中国科学技术大学

University of Science and Technology of China

数据仓库与数据挖掘

第一讲　概述

September 14, 2017

Outline

课程信息

上课时间	2 至 19 周（周四下午 2：00 开始）
上课地点	合肥 202
上机地点	合肥 404
上机时间	周 10，12，14，16，18 晚上 (18:30-21:30, 共 5 次)
教材	《数据挖掘：概念与技术》（第三版）
助教	待定
slides 下载方式	校研究生院研究生信息平台（第四周开始生效）

为什么要进行数据挖掘

信息技术的发展和进步，产生了新的生产领域：

- 计算机：保存和处理信息的工具
- 计算机网络：共享信息
- 物联网：信息收集能力的扩展
- 移动通信：信息收集和共享的增强

个人信息的收集

如果考虑采用现代信息技术，记录一个人的一生，那么……

- 行踪/工作/学习/娱乐/饮食/财务
- 记录数据的工具
- 记录数据的格式/媒体：文字/数字/视频/音频/表格

Question

数据量大约是多少？1 年 365 天，一生数十年

科大的每年会有多少数据产生？

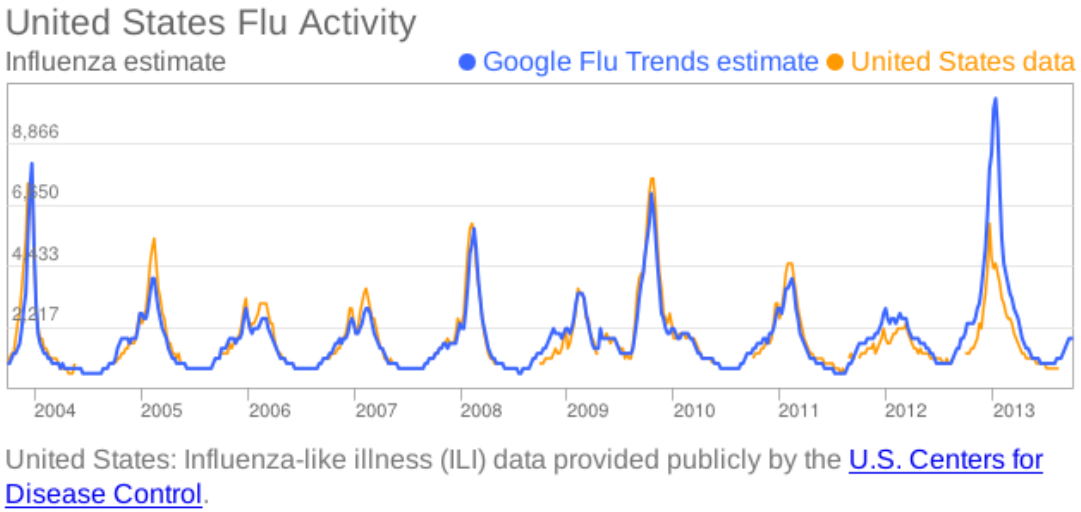
- 每个人每天：1MB
- 假设 1 万人：10GB
- 每年: 3.65TB
- 假设一个人每分钟可以阅读 1KB 的内容，则读完科大每年产生的数据约需要：3.65TB/60/24/365/1KB >6900 年

其它例子

美军在阿富汗大量使用无人侦察机，大型企业

奇思妙想：搜索引擎的搜索记录

- 每天搜索/查询次数超过 30 亿
- 每次搜索/查询被视为一条“事务”，包括搜索来源的 IP，时间，关键词，返回链接，被点击链接等
- 记录这些数据干什么？一个例子：Google Flu Trends
  - 在流感爆发的季节里，人们搜索流感相关的信息就会比较多; 能不能通过对这些流感相关的搜索记录进行统计？
  - GFT 做了，和传统方法，来自 CDC 的数据进行了对比，获得的结果发表在《Nature》上:



过高预计

google.org 流感趋势

语言： 中文（中国）

Google.org 首页

登革热流行趋势

流感趋势

首页

请选择国家/地区

它的工作原理是怎样的？

常见问题解答

它的工作原理是怎样的？

我们发现，某些搜索字词非常有助于了解流感疫情。Google 流感趋势会根据汇总的 Google 搜索数据，近乎实时地对全球当前的流感疫情进行估测。

全球每星期会有数以百万计的用户在网上搜索健康信息。正如您所预料的那样，在流感季节，与流感有关的搜索会明显增多；到了过敏季节，与过敏有关的搜索会显著上升；而到了夏季，与晒伤有关的搜索又会大幅增加。所有这些现象均可通过 Google 搜索解析进行研究。 但是，搜索查询趋势能否为实际现象建立一个准确可靠的模式而提供依据呢？

我们发现，搜索流感相关主题的人数与实际患有流感症状的人数之间存在着密切的关系。当然，并非每个搜索"流感"的人都真的患有流感，但当我们将与流感有关的搜索查询汇总到一起时，便可以找到一种模式。我们将自己统计的查询数量与传统流感监测系统的数据进行了对比，结果发现许多搜索查询在流感季节确实会明显增多。通过对这些搜索查询的出现次数进行统计，我们便可以估测出世界上不同国家和地区的流感传播情况。我们的研究结果已发表在美国《自然》杂志上。

历史估测数据 查看以下国家/地区的数据： 美国

美国流感疫情

流感估测病例数 ● Google 流感趋势估测数据 ● 美国的数据

美国：流感样疾病 (ILI) 数据由美国疾病控制中心提供。

这些图表显示了根据历史查询所得的不同国家和地区的流感估测结果，以及这些结果与官方的流感监测数据的对比。从图中可以看出，根据与流感相关的 Google 搜索查询所得到的估测结果，与以往的流感疫情指示线非常接近。当然，过去的表现并不能保证以后的结果一定准确。

我们为何要不厌其烦地根据汇总的搜索查询来估测疫情呢？传统的流感监测系统非常重要，但大多数卫生机构都只关注单个国家或地区，而且每隔一个星期才会更新一次估测数据。目前，我们已在全球的若干国家/地区推出 Google 流感趋势，此产品的数据每天都在更新，可为现有的流感监测系统提供有益的补充。

对于流行病学家而言，这个系统的开发是他们所喜闻乐见的，因为如果能及早监测到疫情爆发的征兆，就可以显著减少患病人数。当一株新的流感病毒在特定条件下形成后，一旦在全球范围内流行，就可能会夺去数百万人的生命（例如，1918 年就发生过这种灾难）。我们最新的流感估测结果可帮助公共卫生官员和专业卫生技术人员更好地应对季节性流感。

流感趋势试验版

我们发现，在我们推出流感趋势试验版的国家/地区中，根据汇总的流感相关查询所得到的季节性曲线与实际的流感疫情水平非常接近。通常，这些估测结果并没有与官方的流感监测数据进行对比。这些试验性的估测结果可从流感趋势试验版的相应国家/地区页面下载。

保护用户的隐私权

Google 深知用户对我们的信任，而且致力于保护用户的隐私。我们以匿名方式统计某些搜索查询在每星期的出现次数，因此 Google 流感趋势绝不会被用来确定个人用户的身份。我们依据的是长期以来发送至 Google 的数以百万计的搜索查询，而我们观察到的数据模式只适合 Google 搜索用户这个庞大的群体。要详细了解这些数据的使用情况以及 Google 如何保护用户的隐私权，请访问我们的隐私权中心。

详细了解 Google 流感趋势背后的研究过程：

阅读美国《自然》杂志发表的文章"运用搜索引擎查询数据检测流感疫情"(Detecting influenza epidemics using search engine query data)
HTML | PDF

下载 Google 流感趋势的全球估测结果

让人憎恨/喜爱的数据利用

思考 1:
手机软件泄漏个人隐私: 位置信息，通话记录，软件开发者用这些信息干什么？

思考 2:
QQ，Baidu，Google，360 软件为什么免费？这些企业能进一步改进吗？

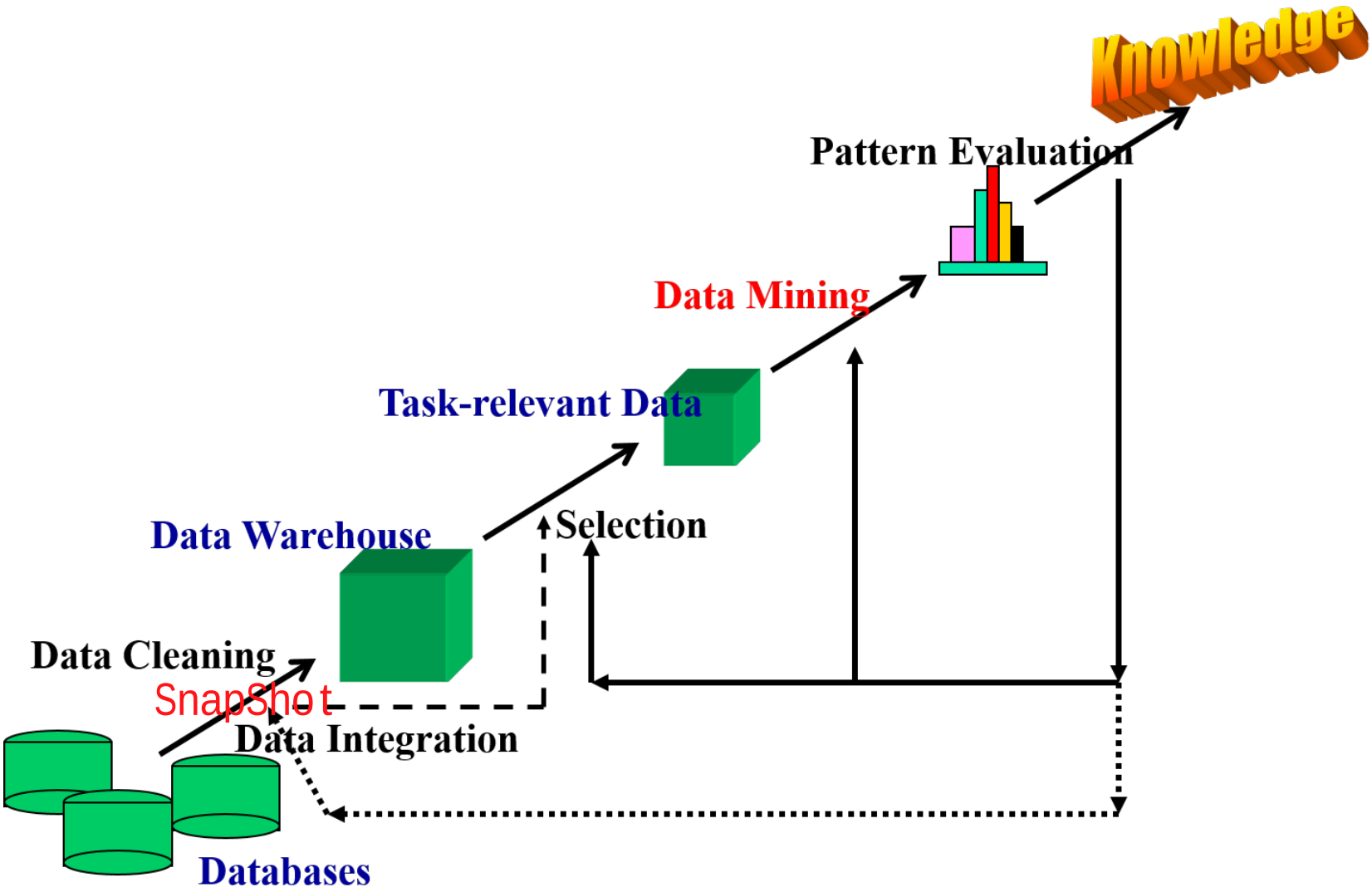
What is Data Mining?

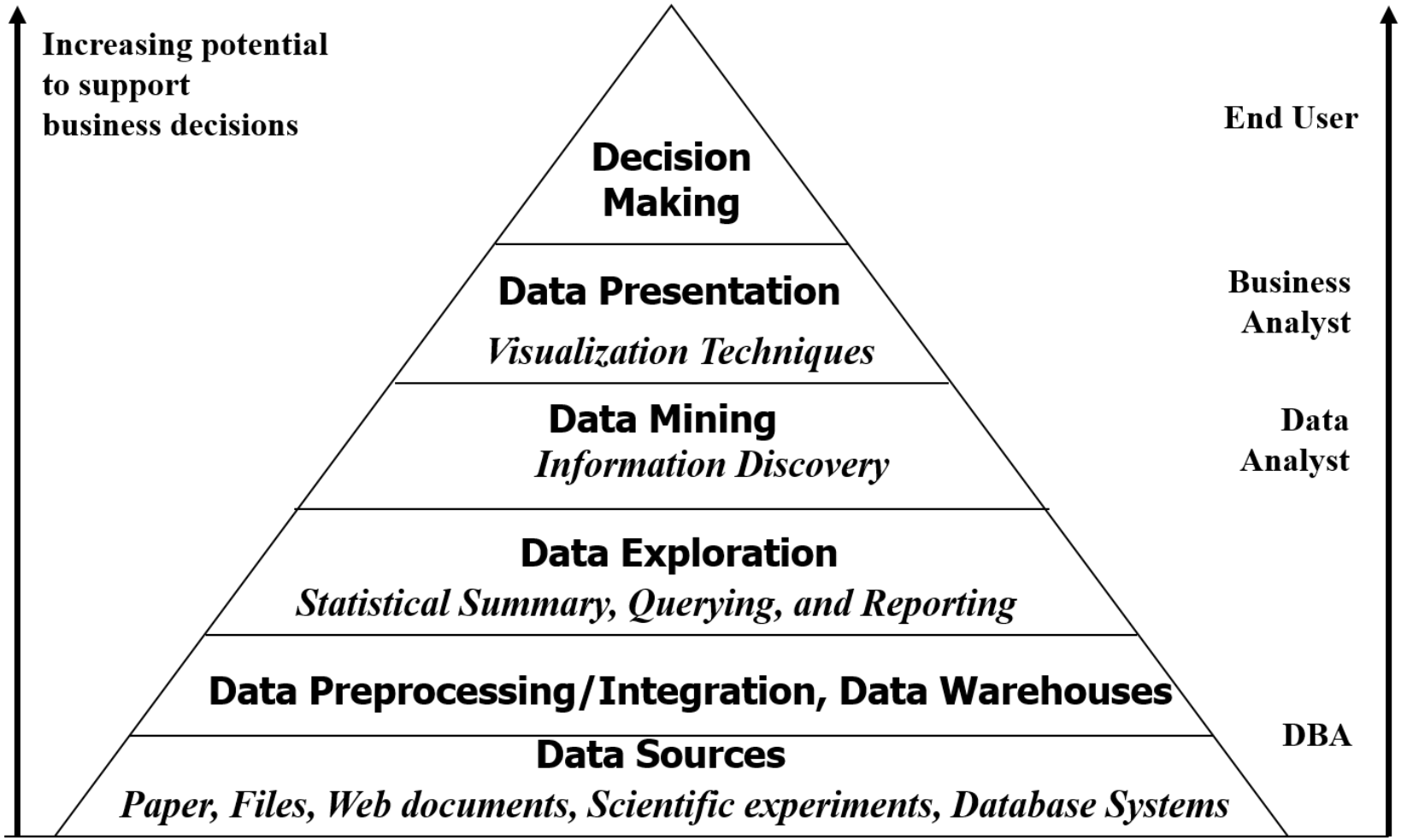
四个概念

1. 数据挖掘：Data Mining
2. 知识发现：Knowledge Discovery
3. 机器学习：Machine Leaning
4. KDD：Knowledge Discovery in Database

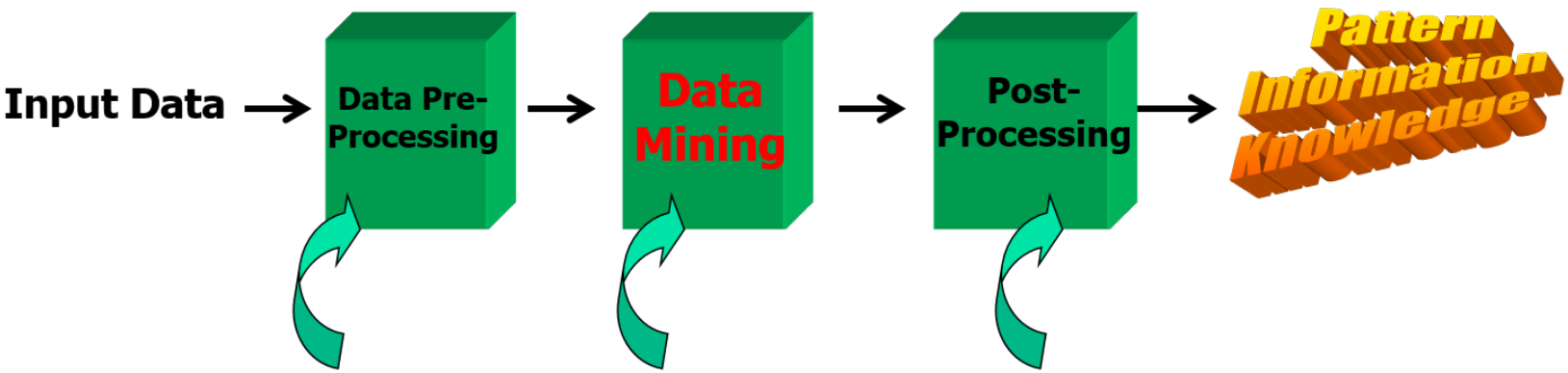
大数据分析

KDD 中的 DM


Knowledge
Pattern Evaluation
Data Mining
Task-relevant Data
Data Warehouse
Selection
Data Cleaning
SnapShot
Data Integration
Databases


Increasing potential
to support
business decisions
Decision
Making
Data Presentation
Visualization Techniques
Data Mining
Information Discovery
Data Exploration
Statistical Summary, Querying, and Reporting
Data Preprocessing/Integration, Data Warehouses
Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems
End User
Business
Analyst
Data
Analyst
DBA

ML 中的 DM



Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
... ...

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

What do you think?

我对 ML 和 DM 的理解

价格 数量 / 克数	500	1000	2000	3000	5000	10000	20000	30000
105	330	350	390	430	530	780	1430	2010
128	335	355	400	445	560	885	1540	2180
157	340	370	425	480	610	1000	1765	2510
200	355	390	★	540	705	1180	2120	3030
250	375	420	510	600	805	1375	2510	3600
300	390	450	560	670	910	1590	2930	4200

DM 是指上表中有未填写的格子（红色星号），找合适“内容”填进去

价格 数量 / 克数	500	1000	2000	3000	5000	10000	20000	30000
105	330	350	390	430	530	780	1430	2010
128	335	355	400	445	560	885	1540	2180
157	340	370	425	480	610	1000	1765	2510
200	355	390	465	540	705	1180	2120	3030
250	375	420	510	600	805	1375	2510	3600
300	390	450	560	670	910	1590	2930	4200

机器学习/ML 是指上表不是一张完整的表，把表补全。
表有多少行？哪些行存在未填写的格子？

从 4 个不同的角度去理解 DM

要挖掘的数据

- 数据库：关系，面向对象，异构，遗产
- 数据仓库
- 事物数据库
- 流式数据，时空数据库，时间序列
- 文本和 web
- 多媒体数据
- 图、社交网络和信息网络

挖掘的方法

- 流式计算/云计算/云存储
- 数据仓库
- 机器学习
- 统计学
- 模式识别
- 可视化
- 高性能计算

挖掘出的知识

- 特征
- 区别
- 关联
- 类别
- 聚类
- 趋势/偏离
- 孤立点分析

应用的领域

- 零售业/营销
- 通信行业/互联网/信息网络
- 银行业
- 欺诈分析
- 生物信息学
- 股票市场分析
- 信息网络

A Brief History of DM Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases, "Knowledge Discovery in Databases" (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases, "Advances in Knowledge Discovery and Data Mining" (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 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
- More conferences on data mining. PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

数据挖掘的主要期刊和会议

- 会议
  - 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)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR
- 期刊
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD
- Where to find the references?
  - Google/Google Scholar
  - siteseer
  - DBLP

数据挖掘存在的主要问题

用户交互

- 交互式挖掘
- 背景知识的嵌入
- 挖掘结果的可视化

效率和可扩展性

- 挖掘算法的效率和可扩展性
- 并行/分布式/流式计算/增量式挖掘算法

数据挖掘面临的社会问题

- 社会影响
- 隐私保护和数据挖掘
- 不可见数据挖掘

数据类型的多样性

- 复杂数据类型的处理
- 动态/网络化/全球数据仓库/大数据

挖掘方法和技术

- 各种新知识的挖掘
- 多维空间/高维空间的处理方法
- 融合多学科/领域的方法技巧
- 利用网络化环境来提高挖掘的能力
- 噪声数据/不完全数据/不确定性的处理
- 有约束的挖掘

大数据的定义

- 数据分析的前沿技术, 定义多种多样，有些许差别，比较公认的是认为其具备特点 4 “V”
- 4 个 “V”，Volume（数据体量巨大，从 TB 级别，跃升到 PB 级别）、Velocity（处理速度快，1 秒定律，可从各种类型的数据中快速获得高价值的信息）、Variety（数据类型繁多）、value（只要合理利用数据并对其进行正确、准确的分析，将会带来很高的价值回报）
- 相关技术：并行数据库，MapReduce, 分布式软件框架 Hardoop，云计算

应用：
GFT 谷歌流感预测，Target 超市针对孕妇的产品推销等等