Data Analytics 1

Vikram Pudi
vikram@iiit.ac.in
IIIT Hyderabad

Data Systems Evolution

- Traditional Database Systems
  - Indexing
  - Query languages
  - Query optimization
  - Transaction processing
  - Recovery …
- Relational / SQL

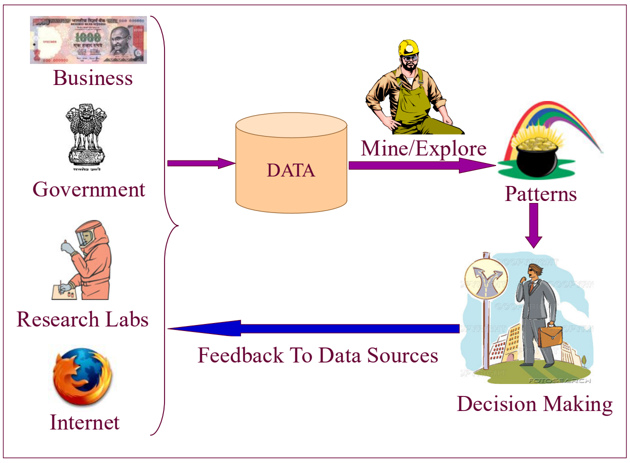
Post-Relational Revolution

- New organizations of data
  - Object oriented (OO) [Zope] and object-relational (OR) systems [SQLAlchemy]
  - Semi-structured [XML, JSON] and unstructured data (Text)
  - Vertical / Column stores [Cassandra]
  - Unnormalized relations: Document Databases [MongoDB]
  - Key-value Stores [Redis]
  - Graph Databases [Neo4j]
- New functionality
  - Distribution & Heterogeneity (multi-databases, interoperability)
  - Active databases (triggers) and deduction
  - ERP packages (application-oriented tasks common to many organizations)
  - ***Data analysis*** (data warehouses and ***data mining***)
- More complex data domains (e.g., design, geography, molecular biology, social networks)
- Relaxation of ACID test for DBMS

Data Mining

= Automated discovery of *interesting patterns* in large datasets

- Researchers identified several kinds of interesting patterns in an adhoc manner
  - classification and regression models, clusters, association rules, frequent patterns, sequential patterns, time-series patterns, summaries, cyclic patterns, hierarchical patterns, max-patterns, closed patterns, multi-dimensional patterns, etc.



Business

Government

Research Labs

Internet

DATA

Mine/Explore

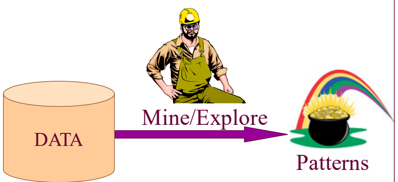
Patterns

Decision Making

Feedback To Data Sources

Data Science / KDD Life-cycle

- Domain understanding
- Data Preprocessing
  - Data integration
  - Cleaning
  - Selection
  - Transformation
- *Data Mining*
- Post Mining
  - Presentation / Visualization
  - Evaluation
  - Decision making



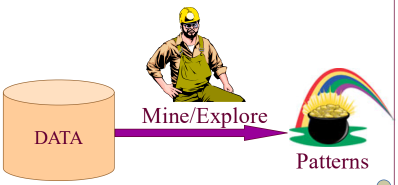
DATA

Mine/Explore

Patterns

Data Science / KDD Life-cycle

- Domain understanding
- Data Preprocessing
  - Data integration
  - Cleaning
  - Selection
  - Transformation
- *Data Mining*
- Post Mining
  - Presentation / Visualization
  - Evaluation
  - Decision making



Many knowledge discovery applications need *exact, interpretable* knowledge for decision making

Types of Patterns

- Associations
  - *Coffee* buyers usually also purchase *sugar*
- Clustering
  - Segments of customers requiring different promotion strategies
- Classification
  - Customers expected to be *loyal*

Take Home

- Data mining is a mature field
- Don't waste time developing new algorithms for core tasks
- Focus on applications to challenging kinds of data
  - Streams, Distributed data, Multimedia, Web, …
- Most effort is in how to map domain problems to data mining problems
- And how to make sense of the output.

Grading Plan (Tentative)

Normal Semester	Online Semester
10% Assignments	20% Assignments
30% Mid	20% Quizzes
25% Endsem	25% Endsem
30% Projects	35% Projects