多源异构大数据的机器学习关键技术研究进展

徐增林
zenglin@gmail.com
电子科技大学
大数据研究中心
计算机科学与工程学院

统计机器智能与学习实验室

统计机器智能与学习实验室 **(Statistical Machine Intelligence & LEarning, SMILE)**

- 网址：http://bigdatalab.weebly.com

研究目标：

- **复杂多源异构数据处理技术**：分类、聚类、半监督学习、多核学习、特征选择、多任务学习、多视角学习、集成学习、网络分析、张量分析
- **统计机器学习理论研究**: 近似算法、随机投影算法、稀疏学习等的理论
- **贝叶斯图模型研究**：高斯过程、主题模型、隐变量模型
- **机器学习的优化与推断研究**：最优化算法、Variational Inference、先进采样算法、混合算法
- **机器学习大数据平台研究**：在线学习、分布式学习
- 机器学习在社会网络、神经信息学、健康、安全等领域的应用

大数据挖掘与推理研究所

大数据挖掘与推理研究所（Institute of Big Data Mining and Reasoning)@电子科大大数据研究中心

研究目标：

- 异构多源大数据处理与建模
  - 实时数据处理、多源数据处理、时间空间数据分析、复杂网络数据分析、金融大数据建模、媒体大数据建模、医学大数据建模、移动大数据建模
- 大数据智能计算与分析技术
  - 分布式大数据查询技术、先进机器学习与数据挖掘理论研究、并行化机器学习和数据挖掘算法研究、随机化算法与在线学习、社会网络分析、Web挖掘与检索、商业智能、排名与推荐算法、深度学习算法、大数据降维技术
- 大数据分布式计算模型与系统
  - 大数据分析平台Hadoop/Spark性能优化与功能增强、大数据机器学习平台研究、面向行业应用（如医疗、教育、安全、移动数据）的大数据分析与学习平台设计等
- 大数据知识表示与推理技术研究
  - 大型本体知识库构建方法和本体映射等知识深层理解的关键处理算法、知识的深层表示、大型知识库上逻辑推理机制和机器学习

大数据挖掘与推理研究所

电子科大大数据研究中心
大数据挖掘与推理研究所（Institute of Big Data Mining and Reasoning）

主要人员：

- 周涛（大数据中心主任、优青、拔尖、教授）
- 申洪涛（大媒体计算中心主任，千人计划入选者）
- 徐增林（青年千人计划入选者，教授）
- 符红光（863子课题负责人）
- 邵俊明（校百人、教授）
- 邵杰（校百人、教授）
- 杨阳（校百人、教授）
- 尚明生（教授）

加入我们

中组部“青年千人计划”入选者徐增林教授团队，因科研和教学工作需要，面向海内外诚聘优秀青年学者加盟。团队的研究着重于机器学习、统计学习、数据挖掘技术及其在社会网络分析、医学图像处理、空间安全数据分析、神经信息学等方面的应用。

详情： http://bigdatalab.weebly.com/，人力资源部 http://www.hr.uestc.edu.cn/hr/info/9481.htm

- 研究助理/博士生/硕士生
- 特聘教授/特聘副教授/骨干教师/在职和脱产博士后

在研项目：

1. 运维大数据平台设计与实现
2. 医疗大数据分析平台设计与实现
3. 基于异构计算的大数据平台设计与实现

报告提纲

大数据的发展

大数据分析面临的挑战

大数据机器学习算法与平台

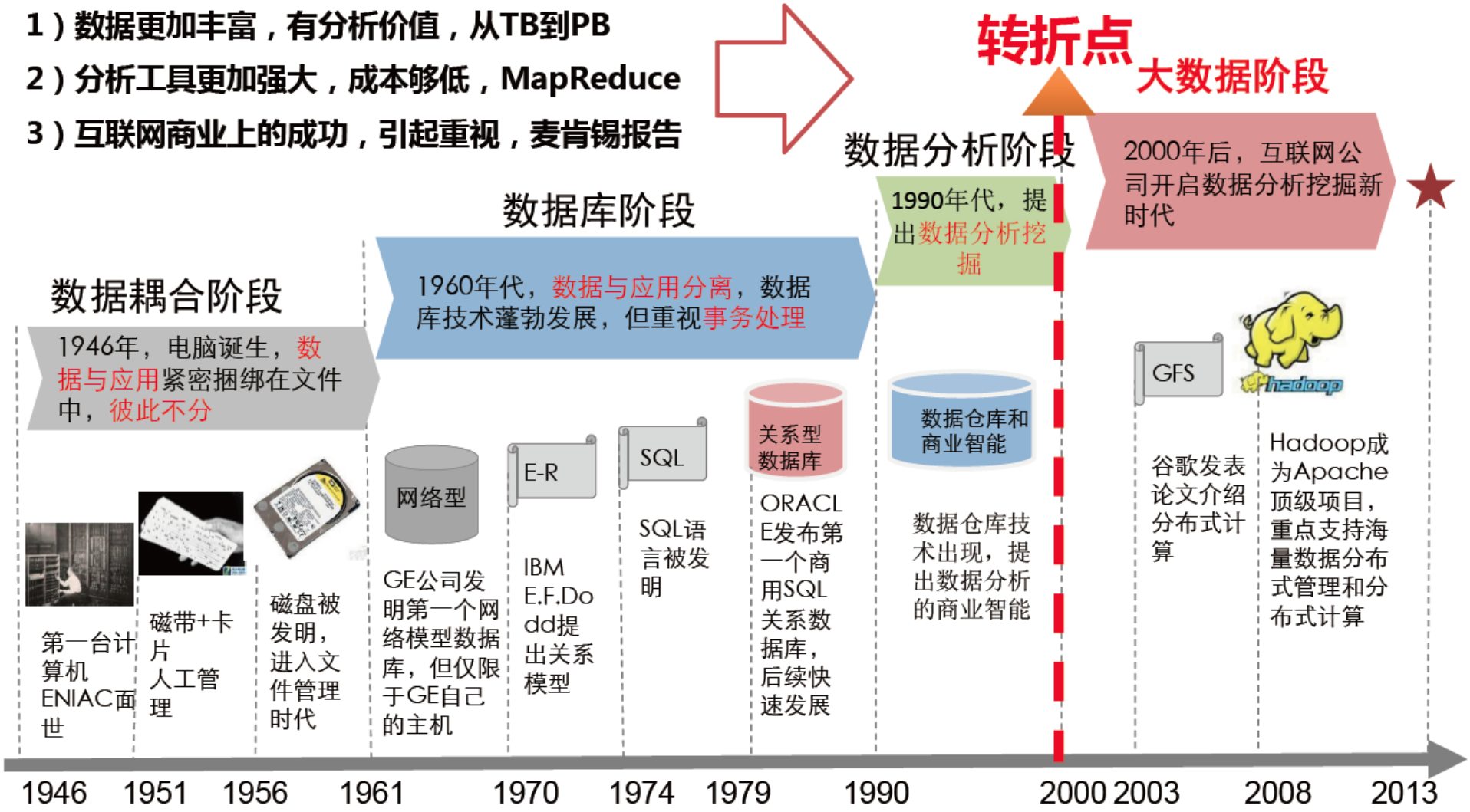
大数据发展历史

信息化从重视流程电子化到重视数据资产化转变。

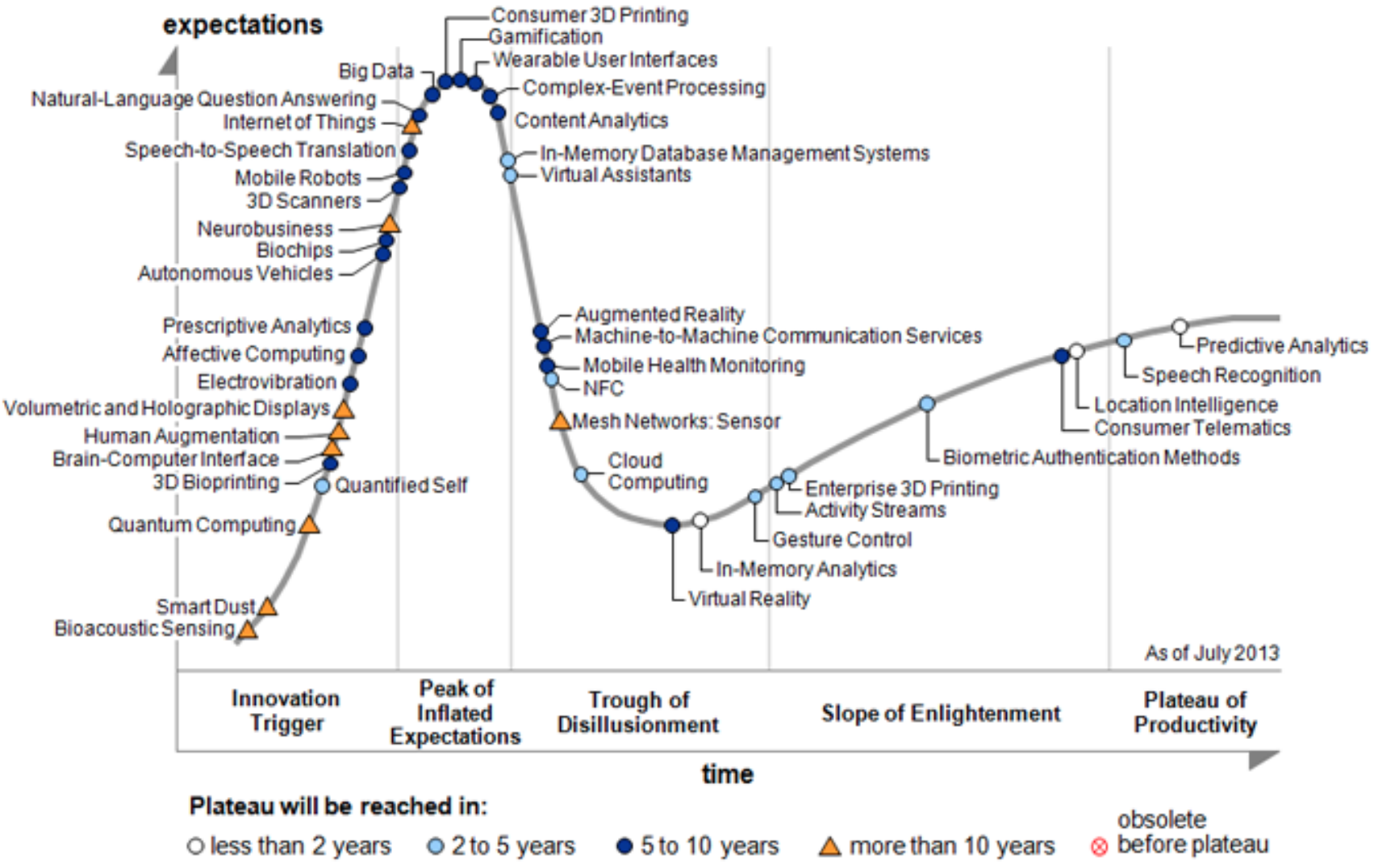
1）数据更加丰富，有分析价值，从TB到PB

2）分析工具更加强大，成本够低，MapReduce

3）互联网商业上的成功，引起重视，麦肯锡报告



大数据在计算机科学中处于最前沿



大数据维护安全

大数据改变未来战争

- 美国大数据研究和发展计划、欧盟Horizon 2020计划都把大数据提到了国家安全战略层面
- 数字主权是继海、陆、空、天四空间之后另一个大国博弈的空间
- 基于海量数据分析决策的“近传感器计算”将成为未来战争的典型形态

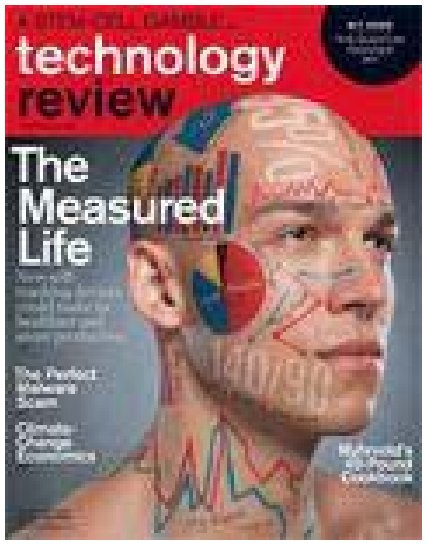
大数据摧毁暴力恐怖

- 阿富汗反恐战争中针对每股恐怖分子的全方位情报侦监系统每天产生数据量平均达到53T
- 美国国家安全局局长亚历山大在众议院特设情报委员会听证会时指出，通过“棱镜”等监视项目所获得的情报数据及相应分析，美国政府至少防止和挫败了50起恐怖袭击事件

大数据维护公共安全

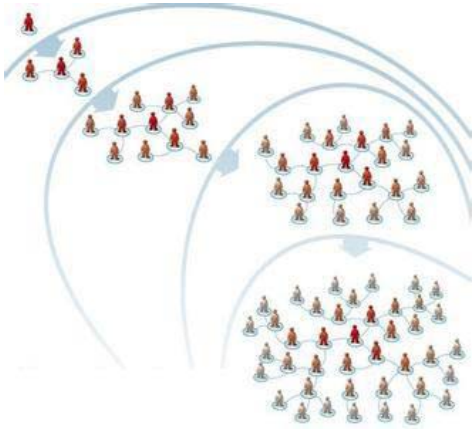
- 通过大数据采集分析，监测异常行为，发现和制止不法行为
- 通过大数据分析，提高犯罪行为实施前的预防能力和实施后的出警效率
- 通过大数据分析，提高刑侦队伍的破案率

大数据改善民生



大数据辅助健康管理

- 个人基因测序数据可以对已患疾病进行针对性治疗，对可能疾患进行提前预防
- 非干预穿戴设备通过实时采集脉搏、血压、体表导电率、压力等等指标对预警突发疾病、实时监控个体健康情况，为残疾人、老年人、婴幼儿和特定疾病患者提供实时的个性化服务
- 通过对诊疗过程数据的分析，可以为初级医院疑难病例的治疗提供智能决策辅助、发现患者骗保行为、监测医院、诊室甚至个别医生不正常的过度医疗和用药行为



大数据实现个性教育

- 加拿大Student Success Systems 基于学生个体数据分析给出发展状况评估、学业成长预测和个性化引导方案
- 美国DreamboxLearning 和MyLab根据不同学生在线学习的情况，设计个性化自适应的学习方案
- 大数据最终帮助形成在定量化基础上的教学引导和教学管理

报告提纲

大数据的发展

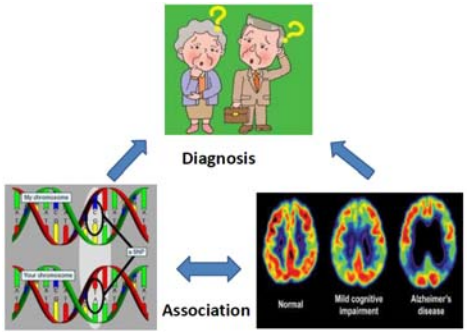
大数据分析面临的挑战

大数据机器学习平台

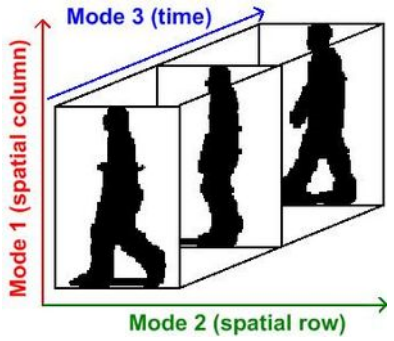
挑战

1. 数据量大且复杂，而分类数据太少，如何充分利用对未分类数据的质量分析来提高分类算法性能？

2. 多源异构数据语义丰富，如何构建融合多源数据的泛化模型？ 或发现多源数据间的关联关系？



3. 复杂数据对象存在多个方面，如何通过数据分析来刻画多个方面之间的相互关系？



挑战一: 未分类数据多样性

未分类数据具有无序性，**分布多样性**等特点-- 相同分布 或弱相关、结构相似、有杂质、高位等。

已分类数据

大象

犀牛

大量未分类数据

研究意义

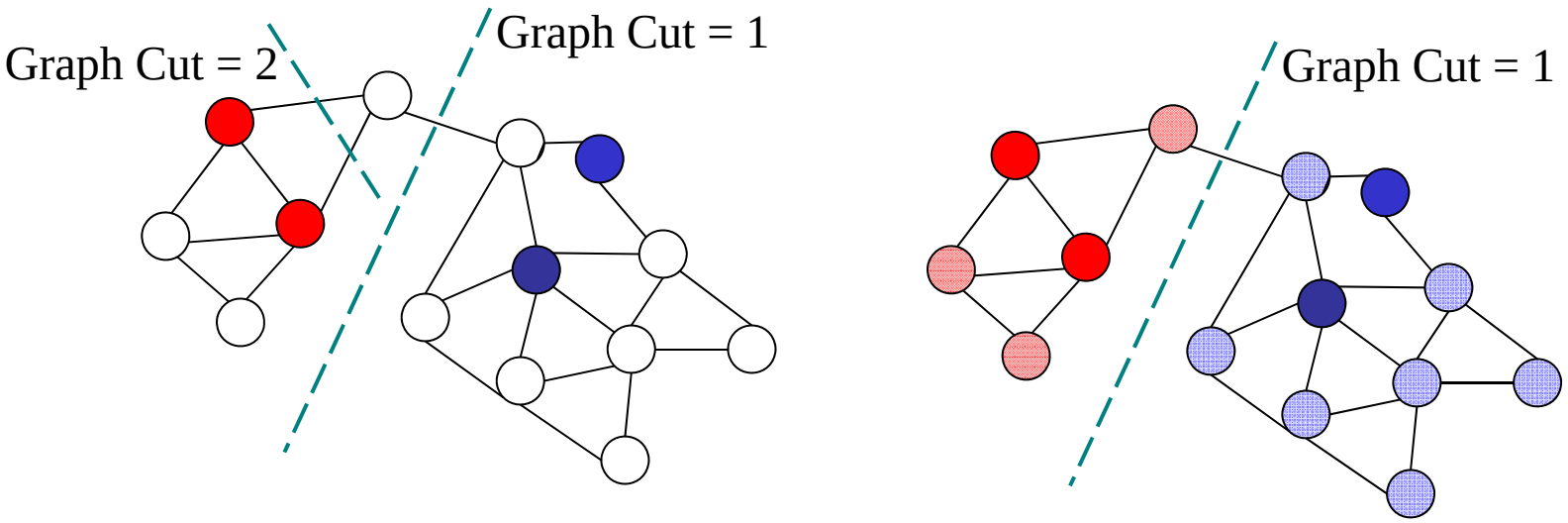
- 提高数据分类的准确率
- 节省专家对数据标记的成本

难点所在

- 未分类数据的复杂性和多样性
- 数据的高维度

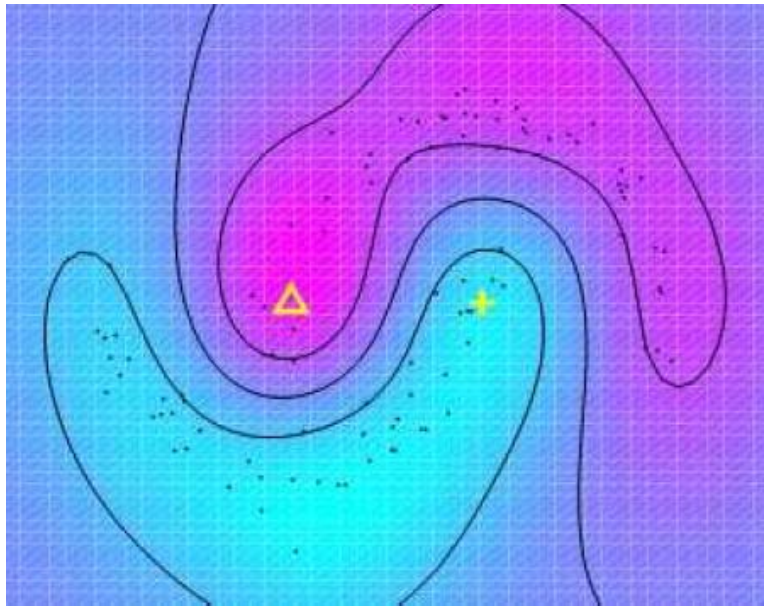
半监督学习示例：基于图的算法

- 利用图的性质(Graph Laplacian)对图进行分割：



半监督学习示例：半监督支持向量机

- S3VM的原理是在照顾已分类数据的情况下，保证相对于未分类数据的决策面边界最大，且决策面应尽量穿过低密度区域。



未分类数据分布多样性建模

- 相同分布 → Semi-supervised Learning
  - Xu Z., et al (2007), Efficient convex relaxation for transductive support vector machine. NIPS
- 分布有差异 → Covariance-shifting
- 存在弱相关关系 → Adaptive Regularization
  - Xu Z., et al (2009), Adaptive regularization for transductive support vector machine. NIPS.
- 结构上存在相似关系 → Self-taught Learning
  - Huang K., Xu Z., et al.(2009), Supervised self-taught learning: Actively transferring knowledge from unlabeled data, IJCNN .
- 好的数据与不相关数据的混合 → Generalized semi-supervised learning
  - Huang K., Xu Z., et al (2008), Semi-supervised learning from general unlabeled data. ICDM.
- Lable不足，无unlabeled data→ Active Semi-supervised Learning
  - Xu Z., et al (2008). Semi-supervised text categorization by active search. CIKM.
- 维度太高 → Semi-supervise d feature selection
  - Xu Z., et al (2010), Discriminative semi-supervised feature selection via manifold regularization. IEEE TNNLS.

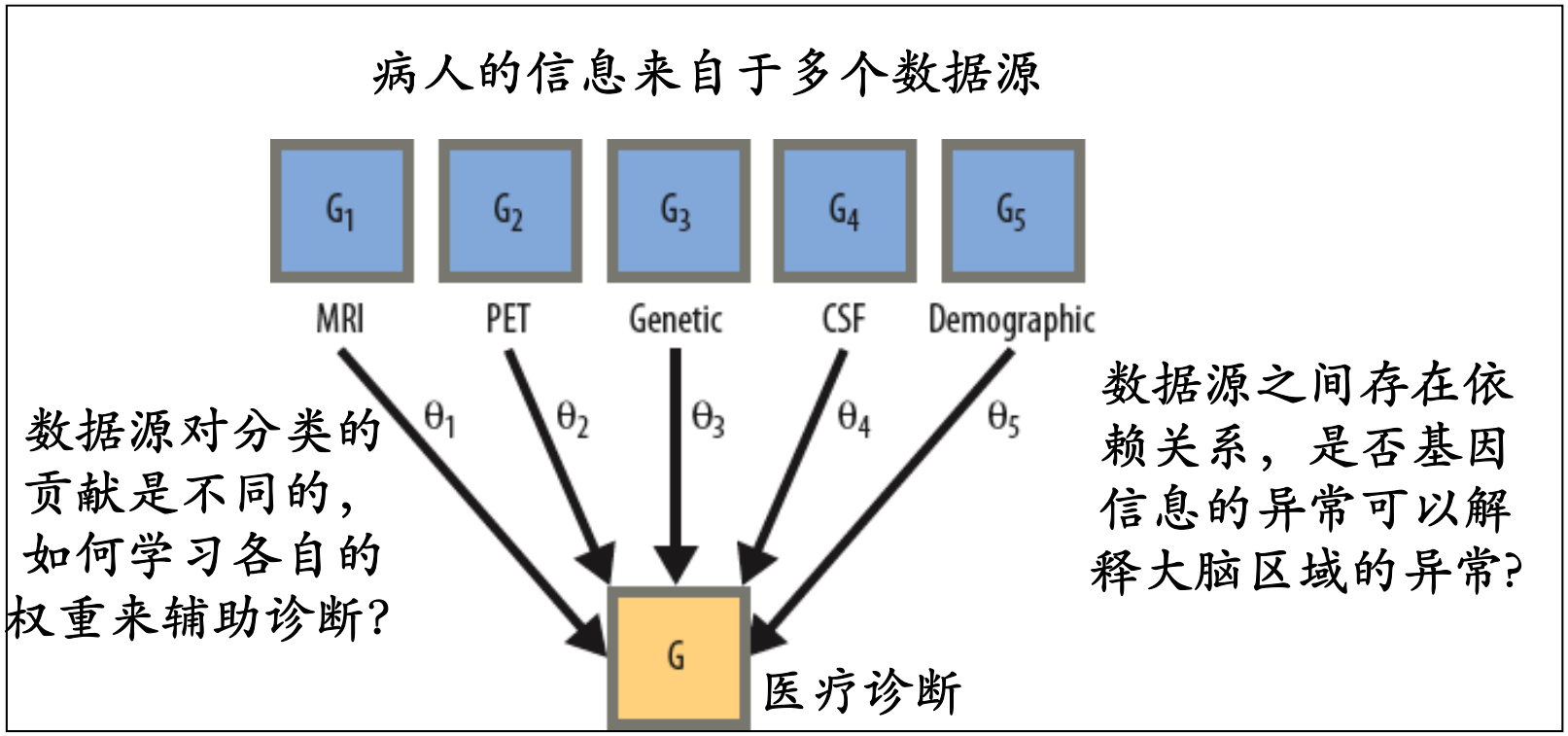
挑战二: 多源异构数据建模

研究意义

- 结合多个数据源的互补信息来提高数据分类的准确率
- 发现数据源之间的关联关系

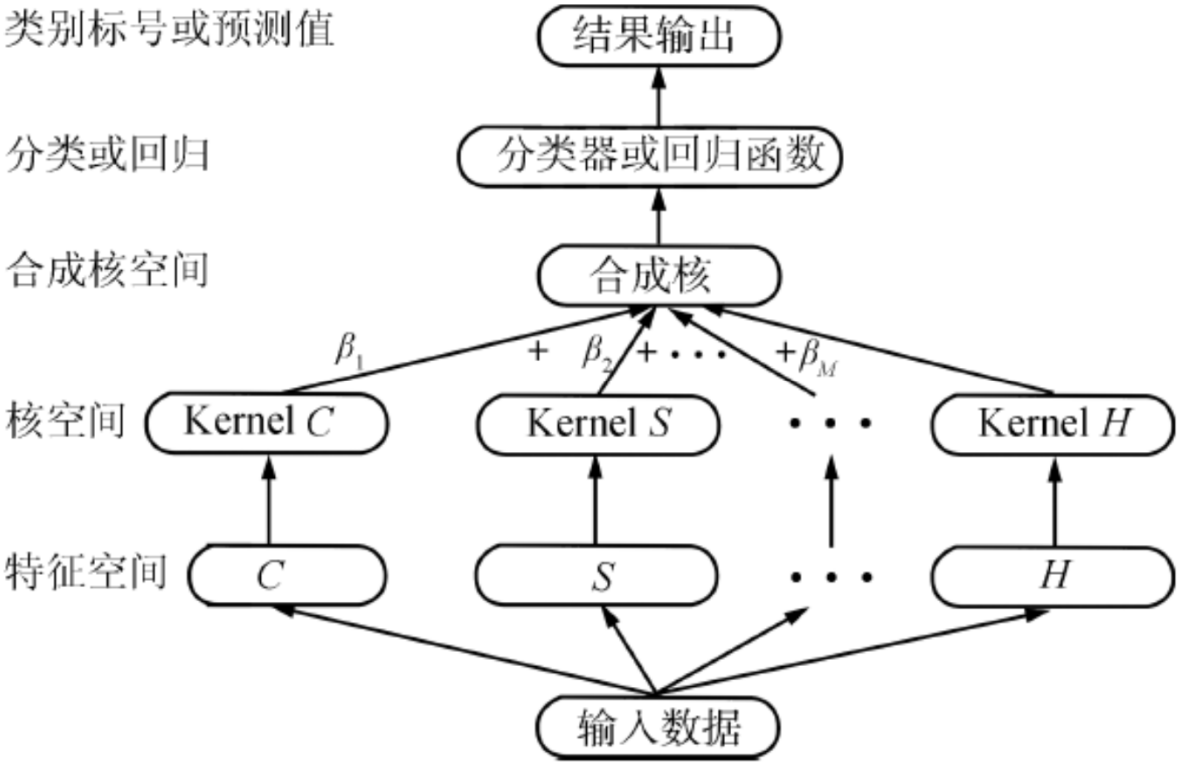
难点所在

- 多数据源的组合方式的多样性
- 数据源的异构性和不确定性



多视角学习样例：多核学习

多核学习算法 -- 学习数据源（子空间）之间的权重



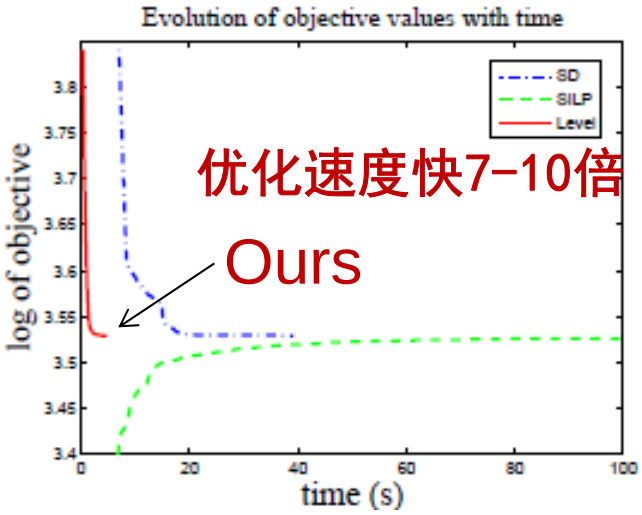
Sun et al (2010), ACTA AUTOMATICA SINICA

代表性工作：多核学习优化算法

多核学习算法优化

问题：优化过程中，传统方法或者没有对过去的梯度进行正则化，或者没有使用历史梯度。

方案：提出了一种基于Level Set的快速多核学习算法，其利用历史梯度，并将当前解投影到Level Set 当中来进行正则化。

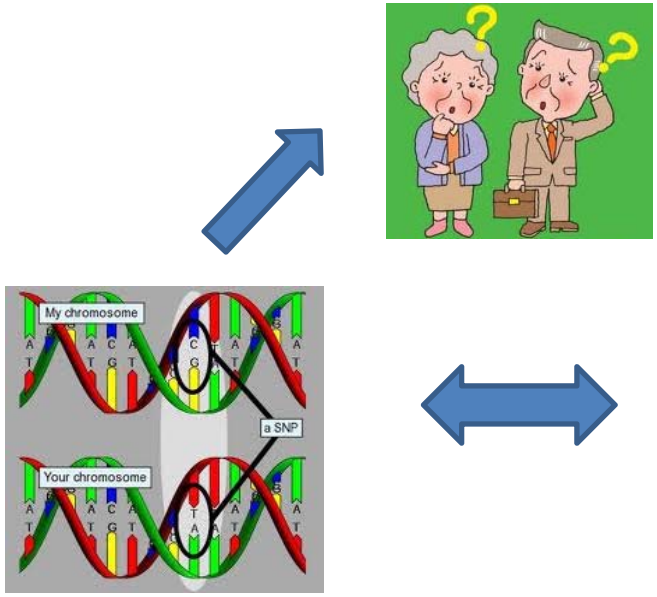


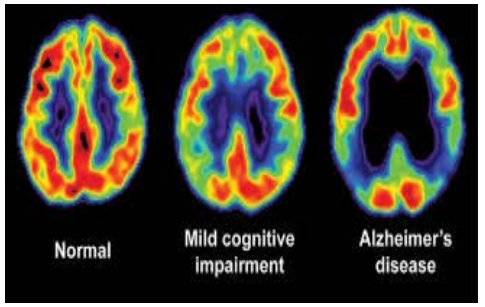
- 稀疏泛化多核学习
- 使用Group Lasso和多核学习之间的等价关系

Z. Xu, R. Jin., et al (2009), *NIPS*
Xu Z., et al (2010), *ICML*
Yang H., Xu Z., et al (2011), IEEE TNNLS

应用：Alzheimer疾病的关联分析

提出一个异构多视角学习算法。该算法基于隐变量模型，对数据源之间的共性和差异性进行建模。





Genetic variations (discrete)

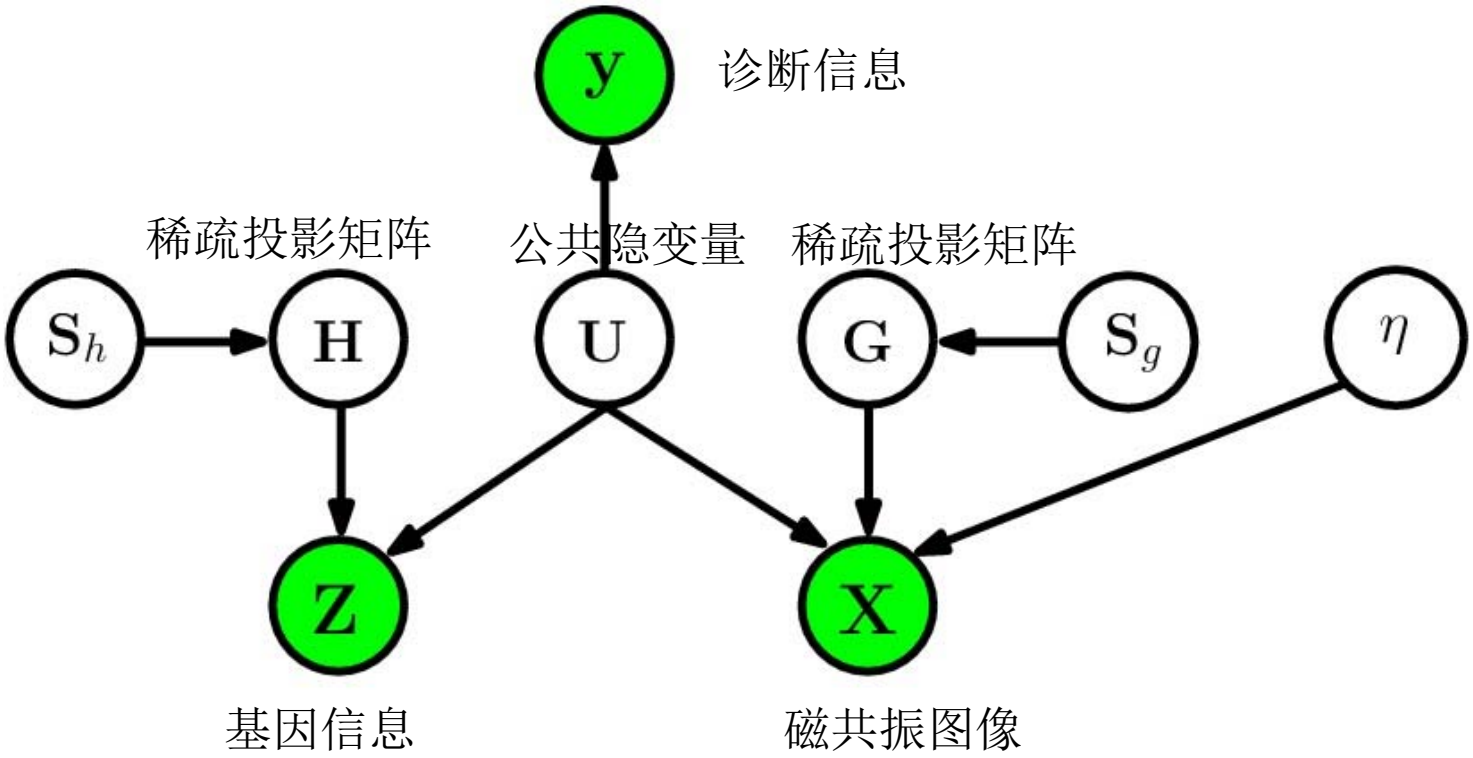
Intermediate phenotypes (continuous)

Zhe S., Xu Z., et al (2014), *PSB*

Zhe S., Xu Z., et al (2015), *AAAI*

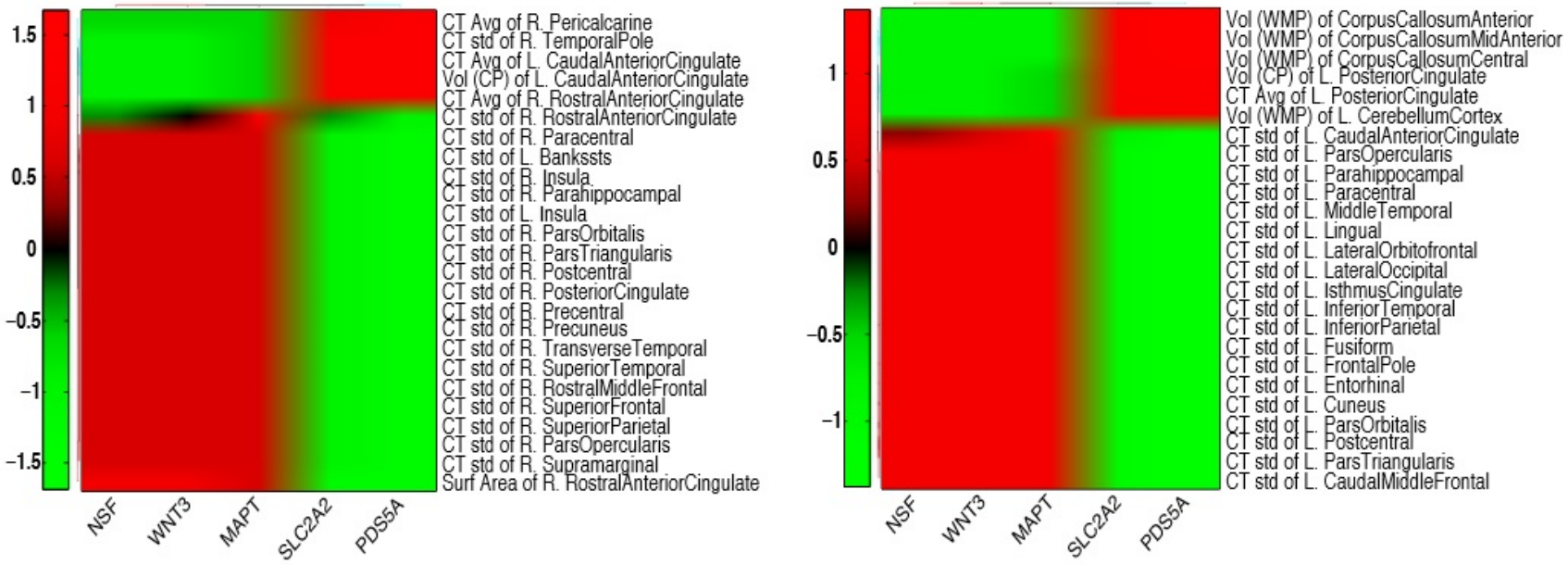
应用：Alzheimer疾病

图模型



应用：Alzheimer疾病

- 大脑区域与基因的相关关系



ADNI 数据库

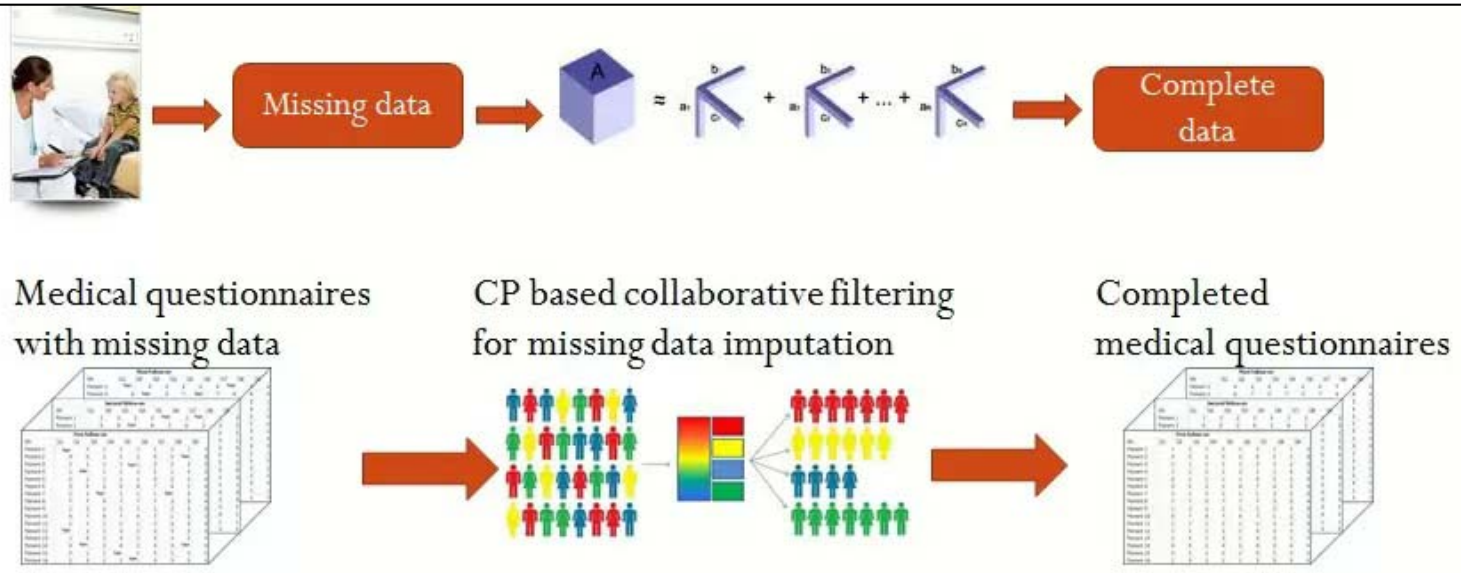
挑战三: 多元关系建模

研究意义

- 现实世界存在大量多元关系
- 张量分析抽象描述数据多个方面之间的交互机制

难点所在

- 多方面交互机制的不确定性
- 数据的异构性和复杂性



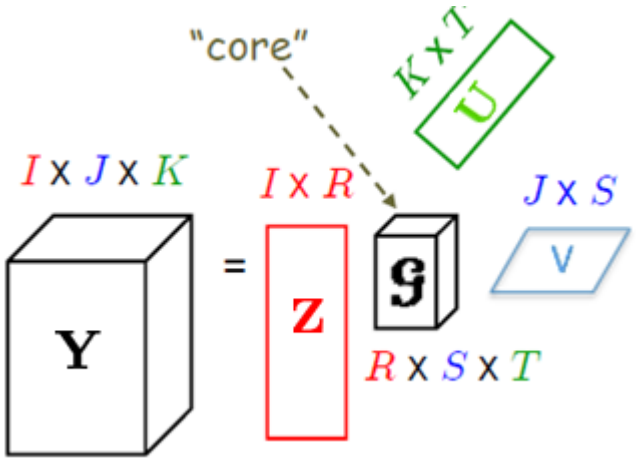
（参与者，问题，答案选项）形成的张量

张量因子分析

完整的张量

代表性工作：非线性张量分解

张量分解算法



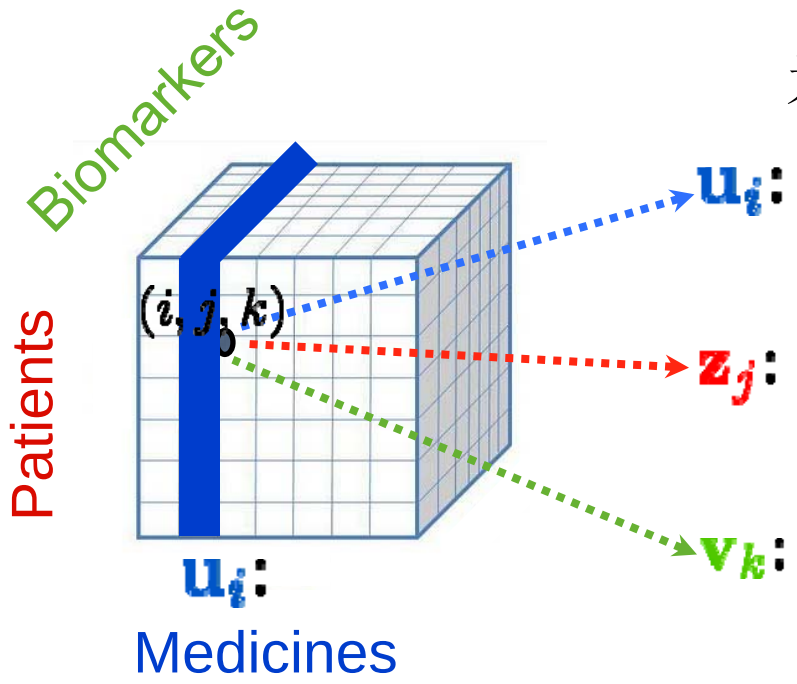
问题：传统方法多为线性分解算法，且不能处理非连续数据

方案：**提出了一种基于高斯过程隐变量模型的张量分解算法，**能处理不同数据类型及缺失值。

Z Xu ., et al (2012), ICML， Chen S.., Xu Z., et al (2013), NIPS

代表性工作：非线性张量分解

定义在张量上的稀疏隐高斯过程



元素 (i, j, k)

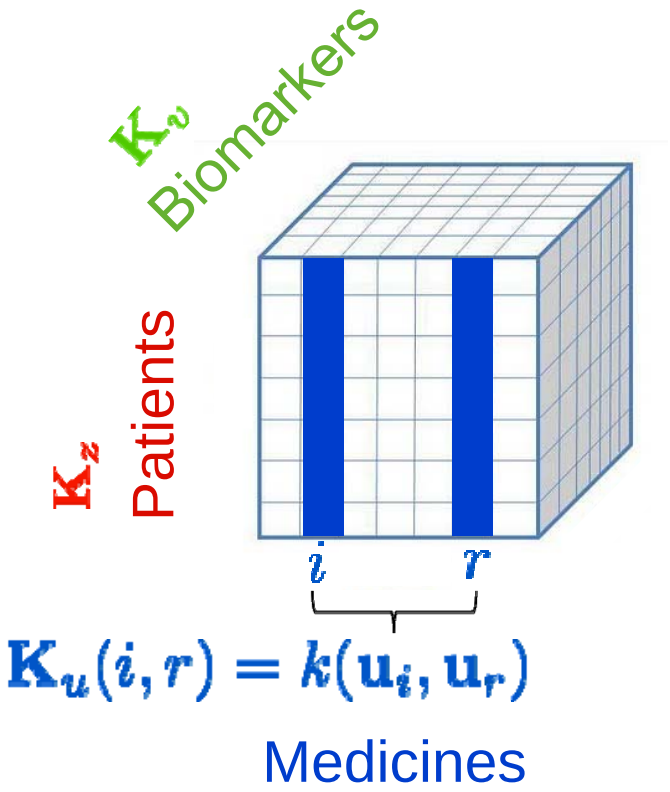
$\mathbf{u}_i$: 药品组的隐权重向量

$\mathbf{z}_j$: 病人组的隐权重向量

$\mathbf{v}_k$: 生物特征组的隐权重向量

代表性工作：非线性张量分解

协方差矩阵表示相似度



- 每一维采用单独的协方差/核函数
- 隐向量越相似，协方差越大

药品i 和r之间的非线性关系

代表性工作：非线性张量分解

张量上的高斯过程

- 无限张量空间中的随机过程
- 任意确定大小的张量的分布都是基于张量的高斯分布

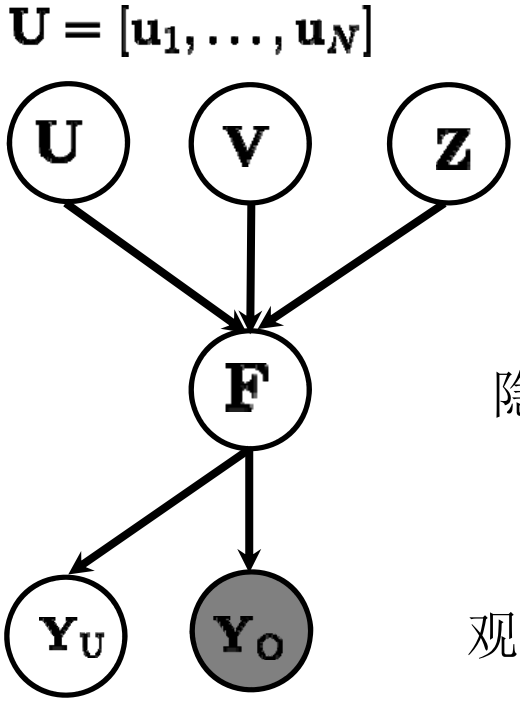
Tensor

$$\mathcal{N}(\mathbf{F}|\mathbf{0};\mathbf{K}_u,\mathbf{K}_v,\mathbf{K}_z) = (2\pi)^{-\frac{3N}{2}} \prod_{s=u,v,z} |\mathbf{K}_s|^{-\frac{N^2}{2}}$$

$$\cdot \exp\{-\frac{1}{2}\|(\mathbf{F} \times_1 \mathbf{K}_u^{-1} \times_2 \mathbf{K}_v^{-1} \times_3 \mathbf{K}_z^{-1}) \circ \mathbf{F}\|^2\}$$

代表性工作：非线性张量分解

图模型表示



稀疏因子

$$\mathbf{u}_i \sim \exp(-\lambda|\mathbf{u}_i|)$$

同样对 $\mathbf{Z}$和 $\mathbf{V}$采样

隐张量

$$\mathbf{F} \sim \mathcal{N}(0; \mathbf{K}_u, \mathbf{K}_v, \mathbf{K}_z)$$

观察数据

$$\mathbf{Y}_O \sim \prod_{(i,j,k)\in O} p(y_{ijk}|f_{ijk})$$

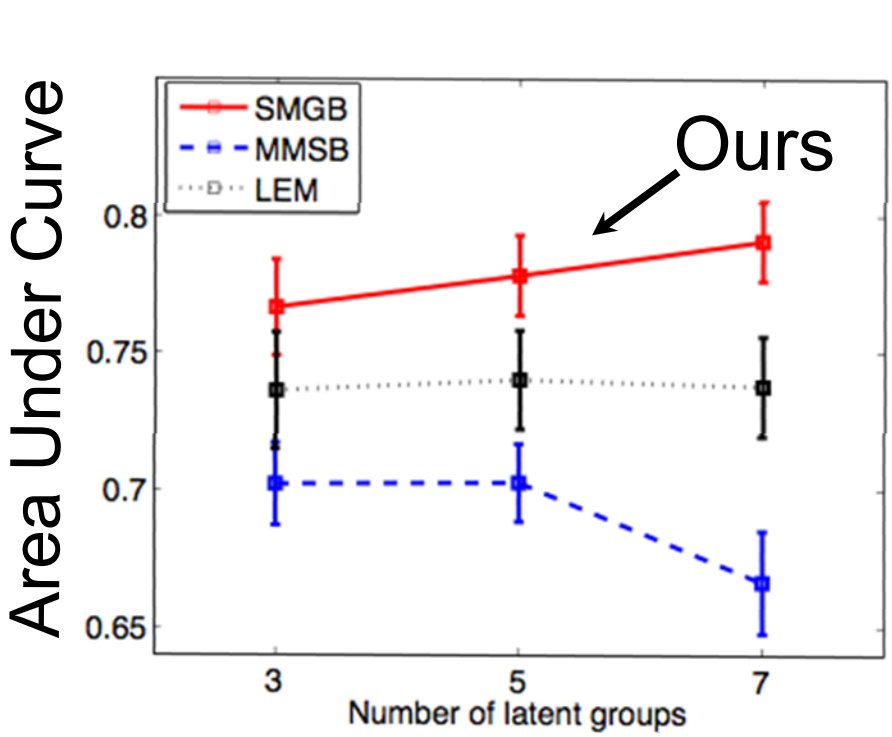
未知数据

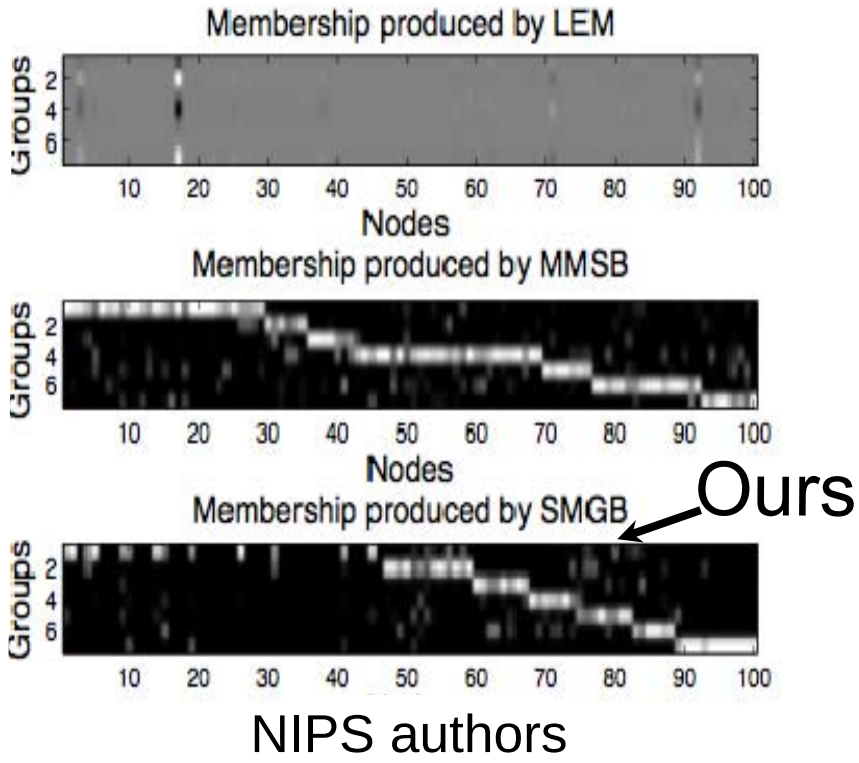
$p(y_{ijk}|f_{ijk})$: 高斯→ 连续数据

Probit → 二进制/有序数据

Possion → 数量

2D应用: Coauthor 网络

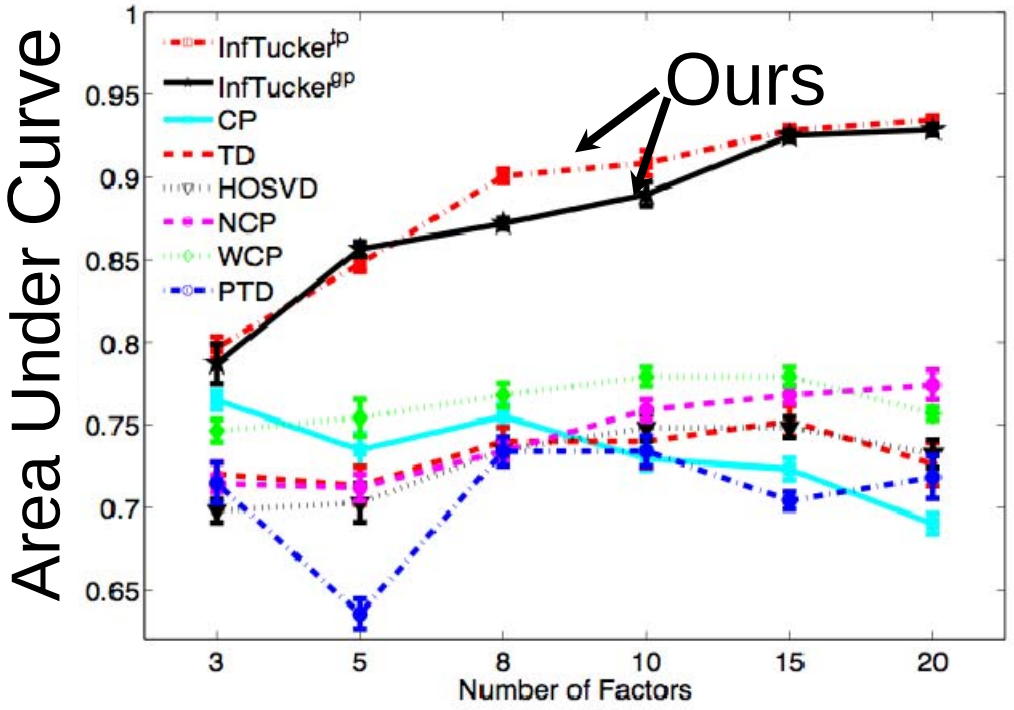




Co-authorship 数据库: co-authorship 链接统计于NIPS 1-17中100位合作最多的作者.

3D应用: 安然(Enron) 邮件

- **安然数集:**
  - **2001**年 破产前安然公司高级管理层的往来邮件.
- **3D** 张量表达:
  - 发送方-接收方-邮件主题



报告提纲

大数据的发展

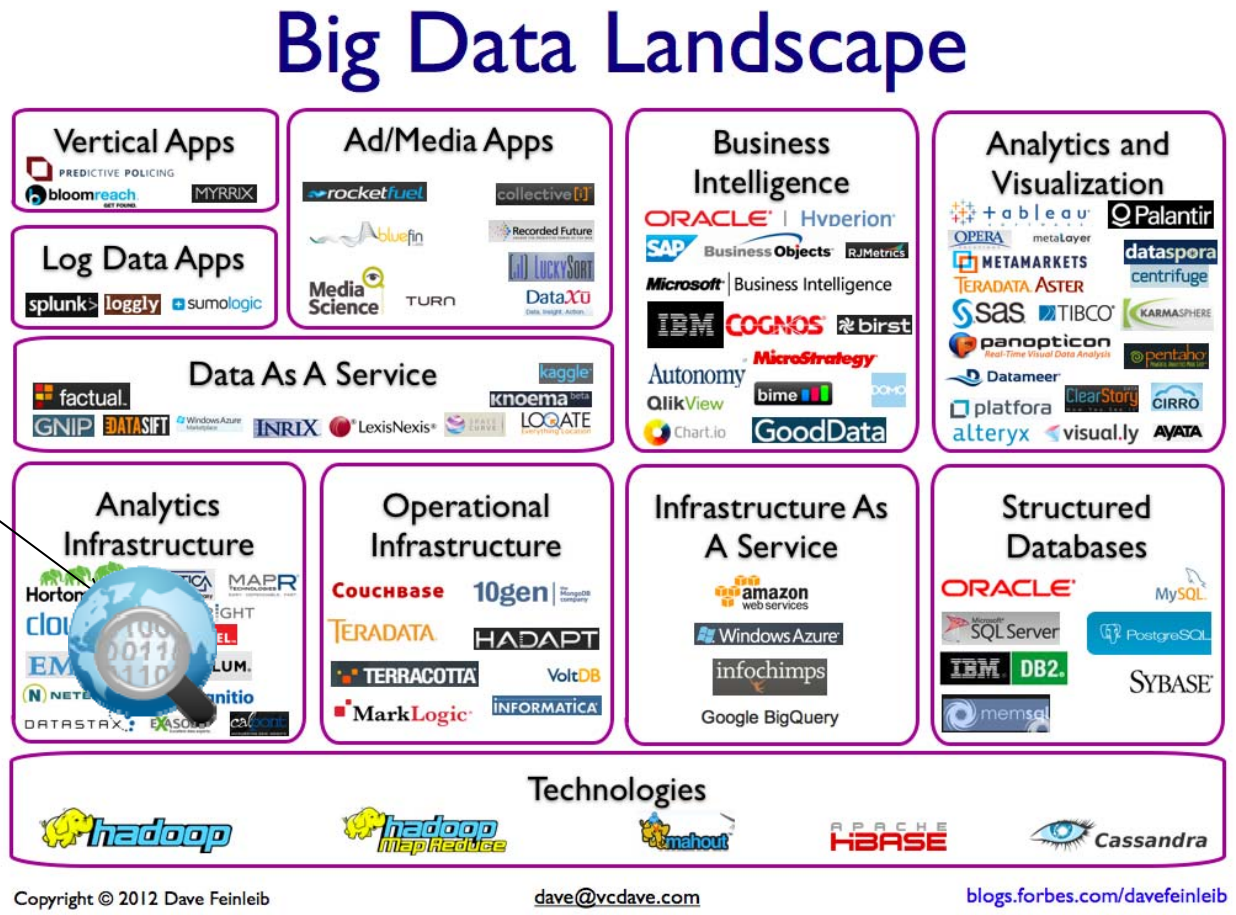
分析大数据面临的挑战

大数据机器学习平台

大数据机器学习平台核心技术研究

机器学习

统计学



国际上高校与企业
在机器学习平台方面研究的最新进展

- Single Machine: GraphChi, TurboGraph
- Map-Reduce
- GraphLab
- Skytree
- Spark

大数据机器学习平台核心技术研究正刚刚开始！

Hadoop

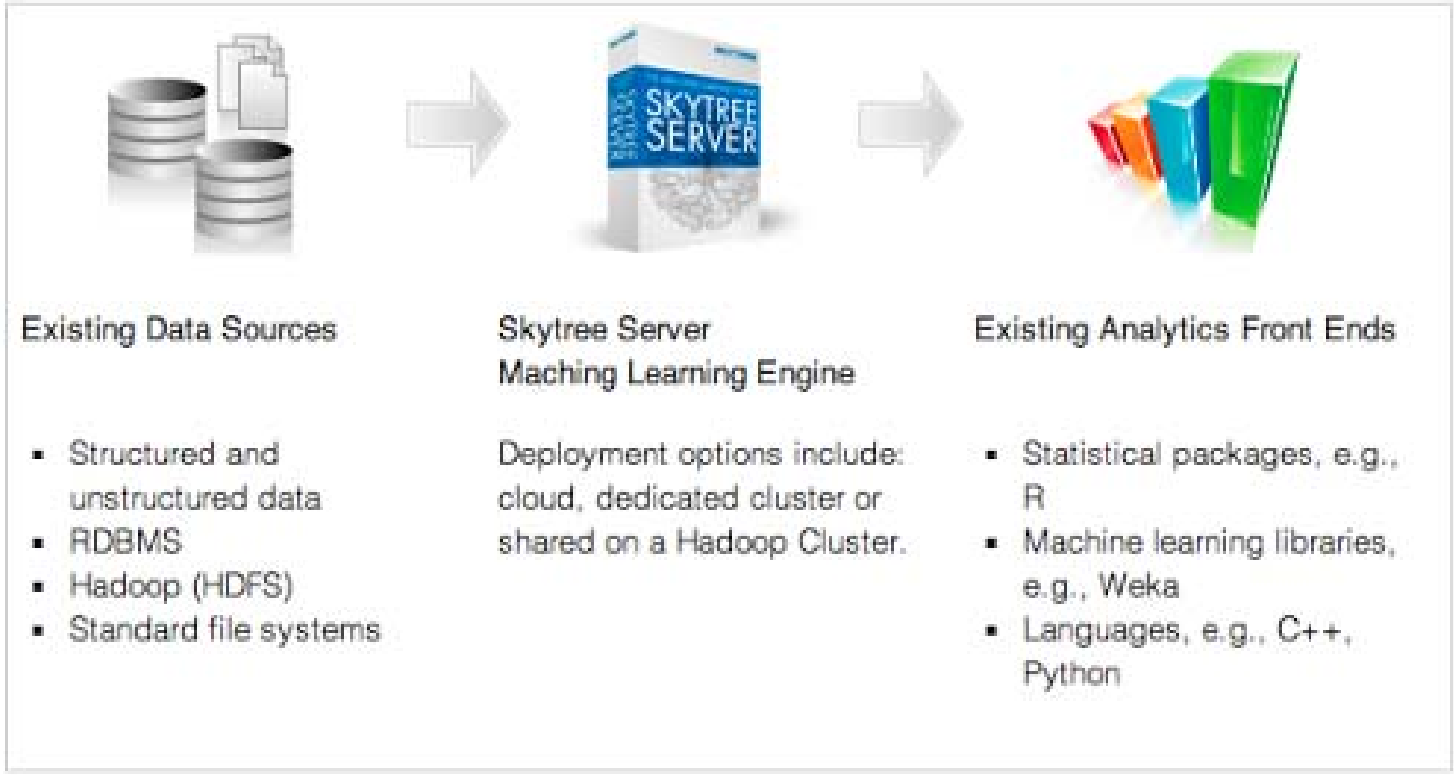
Hadoop是一个由Apache基金会所开发的分布式并行计算框架。其框架最核心的设计是：MapReduce算法和分布式文件系统HDFS。

Hadoop主要的特点：

- 扩容能力：能可靠地存储和处理千兆字节（PB）数据。
- 成本低：可以通过普通机器组成的服务器群来分发以及处理数据。服务器群总计可达数千个节点。
- 高效率：通过分发数据，Hadoop可以在数据所在的节点上并行处理数据。
- 可靠性：Hadoop能自动维护数据的多份复制，并且在任务失败后能自动地重新部署计算任务。

Skytree

SKYTREE BIG DATA ANALYTICS



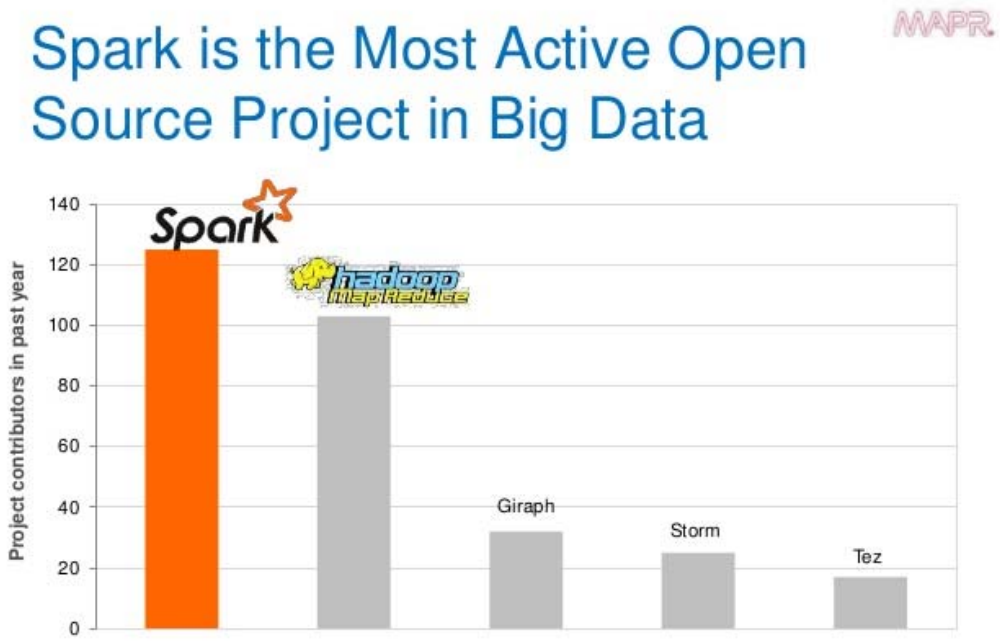
Skytree是Skytree公司开发的机器学习平台，它结合先进的机器学习算法，为企业提供大数据高级分析，目前，它已用于推荐系统，异常识别，预测分析，聚类，市场细分，以及相似性搜索。

Hadoop 缺点

- 效率低下：任务内串行、链式浪费严重、中间结果不可分享
- 不友好算法：
  - 数据连接操作
  - 基于图的算法
  - 需要多轮迭代、循环的算法（如：矩阵SVD分解）
- 编程复杂

Spark

Spark是UC Berkeley AMP lab所开源的通用并行计算框架，基于map reduce算法实现的分布式计算，拥有Hadoop MapReduce所具有的优点。



Spark与Hadoop对比：

- Spark的中间数据放到内存中，对于迭代运算效率更高。更适合于迭代运算比较多的机器学习和数据挖掘运算。
- Spark比Hadoop更通用：Spark提供的数据集操作类型有很多种，给开发上层应用的用户提供了方便。
- Spark对于增量修改的应用模型不适合。
- Spark：在分布式数据集计算时通过checkpoint来实现容错。
- Spark通过提供丰富的Scala, Java，Python API及交互式Shell来提高可用性。

GraphLab

GraphLab是CMU的Select实验室提出的一个面向大规模机器学习/图计算的分布式内存计算框架。GraphLab构建了四种流行的机器学习算法的并行版本：

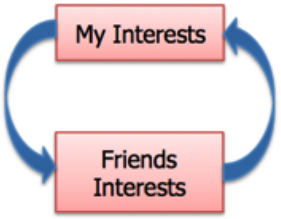
- 一个基于图的数据模型，模型展现了数据和计算过程中的依赖。
- 一组并行访问的模式来保证一系列的并行一致性。
- 一种复杂的模块调度机制。
- 它实现和通过实验评估参数学习和图形算法模型的推论。

Properties of Computation on Graphs

Dependency Graph

Local Updates

Iterative Computation



Map-Reduce for Data-Parallel ML

- Excellent for large data-parallel tasks!

Map Reduce

Feature Extraction

Cross Validation

Computing Sufficient Statistics

GraphLab

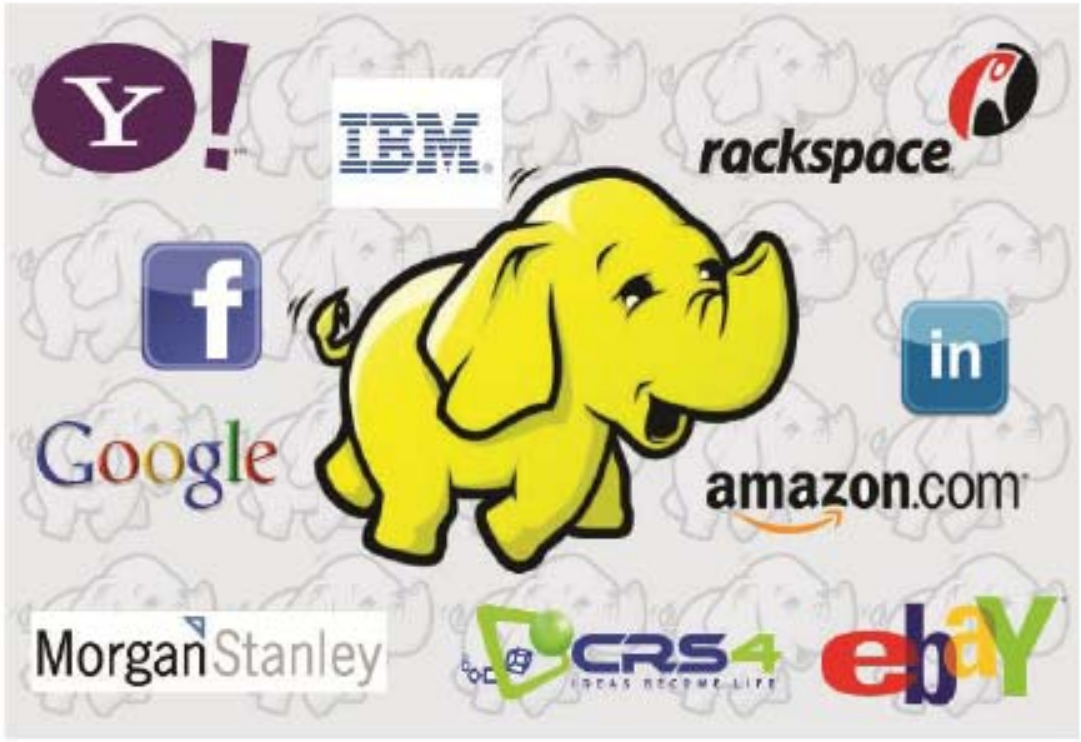
Graphical Models
Gibbs Sampling
Belief Propagation
Variational Opt.

Semi-Supervised Learning
Label Propagation
CoEM

Collaborative Filtering
Tensor Factorization

Data-Mining
PageRank
Triangle Counting

大数据平台应用



SKYTREE
BIG DATA ANALYTICS

Brookfield
RPS

SETI INSTITUTE

eHarmony

adconion
MEDIA GROUP

USGA

CANFAR
Canadian Advanced Network for Astronomical Research

Spark

YAHOO! 淘 Alibaba.com cloudera

GraphLab

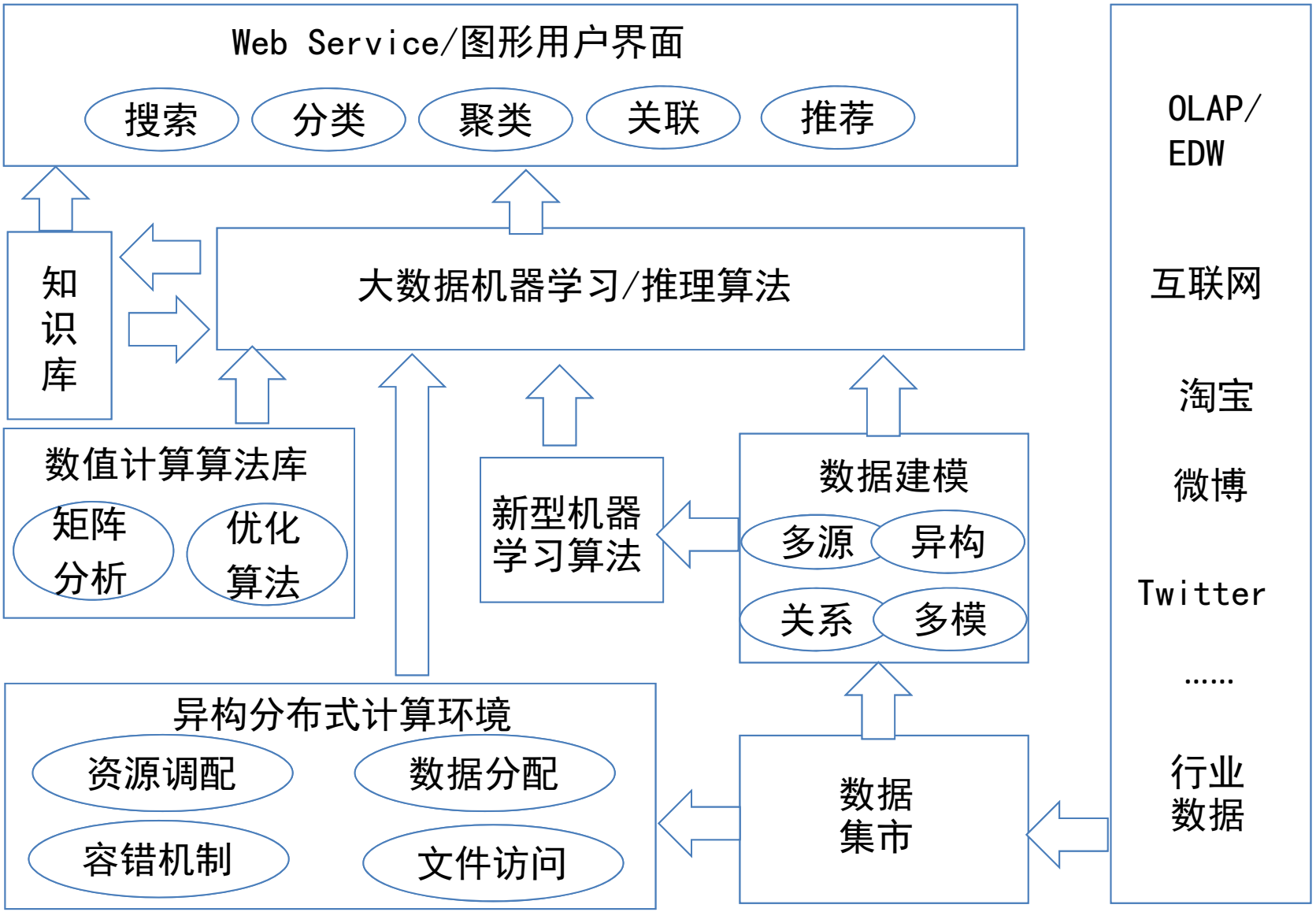
- Retail
- Financial Services
- Oil and Gas
- Health Care
- Marketing and Advertising
- Government

大数据机器学习平台核心技术问题

挑战：如何针对数据学习算法自身特点，设计算法性能最大化的大数据平台？

- **机器学习算法对大数据平台提出了挑战**
  - 数据采样与特征选择
  - 参数验证选择
  - 算法多次迭代
  - 算法性能的理论保证
  - 领域知识的运用
- **适合机器学习任务的大数据基础架构设计与优化**
  - 数据存储及访问
  - 任务分配及节点通信
  - 任务调度优化

大数据分析平台设计


Web Service/图形用户界面
搜索
分类
聚类
关联
推荐
知识库
大数据机器学习/推理算法
数值计算算法库
矩阵分析
优化算法
新型机器学习算法
数据建模
多源
异构
关系
多模
异构分布式计算环境
资源调配
数据分配
容错机制
文件访问
数据集市
OLAP/EDW
互联网
淘宝
微博
Twitter
……
行业数据

Questions?