数据分析架构实例与安全云挖掘

中山大学海量数据与云计算研究中心

北京师范大学珠海研究院

吕威

提纲

- Part 1 数据分析架构实例
  - 数据挖掘例子
  - 数据分析架构实例——网站用户流失预警
  - 开源数据分析软件Weka介绍
- Part 2 大规模数据挖掘（云挖掘Hadoop)
  - Map-Reduce方法
  - Classification (k-NN) 的MapReduce化
- Part 3 安全云挖掘
  - 微分流形在安全云挖掘中的应用(Matlab)

Part 1 数据分析架构实例

数据挖掘例子

网站用户行为分析

Weka介绍

定义、概念

数据分析架构实例

开源软件

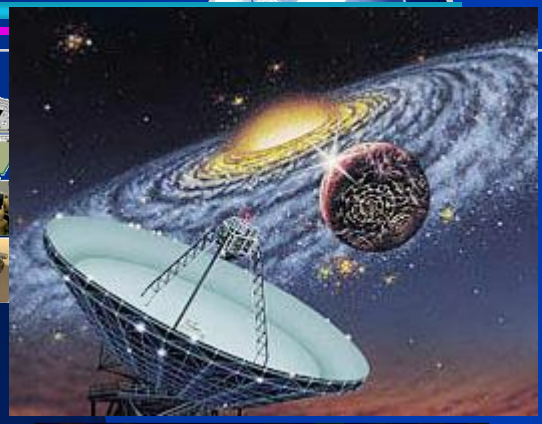
Why Mine Data? Commercial Viewpoint

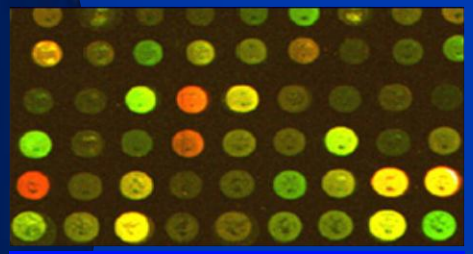
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ **grocery** stores
  - Bank/Credit Card transactions

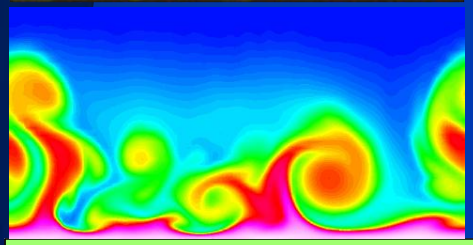
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

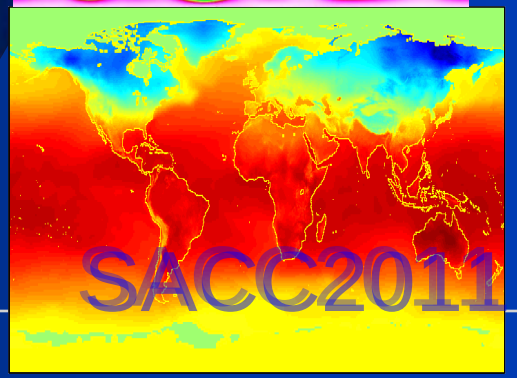
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



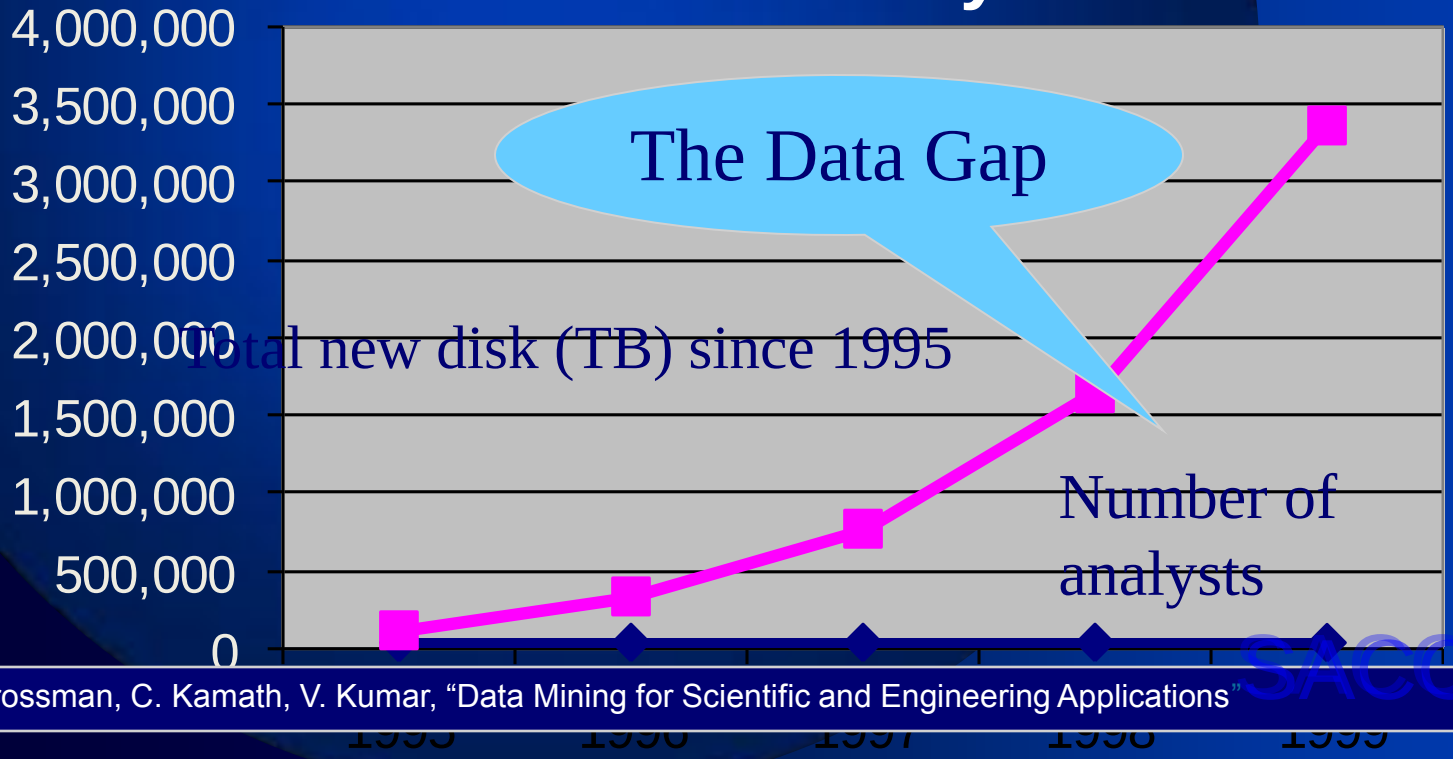






Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all

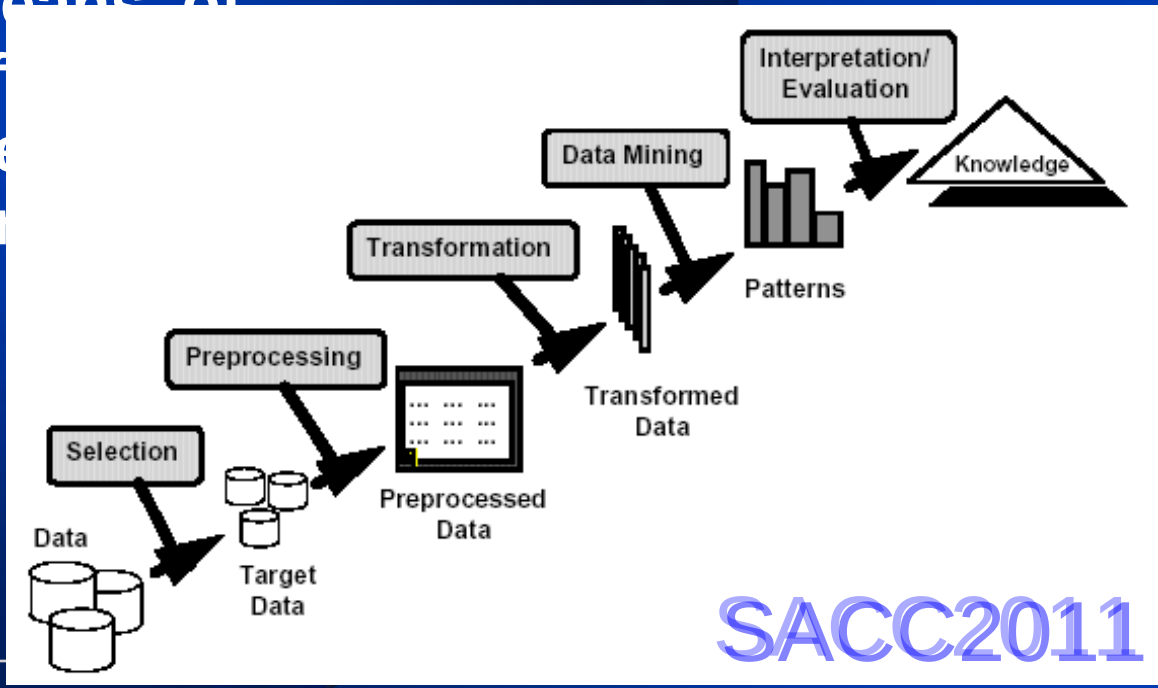


From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

What is Data Mining?

Many Definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of in order to discove meaningful patter



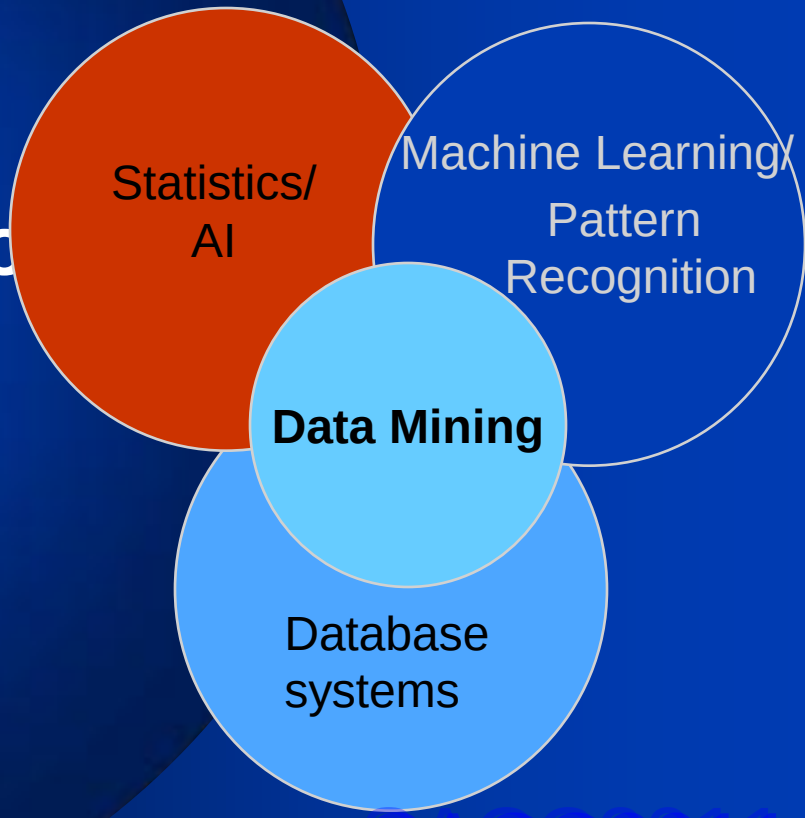
What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional Techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



Data Mining Tasks

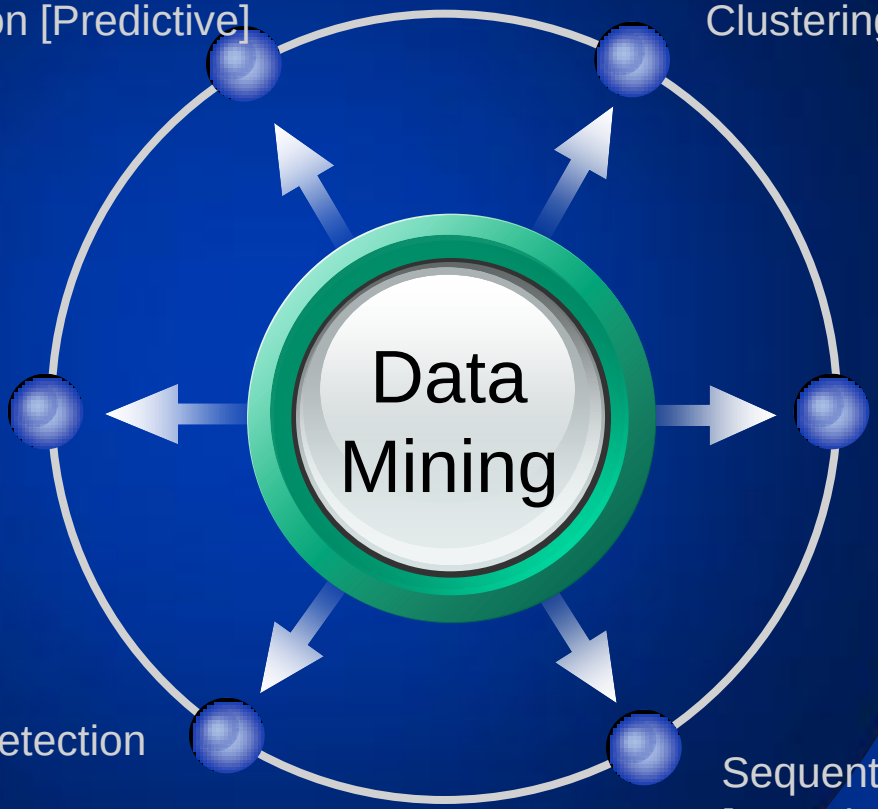
Prediction Methods

Use some variables to predict unknown or future values of other variables.

Description Methods

Find human-interpretable patterns that describe the data.

Data Mining Tasks...



Classification [Predictive]

Clustering [Descriptive]

Regression [Predictive]

Association Rule Discovery [Descriptive]

Deviation Detection [Predictive]

Sequential Pattern Discovery [Descriptive]

数据挖掘例子

1. 超市分析交易数据，安排货架上货物摆布，以提高销售额
2. 信用卡公司分析信用卡历史数据，判断哪些人有风险，哪些人没有
3. 保险公司分析以前的客户记录，决定哪些客户的潜在花费是昂贵的

数据挖掘例子

4

汽车公司分析不同地方人的购买模型，有针对性地发送给客户喜欢的汽车手册

5

广告公司分析人们购买模式，估计他们的收入和孩子数目，作为潜在的市场信息

6

税务局分析不同团体的交所得税的记录，发现异常模型和趋势

Part 1 数据分析架构实例

数据挖掘例子

网站用户行为分析

Weka介绍

定义、概念

数据分析架构实例

开源软件

网站用户行为分析架构实例

- 某网站是游戏门户网站，在多个服务器上运营着多款游戏，每天有大量数据如日志记录等。需根据记录数据进行分析，得出一些有用结果。
- 现在已有各种统计报表，如每日各款游戏点击排名、游戏大厅位置点击排名、各种统计量的饼图、柱图等。
- 希望进一步得到细化分析——数据分析、挖掘

网站用户行为分析

预测模块

奇异点分析

分类模块

网站用户行为分析

聚类模块

关联规则模块

。。。

怎么搭建整个模型呢？

预测模块

用户流失预警

Decision Tree决策树算法、Bayes贝叶斯算法

游戏访问预测

Case Based Reasoning案例推理算法

用户充值预测

K nearest neighbor最近邻算法、最小二乘法

奇异点分析

游戏奇异点分析

Graphical & Statistical-based 图形统计方法

用户流失奇异点分析

Nearest-neighbor based 最近邻方法

用户充值奇异点分析

Density based 密度方法

分类模块

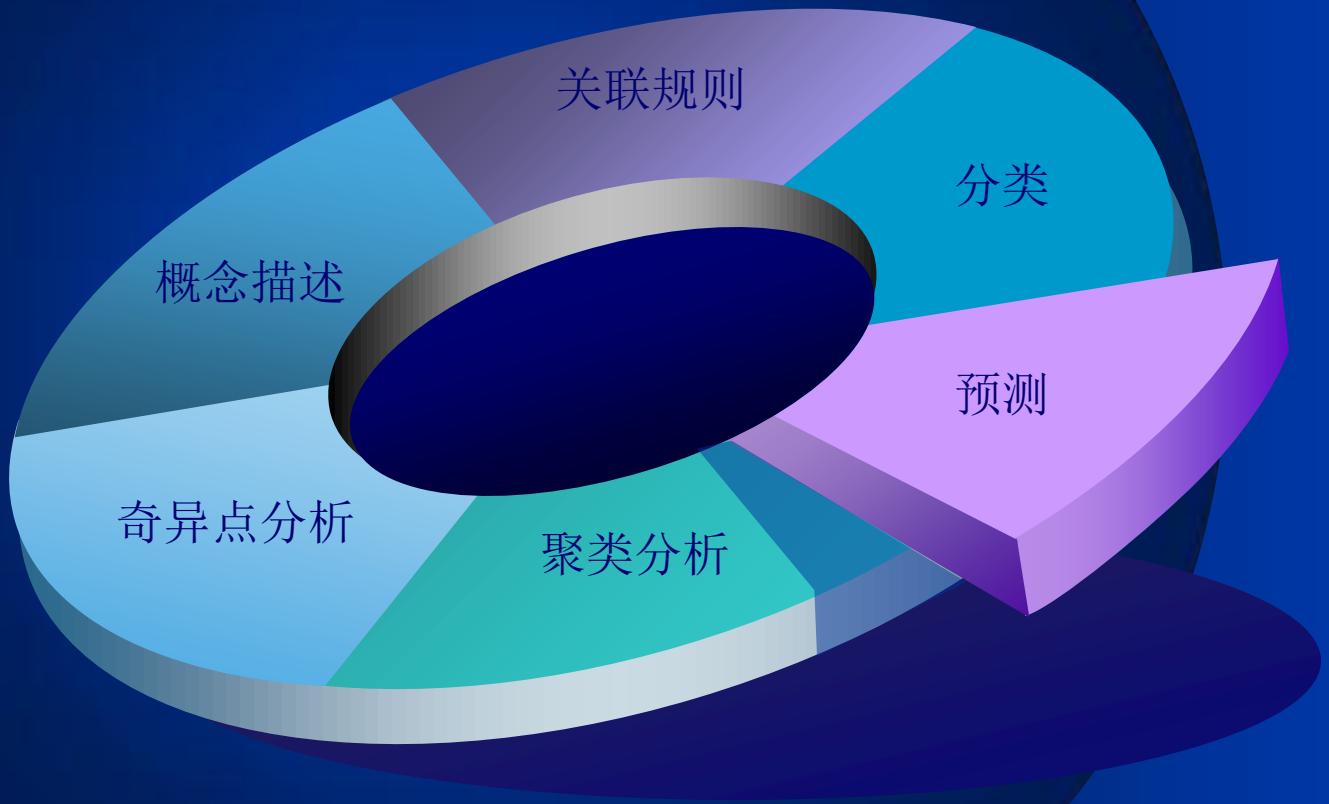
- 游戏分类——Instance-based k, Ibk算法
- 玩家分类——Bayes贝叶斯算法

聚类模块

- 玩家聚类——Kmeans 均值算法
- 游戏聚类——Kmeans 均值算法

- 访问规律——Apriori算法

任务确定



架构目标确定

用户流失预警

流失客户原因分析

其它 模块

用户流失预警——Bayes算法，Ibk算法

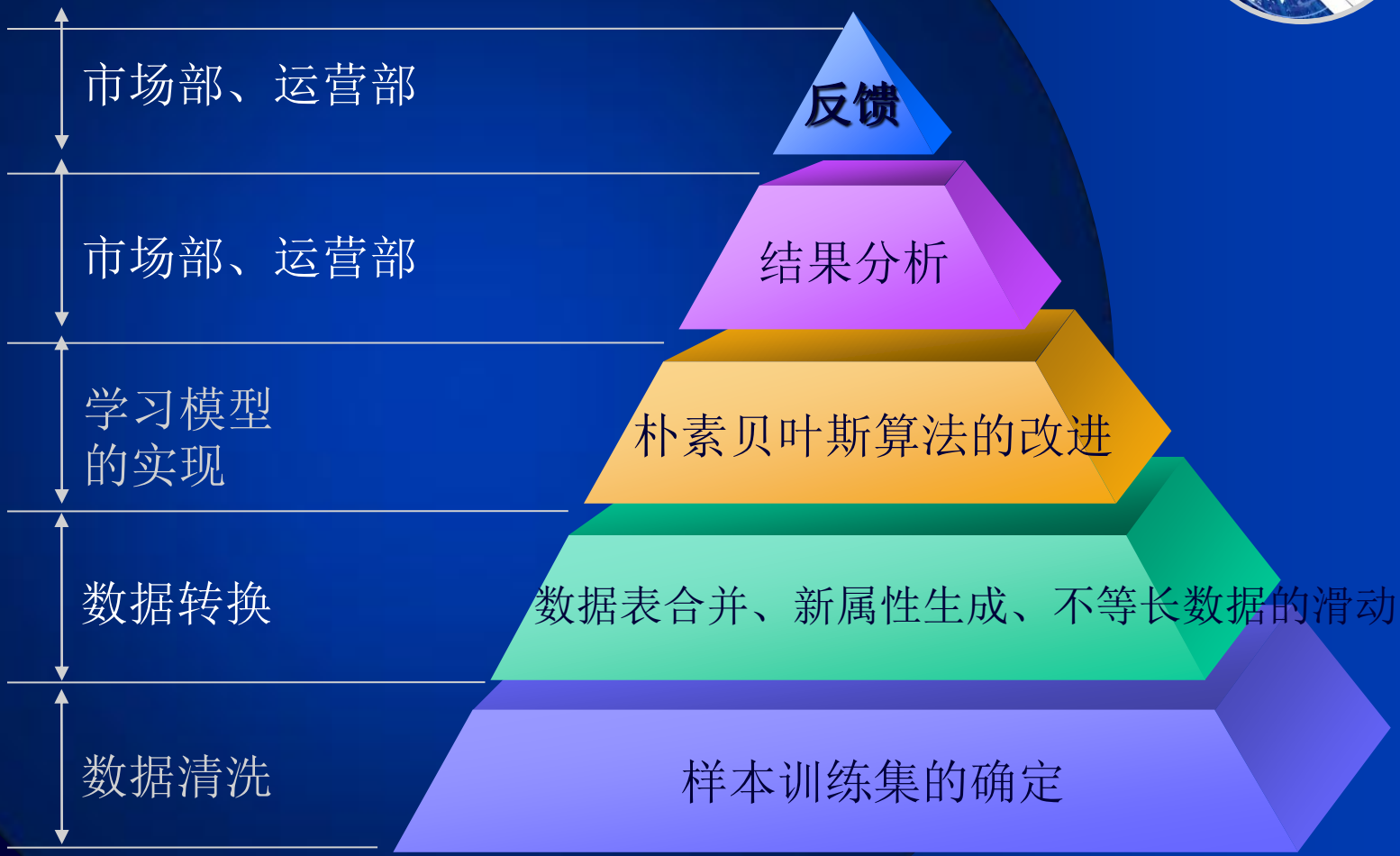
流失客户原因分析——Kmeans 均值聚类算法

其它各模块：可放在长期目标

客户流失预警

- 客户流失分析过程指客户流失逻辑模型的建立过程,包括数据采样、数据分析、模型评估和应用，在一系列分析之后得出客户流失的名单列表、流失的原因、特征和进行流失预警。
- 注意：目前侧重的是预测客户流失，与客户分类应该有一定的区别

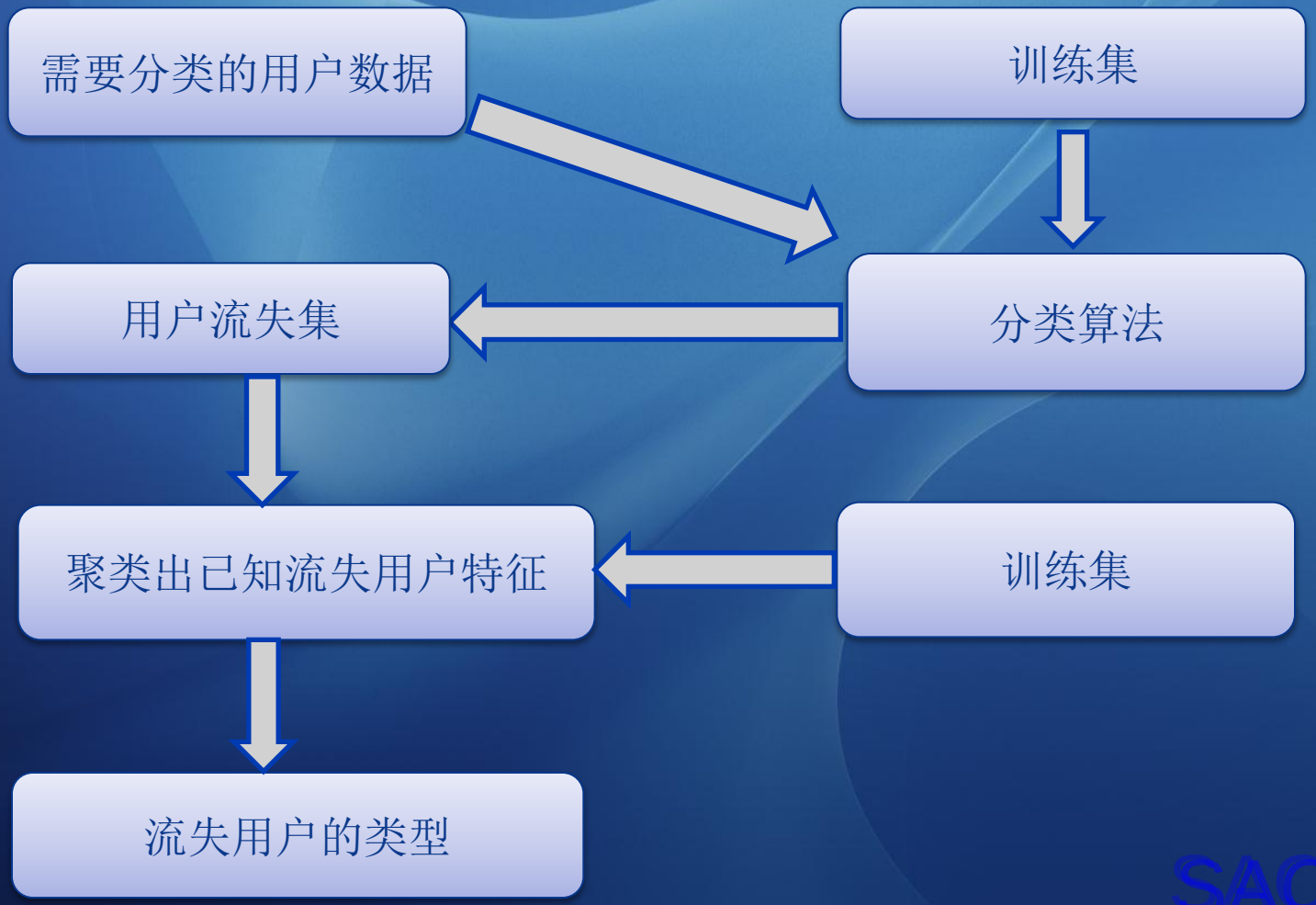
架构过程



架构过程

	数据清洗	数据转换	学习模型实现
方法	小规模数据上先进行尝试	多个表合并成一个大的物理表	简单、可用、快速、好解释
样本数据	某款游戏一个月数据大约150,000条	心跳表中的多条数据将合成一条数据	属性挑选
去除完全无用	大约剩50，000条	生成一些贝叶斯分类算法有用属性	连续属性离散化
活跃用户	剩下15,000条左右	不等长数据处理：滑动窗口	贝叶斯分类
流失用户	大约1/3——5,000条	训练数据集完成：15，000条整齐数据	Kmeans聚类

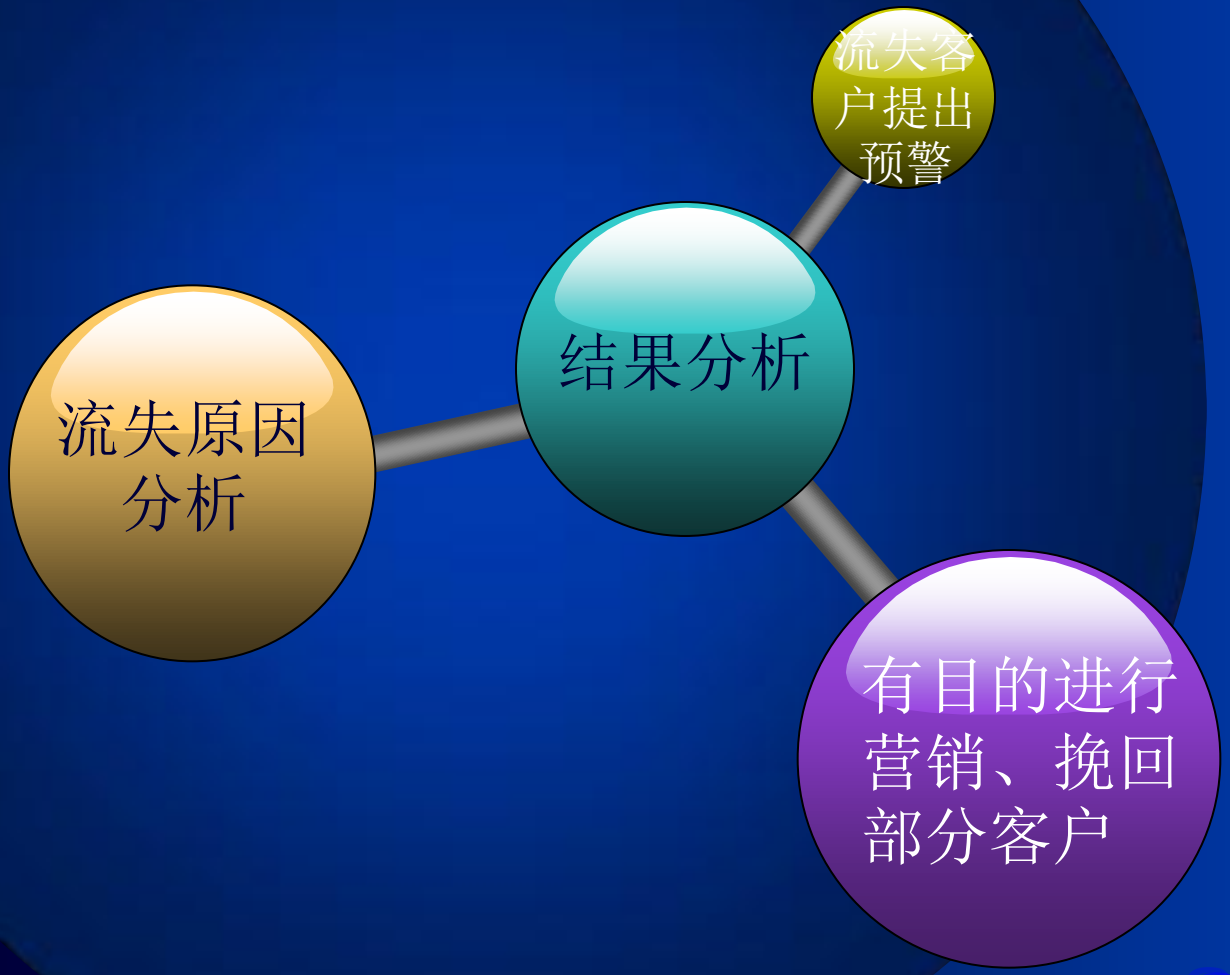
算法框架



算法结果

样本表	客户数目	改进贝叶斯算法准确度
test1	1000	714
test2	1000	736
test3	1000	747
test4	1000	716
test5	1000	762

决策支持



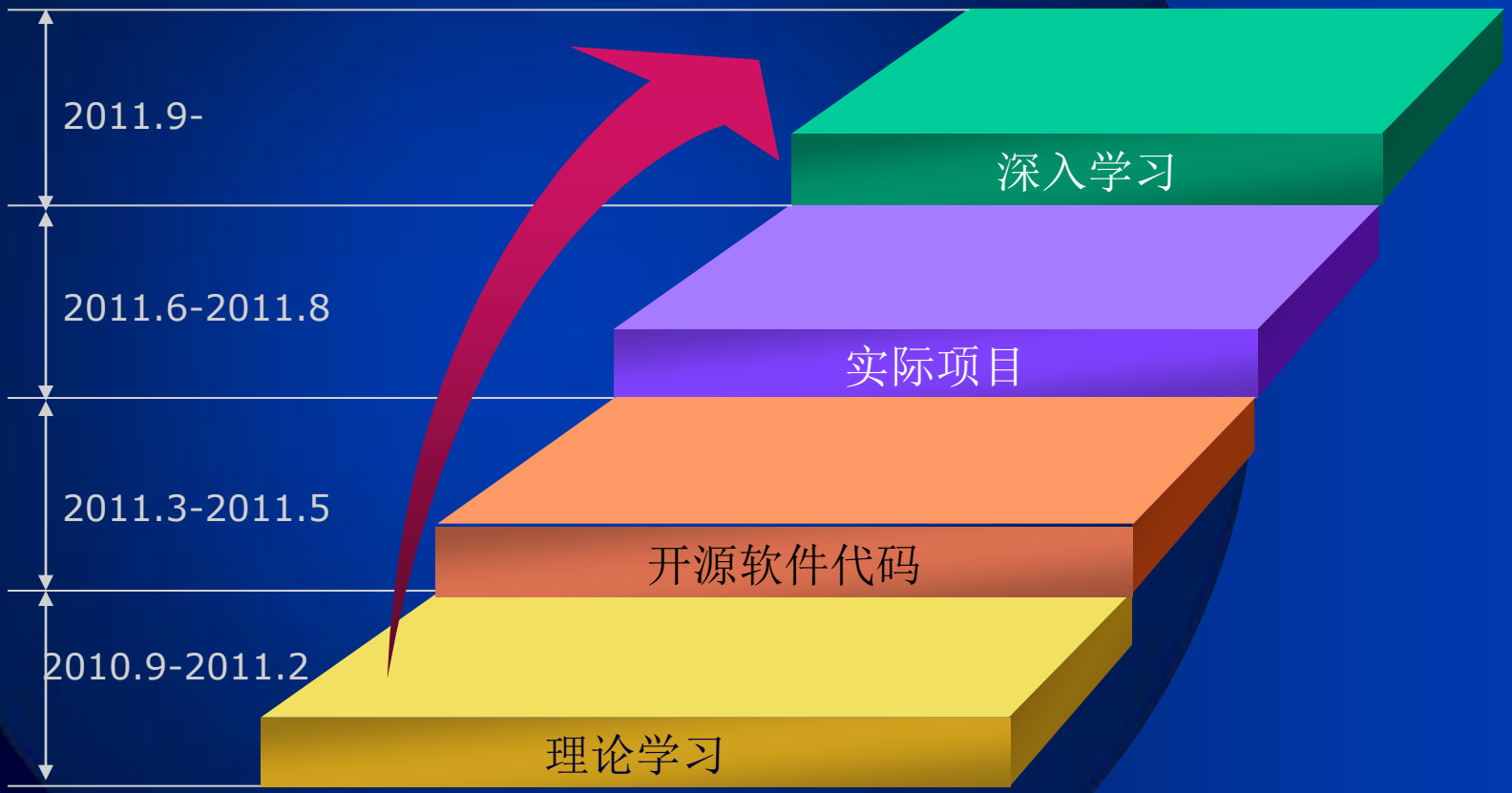
开源数据分析软件Weka介绍

- 开源
- 全面
- 规范
- WEKA的全名是怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），它的源代码可通过http://www.cs.waikato.ac.nz/ml/weka得到

开源数据分析软件Weka介绍

- WEKA作为一个公开的数据挖掘工作平台，集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。如果想自己实现数据挖掘算法的话，可以看一看weka的接口文档。在weka中集成自己的算法甚至借鉴它的方法自己实现可视化工具并不是件很困难的事情。

学生做数据分析项目过程



提纲

- Part 1 数据分析架构实例
  - 数据挖掘例子
  - 数据分析架构实例——网站用户流失预警
  - 开源数据分析软件Weka介绍
- Part 2 大规模数据挖掘（云挖掘Hadoop)
  - Map-Reduce方法
  - Classification (k-NN) 的MapReduce化
- Part 3 安全云挖掘
  - 微分流形在安全云挖掘中的应用(Matlab)

大规模数据挖掘

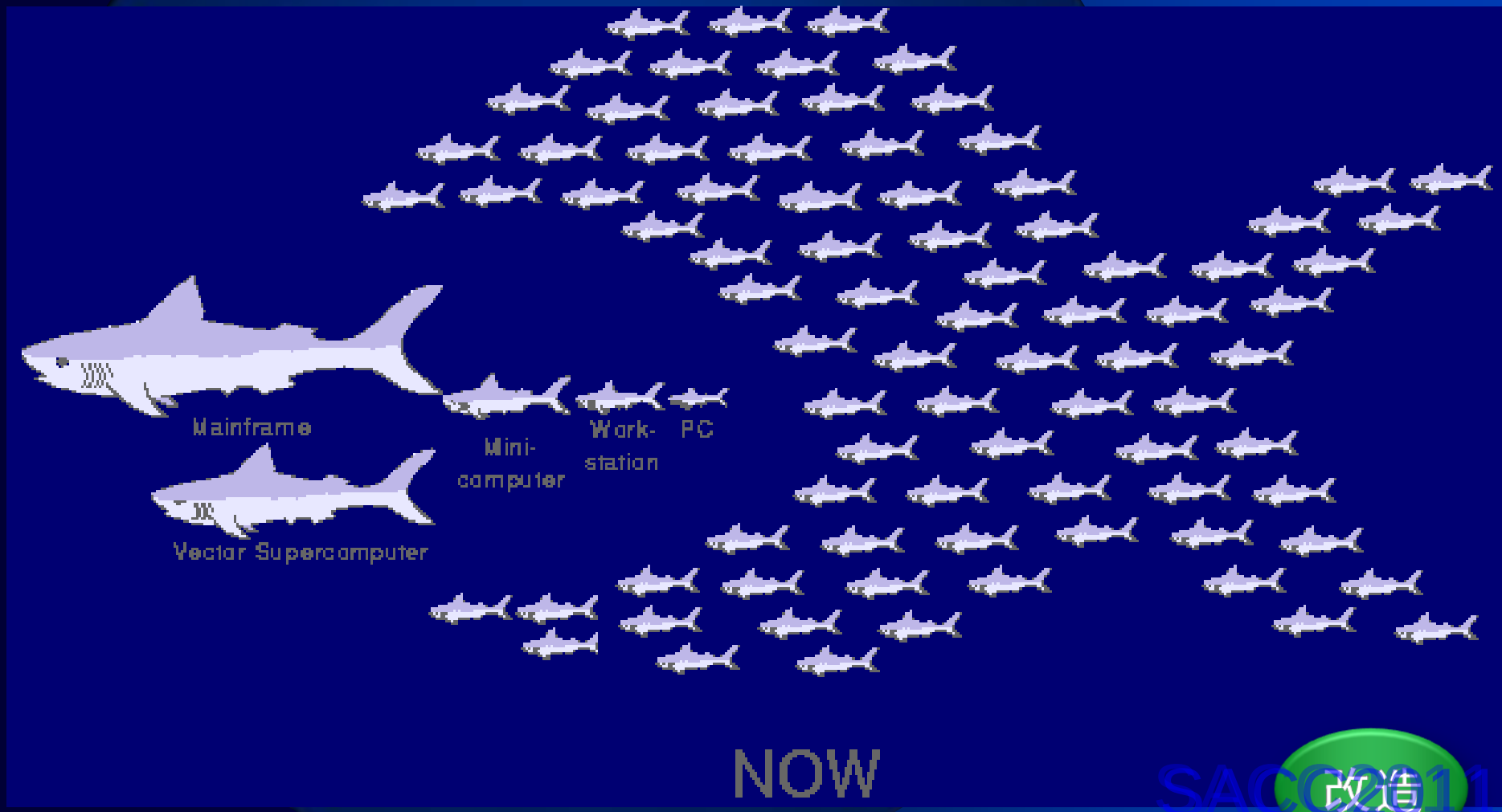
多款游戏、多台服务器

每天独立登陆IP有 600,000~700,000个

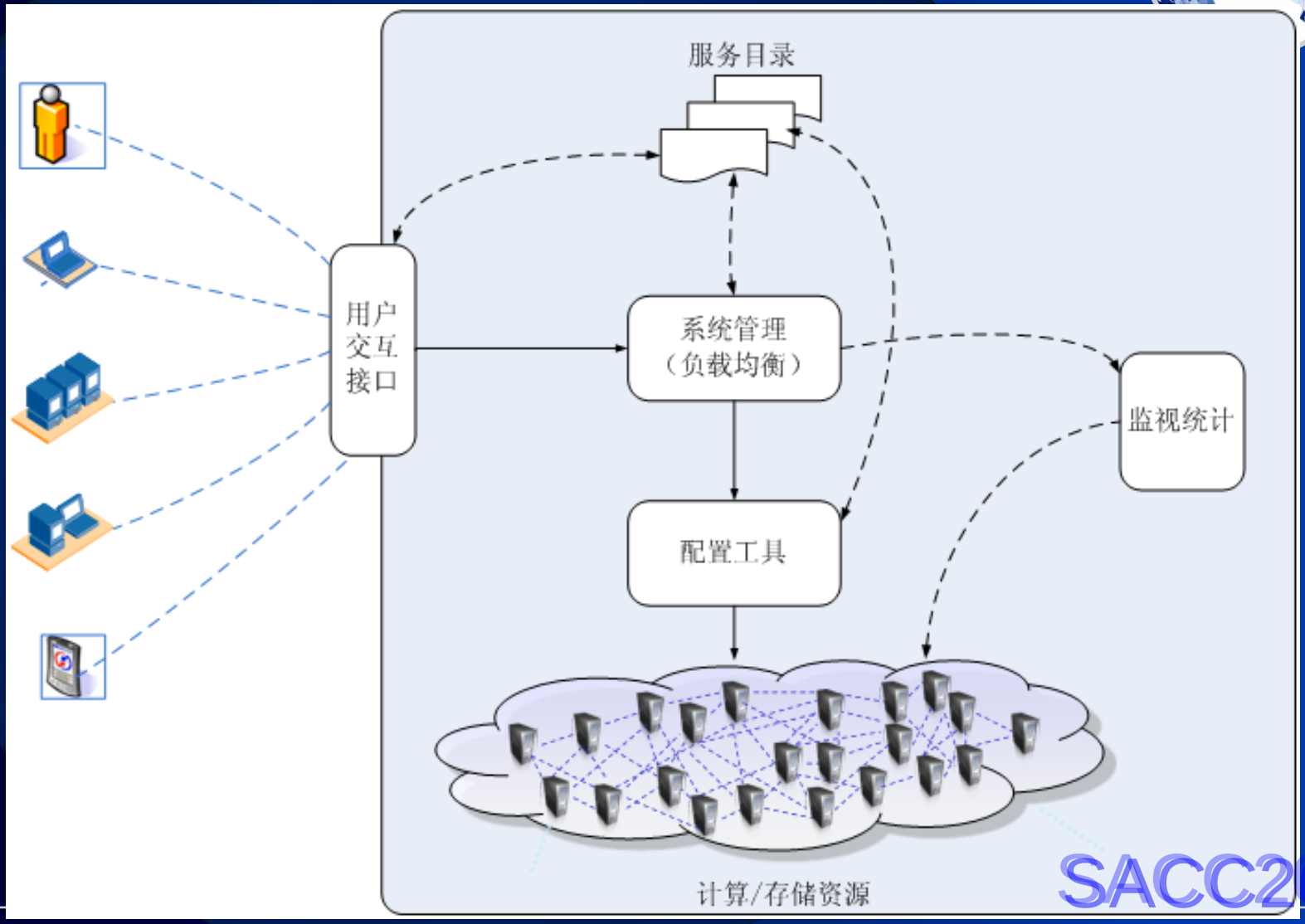
一些数据挖掘算法跑不起来

云化
MapReduce
方法

云计算——网络发展的必然结果



云计算简化实现机制

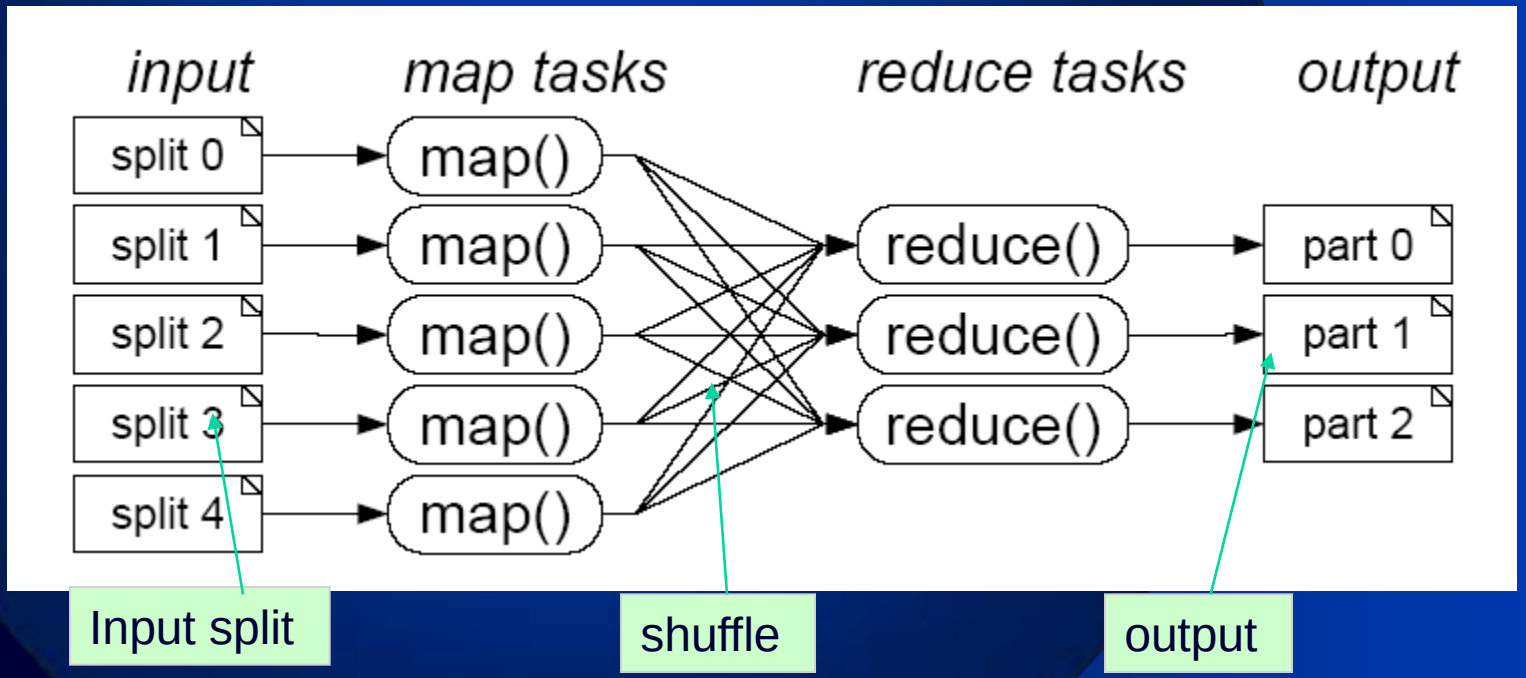


Part 2 大规模数据挖掘（云挖掘 Hadoop)

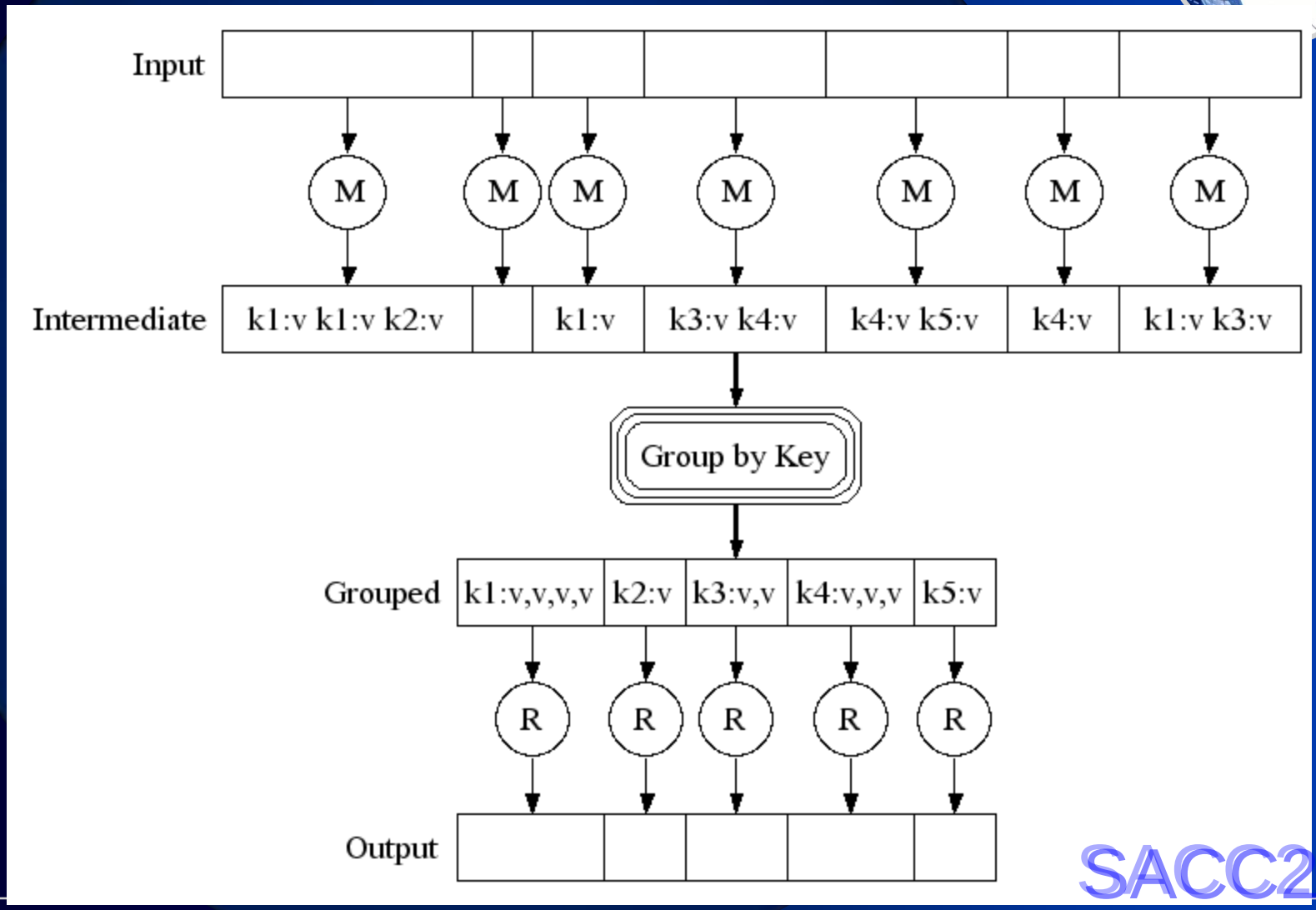
- Map-Reduce方法
- Classification (k-NN)算法的MapReduce化

What's Mapreduce

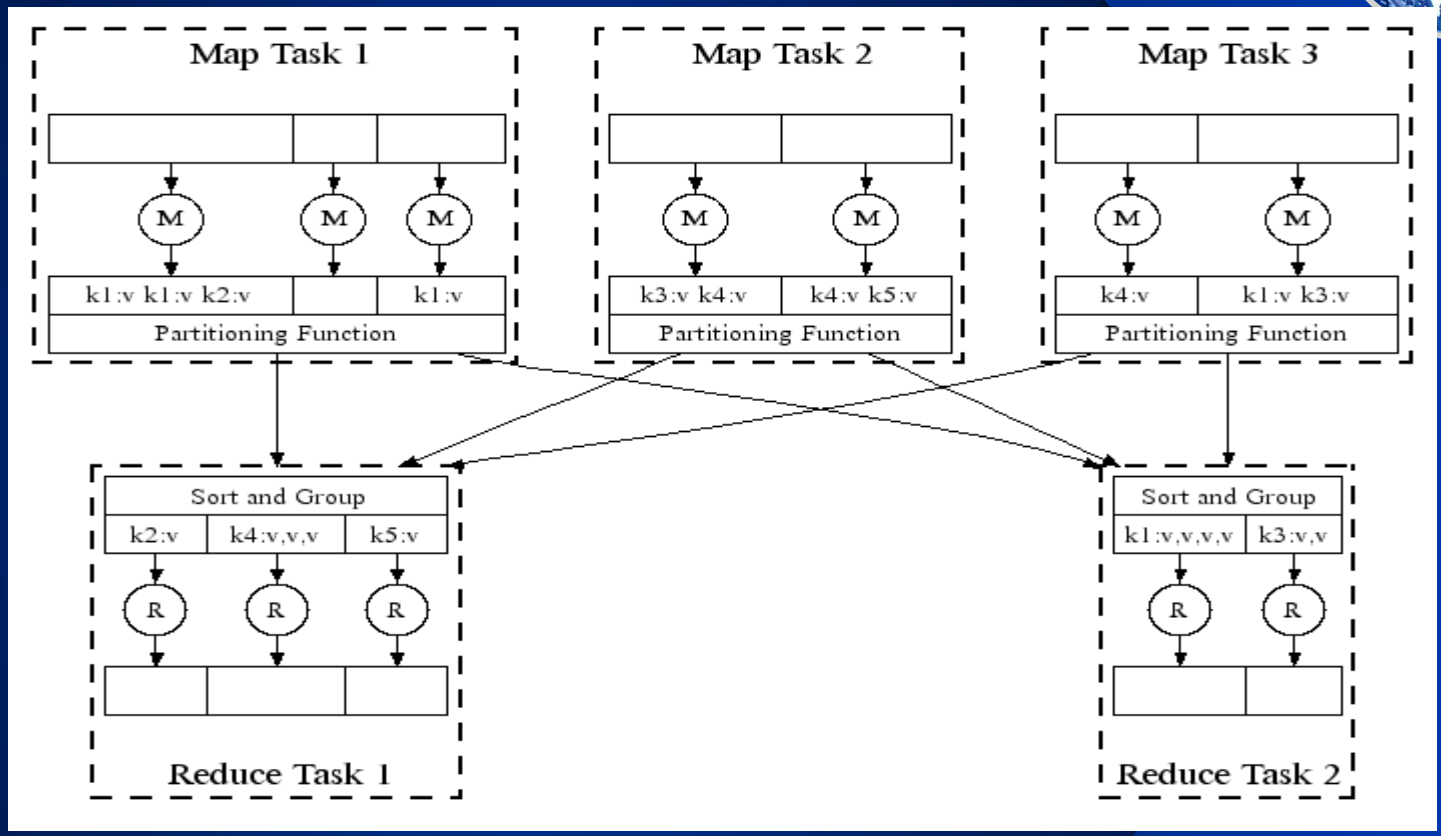
- Parallel/Distributed Computing Programming Model



Shuffle Implementation



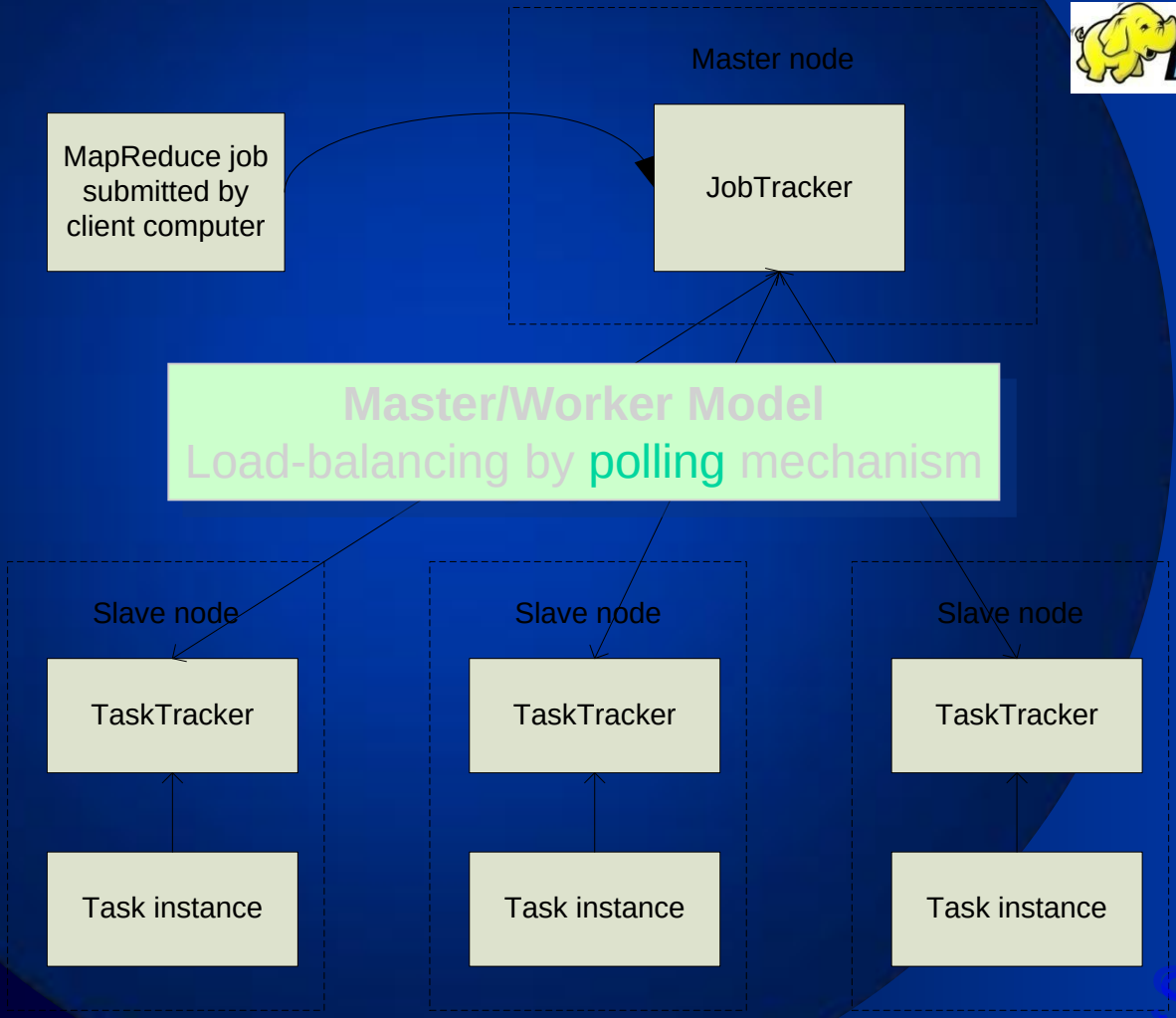
Partition and Sort Group



Partition function: hash(key)%reducer number

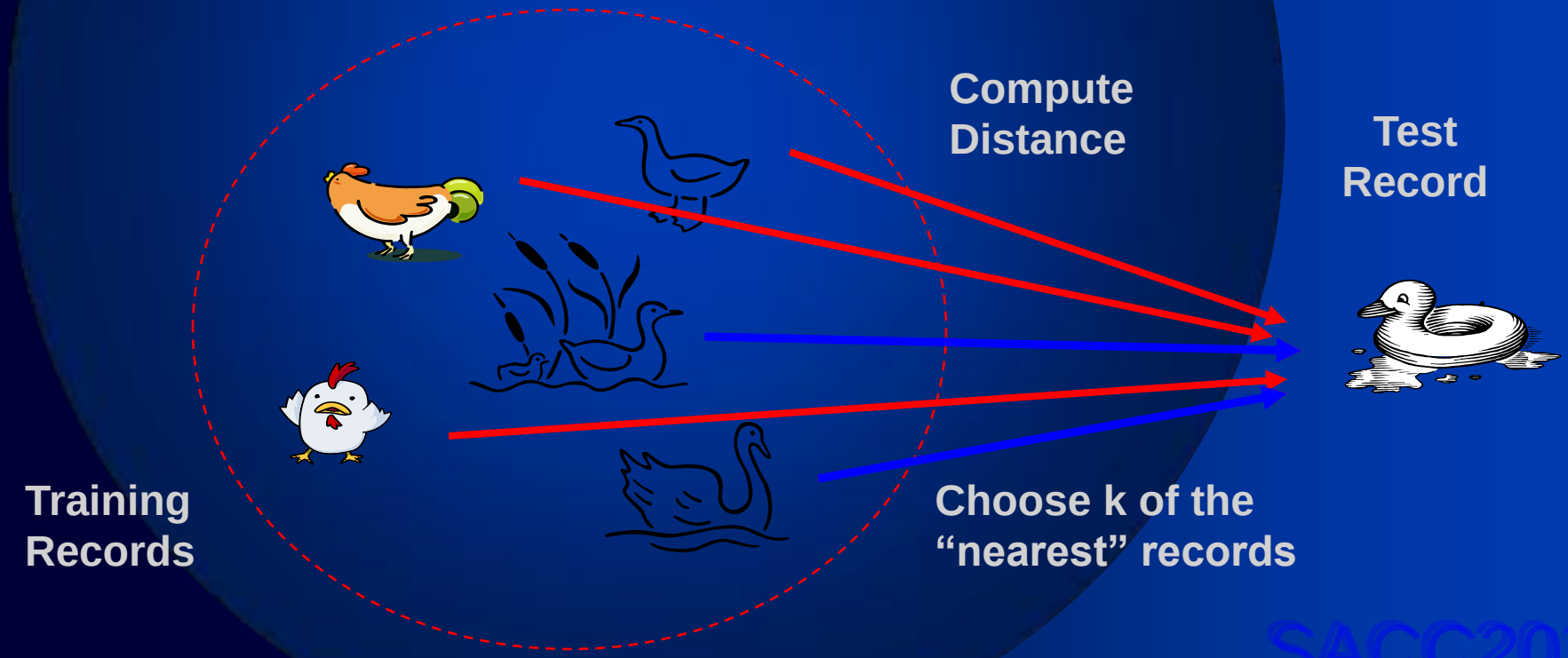
Group function: sort by key

Hadoop MapReduce Architecture

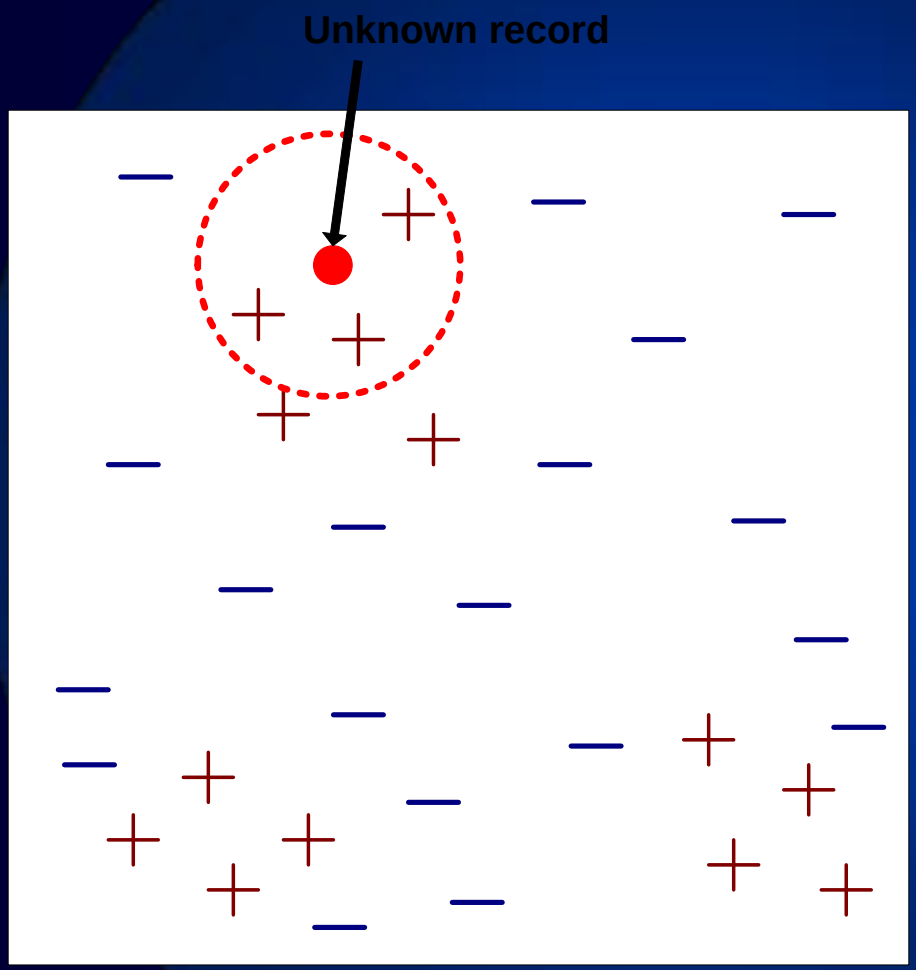


Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck

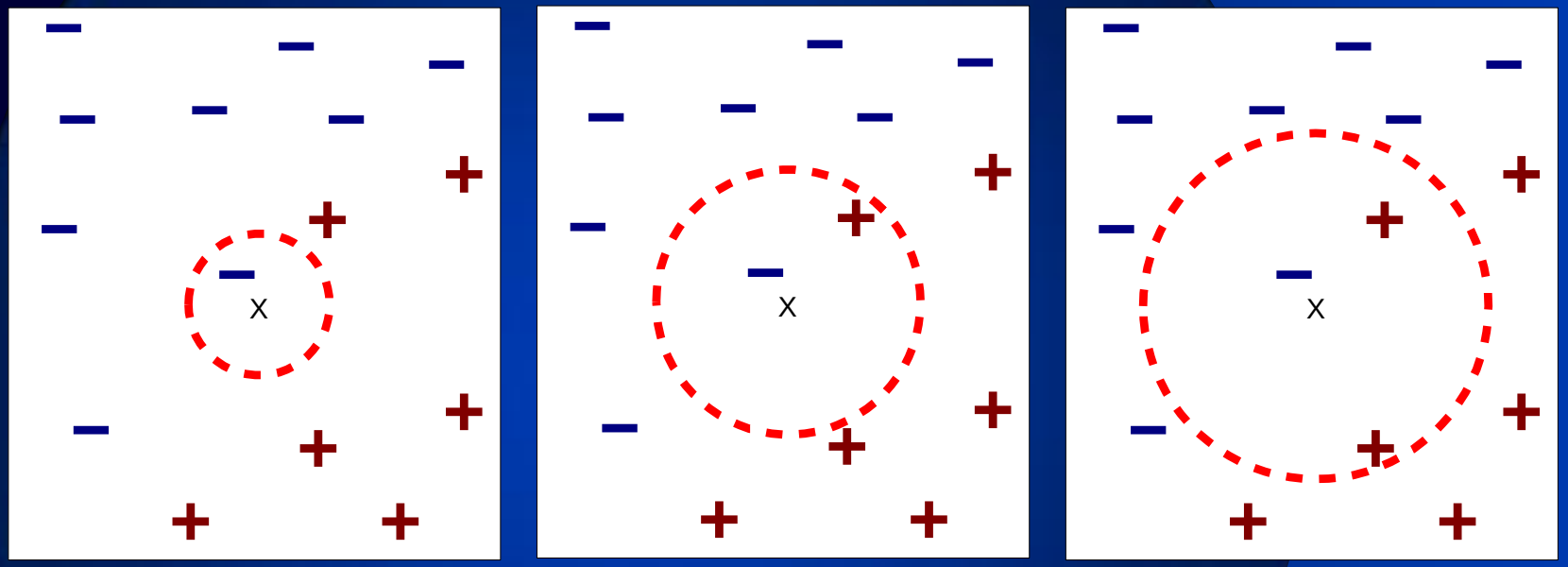


Nearest-Neighbor Classifiers



- Requires three things
  - The set of stored records
  - Distance Metric to compute distance between records
  - The value of *k*, the number of nearest neighbors to retrieve
- To classify an unknown record:
  - Compute distance to other training records
  - Identify *k* nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record (e.g., by taking majority vote)

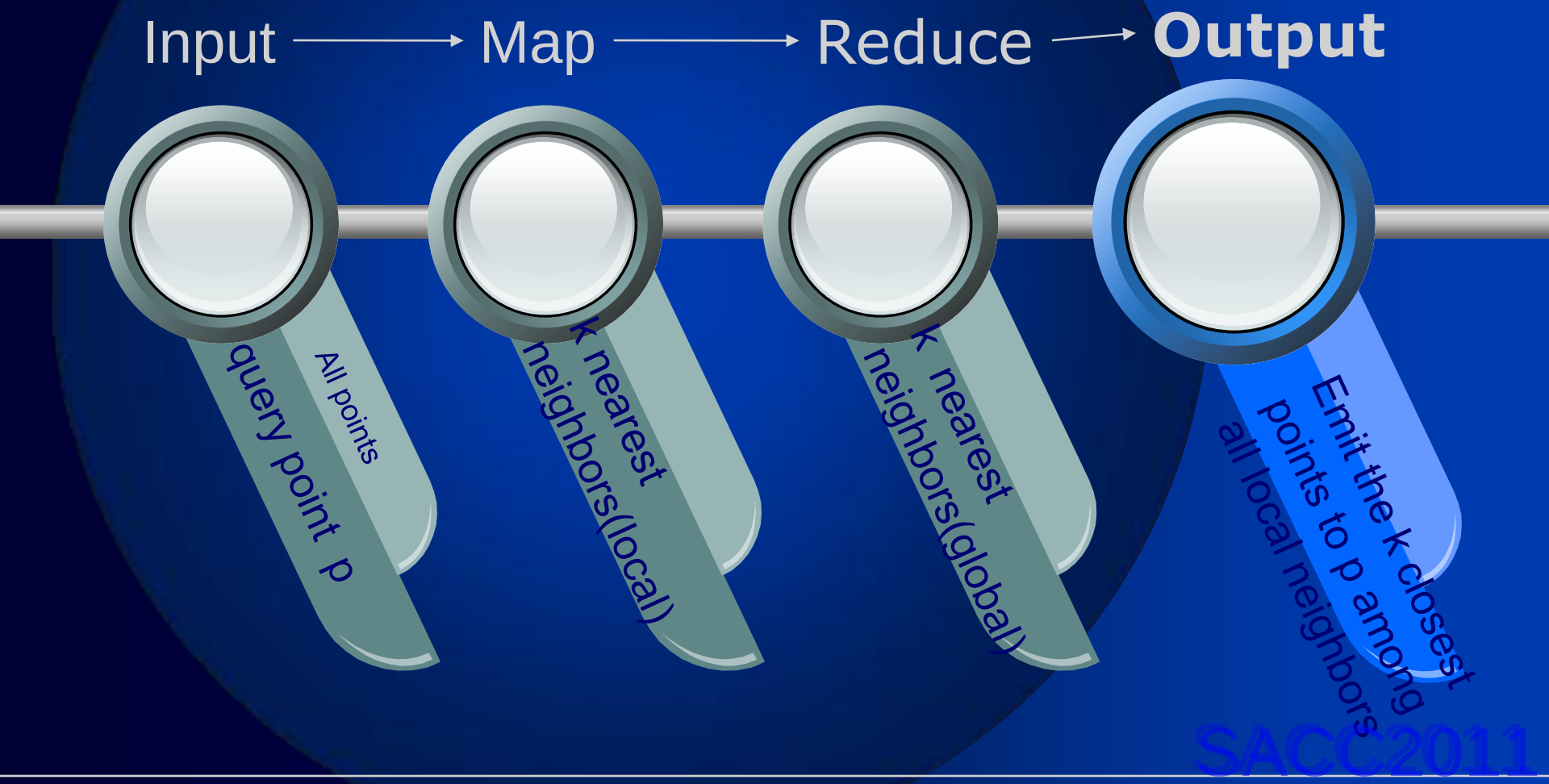
Definition of Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

MapReduce：kNN



MapReduce化算法提高效率

- 单个节点并非跑不出结果：大数据集上需要一天、一周才能出结果。有时候有较高实时要求的任务一小时出结果都太慢
- 利用多个节点进行MapReduce云化，可以利用空置设备同步运行，提高速度，对有较高实时性要求的算法有好处

提纲

- Part 1 数据分析架构实例
  - 数据挖掘的概念与特点
  - 数据分析架构实例——网站用户流失预警
  - 开源数据分析软件Weka介绍
- Part 2 大规模数据挖掘（云挖掘Hadoop)
  - Map-Reduce方法
  - Classification (k-NN) 的MapReduce化
- Part 3 安全云挖掘
  - 微分流形在安全云挖掘中的应用(Matlab)

数据分析带来的隐私保护问题

隐私保护

数据挖掘可以挖掘潜在规律、辅助决策、检测异常模式、恐怖活动和欺诈行为

也可挖掘分析出感兴趣的私人信息。云挖掘中更加涉及到客户端把隐私数据交付给云端进行挖掘，客户对此会产生疑虑。

安全云挖掘

既不泄露隐私，又能保证挖掘结果的大致准确——隐私保护数据挖掘

在客户端向云端传送隐私数据时，可先进行随机化变换、加密

Privacy-preserving Data Mining

Hide sensitive individual data values from the outside world

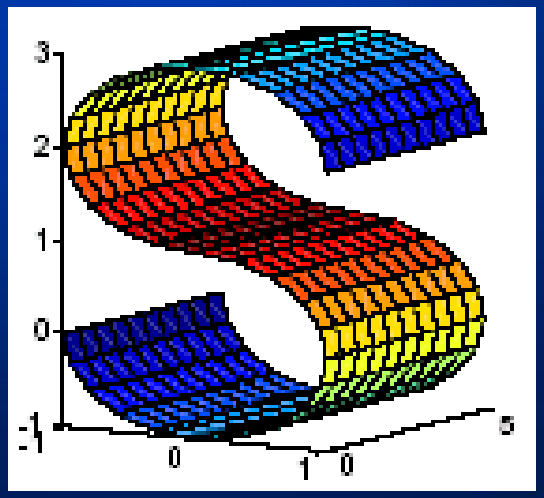
- A Random Rotation Perturbation Approach to Privacy Data Classification
- Deriving Private Information from Randomized Data

- Privacy-Preserving Data mining
- A Framework for High Accuracy Privacy-Preserving Mining

A valid and effcient decision model based on the distorted data can be constructed

微分流形：保持拓扑特性

设 M 是一个Hausdorff 拓扑空间, 若对每一点 $p \in M$,都有 P 的一个开领域 U 和 R^n的一个开子集同胚, 则称 M 为 n 维拓扑流形, 简称为 n 维流形.



几种流形学习算法

1

➢局部线性嵌入(LLE)

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, vol. 290, pp. 2323--2326, 2000.

2

➢等距映射(Isomap)

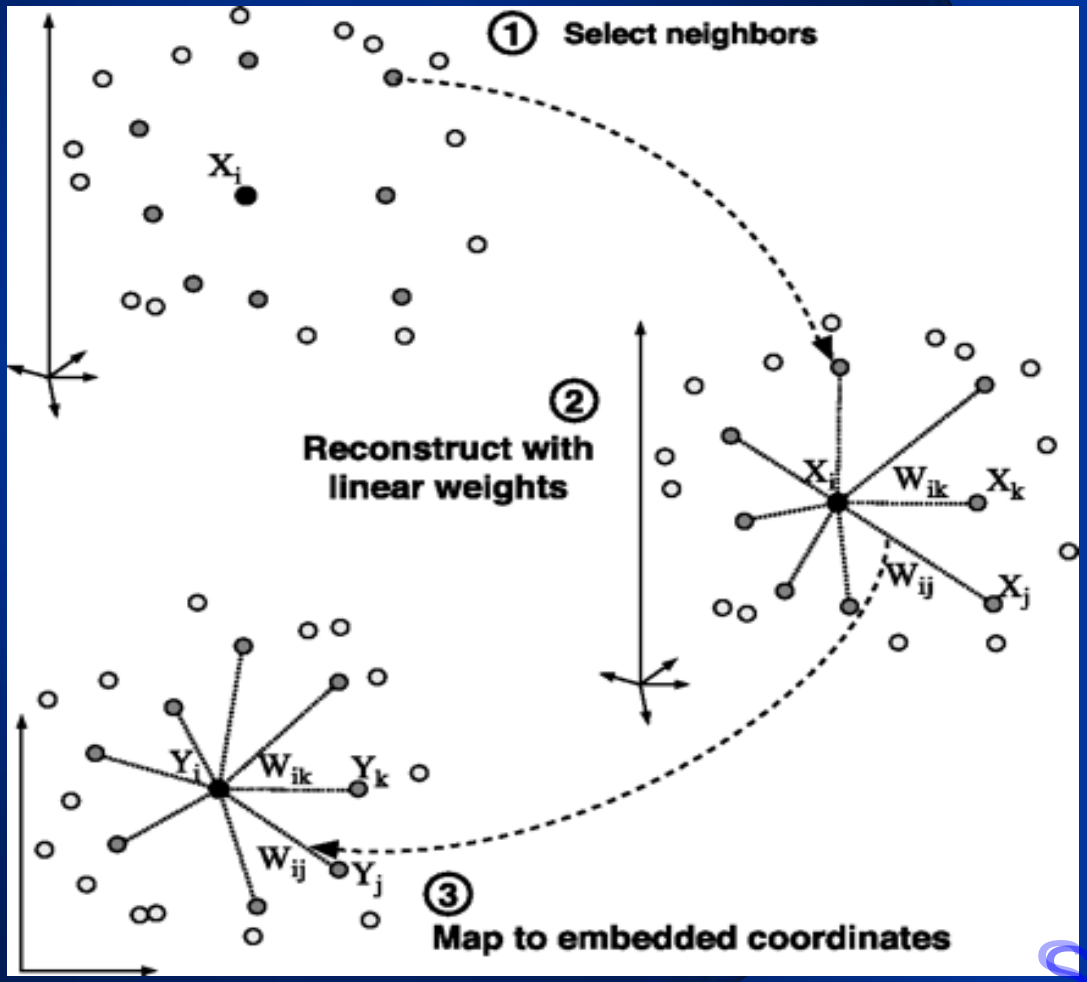
J.B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, vol. 290, pp. 2319--2323, 2000.

3

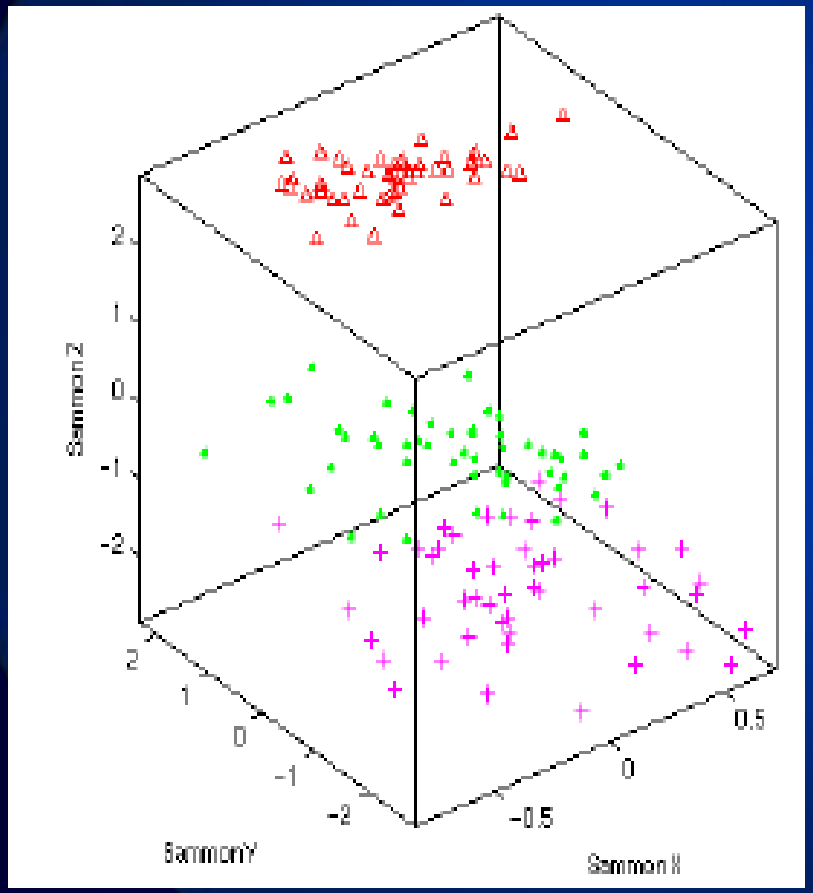
➢拉普拉斯特征映射(Laplacian Eigenmap)

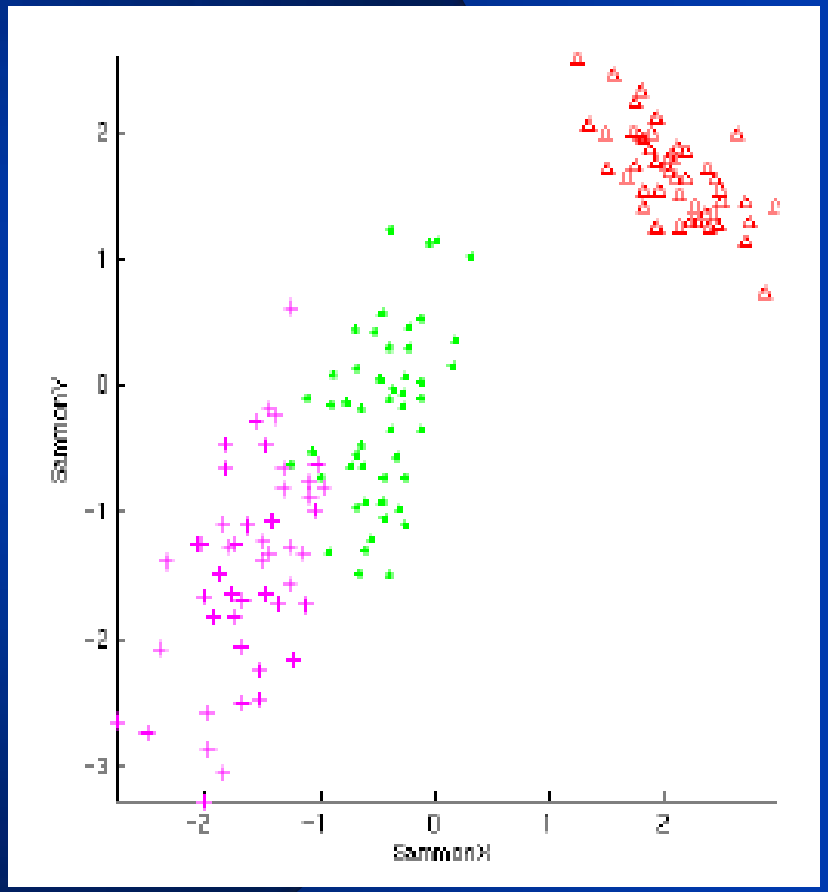
M. Belkin, P. Niyogi, Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, Vol. 15, Issue 6, pp. 1373 –1396, 2003 .

LLE算法示意图



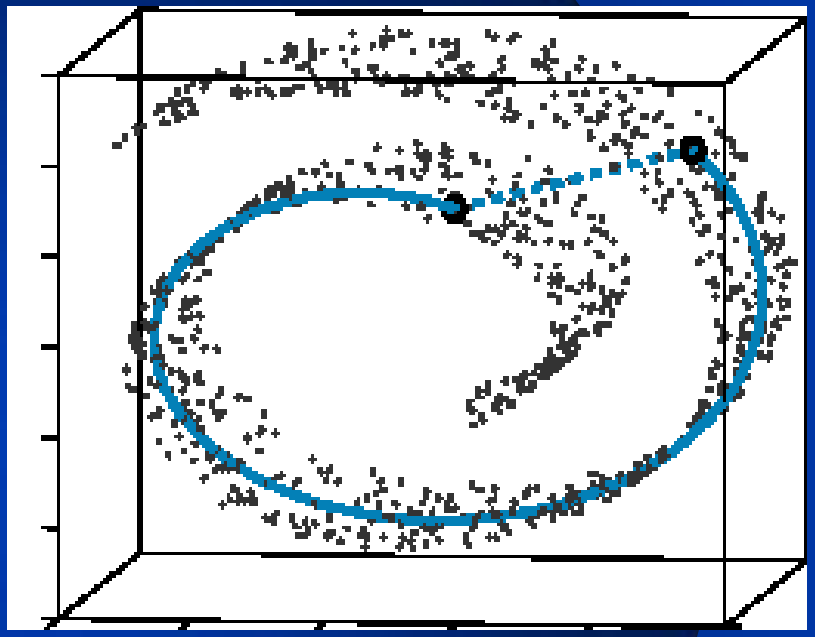
MDS 示意图





Dimensionality Reduction: ISOMAP

By: Tenenbaum, de Silva, Langford (2000)



- Construct a neighbourhood graph
- For each pair of points in the graph, compute the shortest geodesic distances

安全云挖掘

使用微分流形完成了几个隐私保护数据挖掘算法

怎样并行化进行微分流形变换，同时不影响挖掘结果

分析的完整架构

Thank You !

SACC2011