



大数据导论

Introduction to Big Data



第1讲: 绪论

叶允明

计算机科学与技术学院
哈尔滨工业大学（深圳）

课程交流群

- QQ课程群
- Group number : 734659205
- Group name : 大数据导论

助教



- 贾鹏飞
- Email: 20s051024@stu.hit.edu.cn
- Tel: 18463102736



- 赵昕玥
- Email: zhaoxinyue@stu.hit.edu.cn
- 手机: 15118185102



- 陈志豪
- Email: standingbychen@qq.com
- 手机: 13510516506

助教



- 朱启重
- Email: 564621706@qq.com
- 手机: 18266878671



- 陈武桥
- Email: theoarcher2000@gmail.com
- 手机: 15814037160



- 姜昊
- Email: 849974258@qq.com
- 手机: 13613045354

关于这门课程的定位

课程参考资料

- 教案与论文
- 梅宏. 大数据导论. 高等教育出版社, 2018.11.
- 林子雨. 《大数据技术原理与应用(第2版)》. 人民邮电出版社, 2017.
- Pang-Ning Tan, Michael Steinbach, Anuj Karpatne , Vipin Kumar著; 段磊, 张天庆等译. 数据挖掘导论 (原书第2版) . 机械工业出版社, ISBN: 9787111631620, 2019-07-29.
- Jiawei Han, Micheline Kamber, Jian Pei著; 范明, 孟小峰等译. 数据挖掘: 概念与技术. 机械工业出版社, ISBN: 9787111391401, 2012.

课程内容

- 绪论
- 大数据存储与处理框架（Hadoop）
- 数据理解及数据预处理方法
- 大数据的分类与预测算法
- 大数据的聚类与离群点检测算法
- 大数据的关联规则挖掘及其应用

课程形式和要求

- 先修课程：高等数学、代数与几何、概率论与数理统计、高级语言程序设计
- 授课 & 实验
- 最终成绩:
 - 30%小作业
 - 30% 实验
 - 40% 大作业

从该课程你能学到什么？

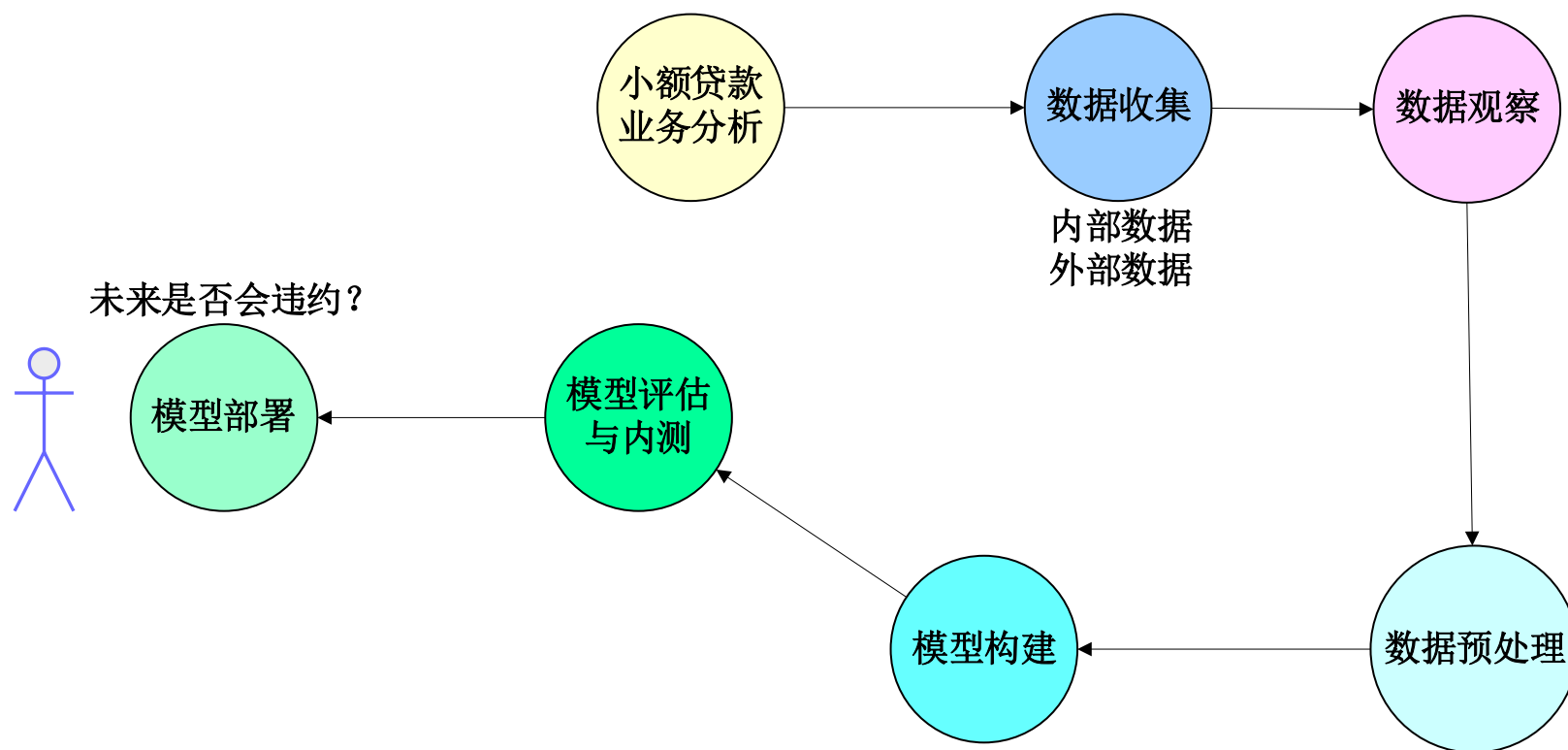
- 如何在实际应用中设计和实现大数据**项目**
 - 大数据项目作为一个过程或工作流的思想 (process or workflow)
- 经典大数据**算法**
 - 包括数据挖掘的经典算法
- 大数据软件**工具**
 - 开源工具、软件产品

从一个简单的大数据应用案例开始

- 一个互联网金融领域的大数据（数据挖掘）应用案例
- 问题：InterFin公司的智能客户准入模型构建及应用

| 序号 | 月收入 | 年龄 | 贷款比 | 未清贷款 | 已抵押资产 | 家属 | 债务比 | 违约 |
|------|-----|----|-----|------|-------|----|-----|-----|
| 1 | 高 | 67 | 低 | 高 | 中 | 中 | 中 | Yes |
| 2 | 中 | 41 | 高 | 低 | 低 | 中 | 中 | Yes |
| 3 | 中 | 46 | 高 | 低 | 中 | 中 | 中 | No |
| 4 | 低 | 49 | 高 | 低 | 低 | 低 | 高 | No |
| 5 | / | 65 | 中 | 低 | 低 | 中 | 高 | No |
| 6 | / | 58 | 低 | 中 | 低 | 高 | 高 | No |
| | | | | | | | | |
| 新客户： | 中 | 55 | 低 | 低 | 中 | 低 | 低 | ? |

InterFin智能客户准入项目的主要流程



第一讲：绪论

- 大数据的历史背景
- 大数据的应用领域
- 数据的定义及其类型
- 大数据技术概况
- 数据的来源及其获取方法简介
- 大数据领域的学习资源

大数据的历史与背景

大数据现象

● 人类社会数字化、信息化和网络化进程的快速发展

➤ 带来了各行各业数据的爆炸性增长!



我国网民数量居世界之首，每天产生的数据量也位于世界前列。

| | |
|-------------|---|
| 淘宝网站 | ◆ 单日数据产生量超过 5万GB ◆ 存储量 4000万GB |
| 百度公司 | ◆ 目前数据总量 10亿GB ◆ 存储网页 1万亿页 ◆ 每天大约要处理 60亿次 搜索请求 |
| 一个8Mbps的摄像头 | ◆ 一小时能产生 3.6GB 的数据 ◆ 一个城市每月产生的数据达 上千万GB |
| 医院 | ◆ 一个病人的CT影像数据量达 几十GB ◆ 全国每年需保存的数据达 上百亿GB |

大数据是什么

“3V” 定义

维基百科给出的定义：

大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。

规模性 (Volume)

多样性 (Variety)

高速性 (Velocity)

价值性 (Value) (IDC)

真实性 (Veracity) (IBM)

“4V” 定义

大数据领域的发展历程

- 大规模数据的处理与分析技术已发展多年，一直是研究热点。但量变会引起质变！
- 2007 年 1 月，图灵奖得主JimGray 指出：科学的发展正在进入“数据密集型科学发现范式”——科学史上的“**第四范式**”
- 《自然》杂志2008年9月出版一个关于大数据的专刊。



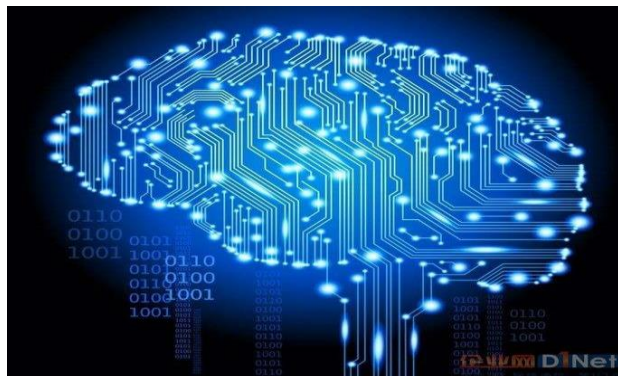
大数据与人工智能

- 国务院：《新一代人工智能发展规划》，国发〔2017〕35号

| 专栏1 基础理论 |
|---|
| 1. <u>大数据智能理论</u> 。研究数据驱动与知识引导相结合的人工智能新方法、以自然语言理解和图像图形为核心的认知计算理论和方法、综合深度推理与创意人工智能理论与方法、非完全信息下智能决策基础理论与框架、数据驱动的通用人工智能数学模型与理论等。 |

- 目前最成功的人工智能应用领域： **大数据智能、大数据机器学习！**

➤ 深度学习需要大数据支撑！



大数据与数据挖掘

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
- 1991-1994 Workshops on Knowledge Discovery in Databases
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD' 95-98)
 - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- ACM Transactions on KDD starting in 2007
- 2008: “大数据”新的术语
- 数据挖掘: 从海量数据中发现“有趣的”的模式或知识
(non-trivial, implicit, previously unknown and potentially useful)

大数据的应用领域

商业智能应用：决策支持

- 数据分析与决策支持

- 市场分析与管埋

- ✓ 精准营销、客户关系管理(CRM)、购物篮分析、交叉销售、市场细分

- 风险分析与管埋

- ✓ 预测（人、财、物）、客户维系、质量控制、竞争分析

- 诈骗检测与异常模式发现

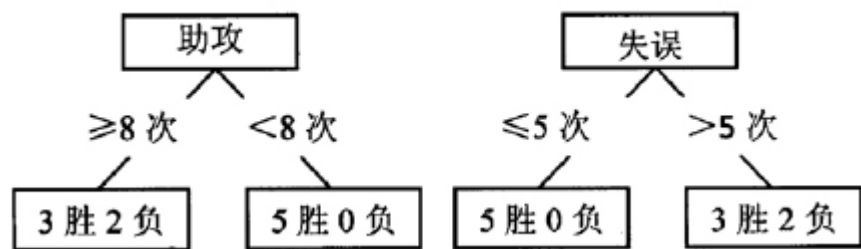
商业智能应用：推荐系统

- 应用领域：电商、信息推荐、电影、音乐等
- 目的：预测用户对商品是否喜欢、喜欢程度、个性化服务



体育应用：篮球针对性训练

- 对运动员成长轨迹进行深度挖掘、建模
- 找出运动员的“短板”与“长版”
- 加强对特长点和薄弱点的训练



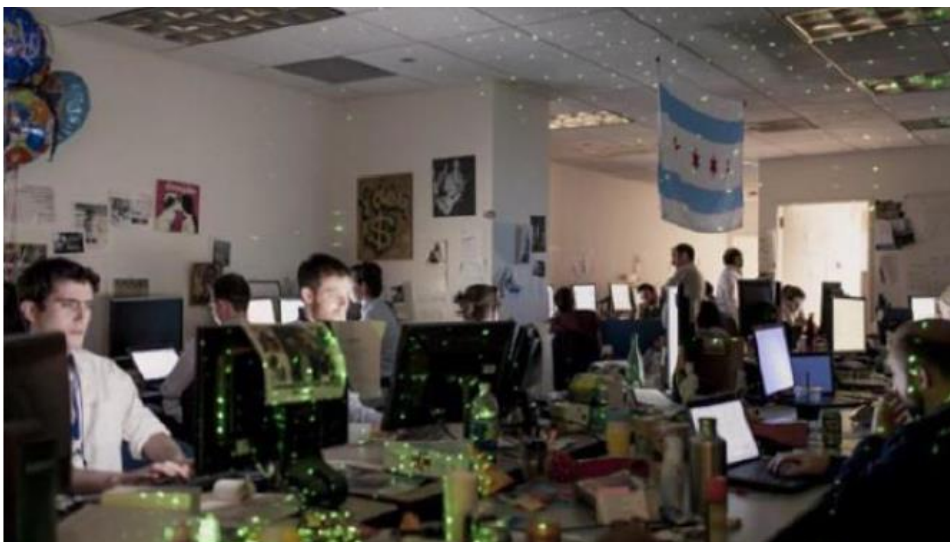
➤ 通过训练加强运动员助攻次数和减少失误率

避免此类情况发生



政治应用：美国总统大选

- 在总统候选人的第一次辩论之后，他们分析出哪些选民将倒戈，为每位选民找出一个最能说服他的理由
- 通过一些复杂的模型来精准定位不同选民，购买了一些冷门节目的广告时段，而没有采用在本地新闻时段购买广告的传统做法，广告效率相比2008年提高了14%
- 向奥巴马推荐，竞选后期应当在什么地方展开活动——那里有很多争取对象
- 借助模型帮助奥巴马筹集到创记录的10亿美元



大数据团队



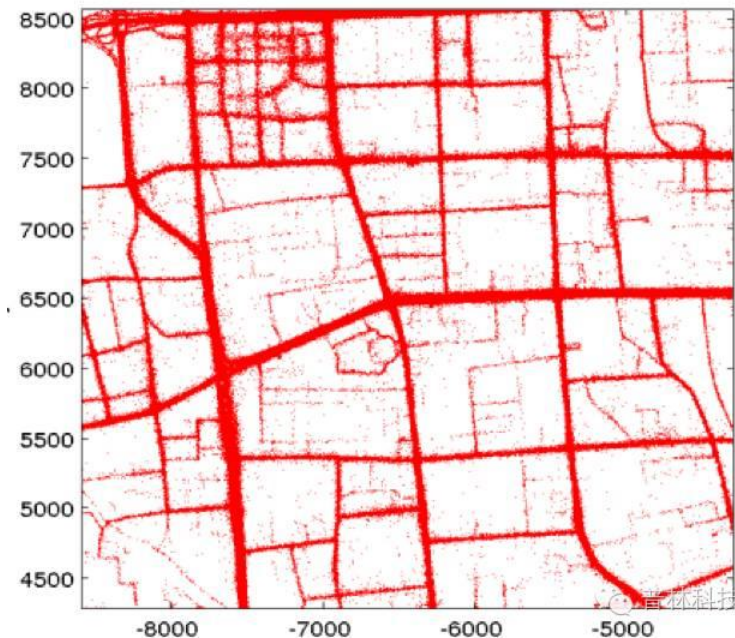
交通应用：拥堵预测

● 建立历史交通数据库

- 道路信息：从GIS数据库中导出北京市路网数据，包括道路的起点终点，中轴线经纬度，道路等级，车道数目等等。
- 车辆信息：数据来源是北京市6万辆出租车每天的GPS数据，出租车每50s生成一条GPS信息。

● 未来时间段车速预测

- 影响交通因素：天气状况，车辆数量，交通事故等，但车速可以包含以上信息。
- 根据历史数据中最相似的情况，从而进行预测，最相似的车速曲线，未来时刻的变化也可能相似。



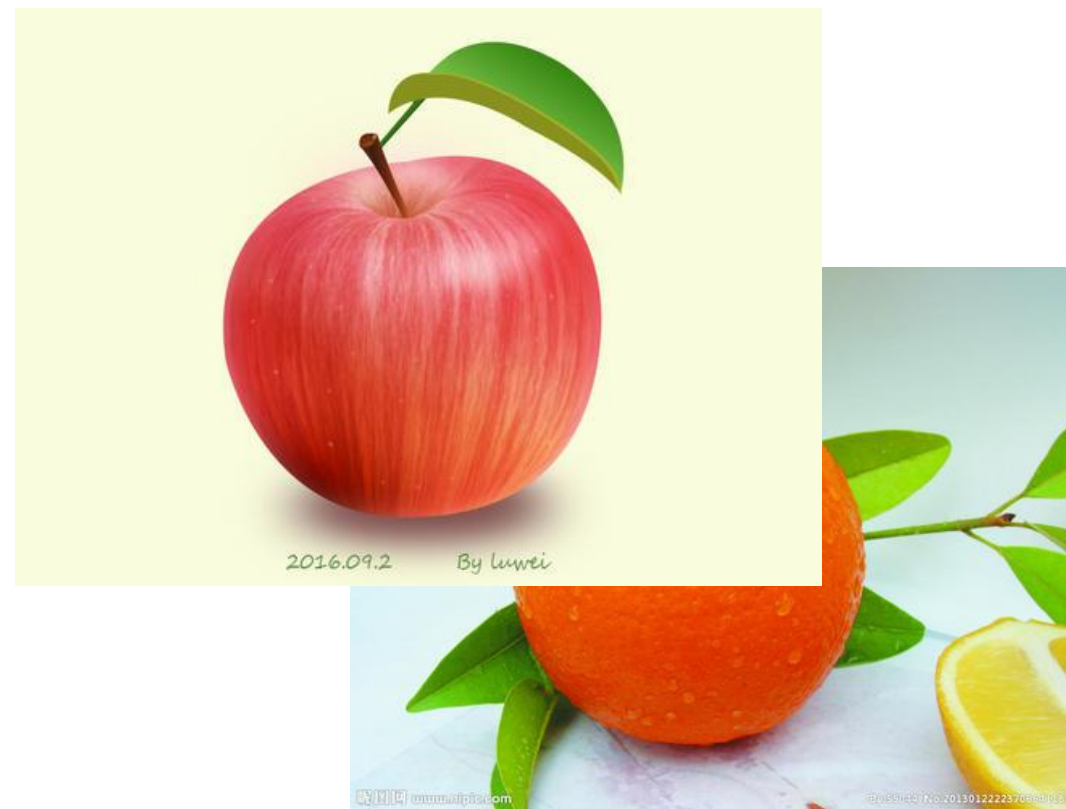
什么是数据？

数据的基本概念

- 数据是对客观事物的观测（测量）或描述而得到的符号或数字集合。

| 序号 | 姓名 | 性别 | 年龄段 | 职业 | 消费收入比 | 剩余信贷比 | 历史贷款 | 房产 | 已抵押资产 | 家属 | 违约 |
|----|----|----|-----|------|-------|-------|------|----|-------|----|----|
| 1 | 张三 | 男 | 中年 | 个体商人 | 居高 | 高 | 3 | 2 | 5 | 2 | 0 |
| 2 | 李四 | 女 | 中年 | 教师 | 一般 | 低 | 0 | 1 | 0 | 1 | 0 |
| 3 | 梁五 | 男 | 青年 | 自由职业 | 超出 | 中 | 1 | 0 | 0 | 0 | 1 |
| 4 | 王六 | 男 | 老年 | 退休 | 正常 | 低 | 2 | 1 | 3 | 0 | 0 |
| 5 | 张七 | 男 | 中年 | 司机 | 一般 | 较高 | 1 | 0 | 0 | 0 | 1 |
| 6 | 陈八 | 女 | 中年 | 建筑师 | 一般 | 低 | 0 | 1 | 2 | 2 | 0 |

客户数据



图像数据

数据对象的概念

- 原始数据通常是一个包含多个数据对象（data object）的集合，每个数据对象通常对应于一个具有完整语义信息的事物，是分析事物的基本单位。

| 序号 | 姓名 | 性别 | 年龄段 | 职业 | 消费收入比 | 剩余信贷比 | 历史贷款 | 房产 | 已抵押资产 | 家属 | 违约 |
|----|----|----|-----|------|-------|-------|------|----|-------|----|----|
| 1 | 张三 | 男 | 中年 | 个体商人 | 居高 | 高 | 3 | 2 | 5 | 2 | 0 |
| 2 | 李四 | 女 | 中年 | 教师 | 一般 | 低 | 0 | 1 | 0 | 1 | 0 |
| 3 | 梁五 | 男 | 青年 | 自由职业 | 超出 | 中 | 1 | 0 | 0 | 0 | 1 |
| 4 | 王六 | 男 | 老年 | 退休 | 正常 | 低 | 2 | 1 | 3 | 0 | 0 |
| 5 | 张七 | 男 | 中年 | 司机 | 一般 | 较高 | 1 | 0 | 0 | 0 | 1 |
| 6 | 陈八 | 女 | 中年 | 建筑师 | 一般 | 低 | 0 | 1 | 2 | 2 | 0 |



不同类型的数据

- 记录数据
 - 关系表数据
 - 事务数据 (Transaction Data)
- 多媒体数据：声、图、文
- 时空数据
 - 空间数据 (Spatial Data)
 - 时间数据 (Temporal Data)
- (关系) 图数据

记录数据

- 数据是记录的汇集，每个记录包含固定的属性集

| <i>Tid</i> | Refund | Marital Status | Taxable Income | Cheat |
|------------|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

事务数据

- 一种特殊类型的记录数据，其中
 - 每条记录（事务）涉及一系列的项
 - 考虑一个杂货店，顾客一次购物所购买的商品的集合构成一个事务，而购买的商品是项。

| <i>TID</i> | <i>Items</i> |
|-------------------|----------------------------------|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

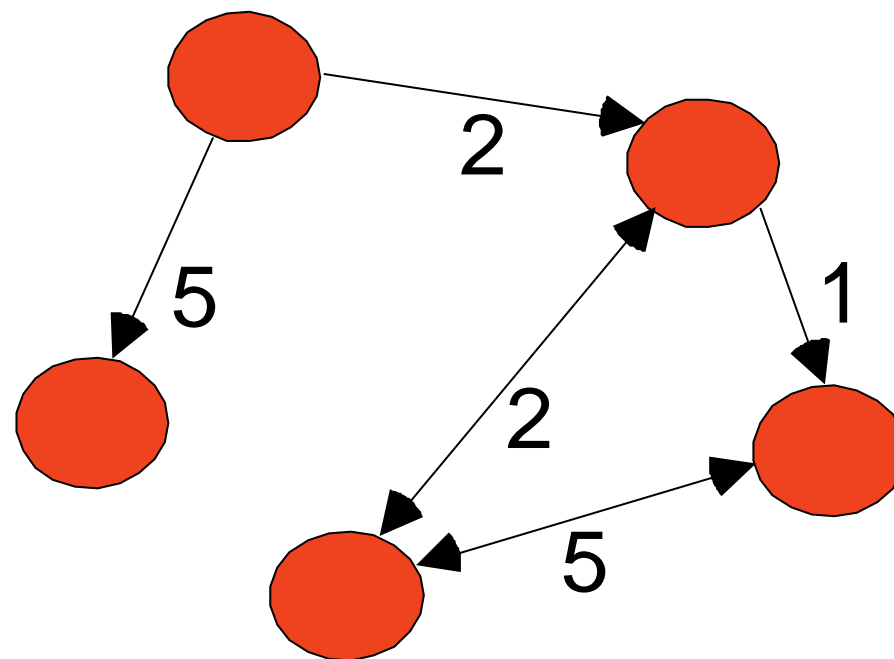
文档数据

- 每一个文档都是一个“术语”向量
 - 每个术语是向量的一个组成部分（属性）
 - 每个组成部分的值是相应术语在文档中出现的次数

| | team | coach | play | ball | score | game | win | lost | timeout | season |
|------------|------|-------|------|------|-------|------|-----|------|---------|--------|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

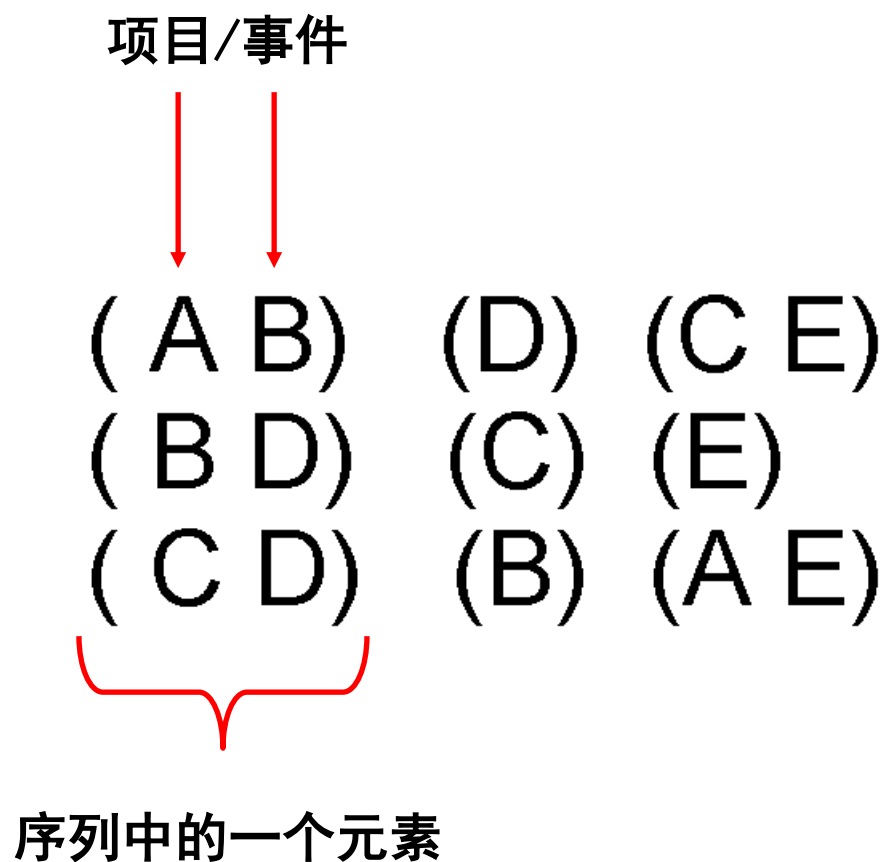
关系图数据

- 网页链接图
- 社交网络
- 文献引用图
-



事件序列数据

- 事务序列



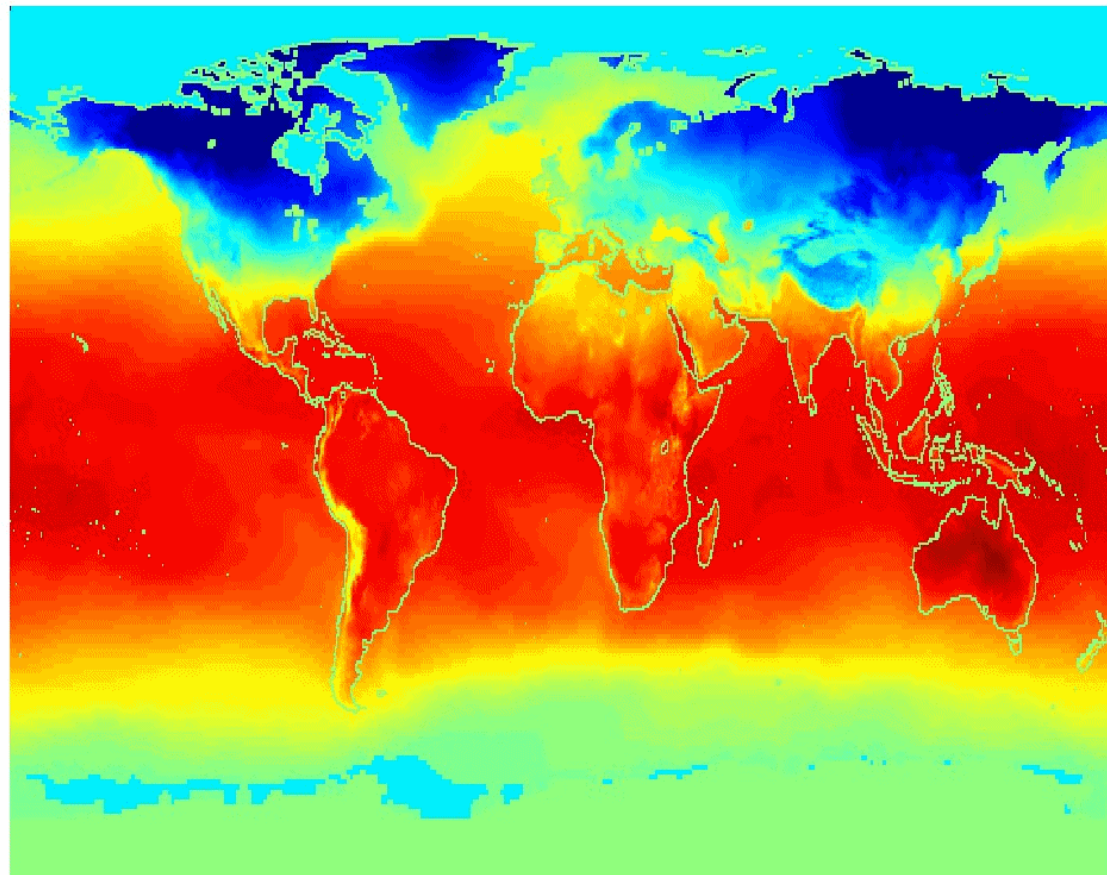
基因序列数据

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCCGCCCCGCGCCGTC
GAGAAGGGCCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

时空数据

Jan

陆地和海洋的
月平均温度



大数据的核心问题

- **核心挑战：**具有**多源、异构、信息碎片化、不确定性**的特征
- **“关联”：**发现多源、异构的碎片化信息之间的**关联关系**

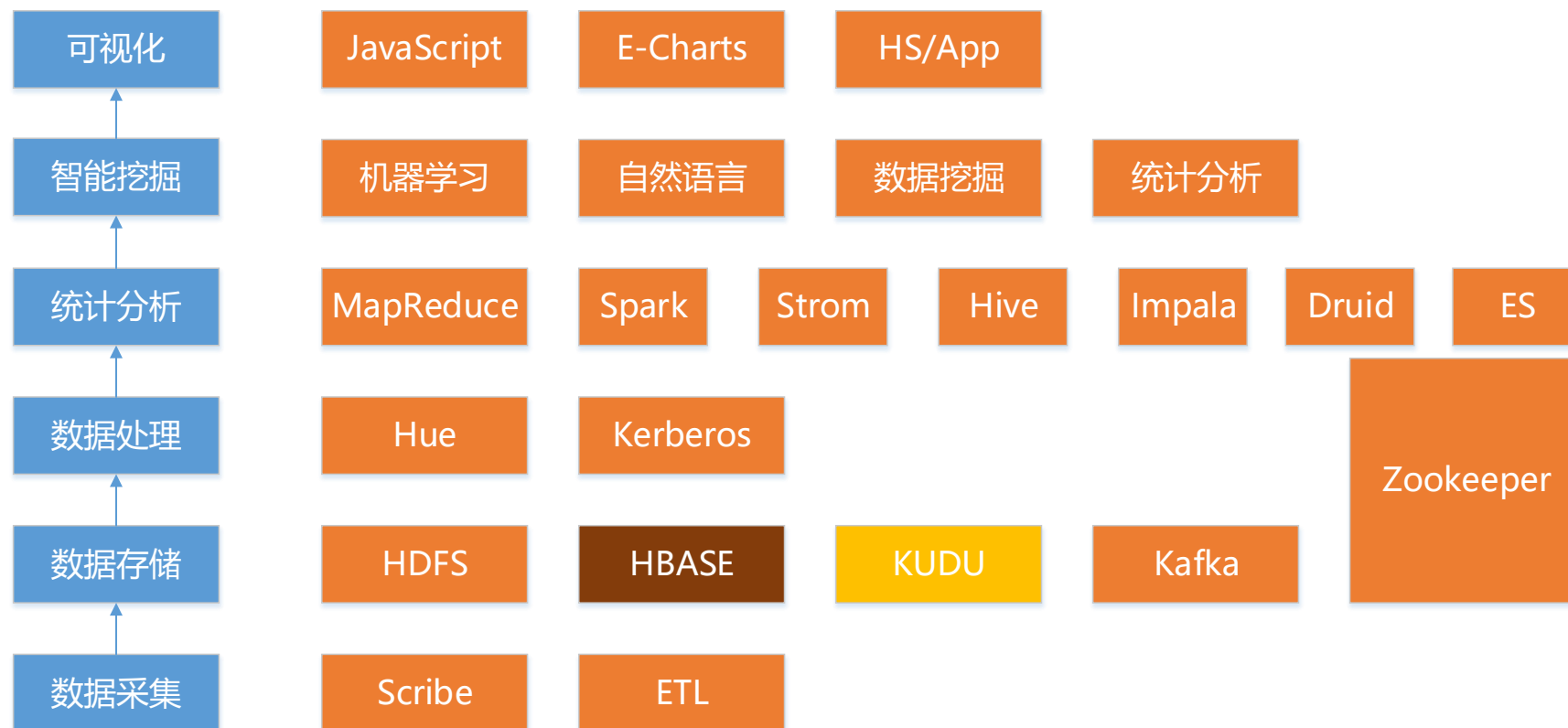


图片引自http://blog.sina.com.cn/s/blog_6773d7b90100jnsd.html

大数据技术概况

大数据技术体系

- 数据采集、数据存储、数据处理、统计分析、智能挖掘、可视化



大数据统计分析技术

- 条件查询

- SQL语言查询（或类SQL）

- 聚合统计

- 按地区汇总销售量
- 按时间维度汇总

| | 江苏 | 上海 | 北京 | 汇总 |
|----|------|-----|-----|------|
| 电器 | 940 | 450 | 340 | 1730 |
| 服装 | 830 | 350 | 270 | 1450 |
| 汇总 | 1770 | 800 | 610 | 3180 |

- 复杂报表

- 多维度、多层次统计分析：联机分析处理
(OLAP)

主要技术挑战：海量数据的检索性能！

从统计分析到智能挖掘

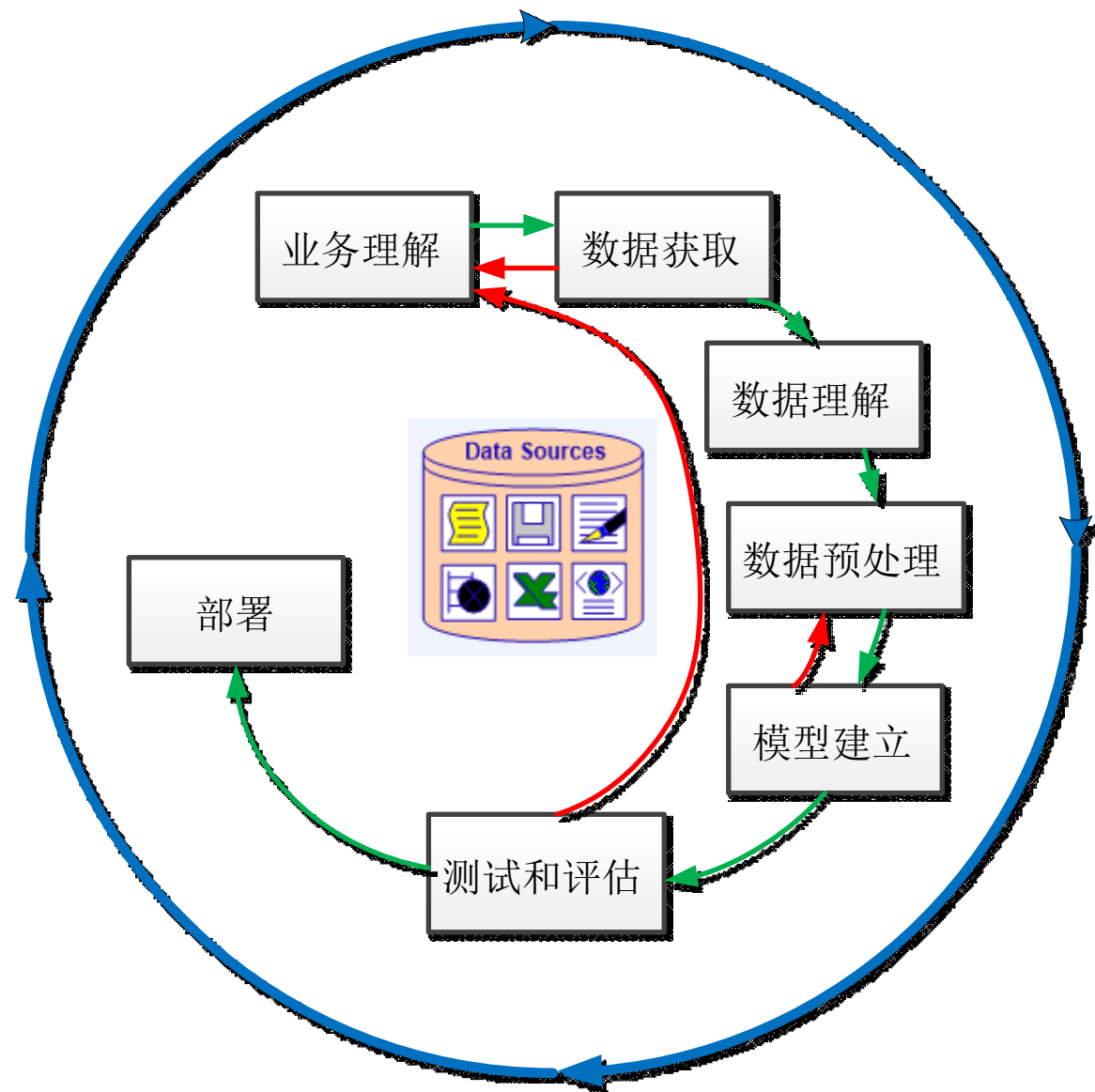
数据挖掘： Data Mining！



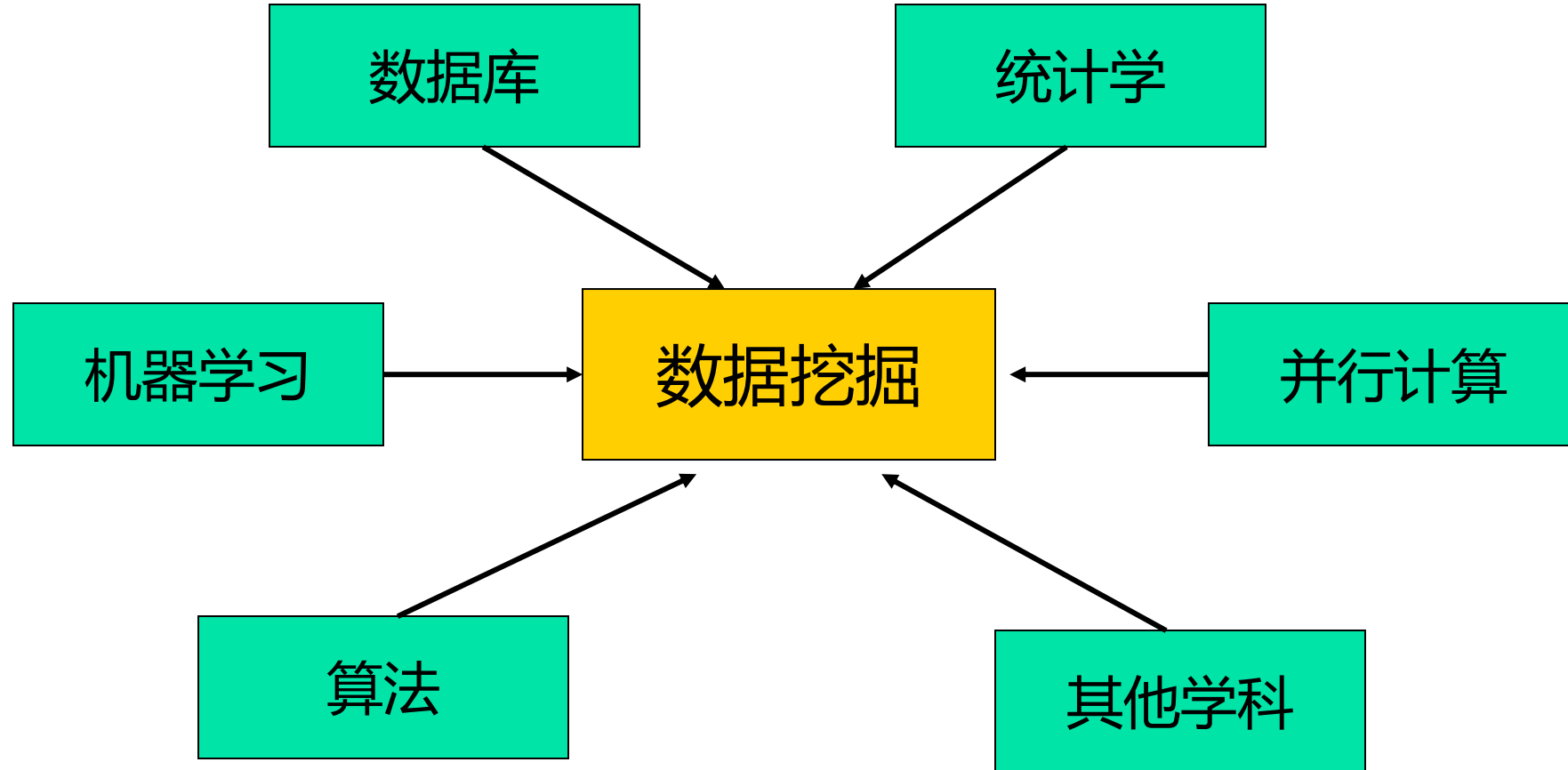
数据挖掘是什么？

- 数据挖掘（从数据中发现知识）
 - 从大量数据中提取有趣的（非平凡的，隐含的，以前未知的和潜在有用的）模式（pattern）或知识
- 替代名称
 - 数据库中的知识发现（Knowledge discovery in Databases, KDD）
 - 知识抽取（knowledge extraction）、模式挖掘（pattern mining）等
- 哪些数据处理和分析任务不是“数据挖掘”
 - 查询处理
 - 专家系统或小型ML /统计程序

实际数据挖掘项目的过程模型



数据挖掘：多学科融合



常见数据挖掘任务

- 多维概念描述：特征化概括和对比区分
- 关联规则挖掘
- 分类和回归预测
- 聚类分析与离群点检测
- 推荐系统：如协同过滤
- 趋势和演变分析
 - 子图模式挖掘、周期性分析

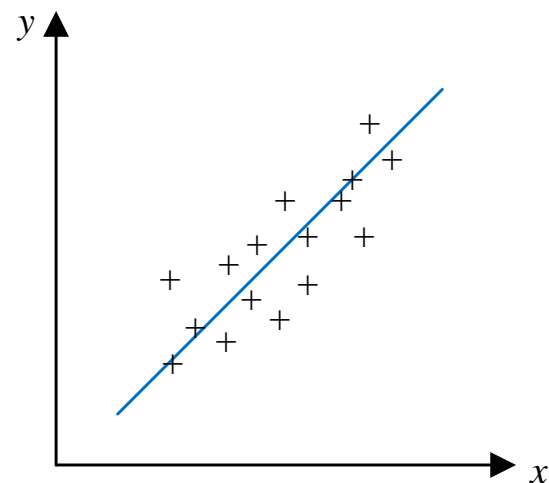
分类与回归

- 分类(classification): 预测给定数据对象的类别 (class, 离散值)
- 回归(regression): 预测给定数据对象对应的目标值 (连续值)

$$y = f(\mathbf{x}), \quad \text{其中 } \mathbf{x} \in \mathbf{D}$$



分类



回归

分类示例

分类属性

分类属性

连续属性

类别

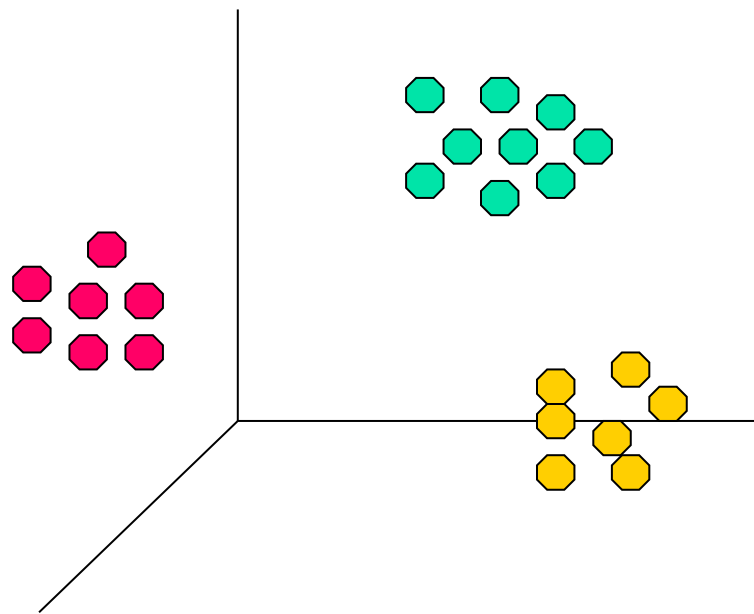
| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

| Refund | Marital Status | Taxable Income | Cheat |
|--------|----------------|----------------|-------|
| No | Single | 75K | ? |
| Yes | Married | 50K | ? |
| No | Married | 150K | ? |
| Yes | Divorced | 90K | ? |
| No | Single | 40K | ? |
| No | Married | 80K | ? |



聚类分析 (cluster analysis)

- 给定一个数据对象集合，以及数据对象之间的相似性度量，找到这样的一组簇 (cluster)：
 - 同一个簇中的数据点彼此更相似，不同簇中的数据对象彼此不太相似。



聚类分析示例


Baidu 百度

苹果

找到相关图片约6581张

版权 高清 最新 动图 1024x768 全部颜色

相关搜索: 水果苹果 苹果壁纸高清 苹果壁纸 苹果Logo 苹果素描 苹果11图片 苹果手机 苹果12真实图片 苹果 apple watch 苹果七 苹果八 苹果的照片 苹果所有型号手机 iPhone 8



The grid contains the following images (row by row, left to right):

- Row 1: A tree full of red apples; an iPhone 11; a tree with red apples; a close-up of red apples; the Apple logo on a black background; two red apples on a branch; a single red apple with water droplets.
- Row 2: A close-up of a white Apple logo on a dark surface; a tree with red apples; a hand holding an iPhone 11; a close-up of red apples; a close-up of red apples; a close-up of red apples; a spiral galaxy in space.
- Row 3: A close-up of a red apple; two yellow-green apples on a branch; a tree with red apples; a close-up of red apples; an iPad displaying a surfing app; a large pile of red apples; a tree with red apples.
- Row 4: A blue and white abstract pattern with a red Apple logo outline; a close-up of a red apple; a tree with red apples; a bowl of diced yellow fruit; a tree with red apples; an Apple iMac; a basket of red apples.

关联规则挖掘

- 挖掘事物之间的关联关系
- 给定一组记录（数据对象），每个记录包含来自给定集合的一些项目（Item）
- 生成项集（itemset）之间的关联规则：

$$X \longrightarrow Y$$

| <i>TID</i> | <i>Items</i> |
|------------|---------------------------|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

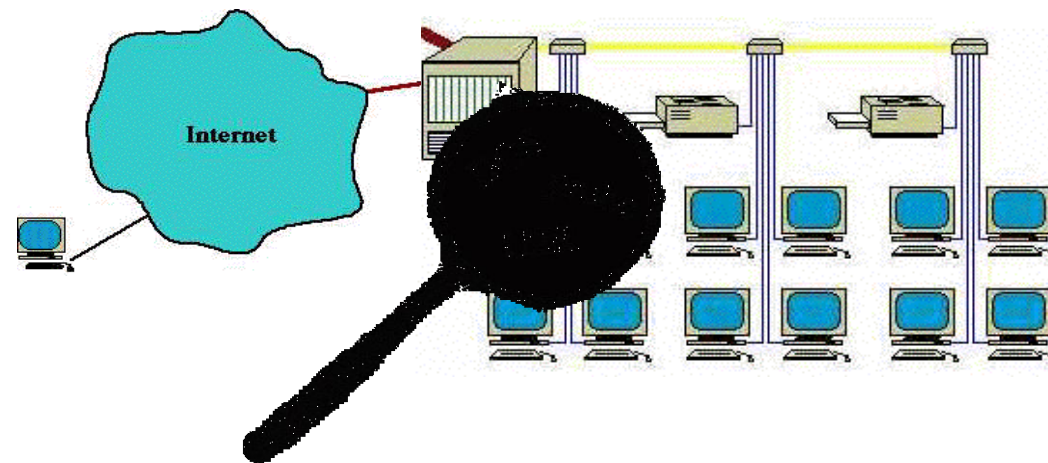
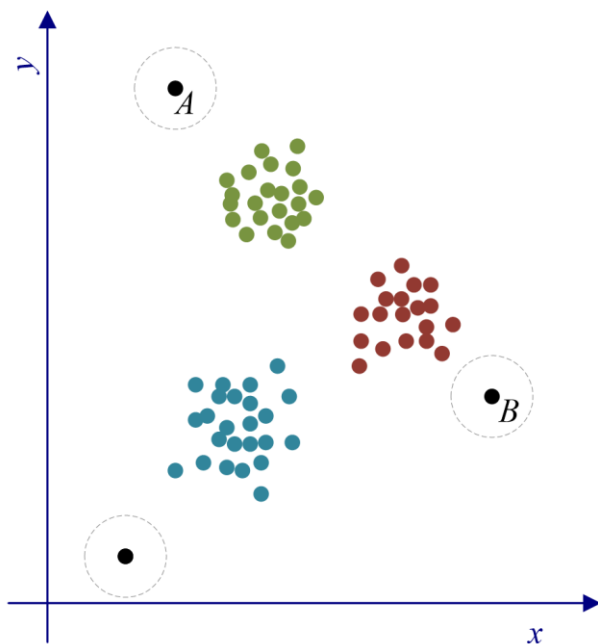
挖掘的关联规则：

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

离群点/异常检测

- 检测与正常行为之间存在的显著偏差
- 应用：
 - 信用卡欺诈检测、网络入侵检测等



推荐系统：协同过滤

- 给定用户偏好的数据库，预测新用户的偏好
- 示例：预测你喜欢的新电影，根据
 - 你过去的偏好
 - 其他有相同偏好的人，以及他们对新电影的偏好

| |  |  |  |  |  |
|---|---|--|---|---|---|
|  | 5 | | | 4 | |
|  | | 1 | 2 | 3 | 3 |
|  | 4 | | 4 | | |
|  | | 3 | | | |
|  | | | | 2 | 1 |

推荐系统的成功应用案例：今日头条

- 根据浏览历史推荐新闻

今日头条

推荐

阳光宽频

热点

图片

科技

娱乐

游戏

体育

汽车

财经

搞笑

更多



习近平：欢迎塞内加尔成为第一个同中国签署“一带一路”合作文件的西非国家

国际 人民网 · 25评论 · 刚刚



习近平在南非媒体发表署名文章

国际 新华网 · 1评论 · 刚刚

要闻

社会

娱乐

体育

军事

明星

为您推荐了10篇文章



习近平：欢迎塞内加尔成为第一个同中国签署“一带一路”合作文件的西非国家

国际 人民网 · 26评论 · 刚刚



习近平在南非媒体发表署名文章

国际 新华网 · 1评论 · 刚刚

中国陆军首度军长大考，释放出什么信号？

军事 上观新闻 · 20评论 · 刚刚



事业单位合并后，有职称的人员应该如何安置？

社会 悟空问答 · 刚刚



面对美国颠倒黑白，华春莹的这些回应太精彩！

国际 海外网 · 25评论 · 刚刚

数据的来源及获取方法

内部数据

- 公司等组织机构从内部的团体、员工、用户或信息中获取的数据：
 - 例：淘宝的用户购买纪录

| 宝贝地址添加 | | | | | |
|---------------------------|--|---------------|-------------|-----------|----------|
| 请点击此处输入买家的旺旺号，可搜索该买家的购买记录 | | 开始 | 暂停 | 继续 | 停止 |
| 统计指定时间内总金额：2013年 5月11日 | | 到 | 2013年11月11日 | 统计金额 | 总金额：{0}元 |
| 所属买家 | 购买产品 | 购买日期 | 购买金额 | 连接 | |
| pants88 | 大力水手 奥丽薇olive oyl女包R5172-20 R5172-28单... | 2012/09/17... | 139.0 | http://tr | |
| pants88 | 韩版专柜正品奥莉薇新款奥丽薇Olive 二折长款钱包R2... | 2012/09/17... | 79.0 | http://tr | |
| pants88 | 大力水手专柜正品奥丽薇女士拉链钱包长款钱包R2941-... | 2012/09/17... | 54.0 | http://tr | |
| pants88 | 韩版新款大力水手奥丽薇女包手提斜挎两用包R5320-37... | 2012/09/15... | 139.0 | http://tr | |
| pants88 | 韩版 新款奥丽薇专柜正品奥莉薇女 单肩包 R5396-24... | 2012/08/31... | 134.0 | http://tr | |
| pants88 | 12专柜正品奥莉薇OLIVE OYL奥丽薇单肩包女包包R5392... | 2012/08/30... | 139.0 | http://tr | |
| pants88 | 正品大力水手奥丽薇女士钱包 R2896-20三折长款钱夹 ... | 2012/08/24... | 54.0 | http://tr | |
| pants88 | 59元 七匹狼 正品 钱包 男士 真皮 短款 潮 牛皮 韩... | 2012/08/24... | 118.0 | http://it | |
| pants88 | 大热卖双人两用跳舞毯 包邮高清电视电脑二合一双人... | 2012/08/22... | 148.0 | http://it | |
| 小妮芭 | zara女装代购 2013欧美秋冬大码条纹T恤 简约纯棉修... | 2013/10/24... | 198 | http://tr | |
| 小妮芭 | 妖精的口袋【ELF SACK】花的安眠曲~秋装提花玫瑰复... | 2013/10/22... | 159 | http://tr | |
| 小妮芭 | 包邮正品 红蜻蜓六/红蜻蜓标准版福娃 赠主题鼠标垫 | 2013/08/20 | 79 | http://it | |

外部数据

- 互联网开放的数据

- 例：开放的数据平台；网络爬虫等



外部数据

- 合作的组织机构的交流数据
 - 例：腾讯和京东的合作，共享某些用户行为数据



内部数据获取方法

- 问卷调查
 - 如课程评价
- 员工信息登记
 - 如腾讯员工信息
- 公司对应业务的数据记录等
 - 如淘宝的购买纪录



一些简单的外部数据获取方法

- 与合作机构的数据交换
- 网络上直接提供的下载 (Download)
- 网络爬虫 (spider, crawler)

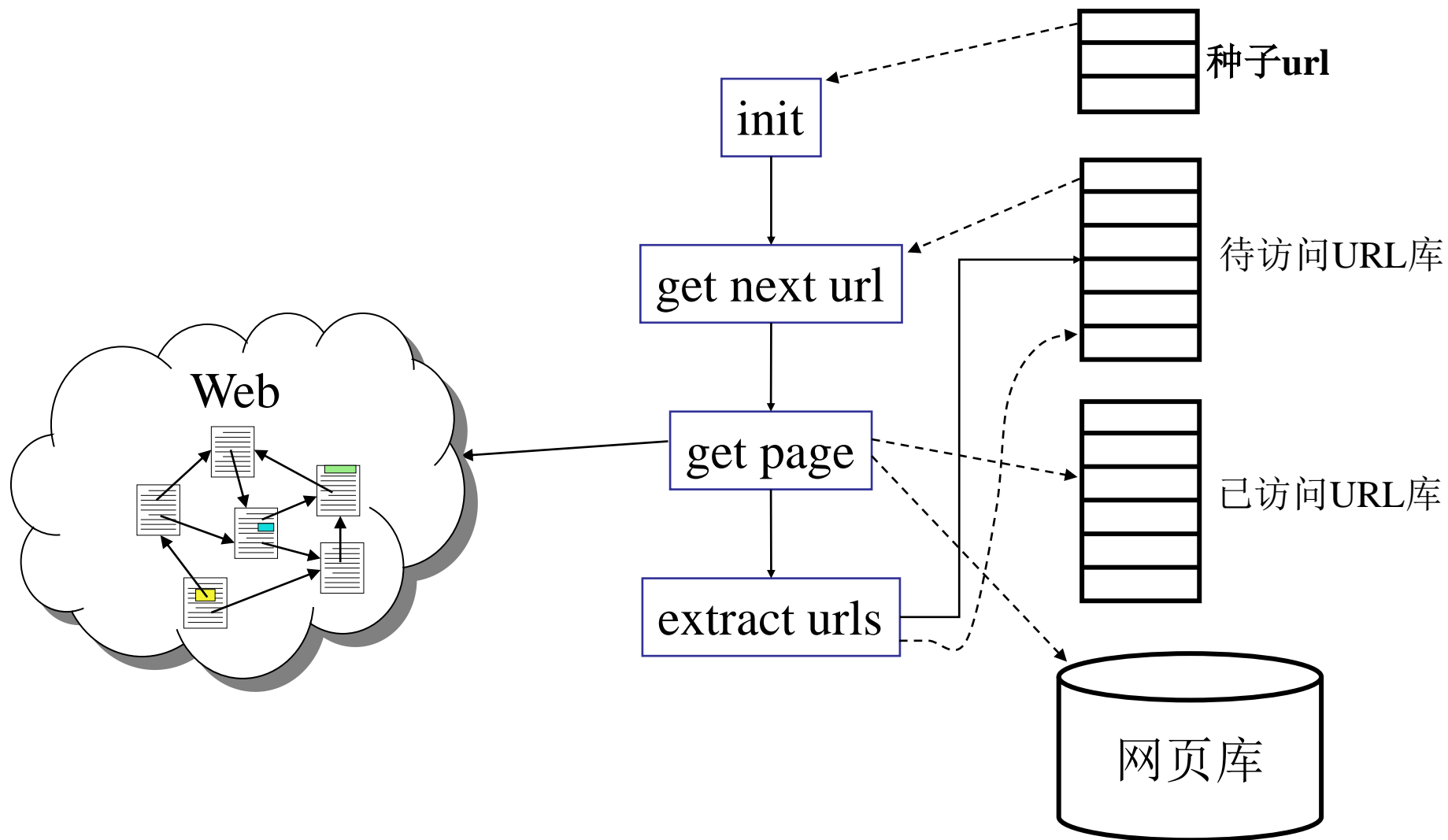


Stanford Large Network Dataset Collection

- [Social networks](#) : online social networks, edges represent interactions between people
- [Networks with ground-truth communities](#) : ground-truth network communities in social and information networks
- [Communication networks](#) : email communication networks with edges representing communication
- [Citation networks](#) : nodes represent papers, edges represent citations
- [Collaboration networks](#) : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- [Web graphs](#) : nodes represent webpages and edges are hyperlinks
- [Amazon networks](#) : nodes represent products and edges link commonly co-purchased products
- [Internet networks](#) : nodes represent computers and edges communication
- [Road networks](#) : nodes represent intersections and edges roads connecting the intersections
- [Autonomous systems](#) : graphs of the internet
- [Signed networks](#) : networks with positive and negative edges (friend/foe, trust/distrust)
- [Location-based online social networks](#) : Social networks with geographic check-ins
- [Wikipedia networks, articles, and metadata](#) : Talk, editing, voting, and article data from Wikipedia
- [Temporal networks](#) : networks where edges have timestamps
- [Twitter and Memetracker](#) : Memetracker phrases, links and 467 million Tweets
- [Online communities](#) : Data from online communities such as Reddit and Flickr
- [Online reviews](#) : Data from online review systems such as BeerAdvocate and Amazon

SNAP networks are also available from [SuiteSparse Matrix Collection](#) by [Tim Davis](#).

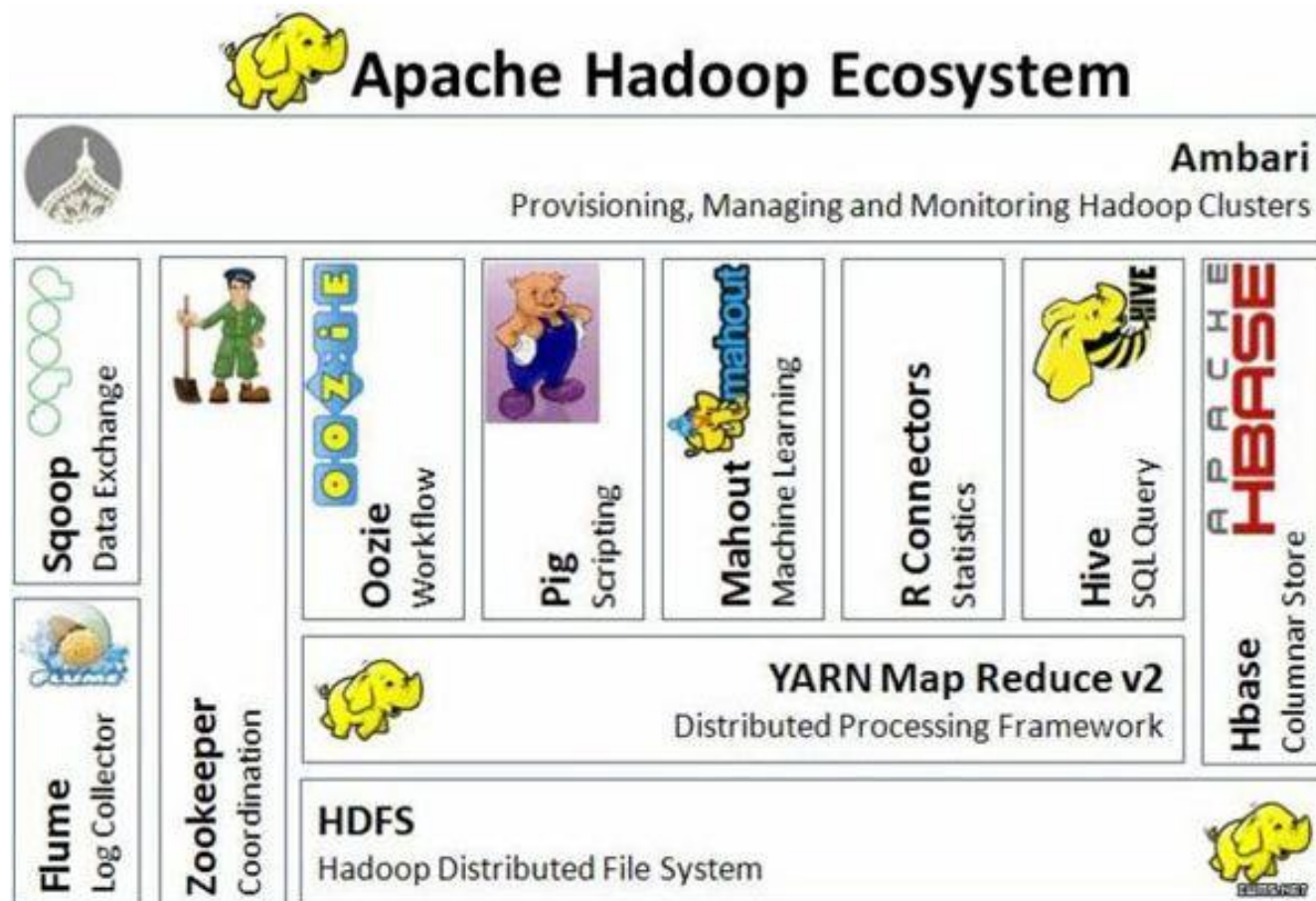
网络爬虫的工作流程



大数据的学习资源

大数据存储与分析系统

- 关系数据库 (SQL) : MySQL, Oracle,
- NoSQL: Not Only SQL
 - Key-value数据库
 - Redis、mongodb
 - 图数据库: Neo4J
 -



数据挖掘系统

- 商业化系统
 - SAS Enterprise Miner
 - SPSS Clementine
 - Insightful Miner
 - Oracle/SQL Server提供的数据挖掘工具
 -
- 开源系统
 - Weka
 - Knime
 - Alphaminer
 - Mahout
 - Spark

大数据领域的重要国际会议和期刊

● Conferences

- ACM SIGMOD
- VLDB
- ICDE
- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (**KDD**)
- SIAM Data Mining Conf. (**SDM**)
- (IEEE) Int. Conf. on Data Mining (**ICDM**)
- PAKDD, PKDD

■ Other related conferences

- WWW, SIGIR
- ICML, CVPR, NIPS, IJCAI

■ Journals

- IEEE Trans. On Knowledge and Data Eng. (TKDE)
- ACM Transactions on Information Systems
- ACM Transactions on Database Systems
- The VLDB Journal
- ACM Trans. on KDD

Thank You for Your Attention

Contact me at: yym@hit.edu.cn

Tel: 26033008, 13760196623

Address: Rm.1402, H# Building

本PPT仅供学习参考，其中图片、数据等引用如有版权要求请及时联系