

# Data Challenges for Data Mining

---

Ramamohanarao Kotagiri

Professor of Computer Science

Department of Computer Science and  
Software Engineering

The University of Melbourne

Australia

# Data Challenges for Data Mining

---

KDD 2006 panel proposes the following criteria for a good grand challenge problem for data mining.

- The problem is hard -- very difficult to solve given the current state of the art
  - Involves data mining: data mining plays an important role in solving the problem
  - Based on a large, publicly available data set
  - There is a specific goal: it is clear when the problem is solved
  - Problem is interesting to researchers and understandable to the public, and preferably stated in one sentence.
  - There is significant public benefit if it is solved.
-

# Data Challenges for Data Mining

---

Grand challenge in most real data have the following issues

- Unhelpful statistical properties
  - Complex data structures
  - Lack of parametric modelling
  - Scaling issues
  - Validation methods
  - Lack of domain specific knowledge
  - Combining domain specific knowledge with data mining
-

# Data Challenges for Data Mining

---

## Unhelpful Statistical Properties

- Sparse data
  - Not representative
  - Uncertain
  - Noisy
  - Missing values
  - Highly imbalanced
  - Very high dimensional
  - Few data points
-

# Data Challenges for Data Mining

---

## Complex data structures

- Relational
  - Unstructured (e.g. Text)
  - Semistructred (e.g. XML)
  - Graphs (e.g. Protein-Protein interaction)
  - Spatial (e.g. GIS)
  - 3D (e.g. Tertiary structure of proteins)
  - Video and Audio
  - Time series
  - Sequences
  - Other
-

# Data Challenges for Data Mining

---

Lack of parametric modelling techniques

- What is the best model to use?
  - How to estimate the parameters?
  - What are the suitable similarity functions?
-

# Data Challenges for Data Mining

---

## Scaling issues

- Can we handle petabyte scale data efficiently?
  - We may need data base type technology for data mining!
  - Can we do ad hoc data mining queries on such large scale data?
  - What are the best sampling techniques for such large scale data?
  - How do we parallelise algorithms for such ad hoc data mining queries?
  - Computational complexities
  - Dimensional reduction
-

# Data Challenges for Data Mining

---

## Validation methods

- How do we validate discovered patterns/knowledge?
  - Can the validation be done in automatically?
  - Can we have validation techniques which are largely domain independent and useful?
-

# Data Challenges for Data Mining

---

Lack of domain specific knowledge

- How a DM can work in different domains?
  - Role of DM in terms recognition by their interdisciplinary researchers
-

# Data Challenges for Data Mining

---

Combining domain specific knowledge with data mining

- How can we easily incorporate domain specific knowledge into data mining techniques
  - How can we incrementally update domain specific knowledge into data mining techniques?
-