天文数据分析（三）

国家天文台 赵永恒

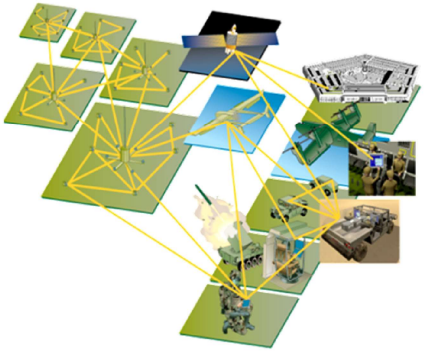
大数据分析方法

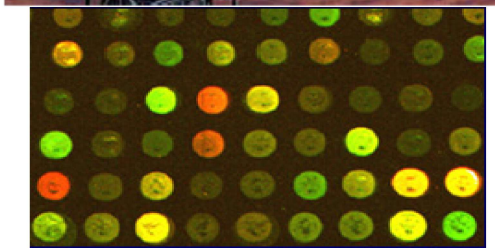
为什么挖掘数据？商业观点

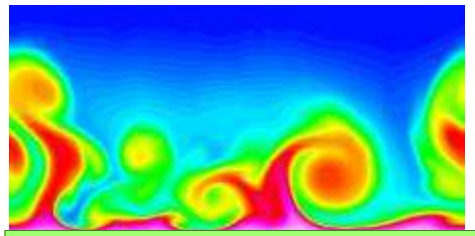
- 大量数据的收集和存储
  - 网络数据、电子商务数据
  - 商店的销售额
  - 银行/信用卡交易
- 计算机价格越来越便宜，
  但功能越来越强大
- 竞争压力日益增加
  - 提供更好的定制服务

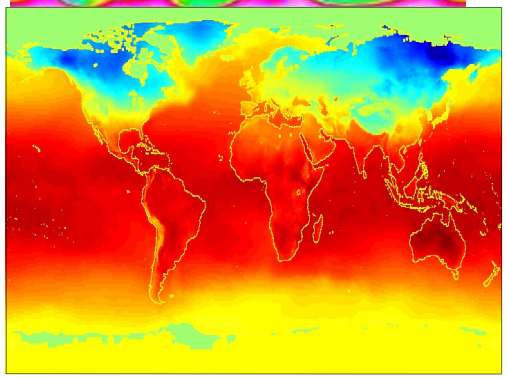
为什么挖掘数据？科学观点

- 高速的数据收集和存储（GB/小时）
  - 卫星遥感数据
  - 望远镜巡天数据
  - 微阵列产生的基因表达数据
  - 科学数值模拟数据（TB）
- 传统技术对数据处理已不可行
- 数据挖掘正好帮助科学家
  - 对数据分类和分割
  - 推理和假设



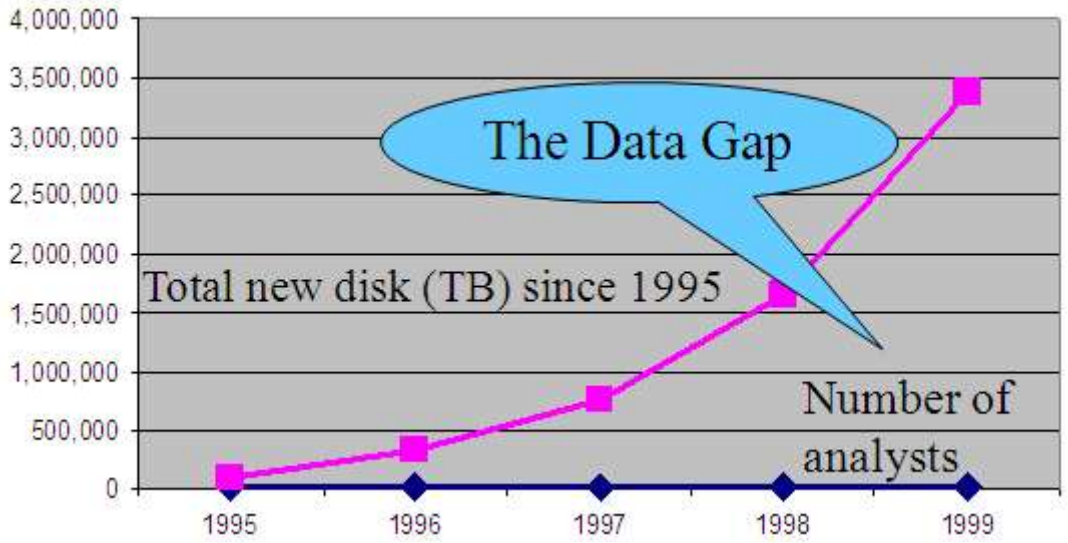






数据挖掘：动机

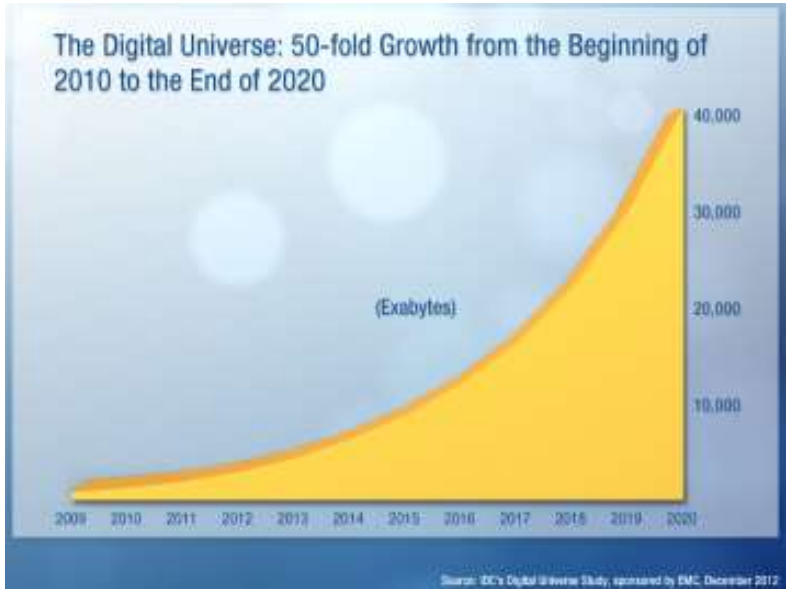
- 大量信息隐藏在数据中不易被发现
- 人们花时间能发现一些有用信息
- 大部分数据无人问津



数据究竟有多少?

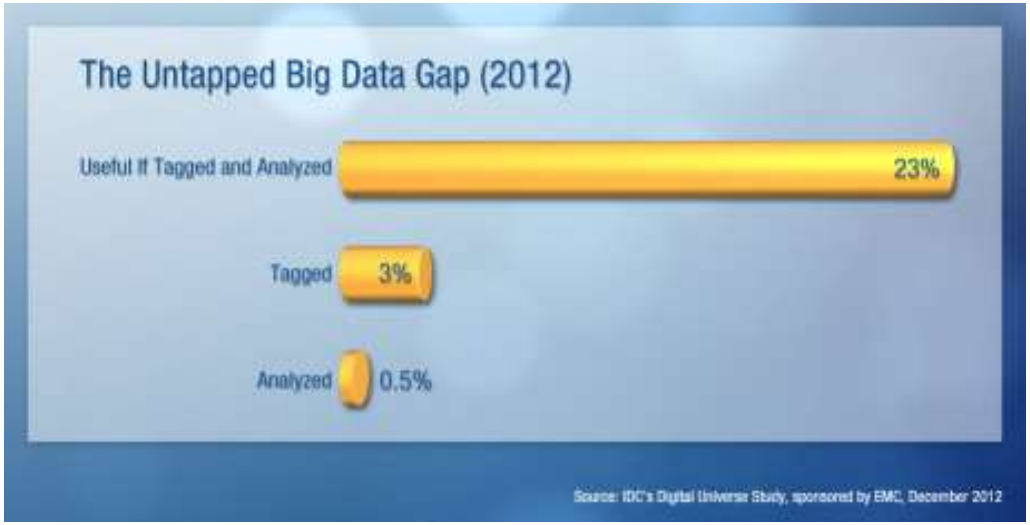
- Google: ~20-30 PB a day
- Wayback Machine has ~4 PB + 100-200 TB/month
- Facebook: ~3 PB of user data + 25 TB/day
- eBay: ~7 PB of user data + 50 TB/day
- CERN's Large Hydron Collider generates 15 PB a year

大数据时代



IDC 预测: 从2005年2020年, 数字宇宙每两年增长一倍，从30 exabytes涨到40,000 exabytes或者到2020年人均数据量 5,200 GB.

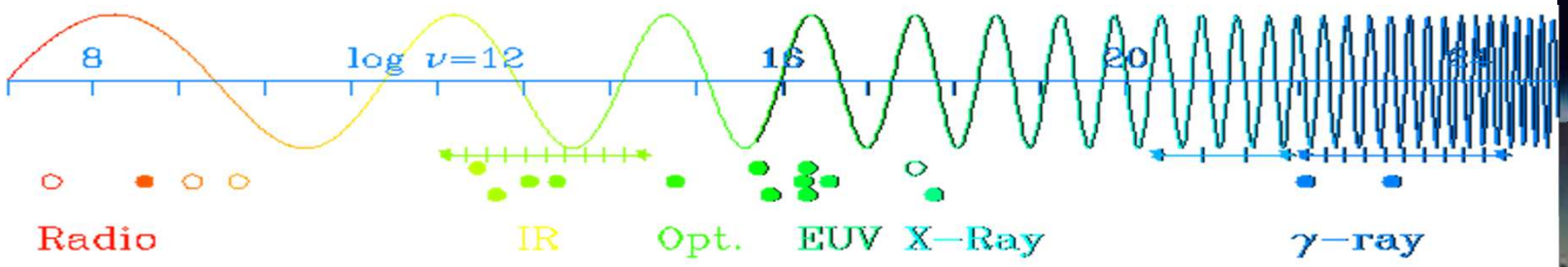
未标注数据缺口:
大部分有用的数据没有标注或分析--部分来自技术的缺乏。

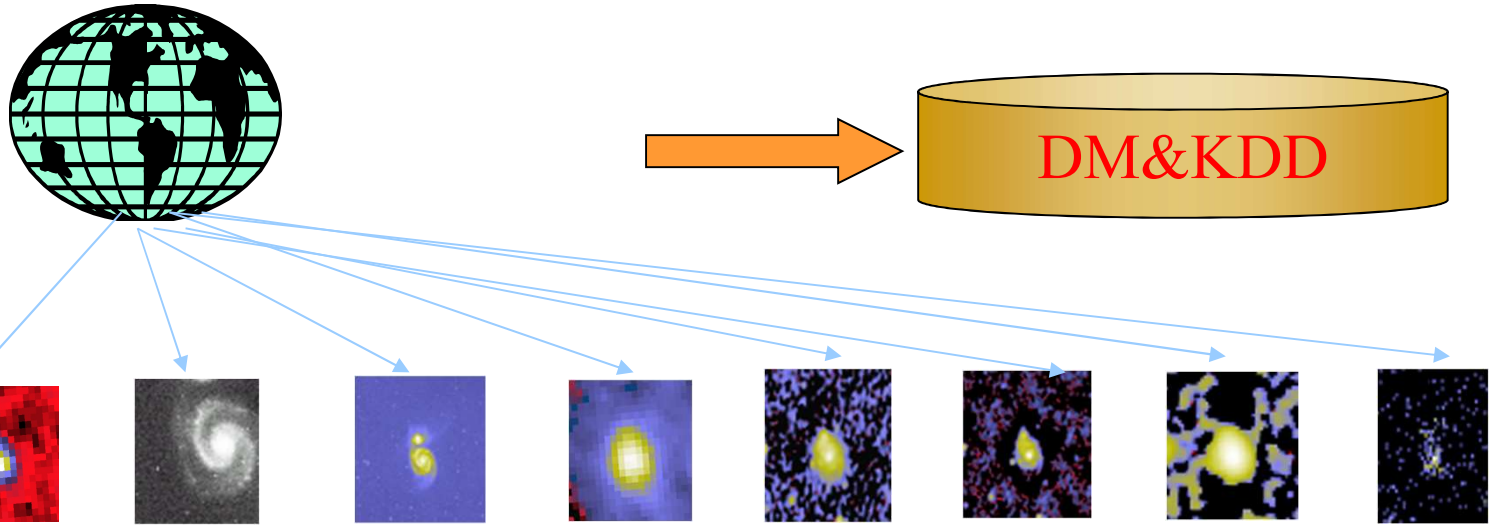


大数据-巨信息量-全波段天文时代

Astronomy facing "data avalanche"

Necessity Is the Mother of Invention

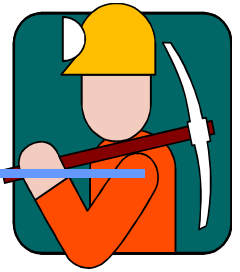




天文大数据

巡天项目	巡天项目全称	运行状态	数据量
DPOSS	The Palomar Digital Sky Survey	Finished	3 TB
2MASS	The Two Micron All-Sky Survey	Finished	10 TB
GBT	Green Bank Telescope	Finished	20 PB
GALEX	The Galaxy Evolution Explorer	Operating	30 TB
SDSS	The Sloan Digital Sky Survey	Operating	40 TB
SkyMapper	Southern Sky Survey	Operating	500 TB
PanSTARRS	The Panoramic Survey Telescope and Rapid Response System	Operating	~ 40 PB expected
LSST	The Large Synoptic Survey Telescope	In Plan	~ 200 PB expected
SKA	The Square Kilometer Array	In Plan	~ 4.6 EB expected

数据挖掘的定义



数据挖掘（数据库中的知识发现）：

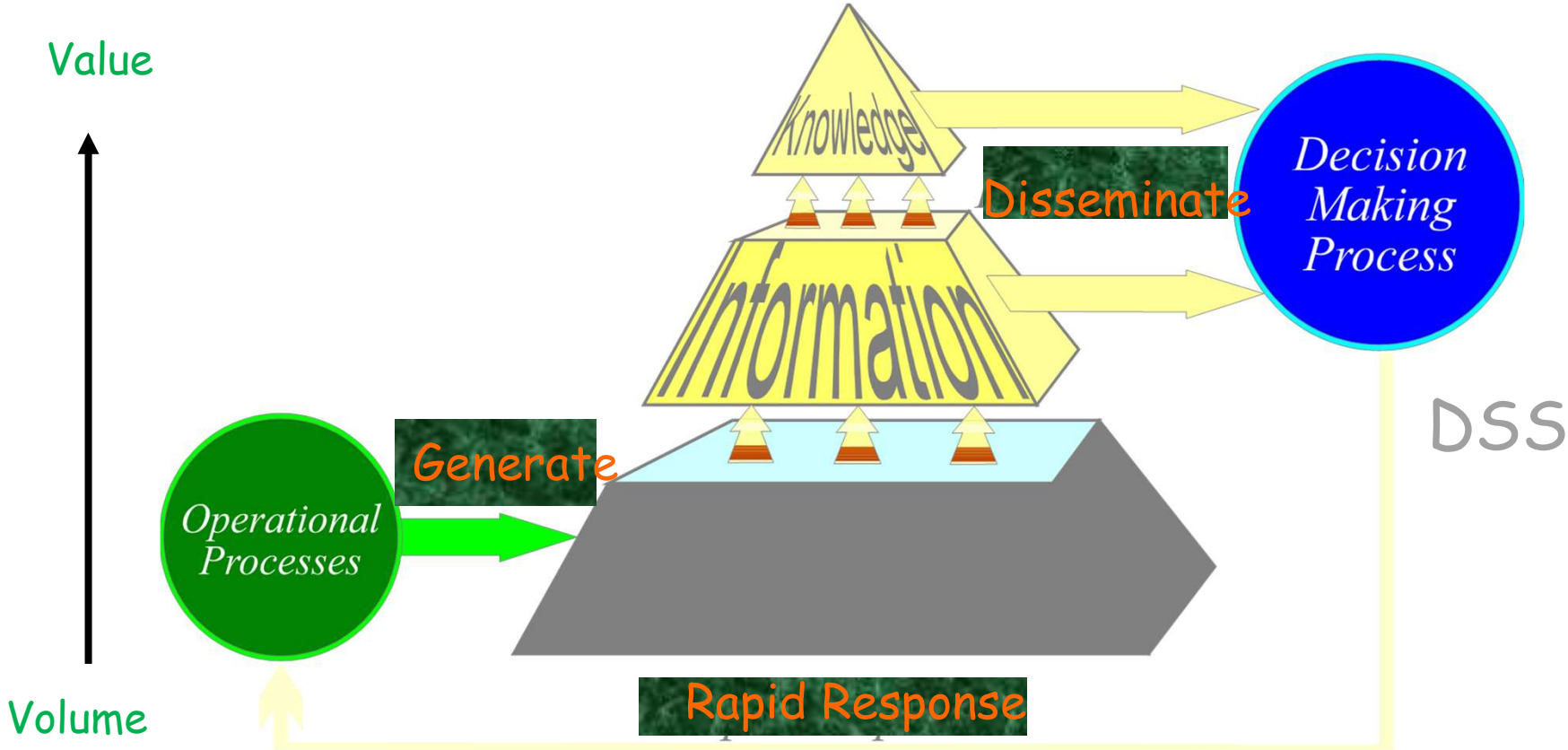
从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的模式或知识的过程。

数据挖掘别名：

数据库中的知识发现、数据考古、知识提炼、数据捕捞、信息收获、数据/模式分析

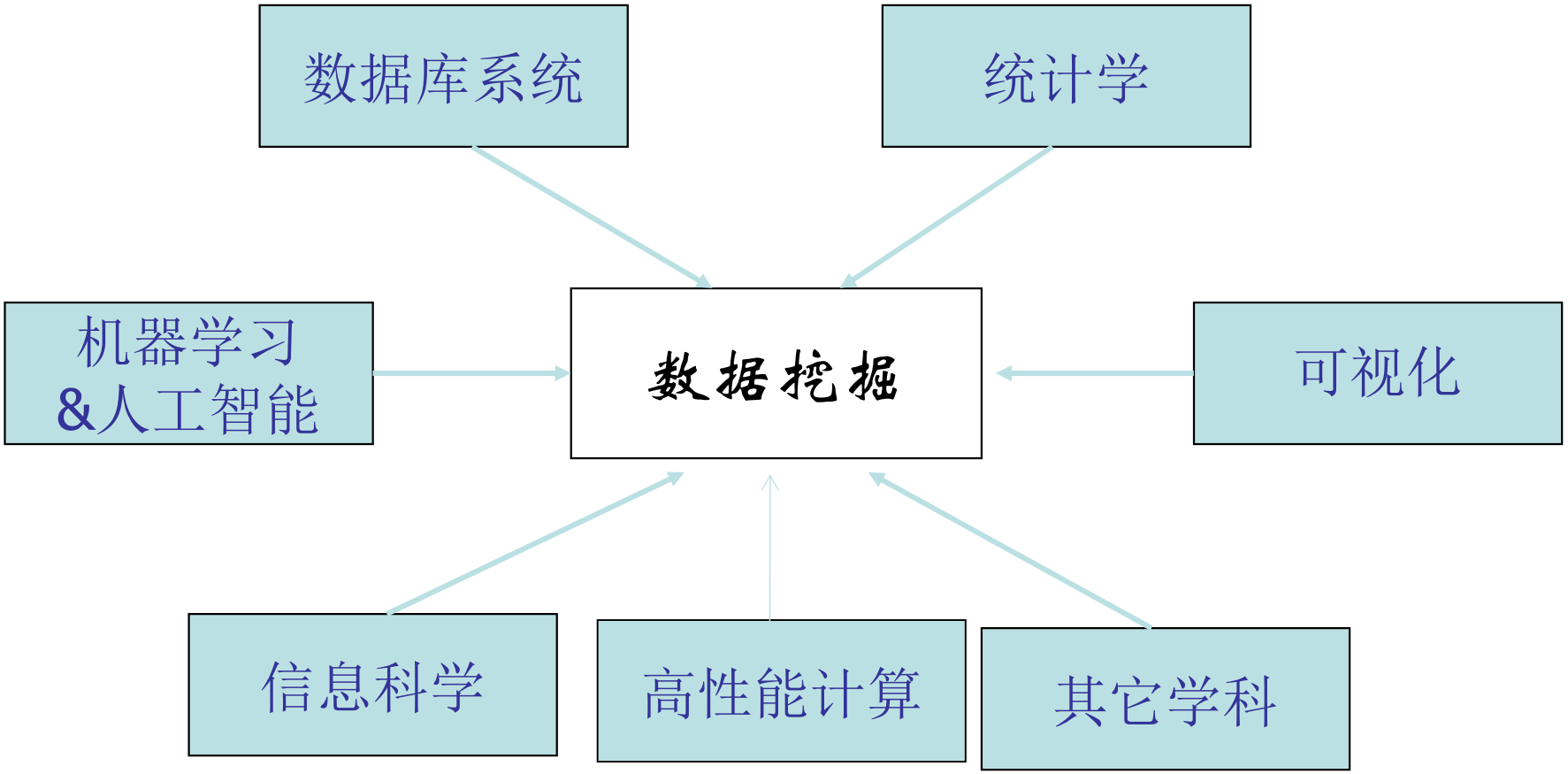
数据挖掘常用于工程界，数据库中的知识发现常用于科学界。

数据挖掘的优点



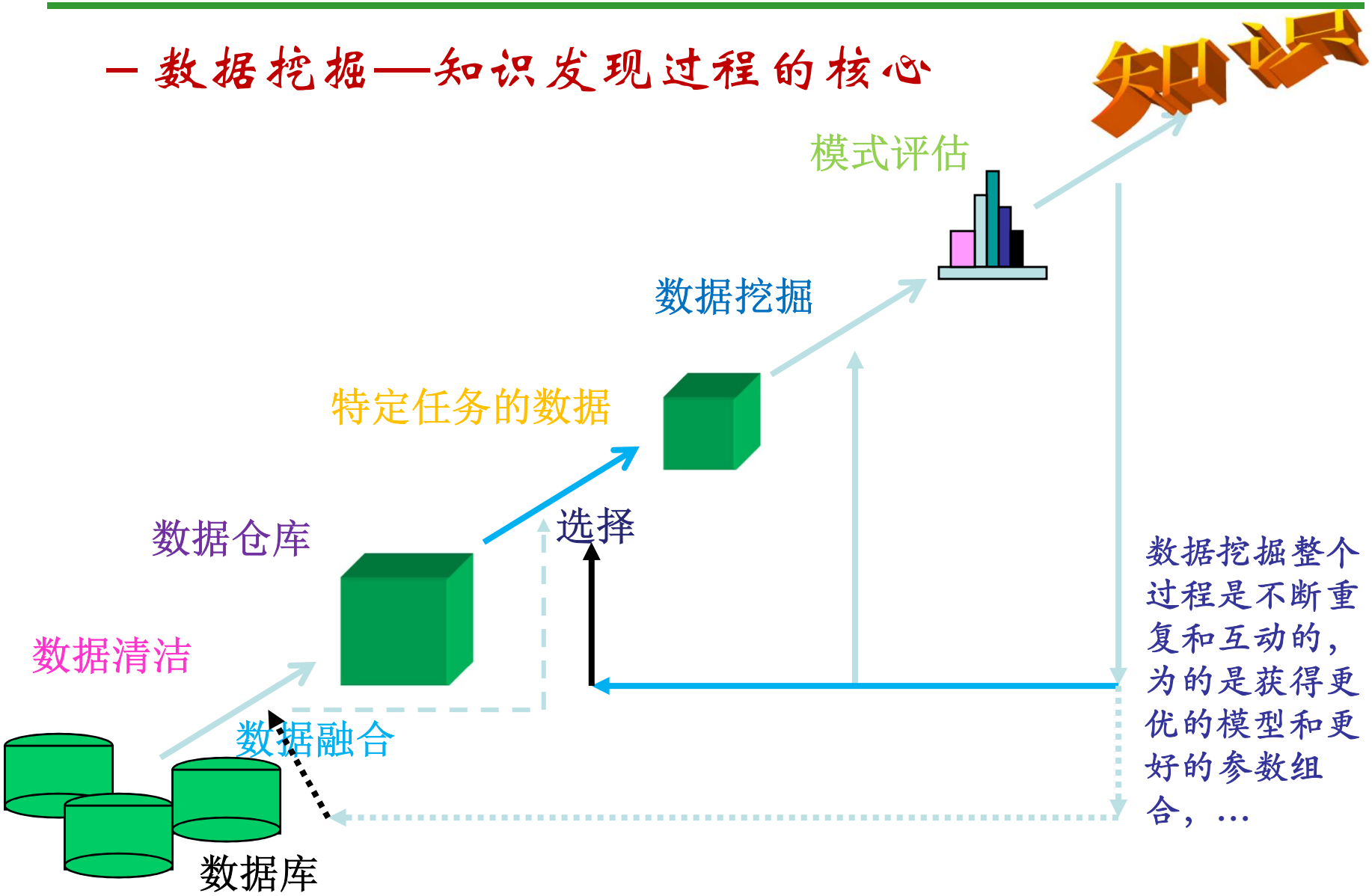
EDP：电子数据处理
MIS：管理信息系统
DSS：决策支持系统

数据挖掘是多学科发展的产物

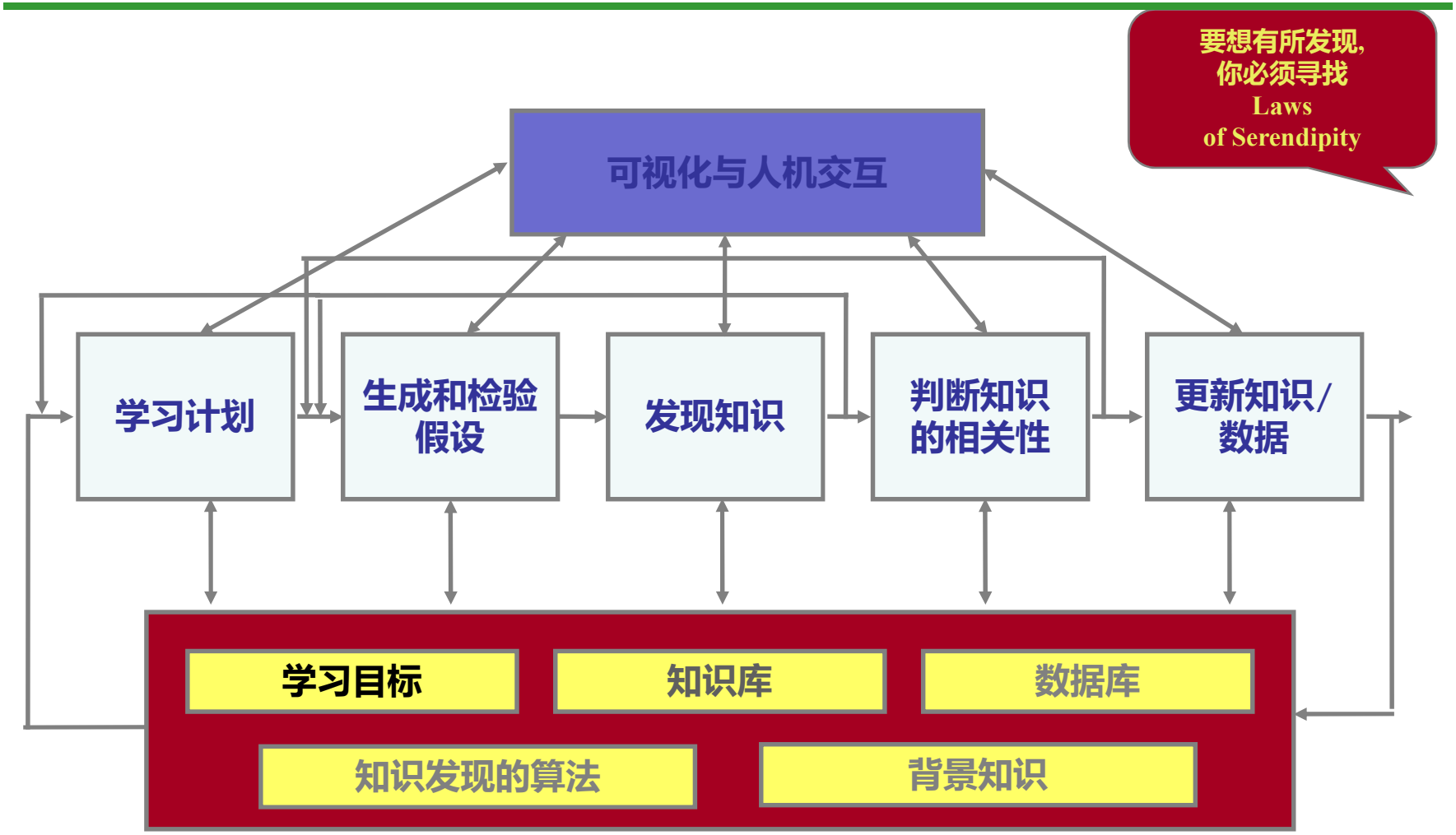


数据挖掘的过程

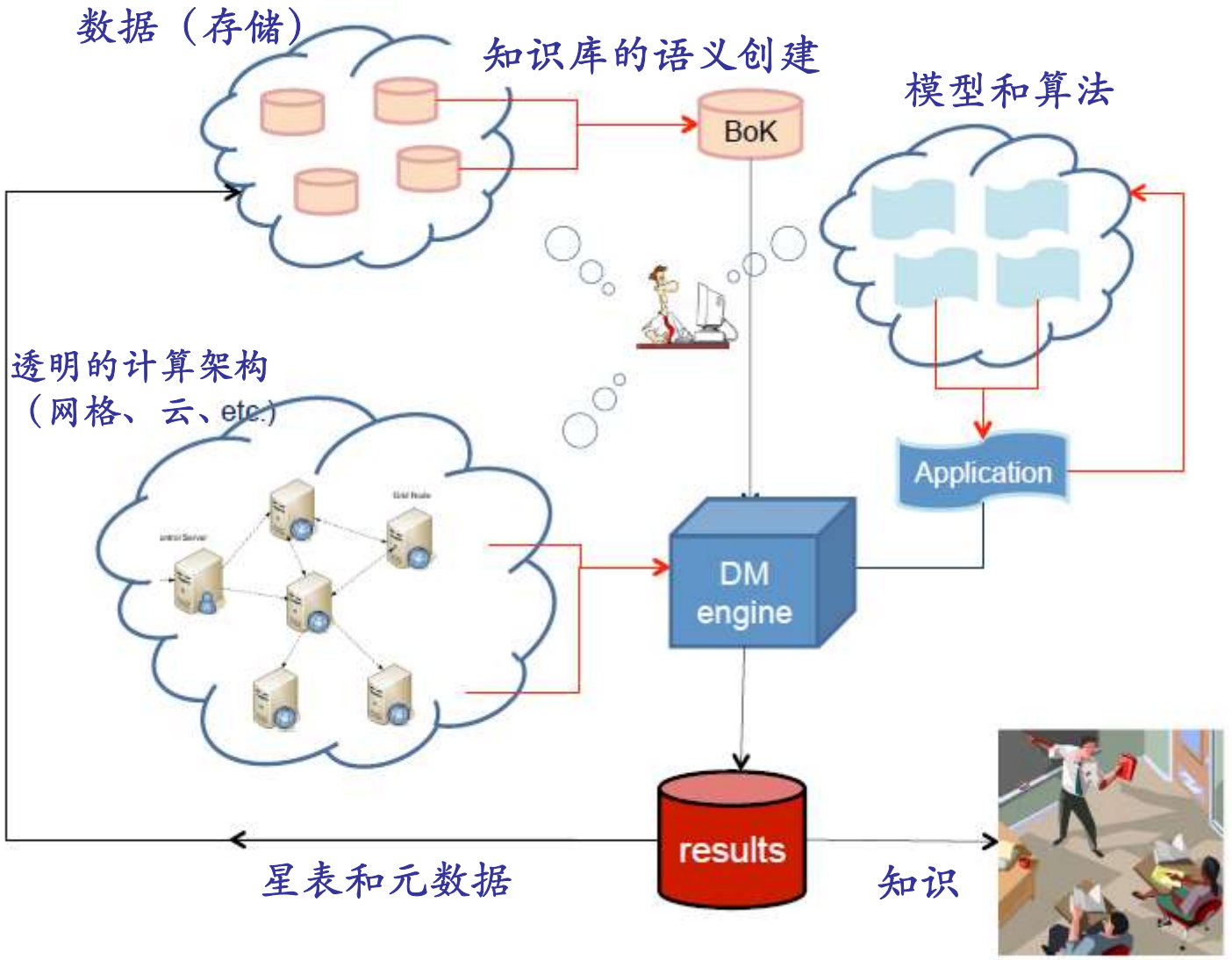
– 数据挖掘—知识发现过程的核心



有效的数据挖掘


要想有所发现,
你必须寻找
Laws
of Serendipity
可视化与人机交互
学习计划
生成和检验
假设
发现知识
判断知识
的相关性
更新知识/
数据
学习目标
知识库
数据库
知识发现的算法
背景知识

有效的数据挖掘分解图


数据（存储）
知识库的语义创建
BoK
模型和算法
透明的计算架构
（网格、云、etc.）
Application
DM
engine
results
星表和元数据
知识

数据挖掘的历史

- 数据库中的知识发现工作组发端于1989年
  - 现在由ACM的数据挖掘和知识发现（ SIGKDD ）专委会主办
  - IEEE 会议系列从2001年开始承办
- 关键的奠基者 /技术贡献者:
  - Usama Fayyad, JPL (then Microsoft, then his own company, Digimine, now Yahoo! Research labs)
  - Gregory Piatetsky-Shapiro (then GTE, now his own data mining consulting company, Knowledge Stream Partners)
  - Rakesh Agrawal (IBM Research)

“数据挖掘”一词至少从1983年开始流行，不过在统计领域对之持轻蔑的态度。

数据挖掘社区的发展历史

- 1989 IJCAI 数据库中的知识发现工作组 (Piatetsky-Shapiro)
  - 数据库中的知识发现(G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994数据库中的知识发现工作组
  - 知识发现和数据挖掘进展 (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998数据库中的知识发现和数据挖掘会议 (KDD'95-98)
  - 数据挖掘和知识发现期刊 (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 会议s, and SIGKDD Explorations
- 更多关于数据挖掘的会议
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.

数据挖掘：对哪些数据？

- 各种数据集合
  - 不动产数据
  - 多媒体数据 (通常高维)
  - 空间数据 (含有空间信息，如地图、卫星图像数据、天文数据等)
  - 时间序列数据 (与时间相关; 常常动态变化)
- 万维网数据
  - 基本上是大量的异构分布数据
  - 需要新的或另外的工具和技术
    - 信息检索、过滤、抽取
    - 一些机构帮助浏览和过滤
    - 网页内容、使用及相关的结构的挖掘工具
  - 社交网
    - 用户产生的元数据、社交网络、共享资源等.

数据挖掘的分类

- **广义上分：**
  - -- 描述性数据挖掘
  - -- 预测性数据挖掘
- **不同的角度，不同的分类**
  - -- 挖掘的数据类型
  - -- 要发现的知识类型
  - -- 使用技术的类型
  - -- 应用的领域

数据挖掘的任务

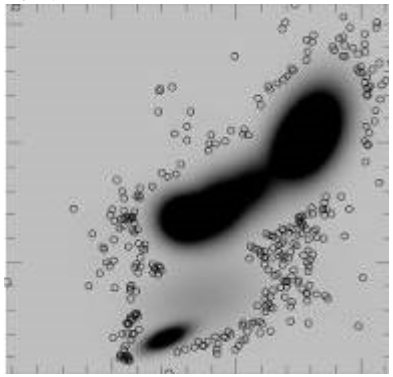
- 分类分析
- 回归分析
- 时序数据分析
- 数据总结
- 聚类分析
- 关联规则分析
- 序列模式分析
- 依赖关系分析
- 偏差分析

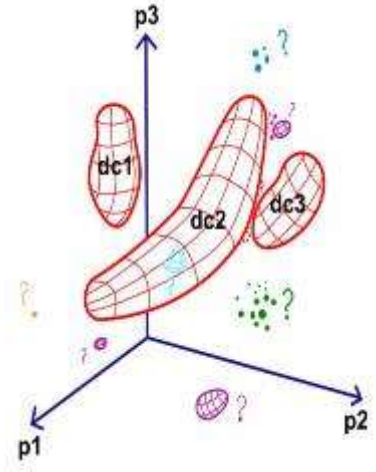
预测性

描述性

多种挖掘任务：

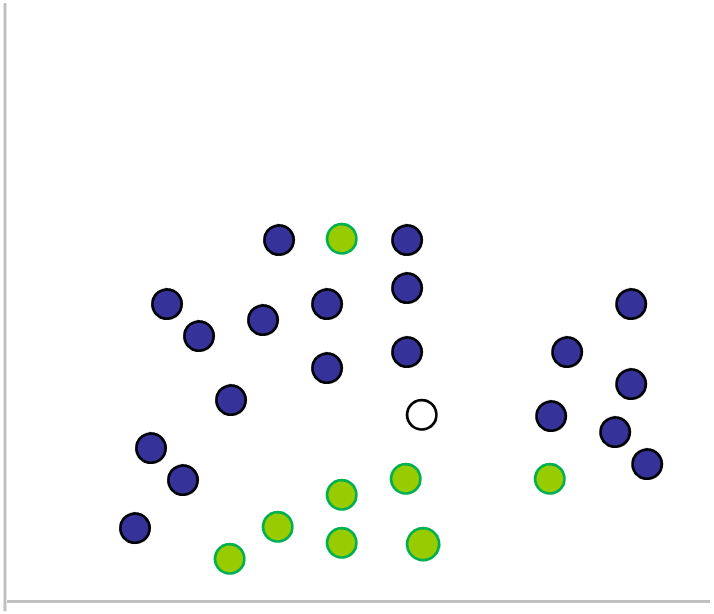
- 经常交叉相关
- 每一种任务经常尝试多种不同的算法和技术
- 每一种任务可以通过不同的知识发现过程来实现





分类分析（known knowns）

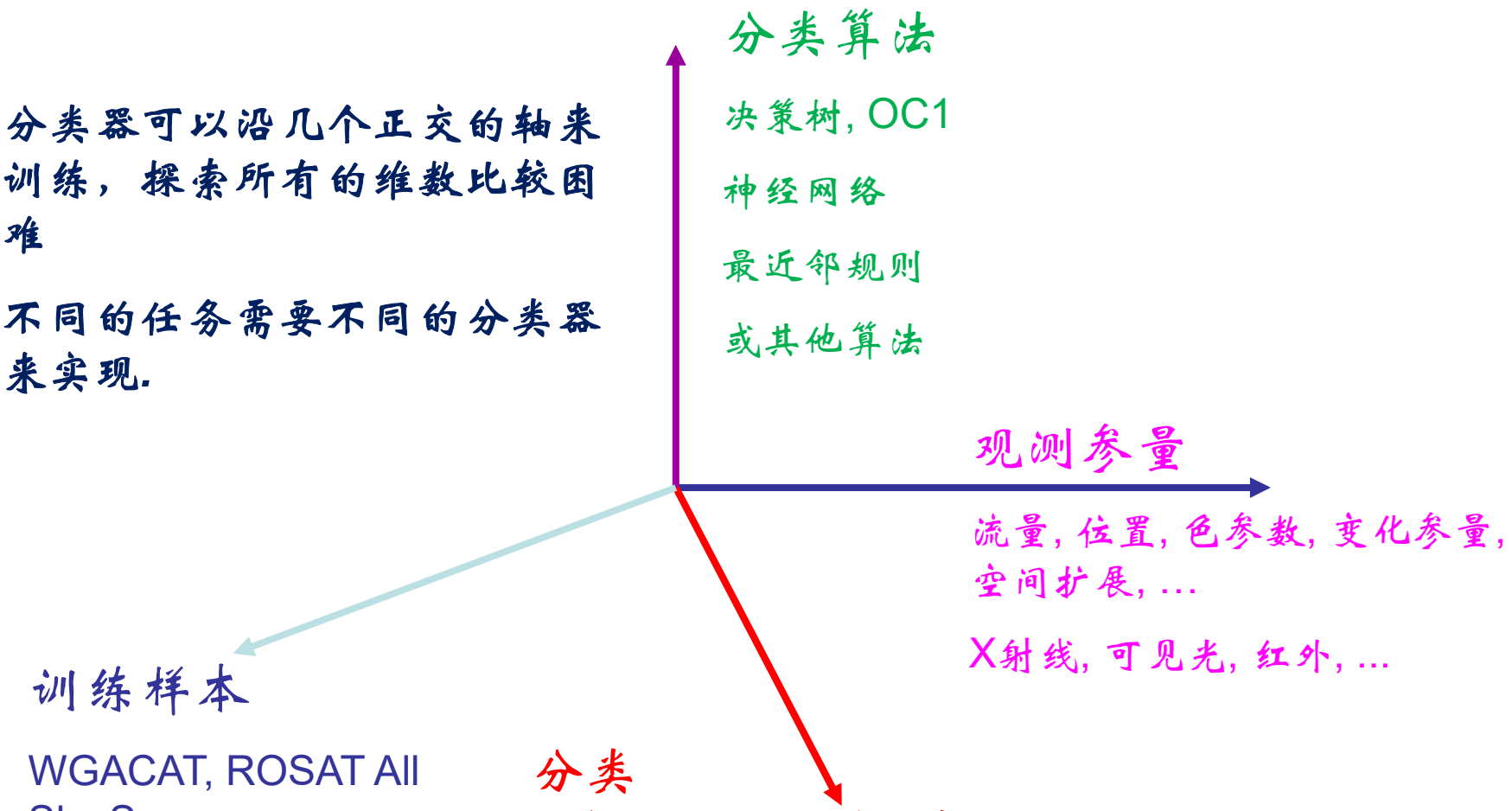
- 定义
  - 按照某种规则，新的数据被划分到已知类别中的一类。
  - 这个规则是通过具有标签的数据进行监督学习获得的。
- 应用
  - 恒星分成不同的光谱型，星系按哈勃或形态分类，活动星系核进一步细分，等等
- 方法
  - 神经网络
  - 决策树
  - Naïve Bayesian Networks
  - 支持矢量机
  - 学习矢量量化
  - 遗传算法
  - .....



采用何种分类器?

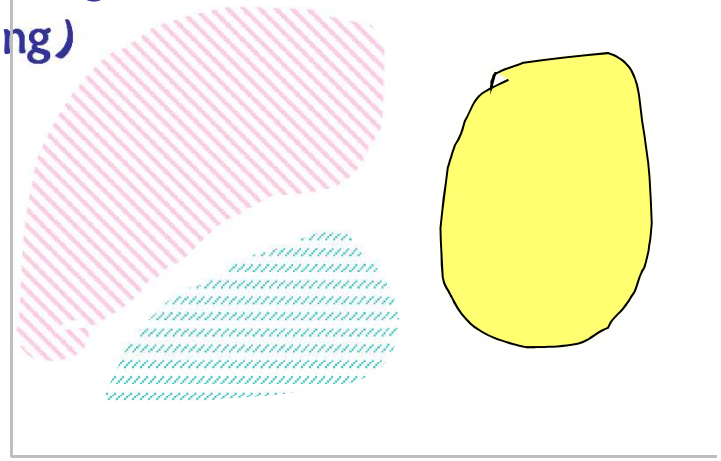
分类器可以沿几个正交的轴来训练，探索所有的维数比较困难

不同的任务需要不同的分类器来实现.

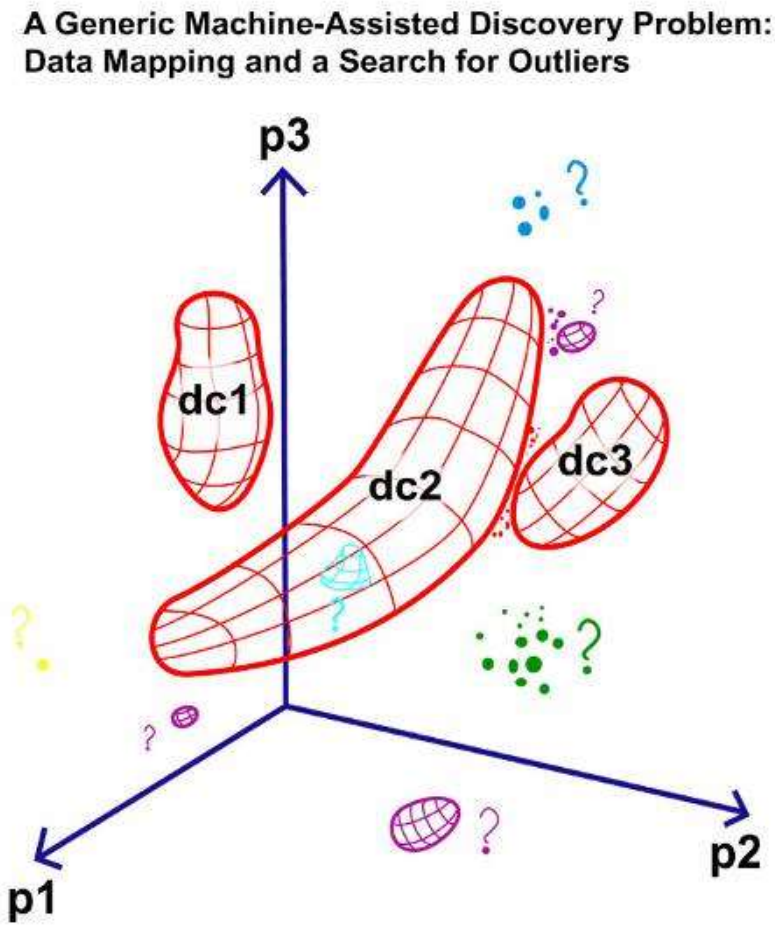


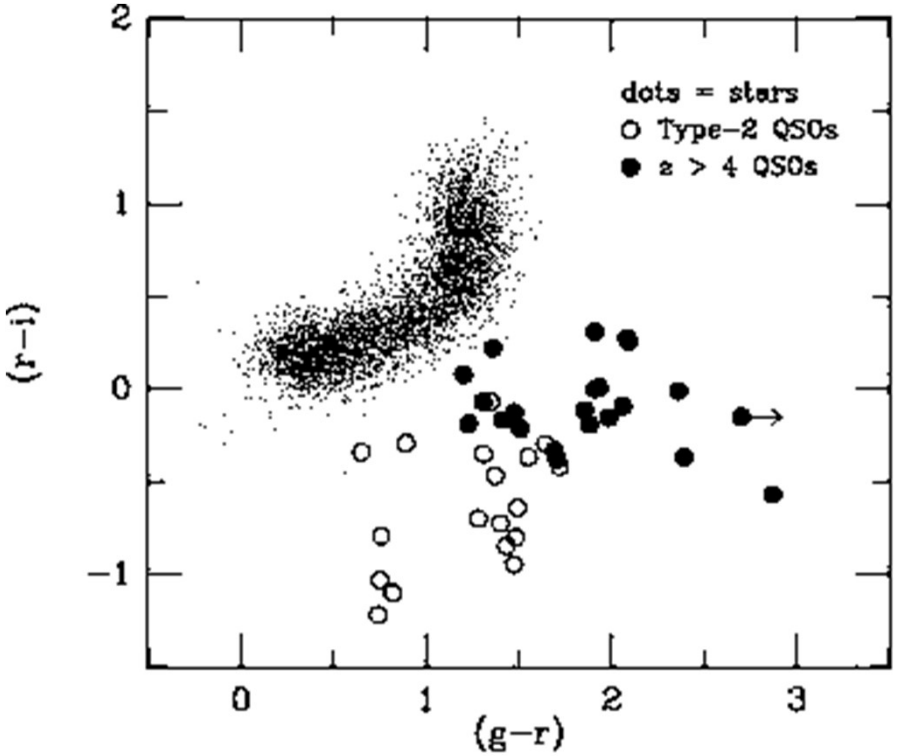
聚类分析（uknown uknowns）

- 定义：
  - 按照某种规律聚在一起的称为一类。
  - 所用的数据是无标签的，通过非监督的学习方式训练数据，类间的差异尽可能地大，而类内的差异尽可能地小。
- 应用：
  - SDSS的双色图恒星聚在一块如香蕉状，类星体则偏离该区域。
- 方法：
  - K均值聚类
  - Hierarchical clustering
  - 预期最大算法（Expectation Maximization algorithm）
  - 高斯混合模型（Gaussian mixture modeling）
  - 主成分分析
  - ……
- 优越性
  - 新的概念(Concept discovery)
  - 点滴知识(Bootstrapping knowledge)



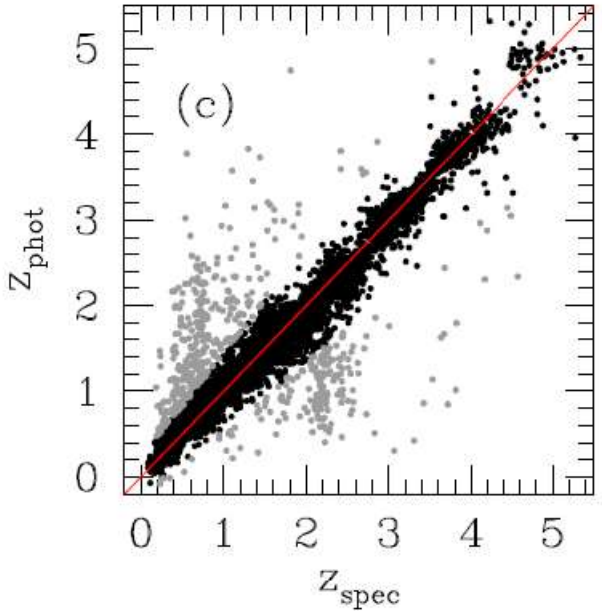
聚类分析


A Generic Machine-Assisted Discovery Problem:
Data Mapping and a Search for Outliers
p3
p1
p2
dc1
dc2
dc3


dots = stars
Type-2 QSOs
z > 4 QSOs
(r−i)
(g−r)
2
1
0
−1
0
1
2
3

回归分析（known uknowns）

- 定义
  - 将一个连续应变量建模为一个或多个预测元的函数。
  - 这个规则是通过具有标签的数据进行监督学习获得的。
- 应用
  - 恒星物理参数（Teff、g、[Fe/H]）的测量，星系和类星体的测光红移，等等
- 方法
  - 神经网络
  - 决策树
  - kNN
  - 支持矢量机
  - 核回归
  - .....



分类分析与回归分析

- 分类分析:
  - **预测种类标签(离散的或名词的)**
  - **基于训练数据集和种类标签来构建模型，用于对新数据分类**
- 回归分析:
  - **预测值是连续变量**
  - **基于连续的标签创建模型，预测未知或缺值**

- ***典型应用***
  - **信用卡批准**
  - **目标市场**
  - **医疗诊断**
  - **诊治疗效分析**

分类分析与聚类分析

- **监督学习（分类）**

  --监督：训练数据有标签，表明数据记录的类型

  -- 新数据分类是基于训练样本

- **非监督学习（聚类）**

  -- 训练样本无标签

  --给定一组观测量、测量量等，目的是找到其所属已经存在类别

模型评估方法（1）

- **训练和评估模型，数据集通常分成三份：训练集、测试集、评估集**
  - 训练集
    用于建立初始模型
  - 测试集
    调整模型，以便建立更加通用的模型，思想是防止过度训练
  - 评估集
    评估模型的性能

模型评估方法（2）

- **样本外检测（ out of sample testing ）**

  随机从样本中选出部分作为验证数据，其余作为训练数据，训练集与测试集不交叉，测试集一般少于原样本的三分之一。

- **交叉验证(cross validation)**
  - **K-折交叉验证（k-fold cross validation）**
  - **留一验证(leave-one-out testing)**

- **Bootstrap方法**

  是一种有放回的抽样统计过程。

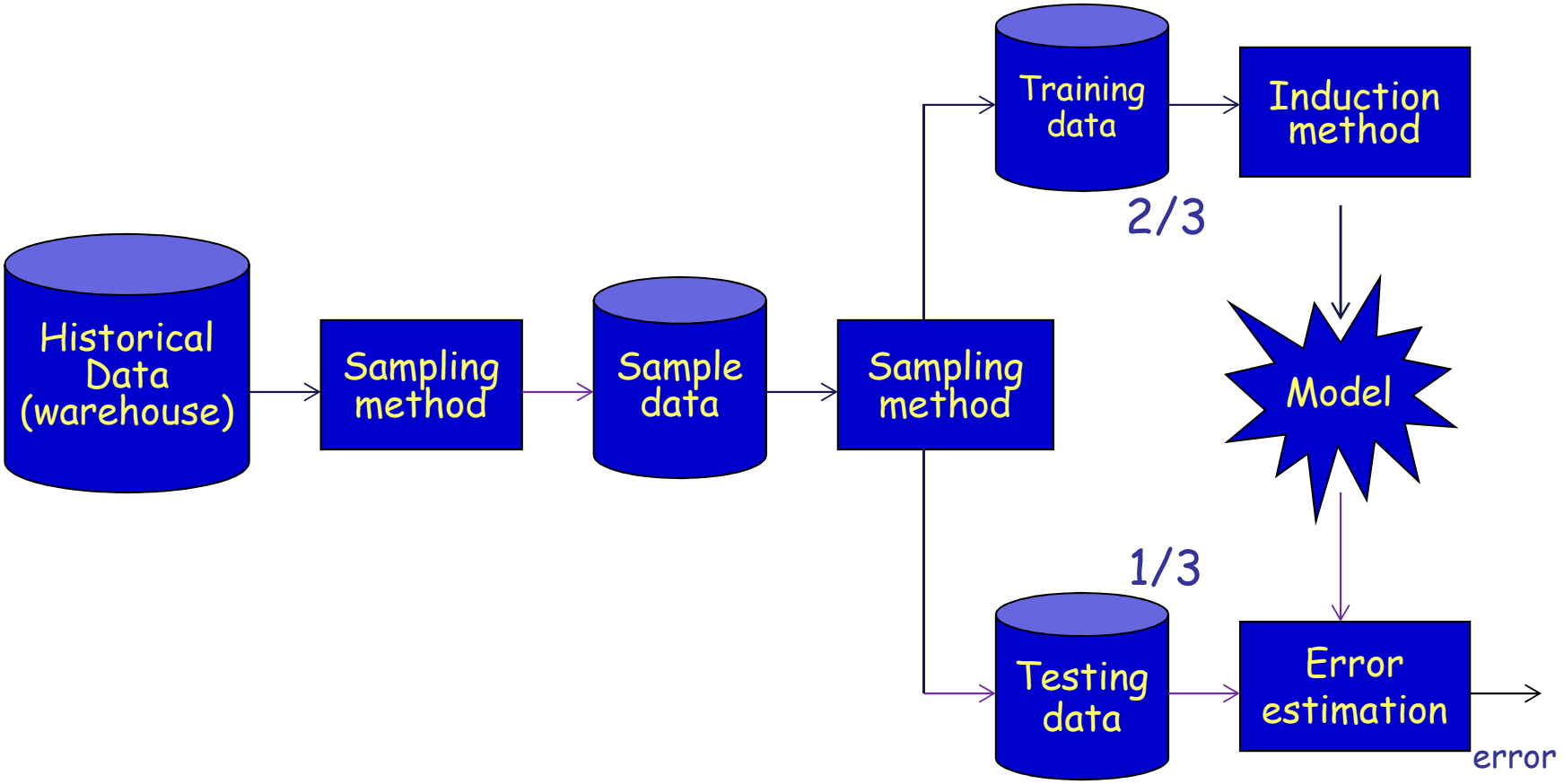
  条件：样本足够多，致使抽样能够体现原始样本

  思想：从一个数据集中有放回的抽样，形成训练集

  优点：是小数据集的错误率估计的最好方法

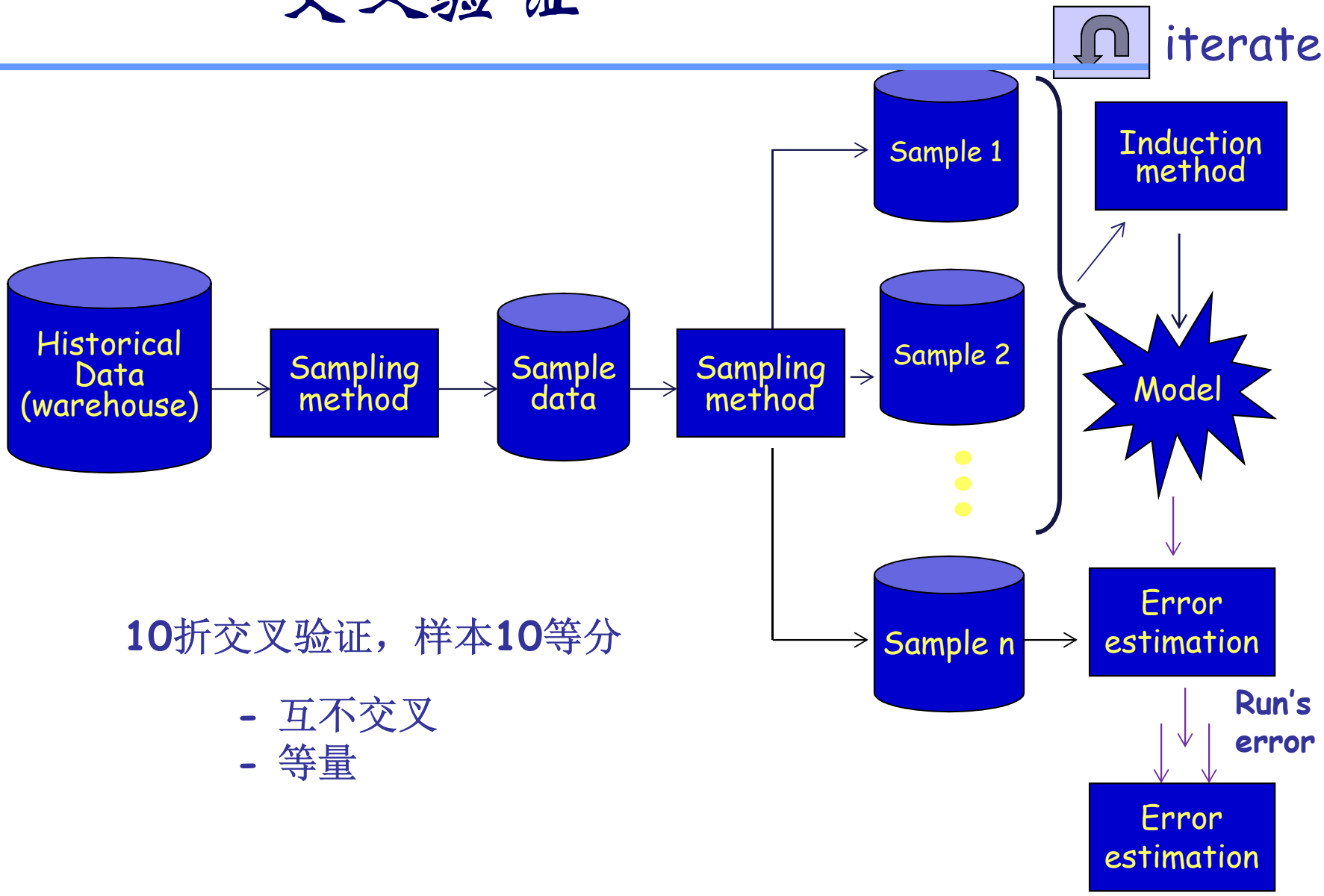
  缺点：一个随机的数据集，被等量分成两类，因此真实的错　　误率为0.5

样本外检验



检测样本的评估质量依赖于检测样本是否具有代表性和独立假设的有效性。

交叉验证



10折交叉验证，样本10等分

- 互不交叉
- 等量

分类模型有效性的评估方法

- **混淆矩阵（confusion matrix）**
  用来反映某一个分类模型的分类结果的，其中行代表的是真实的类，列代表的是模型的分类。
- **精确率Precision**
  检索出来的条目中有多少是准确的。
- **召回率Recall（完备性）**
  所有准确条目有多少被检索出来了。
- **F-measure**
  是精确率和召回率的加权平均。
  当F-measure较高时，说明实验方法比较理想。(非平衡样本有效)

	Predicted positive class	Predicted negative class
Actual positive class	TP (true positive)	FN (false negative)
Actual negative class	FP (false positive)	TN (true negative)

$$\text{Accuracy} \quad (\text{Acc}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{True Positive Rate} \quad (\text{Acc}^+) = \frac{\text{TP}}{\text{TP} + \text{FN}} = \text{Recall}$$

$$\text{True Negative Rate} \quad (\text{Acc}^-) = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

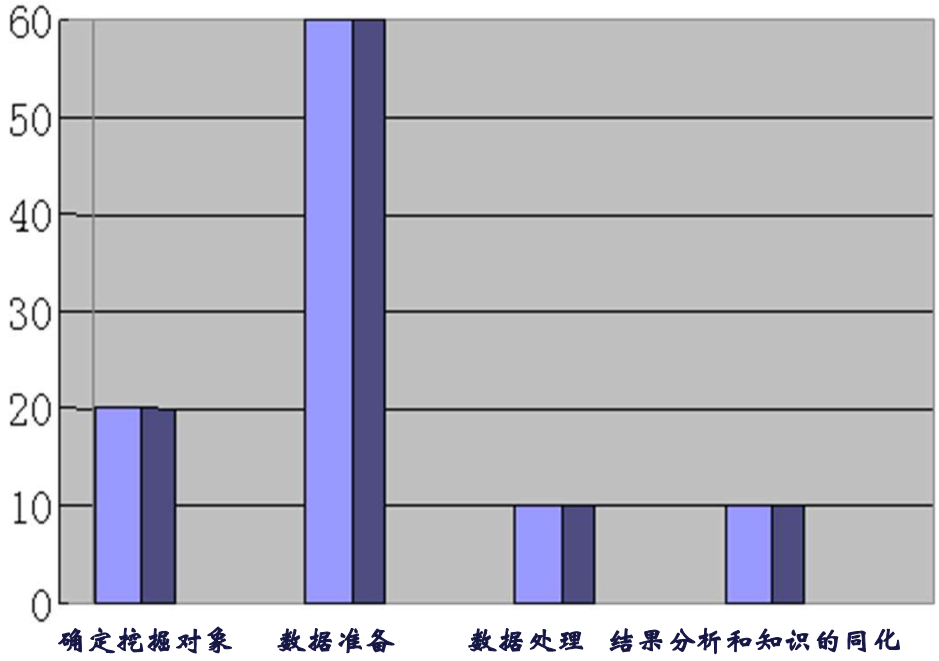
$$F - measure \quad (\text{FM}) = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{G} - \text{mean} \quad (\text{GM}) = (\text{Acc}^- \times \text{Acc}^+)^{\frac{1}{2}}$$

$$\text{WeightedAccuracy} \quad (\text{WA}) = \beta \times \text{Acc}^+ + (1 - \beta) \times \text{Acc}^-$$

数据挖掘前提：数据准备

- 数据的清洗
  - -- 预处理数据，去掉噪声，处理缺值数据
- 相关性分析(特征选择)
  - -- 去掉不相关或冗余变量
- 数据转换
  - -- 推广和/或规范化数据



如何从众多的算法中挑选出最优或较优的方法？

- 预测准确性
- 速度和可伸缩性
  - -- 建模的时间
  - -- 预测时间
- 健壮性
  - -- 处理噪声和缺值数据
- 可扩展性
  - -- 数据库中的效率
- 可解释性
  - -- 模型和规则的可理解性
- 可控性
  - -- 决策树的大小
  - -- 分类规则的简洁性

竞技场中无常胜将军，
算法常不断推陈出新！
只有更好，没有最好！

常用的分类方法

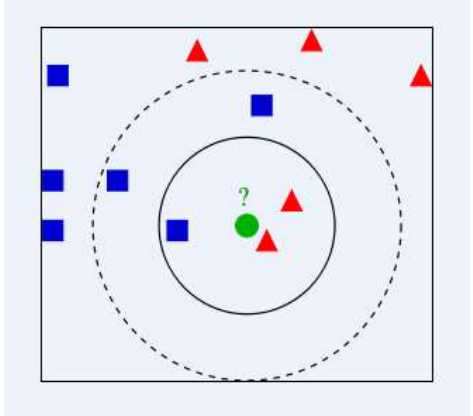
- 基于实例的学习
- 决策树
- 神经网络
- 支持矢量机
- 贝叶斯分类
- 案例推理
- 粗糙集
- 模糊集
- 遗传算法

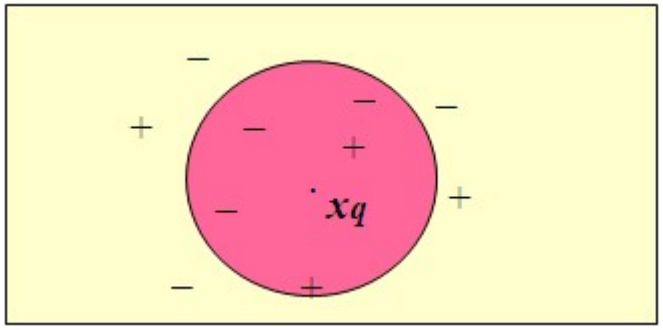
基于实例的学习

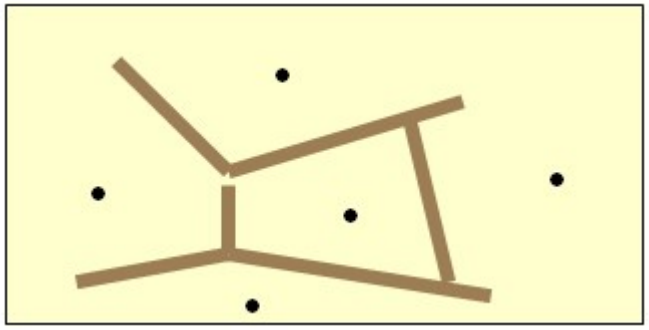
- 基于实例的学习
  - -- 存储训练样本，延迟学习直到有新的实例需要分类，所以又称为懒学习。
- 典型的方法
  - -- k-近邻方法
    - 一个实例作为欧氏空间中的一个点
  - -- 局部加权回归
    - 创建局部近似
  - -- 案例推理
    - 应用符号表示和知识基推理

k近邻方法

- 所有的实例对应n维空间中的点
- 最近邻通常用欧氏空间考察
- 目标函数是离散值或实数
- 对离散值，k-NN返回离预测点最近的k个实例中最频繁的类别
- 沃罗诺伊图：决策平面是通过训练样本的一组典型的1-NN来构成







k近邻方法

- K-NN预测连续值
  -- 计算k个最近邻的平均值

- 距离加权的近邻算法

$$\hat{f}_{kern}(x,h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

  -- 按照它们距预测点的距离加权，越近的权重越大。

- 由于对k个近邻取平均，可以容忍噪声数据

- 维数灾难：邻居之间的距离会被不相关的变量主导。
  -- 为此，去掉不相关和冗余的变量是很必要的

常见的决策树方法

- ID3
- C4.5
- CART
- IBM IntelligentMiner
- Random Forest

可伸缩的决策树方法

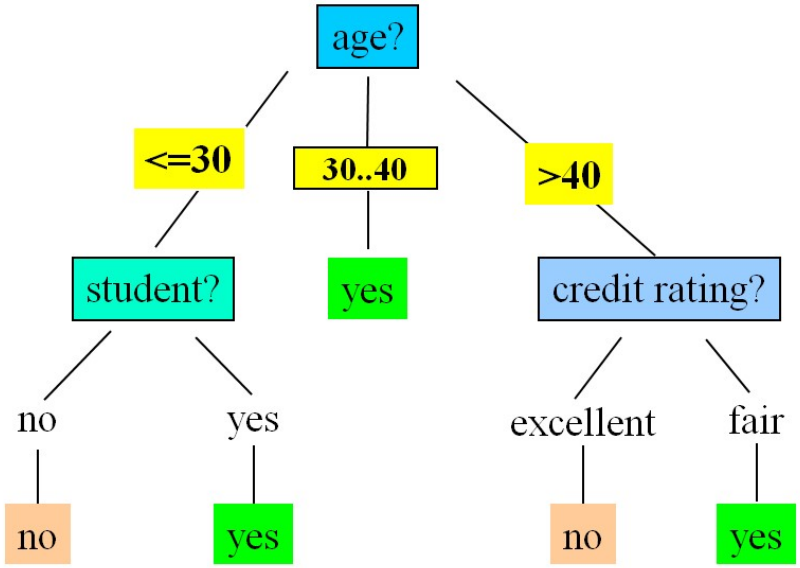
- SLIQ (EDBT’96 — Mehta et al.)
- SPRINT (VLDB’96 — J. Shafer et al.)
- PUBLIC (VLDB’98 — Rastogi & Shim)
- RainForest (VLDB’98 — Gehrke, Ramakrishnan & Ganti)

决策树

- 基础方法（贪婪算法）
  - -- 树是自上而下通过递归各个击破的方式建立的
  - -- 最初，所有的训练集在根部
  - -- 属性是类型值，如果是连续的，要提前离散化处理
  - -- 样本按照选定属性来递归分割
  - -- 检测属性的选择是基于启发式或统计方法（如：信息增益、基尼指数）
- 停止分割的条件
  - -- 给定节点，所有样本属于同一类
  - -- 没有可以再进一步分割的属性，叶节点分类采取多数决
  - -- 没有样本剩余

决策树提取分类规则

- 提取的知识用IF-THEN规则来表示
- 每条路径的规则是从根节点到叶节点
- 沿着一条路径的每个属性对形成一个组合条件
- 叶节点给出种类预测
- 获得的分类规则，人们易于理解

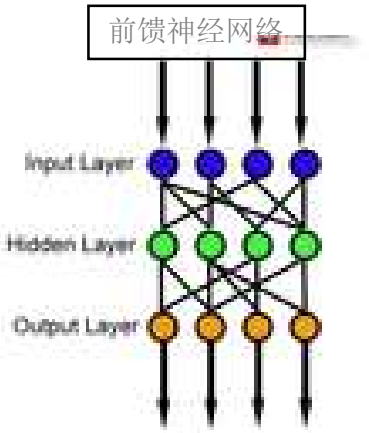


决策树的改进

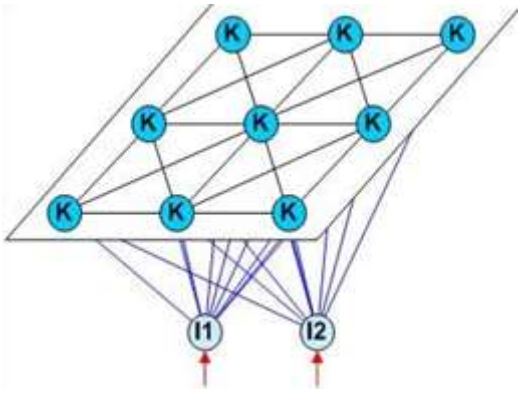
- 支持实型属性
  - -- 动态地定义新的离散值属性，将连续属性对应到离散的数据集合中
- 处理缺值属性
  - -- 用最常见的属性值代替
  - -- 每个可能值的概率大的代替
- 属性重建
  - -- 基于现有的稀疏表示创建新的属性
  - -- 尽可能减少分散、重复和复制

- **决策树的优点**
  - -- 相比其他分类方法，相对较快的学习速度
  - -- 可以获得简单的易于理解的分类规则
  - -- 可以用SQL查询直接与数据库相连
  - -- 与其他数据挖掘方法有相当的分类精度

常用的神经网络方法



- 感知器神经网络
- 线性神经网络
- 递归神经网络
- BP传播神经网络
- 径向基神经网络
- Hopfield神经网络
- 学习矢量量化
- 自组织竞争型神经网络
- Simulink神经网络工具箱



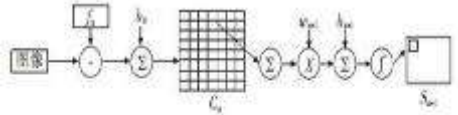


图 4 CNN 中卷积和采样过程

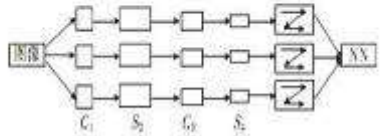
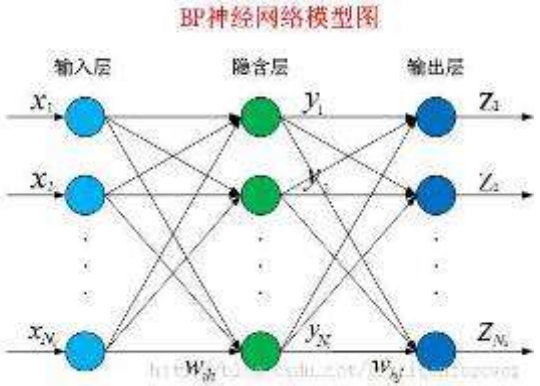


图 5 卷积神经网络的原理



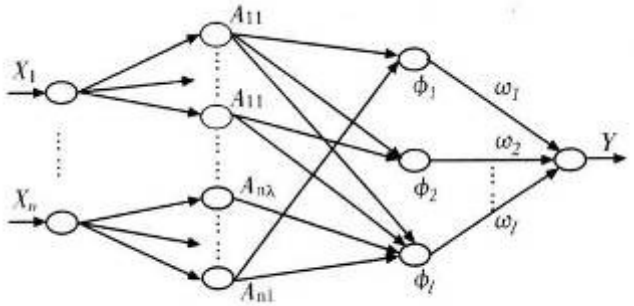
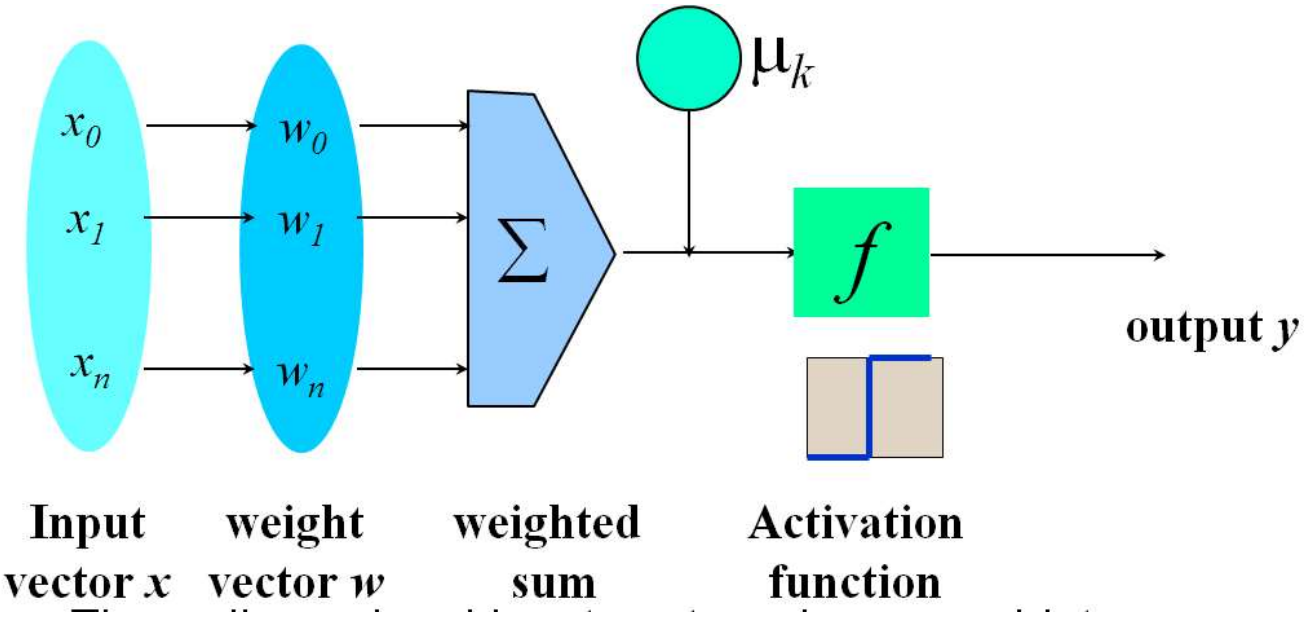


图 1 模糊神经网络的结构

神经网络

- 类似于生物系统（一个很好的学习系统）
- 适合并行，提高计算效率
- 第一个学习算法始于1959年（Rosenblatt），如果提供了目标输出值的单个神经元有固定的输入，可以递增地改变权重以利用感知器的学习规则来产生这些输出



例如

$$y = \frac{1}{1 + e^{-x}}$$

多层神经网络

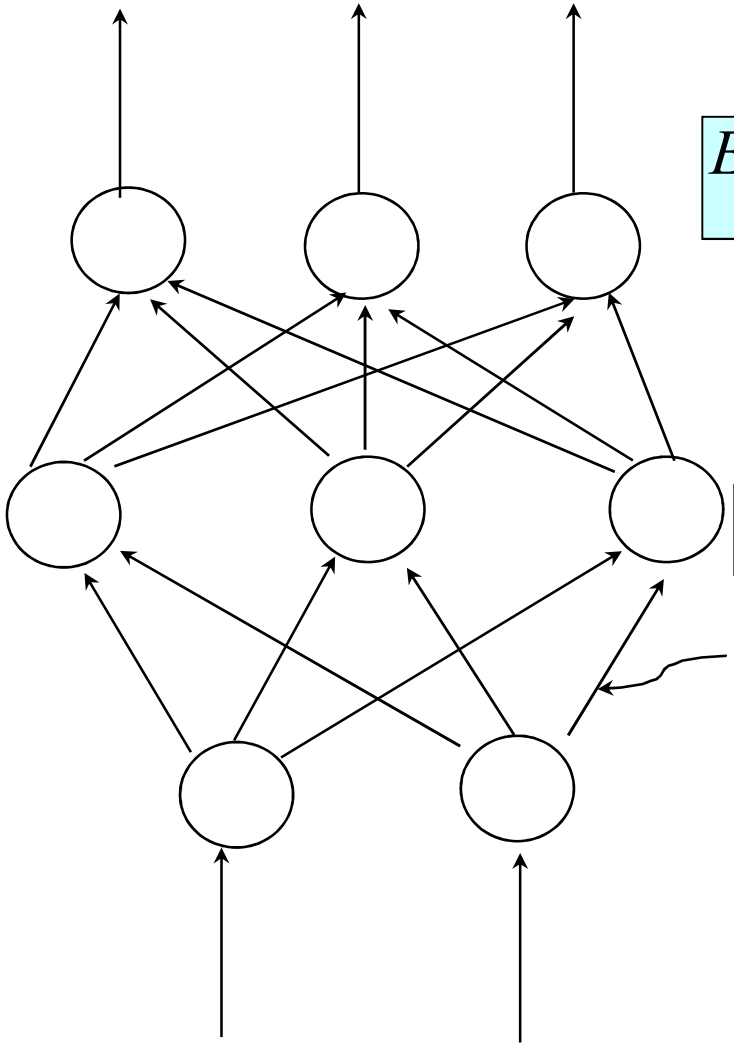
输出向量

输出节点

隐层节点

输入节点

输入向量: x_i



w_{ij}

$$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$$

$$\theta_j = \theta_j + (l)Err_j$$

$$w_{ij} = w_{ij} + (l)Err_j O_i$$

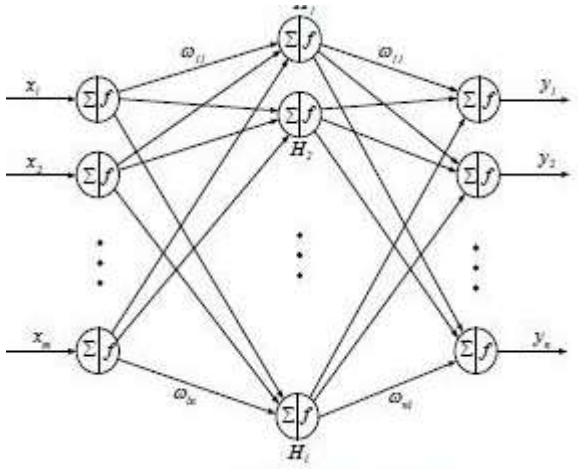
$$Err_j = O_j(1-O_j)(T_j - O_j)$$

$$O_j = \frac{1}{1+e^{-I_j}}$$

$$I_j = \sum_i w_{ij} O_i + \theta_j$$

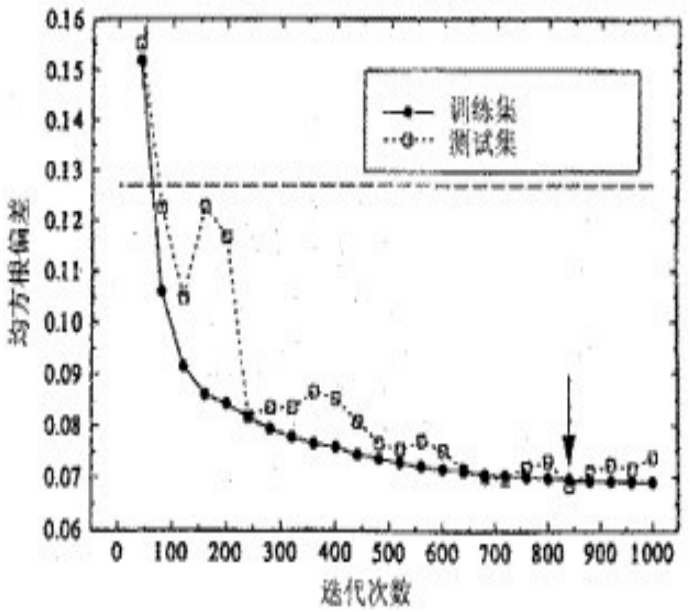
神经网络训练

- 训练的最终目标
  - -- 获得一组权重，使得几乎所有训练样本中的元组正确分类
- 步骤
  - -- 权重起初随机取值
  - -- 将元组一个接一个输入网络
  - -- 对每一个单元
    - * 每个单元的净输入是所有与该单元相连的输入的线性组合
    - * 通过激活函数计算输出值
    - * 计算误差
    - * 更新权重和偏置



网络修剪和规则提取

- 网络修剪
  - -- 完全相连的网络是难于表达的重
  - -- N个输入节点，h个隐层节点，m个输出节点导致h（m+N）个权重
  - -- 修剪：移走那些不影响网络分类精度的链接
- 从训练的网络中提取规则
  - -- 对激活值离散化
  - -- 用类平均值代替单个激活值以维持网络的精度
  - -- 枚举的方法从离散的激活值的输出中找到激活值与输出值之间的规则
  - -- 找到输入与激活值之间的关系
  - -- 结合上述两点找到输入与输出之间的规则



- ***神经网络训练需要注意：***
- 过拟合（ overfitting ）

  使用过多参数，以致太适应训练样本而非一般情况，使用最小最佳支援值避免过拟合
- 乏适（ underfitting ）

  使用太少参数，以致于不适应训练样本，或称拟合不足

神经网络的特征

- **非线性**

非线性关系是自然界的普遍特性。大脑的智慧就是一种非线性现象。人工 神经元处于激活或抑制二种不同的状态，这种行为在数学上表现为一种非线性关系。具有阈值的神经元构成的网络具有更好的性能，可以提高容错性和存储容量。

- **非局限性**

一个神经网络通常由多个神经元广泛连接而成。一个系统的整体行为不仅取决于单个神经元的特征，而且可能主要由单元之间的相互作用、相互连接所决定。通过单元之间的大量连接模拟大脑的非局限性。联想记忆是非局限性的典型例子。

- **非常定性**

人工神经网络具有自适应、自组织、自学习能力。神经网络不但处理的信息可以有各种变化，而且在处理信息的同时，非线性动力系统本身也在不断变化。经常采用迭代过程描写动力系统的演化过程。

- **非凸性**

一个系统的演化方向，在一定条件下将取决于某个特定的状态函数。例如能量函数，它的极值相应于系统比较稳定的状态。非凸性是指这种函数有多个极值，故系统具有多个较稳定的平衡态，这将导致系统演化的多样性。

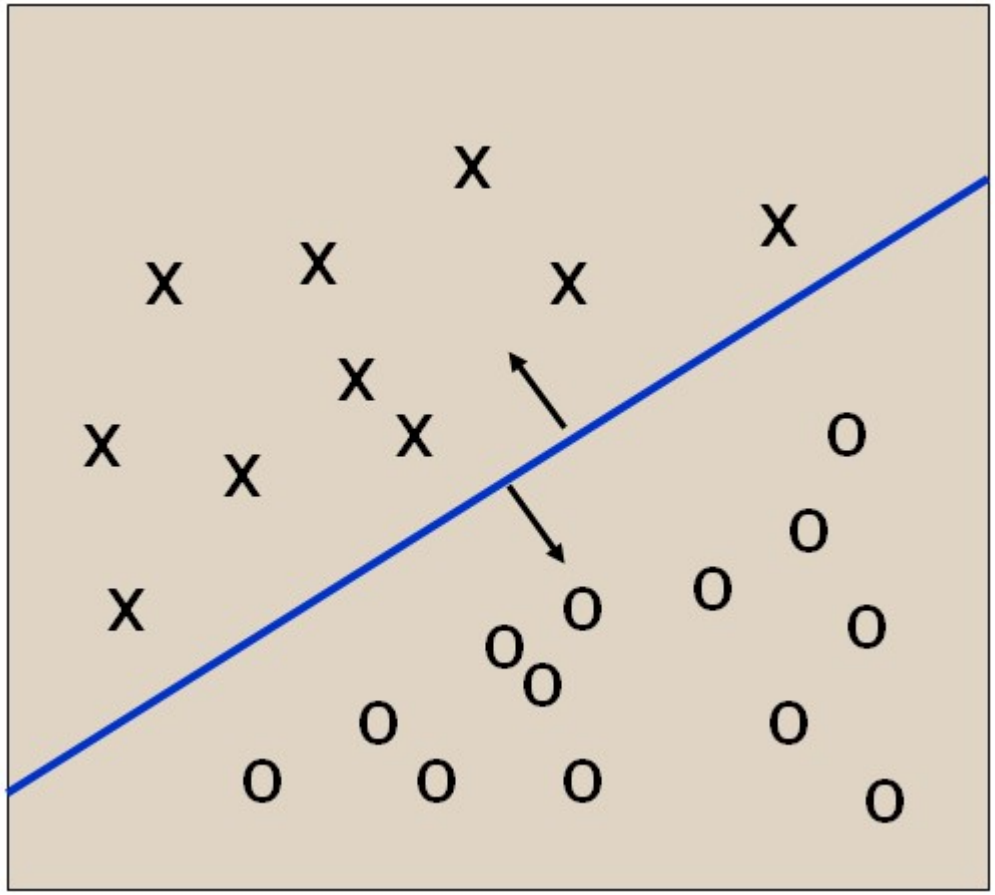
神经网络的分类

- 按网络性能分：连续型与离散型，确定型与随机型
- 按学习方式分：有教师型和无教师型
- 按照突触性质区分：一阶线性和高阶非线性关联网络
- 按网络连接的拓扑结构分类
  -- 层次型结构：输入层、中间层、输出层
  -- 互联型结构：全互联型、局部互联型、稀疏连接型
- 按网络内部的信息流向分类
  -- 前馈型网络和反馈型网络

神经网络的优点

- 并行分布处理
- 高度鲁棒性和容错能力
- 分布存储及学习能力
- 能充分逼近复杂的非线性关系

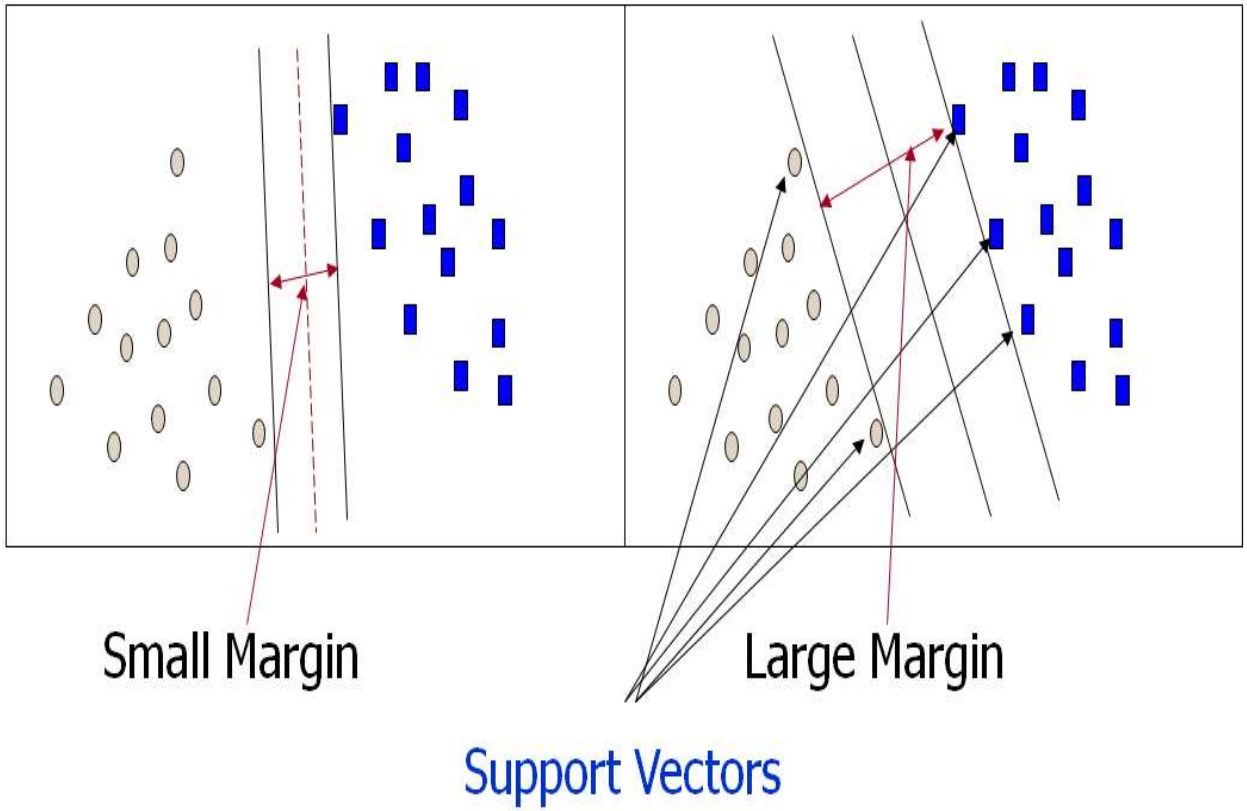
线性分类器



- 二分类问题
- 样本高于蓝线的属于类型“x”
- 样本低于蓝线的属于类型“o”
- 例子：SVM、感知机、概率分类器

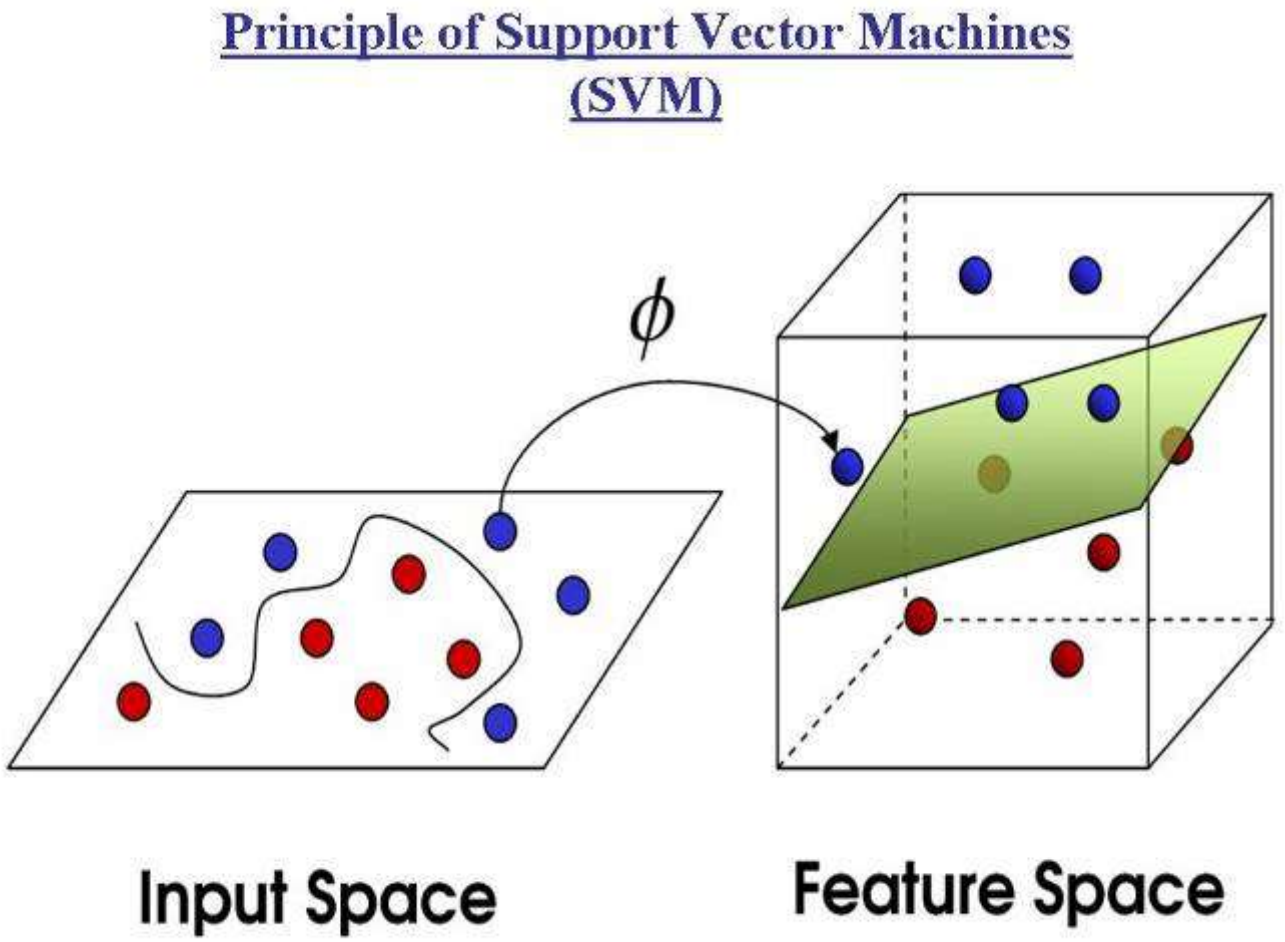
支持矢量机

原理：利用结构风险最小化的原理，即最小化预期风险的上限。通过最大化超平面与任意类训练样本的最小距离或最大化分类边界的距离，从而得到最优超平面。



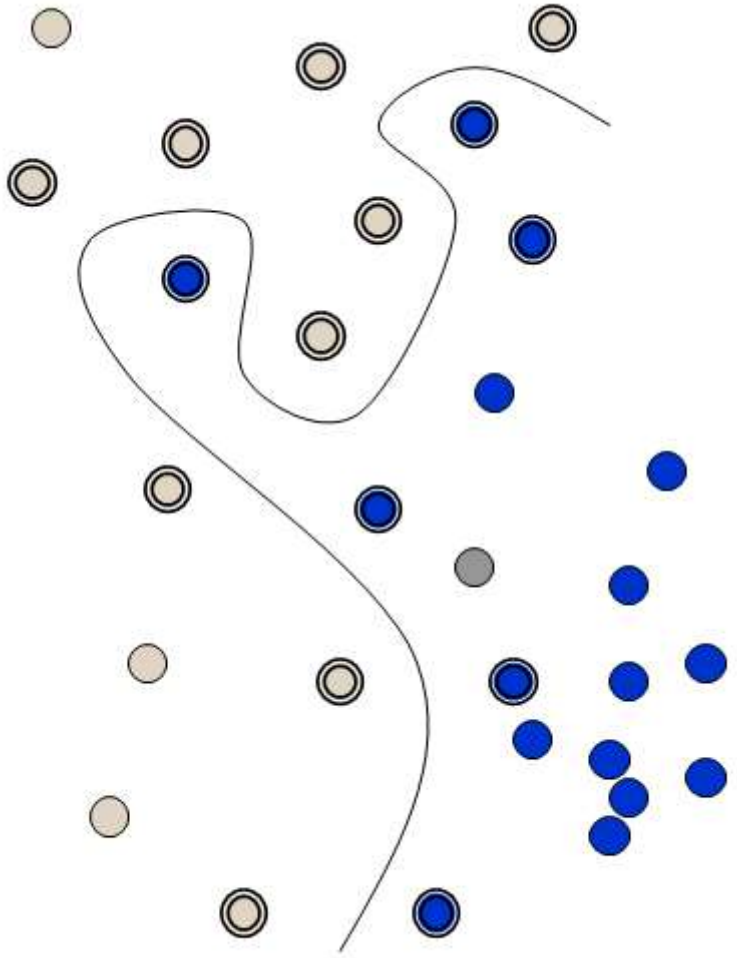
支持矢量机

将数据从低维空间映射到高维空间中，寻找最大间隔分类超平面。



支持矢量机

具有影子的点是支持矢量，很明显它们代表最好的分界。曲线就是分类边界。



支持矢量机分类

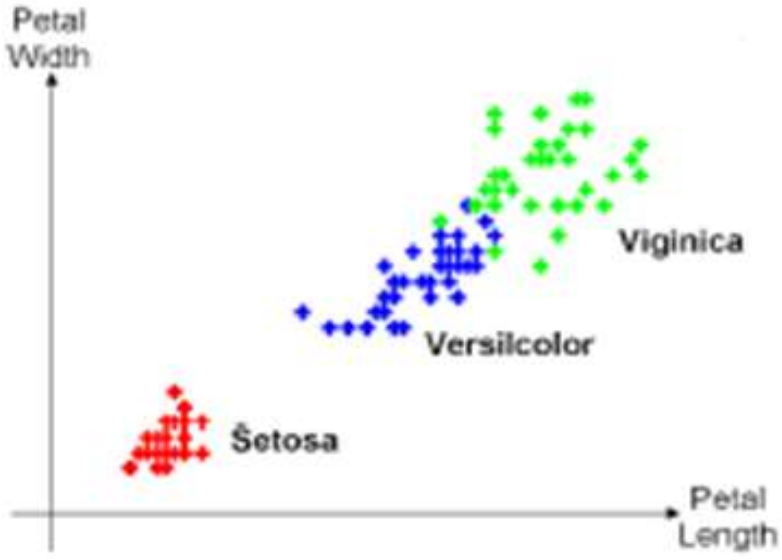
- 线性支持矢量机
- 非线性支持矢量机

  不同的核函数：

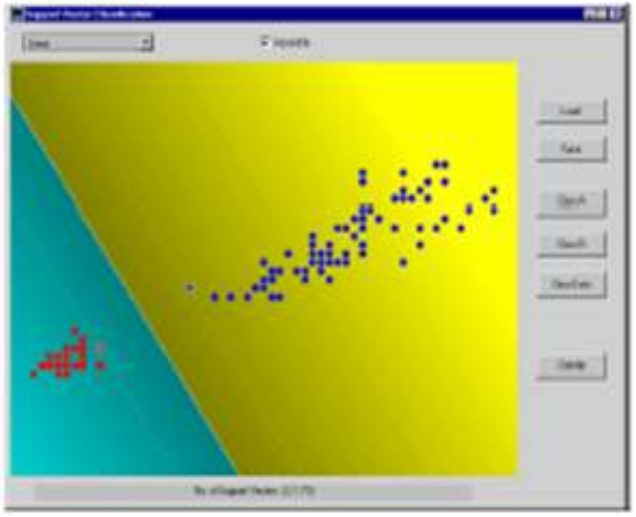
  -- 多项式
  -- 径向基
  -- 多层感知器
  -- 傅里叶级数
  -- 样条函数
  -- 叠加的核函数
  -- 张量积

支持矢量机分类事例

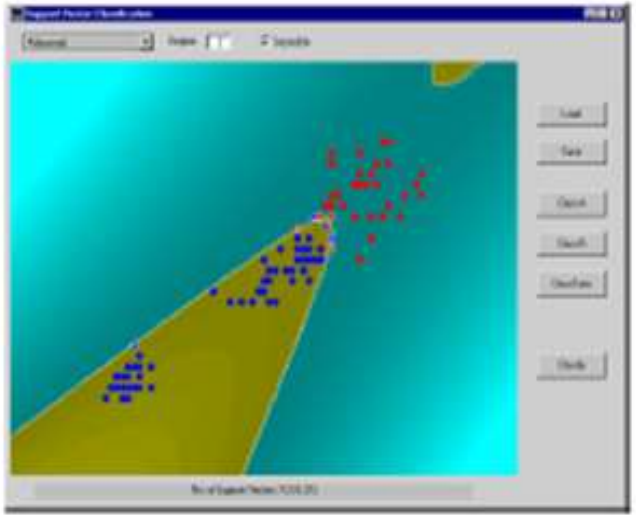
以鸢尾属植物数据分类为例，看一下支持矢量机的工作原理。该数据有4个属性值，为可视化起见，我们只取最主要的两个属性，即花瓣的长度和宽度。



鸢尾属植物数据分布

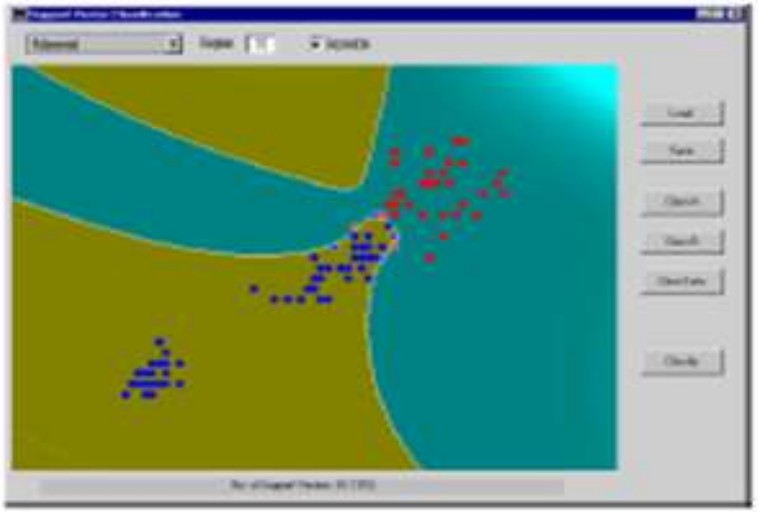


用线性的支持矢量机将 Setosa 数据分出来（$C=\infty$）



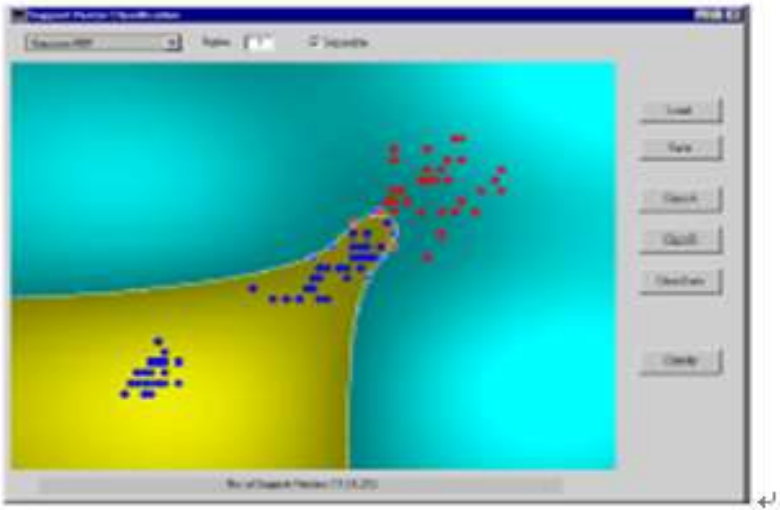
用 2 次多项式的支持矢量机将

Vignica 数据分出来（$C=\infty$）



用 10 次多项式的支持矢量机将 Vignic

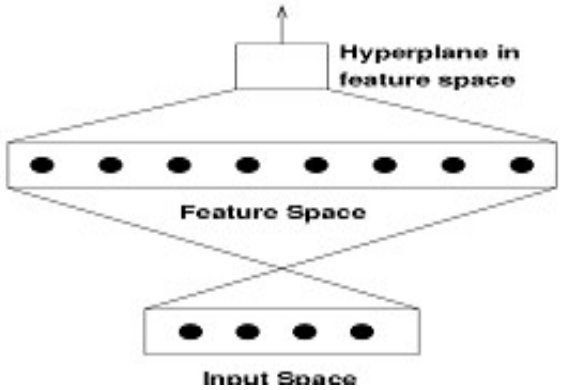
数据分出来（$C=\infty$）



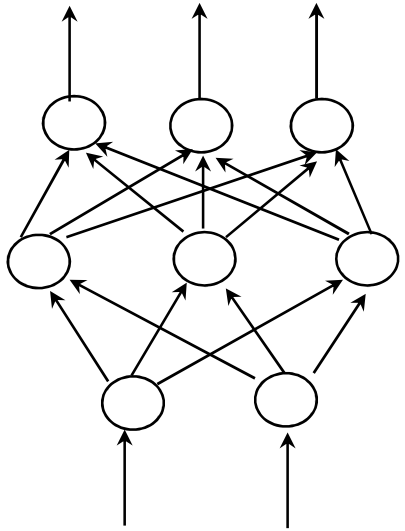
用径向基的支持矢量机将 Vignica

支持矢量机 VS. 神经网络

- 支持矢量机
  - -- 概念新
  - -- 好的推广性
  - -- 学习难，二次规划
  - -- 用核函数可以学习复杂的情形
  - -- 结构风险最小化
  - -- 难并行化



- 神经网络
  - -- 十分古老
  - -- 推广性强但无强的数学基础
  - -- 可以渐进式地学习
  - -- 可以学习复杂的函数（如多层神经网络）
  - -- 经验风险最小化，易陷于局部最小
  - -- 易并行化



聚类分析分类

- 分区算法：

  构建各种分区，然后通过一些标准对其评估

- 层次算法：

  对一组数据或对象使用一些标准分级分解

- 密度为基础的算法：

  基于联系和密度函数

- 网格为基础的算法：

  基于多级粒度结构

- 模型为基础的算法：

  为每个簇创建模型，找到最佳的模型拟合

聚类分析：常见的分区算法

- K均值

  每个类由类中心来表示

- K-medoids 或PAM(Partition around medoids)

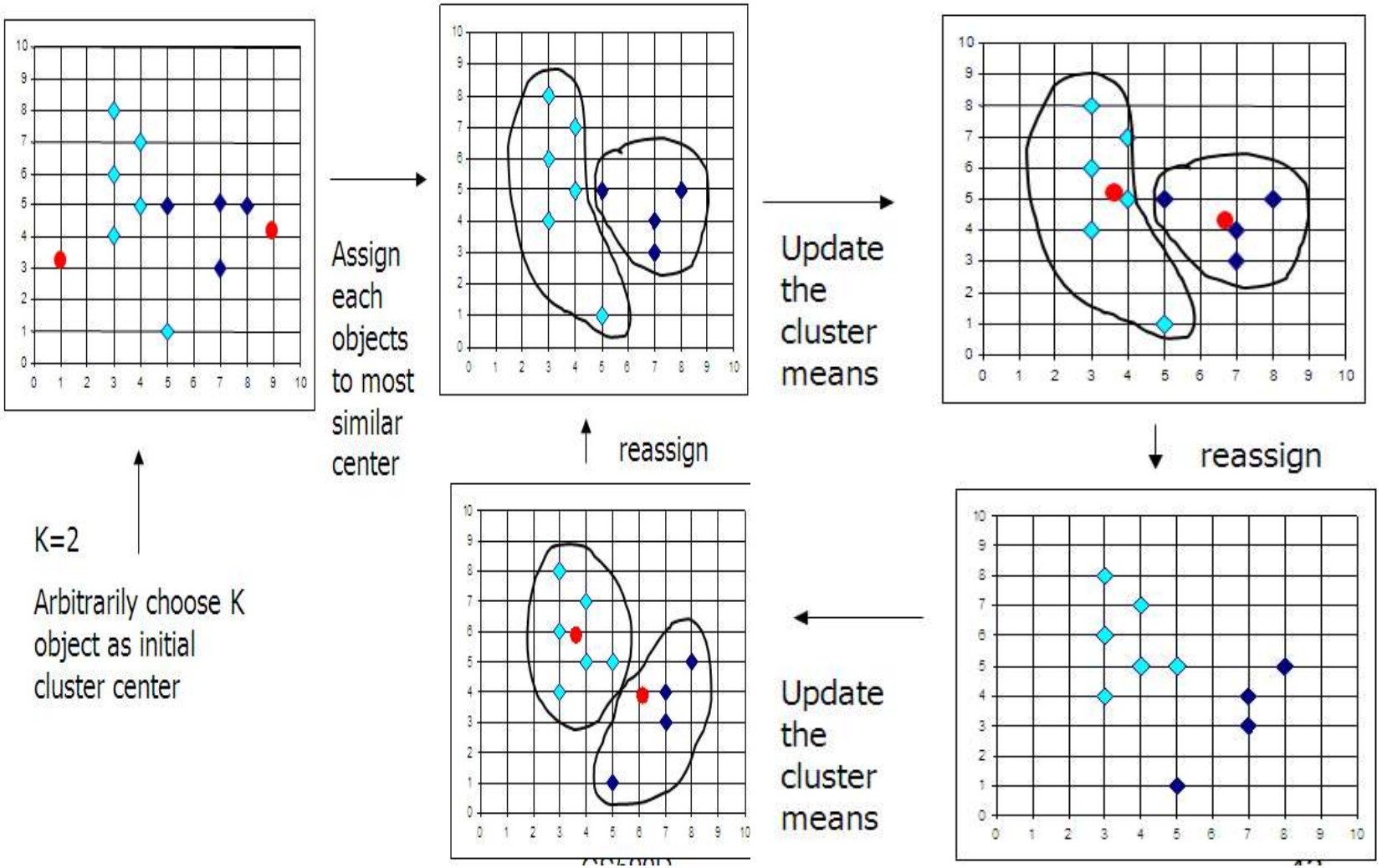
  对每个类由该类中的一个对象来表示

K均值

- 给定K值，k均值需要四步来执行：

  -- 将样本分成k个非空子集
  -- 计算种子点作为目前分割类的中心，每一类的平均中心。
  -- 分配每个对象为最近种子点的类
  -- 返回第2步，当没有新的分类时停止

K均值


K=2
Arbitrarily choose K object as initial cluster center
Assign each objects to most similar center
reassign
Update the cluster means
reassign
Update the cluster means

k均值的优缺点

- 优点：相对而言比较有效，简单易执行
- 点评：易陷入局部最优。全局最优可以用确定性退火和遗传算法来实现
- 缺点

  -- 仅在平均值给定时能用，那么类型数据怎么处理？

  -- 需要预先给定类别数k

  -- 对含噪声和离群数据难于处理

  -- 不适合具有非凸形状的数据聚类

k均值的变种

- 几种k均值的变种：
  - -- 起初k均值的选择
  - -- 相异计算
  - -- 计算类均值的方法
- 处理类型数据：k-modes
  - --用众数代替类的平均值
  - -- 用新的相异计算处理类型数据
  - -- 用基于频率的方法更新类的众数
  - -- 混合类型数据和数值数据：k-prototype方法
- K-Medoids
  - -- 不是拿样本的平均值作为参考点，而是以一个类中最靠近中心位置的点作为参考点

聚类分析：密度为基础的聚类

- 聚类是基于密度的
- 主要特征
  - -- 发现任意形状的簇
  - -- 处理噪声数据
  - -- 一次性处理
  - -- 需要密度参数作为终止参数
- 几个流行的方法
  - -- DBSCAN: Ester, et al. (KDD’96)
  - -- OPTICS: Ankerst, et al (SIGMOD’99).
  - -- DENCLUE: Hinneburg & D. Keim (KDD’98)
  - -- CLIQUE: Agrawal, et al. (SIGMOD’98)

聚类分析：网格为基础的聚类

- 聚类是基于多分辨率网格数据结构

- 几个流行的方法

  -- **STING** (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)

  -- **WaveCluster** by Sheikholeslami, Chatterjee, and Zhang (VLDB’98)，用小波方法聚类

  -- **CLIQUE**: Agrawal, et al. (SIGMOD’98)

数据预处理：降维

- 维灾
  - -- 维数增加，数据变得越来越稀疏
  - -- 数据点之间的距离和密度对聚类和离群数据发现很重要的量，意义逐渐变小
  - -- 子空间的组合呈指数增长
- 降维
  - -- 避免维灾
  - -- 去掉不相关属性和降噪
  - -- 减少时间和空间的浪费
  - -- 有助于可视化
- 降维技巧
  - -- 主分量分析方法
  - -- 特征选择
  - -- 特征重建

消除冗余，简化数据，提高计算效率！

常用的降维方法

- 主成分分析（Principal Component Analysis,PCA)
- 独立成分分析（Idenpendent Component Analysis,ICA ）
- 线性判别分析（Linear Discriminant Analysis,LDA）
- 因子分析（Factor Analysis)
- 多维尺度变换(Multidimensional Scaling,MDS)
- 典型相关分析
- 等距映射（Isomap）
- 局部线性嵌入（Locally Linear Embedding,LLE）
- Laplacian 特征映射（Laplacian Eigenmaps）
- 局部保留投影（Local Preserving Projection,LPP）
- 局部切空间排列（Local Tangent Space Alignment,LTS）
- 最大方差展开（ Maximum Variance Unfolding,MVU）

降维方法分类

- 线性降维：**PCA，LDA，LPP，ICA,FA ,MDS**
- 基于核函数非线性：**KPCA,KICA,KDA,KFDA**
- 非线性流行学习： **Isomap,LLE,Laplacian Eigenmaps,LTSA,MVU**
- 非监督： **PCA,LPP,Isomap,LLE,Laplacian Eigenmaps,LTSA,MVU,ICA,FA,MDS**
- 监督：**LDA**
- 全局： **PCA,LDA,ICA,MDS,FA,Isomap,MVU**
- 局部：**LLE,LPP,LTSA**

Algorithm	Linear	G/L	Supervis	Time
PCA	linear	global	un	0.0321
LDA	linear	global	supervised	0.0029
LPP	linear	local	un	0.0996
Isomap	non	global	un	1.7073
LLE	non	local	un	0.2308
Laplacian	non	local	un	0.1393
LTSA	non	local	un	0.3199
MVU	non	global	un	1.2065

一、样本信息是否利用

- 监督降维方法
- 半监督降维方法
- 非监督降维方法

二、依据所要处理的数据属性类型的不同

- 线性降维方法
- 非线性降维方法

主成分分析（PCA）

对一个样本有 n 个物体、p 个参量 $x_j(j=1,\ldots,p)$，可以找到一组新的正交独立变量，$\xi_1,\ldots,\xi_i,\ldots,\xi_p$，每一个变量是原有变量的线性叠加：

$$\xi_i = a_{i1}x_1 + \cdots + a_{ij}x_j + \cdots a_{ip}x_p$$

确定常数 a_{ij} 使最少数目的新变量可以尽可能地解释样本的变量。那么 ξ_i 称为主分量。

如果原始数据的大多数变量可以仅用 p 个参量中的几个新变量解释，这样我们就找到了原始数据的较简单的描述，以较少的变量对数据分类。有趣的是主分量分析方法表明原始的数据相关，从而导致新的物理观点。若观测的变量不相关，当然也就不会发现主分量。

主成分分析（PCA）

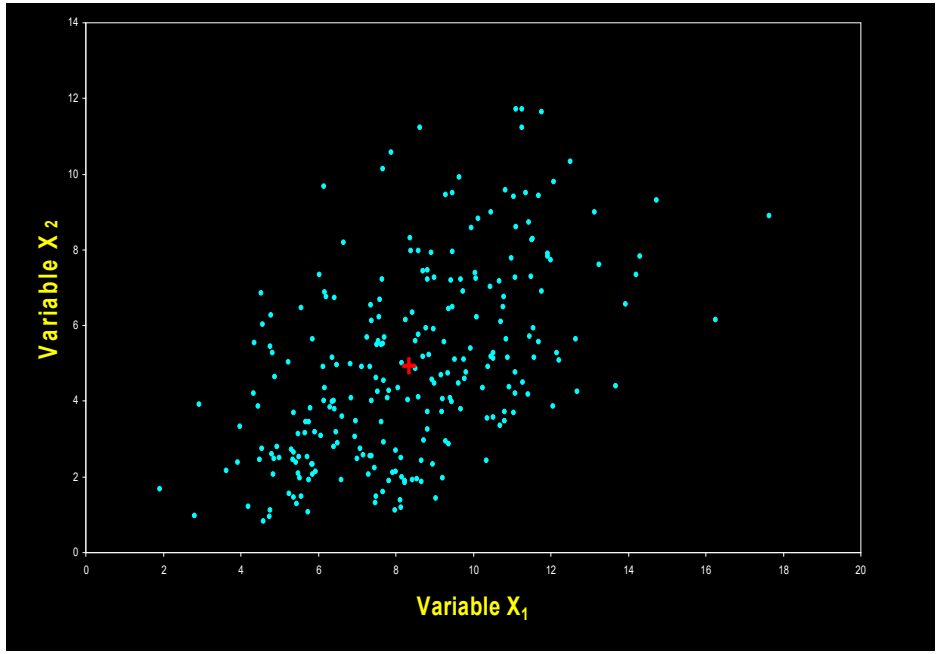
- **PCA** 是广泛应用的数据压缩和降维方法

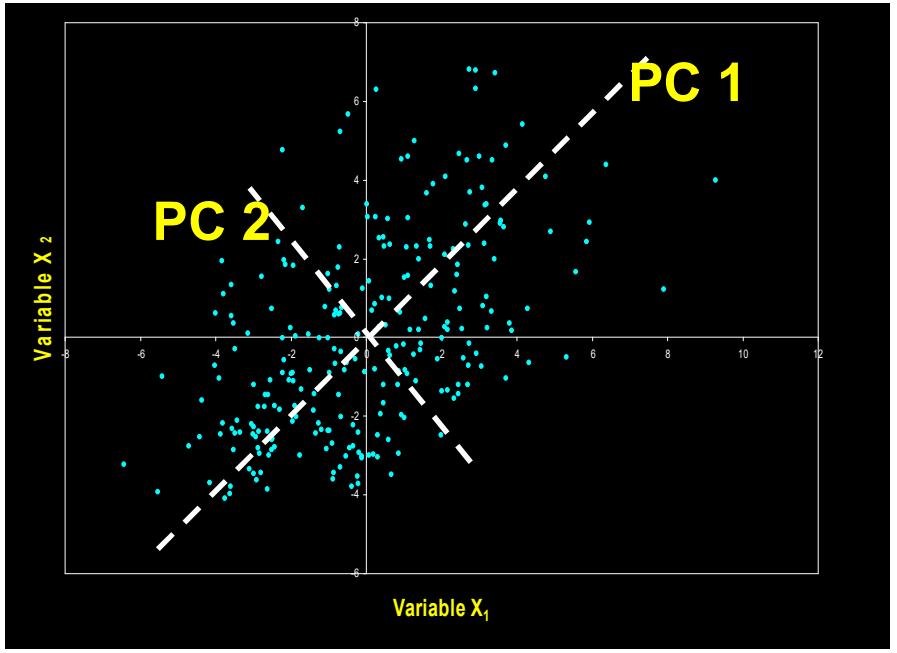
  - **PCA**采用矩阵 ***A***, ***n****个目标，* ***p****个属性*, 这几个属性可以相关，可以沿非相关的轴(主成分或主要轴)来重新构造，即原始***p****个属性的线性叠加*
    - 前**k**个主成分表达了整个样本的变化
    - 剩余的成分可以抛弃，这样在新的低维空间中可以表示原来样本参量的大部分信息

  - **PCA**通过确定协方差矩阵的本征矢量和本征值来实现的。
    - 记住: 两个随机变量的协方差倾向于一起变化

主成分分析的几何解释

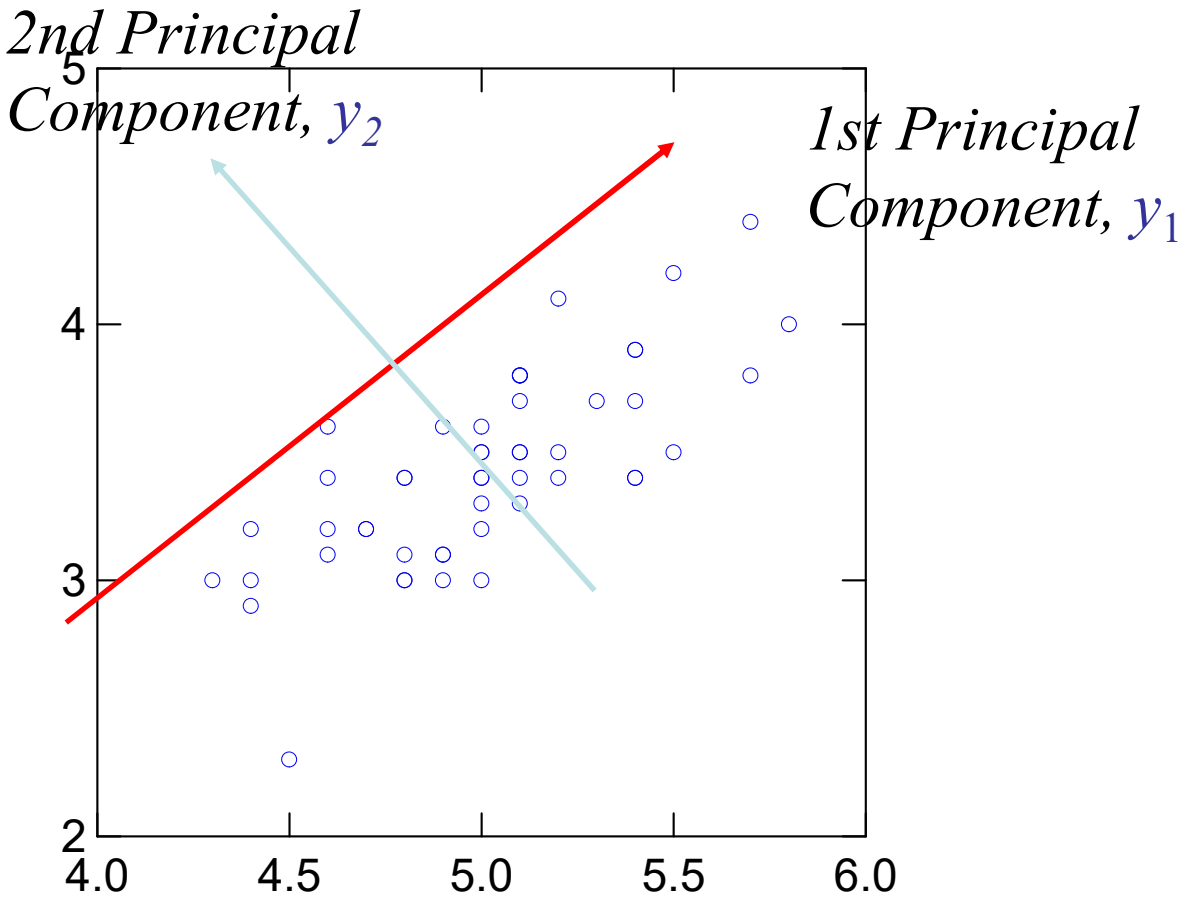
- **目标是旋转p维空间的轴到新的位置，具有如下特征：**
  - 第一主成分描述方差最大的，第二主成分描述方差次大，以此类推…
  - 每一对主成分的协方差为零，即主轴是非相关的



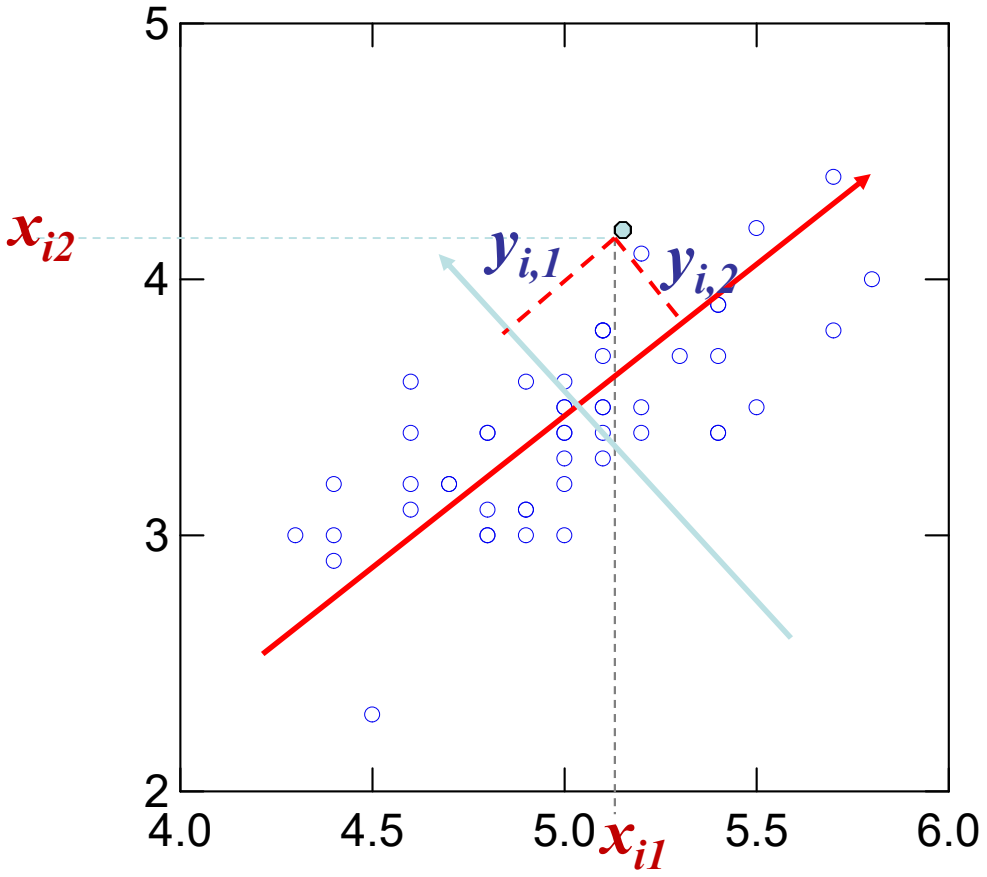


注意: 每一个主轴是原轴的线性叠加

主成分

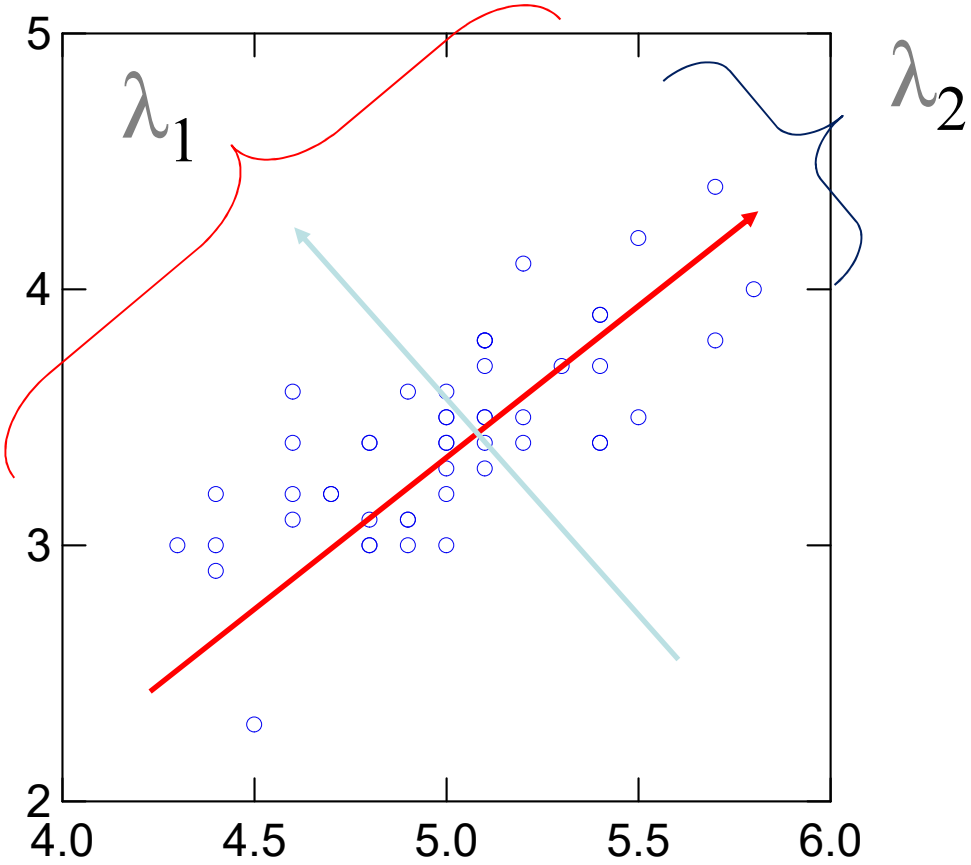

2nd Principal
Component, y_2
1st Principal
Component, y_1
5
4
3
2
4.0
4.5
5.0
5.5
6.0

主成分: Scores



主分量：本征值

本征值代表了沿各个主成分方向的方差



主成分分析的作用

- 降维
- 确定线性混合变量
- 特征提取
- 多维变量的可视化
- 证认隐含变量
- 数据聚类或发现离群数据

局部线性嵌入(LLE)

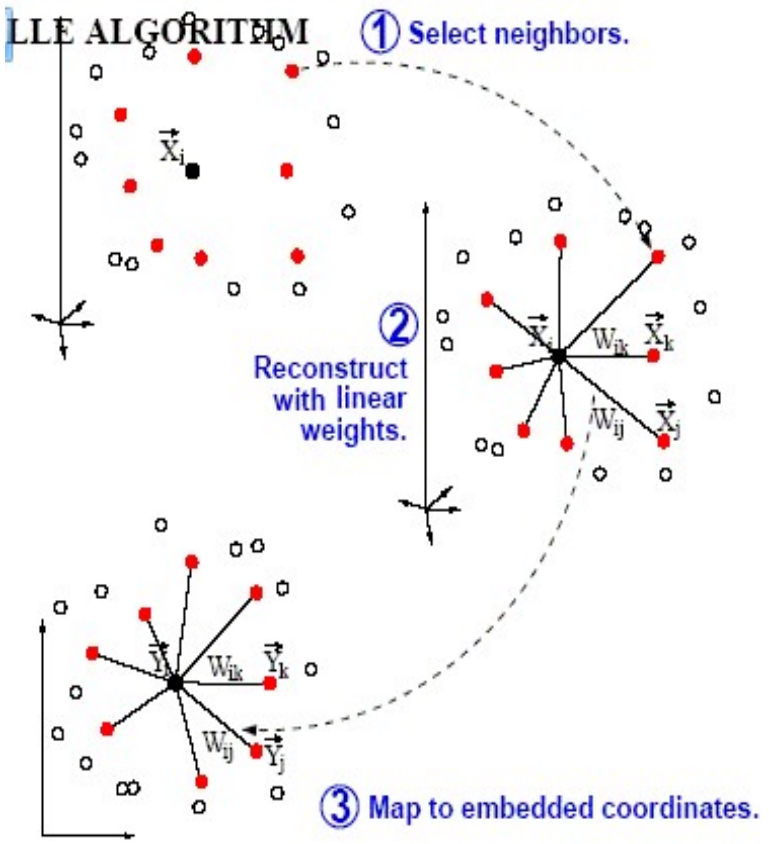
每一个数据点都可以由其近邻点的线性加权组合构 造得到。

- 算法的主要步骤分为三步：

(1)寻找每个样本点的k个近邻点（k是一个预先给定的值）；

(2)由每个样本点的近邻点计算出该样本点的局部重建权值矩阵；

(3)由该样本点的局部重建权值矩阵和其近邻点计算出该样本点的输出值，定义一个误差函数。



PCA应用于恒星光谱分类

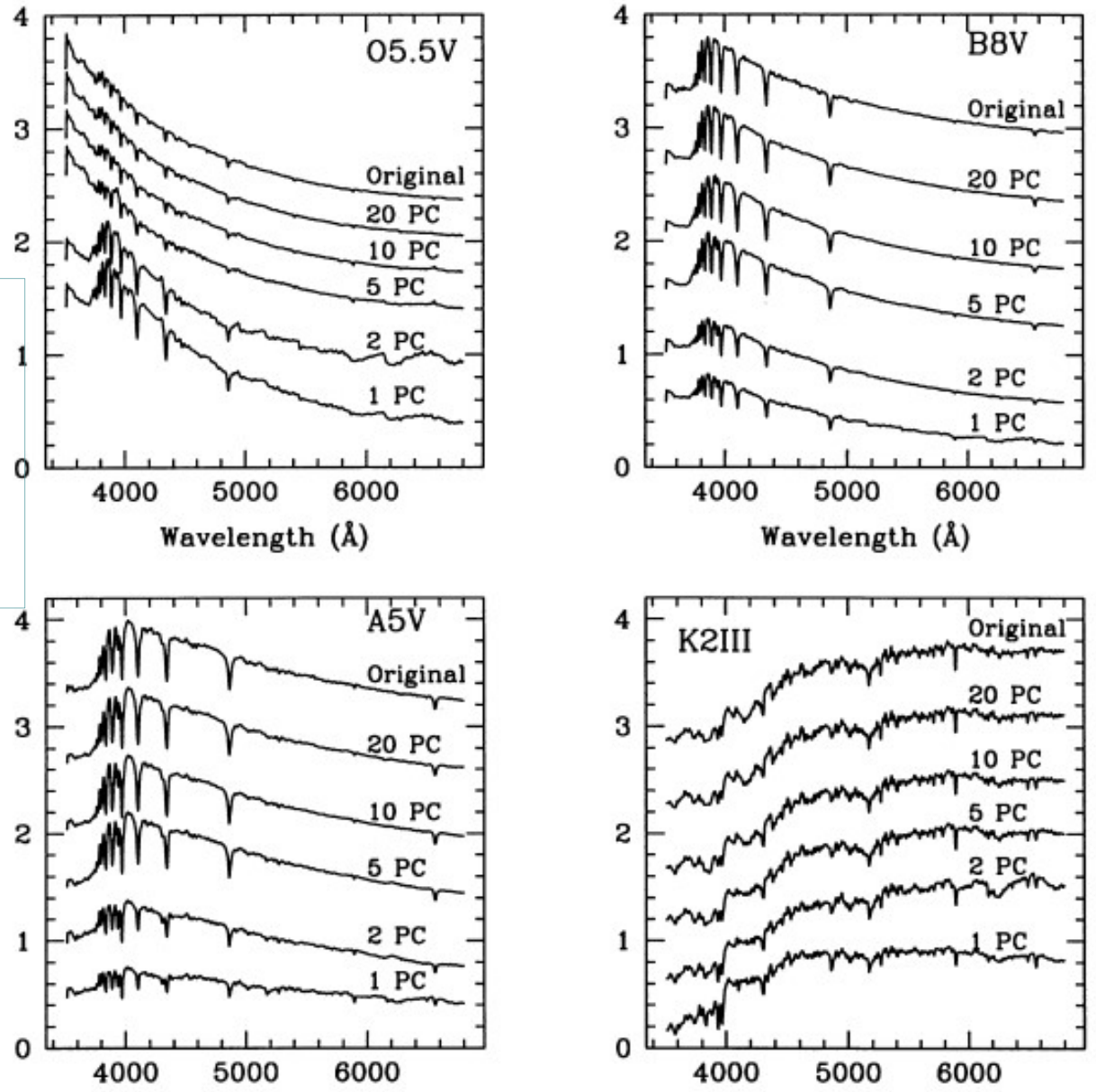
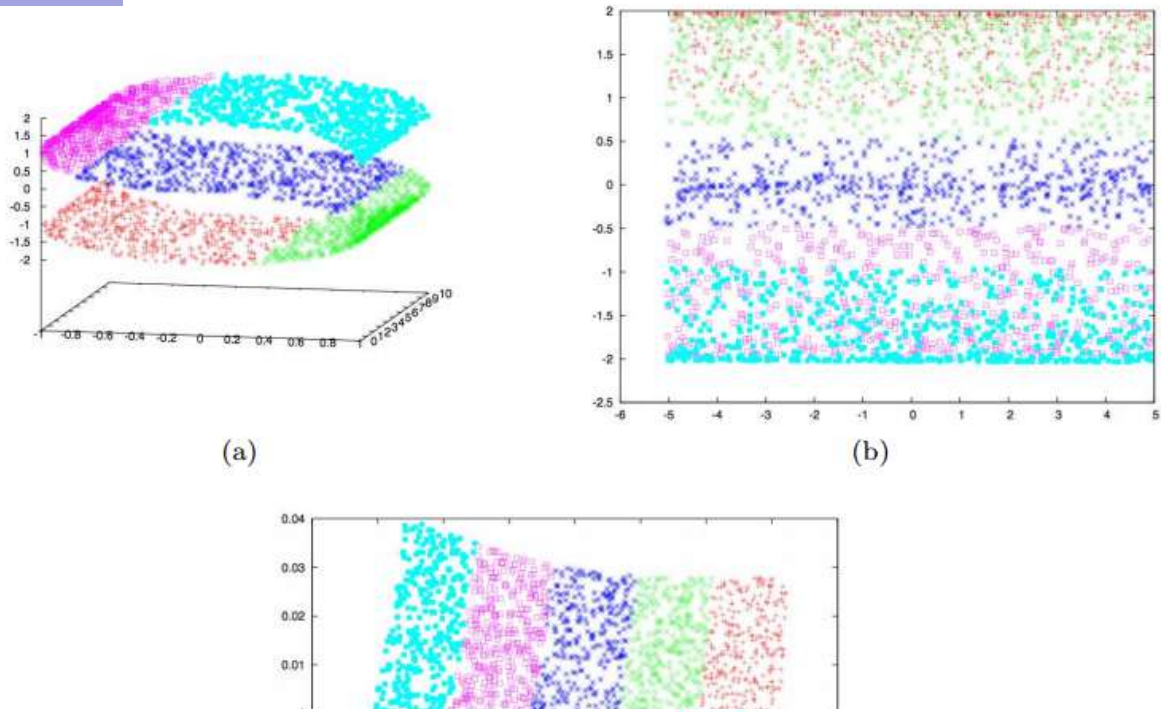


Figure 1. Reconstruction of four different spectral types out of the test spectra using the first 20, 10, five, two and one principal components.

PCA与LLE
应用于恒星光谱分类，LLE优于PCA

(LLE)



(a) (b)

FIG. 1: A demonstration of the advantage of LLE over PCA. Figure 1(a) shows the unprocessed, three-dimensional input data. Figure 1(b) shows the result of a simple, two-dimensional PCA projection of the data. Figure 1(c) shows the result of a two-dimensional LLE projection of the data. Color-coding is consistent between samples. Note that the PCA projection confuses the relationship between points where curvature in Figure 1(a) is strongest. LLE correctly maps the data to its underlying manifold.

LAMOST K巨星：SVM

The Astrophysical Journal, 790:110 (16pp), 2014 August 1 doi:10.1088/0004-637X/790/2/110

THE K GIANT STARS FROM THE LAMOST SURVEY DATA. I. IDENTIFICATION, METALLICITY, AND DISTANCE

Chao Liu[1], Li-Cai Deng[1], Jeffrey L. Carlin[2], Martin C. Smith[3], Jing Li[1,3], Heidi Jo Newberg[2], Shuang Gao[1], Fan Yang[1], Xiang-Xiang Xue[4], Yan Xu[1], Yue-Yang Zhang[1], Yu Xin[1], Yue Wu[1], and Ge Jin[5]
[1] Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Datun Road 20A, Beijing 100012, China; liuchao@nao.cas.cn
[2] Department of Physics, Applied Physics and Astronomy, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180, USA
[3] Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, China
[4] Max Planck Institute for Astronomy, Königstuhl 17, Heidelberg D-69117, Germany
[5] University of Science and Technology of China, Hefei 230026, China
Received 2014 February 19; accepted 2014 June 18; published 2014 July 10

ABSTRACT

We present a support vector machine classifier to identify the K giant stars from the LAMOST survey directly using their spectral line features. The completeness of the identification is about 75% for tests based on LAMOST stellar parameters. The contamination in the identified K giant sample is lower than 2.5%. Applying the classification method to about two million LAMOST spectra observed during the pilot survey and the first year survey, we select

K巨星光谱

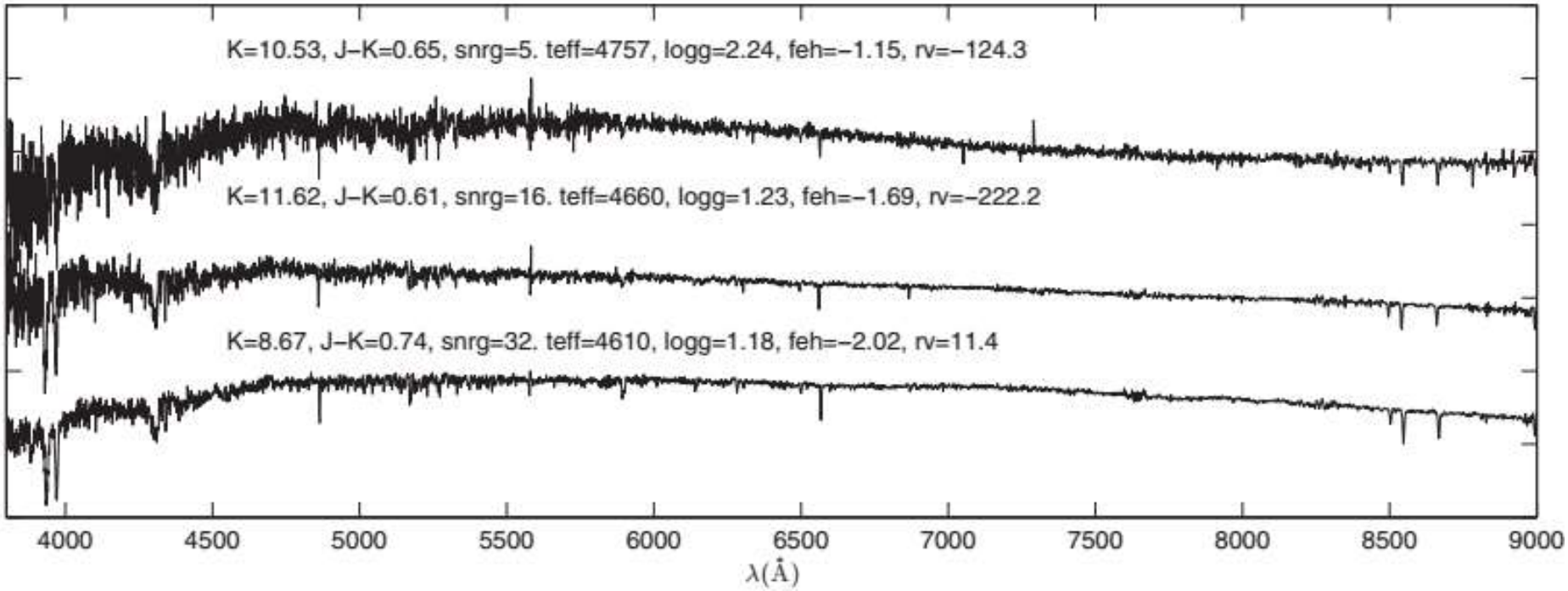

K=10.53, J−K=0.65, snrg=5. teff=4757, logg=2.24, feh=−1.15, rv=−124.3
K=11.62, J−K=0.61, snrg=16. teff=4660, logg=1.23, feh=−1.69, rv=−222.2
K=8.67, J−K=0.74, snrg=32. teff=4610, logg=1.18, feh=−2.02, rv=11.4
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
λ(Å)

Table 1
Line Indexes Definition

Name	Index Bandpass (Å)	Pseudocontinua (Å)	References
H_δ	4083.50–4122.25	4041.60–4079.75	
		4128.50–4161.00	Worthey & Ottaviani (1997)
CN	4143.375–4178.375	4081.375–4118.875	
		4245.375–4285.375	Worthey et al. (1994)
G band	4282.625–4317.625	4267.625–4283.875	
		4320.125–4336.375	Worthey et al. (1994)
H_γ	4319.75–4363.50	4283.50–4319.75	
		4367.25–4419.75	Worthey & Ottaviani (1997)
H_β	4847.875–4876.625	4827.875–4847.875	
		4876.625–4891.625	Worthey et al. (1994)
Mg_1	5069.125–5134.125	4895.125–4957.625	
		5301.125–5366.125	Worthey et al. (1994)
Mg_2	5154.125–5196.625	4895.125–4957.625	
		5301.125–5366.125	Worthey et al. (1994)
Mg_b	5160.125–5192.625	5142.625–5161.375	
		5191.375–5206.375	Worthey et al. (1994)
TiO	6191.375–6273.875	6068.375–6143.375	
		6374.375–6416.875	Worthey et al. (1994)
H_α	6548.00–6578.00	6420.00–6455.00	
		6600.00–6640.00	Cohen et al. (1998)

3. K GIANT SELECTION

3.1. Support Vector Machine Classifier

Support vector machine (SVM) is a machine learning algorithm which is suited for classification (Cortes & Vapnik 1995) and broadly used in astronomy (e.g., Bailer-Jones et al. 2008, 2013; Liu et al. 2012; Saglia et al. 2012). As a supervised algorithm, it needs a set of known data to train the SVM model first. A subset of the training data are selected as the support vectors during the training phase. The support vectors define the linear boundary of classes in a high-dimensional inner product space. When the training process is done, the SVM model is ready for prediction; any data input to the model will be marked as a certain class depending on the region to which the input data are projected.

3.2. Data Preprocessing

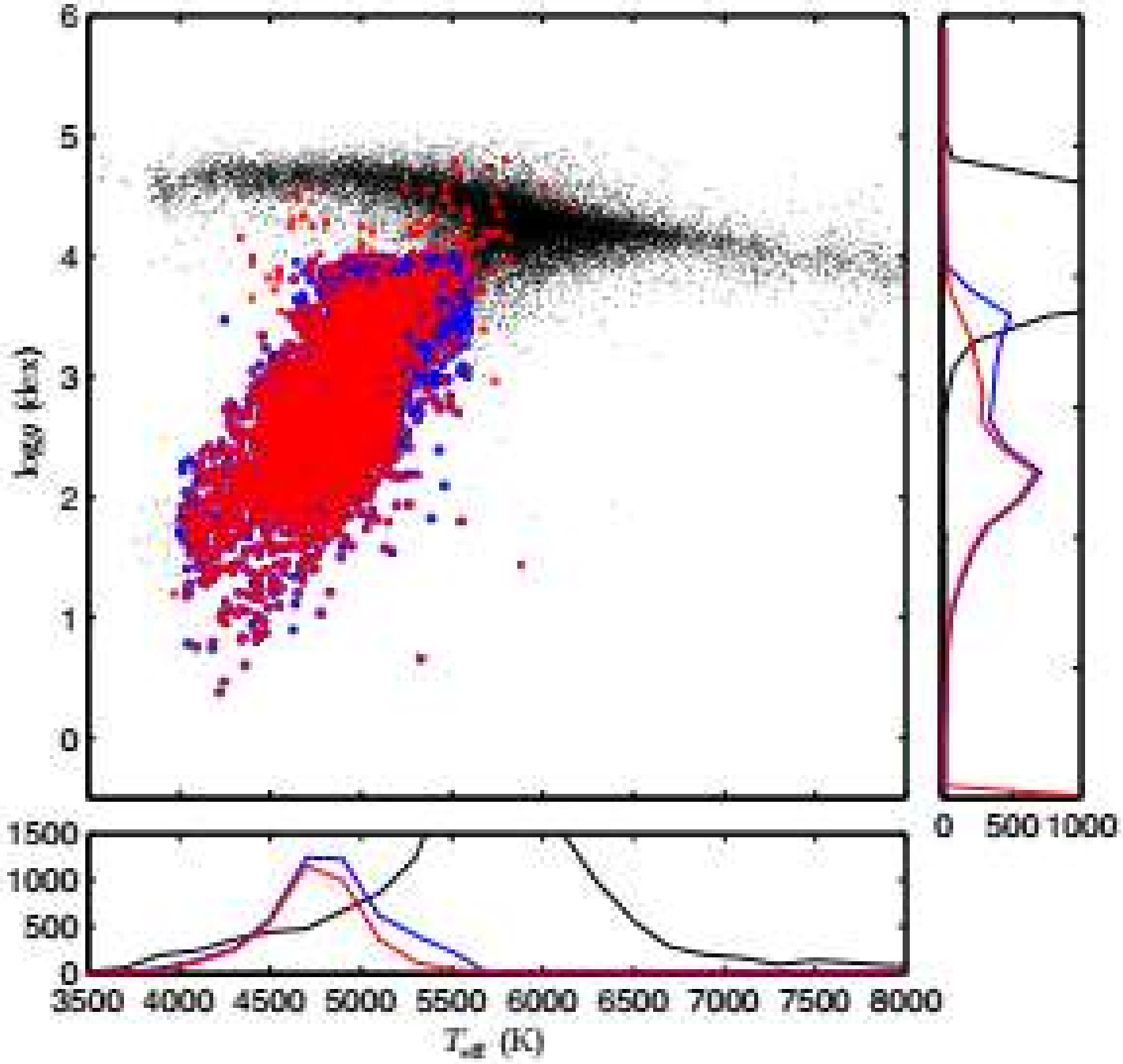
The full spectrum of a star does not carry useful information in every pixel. Some parts of the spectrum that contain strong

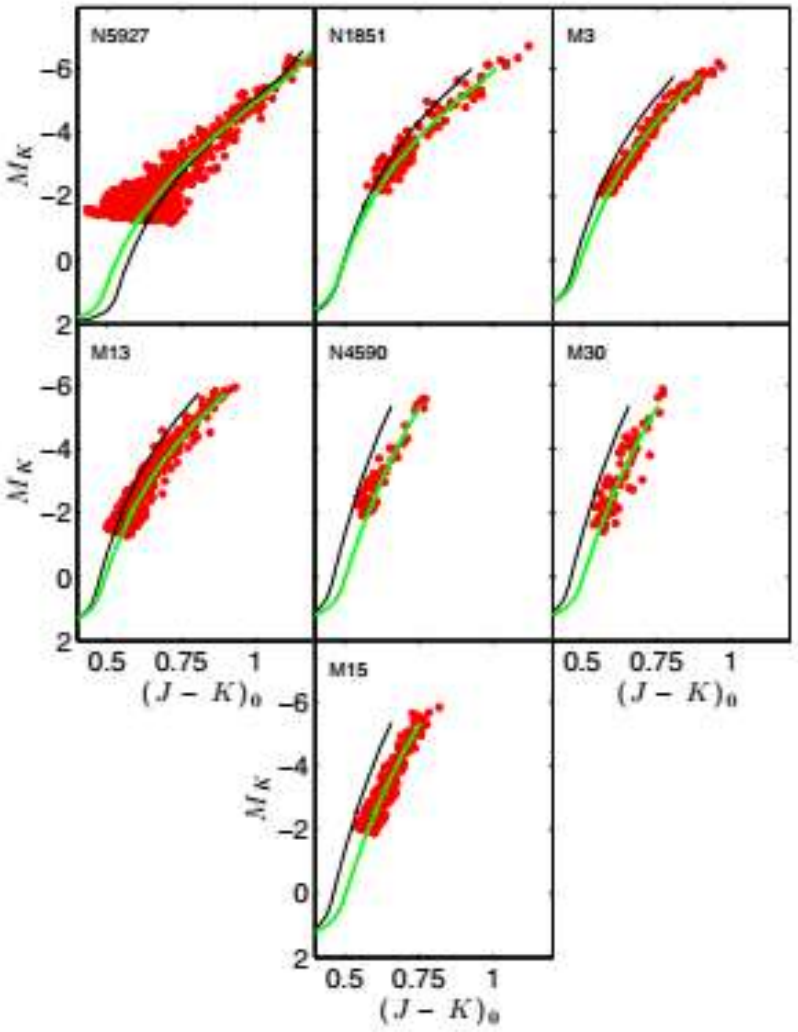
where $f_{\rm cont}(\lambda)$ and $f_{\rm line}(\lambda)$ are the fluxes of the continuum and the spectral line, respectively, both of which are functions of the wavelength λ. The continuum $f_{\rm cont}$ is estimated via linear interpolation of the fluxes located in the pseudocontinuum region on either side of each index bandpass (see Table 1).

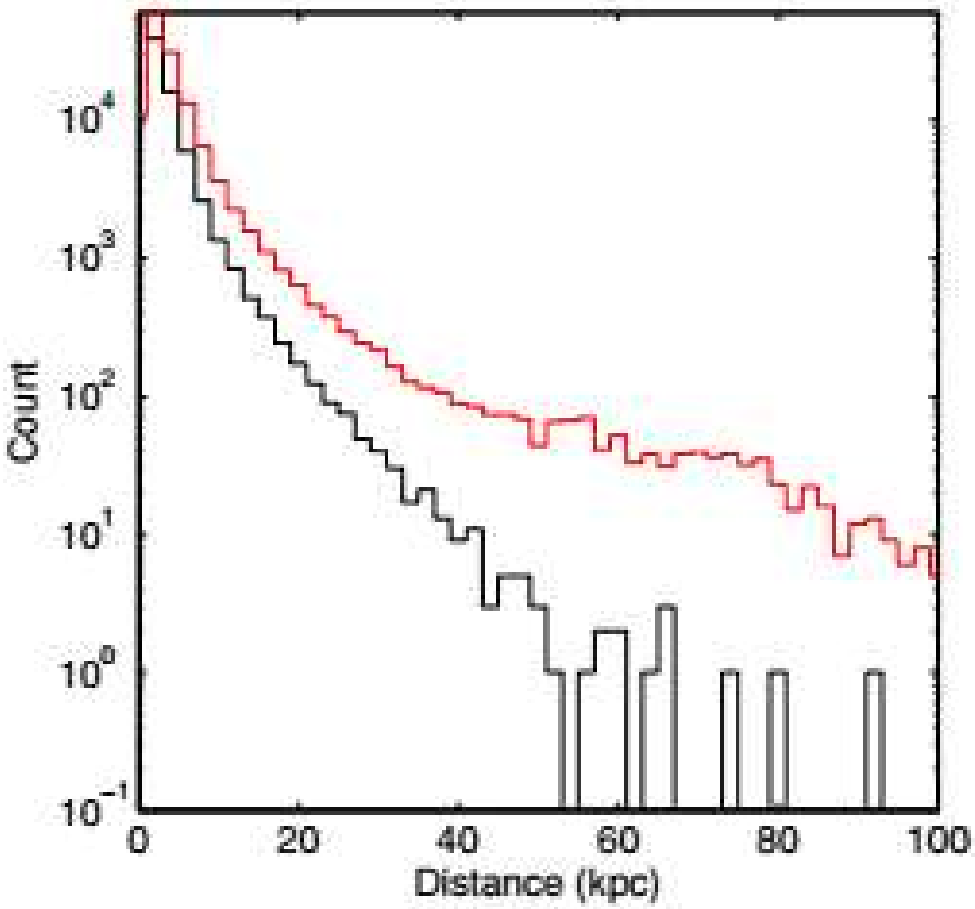
Figure 2 shows the difference between the K giant and non-K giant stars of the MILES library (Sánchez-Blázquez et al. 2006) in the EWs of Mg_b, H_β, and TiO. We can essentially separate the K giant stars from EW_{H_β} versus EW_{Mg_b} with some local overlapping. TiO can help to distinguish the giant from the dwarf stars in some of the overlapped region, particularly at $4 < EW_{Mg_b} < 5$.

3.3. Training of the SVM Classifier

By training with a set of data with known giant/non-giant separation, the parameters within SVM can be properly tuned to obtain the best model for the classification. For this purpose, the training data is defined using a common sample of the LAMOST pilot survey and SDSS DR9 (Ahn et al. 2012) with the following additional criteria: (1) the signal-to-noise ratio for SDSS spectra


6
5
4
3
2
1
0
log g (dex)
0
500
1000
1500
1000
500
0
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
$T_{\rm eff}$ (K)


N5927
N1851
M3
M13
N4590
M30
M15
M_K
$(J-K)_0$
−6
−4
−2
0
2
0.5
0.75
1


Count
Distance (kpc)
10^4
10^3
10^2
10^1
10^0
10^{-1}
0
20
40
60
80
100

LAMOST恒星参数：KPCA方法

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS 464, 3657–3678 (2017) doi:10.1093/mnras/stw2523
Advance Access publication 2016 October 5

Estimating stellar atmospheric parameters, absolute magnitudes and elemental abundances from the LAMOST spectra with Kernel-based principal component analysis

M.-S. Xiang,[1★†] X.-W. Liu,[2,3] J.-R. Shi,[1] H.-B. Yuan,[4] Y. Huang,[2] A.-L. Luo,[1] H.-W. Zhang,[2] Y.-H. Zhao,[1] J.-N. Zhang,[1] J.-J. Ren,[2★] B.-Q. Chen,[2★] C. Wang,[2] J. Li,[5] Z.-Y. Huo,[1] W. Zhang,[1] J.-L. Wang,[1] Y. Zhang,[6] Y.-H. Hou[6] and Y.-F. Wang[6]

[1]*National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, P. R. China*
[2]*Department of Astronomy, Peking University, Beijing 100871, P. R. China*
[3]*Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, P. R. China*
[4]*Department of Astronomy, Beijing Normal University, Beijing 100875, P. R. China*
[5]*Department of Space Science and Astronomy, Hebei Normal University, Shijiazhuang 050024, P. R. China*
[6]*Nanjing Institute of Astronomical Optics & Technology, National Astronomical Observatories, Chinese Academy of Sciences, Nanjing 210042, P. R. China*

3.1 The Kernel-based PCA

A detailed introduction of the Kernel-based PCA can be found in Schölpokf et al. (1998) and Müller et al. (2001). Below, we briefly summarize the algorithm for completeness. Let x_k, $k = 1, \ldots, M$, denote the spectra of m stars, each contains N wavelength pixels. $x_k \in \mathrm{R}^N$. The spectra are normalized such that the sum of all pixel squared values of a given spectrum is unity. Note that the normalization is critical to generate realistic values of the kernel function. To extract data structures with KPCA, we map the spectra into feature space F by a (non-linear) function $\Phi(x_k)$. Then we have the covariance matrix in F,

$$\mathbf{C} = \frac{1}{M} \sum_{i=1}^{M} \Phi(x_i)\Phi(x_i)^{\mathrm{T}}. \tag{1}$$

To calculate the principal components, we solve the Eigenvalue problem below to find Eigenvalue $\lambda > 0$ and Eigenvector $V \neq 0$:

$$\lambda V = \mathbf{C}V = \frac{1}{M} \sum_{i=1}^{M} (\Phi(x_i) \cdot V)\Phi(x_i). \tag{2}$$

All Eigenvectors with non-zero Eigenvalue can be written in the span of $\Phi(x_1),\ldots,\Phi(x_M)$ such that,

$$V = \sum_{j=1}^{M} \alpha_j \Phi(x_j). \tag{3}$$

Multiplying equation (2) by $\Phi(x_k)$ from the left yields

$$\lambda \sum_{j=1}^{M} \alpha_j (\Phi(x_k) \cdot \Phi(x_j))$$

$$= \frac{1}{M} \sum_{j=1}^{M} \alpha_j \sum_{i=1}^{M} (\Phi(x_k) \cdot \Phi(x_i))(\Phi(x_i) \cdot \Phi(x_j)). \tag{4}$$

Defining an $M \times M$ matrix $\mathbf{K}$,

$$K_{ij} := (\Phi(x_i) \cdot \Phi(x_j)), \tag{5}$$

the Eigenvalue problem becomes

$$M\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}. \tag{6}$$

Even in the most realistic cases, the non-linear transformation Φ in general cannot be expressed explicitly. Therefore, instead of calculating the products $(\Phi(x) \cdot \Phi(y))$ in equation (6) directly, we use a kernel representation of the form,

$$k(x, y) = (\Phi(x) \cdot \Phi(y)). \tag{9}$$

Various forms of kernel function, such as polynomial, radial basis functions and sigmoidal, as well as other more complicated kernels, have been validated (Schölpokf et al. 1998; Müller et al. 2001). In this work, we use the Gaussian radial basis functions,

$$k(x, y) = \exp\left(\frac{-\|x - y\|^2}{c}\right), \tag{10}$$

where $\|\cdot\|$ represents the Euclidean norm, $\|x - y\| \equiv \sqrt{(x - y) \cdot (x - y)}$, and c is the width of the kernel. Throughout this paper, we adopt $c = 0.005$, a typical value of the squared Euclidean norm $\|x - y\|^2$ for the LAMOST spectra. In fact, we have examined different values of c (e.g. 0.005, 0.05, 0.5, 1.0, 5.0) making use of both LAMOST–*Kepler* sample stars and member stars of open clusters, and found that 0.005 is an optimal one.

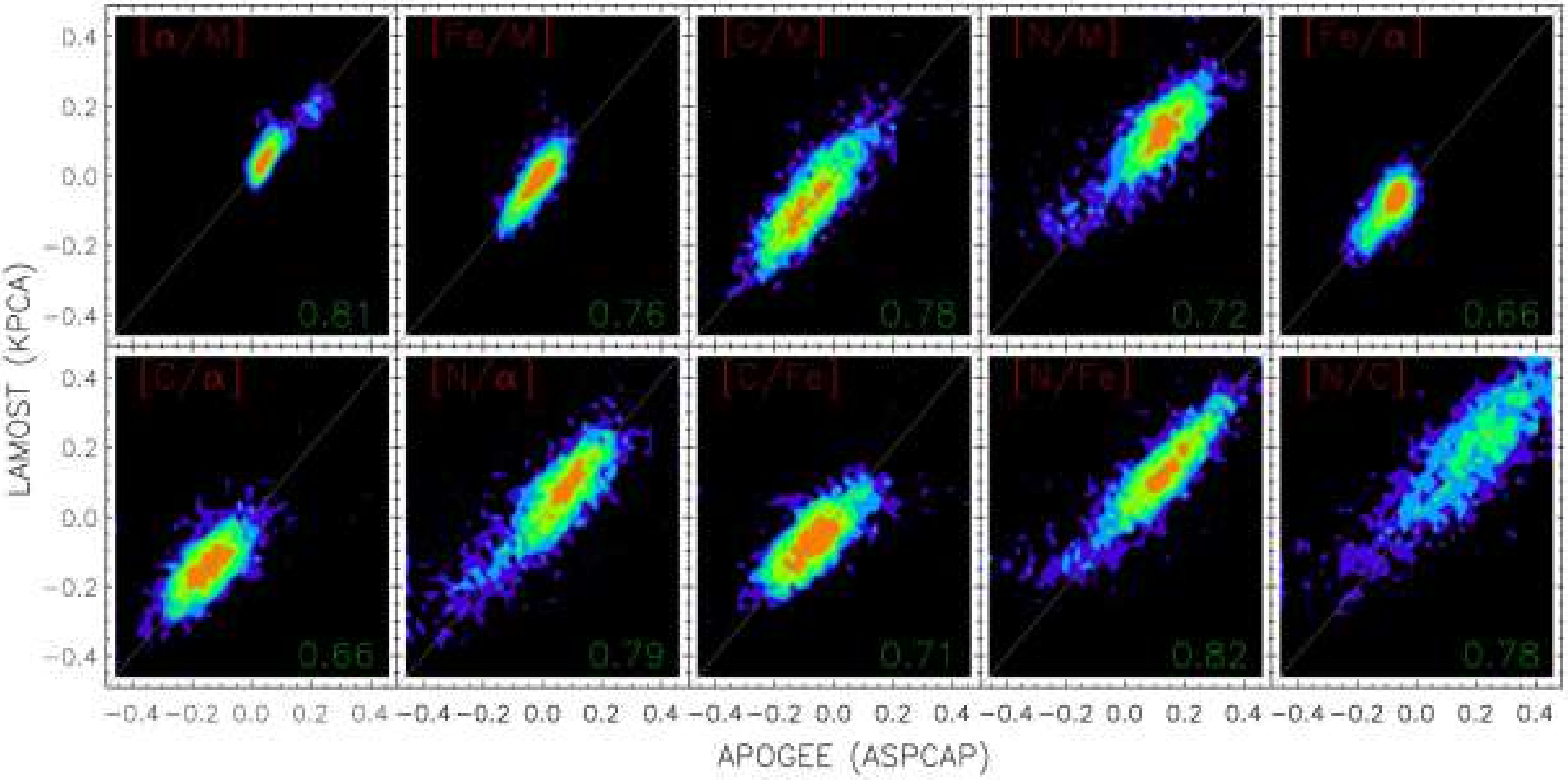
3.2 The regression

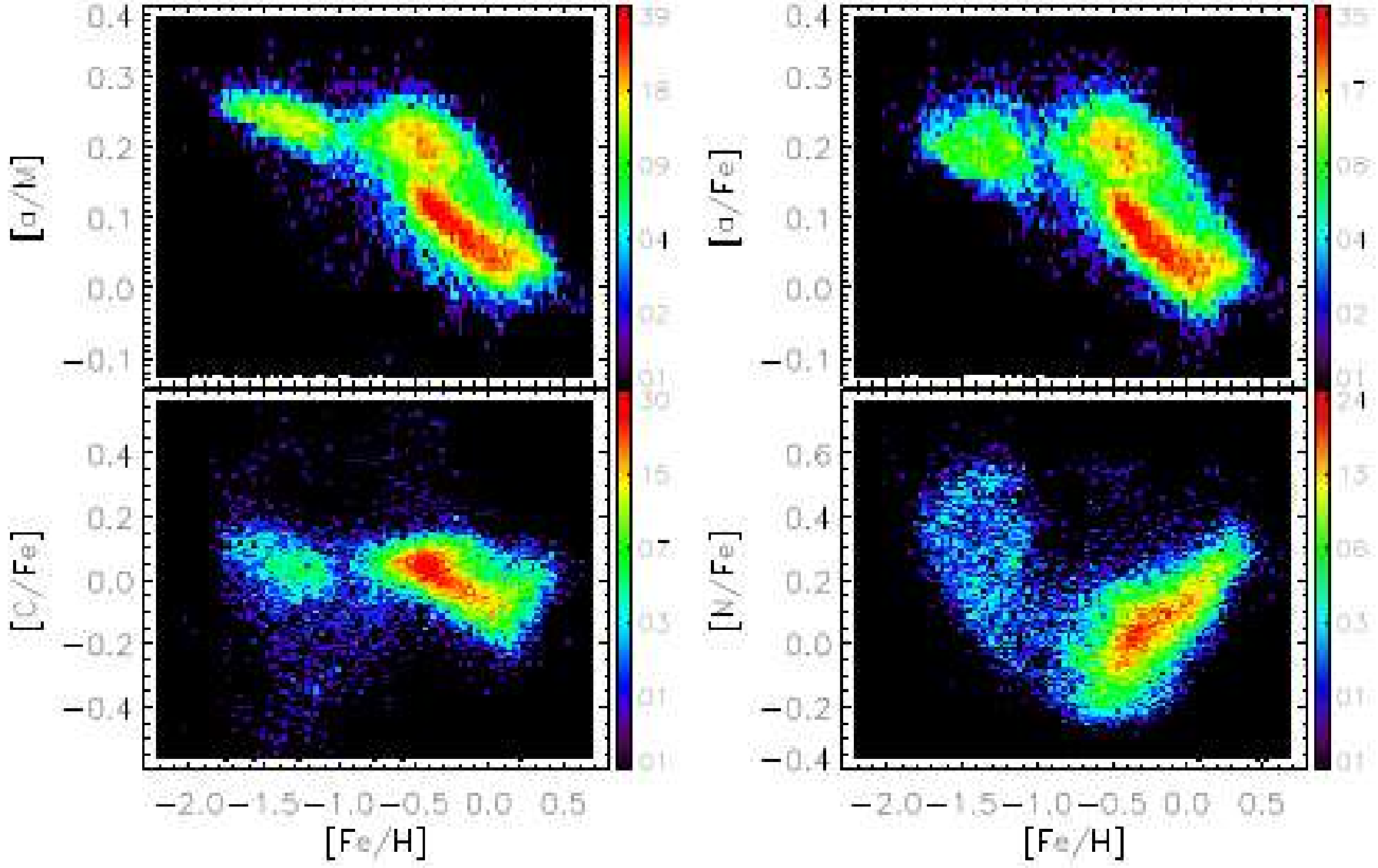
To derive atmospheric parameters from the principal components, we construct a multiple-linear relation between the principal components $\boldsymbol{P}$ and the stellar atmospheric parameters for each parameter y,

$$y = \sum_{i=1}^{N} c_i \boldsymbol{P}_i + c_0, \tag{11}$$

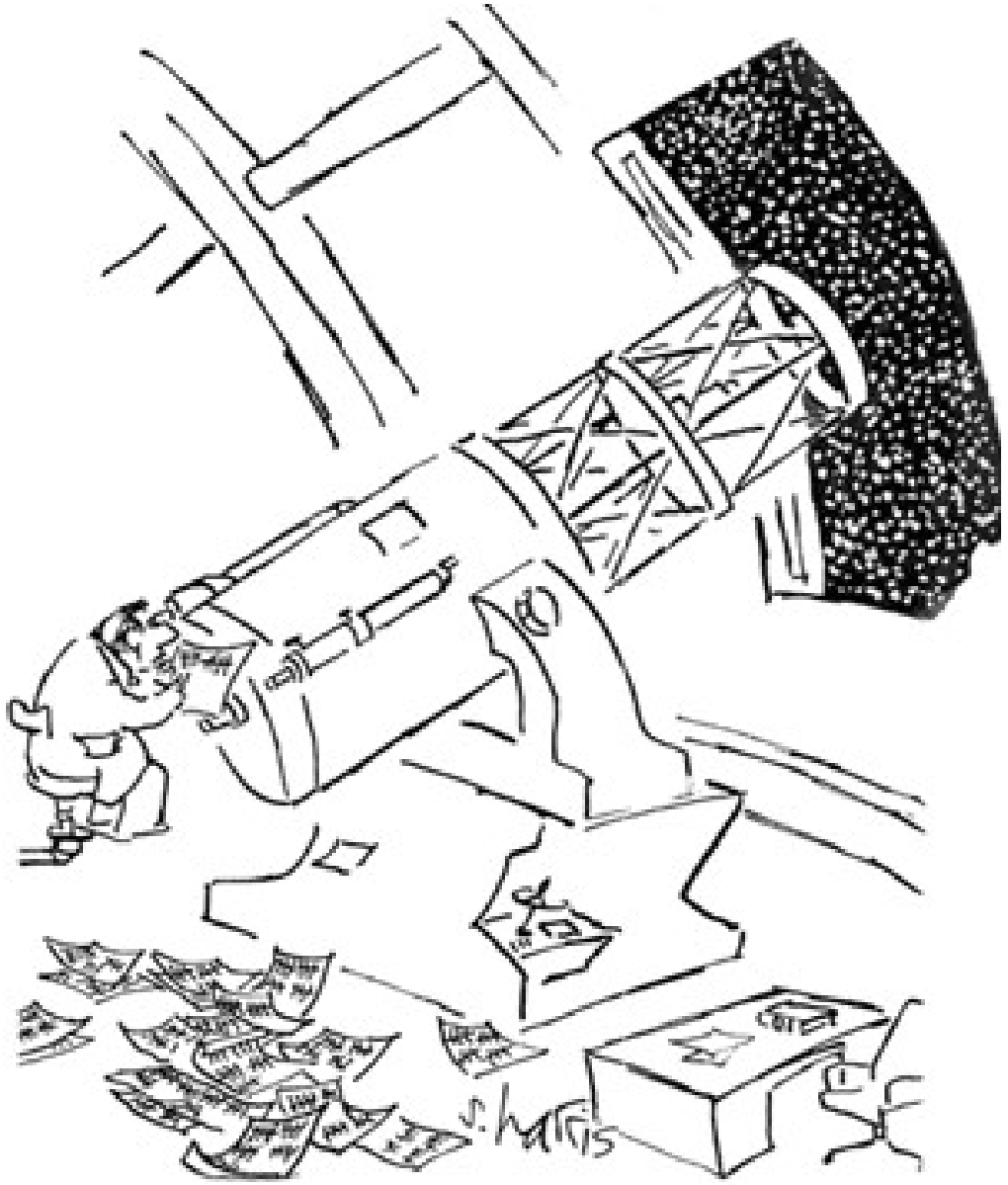
where N is the adopted number of principal components, determined empirically with a brute-force search (cf. Section 4), c_0 is a constant, and y is any one of the parameters $T_{\rm eff}$, log g, [Fe/H], M_V, M_{K_s}, [M/H], [α/M], [α/Fe], [C/H] and [N/H]. The coefficients c_i, $i = 1, \ldots, N$, are determined by a least square multiple-linear fit to a training data set.

When estimating log g values for giant stars using the LAMOST–*Kepler* sample stars as the training set (cf. Section 4.3), $T_{\rm eff}$ and [Fe/H] from LSP3 are adopted as priors. This is carried out by taking log $T_{\rm eff}$ and [Fe/H] as input pixel values of spectral flux. Similarly, when estimating log g values for dwarfs using the MILES library as the training set (cf. Section 4.1), as well as when estimating [M/H], [α/M] and [α/Fe] using the LAMOST–APOGEE stars as the training set (cf. Section 4.4), $T_{\rm eff}$ yielded by LSP3 is adopted as a prior. For the estimation of individual elemental abundances [Fe/H], [C/H] and [N/H] with the LAMOST–APOGEE training set, the LSP3 $T_{\rm eff}$ as well as the KPCA estimated [M/H] are adopted


[α/M]
0.81
[Fe/M]
0.76
[C/M]
0.78
[N/M]
0.72
[Fe/α]
0.66
[C/α]
0.66
[N/α]
0.79
[C/Fe]
0.71
[N/Fe]
0.82
[N/C]
0.78
0.4
0.2
0.0
−0.2
−0.4
LAMOST (KPCA)
−0.4 −0.2 0.0 0.2 0.4
APOGEE (ASPCAP)


[α/M]
[C/Fe]
[α/Fe]
[N/Fe]
[Fe/H]
[Fe/H]

天文学：数据驱动的科学



天文学：是发现驱动的科学

- 驱动发现的因素:

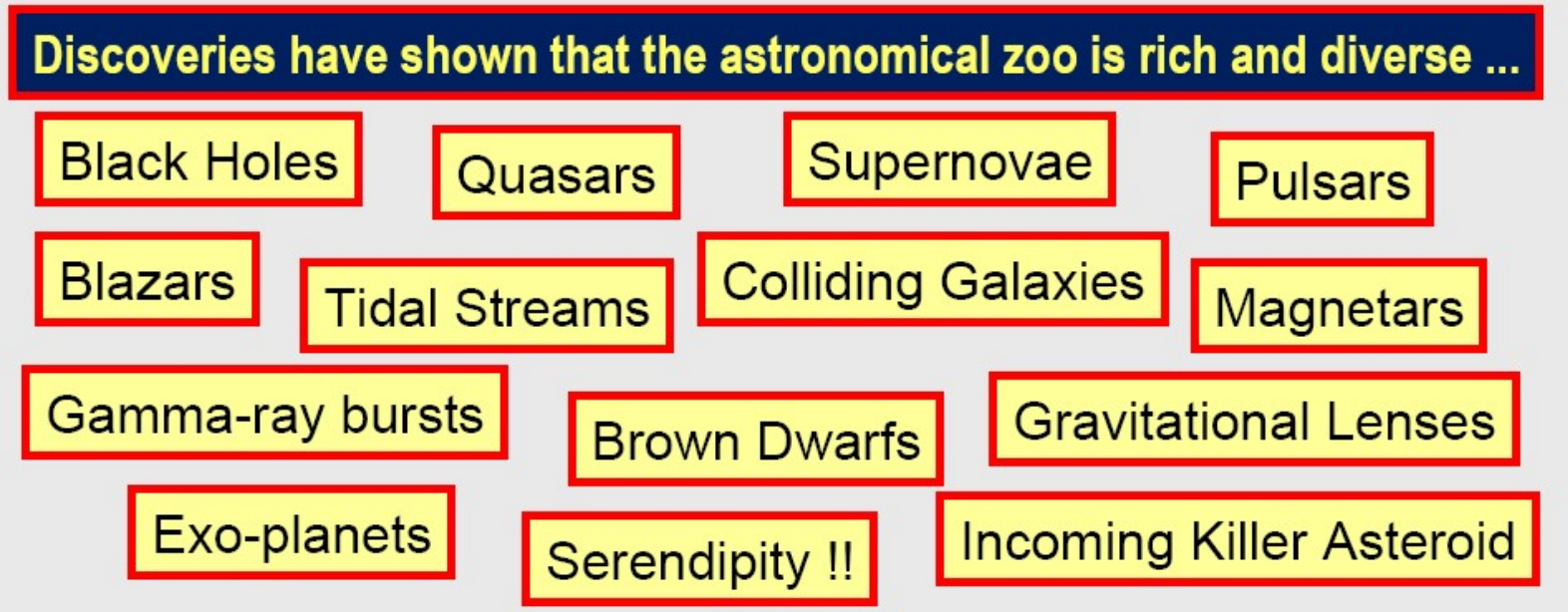
– – 新问题

– – 新的思想

– – 新模型

– – 新理论

– – 更重要的是新数据!



天文学：是发现驱动的科学

- 发现导致:
  - 新的问题
  - 新思想
  - 新模型
  - 新理论
  - 更重要的是…更多的新数据!

- 因此，需要更有效的挖掘和分析算法或工具

大型巡天导致天文学步入一个新的时代

- 大多数数据大的人们无法看
- **这就需要存储技术、网络技术、数据库相关技术和标准等**
- 许多知识被数据的复杂性所掩盖而难以获得
- **大多(不是所有的)经验关系是建立在3维参数空间基础上的，如椭圆和核球星系的基平面。宇宙就是这么简单还是人类认知的偏见？**
- 大部分数据人们是无法直接理解的
- **这就需要数据挖掘、知识发现、数据理解技术、超高维可视化、人工智能/机器帮助的发现**

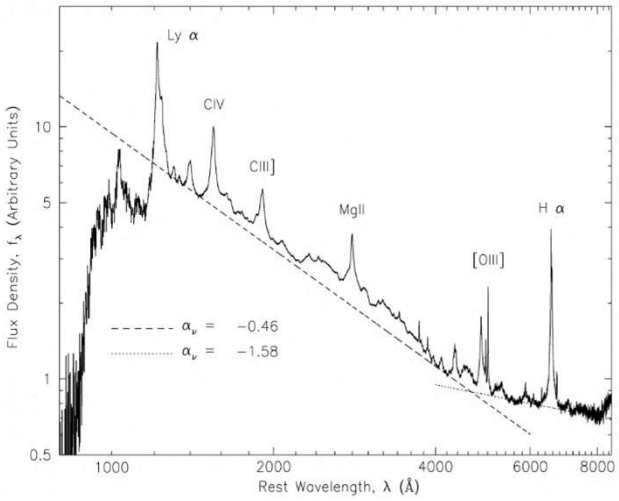
数据挖掘是帮助和加速科学发现过程的利器

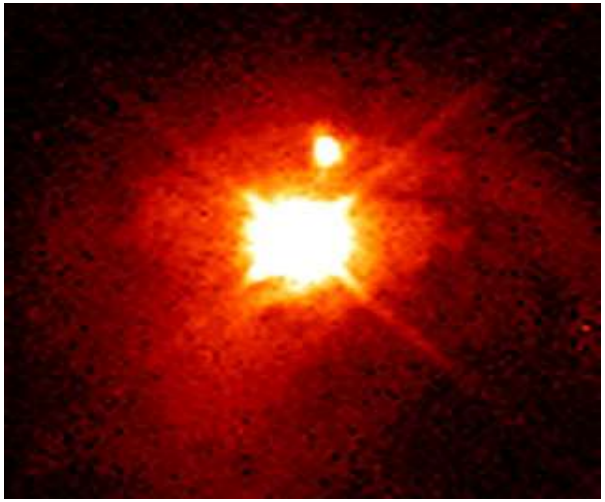
天文数据的特点

- 空间性
- 多波段性
- 海量性
- 非线性
- 异构性
- 缺值性或坏标记
- 分布性
- 高维性
- 时序性
- 开放性

天文数据的常用类型

- 光谱数据
- 图像数据
- 星表数据
- 时序数据
- 模拟数据

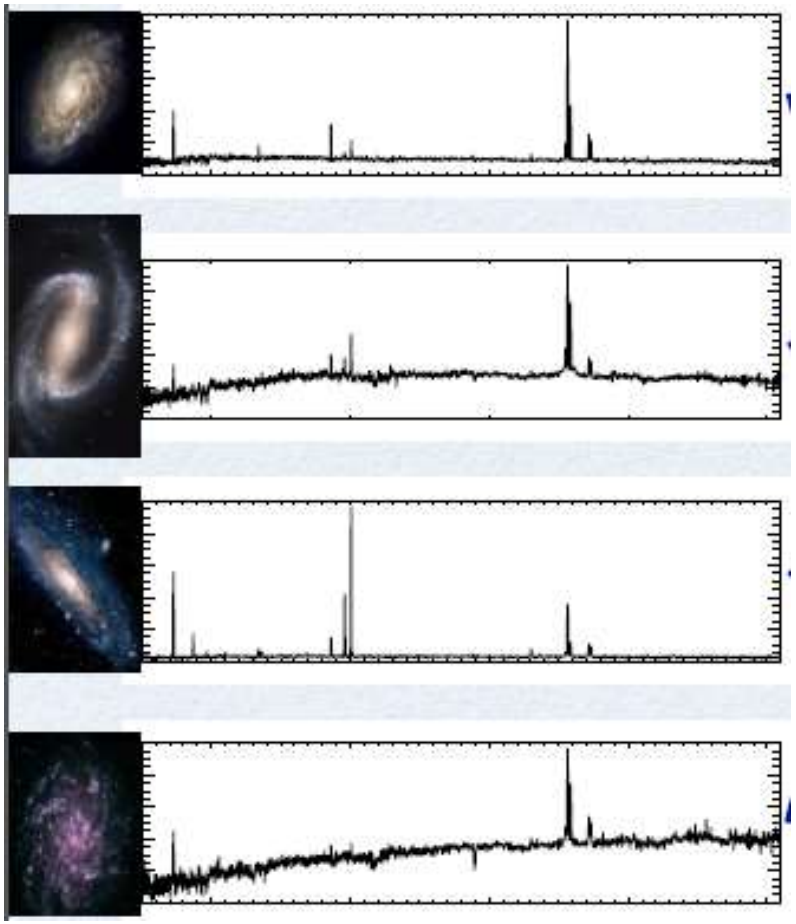




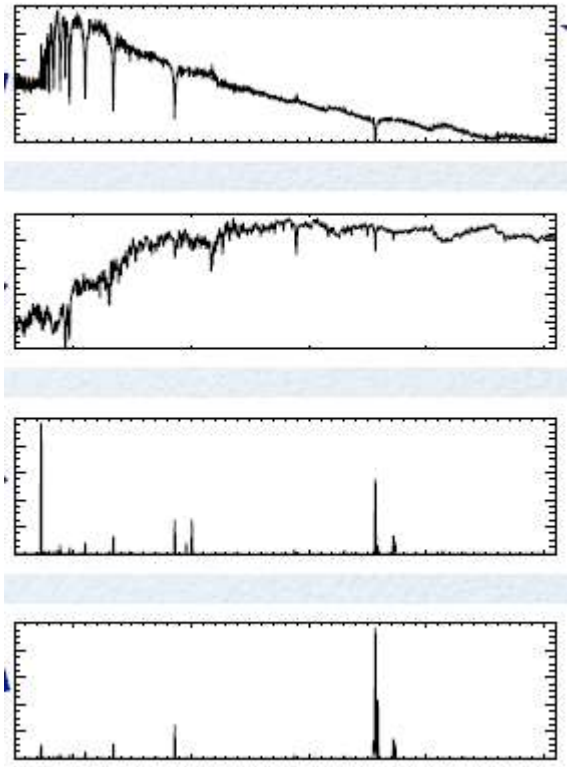
VIII/90/first12 The FIRST Survey Catalog, Version 12Feb16 (Becker+ 2012)
Post annotation The FIRST survey catalog, 12Feb16 Version (946464 rows)

Full	r	RAJ2000	DEJ2000	FIRST	FITS	RAJ2000	DEJ2000
	arcmin	"h:m:s"	"d:m:s"			"h:m:s"	"d:m:s"
1	0.6709	23 59 57.508	-00 00 14.94	J235957.5-000014	FITS	23 59 57.508	-00 00 14.94

星系的图像和光谱、恒星和气体的光谱



星系



早型星

晚型星

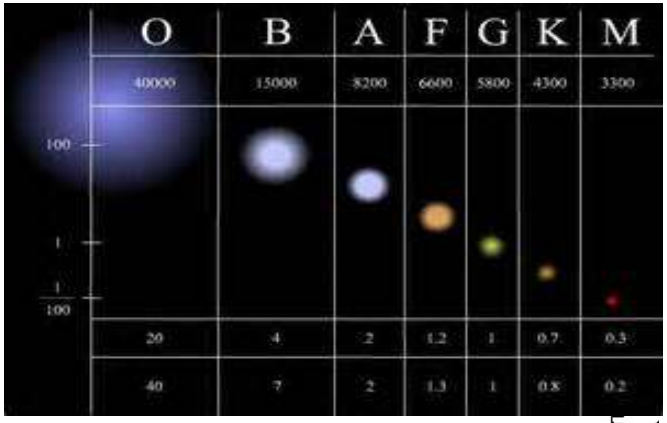
气体

气体

天文中的数据挖掘课题

Tagliaferri et al. 2003	Ball & Brunner 2009
S/G separation	S/G separation
Morphological classification of galaxies *(shapes, spectra)*	Morphological classification of galaxies *(shapes, spectra)*
Spectral classification of stars	Spectral classification of stars
Image segmentation	-----
Noise removal *(grav. waves, pixel lensing, images)*	-----
Photometric redshifts *(galaxies)*	Photometric redshifts *(galaxies, QSO's)*
Search for AGN	Search for AGN and QSO
Variable objects	Time domain
Partition of photometric parameter space for specific group of objects	Partition of photometric parameter space for specific group of objects
Planetary studies (asteroids)	Planetary studies (asteroids)
Solar activity	Solar activity
Interstellar magnetic fields	----
Stellar evolution models	

O B A F G K M L T 天文的应用：恒星光谱分类



恒星光谱序列（温度序列）：

O B A F G K M L T

矮星光谱分类

Normalized Flux (F_λ) + Constan

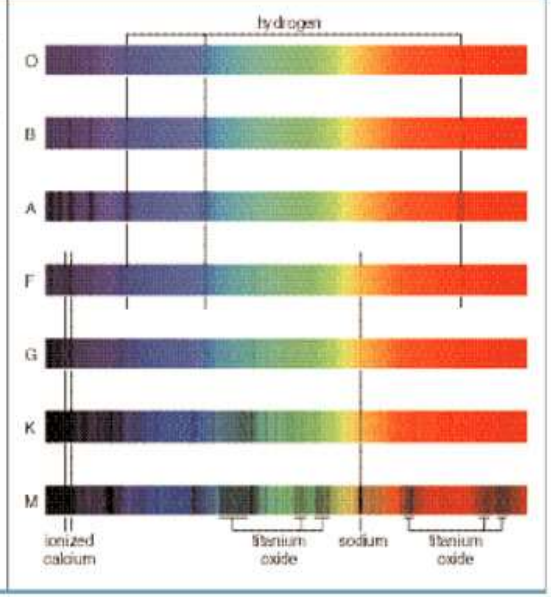
1

1

1

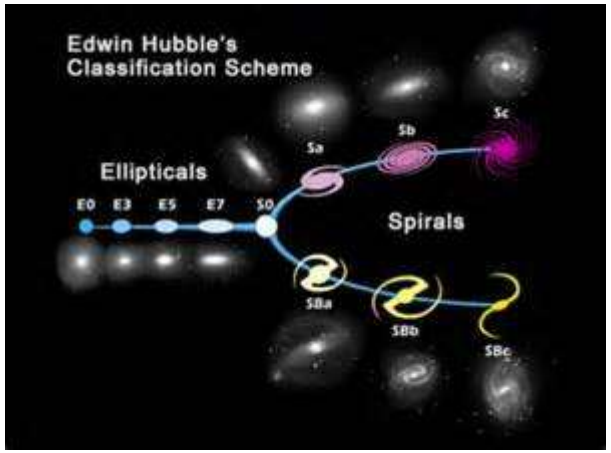
Table 16.1 The Spectral Sequence

Spectral Type	Example(s)	Temperature Range	Key Absorption Line Features	Brightest Wavelength (color)	Typical Spectrum
O	Stars of Orion's Belt	>30,000 K	Lines of ionized helium, weak hydrogen lines	<97 nm (ultraviolet)*	
B	Rigel	30,000 K–10,000 K	Lines of neutral helium, moderate hydrogen lines	97–290 nm (ultraviolet)*	
A	Sirius	10,000 K–7,500 K	Very strong hydrogen lines	290–390 nm (violet)*	
F	Polaris	7,500 K–6,000 K	Moderate hydrogen lines, moderate lines of ionized calcium	390–480 nm (blue)*	
G	Sun, Alpha Centauri A	6,000 K–5,000 K	Weak hydrogen lines, strong lines of ionized calcium	480–580 nm (yellow)	
K	Arcturus	5,000 K–3,500 K	Lines of neutral and singly ionized metals, some molecules	580–830 nm (red)	
M	Betelgeuse, Proxima Centauri	<3,500 K	Molecular lines strong	>830 nm (infrared)	



*All stars above 6,000 K look more or less white to the human eye because they emit plenty of radiation at all visible wavelengths.

天文的应用：星系形态分类




elliptical galaxy

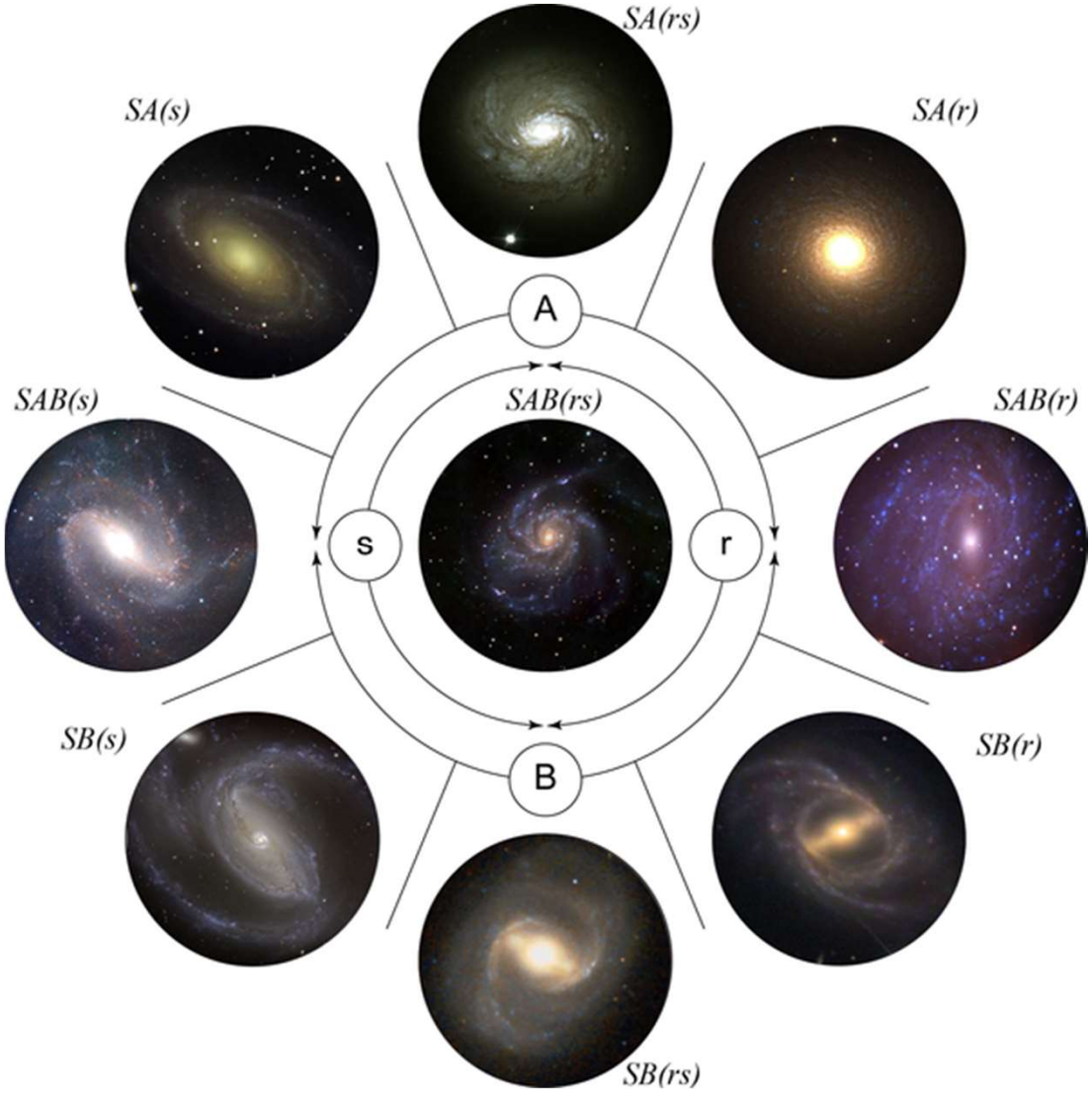

lenticular galaxy

normal spiral galaxy

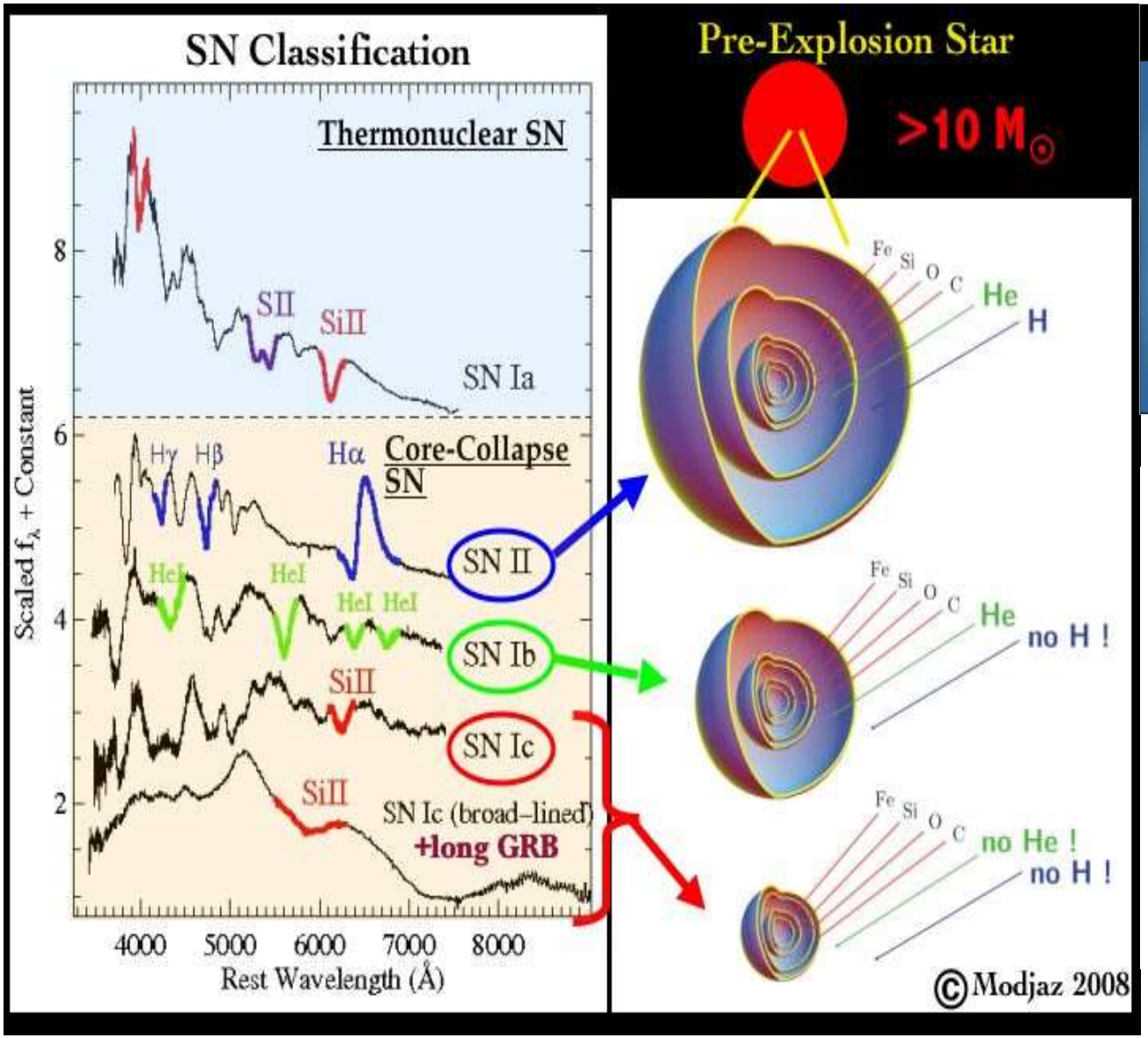

type I irregular galaxy

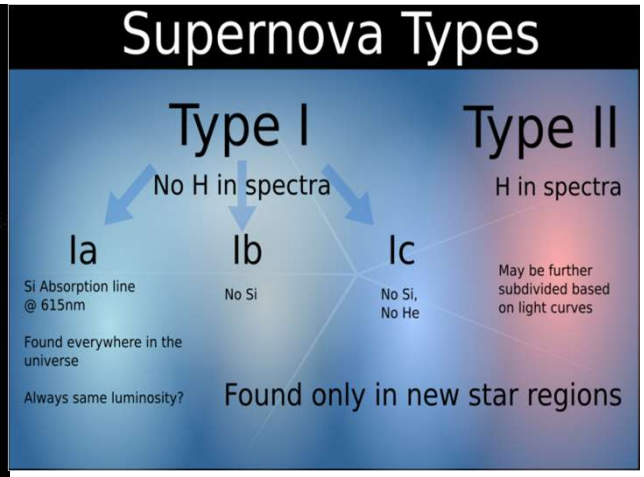

barred spiral galaxy

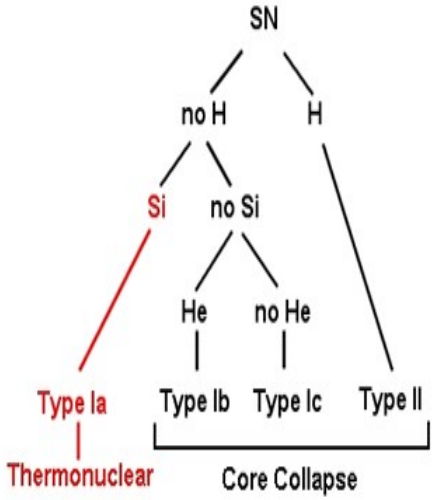

type II irregular galaxy



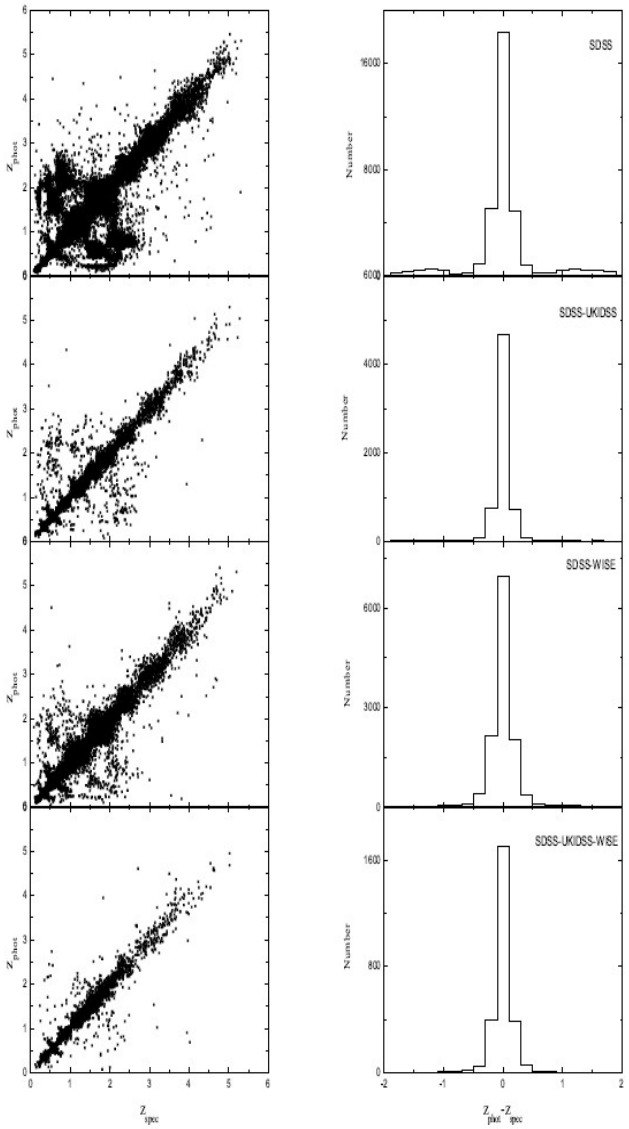
天文的应用：超新星分类


SN Classification
Thermonuclear SN
SII
SiII
SN Ia
Core-Collapse SN
Hγ
Hβ
Hα
HeI
HeI
HeI HeI
SN II
SN Ib
SiII
SN Ic
SiII
SN Ic (broad-lined)
+long GRB
Scaled f_λ + Constant
8
6
4
2
4000
5000
6000
7000
8000
Rest Wavelength (Å)
Pre-Explosion Star
>10 $M_\odot$
Fe
Si
O
C
He
H
no H !
no He !
no H !
©Modjaz 2008

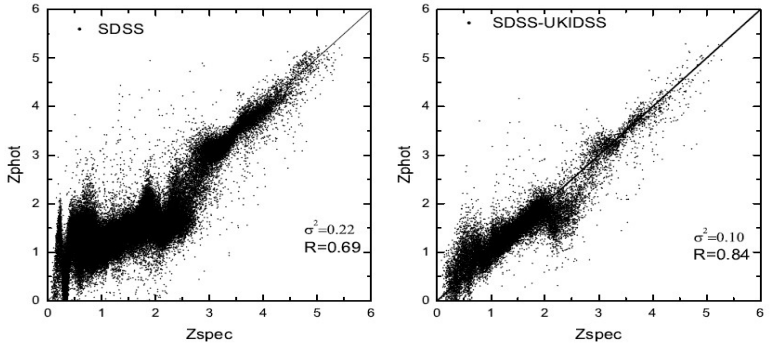

Supernova Types
Type I
No H in spectra
Type II
H in spectra
Ia
Si Absorption line @ 615nm
Found everywhere in the universe
Always same luminosity?
Ib
No Si
Ic
No Si, No He
May be further subdivided based on light curves
Found only in new star regions


SN
no H
H
Si
no Si
He
no He
Type Ia
Type Ib
Type Ic
Type II
Thermonuclear
Core Collapse

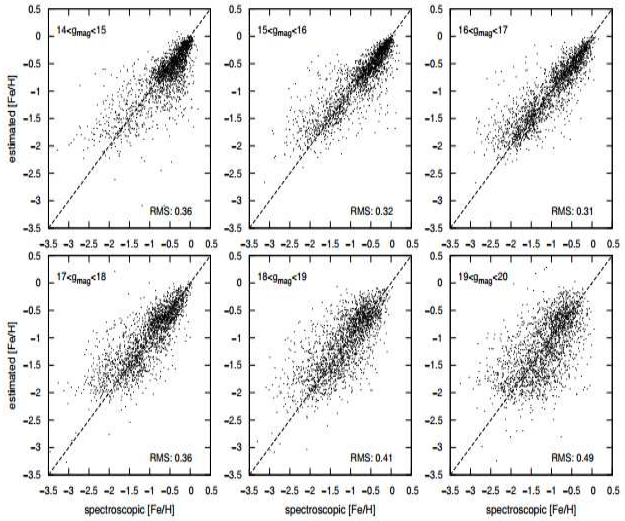
天文的应用：测光红移预测



- 基于多波段数据，应用了K近邻方法预测类星体的测光红移预测，发现随着波段的增多，红移预测精度有所增加。



天文的应用：恒星参数估计



Decin *et al* (2004)

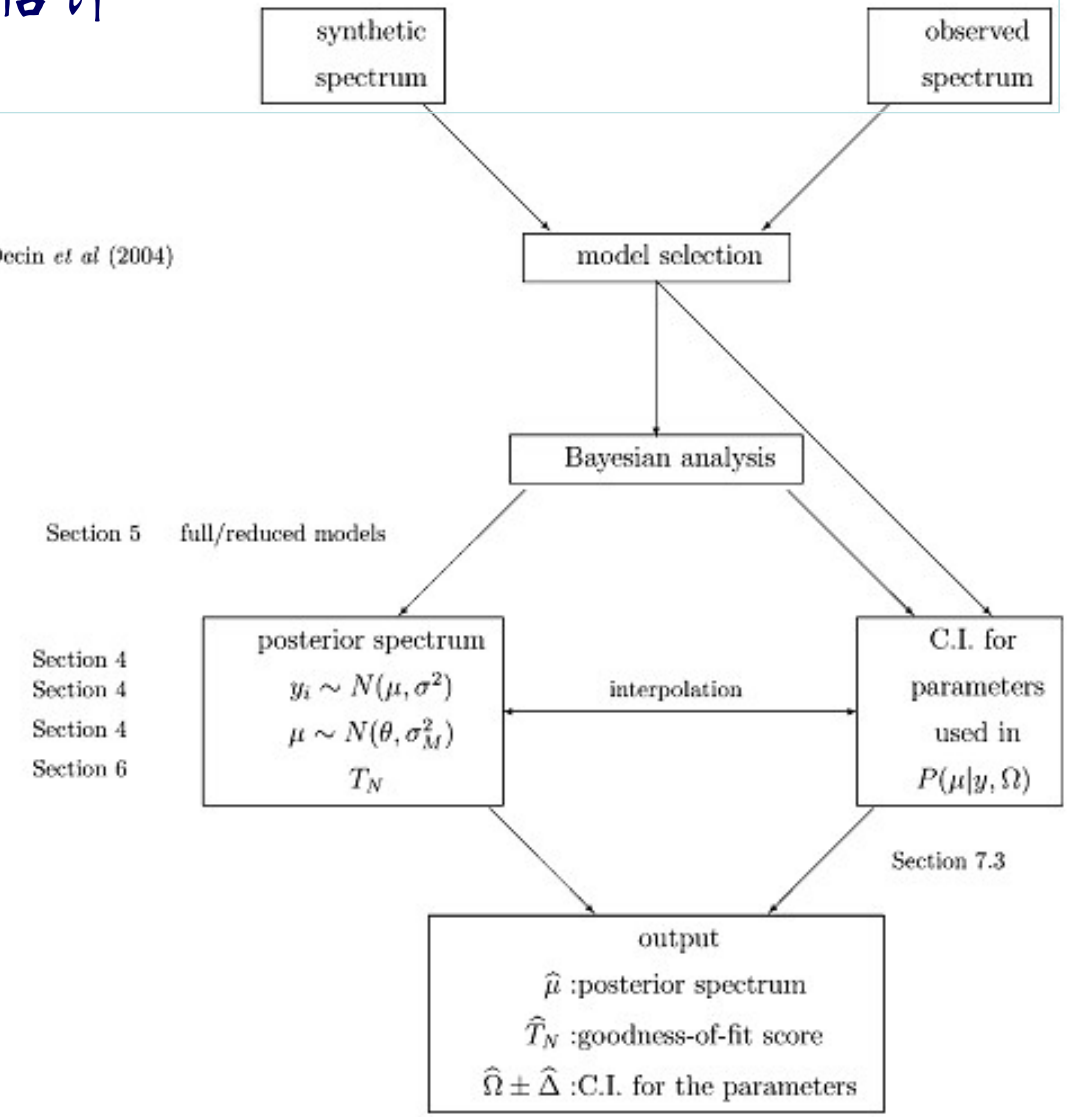


Figure 2. Schematic representation of the model building and selection steps' sequencing.

天文应用：聚类问题

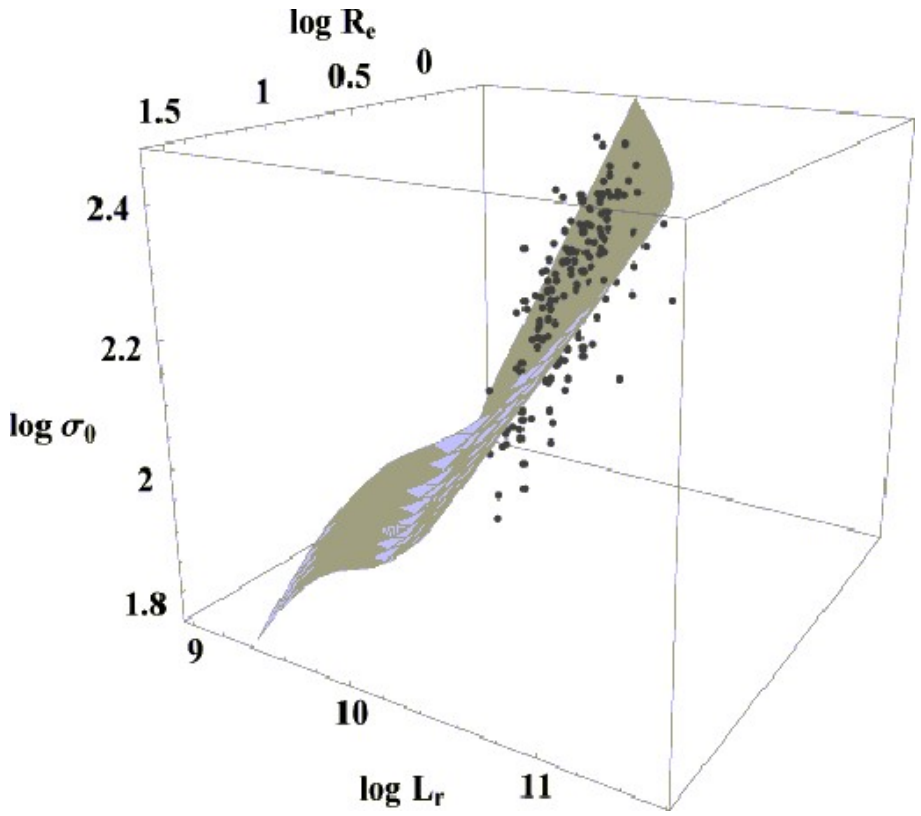
- 聚类问题：
  - 在数据集中查找聚类的天体
  - 统计意义和科学意义上各个类别的重要性是什么？
  - 找“朋友的朋友”或近邻的最优算法？
    - $N > 10^{10}$，如何有效地排序、分类？
    - 维数～1000 – 因此，若干子空间搜索问题
  - 是否存在两点或更高阶的相关性？
    - $N > 10^{10}$，N-point 相关怎么做？
      - 与$N^2 \log N$成正比的算法显然不能用

天文应用：离群探测

- 离群探测：(未知的未知)
  - 找到那些超出我们预期的天体或事件 (不属于已知类别)
  - 这些有可能是真正的科学发现或垃圾
  - 因此，离群探测可用于：
    - 新奇发现 –*Nobel prize?*
    - 异常探测 – *探测系统是否正常工作?*
    - 数据质量保证 – *数据流是否正常工作?*
  - 在1000维空间中或感兴趣的子空间（低维空间）中，如何最优化地探测到离群?
  - 怎样衡量“兴趣度”？

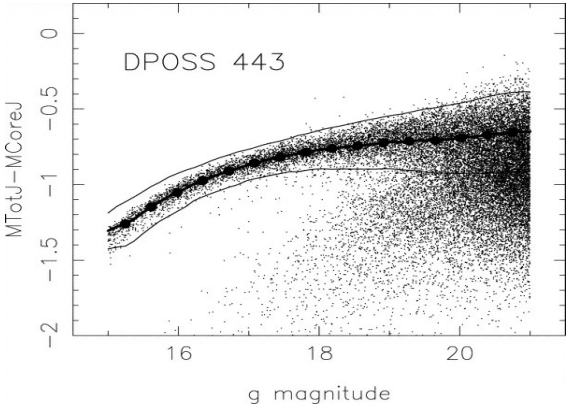
天文应用：降维

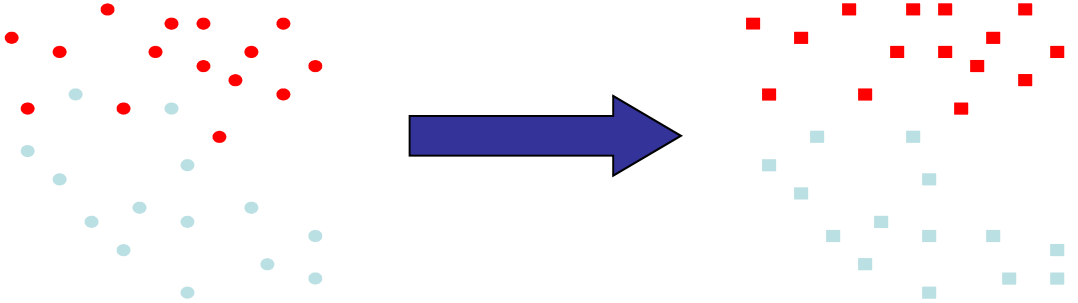
- 降维问题：
  - 寻找相关性和参数的基平面
    - 维数成千上万
      - 维灾！
    - 参数之间的相关性？线性或非线性混合？
    - 本征值或紧致表示是否可以代表整个数据集的性质？



天文应用：叠加与分解

- 叠加和分解问题:
  - 在参数空间中重叠的天体找出它们的所属类别

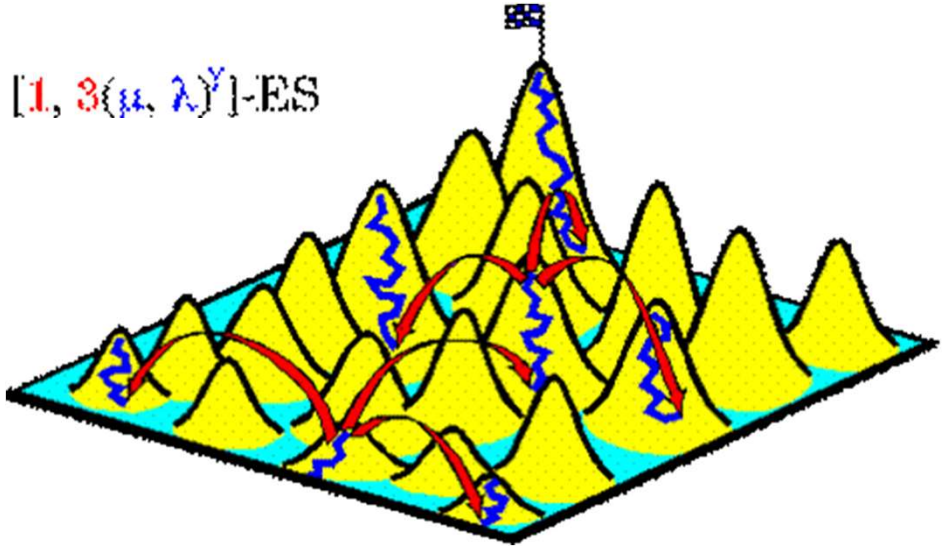




  - 假设10^{10} 天体在1000维空间中重叠怎么办?
  - 如何最优地分解和抽取不同类型的天体?
  - 一些约束条件如何应用?

天文应用：最优化

- 最优化问题:
  - 在高维参数空间中如何找到复杂的多变量函数的最优解（最佳拟合、全局最大似然）



天文应用：时序分析

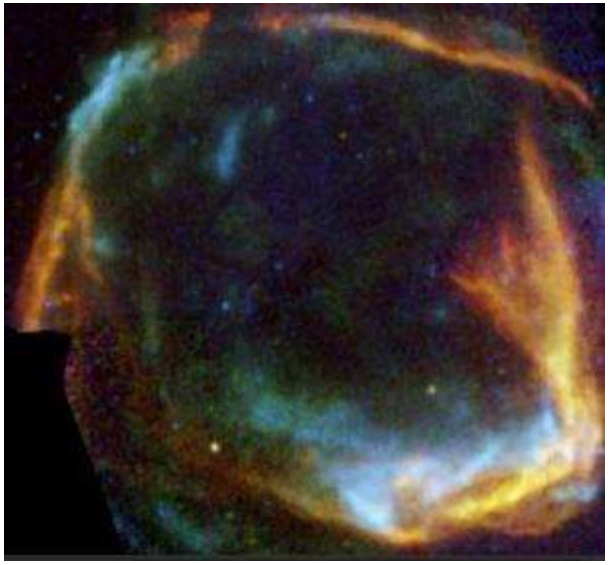
- 变源寻找

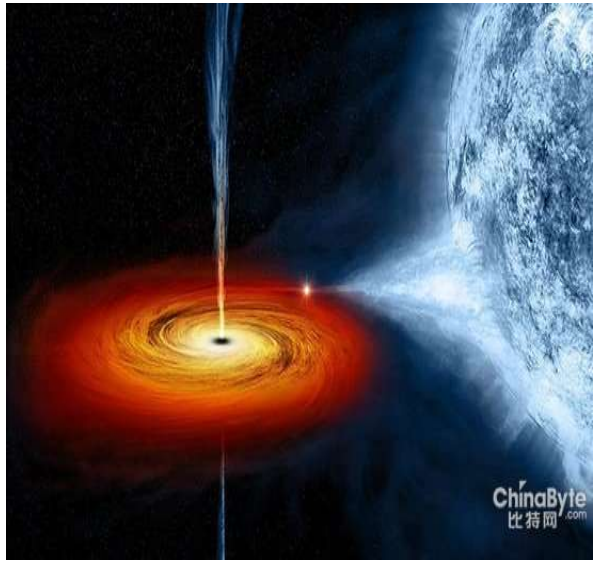
  变星、超新星、类星体、双星、伽玛射线暴等的发现

- 周期寻找

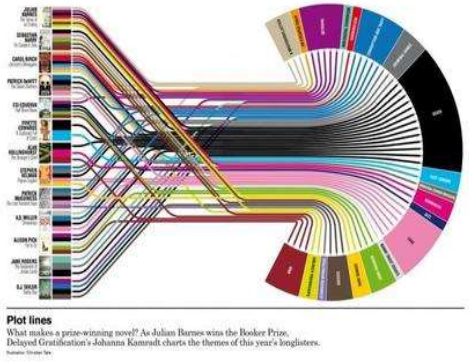
  在时序数据中寻找周期性变化

LSST是未来天文时序研究的最佳试验场

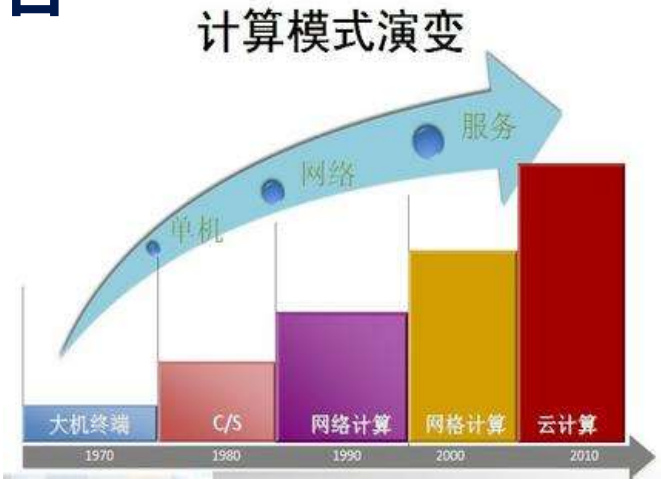




实践数据挖掘



- 线性或非线性
- 高斯或非高斯
- 连续或离散
- 是否存在缺值
- 对比特征和样本数
- 按照数据挖据的任务和特征，选择合适的数据挖掘算法
- 与云计算和云存储结合
- 与数据库结合
- 可视化技术
- 高性能计算结合
- 适合大数据



未来天文数据的挑战

- 统计计算和挖掘方法用于PB和EB量级的可扩张性
- 在海量多维数据空间中同时多点拟合的算法优化
- 用于探索PB级数据的紧致表示的多分辨率、多级、分形、分级方法和结构
- PB量级数据的可视化分析 (包括特征探测, 模型和有趣事件或天体的发现, 相关关系、聚类, 新类型天体的发现, 降维)
- 高维PB级数据的索引和联合存储技巧（树、图、网络拓扑）
- PB级数据库的快速查询和搜索方法

主要挑战

知识发现工具
- 可用性、可扩展性、互动的数据挖掘+可视化
- 机器学习/人工智能和人机交互的发现

社区的认知和职业规划
- 改变科学届/学术届的文化
- 奖励和认可机制

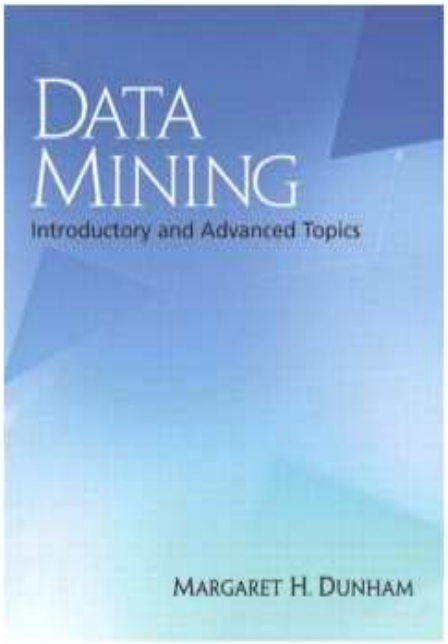
超高维数据空间的可视化
- 优化人类感知和理解
- 可视化的数据探索和发现

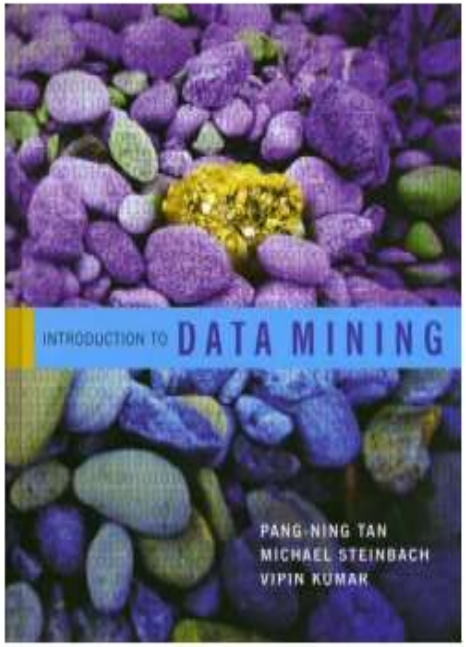
出版和合作的新形式
- 超出论文的范畴;较好的合作工具

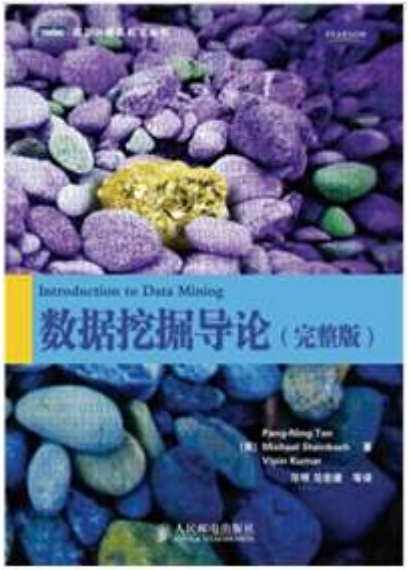
培养新一代的科学家
- 更好地使用在线的学习工具和方法

推荐阅读

Recommended Books








Thank You !