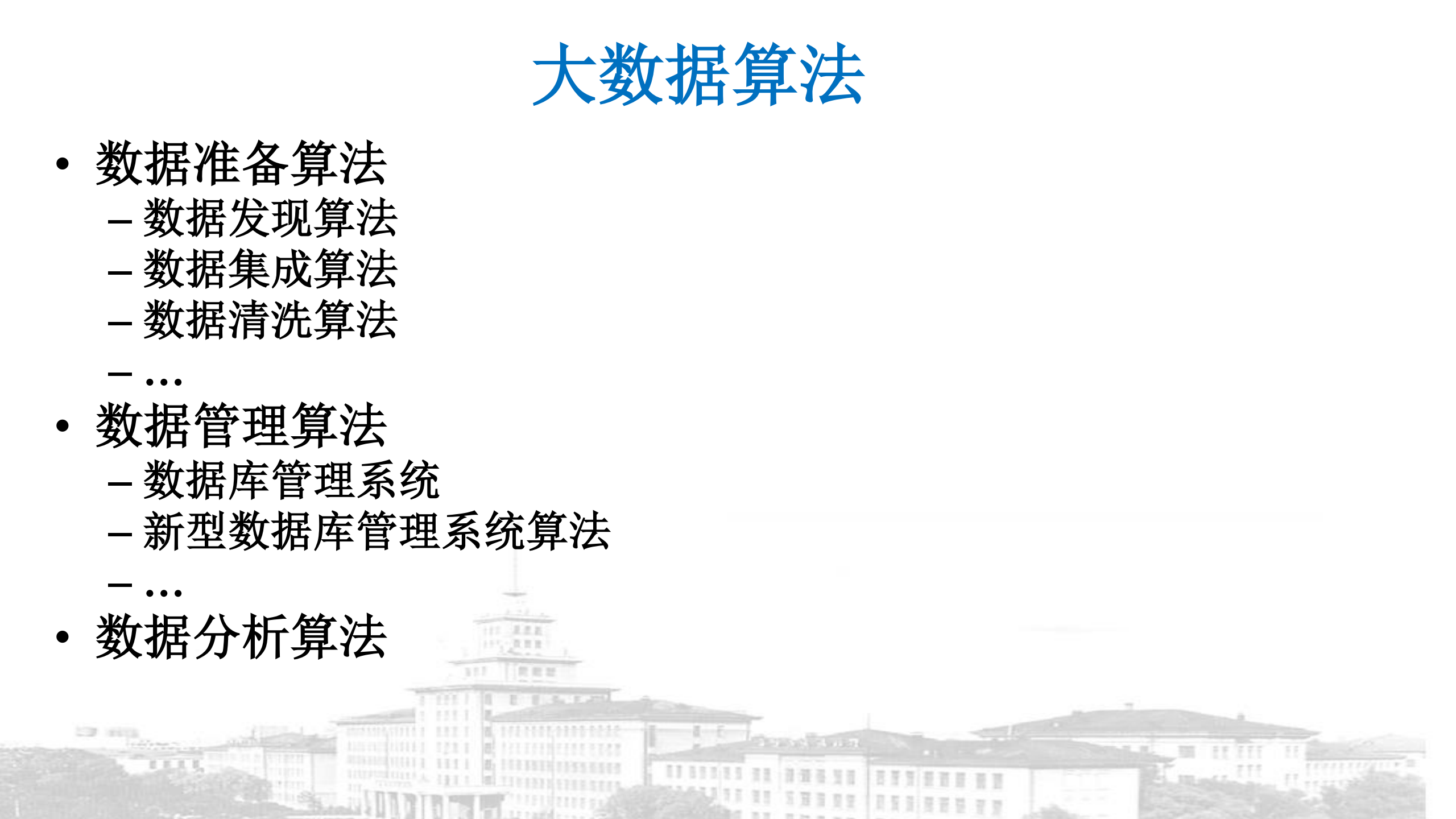
大数据算法

- 精确算法设计方法
- 并行算法
- 近似算法
- 随机算法
- 在线算法/流式算法
- 外存算法
- 面向新型硬件的算法
- 人机结合算法



大数据算法

- 数据准备算法
  - 数据发现算法
  - 数据集成算法
  - 数据清洗算法
  - …
- 数据管理算法
  - 数据库管理系统
  - 新型数据库管理系统算法
  - …
- 数据分析算法



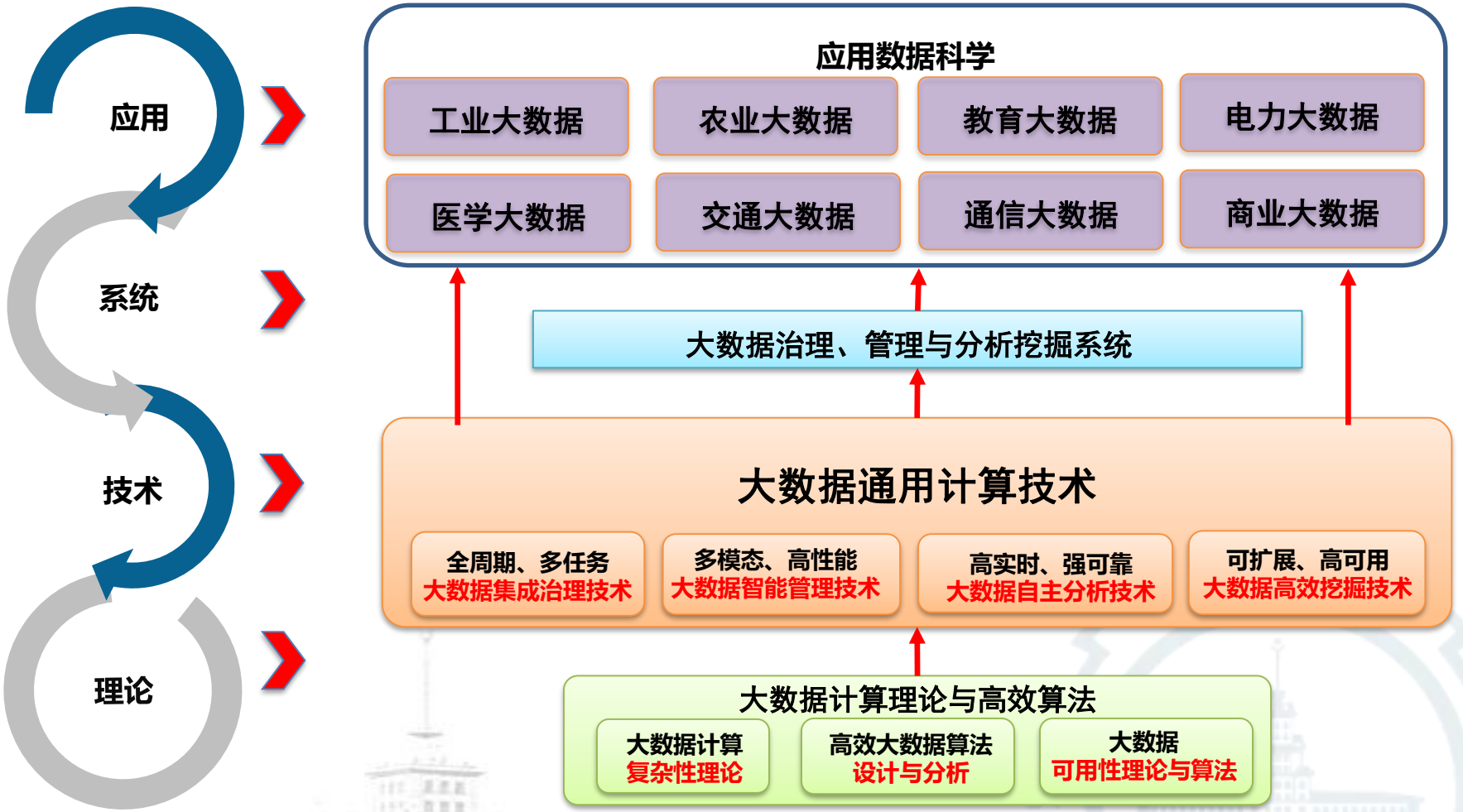
本章内容

10.1近似算法

10.2图论算法

10.3大数据算法

10.4总结



研究成果

- ◆ 研制我国第一个计算机机群系统和机群并行数据库系统，获国家科技进步二等奖
- ◆ 建立了压缩数据“无解压计算”方法学，开辟海量数据计算新途径
- ◆ 开展大数据可用性的基础理论和算法研究，引领我国大数据治理研究
- ◆ 开辟海量不确定图数据计算研究领域，获得教育部自然科学二等奖

算法在大数据时代的应用

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY 1920-2020
海量数据计算研究中心 Massive Data Computing Lab

物理世界正确映射到计算机世界

认知改造世界

海量数据获取 → 海量数据传输 → 海量数据存储 → 海量数据质量 → 海量数据问题求解

如何提升现有计算理论、算法设计方法学、计算系统性能，使之满足大数据计算需要?

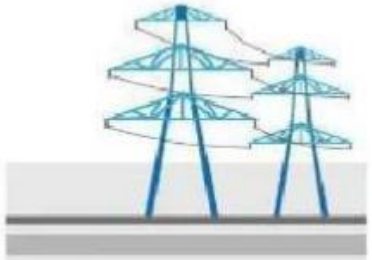
如何提高大数据的质量，确保大数据计算结果的可用性?

如何实现多学科交叉，凝练和解决的各领域的大数据问题?

算法在大数据时代的应用

1%的威力

航空
节约 1% 的燃料
300 亿美元



电力
节约 1% 的燃料
660 亿美元

医疗
系统效率提高 1%
630 亿美元

海量数据
问题求解

铁路
系统效率提高 1%
270 亿美元

石油天然气
资本支出降低 1%
900 亿美元

注：基于全球具体行业节约 1%
资料来源：GE。

算法在大数据时代的应用

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY 1920-2020
海量数据计算研究中心 Massive Data Computing Lab

物理世界正确映射到计算机世界

认知改造世界

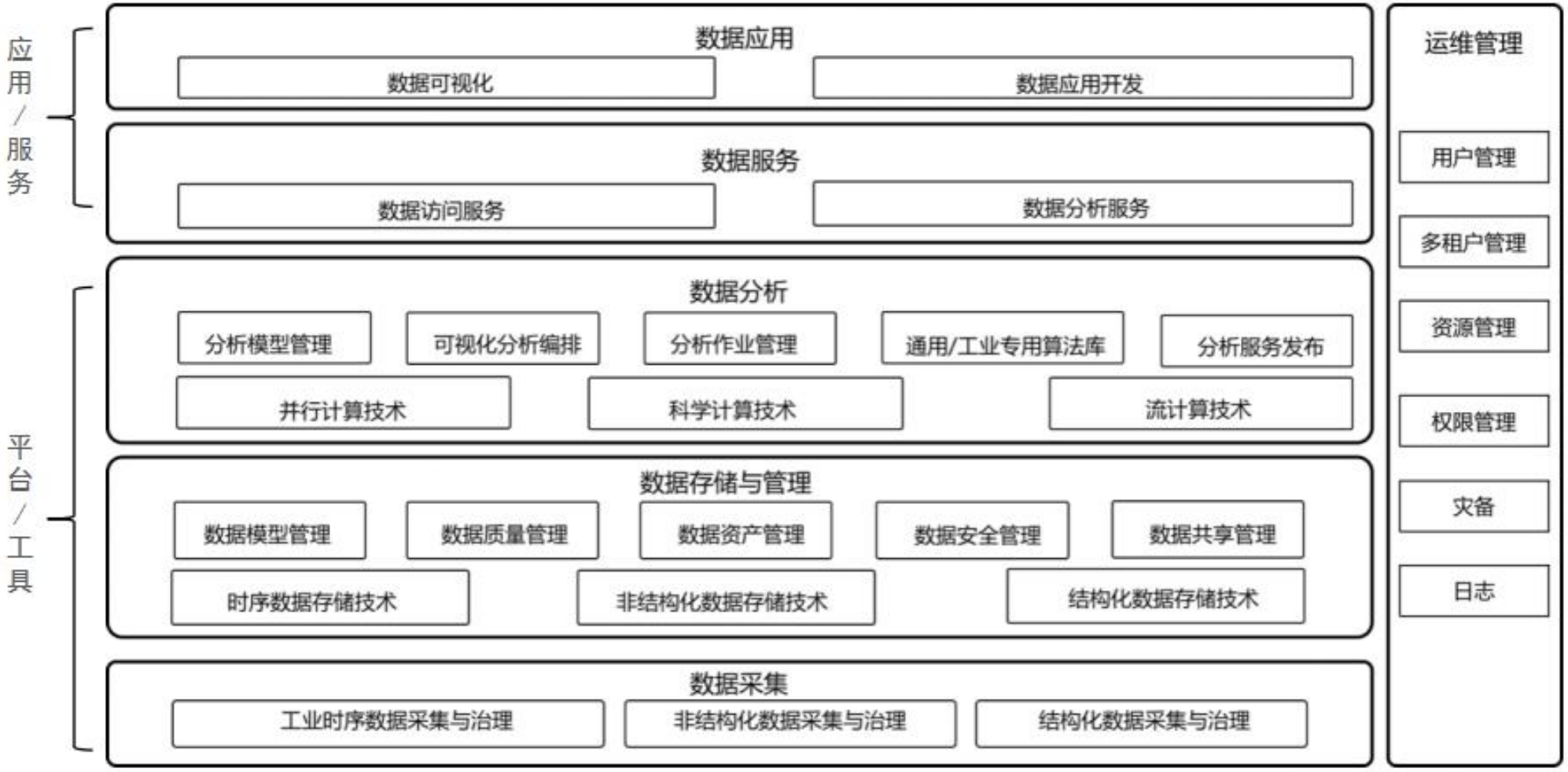
海量数据获取 → 海量数据传输 → 海量数据存储 → 海量数据质量 → 海量数据问题求解

- 大数据计算理论
- 大数据计算问题的理论与算法设计
- 大数据质量管理理论与方法
- 支撑人工智能的大数据技术(DB4AI)
- 智能大数据管理与分析理论与技术(AI4DB)
- 大数据分析与挖掘的理论与技术
- 面向应用(工业、电信、医疗等)的大数据计算管理与分析平台

算法在数据活动中的应用

哈爾濱工業大學 HARBIN INSTITUTE OF TECHNOLOGY 1920-2020
海量数据计算研究中心 Massive Data Computing Lab

- 数据质量管理是得到可靠分析结果的重要保障
  - 数据质量需求定义、评估、分析、提升和监控环节持续改善



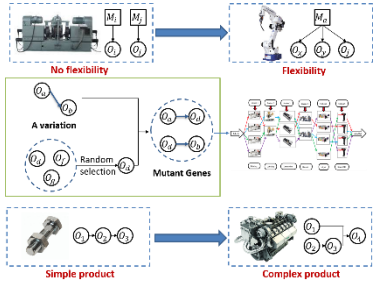
面向医疗、制造、能源、教育等重要领域开展大数据科学与技术方面的研究

智慧医疗

- ✓ 面向互联网资源的医学命名实体识别研究
- ✓ 面向智能导诊的个性化推荐系统
- ✓ 基于复合模型的传染病预警系统
- ✓ 基于大数据的心理健康分析系统
- ✓ 基于医疗数据挖掘的慢性病分析系统
- ✓ 基于半监督学习的疾病预测模型建立方法及装置

智能制造

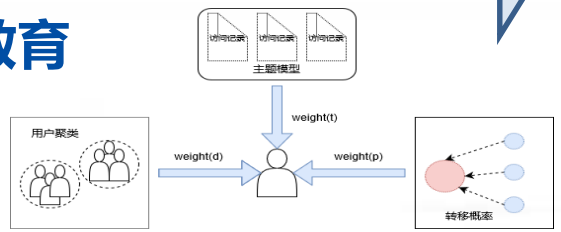
- ✓ 基于相关性分析的工业时序数据异常检测
- ✓ 面向工业大数据异构数据库的监控系统及监控方法
- ✓ 利用时间窗和迁移学习预测工业设备故障
- ✓ 时序数据错误检测与修复研究
- ✓ 面向众包数据清洗的主动学习技术



智慧能源

- ✓ 大电网潮流收敛调整、暂稳应用原型系统
- ✓ 基于机器学习的知识图谱双存储结构
- ✓ 知识图谱上有标签和子结构约束的可达性查询算法
- ✓ 半外存环境下高效强连通分量计算算法

智慧教育



- ✓ 教育大数据分析
- ✓ MOOC课程学生流失的分析与预警
- ✓ 基于LSTM的Mooc学习状态预测
- ✓ 面向大数据的课程架构设计
- ✓ 基于大规模开放在线课程系统的学生信息数据清理
- ✓ 基于学习行为的MOOC课程推荐系统

算法之路仍未停止

读更多的书、开展更多的思考、请教更多的人

学业进步、内心富足！

大数据计算基础理论 | 大数据获取 | 大数据治理 | 大数据管理 | 大数据分析挖掘

哈尔滨工业大学海量数据计算研究中心

MDC
MASSIVE DATA COMPUTING

教育部“海量数据计算理论与技术”创新团队
大数据科学与工程黑龙江省重点实验室
哈尔滨工业大学国际大数据计算研究中心

2024年4月