OntarioTech Science

Kourosh Davoudi
kourosh@uoit.ca

Week 1: Introduction

CSCI 4150U: Data Mining

Welcome to Data Mining !

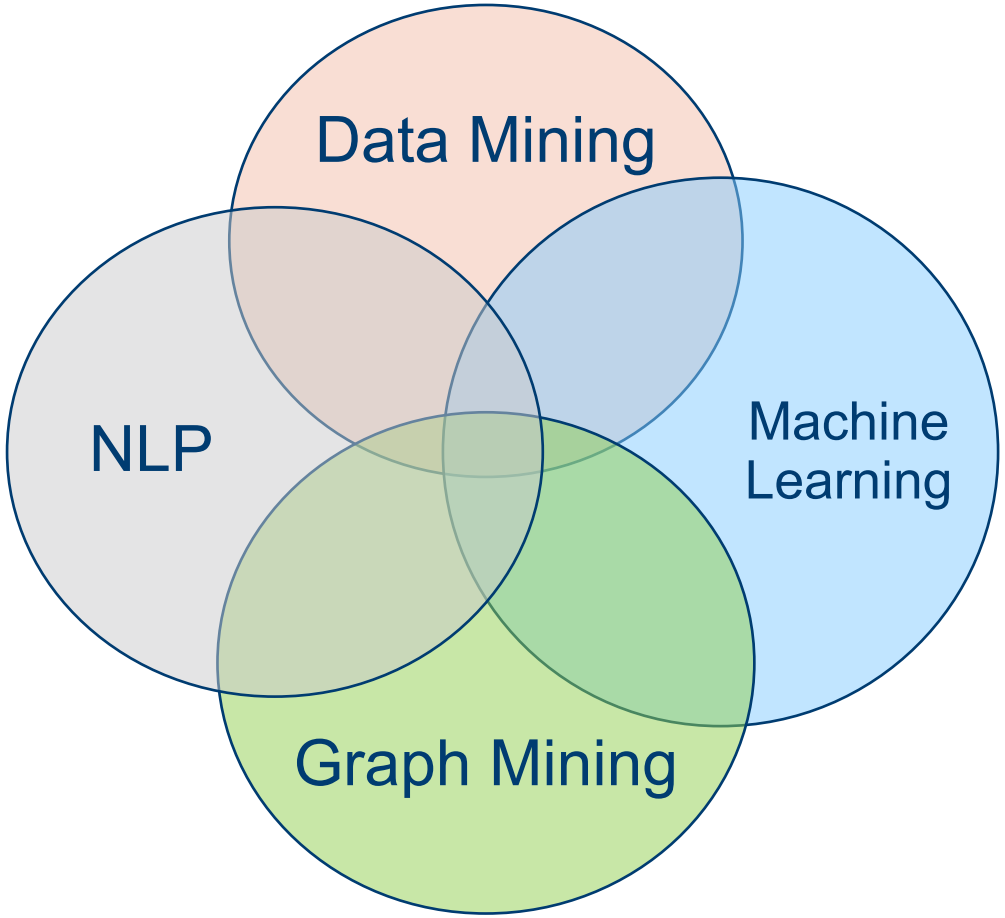
- What we learn this week:
  - Course Description
    - Structure
    - Goal
    - Content
  - Introduction to Data Mining
  - Data

Kourosh Davoudi

- Assistant Professor in Computer Science (Ontario Tech University)
- Postdoctoral Fellowship : (University of Waterloo)
- PhD: Computer Science (York University)
- Previous business partners:

Kourosh Davoudi

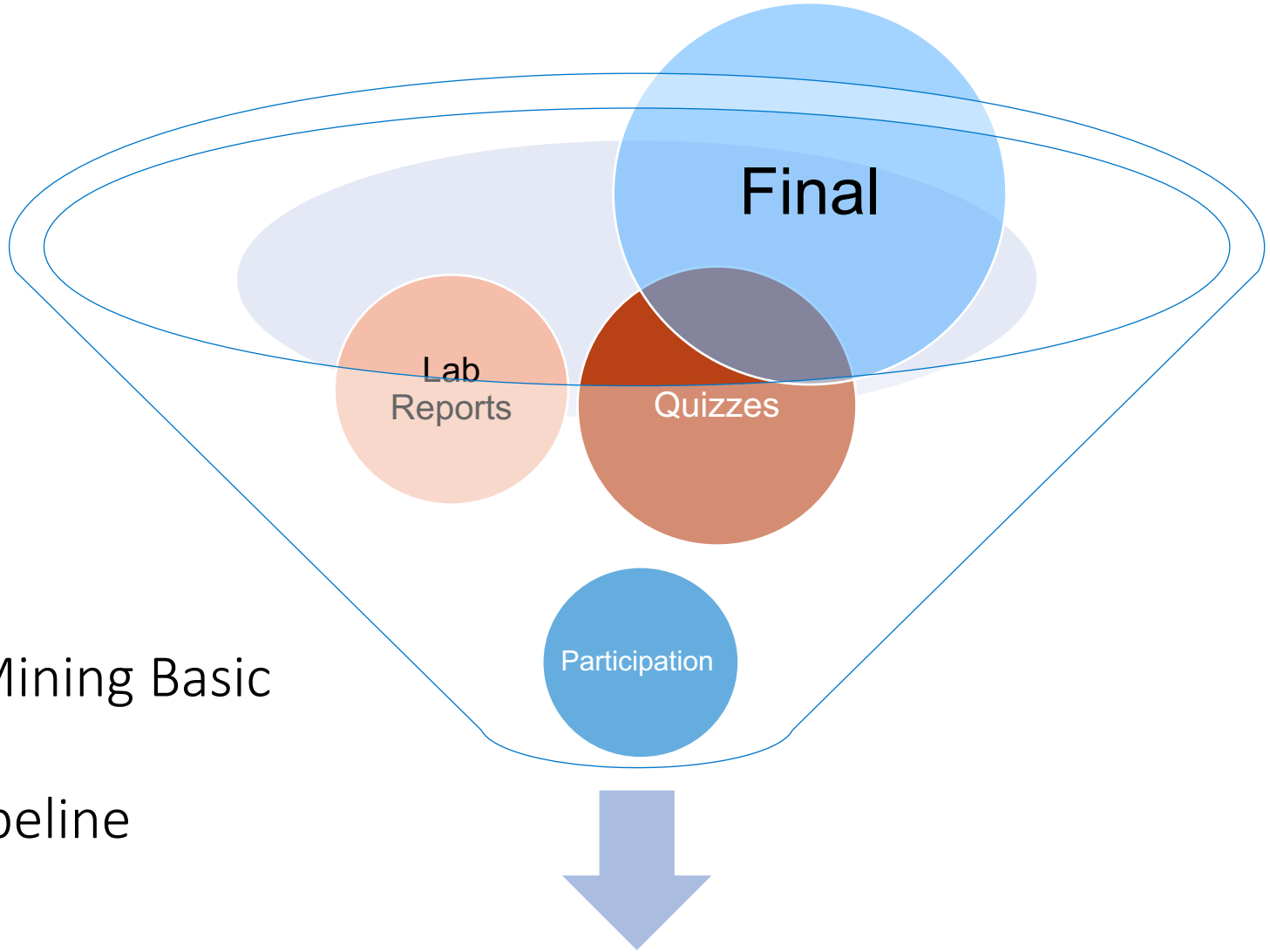
- Area of interest:



How about you?

- How do you like your major?
- What is your favorite course?
- Which jobs in computer science are you interested in?
- What do you expect from this course?
- Which programming languages have you work with?
- …

Course Structure



Course Outcomes

- Understanding the Data Mining Basic Concepts
- Develop a Data Mining Pipeline

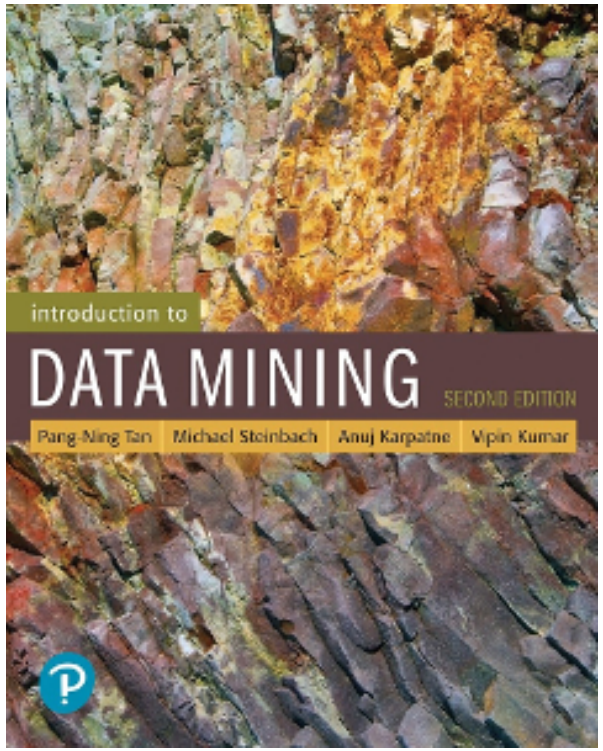
Course Content (Subject to Change)

1. Introduction to data mining
2. Data
3. Data Exploratory Analysis
4. Classification I (Basic Techniques)
5. Classification II (Alternative Techniques)
6. Clustering I (Basic Concepts and Techniques)
7. Clustering II (Advanced Concepts and Algorithms)
8. Anomaly Detection
9. Association Rule Mining

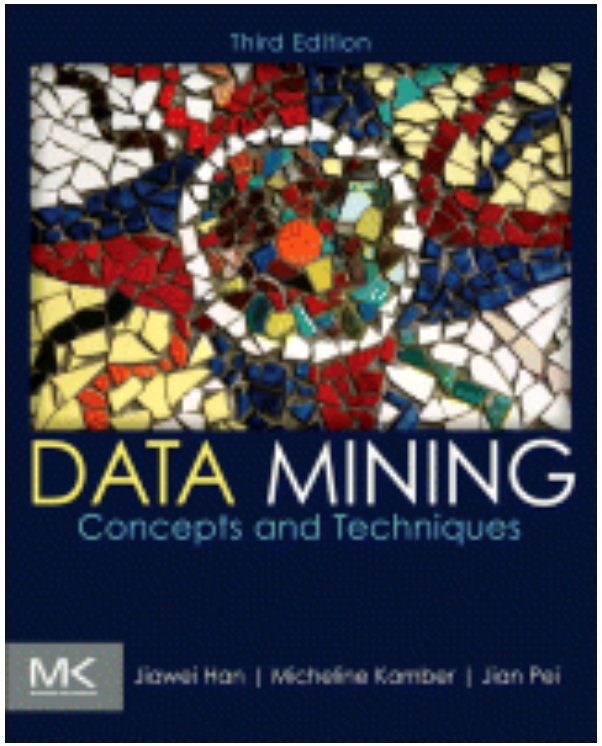
Evaluation Components

Component	Due Date	Weight
Class Activities and Participation		5 %
Quiz I	Feb 8, 2021 (class time)	15 %
Midterm Lab Report I	Feb 26, 2021, 11:59 PM	15 %
Quiz II	Mar 8, 2021 (class time)	15 %
Final Lab Report II	Apr 9, 2021, 11:59 PM	15 %
Final Exam	TBA by the university	35 %

Useful Textbooks



Introduction to Data Mining, 2'nd Edition
Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar

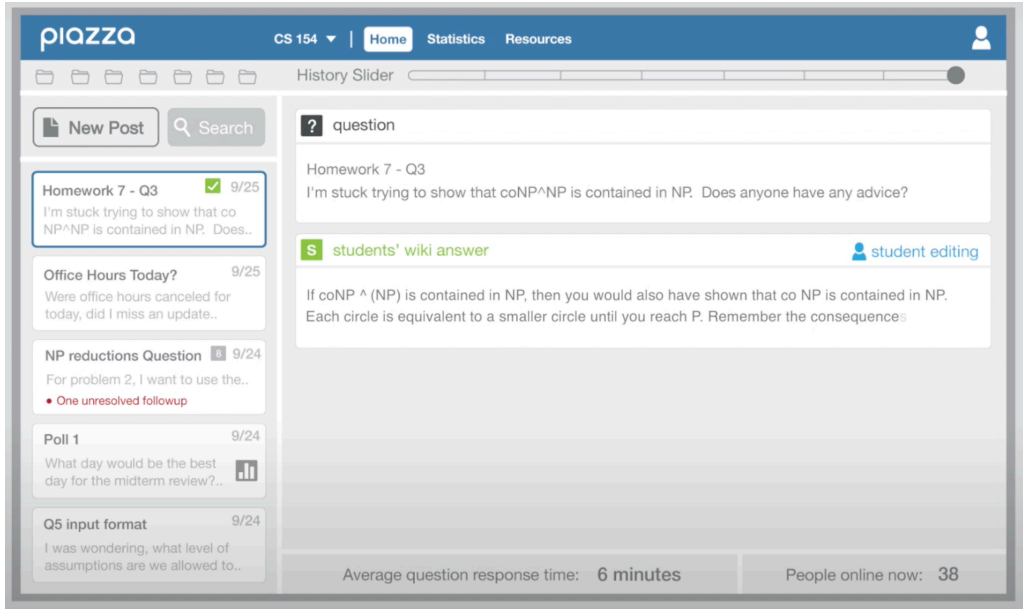


Data Mining: Concepts and Techniques
Jiawei Han, Micheline Kamber and Jian Pei

Communication

Piazza

- Please note that questions about lectures/assignments/exams should be posted to the Piazza



Office Hours and Contacts

Course Instructor:

Dr. Kourosh Davoudi

- **Email:** kourosh@uoit.ca (For official matters. Email subject should be: 4050U)
- **Office Location:** UA 2015
- **Office Hours:** TBA or by appointment (online)

- **Phone:** (905) 721-8668 x 2779
- **Webpage:** http://dmlab.science.uoit.ca/hdavoudi/

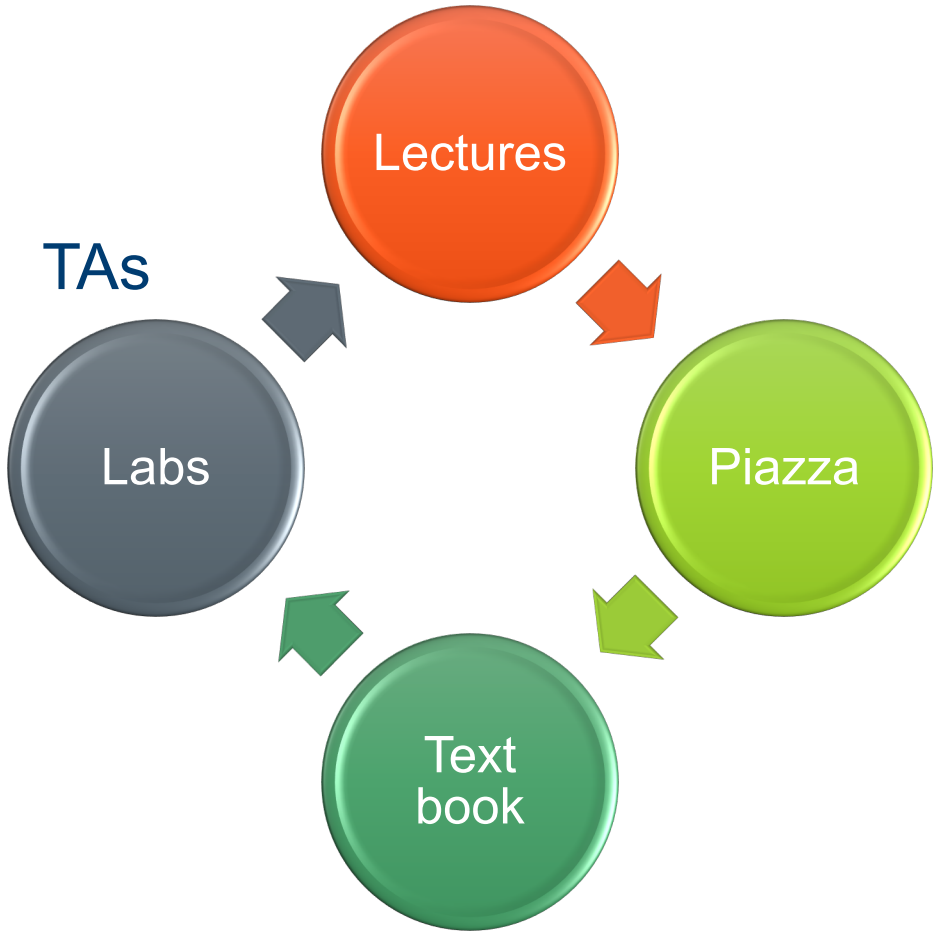
We have great TAs:

Marzieh Najafabadi: Marzieh.AhmadiNajafabadi@ontariotechu.ca

Aref Divshali: Aref.AbedjooyDivshali@ontariotechu.ca

Teaching Philosophy

Instructor
(facilitator)



Some suggestions/comments

- The lectures are fast or slow
- I am here only for a grade/requirement
- I feel that I need some background
- I need help due to pandemic issue
- I have some questions related to the labs

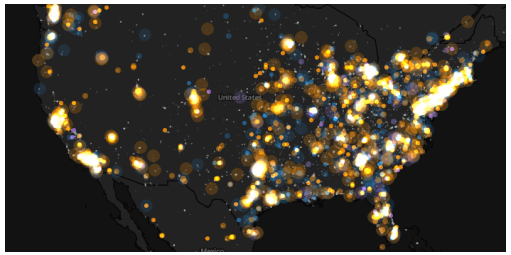
We are here and try our best to facilitate your learning process

Introduction to Data Mining

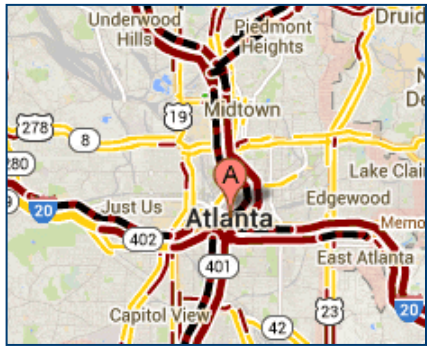
Outline and Learning Outcome

- Why data mining?
- What is data mining?
- Why not use classical data analysis?
- Know about origin of data mining
- Explain data mining tasks

Large-scale Data is Everywhere!



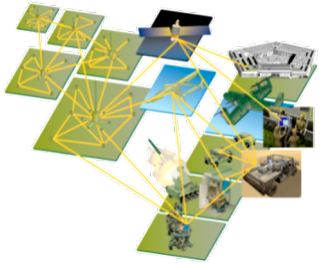
Social Networking: Twitter



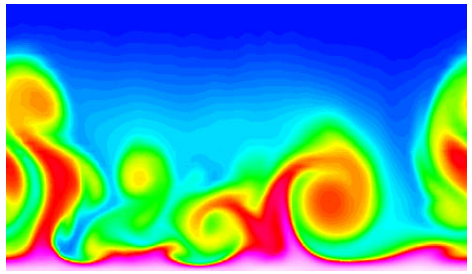
Traffic Patterns

E-Commerce

Cyber Security



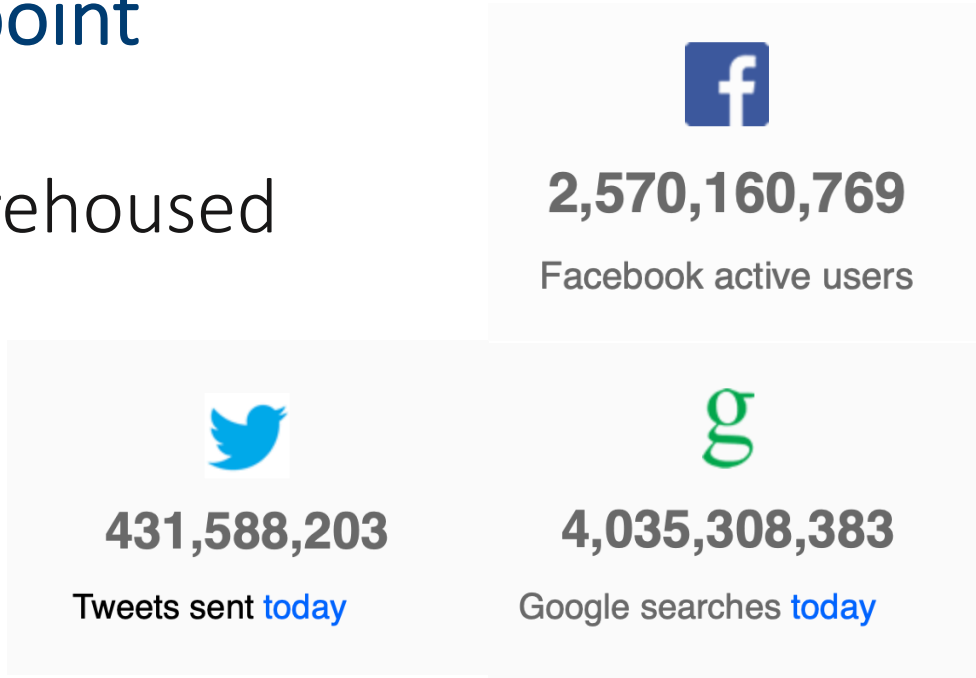
Sensor Networks



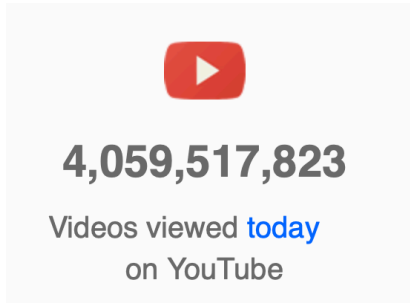
Computational Simulations

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
  - Purchases at department/ grocery stores, e-commerce
  - Bank/Credit Card transactions



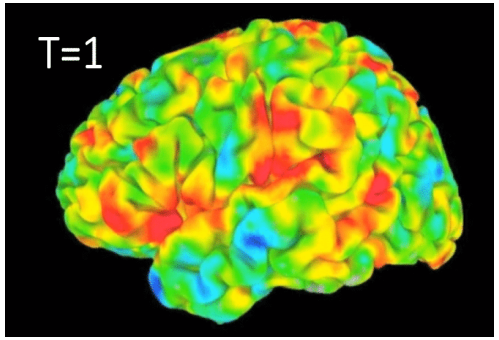
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services
    - e.g. in Customer Relationship Management



https://www.internetlivestats.com

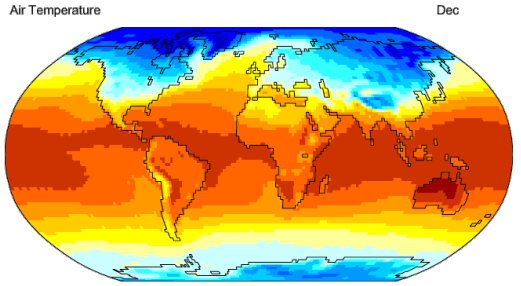
Why Data Mining? Scientific Viewpoint

- Data collected and stored at enormous speeds
  - remote sensors on a satellite
  - telescopes scanning the skies
  - high-throughput biological data
  - scientific simulations

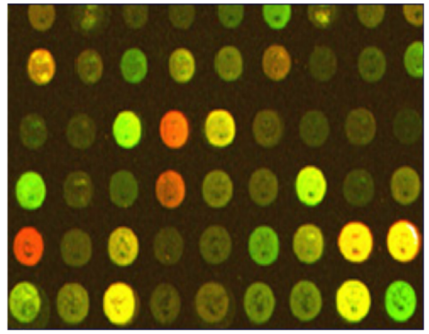


fMRI Data from Brain

Sky Survey Data



Surface Temperature of Earth

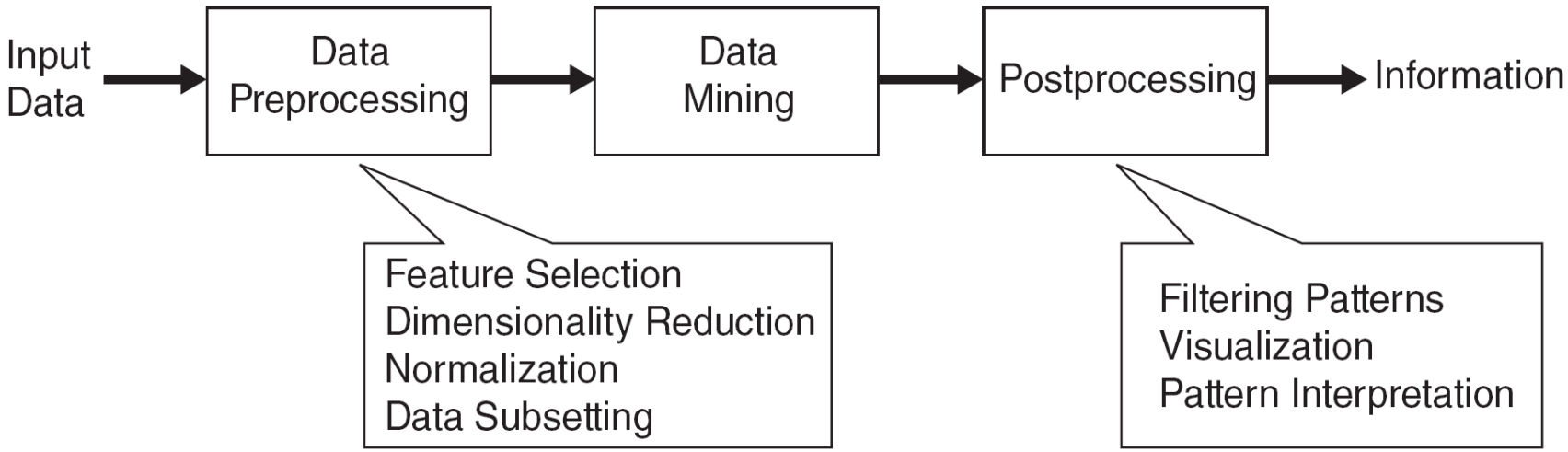


Gene Expression Data

- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



What is (not) Data Mining?

What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

What is Data Mining?

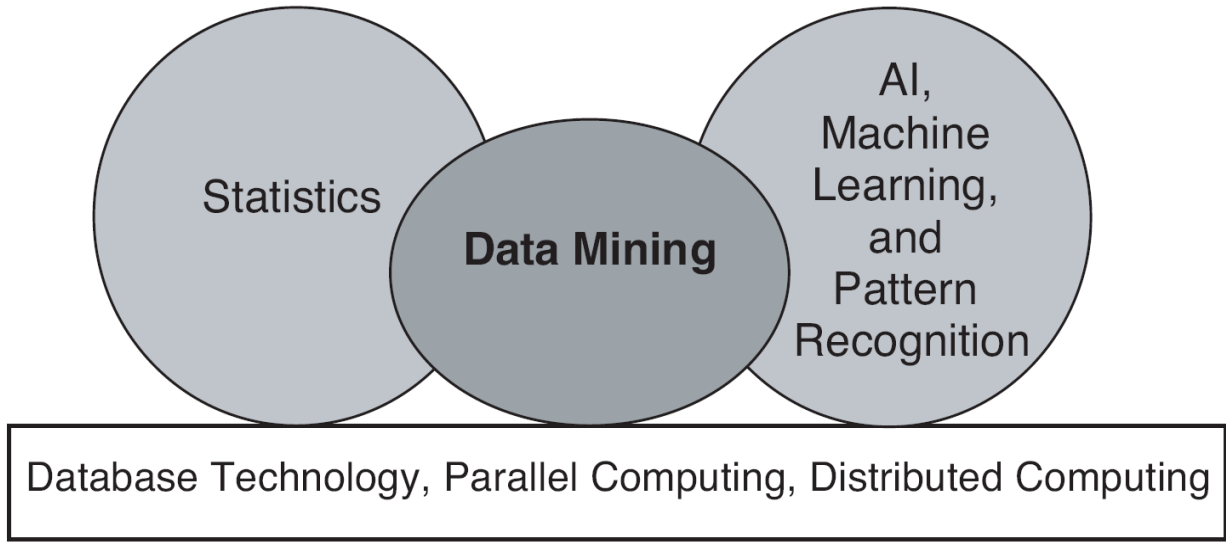
- Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly… in Boston area)
- Group together similar documents returned by search engine

Why not use classical data analysis?

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional Analysis

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed

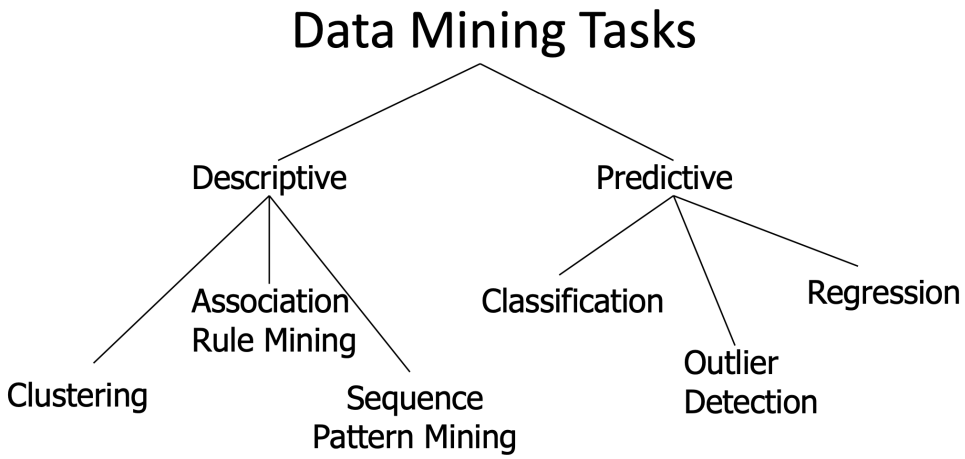


- A key component of the emerging field of data science and data-driven discovery

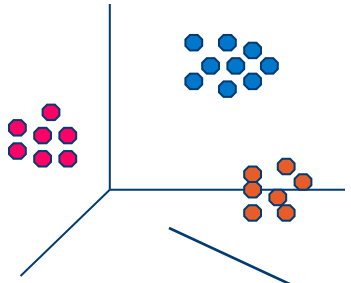
Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.

- Description Methods
  - Find human-interpretable patterns that describe the data.



Data Mining Tasks …

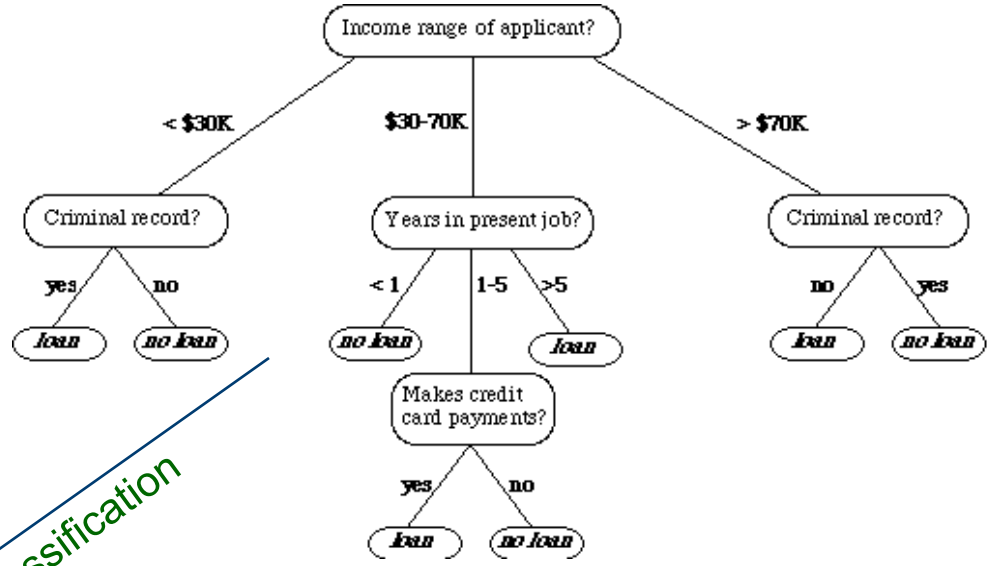


Clustering

Data

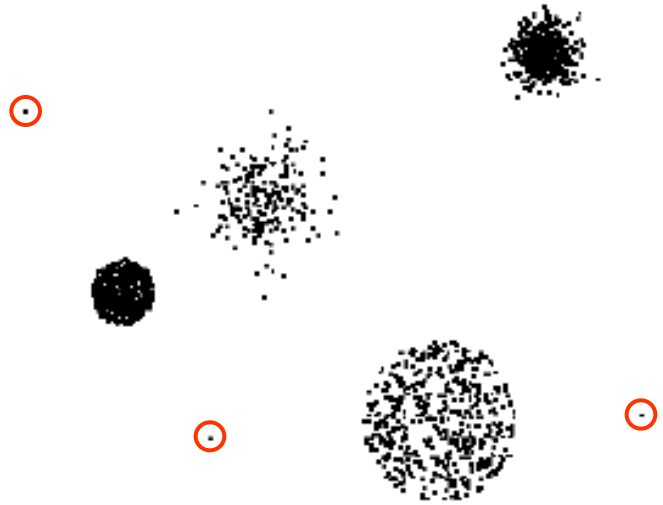
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

classification



Association Rules

Anomaly Detection

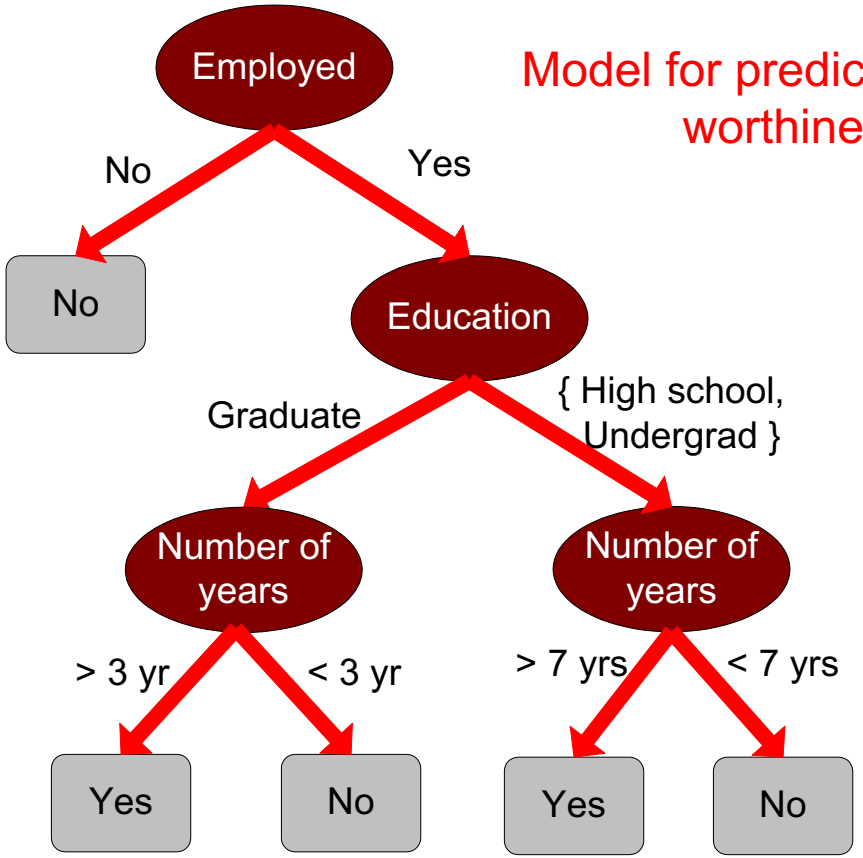


Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…



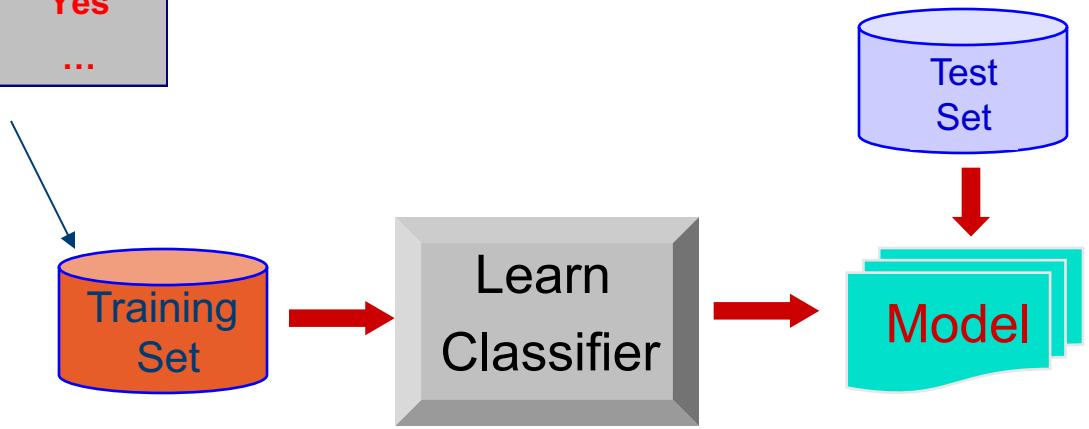
Model for predicting credit worthiness

Classification Example

categorical categorical quantitative class

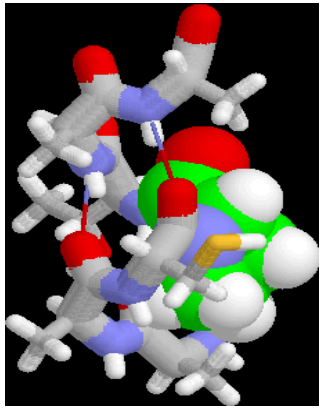
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

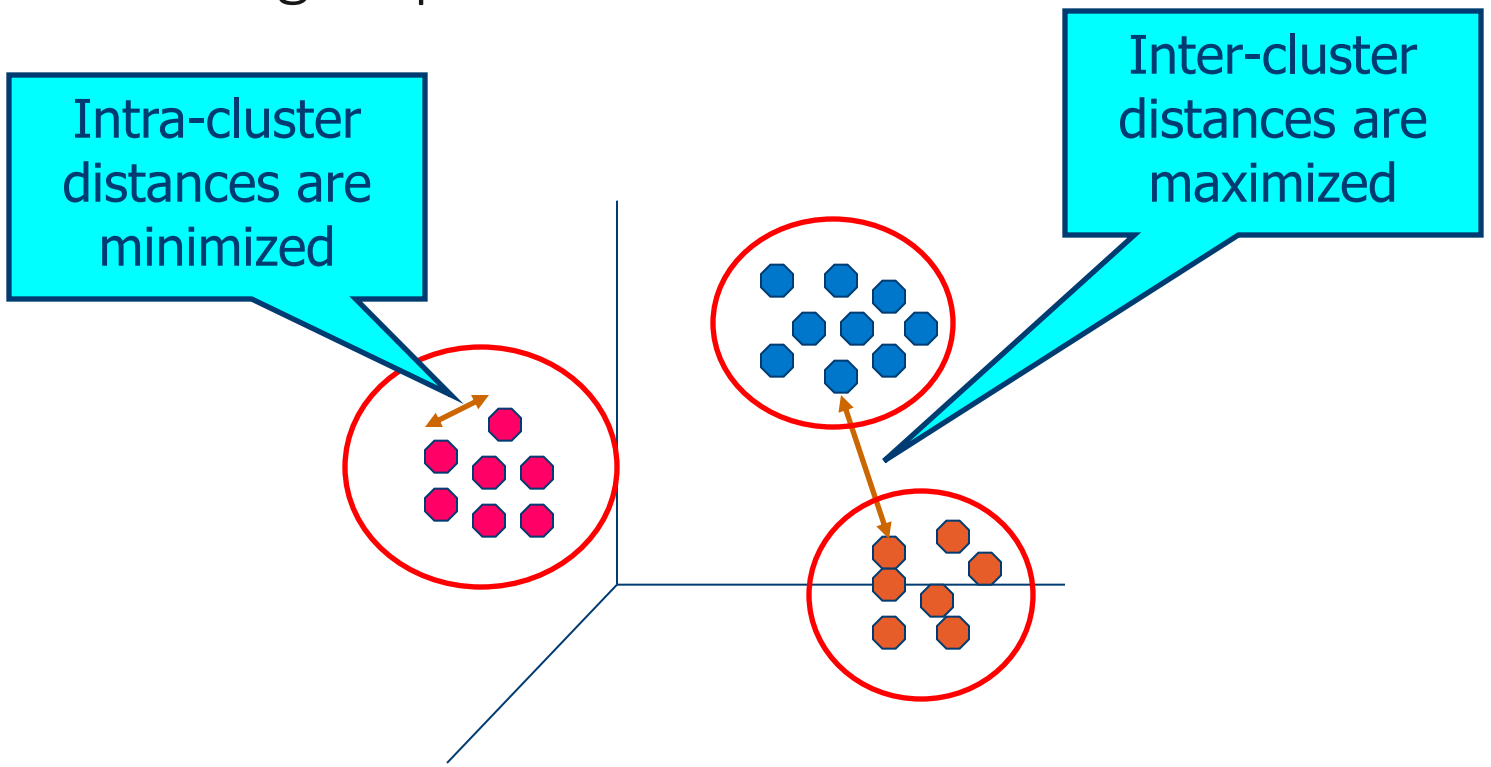


Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- **Examples:**
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

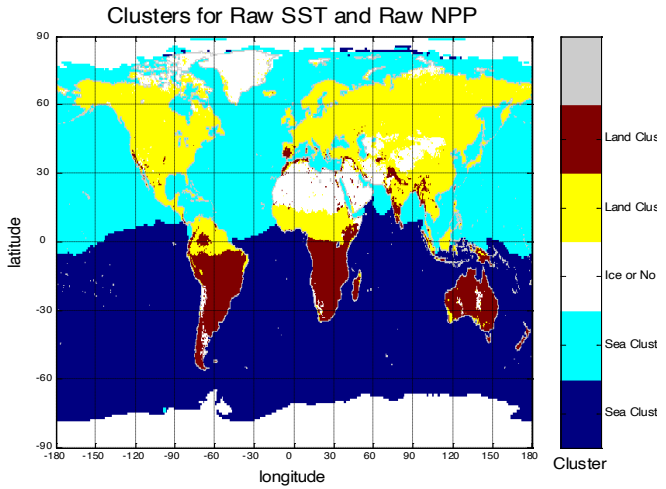
Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

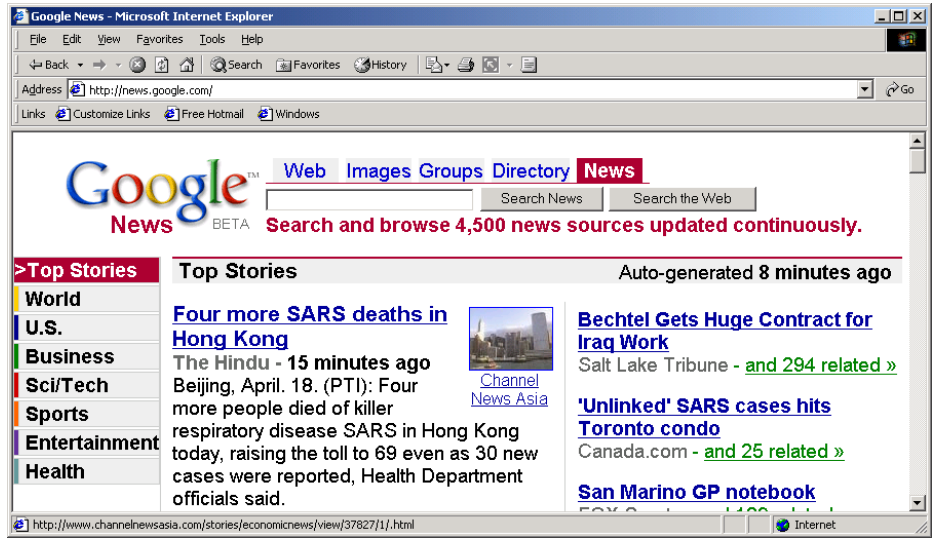


Applications of Cluster Analysis

- Understanding
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- Summarization
  - Reduce the size of large data sets



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

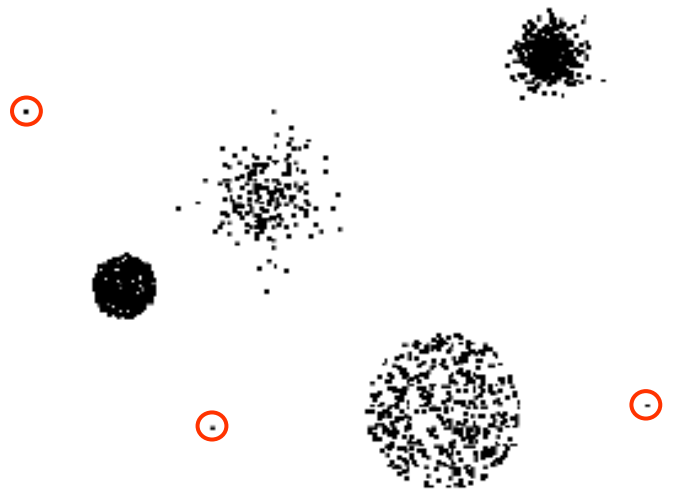
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association Analysis: Applications

- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Deviation/Anomaly/Change Detection



- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.

Class Activity

Which tasks are data mining?

A. Predicting the house price in a an area based on the features
B. Finding companies producing a same product in an area
C. Monitoring the heart rate of a patient for abnormalities
D. Extracting the frequencies of a sound wave
E. Predicting the outcomes of tossing a (fair) pair of dice

3 Things to do:

1. Register in piazza
2. Register in ThurningPoint (via Canvas)
3. Review the Syllabus