

Knowledge Discovery and Data Mining

Unit # 17

DM Courses Available Online

- MIT
 - <http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/lecture-notes/>
- Cleveland State University Ohio
 - <http://academic.csuohio.edu/fuy/EEC%20525/syllabus.html>
- Virginia Tech
 - <http://people.cs.vt.edu/~ramakris/Courses/CS6604/lectures.html>

DM Courses Available Online (Con'td)

- Worcester Polytechnic Institute
 - http://web.cs.wpi.edu/~cs525d/f09/schedule_cs525_f09.html
- Claremont Graduate School
 - <http://web.cgu.edu/Faculty/leroy/content/Courses/2010-Fall-DM/Syllabus-Fall-2010-DM.html>
- Central Connecticut State University
 - http://www.cs.ccsu.edu/~markov/ccsu_courses/580Syllabus.html
- Temple University
 - http://www.cs.ccsu.edu/~markov/ccsu_courses/580Syllabus.html

Course Objectives

- Know the knowledge discovery process
- Understand the different categories of algorithms
- Be able to judge which algorithms fit different problems
- Have practical experience choosing algorithms for a specific problem
- Have practical experience working in technical teams
- Have practical experience executing data mining projects
- Have practical experience using open source data mining software

Course Outline

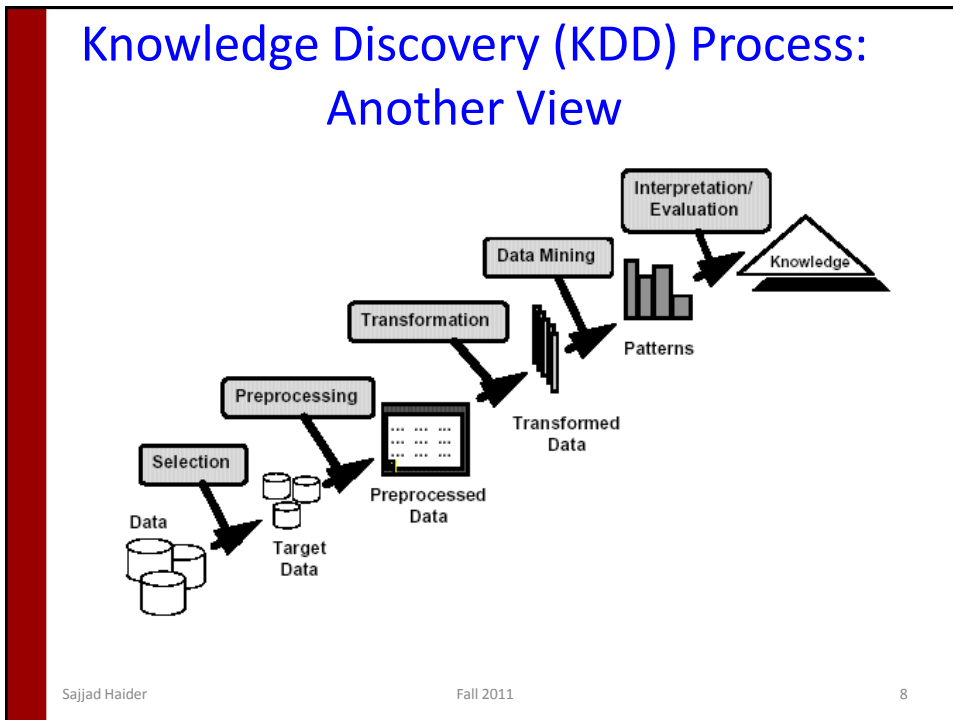
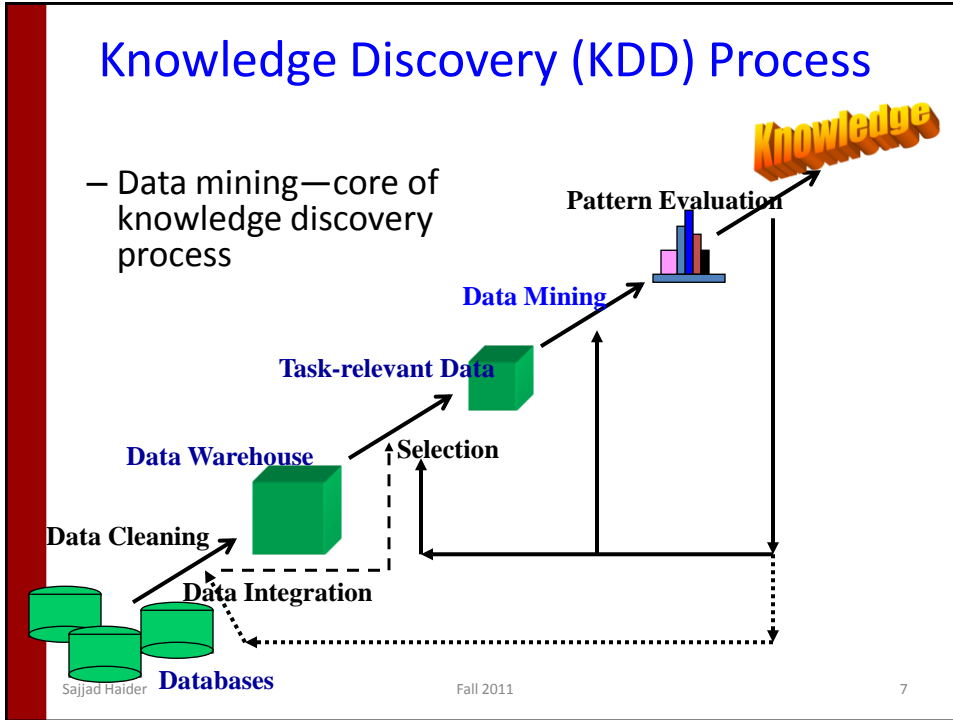
- Data Preparation (Cleansing, Normalization, Transformation)
- Classification
 - Application of different algorithms/techniques
 - Feature Selection/Dimension Reduction
 - Model Evaluation
- Clustering
 - Various Algorithms
 - Model Evaluation (External vs. Internal metrics)
- Association Rules

What Is Data Mining?

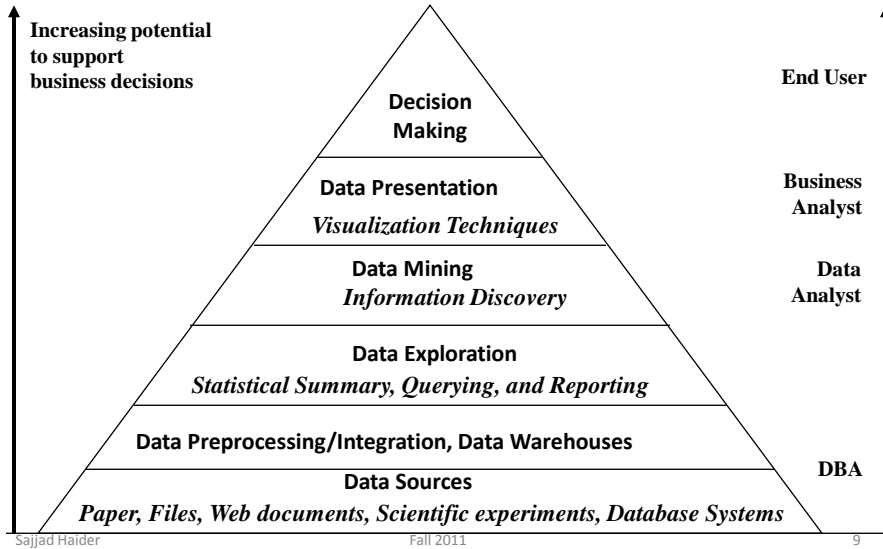


- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
 - Simple search and query processing
 - (Deductive) expert systems



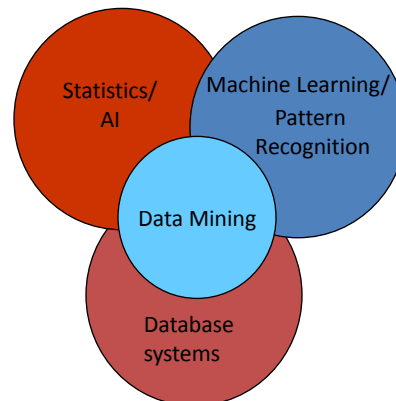


Data Mining and Business Intelligence



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



Are All the “Discovered” Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
 - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
 - A pattern is *interesting* if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
 - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
 - Subjective: based on user’s belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Classification: Application 1

- Direct Marketing
 - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
 - Approach:
 - Use the data for a similar product introduced before.
 - We know which customers decided to buy and which decided otherwise. This *{buy, don’t buy}* decision forms the *class attribute*.
 - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
 - Type of business, where they stay, how much they earn, etc.
 - Use this information as input attributes to learn a classifier model.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

- Fraud Detection
 - Goal: Predict fraudulent cases in credit card transactions.
 - Approach:
 - Use credit card transactions and the information on its account-holder as attributes.
 - When does a customer buy, what does he buy, how often he pays on time, etc
 - Label past transactions as fraud or fair transactions. This forms the class attribute.
 - Learn a model for the class of the transactions.
 - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
 - Goal: To predict whether a customer is likely to be lost to a competitor.
 - Approach:
 - Use detailed record of transactions with each of the past and present customers, to find attributes.
 - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
 - Label the customers as loyal or disloyal.
 - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

Clustering: Application 1

- Market Segmentation:
 - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
 - Approach:
 - Collect different attributes of customers based on their geographical and lifestyle related information.
 - Find clusters of similar customers.
 - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
 - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
 - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
 - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
 - Let the rule discovered be
{Bagels, ... } --> {Potato Chips}
 - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
 - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
 - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
 - Goal: To identify items that are bought together by sufficiently many customers.
 - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
 - A classic rule --
 - If a customer buys diaper and milk, then he is very likely to buy beer.
 - So, don't be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

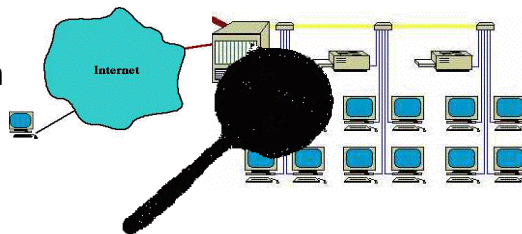
- Inventory Management:
 - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
 - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
 - Credit Card Fraud Detection



- Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day