Introduction 简介

Instructor: Zach Zhizhong ZHOU, Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

数据挖掘的定义

- 从大数据中提取有用信息 Extracting useful information from large data sets. (Hand et al., 2001)

- 数据挖掘是一个使用自动或半自动方法，对大量数据进行探索和分析，试图发现有意义的模式和规则的过程。 Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. (Berry and Linoff, 1997, p. 5)

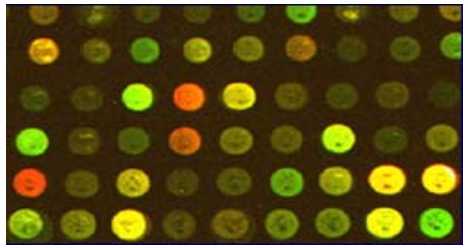
- 数据挖掘是通过仔细检视存储于数据存储库中大量数据，发现有意义的关联关系、模式和趋势的过程数据挖掘使用模式识别技术、统计和数理技术。 [Data Mining is] the process of discovering meaningful correlations, patterns and trends by sifting through large amounts of data stored in repositories. Data mining employs pattern recognition technologies, as well as statistical and mathematical techniques. (Gartner Group 2010)

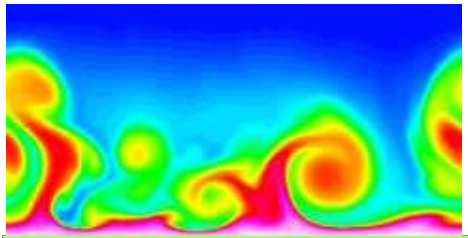
为什么进行数据挖掘（商业观点）

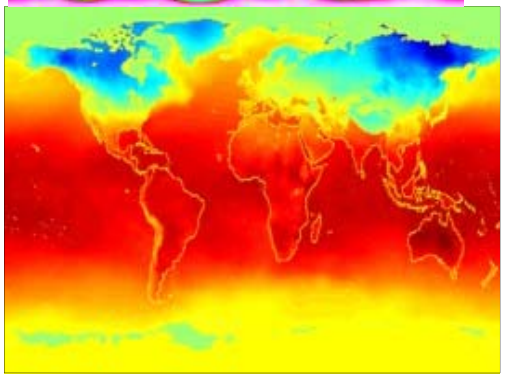
- 大量的数据被收集和存储：
  - 电子商务网站
  - 零售店通过扫码收集的数据
  - 金融数据：信用卡消费记录、股票交易数据等。
- 机器计算的资源更加便宜，功能更强大，更加容易获取。
- 强大的市场竞争压力。

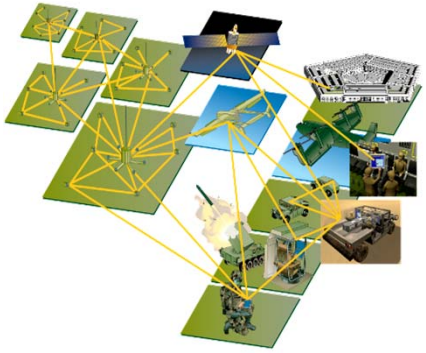
为什么进行数据挖掘（科学观点）

- 大量数据以极高的速度被收集
  - 人造卫星上千个传感器传回的实时数据
  - 电子天文望远镜对星空进行扫描
  - DNA 测序
  - 科学模拟实验产生的大量数据
- 传统数据处理方法不适用于新的数据
- 数据挖掘可以帮助科学家
  - 对数据进行分类和切割
  - 形成科学假说









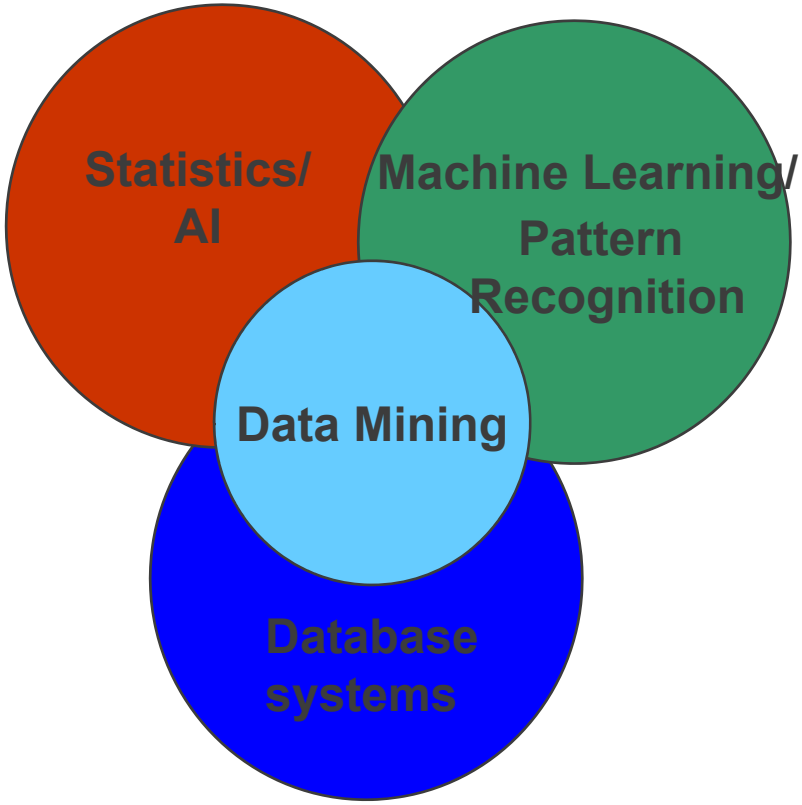
数据挖掘的应用

- 寻找目标客户
- 识别金融欺诈（老鼠仓、盗用账户、盗用信用卡、诈骗等）
- 商业贷款的风险控制
- 预测美国总统选举结果
- 识别身份证造假（人脸识别等）
- 网上婚恋网站

数据挖掘的起源

□数据挖掘利用了来自如下领域的思想：（1）来自统计学的抽样、估计和假设检验，（2）人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论，（3）其他领域的思想包括最优化、进化计算、信息论、信号处理、可视化和信息检索。

- 传统的数据分析技术可能不适用，原因可能是：（1）海量数据，（2）高维数据，（3）异种数据和复杂数据，（4）分布式数据。



研究：儿时朋友越多 成人后收入越多

□一项研究显示，如果只重视学习根本不让儿童玩耍的话，不但让儿童感到枯燥，而且在他们长大以后可能会更受穷。这项研究总结说，那些在学校中朋友最多的十几岁的孩子在长大后比同班同学收入更高的可能性更大。英国萨塞克斯大学（The University of Sussex） 的经济学家对这项历时15年的跟踪调查作了分析。调查的对象是1990年代中期美国多达上千名十几岁的高中生。到30岁的时候，那些当初在学校中朋友很多，人缘较好的人平均工资要高出其他同学的12.5%。研究人员不但调查他们父母的职业，还看每个学生在学校到底有多少好朋友，而且他们在学校中所扮演的角色，比如谁在同学中最具有影响力等。然后，研究人员把这个结果跟当他们到接近30岁时的数据再做比较。结果发现，那些在高中时平均朋友人数比一般人多出10%的人在到接近30岁时的工资也比同龄人高出近11%。而当初在学校最有影响力的人工资比同班同学则高出12.5%。非常有趣的是，一些其他的因素，比如在班级里的顶尖学生或是智商较高对日后的赚钱能力似乎没有太大的影响。当然，如果根本不学习也将犯了一个严重的错误。而拥有一个良好的教育，比如上大学比朋友多更重要，因为上大学可以使薪水水平提高25%。此外，那些已婚人士比单身人士的工资也高出19%。研究人员说，这显示学校以及教育对人们的深远影响远非学习知识本身。同时，也表明为每个孩子提供免费和公共教育的重要性。而且，它也显示应该在儿童时期给与更多的注意力，以及通过社会活动和各种俱乐部等来发展儿童的社交技能。

□ 问题：该研究有什么缺陷?

数据挖掘与统计学的区别

- 统计学核心问题往往是观察了样本之后如何推断总体，这也常常是数据挖掘所关注的。但数据挖掘常常可以得到数据总体，这种情况下统计推断变得毫无意义。

- 统计学往往基于现有理论构建模型，然后分析变量之间的关系，试图得到因果关系。而数据挖掘并不注重因果关系，模型选择也是次要的，主要考虑的因素是算法和算法准则。如神经网络挖掘技术就没有简单明确的模型。

- 统计分析着眼于验证理论，数据挖掘着眼于发现非预期的但很有价值的信息。

- 在一些情况下，数据挖掘比统计分析更注重实时性。