CSE 591 Data Mining

Data Mining, Data Preparation & Web Mining

New Room: LL271

Huan Liu, CSE, CEAS, ASU

http://www.public.asu.edu/~hliu/cse591.html

CSE 591

- Contents

  Classification, Clustering, Association, Data Warehousing, Web, and Applications

- Format - *A seminar course*

  Paper reading, discussion, project, presentation

- Assessment

  Class participation, project proposal, presentation, exams

Course Format

- Research papers - the main source to be found on the course web site
- You can choose one of the textbooks listed. A reference list is an entering point for you to access related subjects
- Everyone is expected to read the papers and participate in class discussion
- Presenters will be evaluated on the spot

Paper presentation

- Each student will be responsible for one topic. All are expected to read the material(s) before the presentation.
  - What is it about?
  - What are points to discuss and improve?
  - What can we do with it?
- Each presentation is about 35 minutes including discussion, question & answer

Project

- Proposal
  - Proposal presentation, discussion, revision
  - A project should be completed in a semester
- Project
  - Presentation and demo
- Report

Topic Distribution (tentative)

Topics	Classes
Introduction	2
Classification	4
Evaluation	2
Pre-processing	2
Clustering	4
Association	4
Web data (XML, RDF), Mining	4
Project related	4
Real-World Application	2
Data Warehousing	2

Your first assignment

- Think about what you want to accomplish.
- Pick an area of interest and choose a general topic for presentation.
- Registered students: send me an email with CSE591 in the subject (use your frequently used email account so you won't miss important announcement) with your areas of interests.
- Complete the above before the 2nd class.

Introduction

- The need for data mining
- Data mining
- Data warehousing
- Web mining
- Applications

What is data mining

- Data mining is
  - extraction of useful patterns from data sources, e.g., databases, texts, web, image.
  - 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.

Patterns (1)

- Patterns are the relationships and summaries derived through a data mining exercise.
- Patterns must be:
  - valid
  - novel
  - potentially useful
  - understandable

Patterns (2)

- Patterns are used for
  - prediction or classification
  - describing the existing data
  - segmenting the data (e.g., the market)
  - profiling the data (e.g., your customers)
  - etc.

Data (1)

- Data mining typically deals with data that have already been collected for some purpose other than data mining.
- Data miners usually have no influence on data collection strategies.
- Large bodies of data cause new problems: representation, storage, retrieval, analysis, ...

Data (2)

- Even with a very large data set, we are usually faced with just a sample from the population.
- Data exist in many types (continuous, nominal) and forms (credit card usage records, supermarket transactions, government statistics, text, images, medical records, human genome databases, molecular databases).

Some DM tasks

- Classification:

  mining patterns that can classify future data into known classes.

- Association rule mining

  mining any rule of the form $X \rightarrow Y$, where X and Y are sets of data items.

- Clustering

  identifying a set of similarity groups in the data

- Sequential pattern mining:

  A sequential rule: $A \rightarrow B$, says that event A will be immediately followed by event B with a certain confidence

- Deviation detection:

  discovering the most significant changes in data

- Data visualization: using graphical methods to show patterns in data.

Why data mining

- Rapid computerization of businesses produces huge amounts of data
- How to make best use of data?
- A growing realization: knowledge discovered from data can be used for competitive advantage.

- Make use of your data assets
- Many interesting things you want to find cannot be found using database queries
  - "find me people likely to buy my products"
  - "Who are likely to respond to my promotion"
- Fast identify underlying relationships and respond to emerging opportunities

Why now

- The data is abundant.
- The data is being warehoused.
- The computing power is affordable.
- The competitive pressure is strong.
- Data mining tools have become available.

DM fields

- Data mining is an emerging multi-disciplinary field:
  - Statistics
  - Machine learning
  - Databases
  - Visualization
  - OLAP and data warehousing
  - …

Summary

- What is data mining?

  KDD - knowledge discovery in databases: non-trivial extraction of implicit, previously unknown and potentially useful information

- Why do we need data mining?

  Wide use of computer systems - data explosion - knowledge is power - but we're data rich, knowledge lean - actionability ...

Data Warehousing

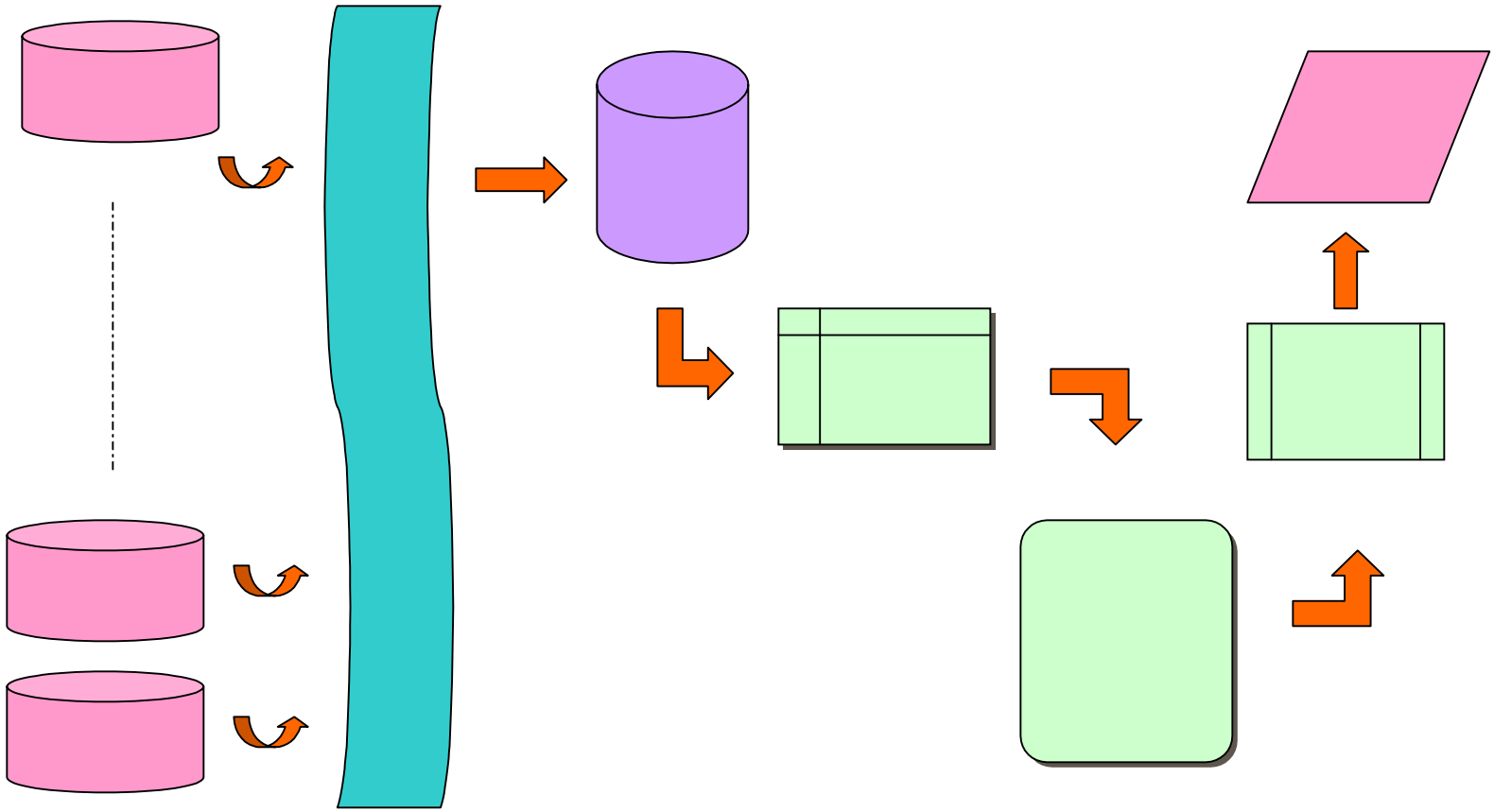
- What is a data warehouse?

  A repository of integrated, analysis-oriented, historical, read-only data, designed for decision support and KDD systems

- Why do we need data warehousing?

  Operational systems were never designed for KDD, they are numerous, of different types, with overlapping/contrary definitions

An Overview of KDD Process (Guess which is which)



Web mining

- The Web is a massive database
- Semi-structured data
- XML and RDF
- Web mining
  - Content
  - Structure
  - Usage