Cosi 104a Introduction to machine learning

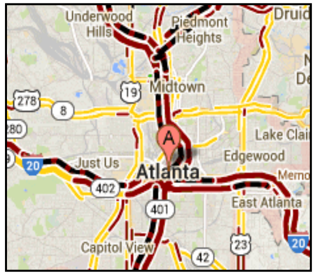
Chapter 2 –Machine Learning Overview

Instructor: Dr. Hongfu Liu
Email: hongfuliu@brandeis.edu

The class notes are a compilation and edition from many sources. The instructor does not claim intellectual property or ownership of the lecture notes.

Large-scale Data is Everywhere!

- **There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies**
- **New mantra**
  - Gather whatever data you can whenever and wherever possible.
- **Expectations**
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.



Traffic Patterns

Biological data

E-Commerce

Cyber Security

Sensor Networks

Social Network

Why Machine Learning? Commercial Viewpoint

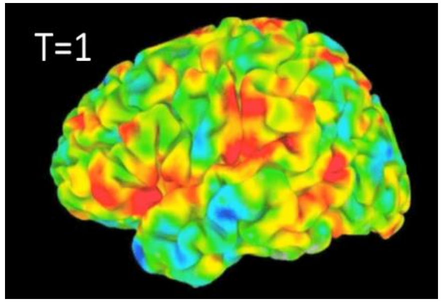
- **Lots of data is being collected and warehoused**
  - Web data
    - Yahoo has Peta Bytes of web data
    - Facebook has billions of active users
  - Purchases at department/grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- **Computers have become cheaper and more powerful**
- **Competitive pressure is strong**
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

Why Machine Learning? Scientific Viewpoint

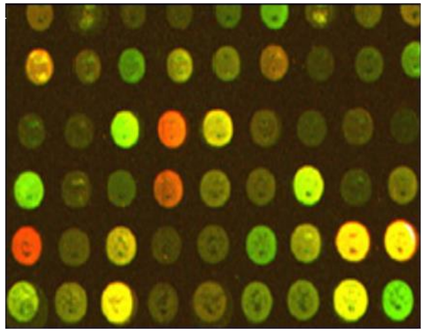
- **Data collected and stored at enormous speeds**
  - Remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data
  - Telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - Scientific simulations
    - Terabytes of data generated in a few hours

- **Machine learning helps scientists**
  - Automated analysis of massive datasets
  - Hypothesis formation

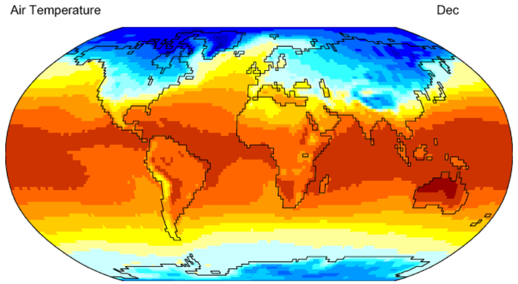
Sky Survey Data



fMRI Data from Brain



Gene Expression Data



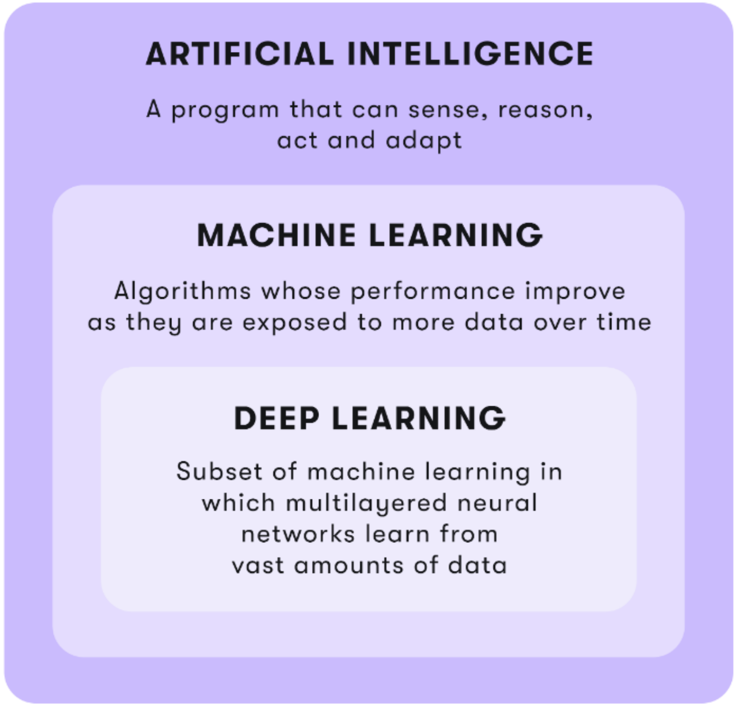
Surface Temperature of Earth

Big Data is popular

- **Data mining solves crucial problems**
  - Improving health care and reducing costs
  - Predicting the impact of climate change
  - Finding alternative/ green energy sources
  - Reducing hunger and poverty by increasing agriculture production
  - Scheduling dangerous good transportation
  - And so on…

What is Machine Learning?

- **No Universal Definitions**
  - Machine learning is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data and thus perform tasks without explicit instructions.



ARTIFICIAL INTELLIGENCE

A program that can sense, reason, act and adapt

MACHINE LEARNING

Algorithms whose performance improve as they are exposed to more data over time

DEEP LEARNING

Subset of machine learning in which multilayered neural networks learn from vast amounts of data

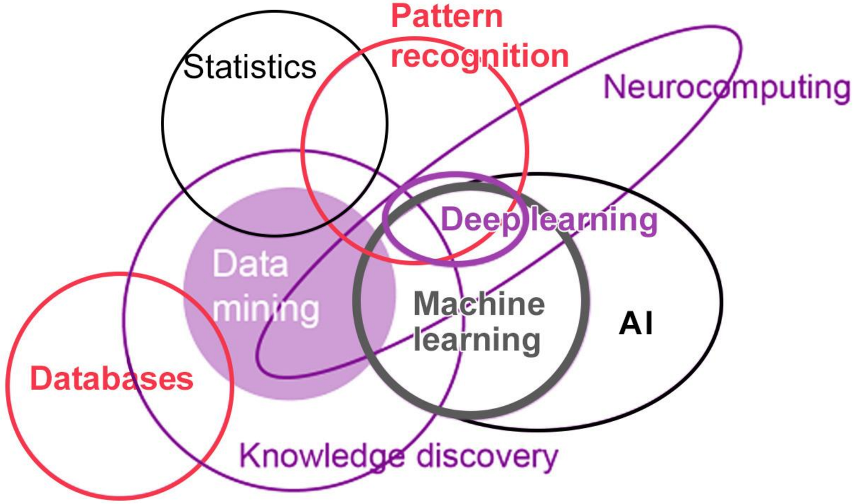
What is NOT machine learning?

- **Some examples**
  - What is not machine learning?
    - Look up phone number in phone directory
    - Query a Web search engine for information about “Amazon”
  - What is machine learning?
    - 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 according to their context (e.g., Amazon rainforest, Amazon.com)
- **Searching vs Mining**

Origins of Machine Learning

- **Draws ideas from AI, pattern recognition, statistics, and database systems**

- **Traditional techniques may be unsuitable due to data that is**
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed

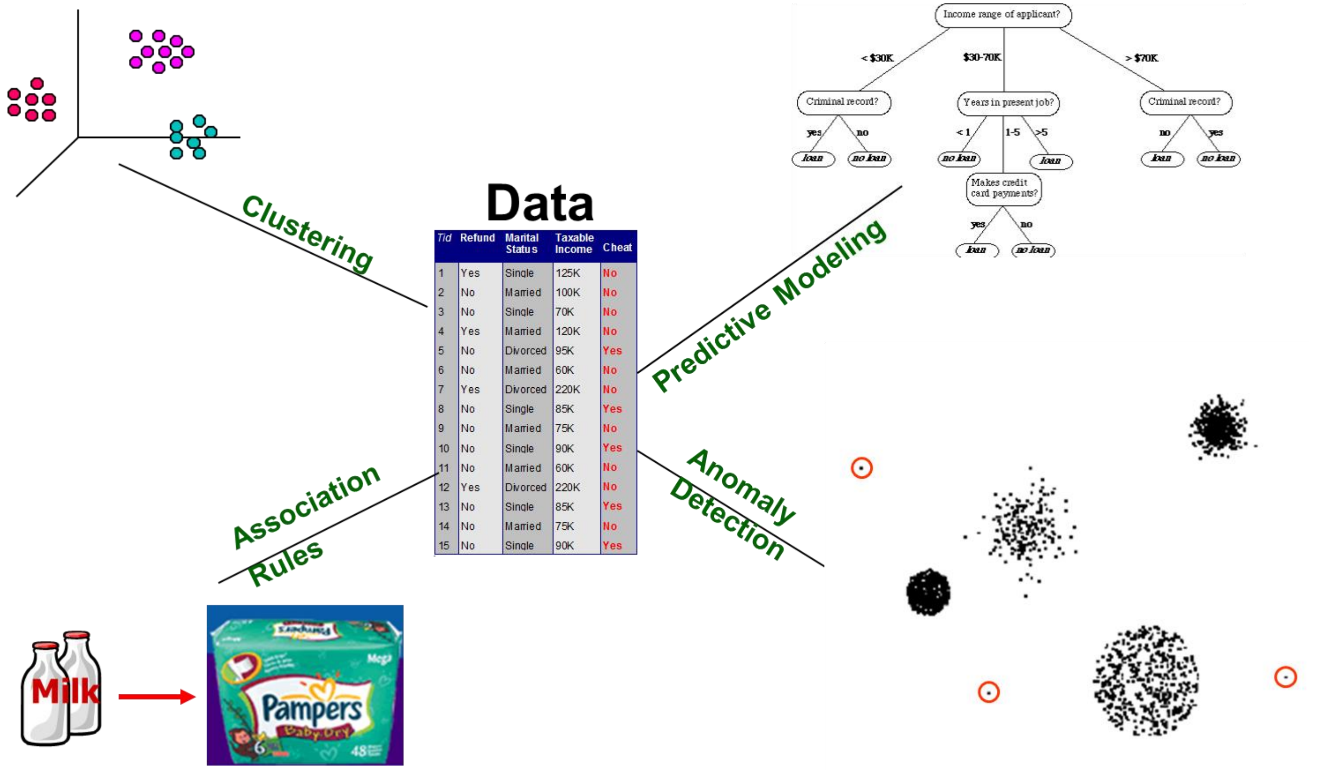


- **A key component of the emerging field of data science and data-driven discovery**

Boundary is blurred

Machine Learning 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.

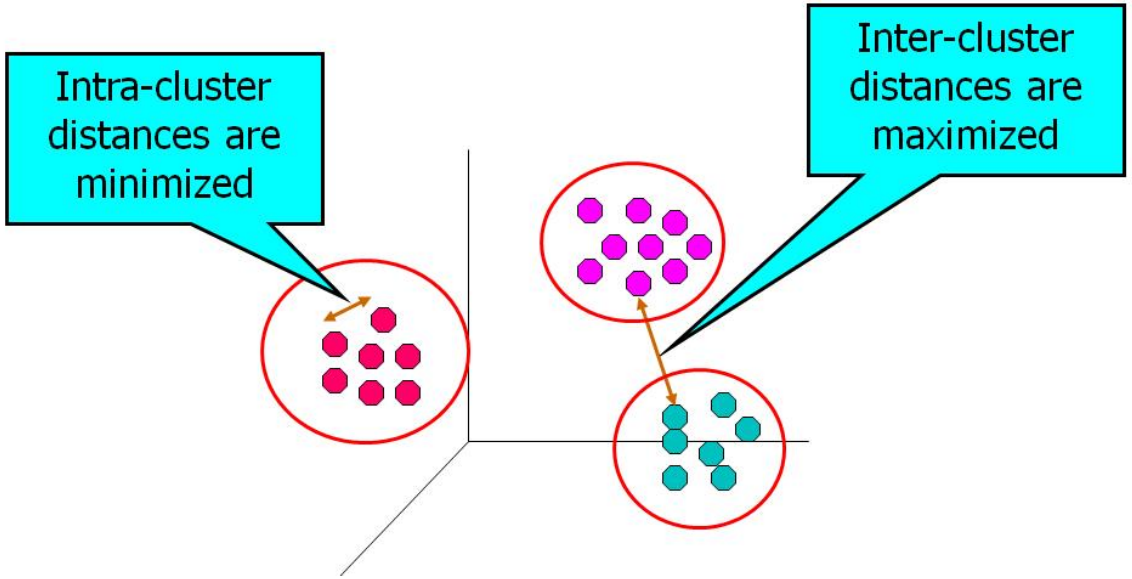


Data

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

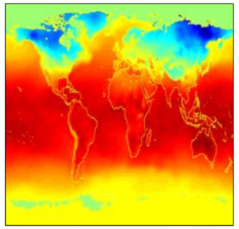
Cluster Analysis

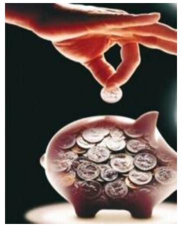
- **Cluster analysis aims to find groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups**



Intra-cluster distances are minimized

Inter-cluster distances are maximized

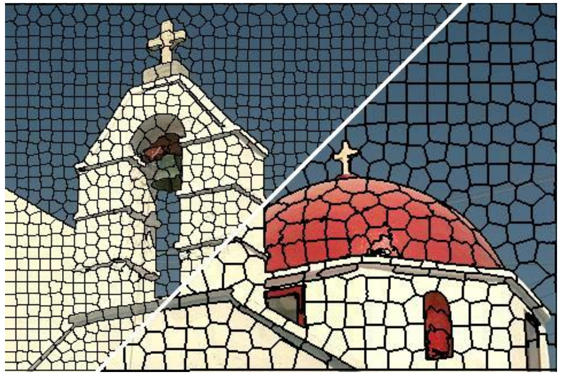




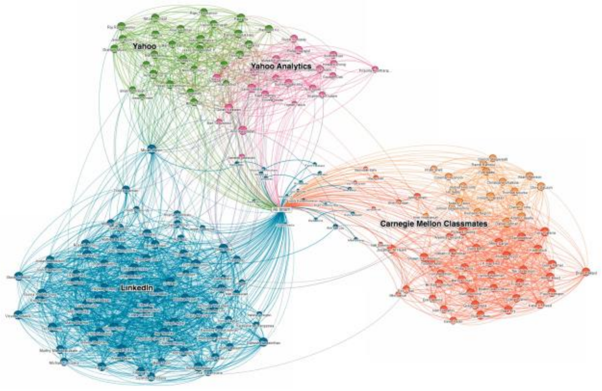
Applications of Cluster Analysis

- **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

City Planning

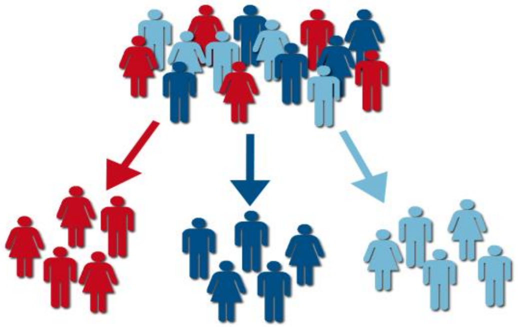


Superpixel



Bigdata Visualization

Recommendation System



Customer Segmentation

Saliency Detection

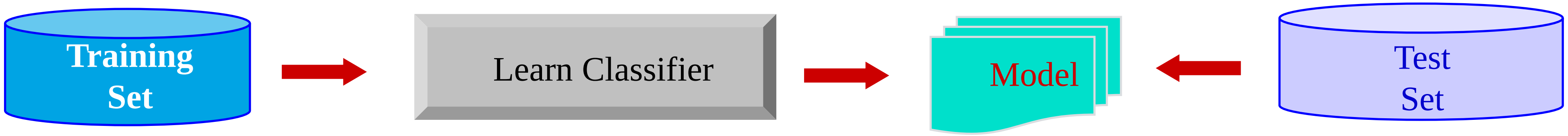
Classification Task

- **Find a model for class attribute as a function of the values of other attributes for predicting new data**

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	?
…	…	…	…	…



Training Set

Learn Classifier

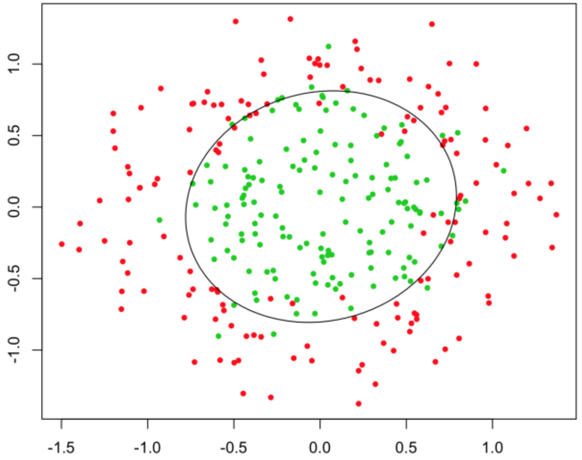
Model

Test Set

Examples of Classification Task

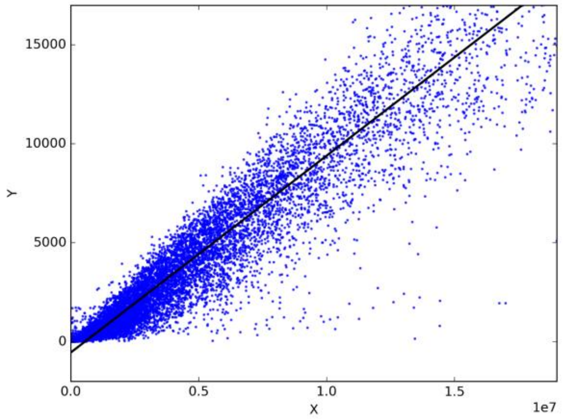
- **Prediction**
  - Classifying credit card transactions as legitimate or fraudulent
  - Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
  - Categorizing news stories as finance, weather, entertainment, sports, etc
  - Identifying intruders in the cyberspace
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

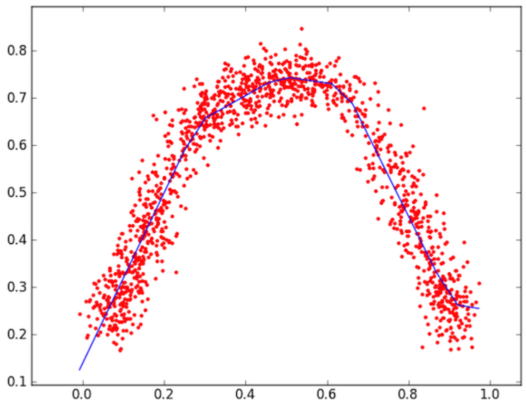
- **Many problems can be formulated as classification task**
  - Recommendation
  - Ranking

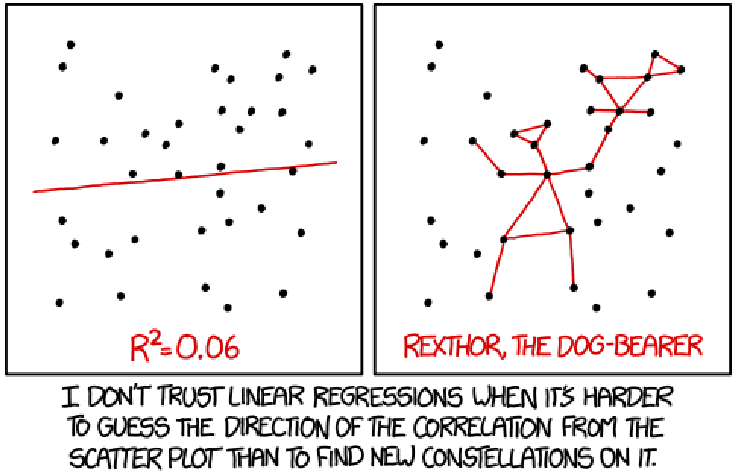


Regression

- **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 advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.






$R^2 = 0.06$
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

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.**

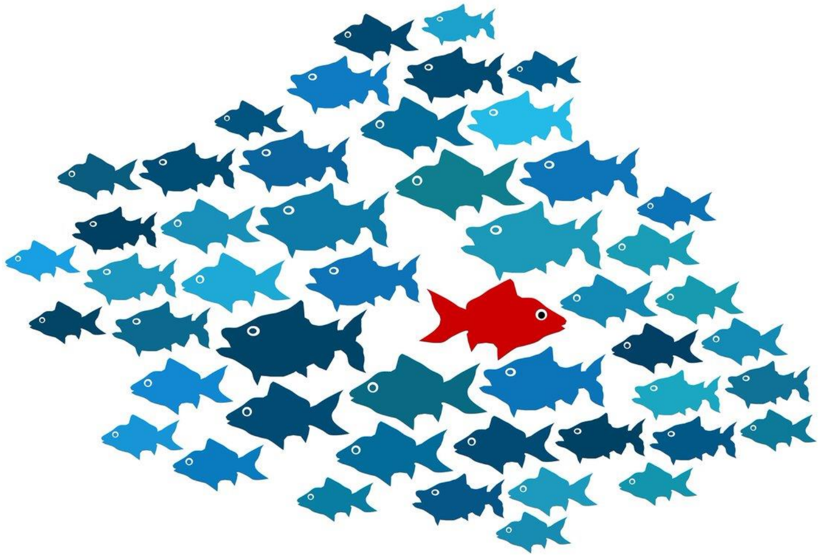
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

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

- **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/Outlier 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
  - High value customer mining



Beyond this

- Air pollution prediction with spatial-temporal data
- Scholarships to needy students
- American presidential election prediction based on twitter
- And so on

Q
&
A

- How to handle big data?