Introduction to Data Mining

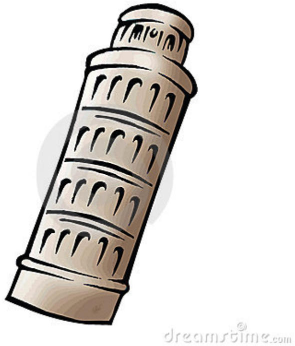
Chiara Renso

KDDLab - ISTI – CNR, Italy

http://www-kdd.isti.cnr.it

email: chiara.renso@isti.cnr.it

Knowledge Discovery and Data Mining Laboratory, ISTI – National Research Council, Italy - Pisa



Analysis of Complex Data:

- Mobility Data Mining
- Complex and Social Networks
- Privacy

Collaboration with UFSC in a EU project called SEEK www.seek-project.eu/

The schedule

Day 1

- Introduction to Data Mining and Data Preprocessing
- Classification
- Practice with Weka

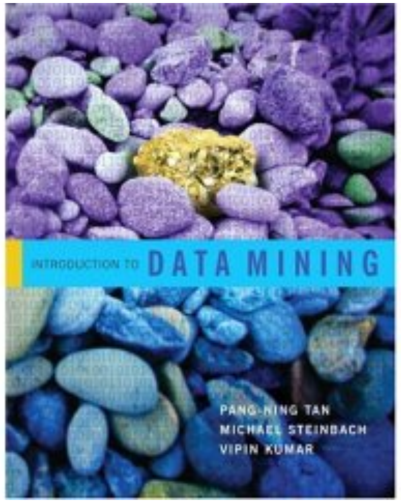
Day 2

- Clustering
- Association analysis
- Practice with Weka

Day 3

- Practice & Advances issues: semantic data mining and trajectory data mining

MATERIAL FOR THE COURSE



Slides have been adapted from the slides associated to the Book: *Introduction to Data Mining*, by Tan, Steinbach, Kumar(1)

http://www-users.cs.umn.edu/~kumar/dmbook/

Some sample book chapters can be downloaded from this site

Practice with the Opensource Weka Tool: Data Mining Software in Java

http://www.cs.waikato.ac.nz/ml/weka/

Weka 3.6 is the latest stable version of Weka,

DATA MINING: INTRODUCTION

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

David Hand, Heikki Mannila & Padhraic Smyth, *Principles of Data Mining, MIT Press, 2001*

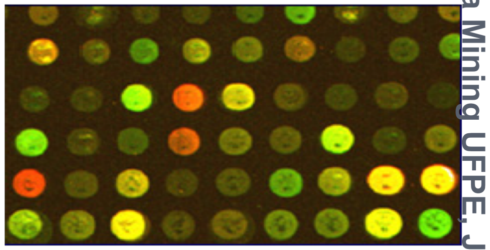
WHY MINE DATA? COMMERCIAL VIEWPOINT

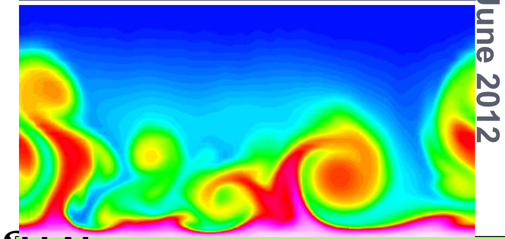
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

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

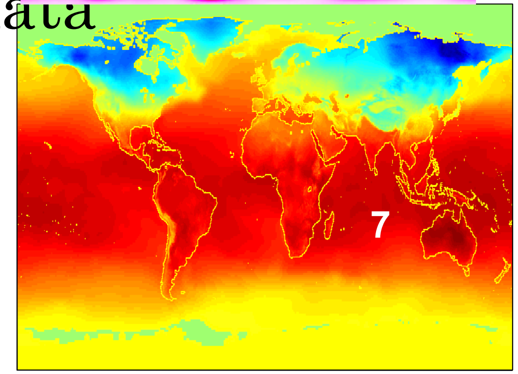
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data



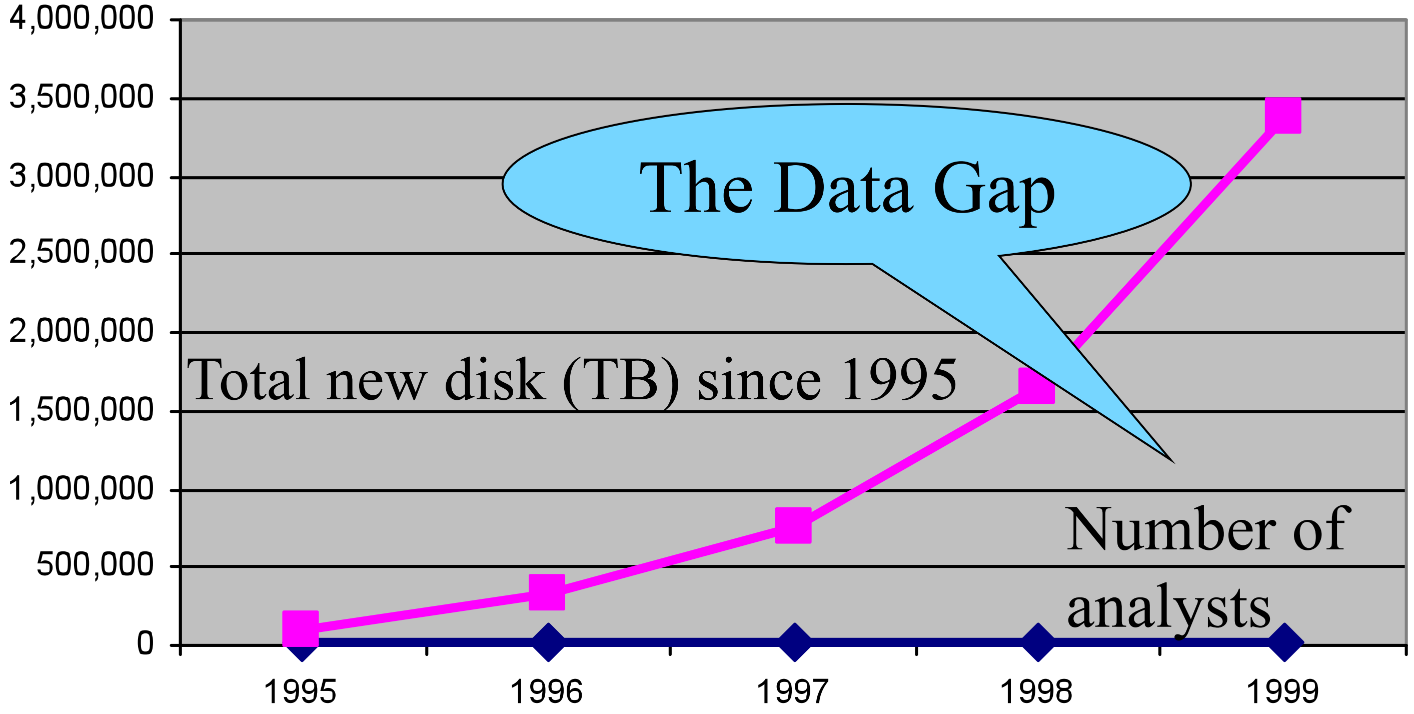


- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation



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

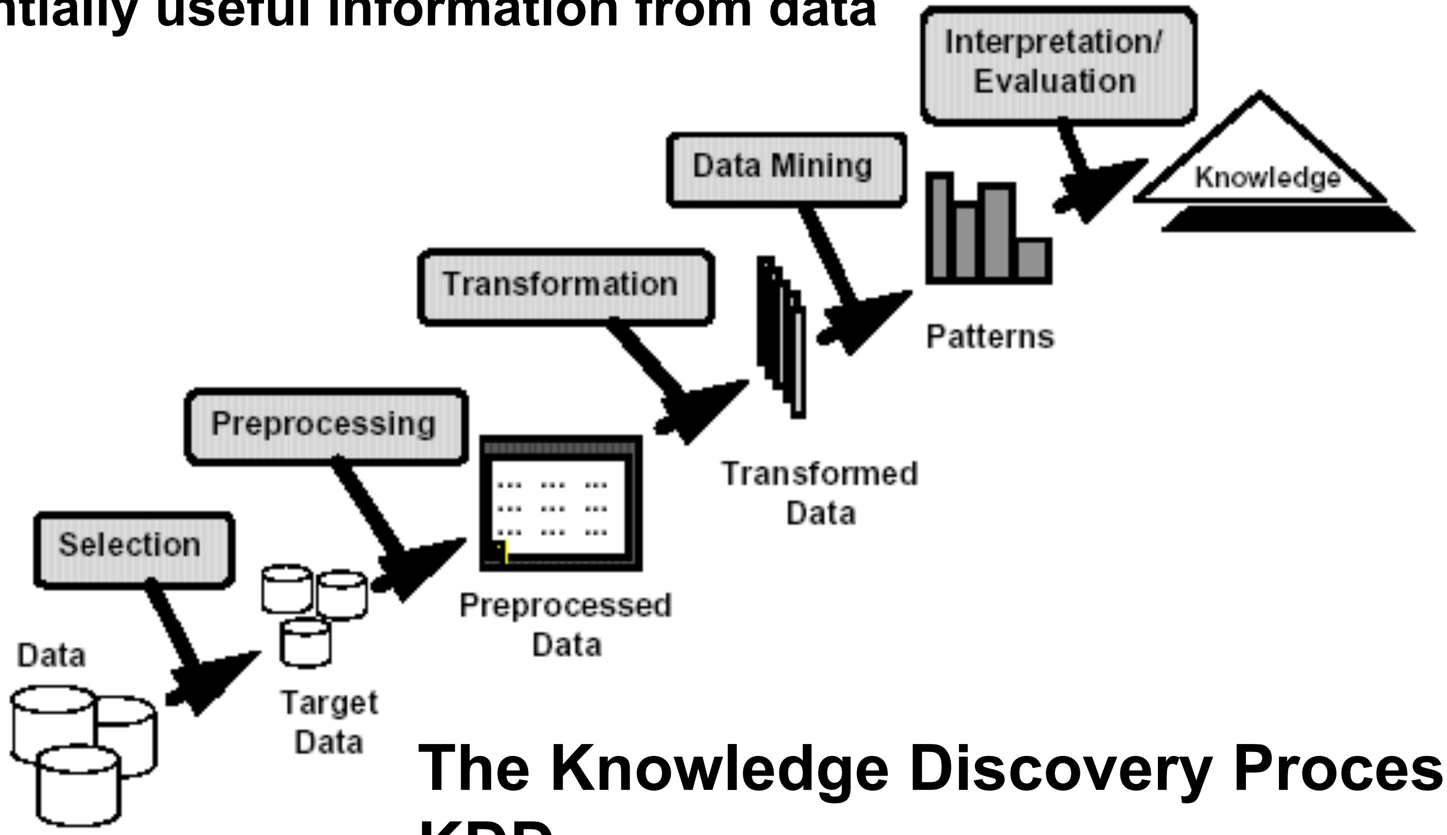


From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”

WHAT IS DATA MINING?

Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

Non-trivial extraction of implicit, previously unknown and potentially useful information from data



The Knowledge Discovery Process - KDD

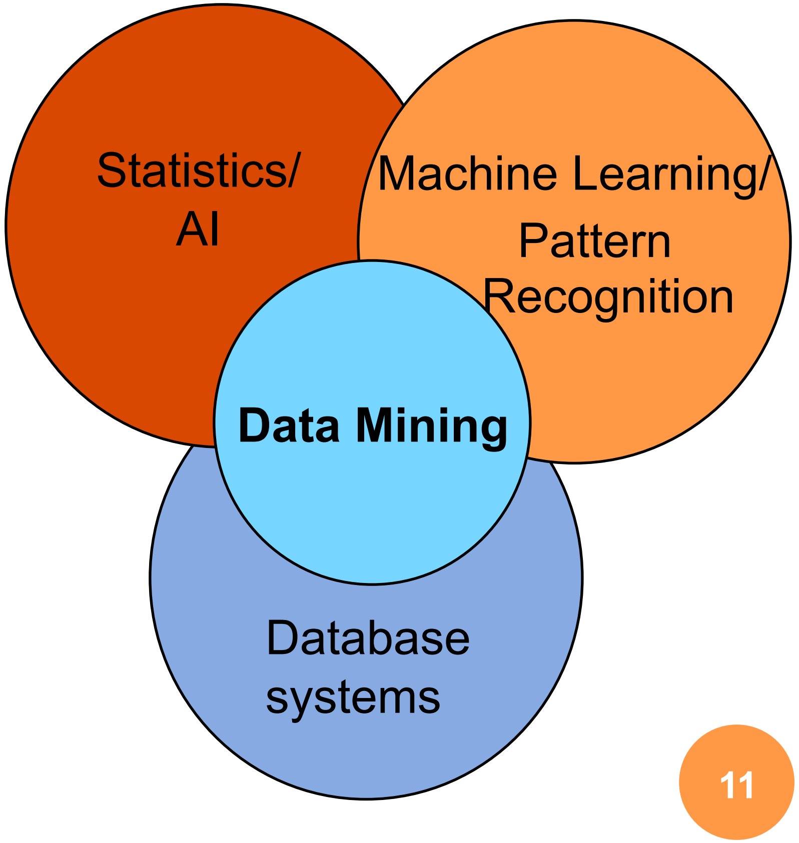
WHAT IS (NOT) DATA MINING?

- **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**

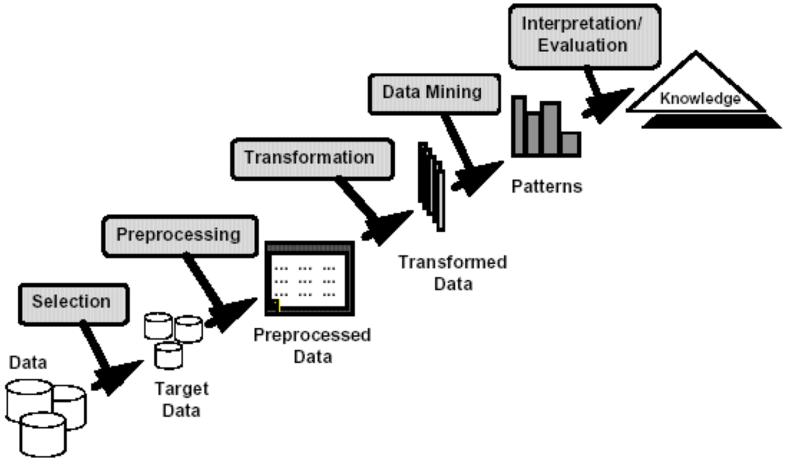


DATA MINING TASKS

- 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]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]



Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data
- ….