Data Mining and Pattern Recognition for Large-Scale Scientific Data

Chandrika Kamath

Center for Applied Scientific Computing

Lawrence Livermore National Laboratory

October 15, 1998

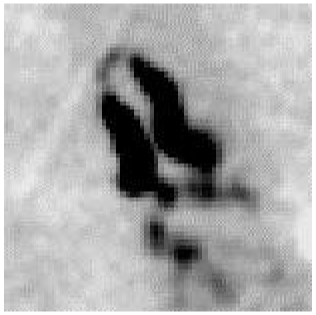
We need an effective way to deal with data overload

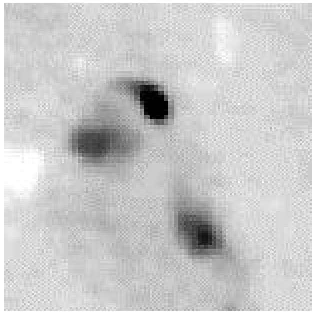
- **Widening gap between data collection capabilities and data analysis abilities**
  - **Data from simulations, experiments, observations**
  - **Terabytes of data, soon to be petabytes**
  - **Complex data (images, time series data)**
- **Manual exploration and analysis is impractical**
  - **Poor utilization of resources**
  - **Potential loss of information**

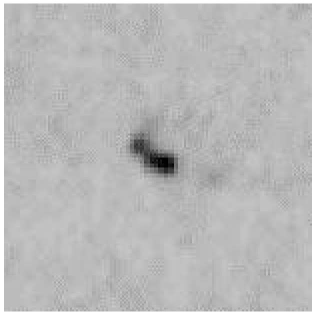
➡ **Need computational tools and techniques to automate the exploration and analysis of large, complex data sets**

What do we mean by the terms Data Mining and Pattern Recognition?

- **Data Mining: Uncovering patterns, associations, anomalies, and statistically significant structures in data**
- **Pattern Recognition: Characterization of patterns in data**
- **Pattern: Arrangement or ordering with an underlying structure**
- **Feature: An extractable measurement or attribute**

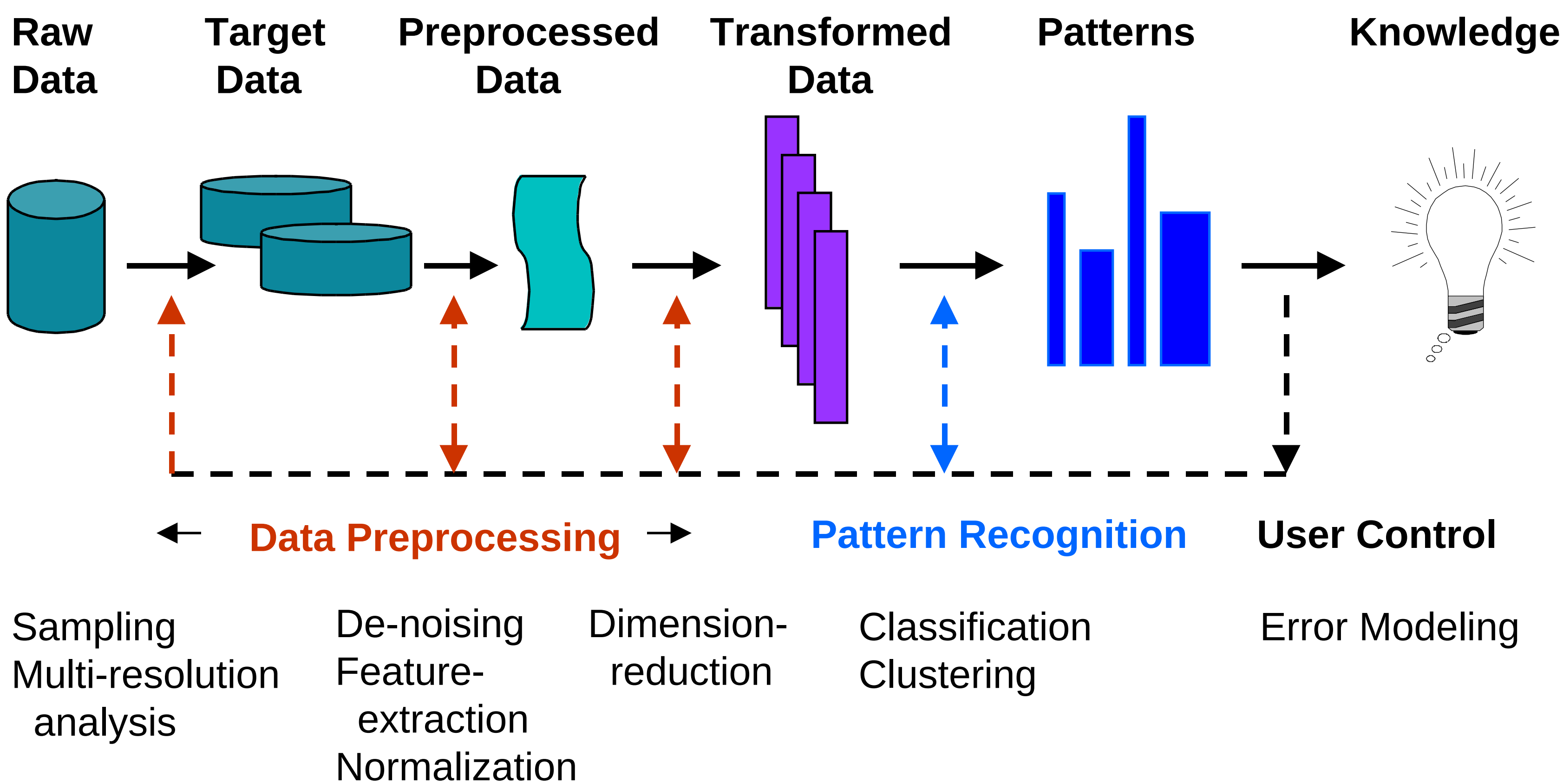






Images of Radio Emitting Galaxies with Bent-Double Morphology

Data Mining: Key steps in an iterative and interactive process

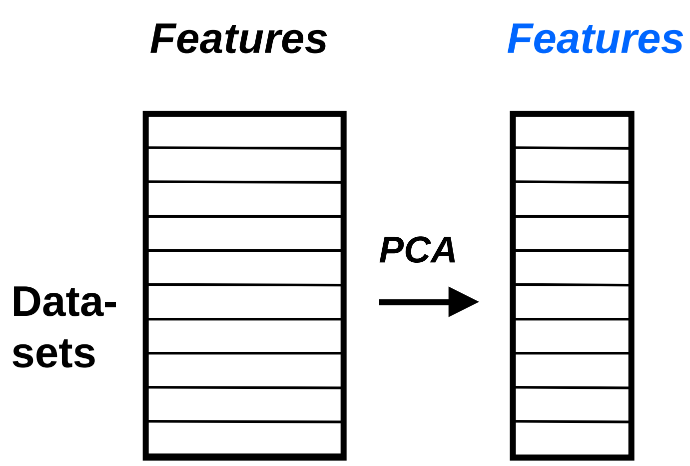


Our research plan for scaling data mining to large and complex data sets

- **Data pre-processing**
  - Implement effective image processing algorithms
  - Investigate the use of multi-resolution analysis
  - Research methods for dimension reduction
- **Pattern recognition algorithms**
  - Consider different algorithms for an application
  - Implement in an object-oriented framework
  - Research ways of making them more effective and efficient
  - Examine accuracy versus computational effort issues
- **Parallel implementation**

Data pre-processing: a time-consuming but critical first step

- **Extraction of features**: image processing and wavelets
  - De-noising
  - Multi-resolution analysis
- **Dimension reduction**: identification of key features
  - Features with greatest variance
  - Principal component analysis

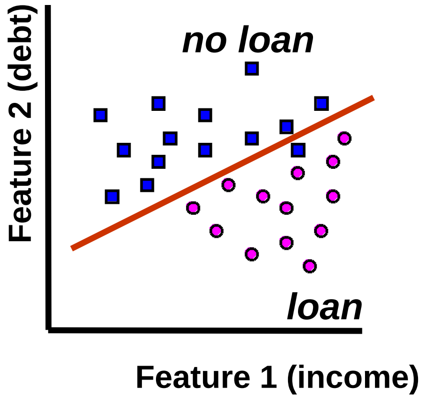


$$A = U \Sigma V^T$$

Pattern Recognition: need for scalable classification and clustering algorithms

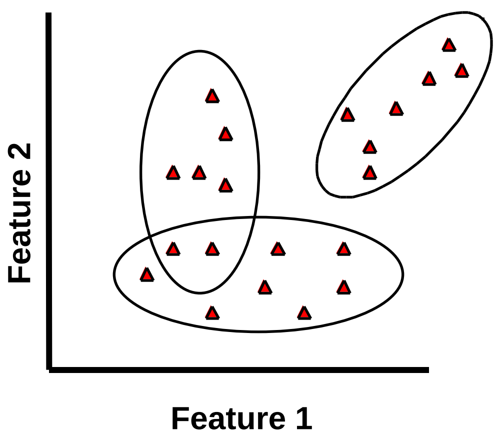
Classification: learn a function to map a data item into one of several predefined classes

- Neural networks
  - Genetic algorithms
  - Simulated annealing



Clustering: a task that identifies a finite set of clusters to describe the data

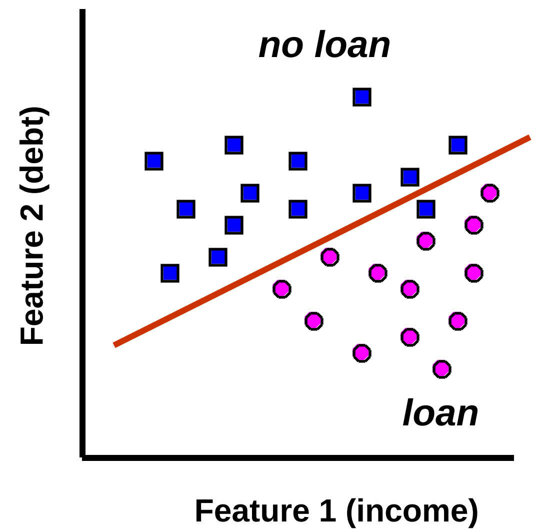
- Graph theoretic techniques
  - Hypergraph partitioning
  - Promising for high dimensional data



Pattern Recognition: need for efficient, accurate, and scalable classifiers

Classification: learning a function that maps a data item into one of several pre-defined classes

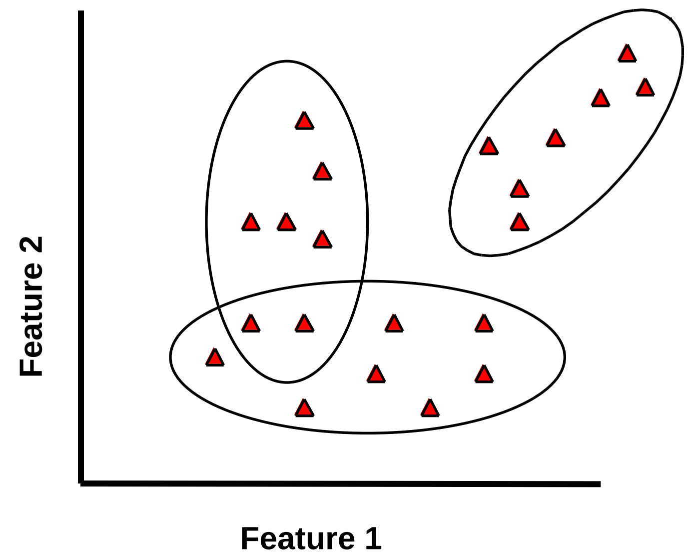
- **Neural networks: avoid local minima**
  - **Genetic algorithms**
  - **Simulated annealing**
- **Decision trees**
  - **attribute selection**
  - **tree pruning**
- **Hybrid algorithms**
  - **techniques for combining classifiers**



Pattern Recognition: need for scalable and interpretable clustering algorithms

Clustering: a descriptive task that seeks to identify a finite set of clusters to describe the data

- **Implement known techniques**
  - k-means
  - fuzzy k-means
  - k-nearest-neighbors
- **Graph theoretic techniques**
  - hypergraph partitioning
  - initial promise for high dimensional data



Large-scale pattern recognition can benefit several applications

- Visualization
- Computational steering
- Computer Security
- Verification and validation
- Global climate modeling
- Astrophysics (MACHO and FIRST)
- ...

➡ **A capability for large-scale pattern recognition will strengthen our ability to perform science by providing an effective way to cope with data overload.**

Sapphire: a multi-disciplinary endeavor

- Core Team (CASC)
  - C. Kamath (PI), C. Baldwin, R. Musick
- Collaborators
  - C. Alcock (IGPP), M. Aufderheide (B Division), R. Becker (UC Davis/LLNL)
- Faculty and students
  - G. Bebis, M. Giamporcaro, R. Karchin, I. Kirby

➡ For more information

http://www.llnl.gov/CASC/sapphire

kamath2@llnl.gov

UCRL-MI-132151: This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48