

# 计算机科学的挑战与方法 -大数据处理技术

主讲教师：怀进鹏

合作教师：邓 婷 沃天宇  
孙海龙 胡春明  
张日崇 马 帅  
李建欣 李 博

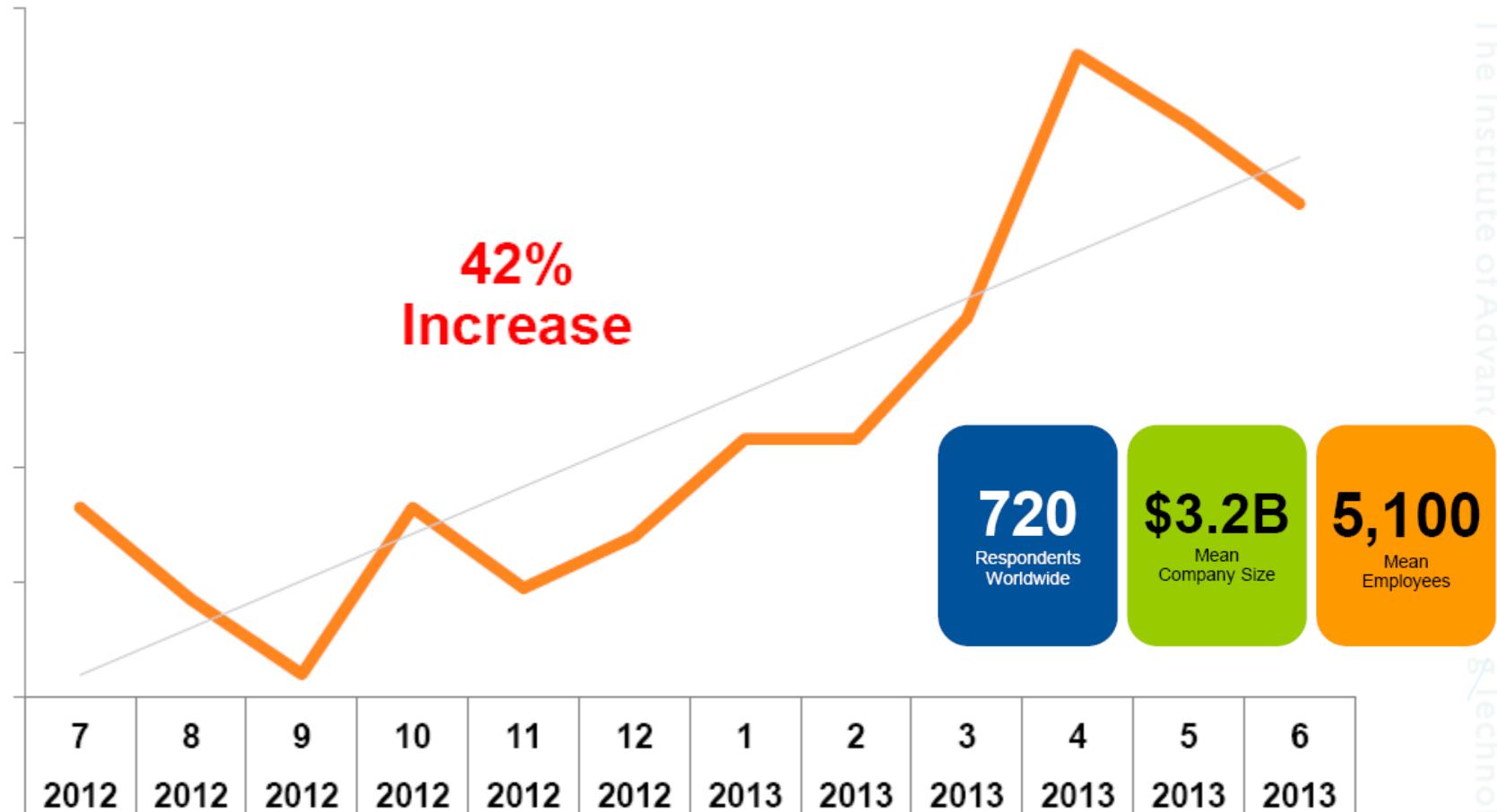
计算机学院  
计算机新技术研究所  
2013年11月16日

# 内容提要

- 大数据的分析现状
- 数据的处理流程及处理技术
- 数据挖掘算法
- 图数据分析处理技术

# Gartner关于业界对Big Data兴趣的分析

## Client Inquiries — Information Management

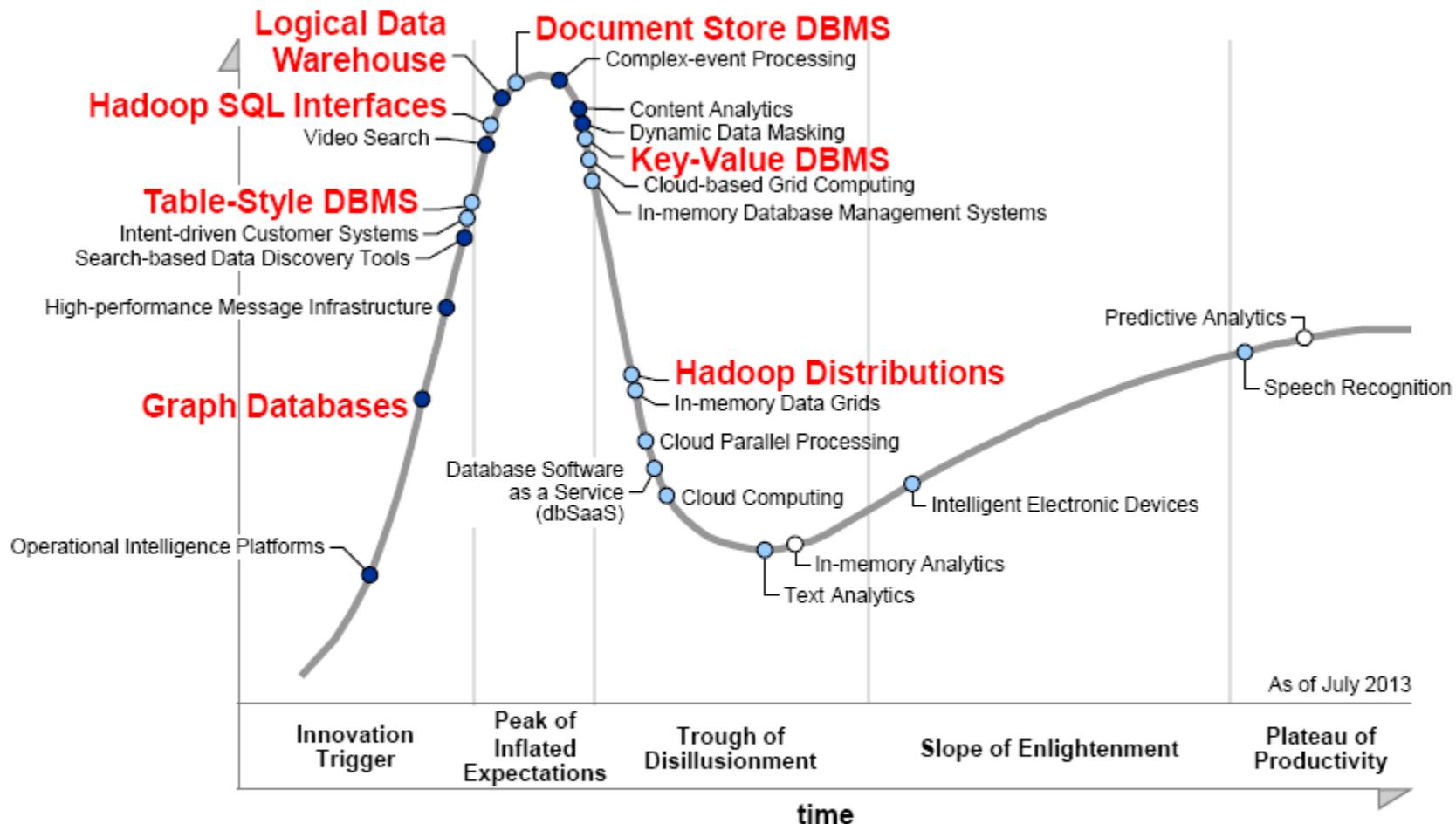


Source: Information Management Team Inquiry Data, July 2012-June 2013

© 2013 Gartner, Inc. and/or its affiliates. All rights reserved.

Gartner

# Gartner关于Big Data处理技术的分析



Plateau will be reached in:

○ less than 2 years    ● 2 to 5 years    ● 5 to 10 years    ▲ more than 10 years    ✗ obsolete  
 ✗ before plateau

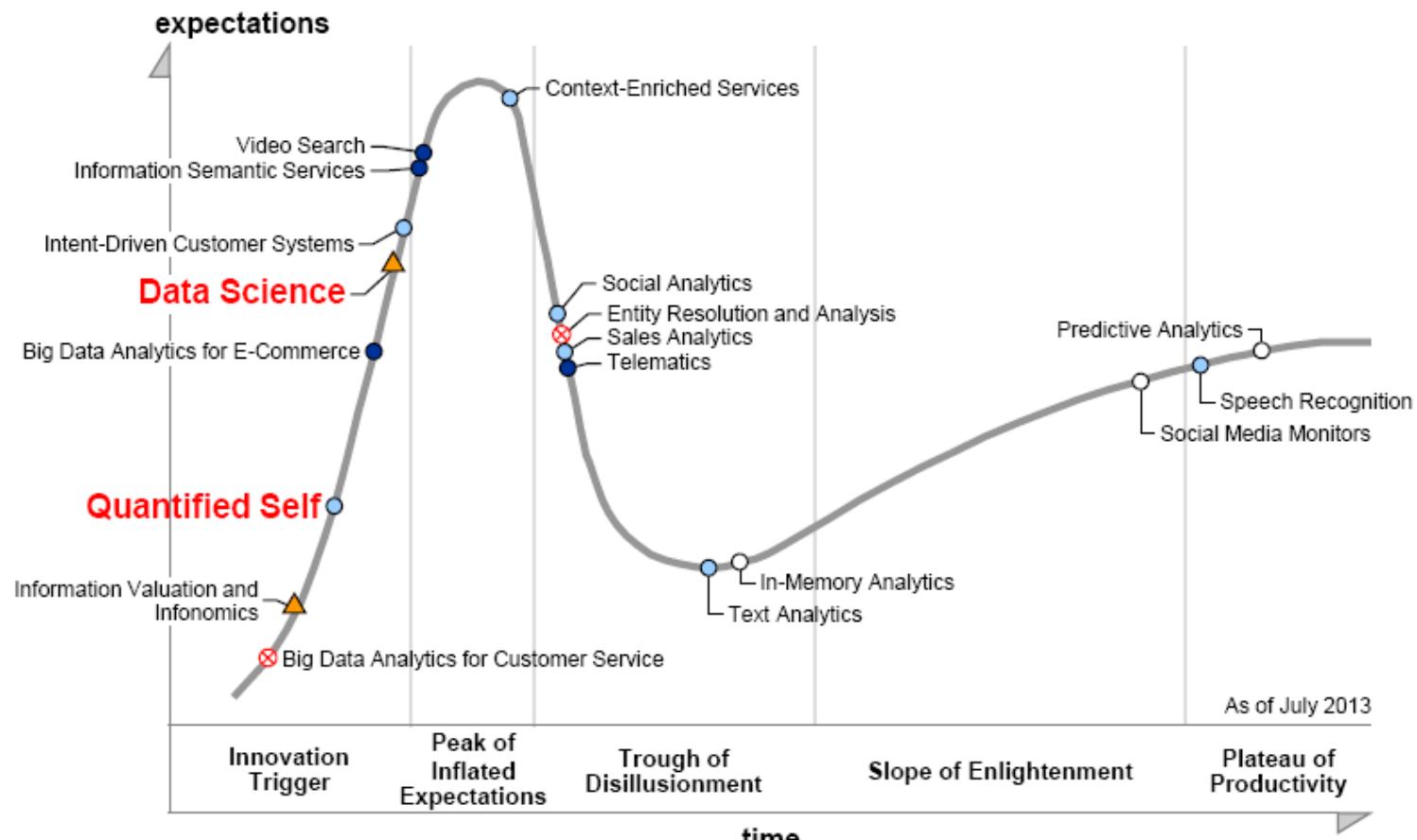
#GartnerSYM

© 2013 Gartner, Inc. and/or its affiliates. All rights reserved.

Source: Hype Cycle for Big Data, 2013, 31 July 2013 (G00252431)

**Gartner**

# Gartner关于Big Data处理技术的分析



Plateau will be reached in:

- less than 2 years     2 to 5 years     5 to 10 years     more than 10 years     before plateau

#GartnerSYM

© 2013 Gartner, Inc. and/or its affiliates. All rights reserved.

Source: Hype Cycle for Big Data, 2013, 31 July 2013 (G00252431)

**Gartner**

# 内容提要

- 大数据的分析现状
- 数据的处理流程及处理技术
- 数据挖掘算法
- 图数据分析处理技术

# 数据的处理流程

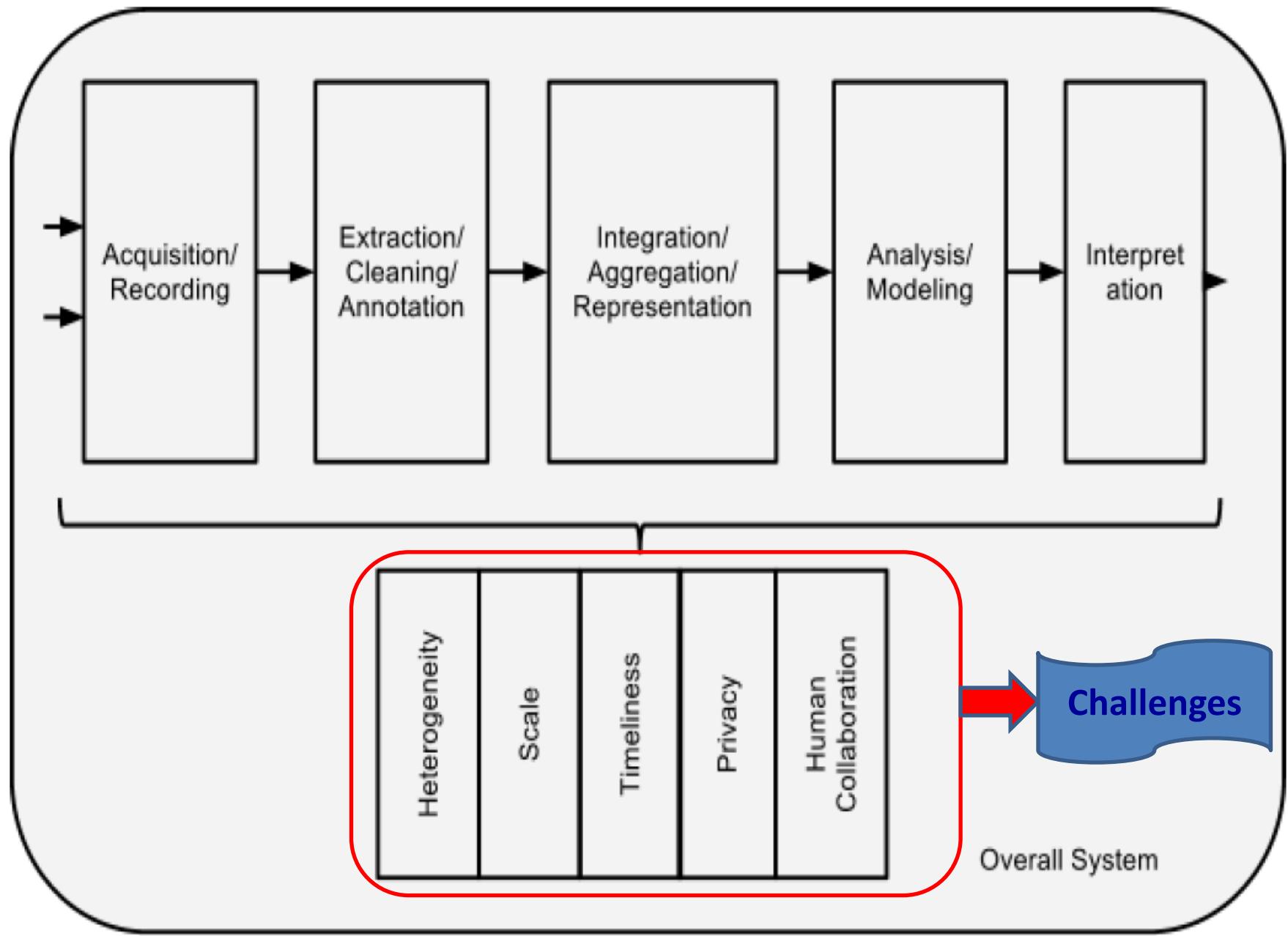
## Challenges and Opportunities with Big Data

- A community white paper developed by leading researchers across US

Divyakant Agrawal, UC Santa Barbara  
Philip Bernstein, Microsoft  
Elisa Bertino, Purdue Univ.  
Susan Davidson, Univ. of Pennsylvania  
Umeshwar Dayal, HP  
Michael Franklin, UC Berkeley  
Johannes Gehrke, Cornell Univ.  
Laura Haas, IBM  
Alon Halevy, Google  
Jiawei Han, UIUC  
Alexandros Labrinidis, Univ. of Pittsburgh

Sam Madden, MIT  
Yannis Papakonstantinou, UC San Diego  
Jignesh M. Patel, Univ. of Wisconsin  
Raghu Ramakrishnan, Yahoo!  
Kenneth Ross, Columbia Univ.  
Cyrus Shahabi, Univ. of Southern California  
Dan Suciu, Univ. of Washington  
Shiv Vaithyanathan, IBM  
Jennifer Widom, Stanford Univ

A result of remote conversation lasted about 3 months (Nov. 2011 ~ Feb. 2012)



# 大数据处理技术分析

## ■ 数据采集

- ETL工具、爬虫、传感器

## ■ 数据存储

- 文件系统、关系数据库、图数据库；NoSQL（hadoop）；

## ■ 数据分析

- NLP、统计、数据挖掘、机器学习、数据库

## ■ 数据展现

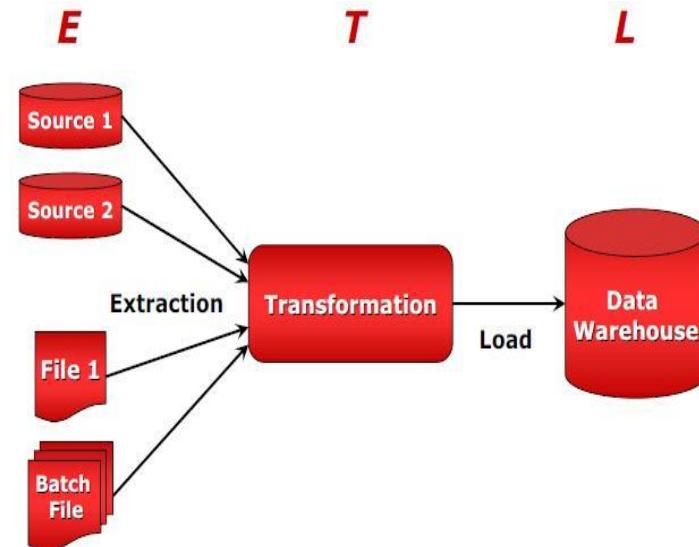
## ■ 数据类别

- 类型（结构、）
- 行业（医疗、社交）

# 数据采集-ETL

## ■ Extract, Transform and Load (ETL)

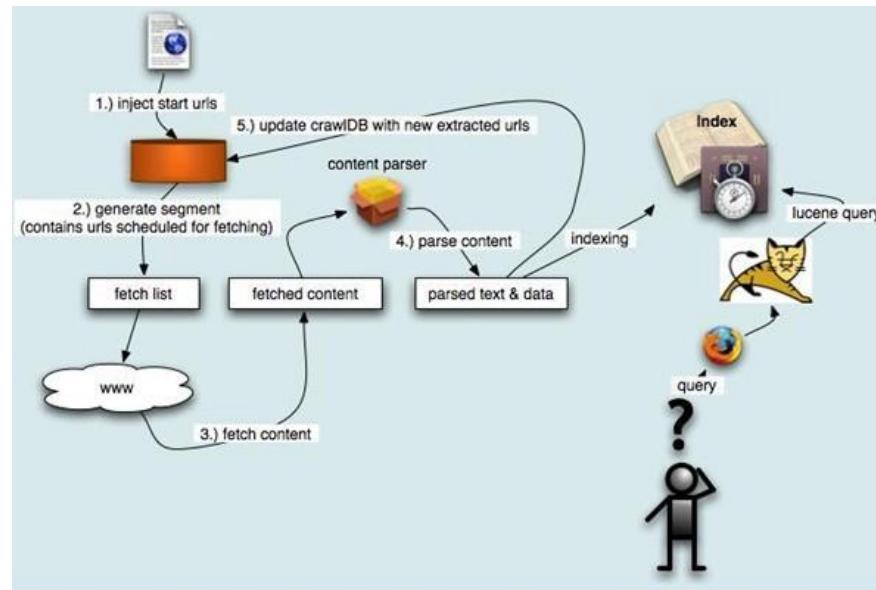
- ETL按照统一的规则集成并提高数据的价值，是负责完成数据从数据源向目标数据仓库转化的过程。



图片来源：<http://igorportela.com/extract-transform-and-load-etl/>

# 数据采集-爬虫

- 网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列，直到满足系统的一系列停止条件。



图片来源: <http://blog.csdn.net/pipi521520/article/details/5599919>

# 数据采集-传感器

- 数据采集是指从传感器和其它待测设备等模拟和数字被测单元中自动采非电量或者电量信号,送到上位机中进行分析，处理。



图片来源<http://www.acurite.com/sensor-based-forecasting>

# 数据存储

## ■ 文件系统

- 文件数据库又叫嵌入式数据库，将整个数据库的内容保存在单个索引文件中，以便于数据库的发布。

## ■ 关系数据库

- 关系数据库，是建立在关系模型基础上的数据库，借助于集合代数等数学概念和方法来处理数据库中的数据

## ■ 图数据库

- 图数据库的基本含义是以“图”这种数据结构存储和查询数据。

## ■ NoSQL ( hadoop )

- 非关系型数据库以键值对存储（key-value），它的结构不固定，每一个元组可以有不一样的字段，每个元组可以根据需要增加一些自己的键值对，这样就不会局限于固定的结构，可以减少一些时间和空间的开销。

# 数据处理与分析

## ■ 数据处理：

- 自然语言处理技术

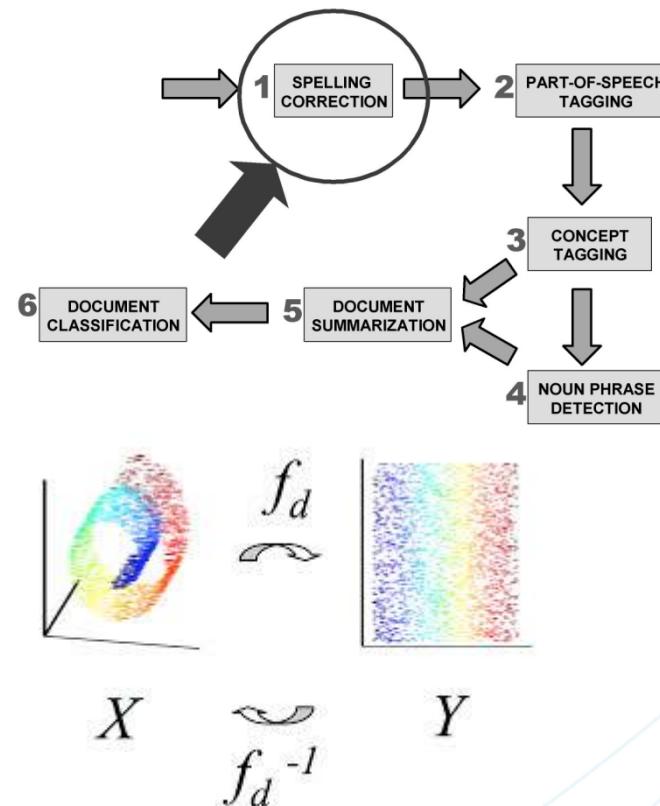
实现人与计算机之间用自然语言进行有效通信的各种理论和方法

- 数据降维技术

将样本点从输入空间通过线性或非线性变换映射到一个低维空间，从而获得一个关于原数据集紧致的低维表示

- 数据清理技术

发现并纠正数据文件中可识别的错误，包括检查数据一致性，处理无效值和缺失值等



# 数据仓库与联机分析处理

- 1988年IBM两位研究人员（Barry Devlin和Paul Murphy）创造性地提出了一个新的术语：数据仓库（Data Warehouse）
- 1992年比尔·恩门出版专著《Building the Data Warehouse》，真正拉开了数据仓库走向大规模应用的序幕，被誉为“数据仓库之父”



“数据仓库是一个面向主题的、集成的、相对稳定、反映历史变化的数据集合，用于支持管理中的决策制定”

- 数据仓库与数据库的主要区别：
  - 数据仓库以**数据分析、决策支持**为目的来组织存储数据
  - 数据库的主要目的是**为系统保存、查询数据**

# 数据挖掘

## ■ 数据挖掘算法按挖掘目的分为：

- 关联规则分析
  - 信息自动分类，信息过滤，图像识别等
- 聚类分析
- 异常分析
  - 入侵检测，金融安全等
- 趋势、演化分析
  - 回归，序列模式挖掘



数据挖掘：在你的数据中搜索知识（有趣的模式）。

# 大数据的应用—决策支持

- 1947年，赫伯特·西蒙在著作《行政组织的决策过程》中指出“人类的理性是有限的，因此所有的决策都是基于有限理论（bounded rationality）的结果”，并指出“如果能利用存储在计算机里的信息来辅助决策，人类理性的范围将会扩大，决策的质量就能提高”
- 预测“在后工业时代，也就是信息时代，人类社会面临的中心问题将从如何提高生产庇转变为如何更好地利用信息来辅助决策”

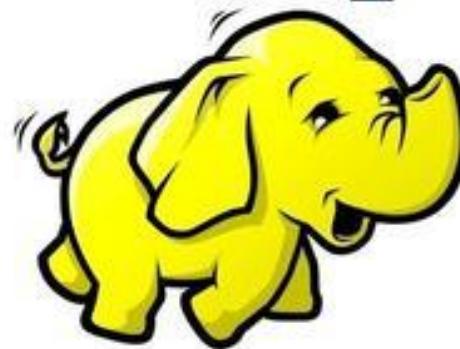


1975年图灵奖  
1978年诺贝尔经济学奖  
1993年美国心理协会终身成就奖

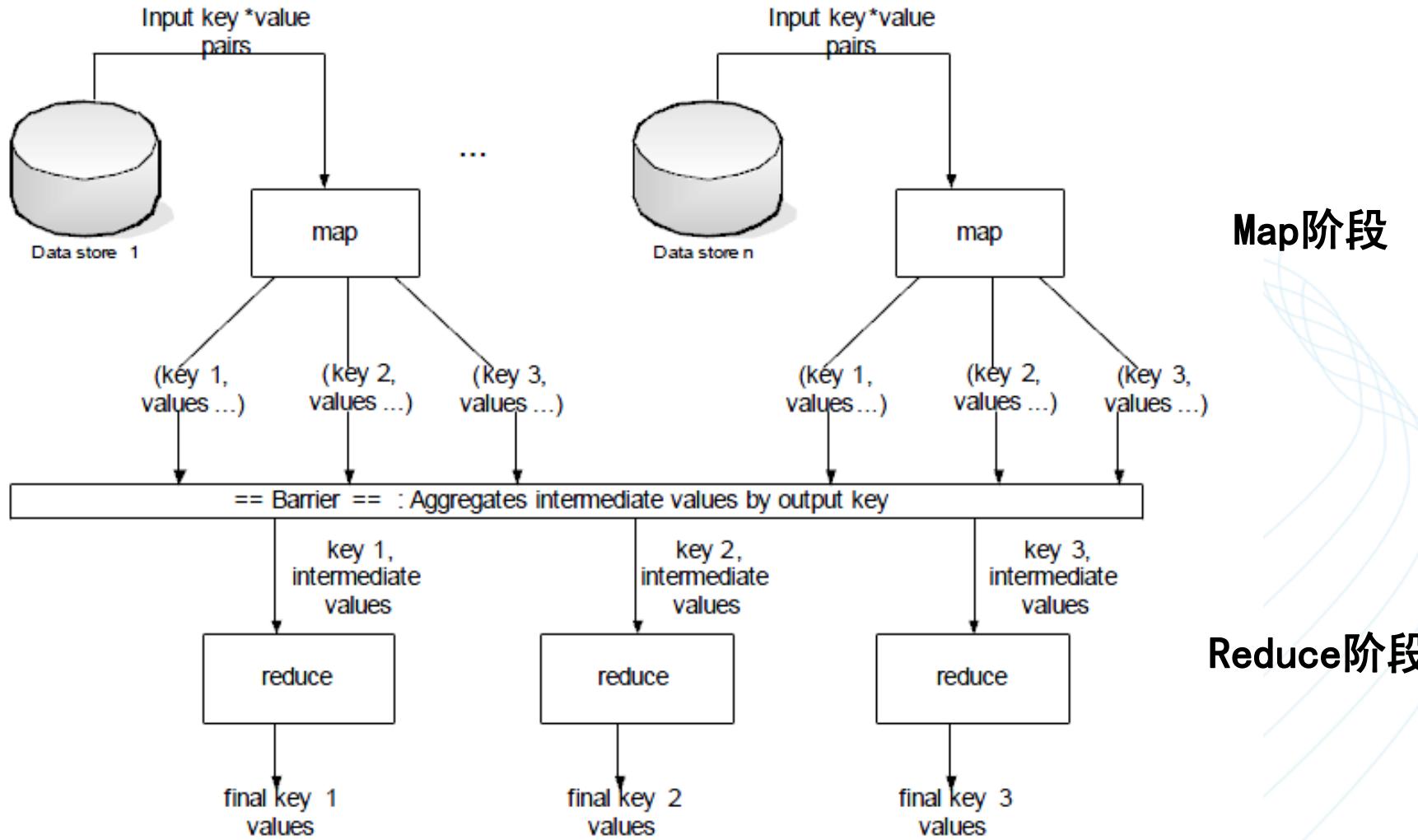
# MapReduce/Hadoop and Beyond

- 由Google提出的一个用于大数据处理的系统
  - Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, OSDI 2004.
- Apache开源社会项目： Hadoop
- 主要的思想来自于functional programming

**hadoop**



# MapReduce/Hadoop and Beyond



# MapReduce/Hadoop and Beyond

## ● MapReduce/Hadoop的局限性

- 比较底层的编程模型
- 对实时处理和递归处理支持不够
- 适合处理具有“局部性”的数据

## ● Beyond MapReduce

- 高层编程语言: Hive (Facebook), Pig (Yahoo!) 等...
- 流式计算: S4 (Yahoo!), Storm (Twitter), Spark (UC Berkeley AMP lab)
- 支持递归的系统: Google Pregel
- 其他技术。 . .

# 大数据可视化

## ■ 数据可视化

- 主要旨在借助于图形化手段，清晰有效地传达与沟通信息。
- 美学形式与功能齐头并进；通过直观地传达关键的方面与特征，实现对于相当稀疏而又复杂的数据集的深入洞察。

## ■ 数据可视化的分类 ( Frits H. Post, Gregory M. Nielson and Georges-Pierre Bonneau (2002). *Data Visualization: The State of the Art.* )

- 可视化算法与技术方法
- 立体可视化
- 信息可视化
- 多分辨率方法
- 建模技术方法
- 交互技术方法与体系架构



核医学成像



螺旋星云可见光图像

# 大数据类别

## ■ 数据类型

- 结构化数据：
  - 关系数据等：数据的查询、统计、更新等操作效率低。
- 半结构化数据：
  - XML、**图数据**等：转换为结构化存储或者按照非结构化存储。
- 非结构化数据：
  - 图片、视频、word、pdf、ppt等：不利于检索、查询和存储

## ■ 行业数据

- 大规模的电子商务数据
- **社会数据（社会网络，互联网等），是一类重要的图数据**
- 移动数据(呼叫详细记录、RFID、传感器网络)
- 医疗数据
- 天文学，大气科学，基因组学，生物地球化学，生物和其他复杂和/或跨学科的科研数据

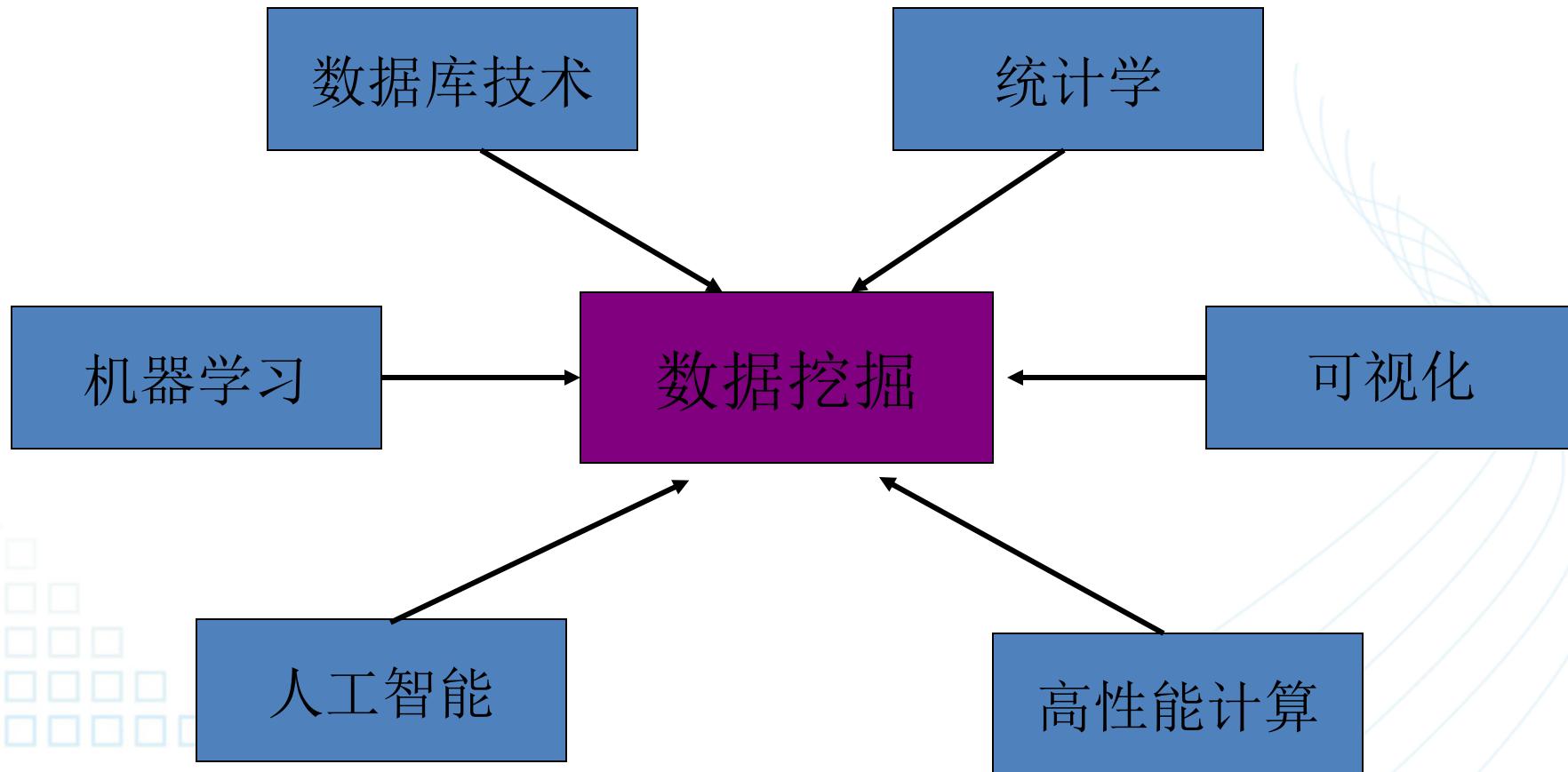
# 内容提要

- 大数据的分析现状
- 数据的处理流程及处理技术
- **数据挖掘算法**
- 图数据分析处理技术

# 数据挖掘的定义

- **数据挖掘是从大量数据中提取或“挖掘”知识。**
- 技术上的定义：数据挖掘（Data Mining）就是从**大量的、不完全的、有噪声的、模糊的、随机的**实际应用数据中，提取**隐含在其中的、人们事先不知道的、但又是潜在有用**的信息和知识的过程。
  -
- 商业角度定义：数据挖掘是一种新的商业信息处理技术，其主要特点是对商业数据库中的大量业务数据进行抽取、转换、分析和其他模型化处理，从中提取辅助商业决策的关键性数据。
- 所谓**基于数据库的知识发现**（KDD）是指从大量数据中提取有效的、新颖的、潜在有用的、最终可被理解的模式的非平凡过程。

# 数据挖掘是多学科的产物



# 统计分析和数据挖掘的区别-孙悟空跟二郎神打仗



- **统计分析：**两人斗争4567次，其中孙悟空赢3456次。另外，孙悟空斗牛魔王，胜率是89%，二郎神斗牛魔王胜率是71%。
- **关联分析：**计算死自动找到出身、教育、经验、单身四个因素。得出结论是孙悟空赢。

贫苦出身的孩子一般比皇亲国戚功夫练得刻苦；  
打架经验丰富的人因为擅长利用环境而机会更多；  
在都遇得到明师的情况下，贫苦出身的孩子功夫可能会高些；  
单身的人功夫总比同样环境非单身的高。

# 数据挖掘与统计学

统计学和数据挖掘有着共同的目标：**发现数据中的结构。**

事实上，由于它们的目标相似，一些人（尤其是统计学家）认为数据挖掘是统计学的分支。这是一个不切合实际的看法。因为数据挖掘还应用了其它领域的思想、工具和方法，尤其是计算机学科，例如数据库技术和机器学习，而且它所关注的某些领域和统计学家所关注的有很大不同。

# 数据挖掘被关注的原因

数据挖掘引起了信息产业界的极大关注，其主要原因是存在大量数据，可以广泛使用，并且迫切需要**将这些数据转换成有用的信息和知识**。获取的信息和知识可以广泛用于各种应用，包括商务管理、生产控制、市场分析、工程设计和科学探索等。

数据挖掘是信息技术自然进化的结果

- 数据库、数据仓库和Internet等信息技术的发展。
- 计算机性能的提高和先进的体系结构的发展。
- 统计学和人工智能等方法在数据分析中的研究和应用。

# 四个概念的不同



数据: 原始的, 未解释的信号或者符号, 如: 1

信息: 有一定解释或意义的数据, 如:  
S.O.S

知识: 综合信息形成的观点和普适性的理论

智慧: 能够综合知识和经验用以生存计划的人类思维的结晶

# 数据挖掘视为数据库中知识发现过程基本步骤的主要环节



# 数据挖掘的应用

电信：流失

银行：聚类（细分），交叉销售

百货公司/超市：购物篮分析（关联规则）

保险：细分，交叉销售，流失（原因分析）

信用卡：欺诈探测，细分

电子商务：网站日志分析

税务部门：偷漏税行为探测

警察机关：犯罪行为分析

医学：医疗保健

# 数据挖掘应用实例：市场分析和管理

## ■ 顾客形象

- ◆ 数据挖掘可以告诉你什麼样的顾客会买什麼样的产品（聚类或分类）

## ■ 识别顾客需求

- ◆ 保证为不同的顾客提供了最好的产品
- ◆ 使用预测手段去发现什麼因素会吸引新的顾客。

## ■ 提供汇总信息

- ◆ 各种各样的多方位汇总信息
- ◆ 统计的汇总信息（数据中心的趋势和变化）

# 数据挖掘应用实例：欺骗性检测和管理

## ■ 应用

- ◆ 广泛应用于医疗系统, 零售系统, 信用卡服务, 电信(电话卡欺骗行为), 等等.

## ■ 实现途径

- ◆ 利用历史性数据建立欺骗性行为模型并使用数据挖掘帮助识别同类例子

## ■ 具体事例

- ◆ 汽车保险: 检测出那些故意制造车祸而索取保险金的人
- ◆ 来路不明钱财的追踪: 发现可疑钱财交易(美国财政部的财政犯罪执行网)
- ◆ 医疗保险: 检测出潜在的病人, 呼叫医生和证明人

# 数据挖掘应用实例：欺骗性检测和管理

## ■ 发现不正确的医学治疗

- ◆ 澳大利亚医疗保险协会证明在许多情况下全面审查测试是很需要的

## ■ 检测电话错误

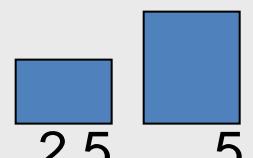
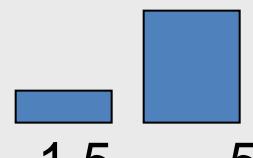
- ◆ 电话呼叫模式：呼叫目的地，持续时间，每天或每周的次数。分析与预期标准相背离的模式

## ■ 零售

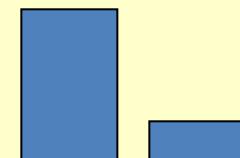
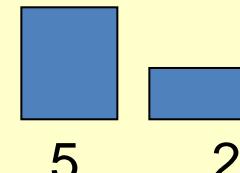
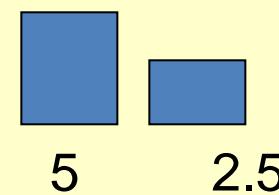
- ◆ 分析家估计38%的零售收缩缘于雇员的不诚实。

# 数据挖掘实例：分类问题

Examples of  
class A

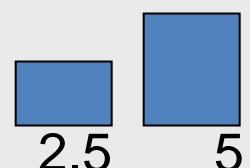
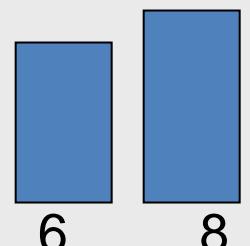


Examples of  
class B

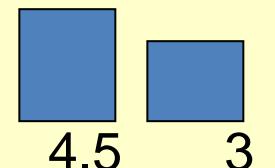
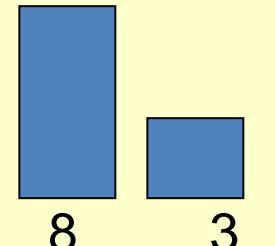
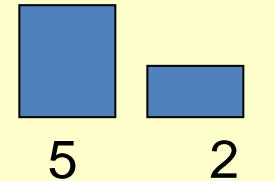
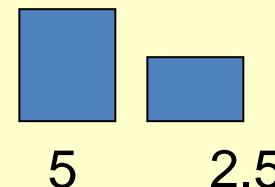


# 数据挖掘实例：分类问题

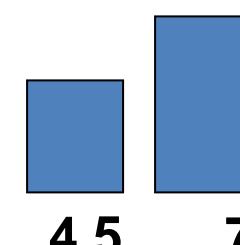
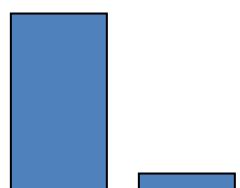
Examples of class A



Examples of class B

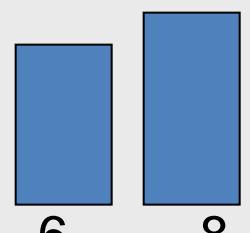


问题：如何分类？

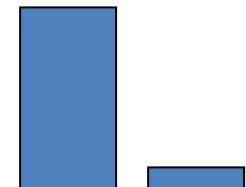
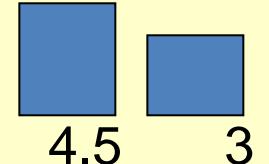
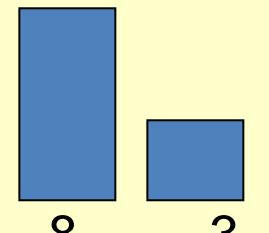
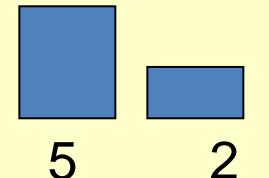
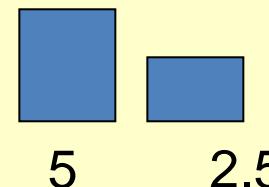


# 数据挖掘实例：分类问题

Examples of class A

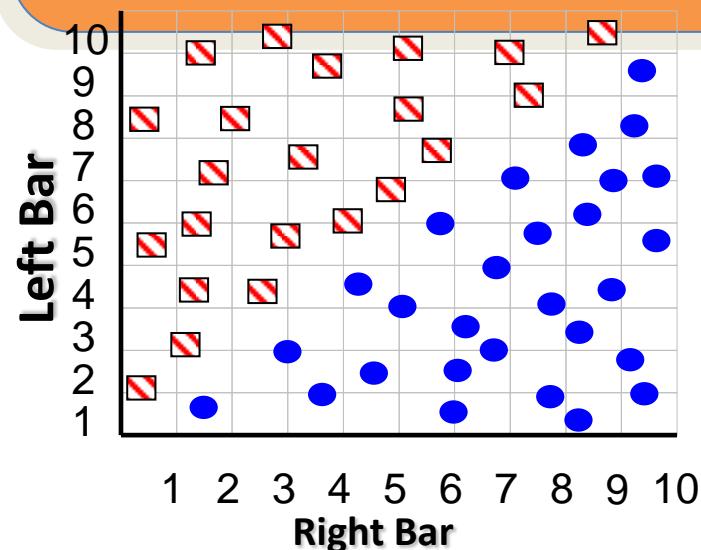


Examples of class B



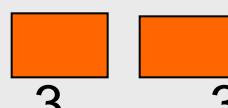
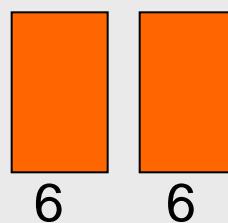
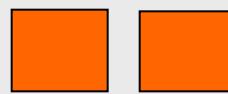
分类规则：

If the left bar is smaller than the right bar, it is an A  
otherwise it is a B.

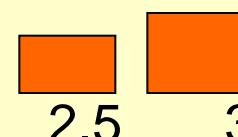
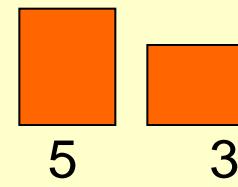
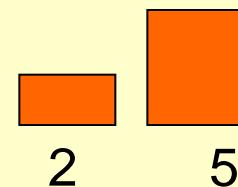
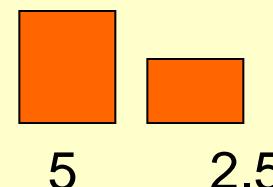


# 数据挖掘实例：分类问题

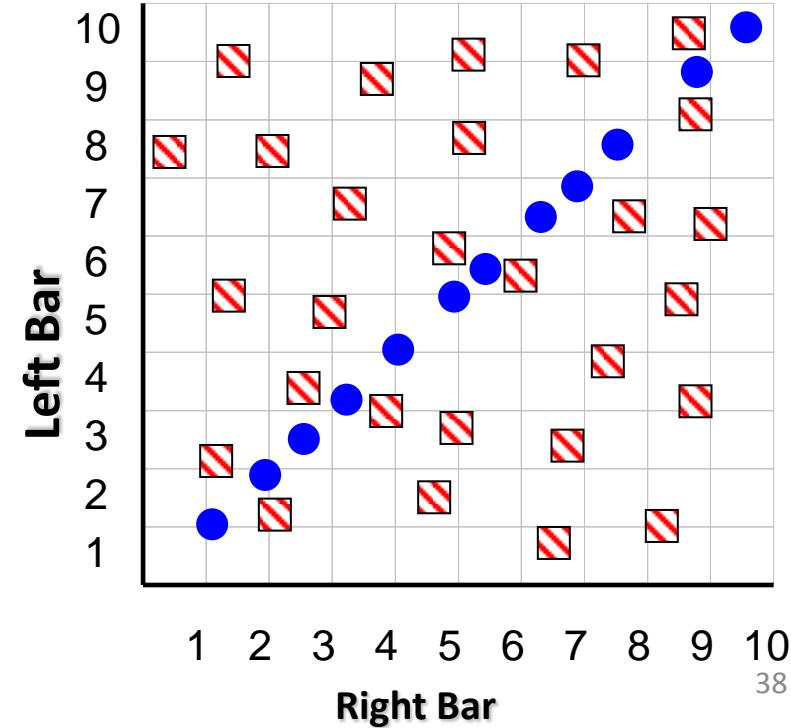
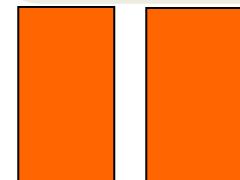
Examples of  
class A



Examples of  
class B

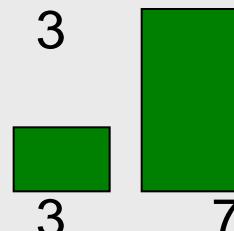
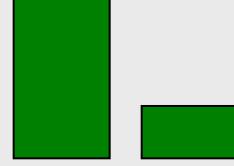
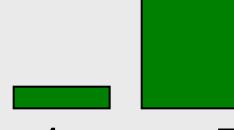
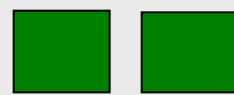


if the two bars are equal sizes, it is an A.  
Otherwise it is a B.

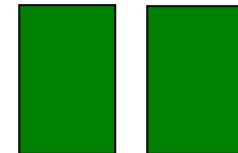
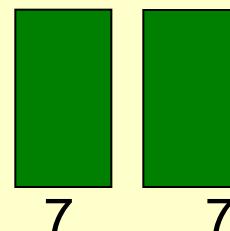
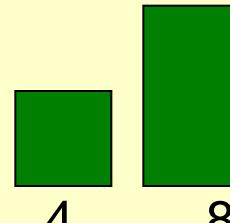
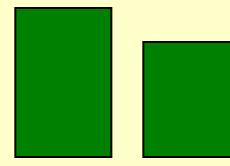
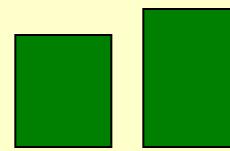


# 数据挖掘实例：分类问题

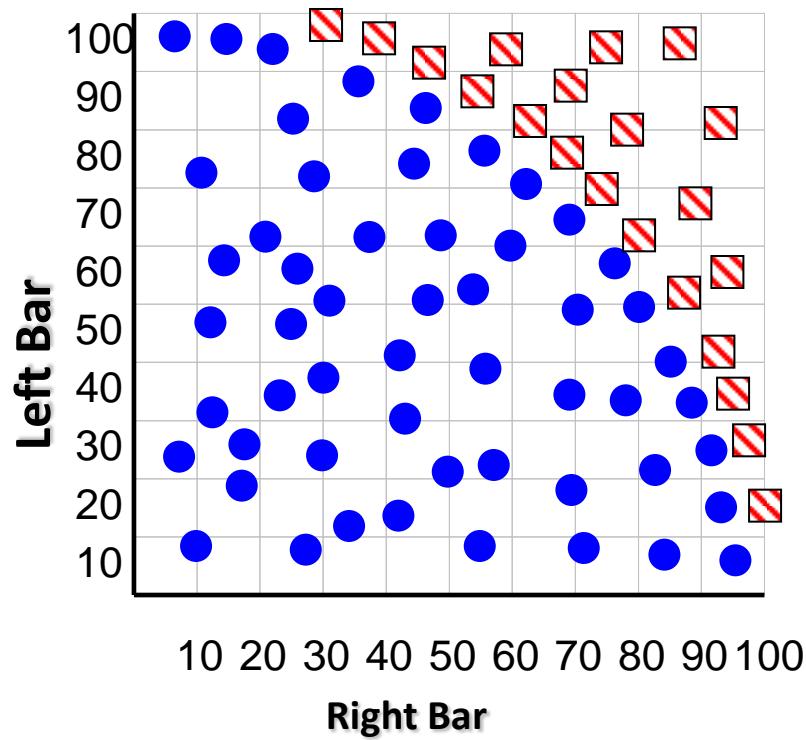
Examples of  
class A



Examples of  
class B



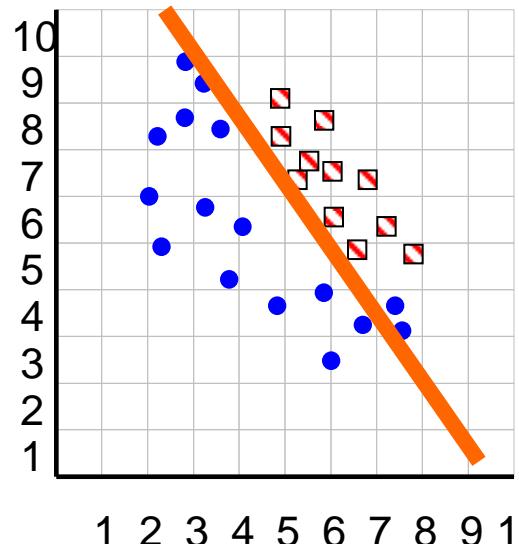
if the square of the  
sum of the two bars  
is less than or equal  
to 100, it is an A.  
Otherwise it is a B.



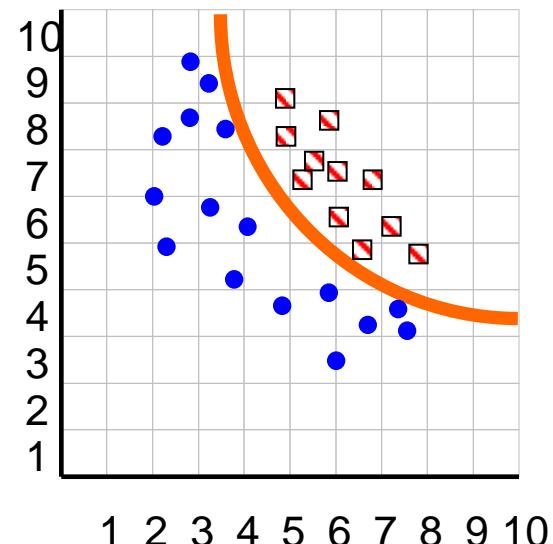
# 数据挖掘目标

目标: 找到 $f(x)$ 拟合已观测数据!

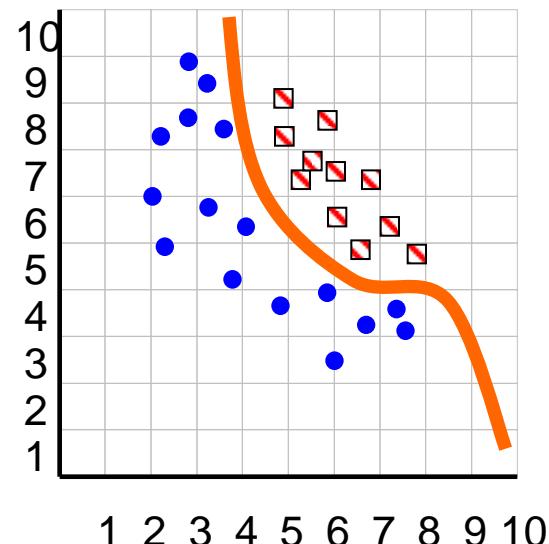
Accuracy = 94%



Accuracy = 100%



Accuracy = 100%



# 数据挖掘：分类问题

## ■ 按训练方式，机器学习可分为：

- (1) **有监督的学习**；有训练样本，学习机通过学习获得训练样本包含的知识，并用其作为判断测试样本的类别的依据。
- (2) **无监督的学习**：无训练样本，仅根据测试样本的在特征空间分布情况判断其类别。
- (3) **半监督的学习**：有少量训练样本，学习机以从训练样本获得的知识为基础，结合测试样本的分布情况逐步修正已有知识，并判断测试样本的类别。
- (4) **强化学习**：没有训练样本，但有对学习机每一步是否更接近目标的奖惩措施。

# 数据挖掘方法实例

- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

# 关联规则挖掘

- 关联规则挖掘发现大量数据中项集之间有趣的关联或相关联系。设  $I = \{i_1, i_2, \dots, i_m\}$  是项的集合。设任务相关的数据  $D$  是数据库事务的集合，其中每个事务  $T$  是项的集合，使得  $T \subseteq I$ 。设  $A$  是一个项集，事务  $T$  包含  $A$  当且仅当  $A \subseteq T$ 。



事先已知



事先未知

- 超市中什么产品会一起购买? — 啤酒和尿布
- 在买了一台PC之后下一步会购买?
- 哪种DNA对这种药物敏感?

定义：从数据集中找出对象或项集之间同时发生的关联或顺序关系。

# 数据挖掘方法实例

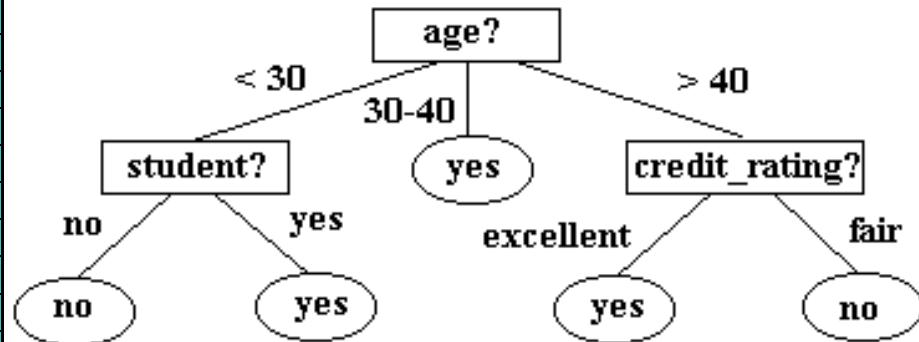
- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

# 数据挖掘方法-决策树

- 决策树学习是归纳推理算法。它是一种逼近离散函数的方法，且对噪声数据有很好的健壮性。在这种方法中学习到的知识被表示为决策树，决策树也能再被表示为多个if-then的规则，以提高可读性。

实例：通过年龄、收入、是否为学生、信用记录来预测用户是否会购买电脑

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no



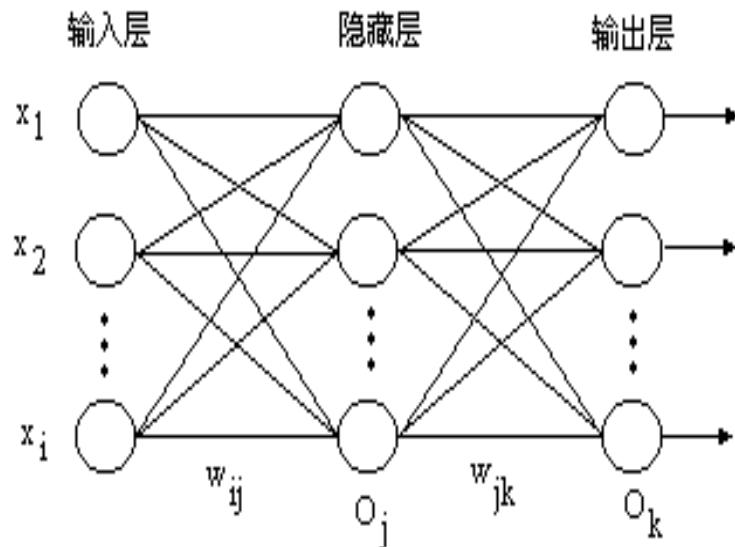
# 数据挖掘方法实例

- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

# 数据挖掘方法-神经网络

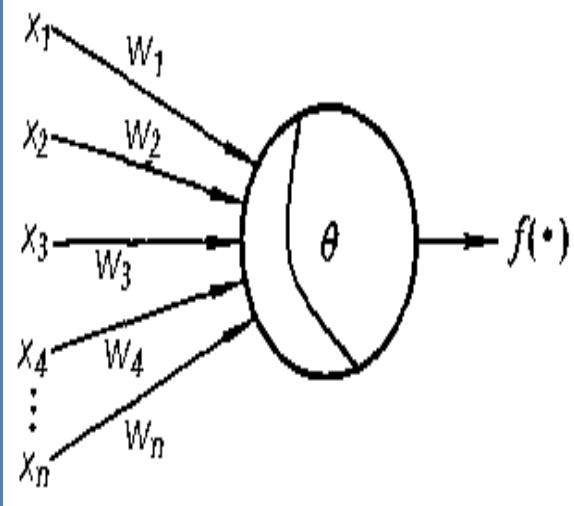
- 人工神经网络（Artificial Neural Networks）提供了一种普遍而且实用的方法，来从样例中学习值为实数、离散或向量的函数。
- 反向传播（Back Propagation）这样的算法使用梯度下降来调节网络参数以最佳拟合由输入/输出对组成的训练集合。
- BP网络的学习方法和目标：对网络的连接权值进行调整，使得对任一输入都能得到所期望的输出。





每个神经元都是一个结构相似的独立单元，它接受前一层传来的数据，并将这些数据的加权和输入非线性作用函数中，最后将非线性作用函数的输出结果传递给后一层。

常用的非线性作用函数是Sigmoid函数，即 $f(x) = 1 / (1 + e^{-x})$ 。在神经网络模型中，大量神经元节点按一定体系结构连接成网状。神经网络一般都具有输入层，隐层和输出层。



# 数据挖掘方法实例

- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

# 数据挖掘方法-朴素贝叶斯（Naive Bayes）分类器

- 朴素贝叶斯分类器是一种基于贝叶斯理论的分类器。它的特点是以概率形式表达所有形式的不确定，学习和推理都由概率规则实现，学习的结果可以解释为对不同可能的信任程度。
- $P(H)$ 是**先验概率**，或 $H$ 的先验概率。 $P(H|X)$ 是**后验概率**，或条件 $X$ 下， $H$ 的后验概率。后验概率 $P(H|X)$ 比先验概率 $P(H)$ 基于更多的信息。 $P(H)$ 是独立于 $X$ 的。

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} = \frac{P(X \mid H)P(H)}{\sum_H P(H)P(X \mid H)}$$

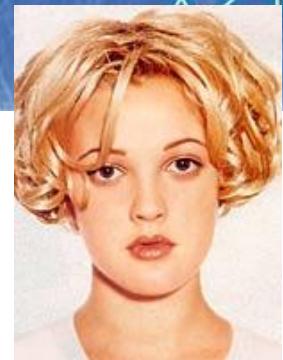
朴素贝叶斯分类能够奏效的前提是， $P(X|H)$  相对比较容易计算。假定 $X$ 表示红色和圆的， $H$ 表示假定 $X$ 是苹果；则 $P(X|H)$ 表示已知苹果，它既红又圆的概率。

假设有如下两个类别：

$$c_1 = \text{male}, \text{ and } c_2 = \text{female}.$$

预测名字为 “drew” 是 male or female, i.e  $p(\text{male} | \text{drew})$  or  $p(\text{female} | \text{drew})$  哪个大？

(Note: “Drew can be a male or female name”)



Drew Barrymore



Drew Carey

给定性别为 “male” ,名字为 “drew”的概率?

$$p(\text{male} | \text{drew}) = \frac{p(\text{drew} | \text{male}) p(\text{male})}{p(\text{drew})}$$

性别为 male 的概率?

名字为 “drew”的概率?  
(actually irrelevant, since it is that same for all classes)



Officer Drew

This is Officer Drew (who arrested me in 1997). Is Officer Drew a **Male** or **Female**?

假设数据库中已有名字和性别的对应数据。可以以此为已观测数据来应用贝叶斯分类器。

$$p(c_j | d) = \frac{p(d | c_j) p(c_j)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male



Officer Drew

$$p(c_j | d) = \frac{p(d | c_j) p(c_j)}{p(d)}$$

$$p(\text{male} | \text{drew}) = \frac{1/3 * 3/8}{3/8}$$

$$= 0.125$$

$\frac{3/8}{3/8}$

$$p(\text{female} | \text{drew}) = \frac{2/5 * 5/8}{3/8}$$

$$= 0.250$$

$\frac{3/8}{3/8}$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Officer Drew 的性  
别更可能为 **Female**.

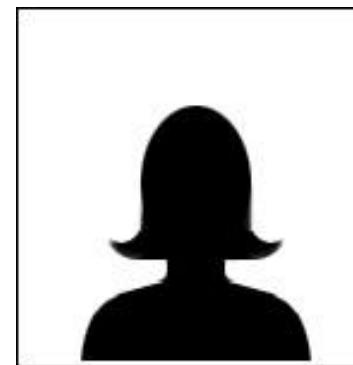
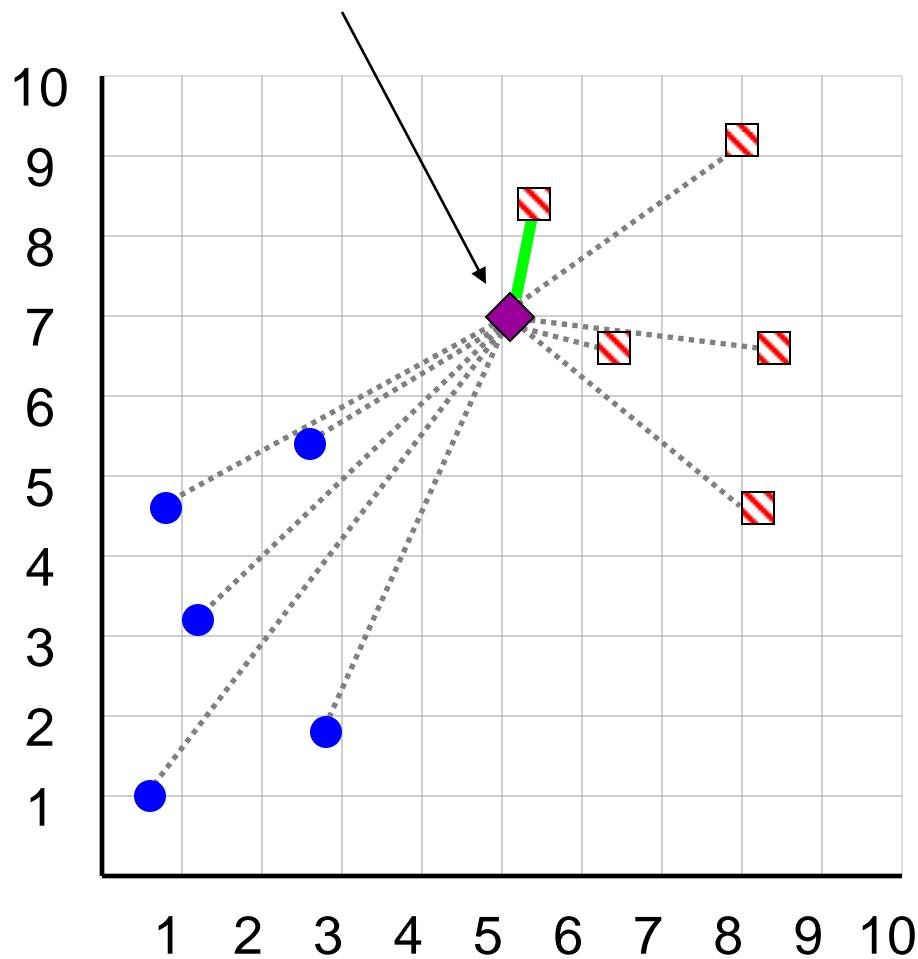
# 数据挖掘方法实例

- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

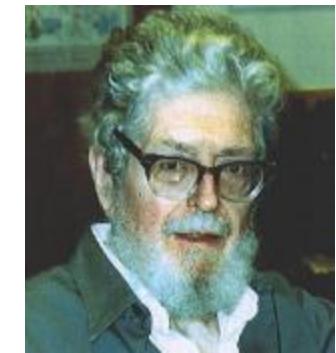
# 数据挖掘方法实例 K-最近邻分类

- K-近邻（K-NN）分类是基于范例的分类方法，它的基本思想是：给定待分类样本后，考虑在训练样本集中与该待分类样本距离最近（最相似）的K个样本，根据这K个样本中大多数样本所属的类别判定待分类样本的类别。
- 它的特例是1- NN，即分类时选出待分类样本的最近邻，并以此最近邻的类标记来判断样本的类。
- K-NN算法的优点在于它有较高的精确程度，研究表明，K-NN的分类效果要明显好于朴素贝叶斯分类、决策树分类。

# 数据挖掘方法实例 K-最近邻分类



Evelyn Fix  
1904-1965



Joe Hodges  
1922-2000

If the **nearest** instance to the  
**previously unseen instance** is a **A**  
class is **A**  
**else**  
class is **B**

■ A  
● B

# 数据挖掘方法实例

- 关联规则
- 决策树
- 人工神经网络
- 朴素贝叶斯分类器
- K近邻分类
- 聚类分析

# 聚类方法分类

- **基于层次的方法：**层次的方法对给定数据集合进行层次的分解。根据层次的分解如何形成，层次的方法可以被分为凝聚或分裂方法。（Chameleon，CURE，BIRCH）
- **基于密度的方法：**只要临近区域的密度超过某个阈值，就继续聚类。避免仅生成球状聚类。（DBSCAN，OPTICS，DENCLUE）
- **基于网格的方法：**基于网格的方法把对象空间量化为有限数目的单元，所有的聚类操作都在这个量化的空间上进行。这种方法的主要优点是它的处理速度很快。（STING，CLIQUE，WaveCluster）
- **基于模型的方法：**为每个簇假设一个模型，发现数据对模型的最好匹配。（COBWEB，CLASSIT，AutoClass）

# 聚类-定义相似度

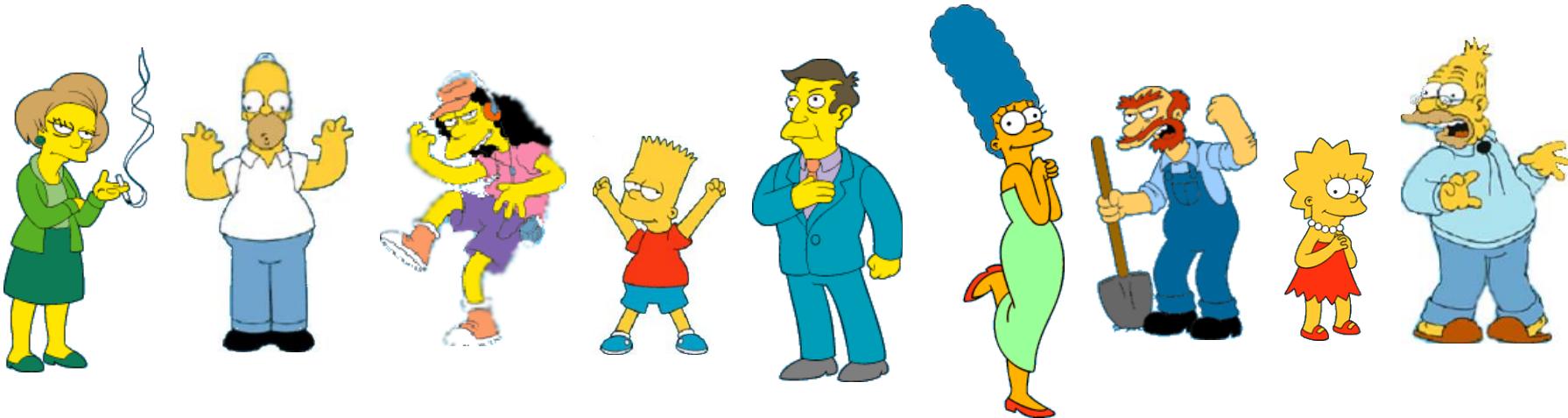
The quality or state of being similar; likeness; resemblance; as, a similarity of features.

Webster's Dictionary

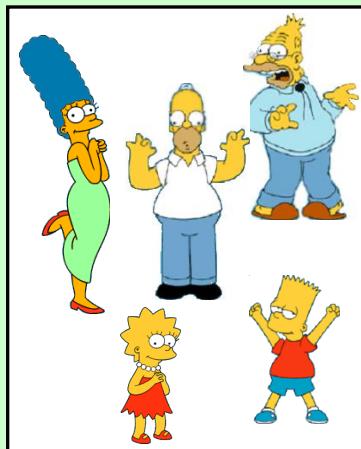


实体的相似度=实体特征的相似度

# 可能的聚类方式



Clustering is subjective !



Simpson's Family

School Employees

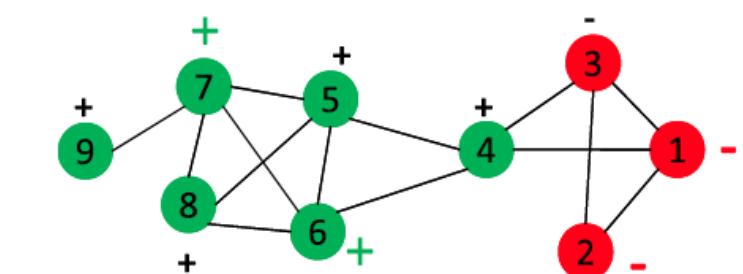
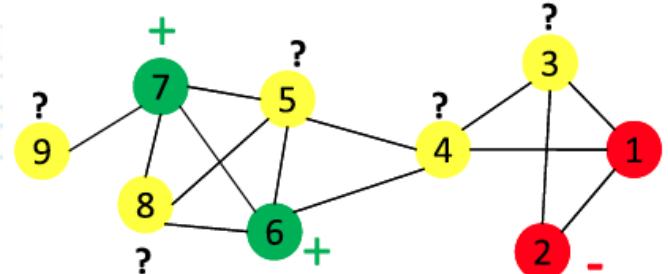
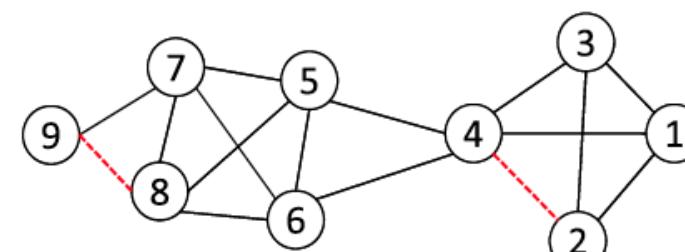
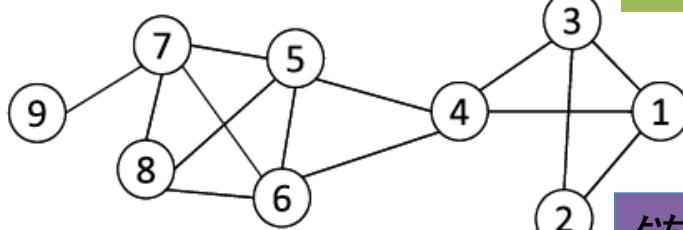
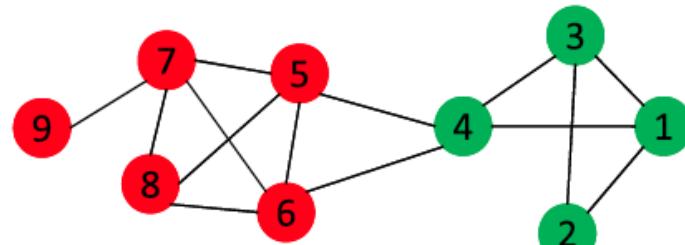
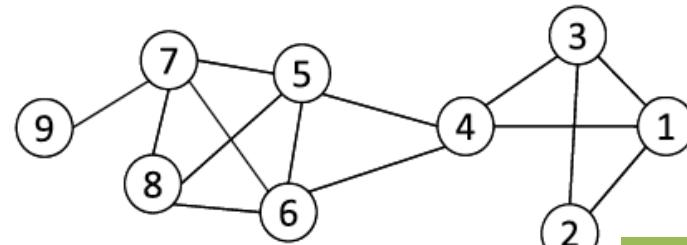
Females

Males

# 社交网络应用：分类、聚类和推荐

## ■ 数据挖掘技术在社交媒体分析中的应用

- Tag suggestion, Product/Friend/Group Recommendation



基于网络的分类

# 性能评估

## 分类精度评价指标

- 理想的分类器应该将所有属于某一类的样本标记为该类；且不将任何一个不属于该类的样本标记为该类。可以采有~~两个~~指标用来评价分类器的性能：准确率（查准率）和召回率（查全率）。对于某一特定类别 $C_i$ ，
- 准确率(P) = 
$$\frac{\text{分类属于 } C_i \text{ 且实际属于 } C_i \text{ 的样本数}}{\text{分类属于 } C_i \text{ 的样本数}}$$
- 召回率(R) = 
$$\frac{\text{分类属于 } C_i \text{ 且实际属于 } C_i \text{ 的样本数}}{\text{实际属于 } C_i \text{ 的样本数}}$$

# 分类精度评价指标（续）

- 对于同一分类器，准确率和查全率的变化趋势通常是相反的，片面追求其中一个指标而完全不顾及另一个是没有意义的。
- 为综合考虑准确率和查全率，可以使用一种能够全面评价分类器性能的指标：F-1。

$$\text{F-1} = \frac{2 \times \text{查准率} \times \text{查全率}}{\text{查准率} + \text{查全率}}$$

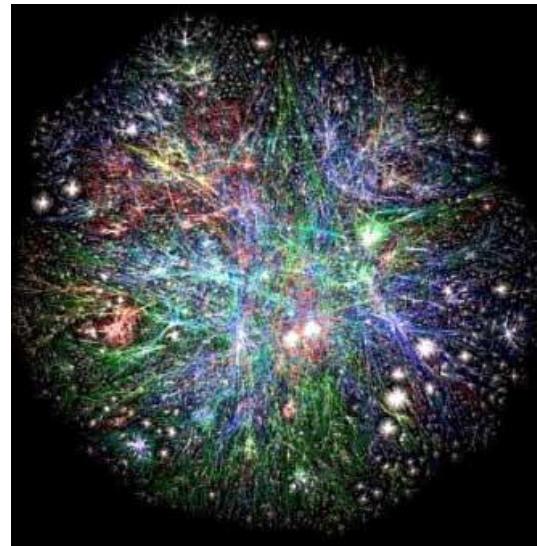
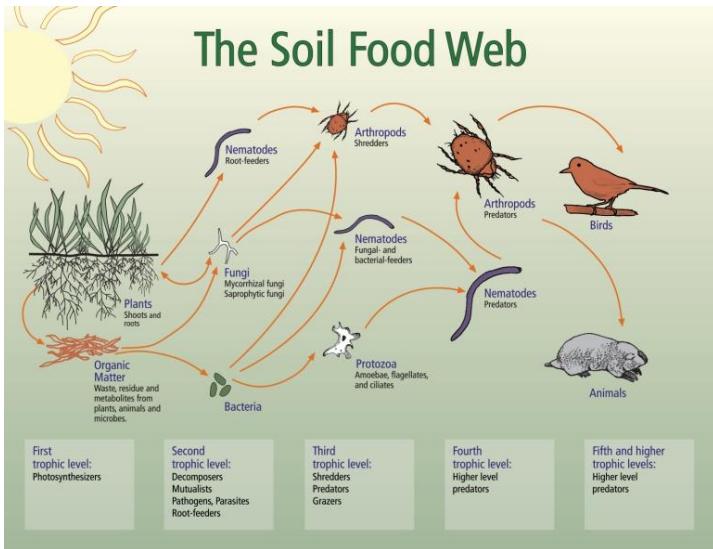
- F-1综合考虑了上述两指标，且偏向于准确率和查全率中较小的一个，只有当准确率和查全率都较大时，F-1指标才会比较大。

# 分类精度评价指标（续）

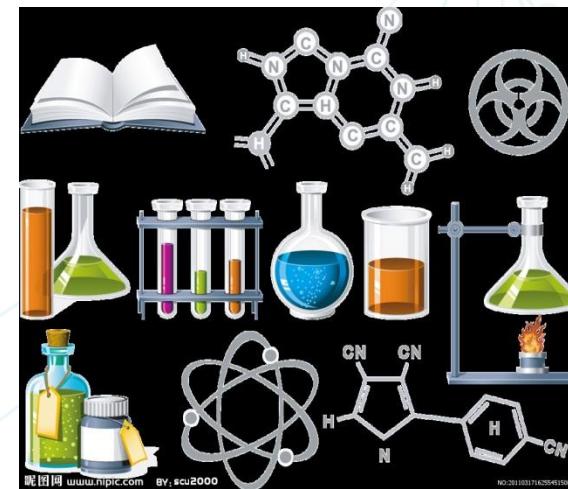
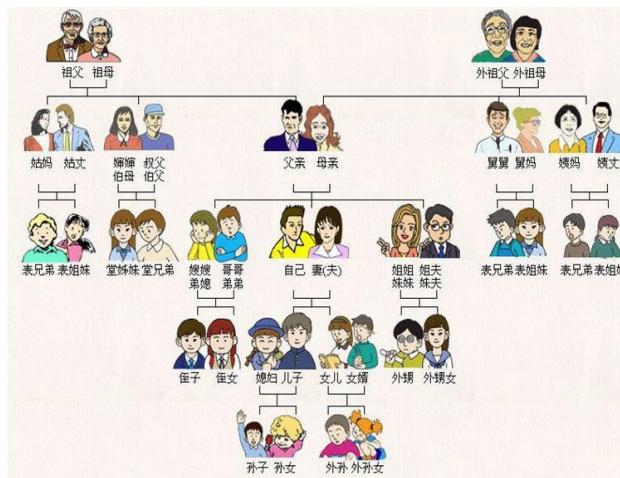
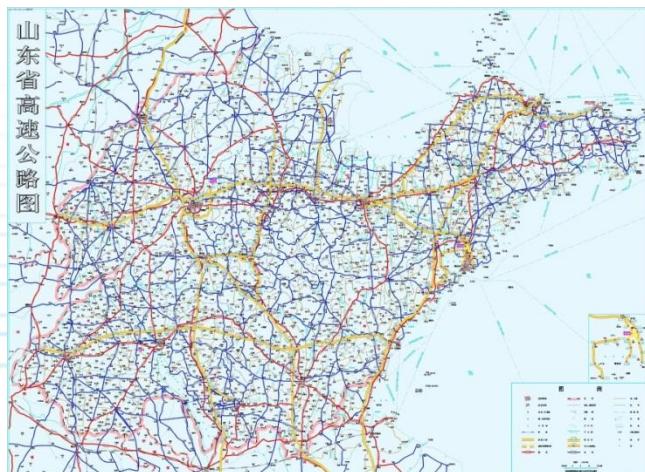
- 多数分类器可以通过调整参数获得不同的准确率和查全率，当分类器的参数调节到正好使准确率和查全率相等时，该值称为P/R无损耗(平衡)点。它也是一种综合考虑准确率和查全率的指标。
- 在综合考虑全部类别的条件下，精确度(Accuracy)也是一个常用的指标，它是指所有分类正确的样本数在所有样本中所占的比例。
- 精确度( $A$ ) = 
$$\frac{\text{所有分类标记正确的样本数}}{\text{全部样本总数}}$$

# 内容提要

- 大数据的分析现状
- 数据的处理流程及处理技术
- 数据挖掘算法
- 图数据分析处理技术

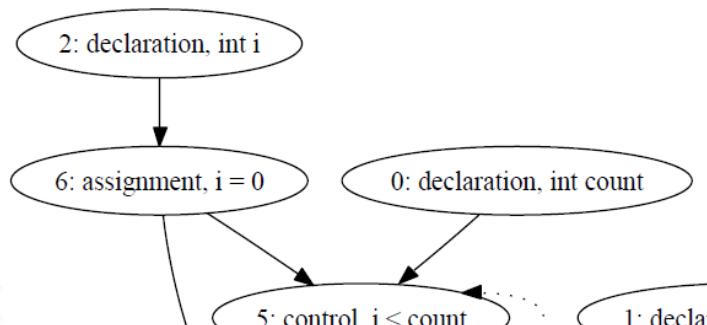


Graphs are **everywhere**, and quite a few are **huge** graphs!



# 图的应用案例—软件剽窃检测 [1]

- 传统的软件剽窃检测工具难以查出隐藏较深的剽窃
- 基于图模式匹配的新型软件剽窃检测工具
  - 将源代码表示为程序依赖图 (program dependence graphs)<sup>[2]</sup>.
  - 工作原理：可以对别人的程序做改动，但程序的逻辑结构难以改变



源代码

```

int sum(int array[], int count)
{
    int i, sum;
    sum = 0;
    for(i = 0; i < count; i++) {
        sum = add(sum, array[i]);
    }
    return sum;
}

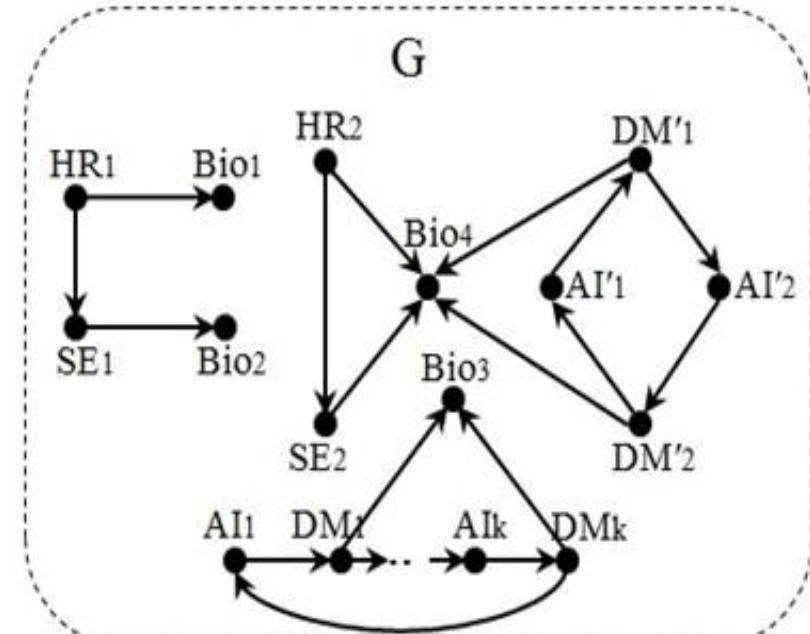
int add(int a, int b)
{
    return a + b;
}
  
```

程序依赖图

# 图的应用案例—推荐系统<sup>[3]</sup>

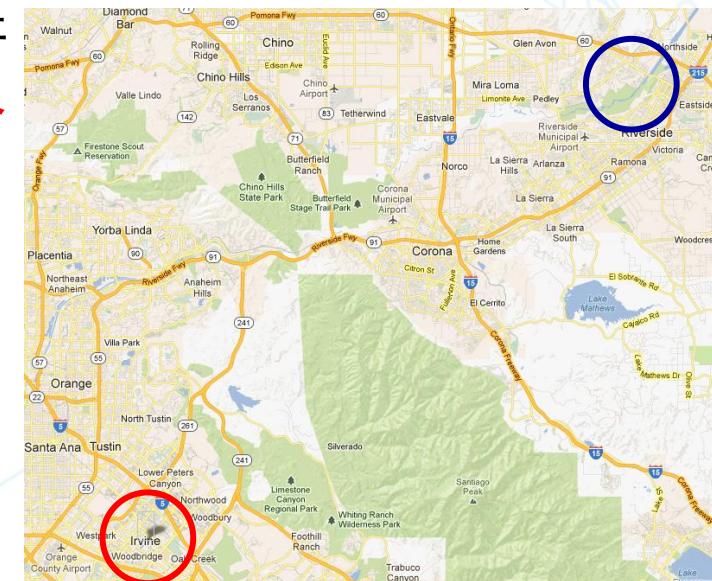
- 推荐在很多新型领域里有着应用。
  - Graph Search 是一种非常有用的推荐工具。

- 一个猎头 A 想找一个生物学家 (Bio) 来帮助一群软件工程师 (SEs) 来分析基因数据
- 猎头将需求抽象成一个查询图 Q, 然后在一个专家推荐网络 G 中查找：
  - ✓ G 中顶点标有一个人的专长
  - ✓ G 中边表示推荐关系, 如  $HR_1$  推荐  $Bio_1$ , and  $AI_1$  推荐  $DM_1$ 。



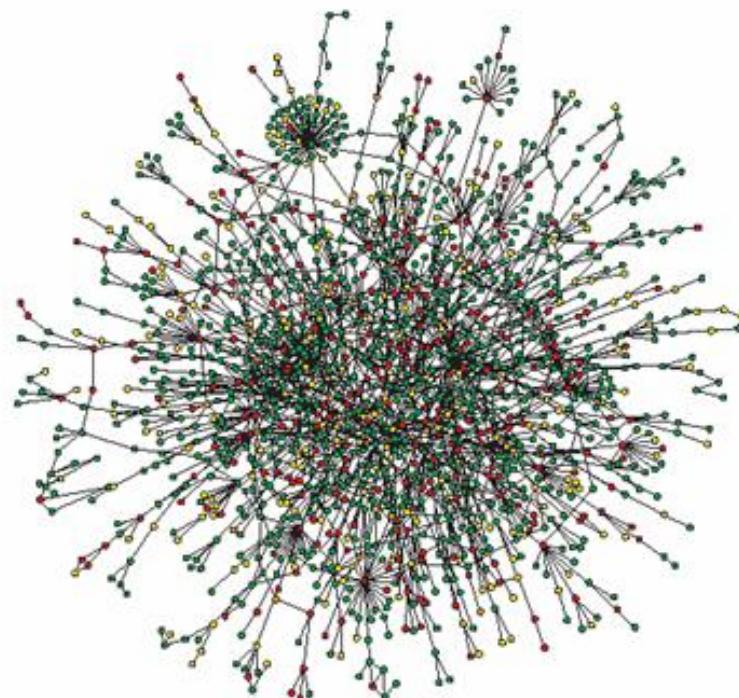
# 图的应用案例—交通路由 [4]

- 由于基于位置服务(Location-Based Services)的广泛应用，Graph search在交通网络中也得到了应用。
- 示例：一位美国司机Mark想从加州的Irvine开车到Riverside。
  - 如果Mark想驾驶自己的小轿车在最短的时间内到达Riverside, 这个问题可以抽象为图上的最短路径问题。通过最短路径算法，可以找出沿着State Route 261可以最快的从 *Irvine, CA* 到达 *Riverside, CA*。
  - 如果Mark向驾驶一辆大卡车运输危险物品，那么有可能一些桥和铁路的交接点是不允许通过的。为了寻找最优的路径，需要用一个模式图来表达约束关系（如正则表达式等）



# 图的应用案例——生物数据分析 [5]

- 很多生物数据可以表示为图，然后用Graph Search的技术来分析生物数据。
  - 蛋白质交互网络（Protein-interaction network，PIN）分析能够提供有机体的功能组织和演化行为非常有价值的内在洞察。



# The need of Graph Search



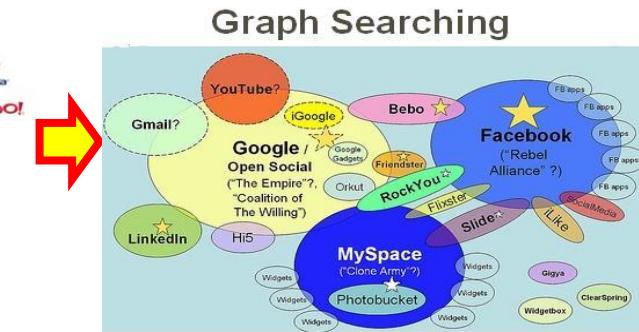
File systems



Databases



World Wide Web



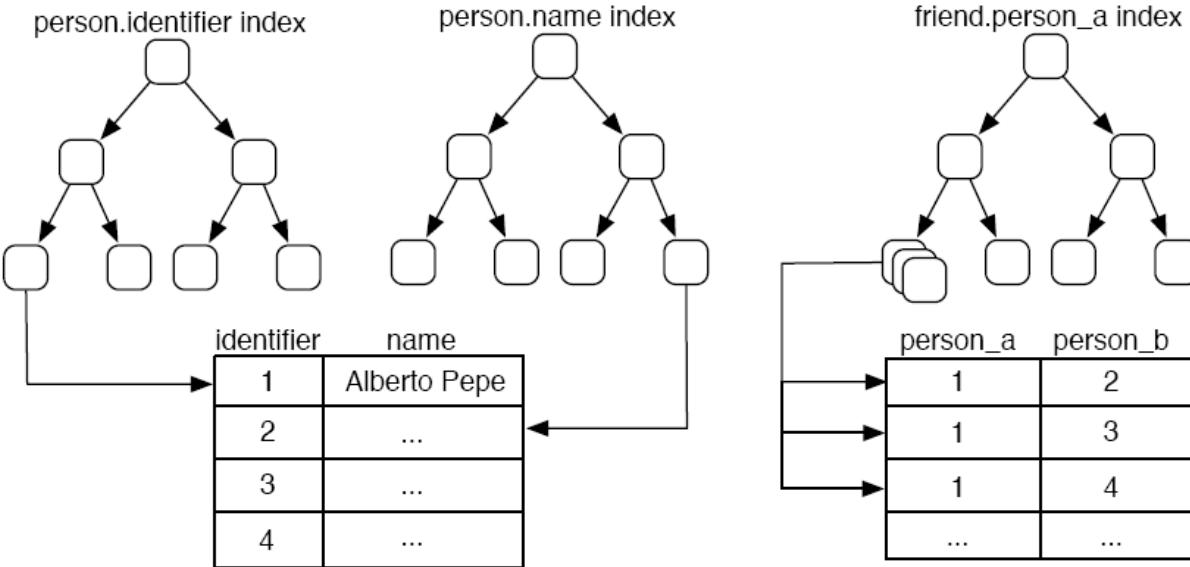
- **File systems - 1960's:** very simple search functionalities
- **Databases - mid 1960's:** SQL language
- **World Wide Web - 1990's:** keyword search engines
- **Social networks - late 1990's:**

Facebook launched “graph search” on 16<sup>th</sup> January, 2013

Assault on Google, Yelp, and LinkedIn with new graph search;  
Yelp was down more than 7%

Graph search is a new paradigm for social computing!

# Graph Search vs. RDBMS<sup>[6]</sup>



**查询:** 找出Alberto Pepe  
所有朋友的名字

## Step 1:

The person.name index  $\rightarrow$  the identifier of Alberto Pepe.  $[O(\log_2 n)]$

## Step 2:

The friend.person index  $\rightarrow$  k friend identifiers.

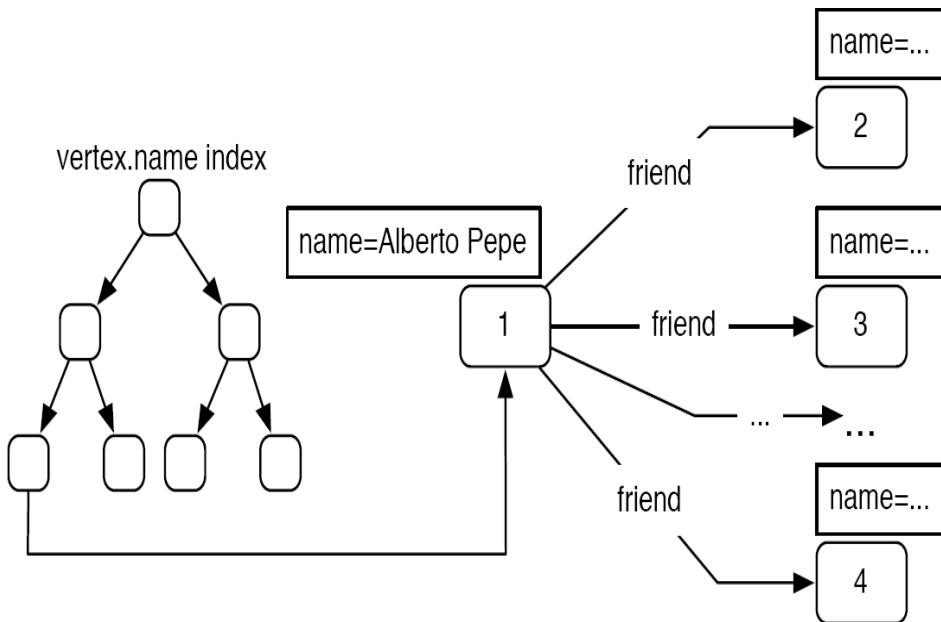
$[O(\log_2 x) : x \ll m]$

## Step 3:

The k friend identifiers  $\rightarrow$  k friend names.

$[O(k \log_2 n)]$

# Graph Search vs. RDBMS<sup>[6]</sup>



查询：找出Alberto Pepe  
所有朋友的名字

## Step 1:

The vertex.name index -> the vertex with the name Alberto Pepe. [O(log2n)]

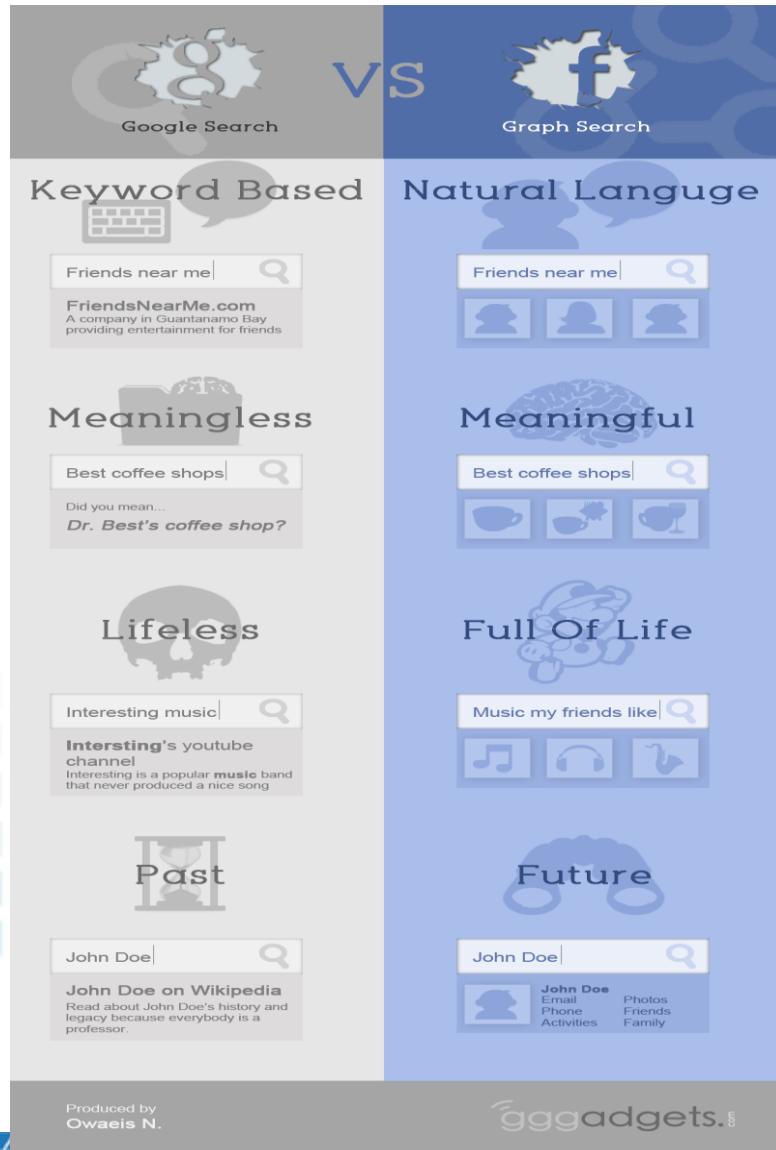
## Step 2:

The vertex returned -> the k friend names.

[O(k + x)]

图搜索要比关系数据库查询效率高的多！

# Graph Search vs. Web Search



It's interesting, and **over the last 10 years, people have been trained on how to use search engines more effectively.**

**Keywords & Search In 2013:  
Interview With A. Goodman & M. Wagner**

## ■ 但是：

- 关键词 vs. 短语、句子
- 简单的网页 vs. 实体
- 没有生命特征 vs. 反应生命特征
- 历史 vs. 未来

# 什么是图搜索(Graph Search)?

我们提出一个统一的定义 [7] (in the name of graph matching):

- 给定模式 $G_p$ 和数据图 $G$ :
  - 检查是否 $G_p$ 匹配 $G$ ; and
  - 找出 $G_p$ 匹配的 $G$ 中的所有子图.

注:

- 两类查询:
  - 布尔查询 (Yes or No)
  - 函数查询, 又可能借助于布尔查询 (作为一个子程序)
- 图包括顶点和边, 通常带有标签
- 通常模式图比较小 (如10个顶点), 数据图很大 (如 $10^8$ 个顶点)

# 国内外研究现状及发展动态分析

## ■ 研究机构和工业界越来越重视

- Microsoft的Dryad项目<sup>[8]</sup>
- Google的MapReduce<sup>[10]</sup>和Pregel<sup>[9]</sup>
- Neo4j: the graph database<sup>[11]</sup>
- UCSB, Edinburgh...
- 北京航空航天大学,北京大学,清华大学, 哈尔滨工业大学, ...

## ■ 大规模动态图的查询技术

- 新型查询语言（PTIME）及其查询算法
- 图数据的分布式查询利用多节点计算能力解决图的规模
- 图的表示（压缩）来进一步解决图的单节点规模问题
- 图数据的分布策略提高查询的性能
- 增量(查询、表示和分布)算法来解决图的动态变化问题

# 图搜索应用及挑战

## ■ 图的表达能力丰富，应用广泛

- VLSI设计<sup>[12]</sup>、蛋白质交互网络<sup>[13]</sup>、模式识别<sup>[14]</sup>
- 程序代码剽窃检测<sup>[15]</sup>，垃圾E-mail检测<sup>[16]</sup>
- 虚拟机映射<sup>[17,18]</sup>、数据清洗<sup>[19]</sup>、网页聚类<sup>[20]</sup>
- 数据库模式匹配<sup>[21]</sup>、社交网络<sup>[22]</sup>、Web网络<sup>[23]</sup>等

## ■ 其中社会图数据是一类典型的大数据

- 图规模越来越大，并且图经常动态变化
  - Facebook用户超过17亿，每天新增用户600K<sup>[24]</sup>
  - Twitter用户超过10亿，每天新增用户300K<sup>[25]</sup>
  - 人人网的用户数量总计已经超过1.2亿<sup>[26]</sup>

## ■ 这些新特性对图的查询技术提出了新的挑战，但目前这方面 系统的研究工作比较欠缺

# 研究内容

- 图的新型查询语言
- 图的动态分布查询
- 图的动态表示方式
- 图的动态分布策略
- 图数据相关系统

# 图的新型查询语言

- 针对某一特定应用的查询语言
  - 最短路径、图的可达性
  - 子图同构、图模拟、强模拟、
- 通用语言：类似SQL查询语言
  - 如：SPARQL Query Language for RDF等

Weiren Yu, Charu Aggarwal, Shuai Ma, and Haixun Wang, On Anomalous Hot Spot Discovery in Graph Streams. ICDM 2013.

# 图的新型查询语言

## ■ 目前的查询语言

- 子图同构查询<sup>[27,28]</sup>

- 保持结构，但计算复杂性过高 (NP-complete)

- 图模拟查询<sup>[29,30]</sup>

- 计算复杂性低 (Quadratic time)，但结构丢失严重

- 图模拟扩展查询<sup>[31,32]</sup>

- 计算复杂性低 (Cubic time)，但结构同样丢失严重

## ■ 一种局部保持的图模拟查询语言-强模拟

- 查询能力介于子图同构(Sub-graph Isomorphism)和图模拟(Graph Simulation)，尽量保持更多的结构

- 计算复杂性保持在PTIME，复杂度不要超过图模拟过多

Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Capturing Topology in Graph Pattern Matching, VLDB 2012.

Shuai Ma, Yang Cao, Wenfei Fan, Jinpeng Huai, and Tianyu Wo, Strong Simulation: Capturing Topology in Graph Pattern Matching, TODS 2014.

# 图的动态表示方式

- 图的不同表示方式所需存储空间差别很大
- 以常用的邻接表和邻接矩阵为例
  - 对**稠密图**, 邻接矩阵没有浪费多少空间, 并且提供 $O(1)$ 的时间复杂度来判断两个顶点是否有边。
  - 对**稀疏图**, 邻接表通常要较好一些。
  - 对一个有100,000个顶点, 200,000条边的图:
    - 采用**邻接矩阵**需要 $10,000 \times 10,000$ 项, 共需要**40G**, 如果每项需要4个字节(整数)
    - 采用**邻接表**需要10,000项来存顶点, 200,000项来存边, 共需要**1.2M**, 如果每项需要4个字节(整数)
    - 前者所需空间是后者的**30,000倍**

# 图的动态表示方式

## ■ 目前图的表示方式

- 很多关于静态图表示的研究[33-54]
- 大部分是面向特定查询的压缩
- 很多是满足社交网络和Web的需求

## ■ 目前动态图的研究

- 动态图算法考虑**查询和更新的Trade-Off**[31, 55-61]
- Thomas Reps增量算法做了系统研究，并提出了系统的**增量算法的复杂性**[59,60,61]

## ■ 大规模图的动态表示

- 面向特定查询的动态图表示方式，
- 更新代价与存储空间的**Trade-Off**（增量算法复杂性等）
  - 社交网络更新频繁、规模大[24,25,26]
- 目前研究比较欠缺，稍微相关的文献[55]研究时态网络的表示

# 图的动态分布策略

## ■ 分布策略对存储空间的影响

- 跨越划分的边需要冗余存储
- 可以抽象为Minimum K-Cut问题
  - NP-complete<sup>[66]</sup>
  - 当K固定，是PTIME的<sup>[64]</sup>
  - 存在近似度为2的近似算法<sup>[65]</sup>

## ■ 分布策略对查询代价的影响

- 多数查询要求Load Balancing
- 跨越划分的边的数目通常与查询代价成正比
- 可以抽象为Balanced Graph Partition问题
  - 将图划分成基本上大体相等的k个部分s
  - NP-complete，当K固定，仍然是NP-complete<sup>[66]</sup>
  - K=2，是图Bisection问题，NP-complete<sup>[66]</sup>，这样当K固定，仍然是NP-complete<sup>[66]</sup>

# 图的动态分布策略

## ■ 分布策略对更新代价的影响

- 更新需要维护图的完整性
  - 边界点和边的维护问题
- 有可能导致数据的重新划分
  - 是个**动态分布式更新**问题，而不是简单的动态图划分

## ■ 分布策略对图表示的影响

- 不同的图表示的更新代价是不一样的
- 需要兼顾图表示（压缩）对分布式更新的影响

# 图的动态分布策略

## ■ 目前数据分布的研究

- 图的划分和动态划分<sup>[64-75]</sup>

- 大量的研究工作，在理论和系统两个方面的研究都很活跃

- 分布式数据库的划分策略<sup>[76]</sup>

## ■ 大规模图的动态分布策略

- 面向特定查询的图的动态分布

- 考虑更新代价与图表示和分布策略三者之间的Trade-Off

- Web网和社交网络更新频繁、规模大<sup>[24,25,26]</sup>

- 虽然（动态）图划分的工作很多，但对分布式动态图的工作和考虑图表示的动态图划分欠缺

# 图的动态分布查询

## ■ 目前研究工作

- 图数据的查询<sup>[63, 80]</sup>
- 并行图查询<sup>[84]</sup>
- 分布式图查询<sup>[79-83]</sup>
- 分布式数据库查询<sup>[76, 77]</sup>

## ■ 大规模动态图数据的分布式查询

- 面向特定查询的动态图分布式查询
- 集成图的表示、分布策略和增量查询方法
- 目前大多研究上面所提到的某一个或两个点，**缺少系统全面的研究**

Shuai Ma, Yang Cao, Jinpeng Huai, and Tianyu Wo, Distributed Graph Pattern Matching. WWW 2012.

# 图的动态分布查询

## ■ 三类图

- Trees
- DAGs
- General graphs with cycles

## ■ 查询语言

- 邻接/可达性查询
- 图模拟查询
- 局部保持的图模拟查询

## ■ 查询语言的查询优化技术

- 建立索引
- 查询重写

# 图数据相关系统

- Neo4J图数据管理系统
- Trinity – MSRA
  - <http://research.microsoft.com/en-us/projects/trinity/>
- Giraph-Google Pregel的开源版
  - <http://giraph.apache.org/>
- 更多参见Graph database
  - [http://en.wikipedia.org/wiki/Graph\\_database](http://en.wikipedia.org/wiki/Graph_database)

系统顶级会议OSDI 2012共录用25篇论文，而其中有三篇是关于图数据的的系统。

# 参考文献(图数据部分)

1. Chao Liu, Chen Chen, Jiawei Han and Philip S. Yu, GPLAG: detection of software plagiarism by program dependence graph analysis. KDD 2006.
2. Ferrante, K. J. Ottenstein, and J. D. Warren. The program dependence graph and its use in optimization. ACM Trans. Program. Lang. Syst., 9(3):319–349, 1987.
3. Shuai Ma, Yang Cao, Jinpeng Huai, and Tianyu Wo, Distributed Graph Pattern Matching, WWW 2012.
4. Rice, M. and Tsotras, V.J., Graph indexing of road networks for shortest path queries with label restrictions, VLDB 2010.
5. David A. Bader and Kamesh Madduri, A graph-theoretic analysis of the human protein-interaction network using multicore parallel algorithms. Parallel Computing 2008.
6. Marko A. Rodriguez, Peter Neubauer: The Graph Traversal Pattern. Graph Data Management 2011: 29-46
7. Shuai Ma, Yang Cao, Tianyu Wo, and Jinpeng Huai, Social Networks and Graph Matching. Communications of CCF, 2012.
8. <http://research.microsoft.com/en-us/projects/dryad/>
9. Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser and Grzegorz Czajkowski, Pregel: a system for large-scale graph processing, SIGMOD 2010.
10. Jeffrey Dean and Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, OSDI 2004.
11. Neo4j: the graph database, <http://neo4j.org/>

12. George Karypis, Rajat Aggarwal, Vipin Kumar and Shashi Shekhar, Multilevel hypergraph partitioning: applications in VLSI domain, IEEE Trans. VLSI Syst. 1999
13. Patrick Durand, Laurent Labarre, Alain Meil, Jean-Louis Divol, Yves Vandenbrouck, Alain Viari and Jerome Wojcik, GenoLink: a graph-based querying and browsing system for investigating the function of genes and proteins, BMC Bioinformatics 2006
14. Donatello Conte, Pasquale Foggia, Carlo Sansone and Mario Vento, Thirty Years Of Graph Matching In Pattern Recognition, IJPRAI 2004
15. Chao Liu, Chen Chen, Jiawei Han and Philip S. Yu, GPLAG: detection of software plagiarism by program dependence graph analysis, KDD 2006
16. Manu Aery and Sharma Chakravarthy, eMailSift: Email Classification Based on Structure and Content, ICDM 2005
17. N. M. Mosharaf Kabir Chowdhury, Muntasir Raihan Rahman and Raouf Boutaba, Virtual Network Embedding with Coordinated Node and Link Mapping, INFOCOM 2009
18. N. M. Mosharaf Kabir Chowdhury and Raouf Boutaba, A survey of network virtualization, Computer Networks 2010
19. Wenfei, Jianzhong Li, Shuai Ma, Hongzhi Wang and Yinghui Wu, Graph Homomorphism Revisited for Graph Matching, PVLDB 2010
20. Adam Schenker, Mark Last, Horst Bunke and Abraham Kandel, Classification Of Web Documents Using Graph Matching, IJPRAI 2004
21. Erhard Rahm and Philip A. Bernstein, A survey of approaches to automatic schema matching, VLDB J. 2001
22. Yuanyuan Tian and Jignesh M. Patel, TALE: A Tool for Approximate Large Graph Matching, ICDE 2008
23. Raghavan and H. Garcia-Molina, Representing Web graphs, ICDE, 2003.

24. <http://www.insidefacebook.com/2009/02/14/facebook-surpasses-175-million-users-continuing-to-grow-by-600k-usersday/>
25. <http://techcrunch.com/2010/04/14/twitter-has-105779710-registered-users-adding-300k-a-day/>
26. <http://news.cnlist.com/CnlistNewsDetail.aspx?tablename=gsbd&GUID={99DE1B2E-0F17-4CB4-8B49-FAE881D02D59}>
27. Brian Gallaghe, Matching structure and semantics: A survey on graph-based pattern matching, AAAI FS., 2006
28. Ullmann, J. R., An Algorithm for Subgraph Isomorphism, JACM, 1976
29. M. R. Henzinger, T. Henzinger and P. Kopke, Computing simulations on finite and infinite graphs, FOCS, 1995
30. Francesco Ranzato and Francesco Tapparo, An efficient simulation algorithm based on abstract interpretation, Information Computation, 2010
31. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, Yinghui Wu and Yunpeng Wu, Graph Pattern Matching: From Intractable to Polynomial Time, PVLDB 2010
32. Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Yinghui Wu, Adding Regular Expressions to Graph Reachability and Pattern Queries, ICDE 2011.
33. S. Navlakha, R. Rastogi, and N. Srivastava, Graph summarization with bounded error, SIGMOD 2008
34. M. Adler and M. Mitzenmacher, Towards compressing Web graphs, DCC 2001

35. K. H. Randall, R. Stata, J. L. Wiener, and R. Wickremesinghe, The link database: Fast access to graphs of the Web, DCC 2002
36. K. Bharat, A. Z. Broder, M. R. Henzinger, P. Kumar, and S. Venkatasubramanian, The connectivity server: Fast access to linkage information on the web, Computer Networks, vol. 30, no. 1-7, 1998.
37. G. Buehrer and K. Chellapilla, A scalable pattern mining approach to Web graph compression with communities, WSDM 2008.
38. K C. Karande, K. Chellapilla, and R. Andersen, Speeding up algorithms on compressed Web graphs, WSDM, 2009
39. K Yasuhito Asano , Yuya Miyawaki , Takao Nishizeki, Efficient Compression of Web Graphs, COCOON 2008
40. K T. Feder and R. Motwani, Clique partitions, graph compression and speeding-up algorithms, J. Comput. Syst. Sci., vol. 51, no. 2, 1995.
41. K Jérémie Barbay, Francisco Claude and Gonzalo Navarro, Compact Rich-Functional Binary Relation Representations, LATIN 2010
42. K Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, Michael Mitzenmacher, Alessandro Panconesi and Prabhakar Raghavan, On compressing social networks, KDD 2009
43. K Hossein Maserrat and Jian Pei, Neighbor query friendly compression of social networks, KDD 2010
44. K Maria Giatsoglou, Symeon Papadopoulos and Athena Vakali, Massive Graph Management for the Web and Web 2.0, New Directions in Web Data Management 1, Springer, 2011.
45. K Nieves R. Brisaboa, Susana Ladra and Gonzalo Navarro, k2-Trees for Compact Web Graph Representation, SPIRE 2009

46. N. Jesper Larsson and Alistair Moffat, Offline Dictionary-Based Compression, DCC1999
47. Paolo Boldi and Sebastiano Vigna, The WebGraph Framework I: Compression Techniques, WWW 2004
48. Paolo Boldi and Sebastiano Vigna, The WebGraph Framework II: Codes For The World-Wide Web, DCC 2004
49. Paolo Boldi, Massimo Santini and Sebastiano Vigna, Permuting Web Graphs, WAW, 2009
50. Alberto Apostolico and Guido Drovandi, Graph Compression by BFS, Algorithms 2009
51. Alireza Mahdian, Hamid Khalili, Ehsan Nourbakhsh and Mohammad Ghodsi, Web Graph Compression by Edge Elimination, DCC 2006
52. Ming-Yang Kao, Neill Occhiogrosso and Shang-Hua Teng, Simple and Efficient Graph Compression Schemes for Dense and Complement Graphs, J. Comb. Optim. 1998
53. Torsten Suel and Jun Yuan, Compressing the Graph Structure of the Web, DCC 2001
54. Tomas Feder, Adam Meyerson, Rajeev Motwani, Liadan O'Callaghan, Rina Panigrahy, Representing Graph Metrics with Fewest Edges, STACS 2003
55. Kamesh Madduri and David A. Bader, Compact graph representations and parallel connectivity algorithms for massive dynamic network analysis, IPDPS 2009

56. C. Demetrescu, I. Finocchi, and G. Italiano, Dynamic Graph Algorithms, Handbook of Graph Theory, J. Yellen e J.L. Gross eds., CRC Press Series, in Discrete Mathematics and Its Applications, 2003.
57. Camil Demetrescu and Giuseppe F. Italiano, Trade-Offs for Dynamic Graph Problems, Encyclopedia of Algorithms, Springer, 2008.
58. Ming-Yang Kao, Encyclopedia of Algorithms, Springer 2008.
59. G. Ramalingam and Thomas Reps, An incremental algorithm for a generalization of the shortest-path problem, J. Algorithms, 1996
60. G. Ramalingam and Thomas Reps, On the computational complexity of dynamic graph problems, TCS 1996
61. G. Ramalingam and Thomas Reps, A Categorized Bibliography on Incremental Computation, POPL 1993
62. Doron Bustan and Orna Grumberg, Simulation-based minimization, ACM Trans. Comput. Log. 2003
63. Elmagarmid, Ahmed K. and Aggarwal, Charu C. and Wang, Haixun , Managing and Mining Graph Data, Advances in Database Systems, Springer 2010.
64. Huzur Saran and Vijay V. Vazirani, Finding k-cuts within Twice the Optimal, FOCS 1991
65. Olivier Goldschmidt and Dorit S. Hochbaum, Polynomial Algorithm for the k-Cut Problem, FOCS1988

66. Garey, M. R.; Johnson, D. S. (1979), Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman
67. Konstantin Andreev and Harald Racke, Balanced Graph Partitioning, Theory Comput. Syst. 2006
68. Per-Olof Fjallstrom, Algorithms for Graph Partitioning: A Survey, Linkoping Electronic Articles in Computer and Information Science, 1998
69. Chao-Wei Ou and Sanjay Ranka, Parallel Incremental Graph Partitioning, IEEE Trans. Parallel Distrib. Syst., 1997.
70. Robert Krauthgamer, Joseph Naor and Roy Schwartz, Partitioning graphs into balanced components, SODA 2009
71. David A. Bader and Kamesh Madduri, SNAP, Small-world Network Analysis and Partitioning: An open-source parallel graph framework for the exploration of large-scale networks, IPDPS 2008
72. Amine Abou-Rjeili and George Karypis, Multilevel algorithms for partitioning power-law graphs, IPDPS 2006
73. Guy Even, Joseph Naor, Satish Rao and Baruch Schieber, Fast Approximate Graph Partitioning Algorithms, SIAM J. Comput. 1999
74. Bruce Hendrickson and Tamara G. Kolda, Graph partitioning models for parallel computing, Parallel Computing 2000

75. George Karypis and Vipin Kumar, A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs, SIAM Journal on Scientific Computing, 1998
76. M. T. Ozs and P. Valduriez, Principles of Distributed Databases, Prentice-Hall
77. Donald Kossmann, The State of the art in distributed query processing, ACM Comput. Surv. 2000
78. Huahai He and Ambuj K. Singh, Graphs-at-a-time: query language and access methods for graph databases, SIGMOD 2000
79. Guozhu Dong, On Distributed Processibility of Datalog Queries by Decomposing Databases, SIGMOD 1989
80. Maurice A. W. Houtsma, Peter M. G. Apers and Stefano Ceri, Complex Transitive Closure Queries on a Fragmented Graph, ICDT 1990
81. Patrick Kling, M. Tamer Ozs and Khuzaima Daudjee, Generating Efficient Execution Plans for Vertically Partitioned XML Databases, PVLDB 2010
82. Gao Cong, Wenfei Fan and Anastasios Kementsietsidis, Distributed query evaluation with performance guarantees, SIGMOD 2007
83. Peter Buneman, Gao Cong, Wenfei Fan and Anastasios Kementsietsidis, Using Partial Evaluation in Distributed Query Evaluation, VLDB 2006
84. Andrew Lumsdaine, Douglas Gregor, Bruce Hendrickson and Jonathan W. Berry, Challenges in Parallel Graph Processing, Parallel Processing Letters, 2007.

# 谢谢！