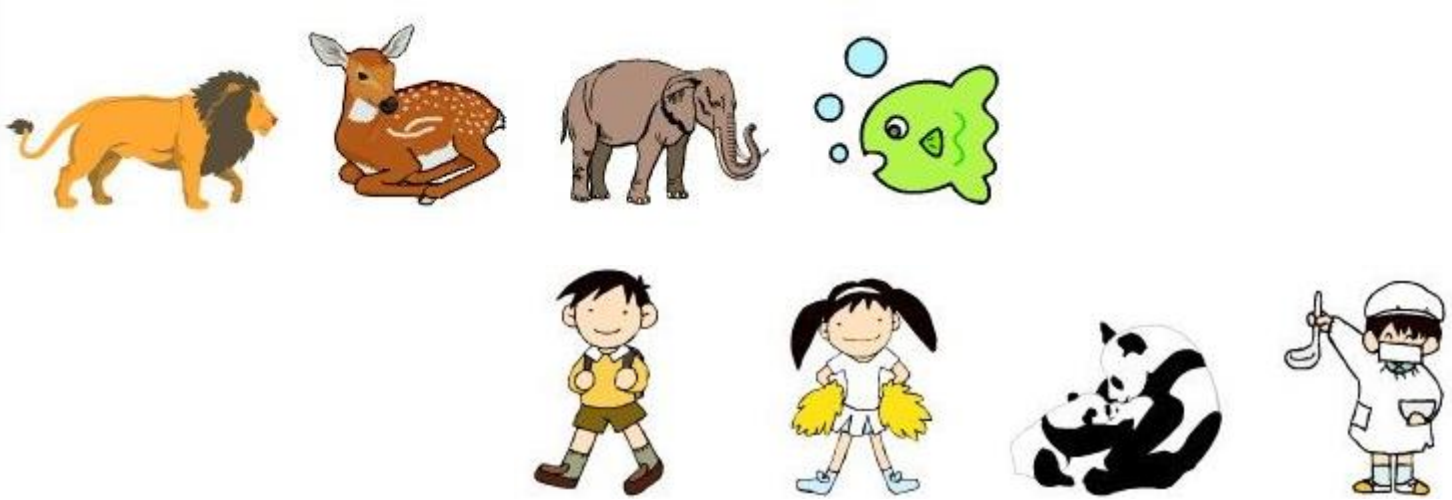
数据挖掘能做什么

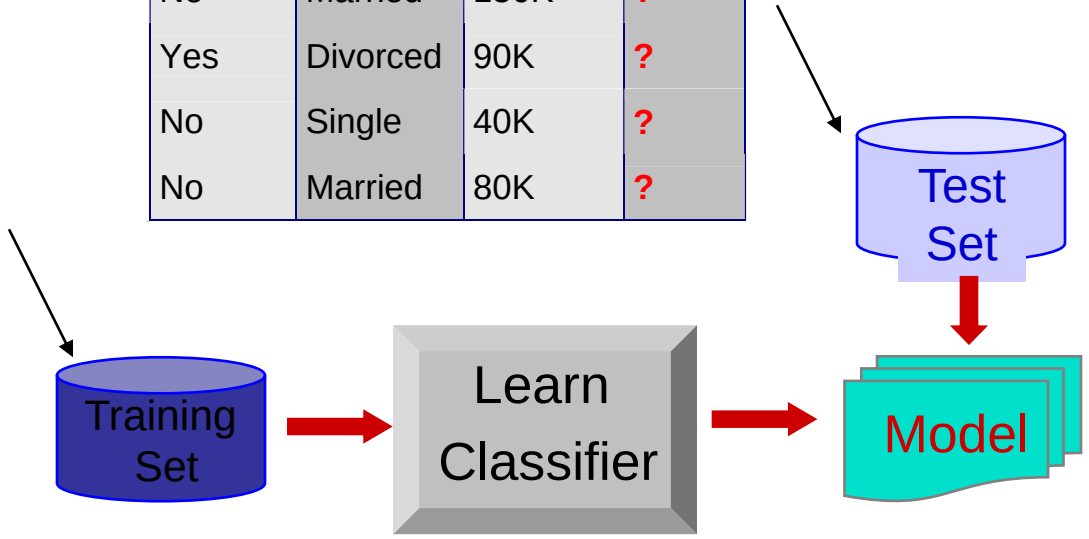
- 分类 （Classification）
  - 应用： 评估信用卡申请者的风险等级——低、中、高
  - 方法：使用已知分类的实例建立分类模型，对未知分类的实例进行分类



分类例子—鉴别信用卡欺诈

	categorical	categorical	continuous	class
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	?



数据挖掘能做什么

- 估值（Estimation）
  - 应用：根据购买模式，估计一个家庭的孩子个数、收入或财产
  - 估值类似于分类，不同之处在于
    - 分类的输出是离散量，估值输出为连续值
    - 分类的类别数确定，估值的量是不确定的

数据挖掘能做什么

- 预测（Prediction）
  - 应用：预测明天上证指数的收盘价
  - 方法：通过分类或估值得出预测模型，用该模型对未知变量的预测

数据挖掘能做什么

- 相关分析（association analysis）
  - 应用：**购物篮分析（Market Basket Analysis）**——超市中，客户在购买A的同时，是否经常会购买B或隔一段时间后，会购买B。目的是找到零售产品之间有趣的关系
  - 方法：生成关联规则，表达客户购买行为的关联关系

交易号	产品
T01	啤酒
T01	尿布
T02	啤酒
T02	尿布
T03	尿布

数据挖掘能做什么

- 聚类（Clustering）
  - 应用：在信用卡公司，发现输入属性的一个集合，来区分接受寿险促销和未接受促销的持卡人。
  - 方法：对实例分组，把相似的实例放在一个聚类中，发现最能区分各聚类的典型属性，使用这些属性开发预测未来结果的模型

应用领域

Industries / Fields where you applied Analytics / Data Mining in 2012?	
CRM/Consumer analytics (56)	28.6%
Health care/ HR (32)	16.3%
Retail (29)	14.8%
Banking (28)	14.3%
Education (28)	14.3%
Advertising (26)	13.3%
Fraud Detection (25)	12.8%
Social Media / Social Networks (24)	12.2%
Science (23)	11.7%
Finance (20)	10.2%
Direct Marketing/ Fundraising (19)	9.7%
Search / Web content mining (16)	8.2%
Biotech/Genomics (15)	7.7%
Insurance (15)	7.7%
Credit Scoring (14)	7.1%
Manufacturing (14)	7.1%
Medical/ Pharma (13)	6.6%
Telecom / Cable (13)	6.6%
Web usage mining (13)	6.6%
Software (11)	5.6%
Ecommerce (10)	5.1%
Government/Military (10)	5.1%
Entertainment/ Music/ TV/Movies (9)	4.6%
Investment / Stocks (8)	4.1%
Security / Anti-terrorism (7)	3.6%
Travel / Hospitality (6)	3.1%
Social Policy/Survey analysis (2)	1.0%
Junk email / Anti-spam (1)	0.5%
Other (20)	10.2%

图1.6 www.kdnuggets.com网站公布的2012年数据挖掘的应用领域

成功案例

- **除了最著名的沃尔玛的尿布和啤酒之外，还有**

（1）Empire Blue Cross公司利用DWT，甄别出虚假开立医疗凭据的医生，节省滥赔支出。

（2）金融犯罪强制网络AI系统（FAIS）使用DWT，识别大型现金交易中可能存在的洗钱行为。

（3）加拿大西门菲沙大学（Simon Fraser）的KDD研究组根据其拥有的十几年的客户数据，进行数据挖掘分析，提出了新的电话收费和管理办法，制定出公司和客户都受益的优惠政策。

（4）美国梅隆（Mellon）银行使用Intelligent Agent数据挖掘工具提高销售和定价金融产品的准确率。

（5）美国西部通信（US West Communications）根据家庭大小、家庭成员平均年龄和所在地特征，使用数据挖掘和数据仓库来确定客户的倾向和需要，从而帮助签约新客户和增加与新客户的交易额。

（6）使用贝叶斯分类数据挖掘技术，萨莎（Sacha）等人成功地通过心肌SPECT图像对心肌灌注进行分类，诊断患者是否患有冠心病。

（7）20世纪Fox公司利用数据挖掘技术分析票房收入来确定在各个市场环境中更容易被接受的演员和故事情节。

（8）科学界普遍认为存在两种γ射线爆。慕克吉（Mukherjee）等人使用统计聚类分析法发现了第三类γ射线爆。

（9）NBA球队使用IBM公司开发的数据挖掘应用软件Advanced Scout系统来优化他们的战术组合。

（10）全球十大视频网站 之一Netflix公司应用大数据的挖掘技术，成功营销热播剧——《纸牌屋》。