



大数据计算基础

第一章 绪 论

王宏志

wangzh@hit.edu.cn

<http://homepage.hit.edu.cn/pages/wang>

- 1 大数据
- 2 大数据的应用
- 3 大数据计算问题求解
- 4 大数据计算工具与技术

- 1 大数据
- 2 大数据的应用
- 3 大数据计算问题求解
- 4 大数据计算工具与技术

[illegible]

什么是大数据？

- 定义1 (Kusnetzky, Dan. What is "Big Data?")
 - 所涉及的数据量规模巨大到无法通过人工，在合理时间内达到截取、管理、处理、并整理成为人类所能解读的信息
- 定义2 (维克托 迈尔-舍恩伯格、肯尼斯 库克耶.“大数据时代”)
 - 不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法
- 定义3 (“大数据” (Big data) 研究机构Gartner)
 - “大数据” 是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产

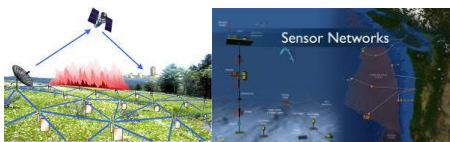


大数据是通过**传统数据库技术**和数据处理工具不能处理的**庞大而复杂**的数据集合。

处处皆是大数据



移动设备



传感网



科学仪器



社交网络



医疗数据



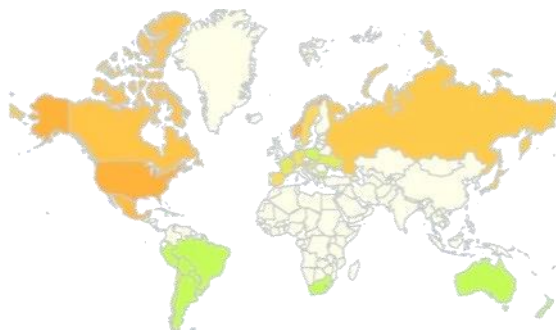
计算机艺术



商业数据

大数据研究意义

传染病预测



智能交通



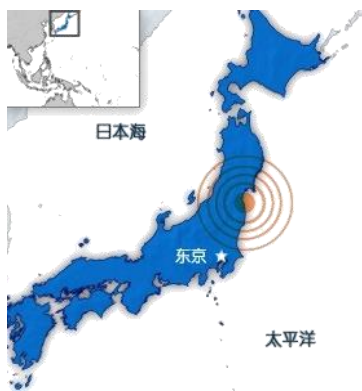
辅助社
会管理

推动科
技进步

大数据
计算

促进民
生改善

海啸实时预警



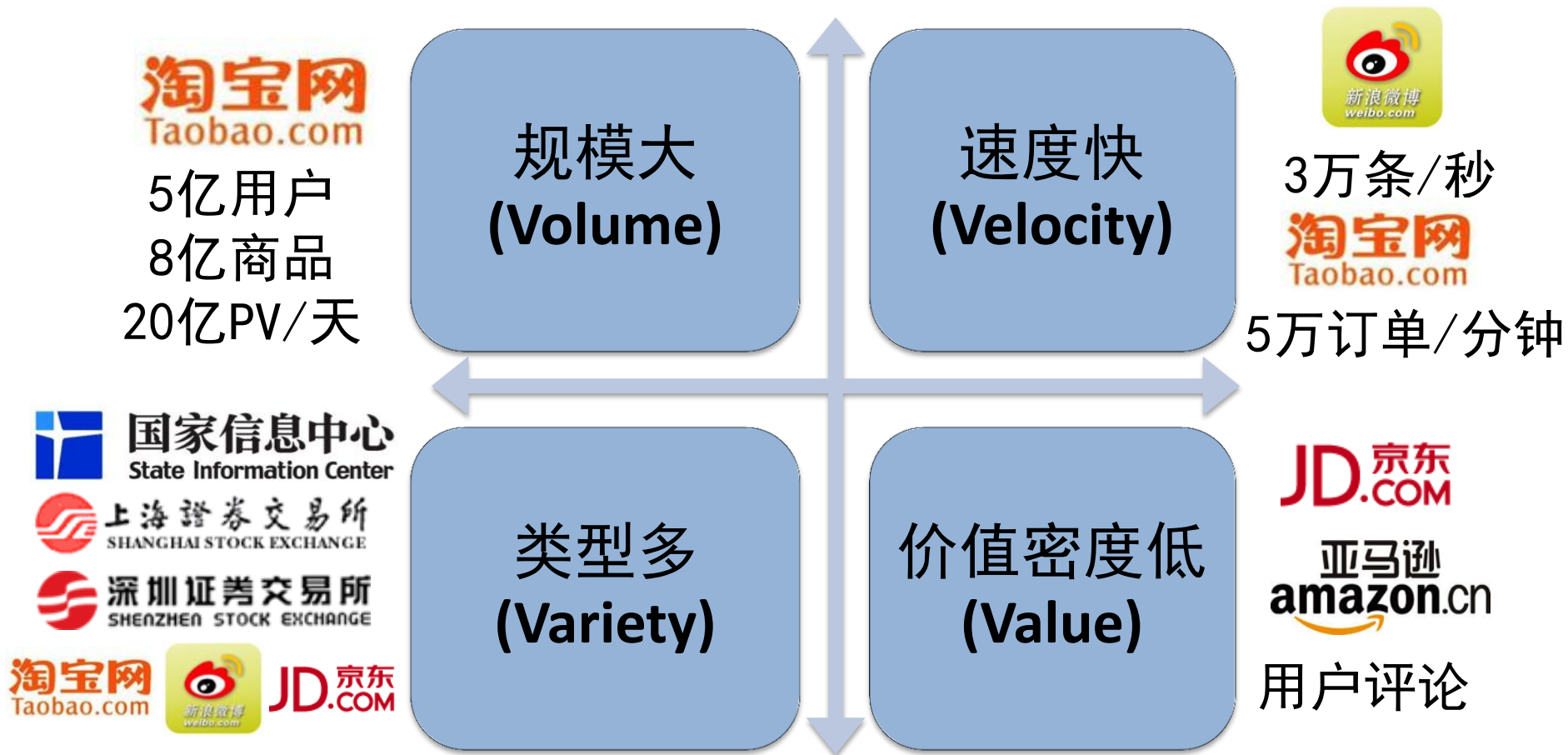
搜索与电子商务

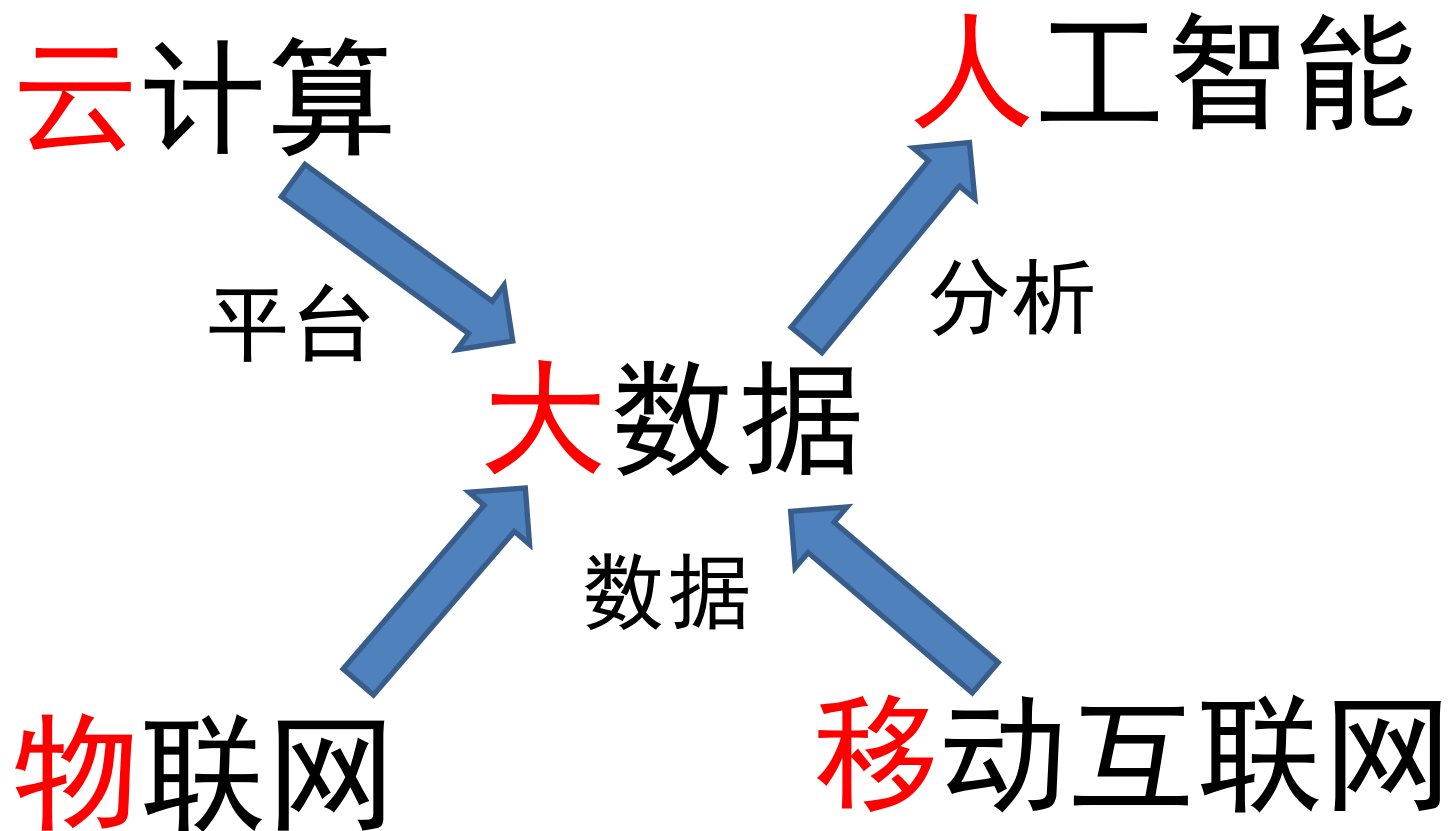
淘宝网
Taobao.com

Baidu 百度

支持商
业决策

大数据的特点





大数据到人工智能



大数据



大数据分析

因子分析
聚类分析
判别分析
关联规则挖掘
回归分析
关联分析
推断统计



知识

规则
模型
知识图谱
策略



人工智能

从大数据到人工智能示例



小冰小娜



Watson自动诊疗



无人驾驶

- 1 大数据
- 2 大数据的应用
- 3 大数据计算问题求解
- 4 大数据计算工具与技术

大数据在科学中的应用

Science Paradigms

- Thousand years ago:
science was **empirical**
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a **computational** branch
simulating complex phenomena
- Today: **data exploration** (eScience)
unify theory, experiment, and simulation
 - Data captured by instruments
or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files
using data management and statistics



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

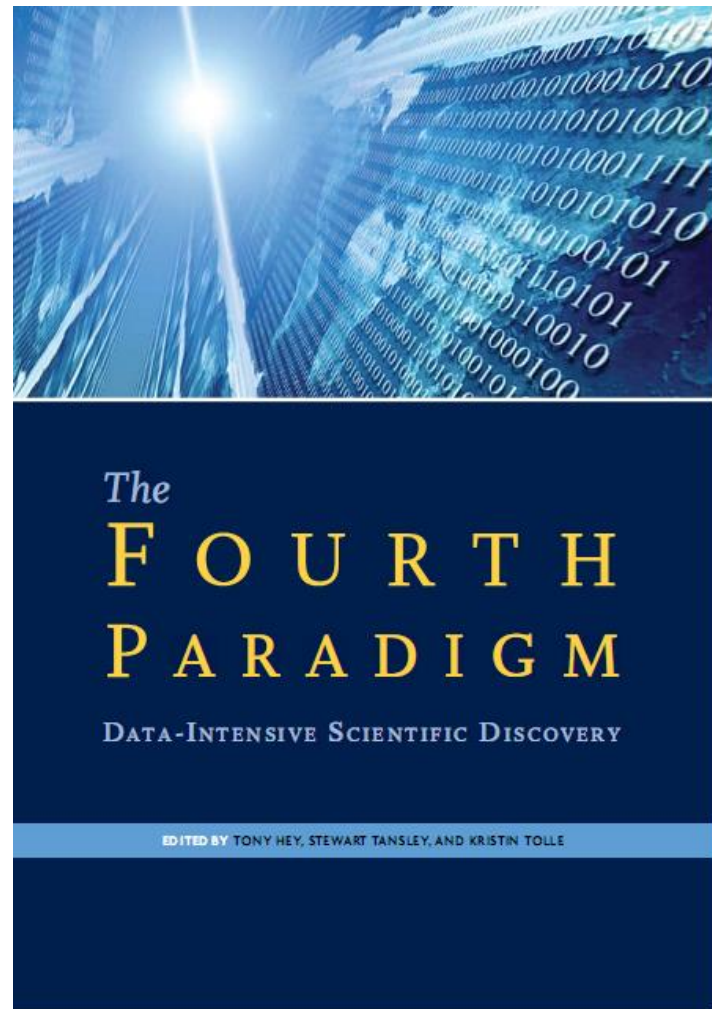


FIGURE 1

第四范式： 数据密集型科学发现

USA:

Microsoft Research 2009



- 科学研究由假设驱动转向基于探索的科学方法
- 过去设问 “我应该设计什么样的实验来验证这个假设？”
- 现在设问 “从这些数据中我能够看到什么？”
- “如果把其他领域的数据溶合进来，能够发现什么？”
- 天文学研究不再用肉眼看望远镜，而是把望远镜观察到的现象以数据形式记录到计算机，对数据进行分析判断

大数据在工业中的应用

长虹集团



- 应用回归分析计算度量参数发现提高生产率的关键参数
 - 空气湿度的方差和生产率呈正相关
 - 气压和生产率正相关
 - 生产车间温度和温度方差在很小的范围内波动
- 利用回归分析建立了控制参数和生产率的关系
- 应用基于规则的分类发现生产时间瓶颈，重新规划流程，提高生产率
- 应用基于关联规则的分类提前发现半完成的有瑕疵面板，避免资源浪费

波音737

- 利用大数据有效实现故障诊断和预测
- 发动机在飞行中每30分钟产生10 TB数据
- 促进实时自适应控制、燃油使用、零件故障预测和飞行员通报



GE能源监测和诊断（M&D）中心

- 收集全球50多个国家上千台GE燃气轮机的数据，每天为客户收集10G的数据
- 分析来自系统内的传感器振动和温度信号的恒定大数据流
- 基于大数据分析将为GE公司对燃气轮机故障诊断和预警提供支撑



imagination at work

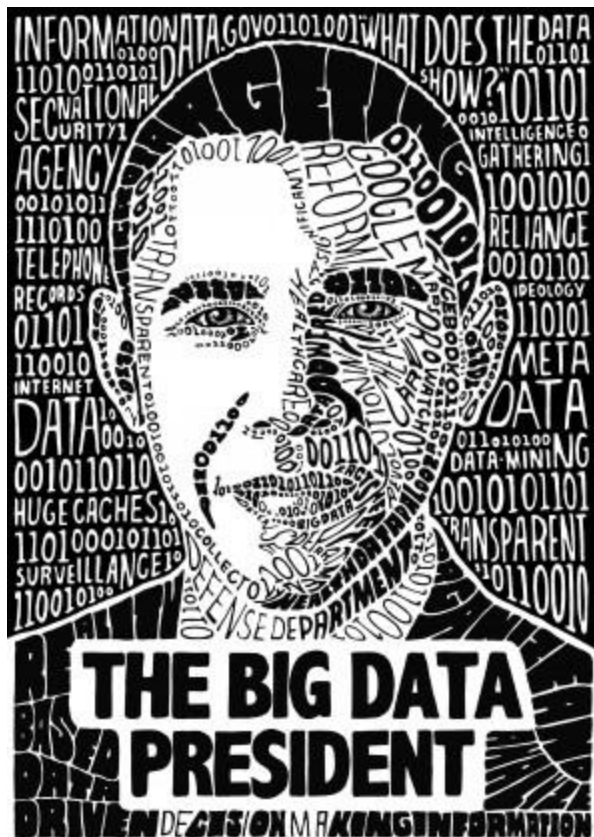
某生物制药企业



- 通过培养生物工程细胞获取最终的药品
- 监视200种以上的变量
 - 确保原料成分的纯净度
 - 确保生产出的药品符合标准
- 在完全相同的两个批次的生产流程当中，最终产量会有50%到100%的差异
 - 将生产过程中紧密相关的步骤结合成簇
 - 分析每个簇中的过程数据
 - 追踪到9个影响产量的变量
 - 发现层析法过程中细胞的接种时间和导电率是影响最终产量的重要原因
 - 优化生产流程
- 从而将疫苗的产量提高了50%，每年在单一疫苗品种上节省的费用达到1000万美元

大数据在社会与经济中的应用

美国大选预测



State-by-State Probabilities

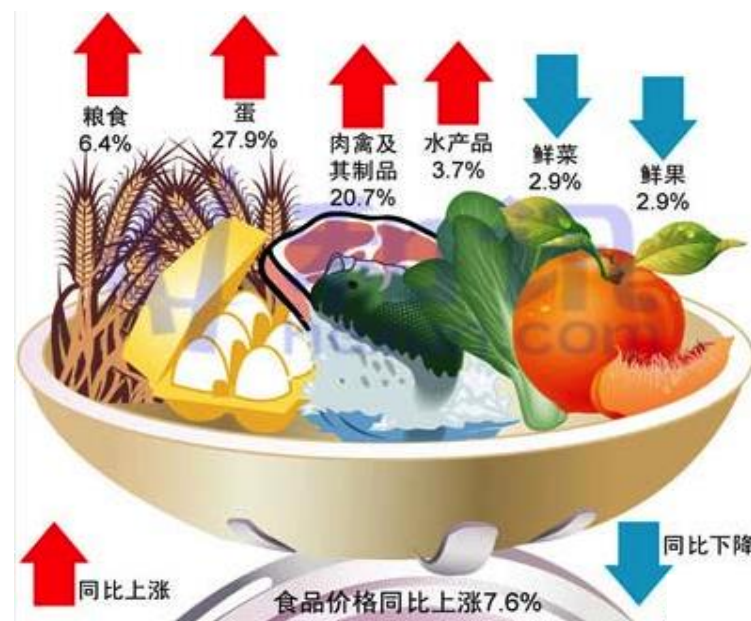


奥斯卡奖预测



- **Best Picture:** *Argo*
- **Best Director:** Steven Spielberg (Rothschild, Brandwatch popular) / David O. Russell (Brandwatch critics)
- **Best Actor:** Daniel-Day Lewis
- **Best Actress:** Jennifer Lawrence (Rothschild, Brandwatch popular) / Jessica Chastain (Brandwatch critics)
- **Best Supporting Actor:** Tommy Lee Jones (Rothschild) / Christoph Waltz (Brandwatch popular) / Robert de Niro (Brandwatch critics)
- **Best Supporting Actress:** Anne Hathaway
- **Best Animated Film:** *Brave*
- **Best Original Song:** Adele's "*Skyfall*"

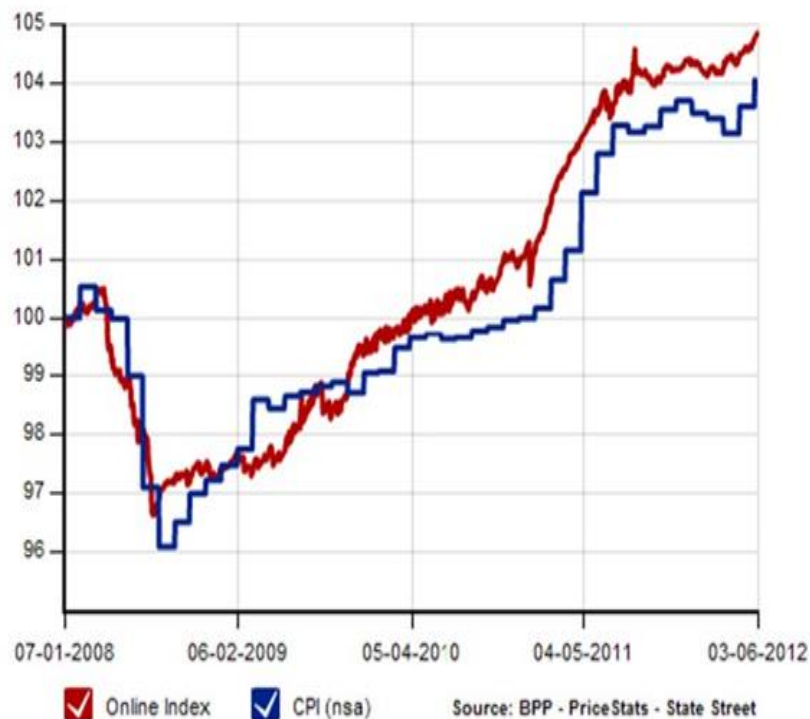
- 居民消费价格指数(CPI)
- CPI意义：
 - 与民生密切相关的国家经济决策重要指标
 - 反映通货膨胀率
- 目前存在问题：
 - “滞后、不科学”——原社科院金融发展室主任易宪容
 - 86%认为CPI与消费感受不符合”——中国政协网



如何准确计算分析CPI → 大数据计算

每日网上价格指数 Daily Online Price Index

- 美国麻省理工学院承担的一项“十亿价格项目”
(Billion Price Project) 是基于学术研究方法对**全世界海量网上零售价格**进行价格指数计算
- 为判断通胀趋势提供信息
- 每天实时收取**50万条**互联网上的商品信息,是美国政府统计收集的**5倍**



每日网上价格指数（2008年7月为基期）与居民消费价格指数的比较

淘宝网络零售价格指数 ISPI

- 基于淘宝网、天猫网、支付宝等网络平台的数据编制
- 大体反映国内网络零售渠道的一般物价变动。包含价格指数系列和实物交易量指数系列
- 分为九大基本分类指数
- 权重为成交金额的比例

iSPI 与 CPI 关系图



来源：国家统计局 阿里研究中心 2011 年 10 月

多点碰撞

2013年5月8日
A市发生盗窃案件



两起案件

不同时间

不同城市

作案手法相似

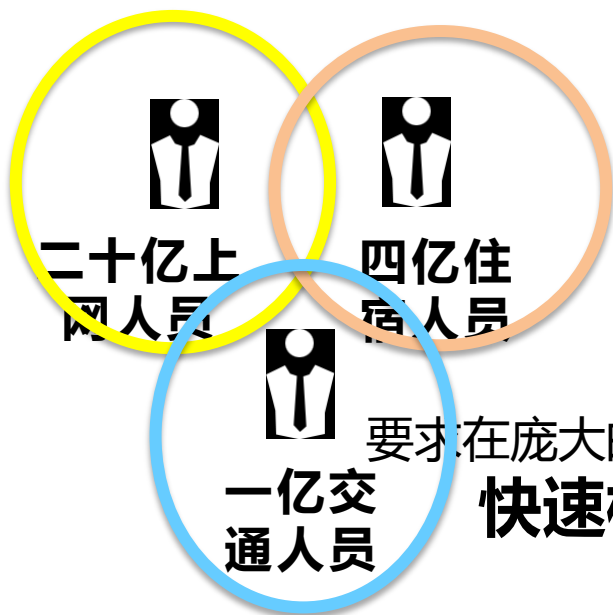
疑似流窜作案

无其它线索

2013年6月3日
B市发生盗窃案件



多点碰撞



首先需找到案发前后三天
在A地出现的人



要求在庞大的数据群中
然后需找到案发前后三天
在B地出现的人

快速检索

在B地出现的人



多点碰撞



将两类数据来回交叉对比
找出案发前后同时在两市出现的人



多点碰撞

对碰撞结果中的人员逐一排查，最后成功锁定嫌疑人



大数据在医疗健康中的应用




基于大数据分析的智能健康系统

23:58 44

Play good recently, every day very substantial. Often and students go out mountain climbing, everyone was very happy. With them every day, we laughed together, play football, play basketball

Save

00:00

44%

1 Do you feel loss of appetite? Or can't help but eat too much?

☐ NO
 ☐ A LITTLE
 ☐ MEDIUM
 ☐ HEAVY

2 Do you suffer from insomnia? Or feel tired and sleepy all day?

☐ NO
 ☐ A LITTLE
 ☐ MEDIUM
 ☐ HEAVY

COMMIT

MOOD

DIARY

SMS

TEST

MOOD

DIARY

SMS

TEST



手机12:11 掌上保健医

患者提供信息提升系统

潜在慢性疾病的患者检测

12:27 100%

掌上保健医

请选择一种慢性疾病

- 糖尿病
- 体征
- 尿毒症
- 消化性溃疡
- 冠心病

确定

- 1 大数据
- 2 大数据的应用
- 3 大数据计算问题求解**
- 4 大数据计算工具与技术

大数据处理技术路线



物理世界正确映射到计算机世界

认知改造世界

大数据获取

大数据传输

大数据存储

大数据质量

大数据问题求解

Intractability

如何提升现有计算管理

论、如何提高大数据的

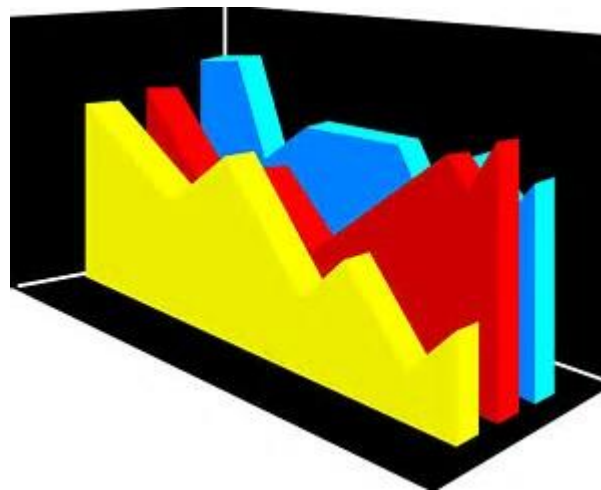
计算

如何实现多学科交叉，凝练和解决的各领域的大数据问题？

大数据计算需要的新思路



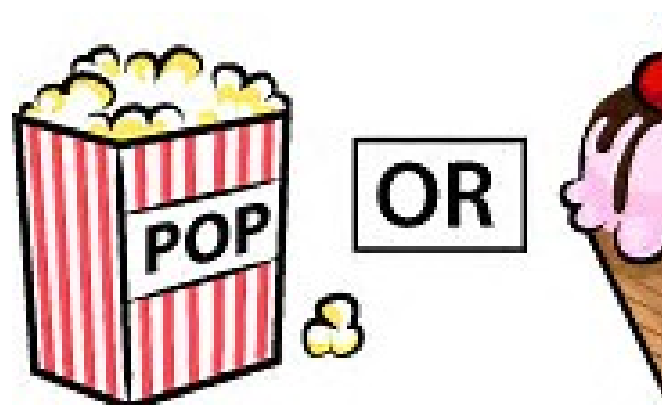
系统



建模

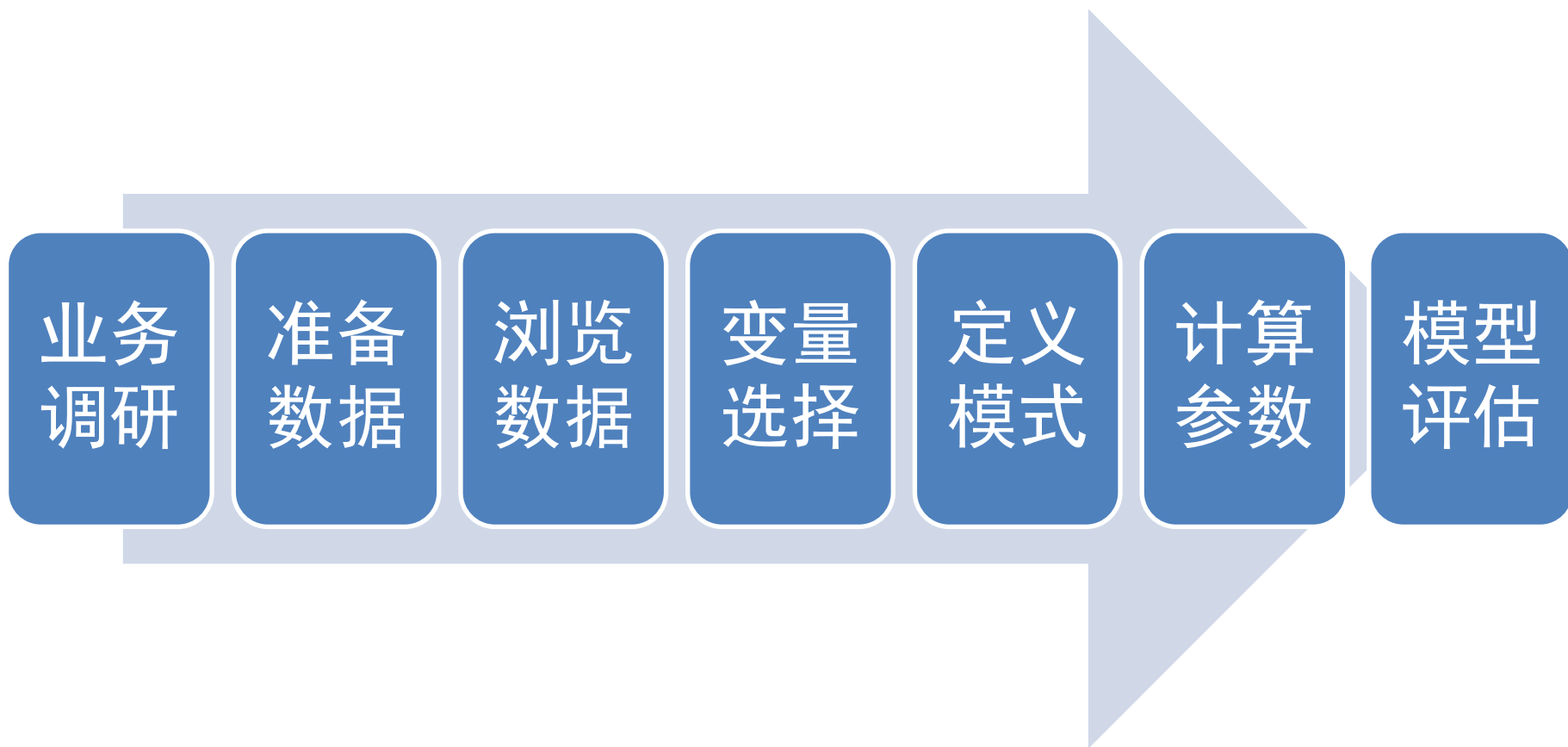


实现

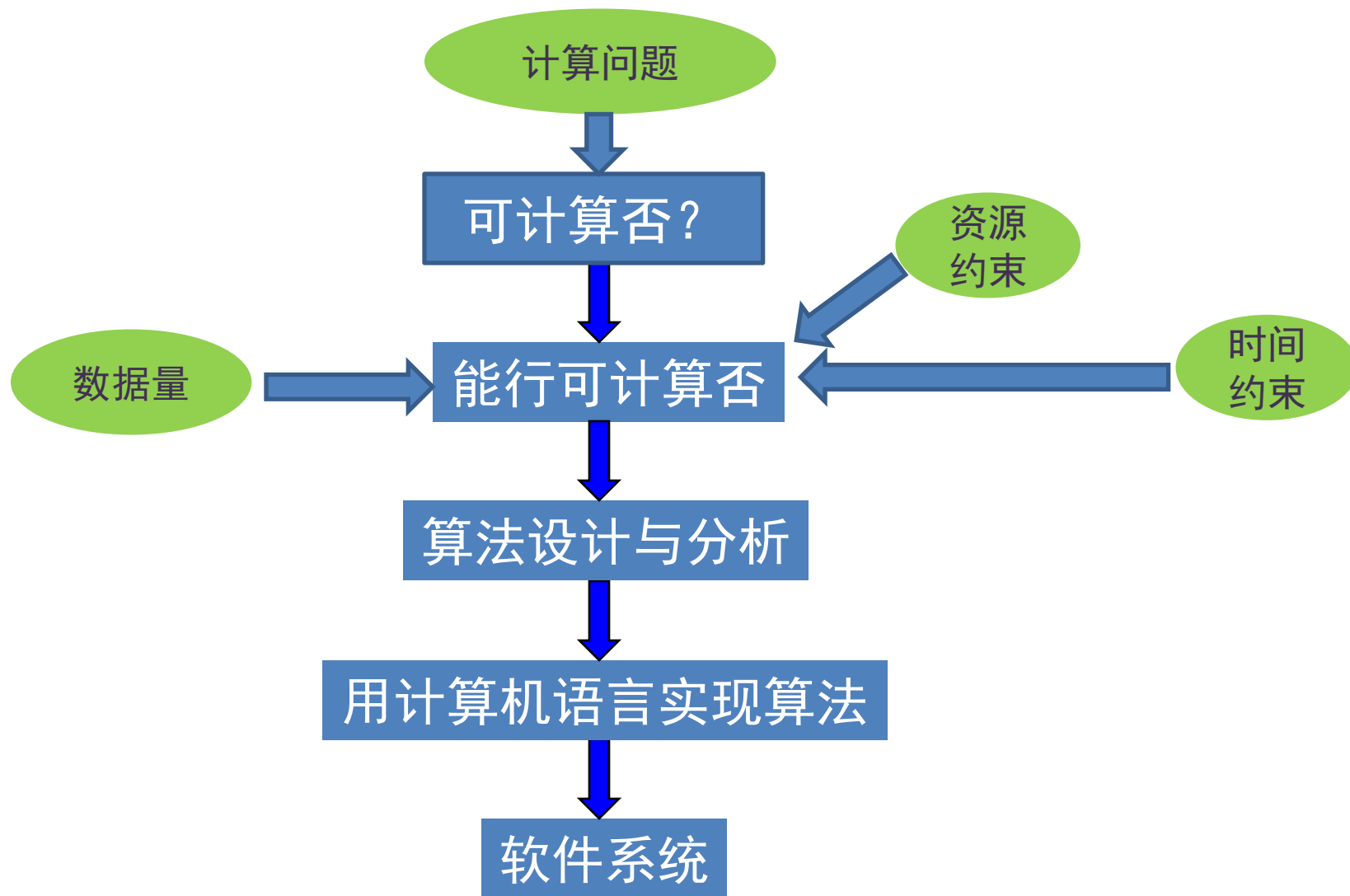


折衷

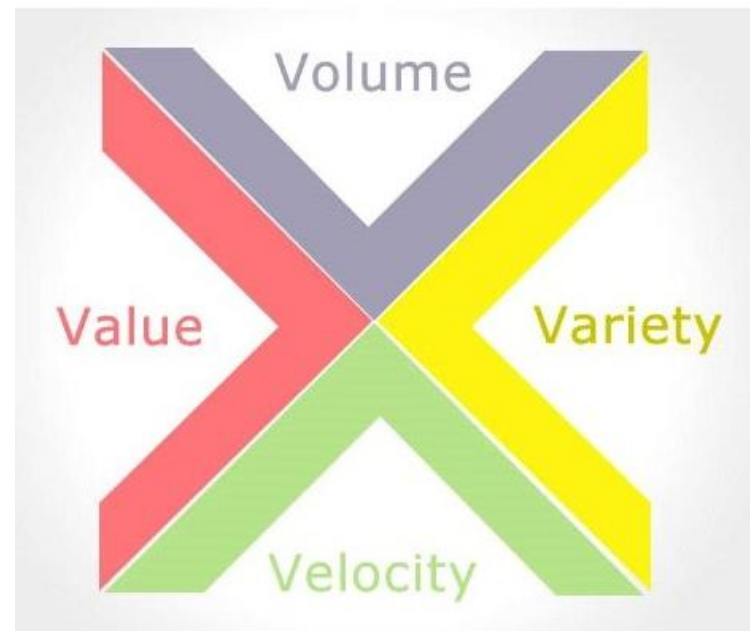
大数据问题求解过程



求解大数据计算问题的过程



- 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



“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

- 1 大数据
- 2 大数据的应用
- 3 大数据计算问题求解
- 4 大数据计算工具与技术

大数据计算工具

数据采集

数据处理

数据分析

数据应用

数据获取



数据仓库及BI一体机



数据分析



分析应用



数据提供



非结构化数据



数据可视化



商业智能工具



数据管控



- 数据为中心的計算框架
 - Hadoop, Spark, Hyracks
- 流处理框架
 - S4, Spark Streaming, Storm, Samza
- 分布式图计算框架
 - Hama, Pregel, GraphEngine, Pregelix, Apache Giraph, Phoebus
- 分布式文件系统
 - HDFS, GFS, GridGain, Seaweed-FS, Tahoe-LAFS, Colossus



- 文档数据库

- Actian Versant, Crate Data, Facebook Apollo, jumboDB, LinkedIn Espresso, MarkLogic, MongoDB, RavenDB, RethinkDB

- 键值存储

- Aerospike, Amazon DynamoDB, Edis, ElephantDB, EventStore, GridDB, LinkedIn Krati, LinkedIn Voldemort, Oracle NoSQL Database, Redis, Riak, Storehaus, Tarantool, TiKV, TreodeDB

- 基于列的存储

- Apache Accumulo, Apache Cassandra, Apache HBase, Facebook HydraBase, Google BigTable, Google Cloud Datastore, Hypertable, InnoDB, Tephra, Twitter Manhattan, Columnar Storage

- 图存储

- ArangoDB, DGraph, Facebook TAO, Google Cayley, Neo4j, OrientDB, Titan, Twitter FlockDB

- NewSQL

NewSQL

- Actian Ingres, Amazon RedShift, BayesDB, CitusDB, Cockroach, Datomic, FoundationDB, Google F1, Google Spanner, H-Store, Haeinsa, HandlerSocket, InfiniSQL, MemSQL, NuoDB, Oracle TimesTen in-Memory Database, Pivotal GemFire XD, SAP HANA, SenseiDB, Sky, SymmetricDS, Map-D, TiDB, VoltDB

- 时间序列数据库

- Cube, Axibase Time Series Database, Heroic, InuxDB, Kairosdb, OpenTSDB, Prometheus, Newts

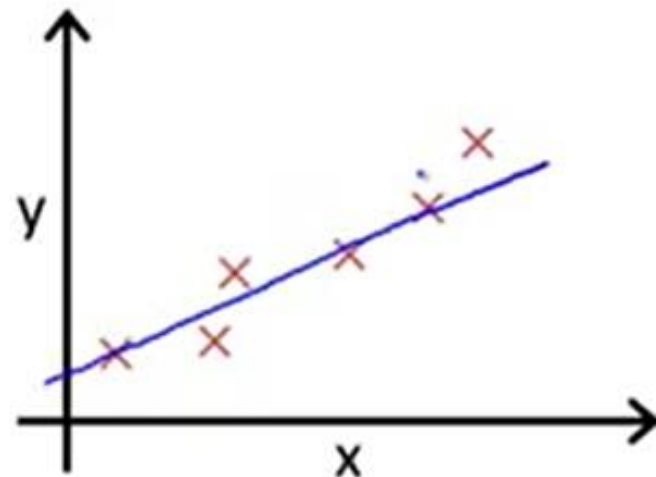


- 嵌入式数据库

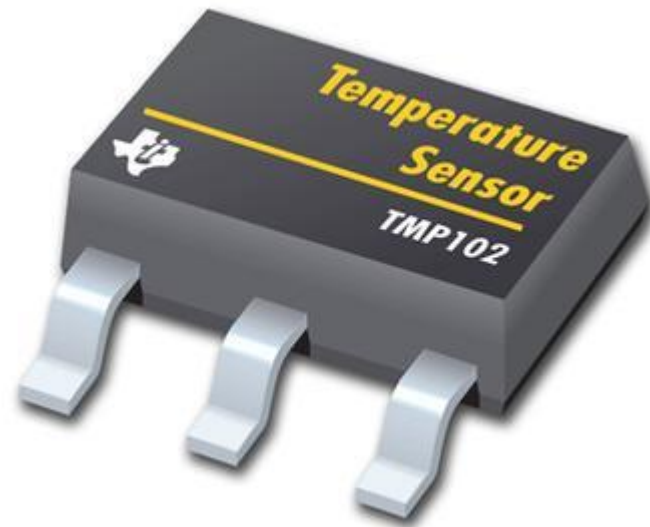
- Actian PSQL, BerkeleyDB, HanoiDB, LevelDB, LMDB, RocksDB



- Apache Mahout, brain, Cloudera Oryx, Concurrent Pattern, convnetjs, Decider, ENCOG, etcML, Etsy Conjecture, Google Sibyl, GraphLab Create, H2O, MLbase, MLPNeuralNet, MonkeyLearn, nupic, PredictionIO, SAMOA, scikit-learn, Spark MLlib, Vowpal Wabbit, WEKA, BidMach, SparkR, Zoomdata, Jethrodata



- Apache Chukwa, Fluentd, LinkedIn Gobblin, StreamSets Data Collector, Apache Nutch



- Apache Ambari, Apache Bigtop, Apache Helix, Apache Mesos, Apache Slider, Apache Whirr, Apache YARN, Brooklyn, Buildoop, Cloudera HUE, Facebook Prism, Google Borg, Google Omega, Hortonworks HOYA, Marathon, Apache Aurora, Apache Falcon, Apache Oozie, qqChronos, LinkedIn Azkaban, Schedoscope, Sparrow, Airflow



- Apache Hadoop Benchmarking, Berkeley SWIM Benchmark, Intel HiBench, PUMA Benchmarking, Yahoo Gridmix3

Benchmark

可视化工具

- Airpal, Arbor, Banana, Bokeh, C3, CartoDB, chartd, Chart.js, Chartist.js, Crossfilter, Cubism, Cytoscape, DC.js, D3, D3.compose, D3Plus, Echarts, Envisionjs, FnordMetric, Freeboard, Gephi, Google Charts, Grafana, Graphite, Highcharts, IPython, Kibana, Matplotlib, Metricsgraphic.js, NVD3, Peity, Plot.ly, Plotly.js, Recline, Redash, Shiny, Sigma.js, Vega, Zeppelin, Zing Charts



– Apache Knox Gateway

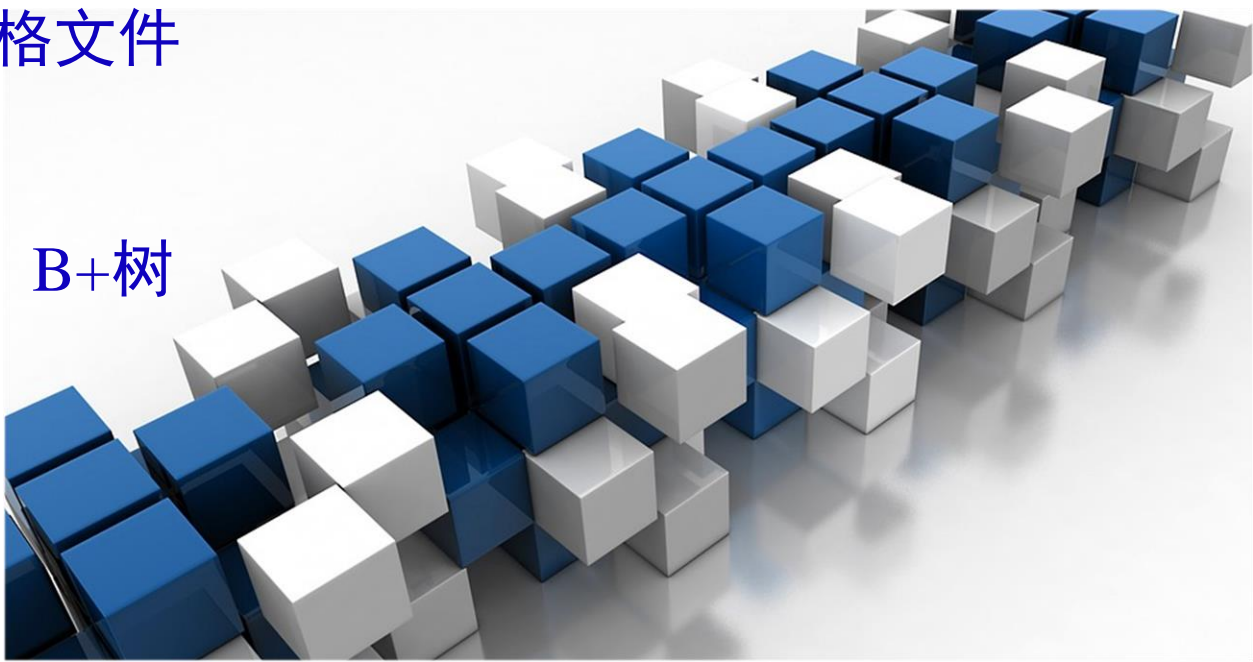


– Apache Sentry



- 大数据程序的特点
 - 数据量过大，无法全部放入内存
 - 避免创建大量的对象
- 需要考虑的
 - 磁盘操作，包括对数据进行加锁、磁盘数据访问方法
 - 内存管理，编写程序以管理内存而不是使用默认的分配方法
- 程序设计方法
 - 基于计算框架的程序设计
 - MapReduce
 - Pregel

- 针对大数据的数据结构
- 随机化的数据结构
 - Min-hash, LSH, 布鲁姆过滤器
- 基于磁盘的数据结构
 - B树, R树, 网格文件
- 并行数据结构
 - 分布式哈希表, B+树

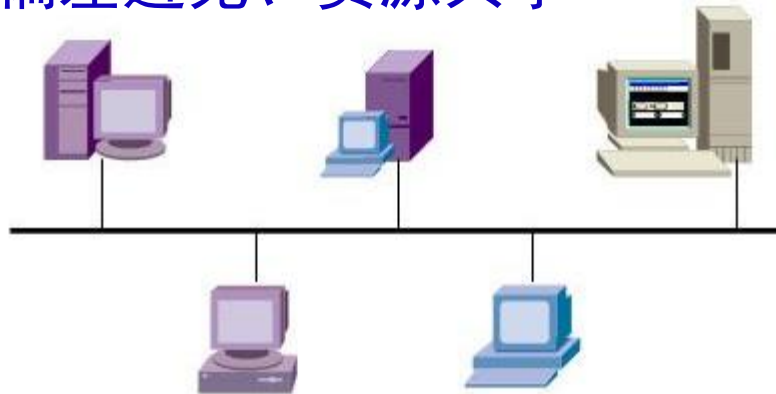


- 大数据算法设计的意识
 - 多项式时间算法往往不能满足要求，需要达到线性和亚线性，可以使用随机算法
 - 主存不足，需要设计外存算法
 - 由于算法的规模过大，需要设计并行算法
 - 由于缺乏知识，需要引入众包技术
- 在效率、资源、准确率之间进行折衷
 - 例如，亚线性算法往往在效率和准确率之间进行折衷

- 在大数据时代，数据管理需要给予额外的关注
- 面向大数据的新技术
 - 有效的数据存储结构
 - 数据库系统中的基本数据操作的高效实现，例如选择、连接、聚集
 - No-SQL和newSQL



- 在网络中传输的不仅是数据，还有计算代码
 - 由于数据量过大，将数据传送到计算端可能无法实现
 - 需要将计算代码传送到数据端
 - 以相应的框架为例，如MapReduce
- 数据中心中的网络
 - 关注云端数据的效率和有效性
 - 使用了容错技术、拥塞控制、偏差避免、资源共享

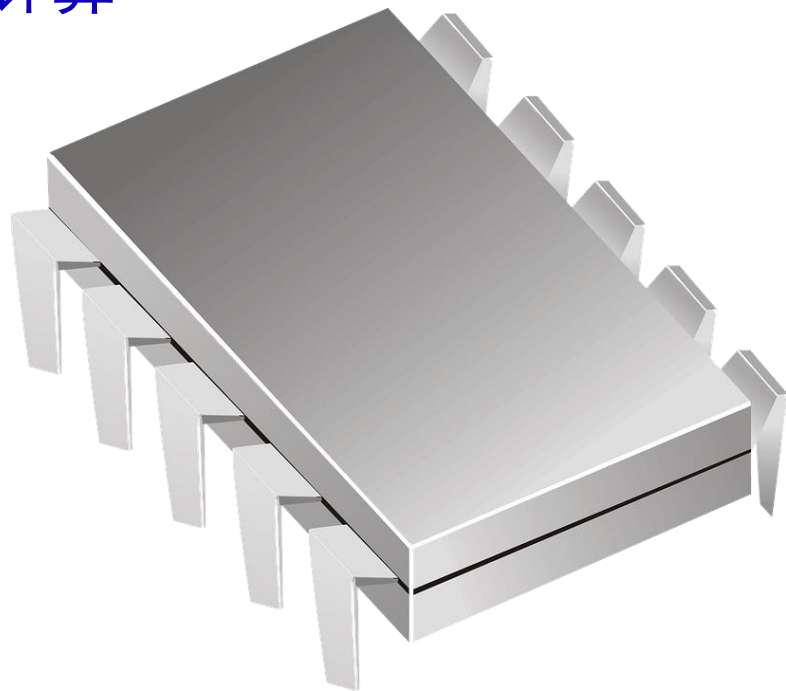


- 面向大数据的编译程序优化技术
 - 针对大数据的编译原则
 - 大数据程序运行实时环境
 - 适应大数据的代码优化技术
- 利用编译知识优化程序
 - 例如，在Java程序设计中，应该避免对象的声明和删除

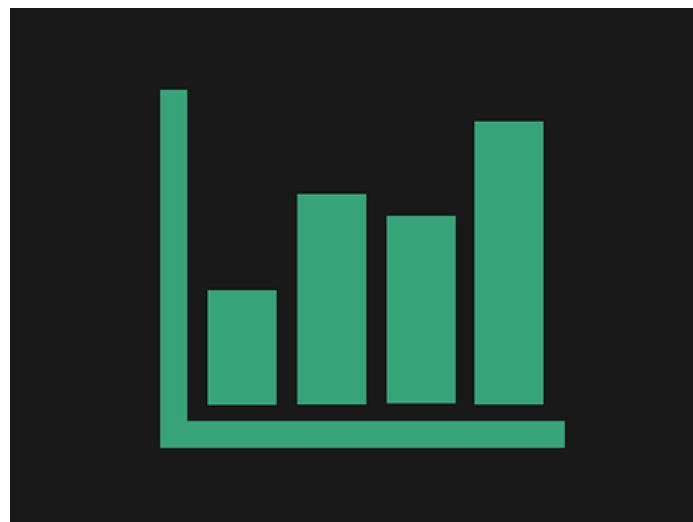
- 大数据计算平台
 - 面向大数据的分布式计算平台原理
 - 灵活性和可扩展性
- 分布式文件系统
 - 单个机器无法管理和使用大规模的数据，因此引入分布式文件系统
 - Cache策略，数据传送，并发控制



- 体系结构帮助理解数据中心化的计算
 - CPU, GPU和FPGA的协作
 - 数据中心新的硬件
 - 节省能源的计算技术
 - 计算单元、内存、外存的协同计算



- 基础是统计和机器学习
- 两个新要点
 - 真实场景下的大数据分析的应用
 - 针对统计和机器学习的可扩展的技术，这主要通过大数据的数据结构和算法来实现



- 重要性
 - 大数据具有容量大、产生速度快、种类多的特点
 - 更加容易产生数据质量问题
 - 劣质数据误导分析结果
- 内容包括
 - 数据质量的度量
 - 数据质量的评估
 - 数据质量问题的检测
 - 数据清洗



- 重要性

- 对于安全和合法的应用来说，安全和隐私是非常重要的
- 安全和隐私已经引起了企业和学术界的重视

- 技术

- 数据安全技术：支持查询和分析的加密算法
- 数据隐私保护技术：k匿名、差分隐私保护

- 资产
 - 数据常常被当作类似于土地和石油一样的资产看待
 - 数据明显有别于土地和石油
- 主要技术包括
 - 数据资产的支配
 - 数据资产的管理技术
 - 数据定价技术

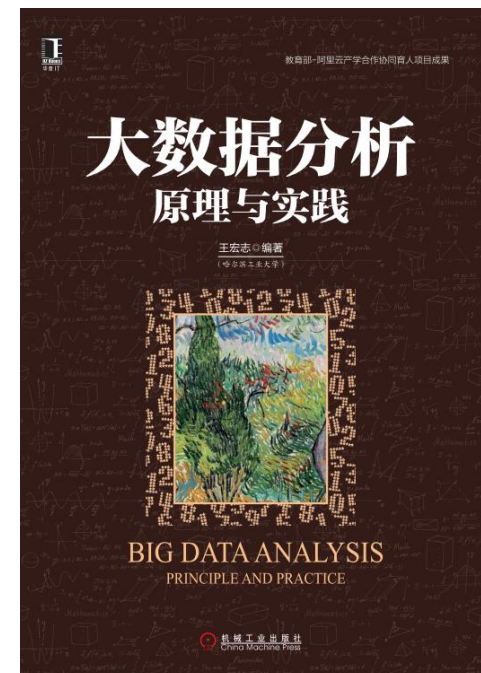
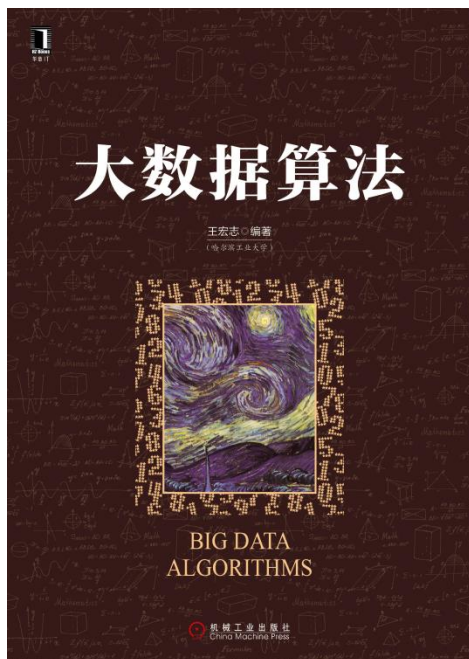


- 大数据计算的(时间、空间、能耗、通信)复杂性理论
- 大数据计算的算法设计方法学
- 大数据质量管理的理论与方法
- 大数据获取的理论与方法
- 大数据问题求解的理论和方法
- 大数据管理与服务平台
- 面向应用的大数据计算理论与算法



大数据计算系统与技术

且听本门课为你娓娓道来



谢谢！

Thanks for attention!

报告人：王宏志