Data Mining: Introduction

Lecture Notes for Chapter 1

Introduction to Data Mining

Fang Zhou

- **Books:**
  - "Introduction to Data Mining" by Pang-Ning Tan, Michael Steinbach, Vipin Kumar

    (https://www-users.cse.umn.edu/~kumar001/dmbook/contents.pdf)
  - "Mining of Massive Datasets" by Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman

- **Lecturer:** Fang Zhou
- **TAs:** Guanyu Lu

- **Lectures:** Week 1- Week 17, Tue, 1pm-2:35pm
- **Practical courses:** Week 4 - Week 15, Tue, 18pm-20:25pm

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
第11节																	
第12节																	
第13节																	

Prerequisites

- Knowledge of basic computer science principles and skills, at a level sufficient to write a reasonably non-trivial computer program
- Good knowledge of Python
- Familiarity with algorithmic analysis
- Familiarity with machine learning

周次	课程内容
1-2	数据类型、预处理
3-6	关联规则
7	项目介绍
8-11	异常检测
12-15	数据分析
16-17	项目汇报

- 课程考核
  - 平时出勤、课堂讨论：10%
  - 平时作业：30%
  - 大作业：60%

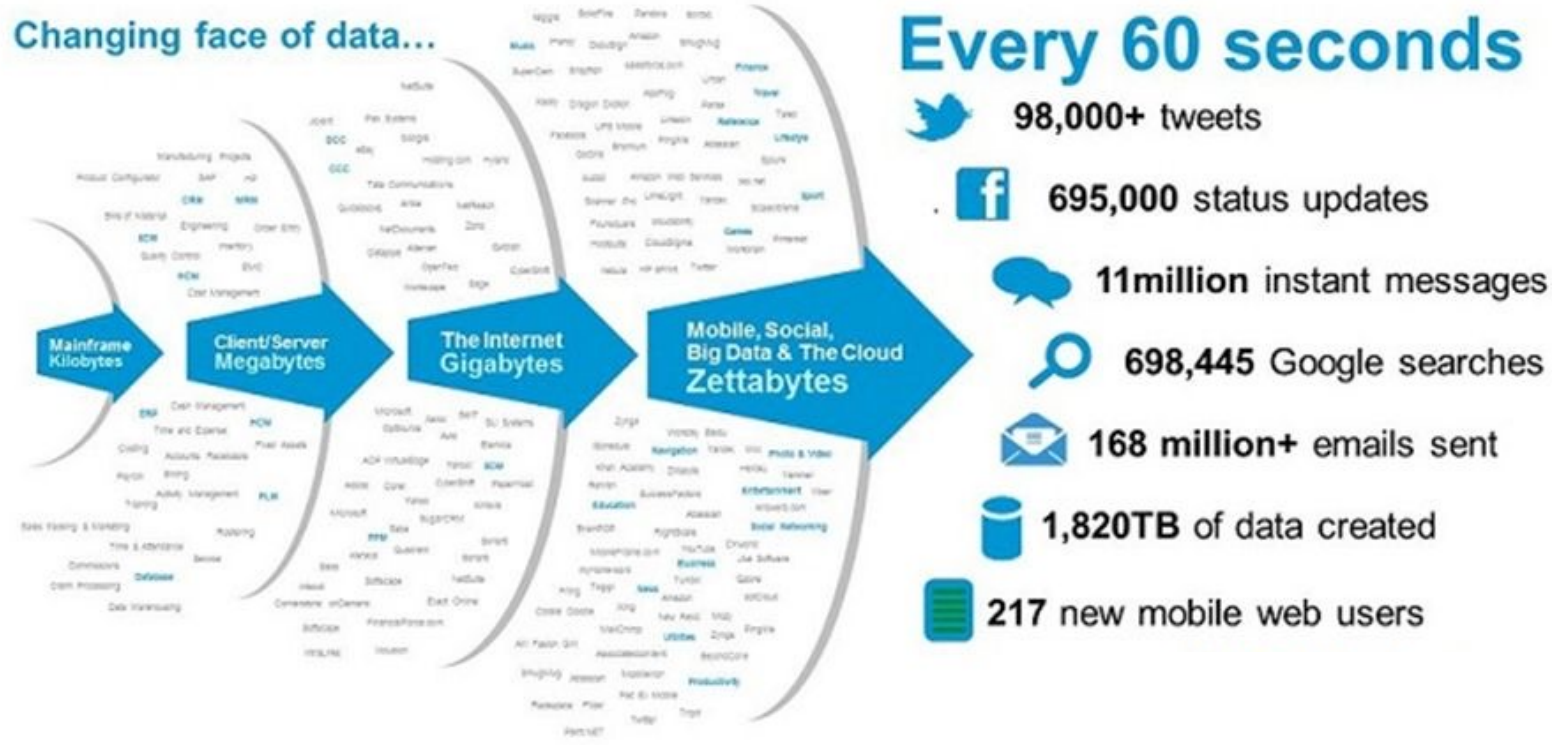
课程目标

- 深刻理解数据的获取、预处理、特征提取和建模过程
- 掌握数据挖掘中关联规则和异常检测基本模型
- 了解异常检测在金融领域中的应用并掌握主要解决方法
- 能够针对应用问题，应用开源软件进行数据处理和分析

Data, Data Everywhere

What is Big Data?

….. Is the exponential growth and availability of data,
both structured and unstructured,
because of the Internet & fast growing technology advancements



Data is power

Data contains value and knowledge

Data Mining

- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - And ANALYZED ← this class

Data Mining ≈ Predictive Analytics ≈ Data Science

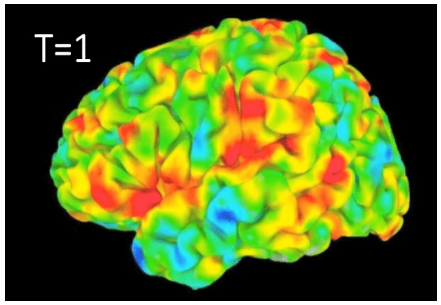
Why Data Mining? Commercial Viewpoint

- 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)

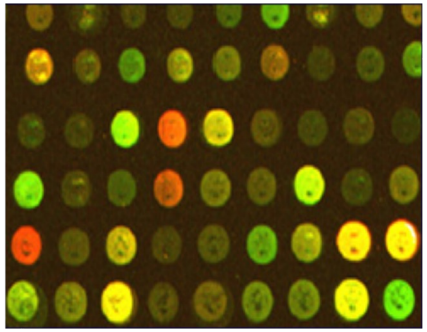
Why Data Mining? Scientific Viewpoint

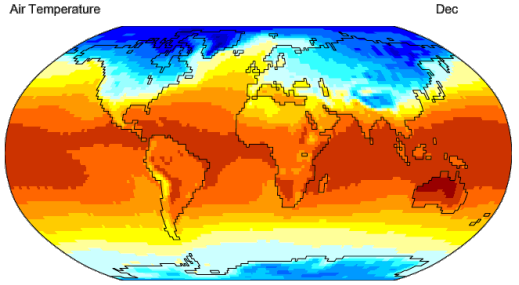
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation



fMRI Data from Brain


Sky Survey Data

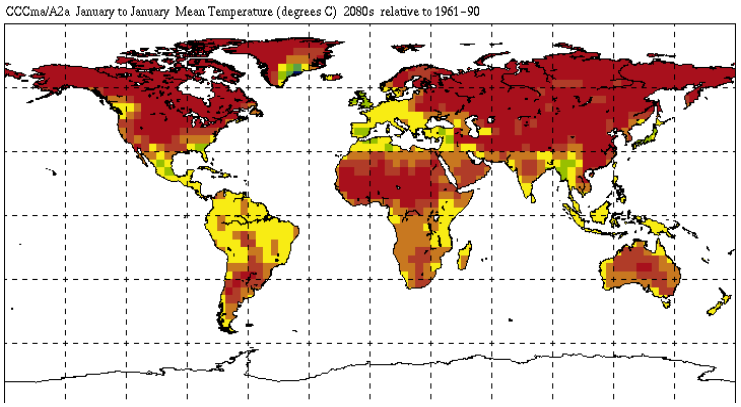

Gene Expression Data



Surface Temperature of Earth

Great Opportunities to Solve Society's Major 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

Good news: Demand for Data Mining

THE DATA SCIENCE / ANALYTICS LANDSCAPE

2,350,000
DSA job listings in 2015

By 2020, DSA job openings are projected to grow
15%

364,000
Additional job listings projected in 2020

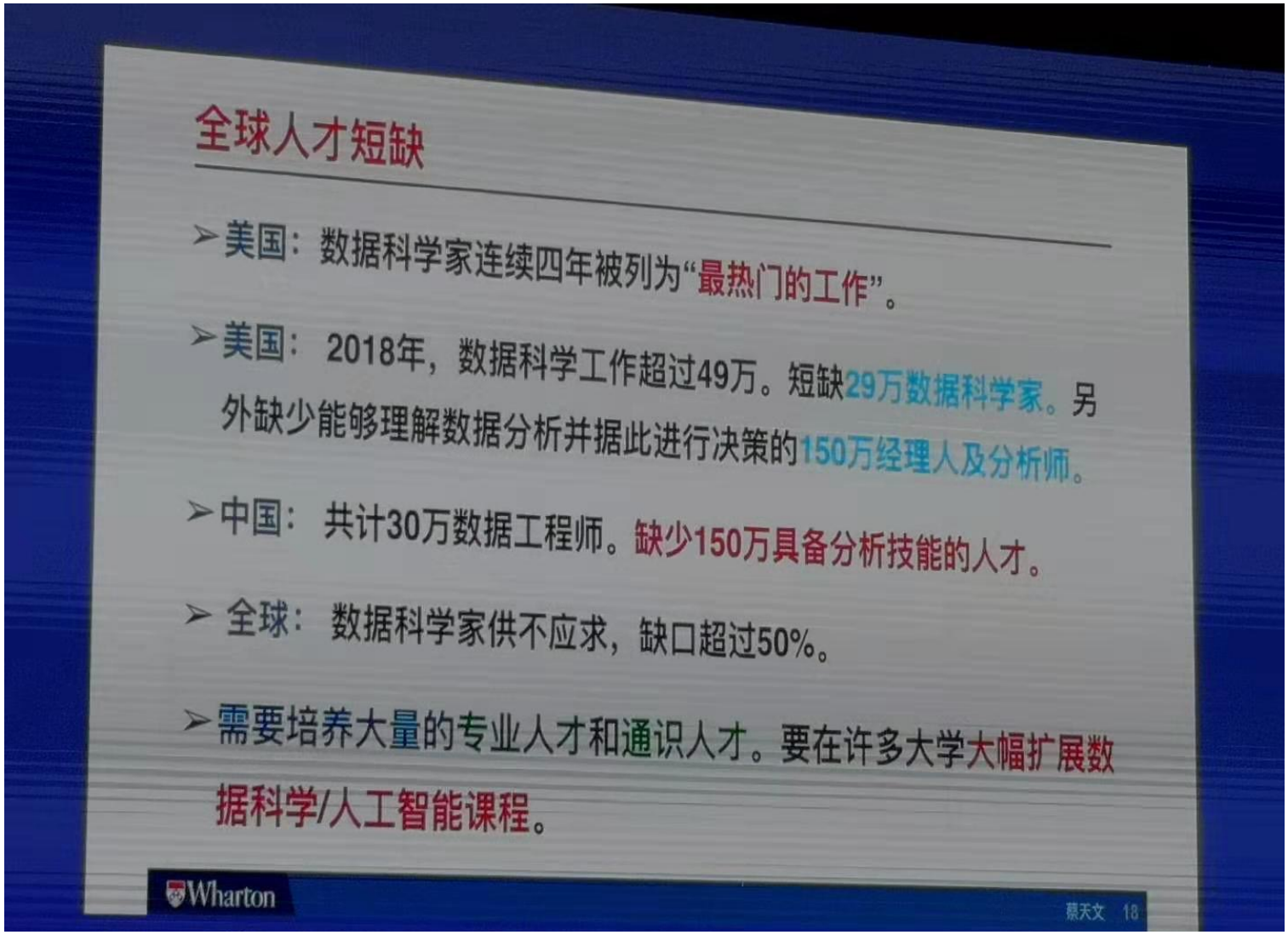
Demand for both Data Scientists and Data Engineers is projected to grow
39%

DSA jobs remain open
5 days
longer than average

DSA jobs advertise average salaries of
$80,265
With a premium over all BA+ jobs of
$8,736

81%
Of DSA jobs require workers with 3-5 years of experience or more

Good news: Demand for Data Mining

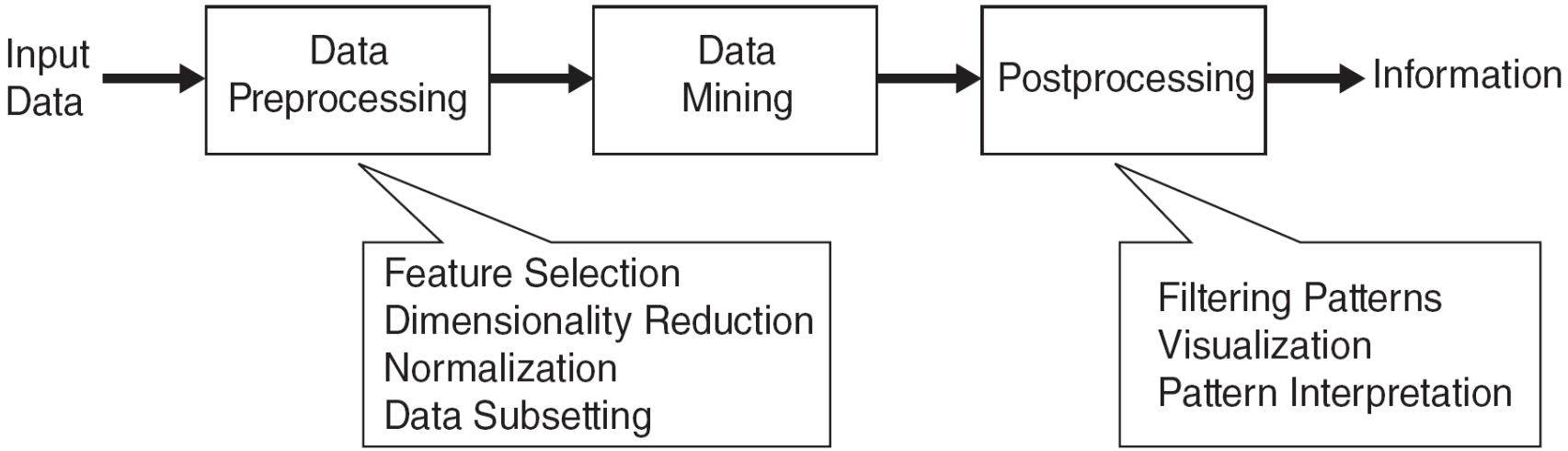


What is Data Mining?

- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

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 Data Mining

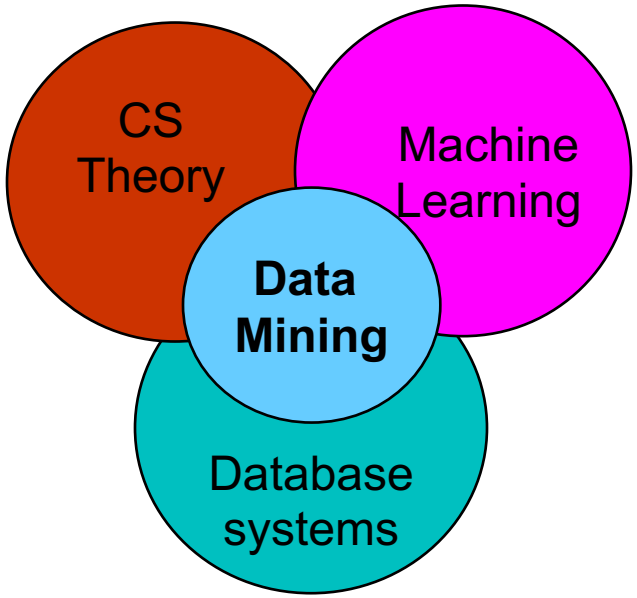
What is (not) Data Mining?

- **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 according to their context (e.g., Amazon rainforest, Amazon.com)

Data Mining: Cultures

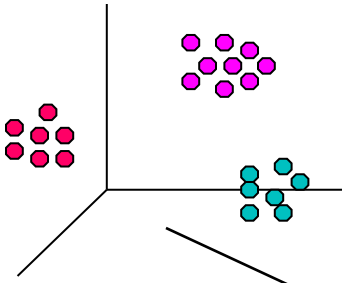
- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms



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 ...

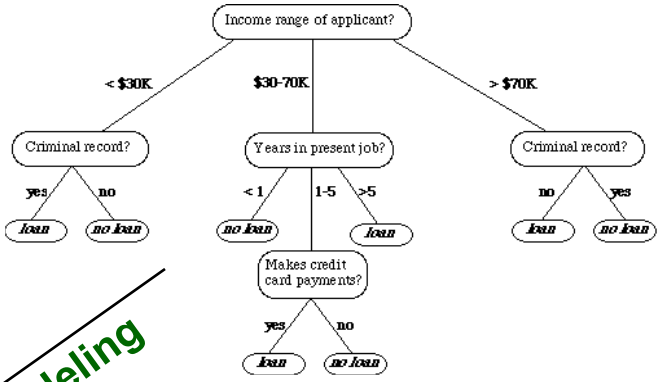


Clustering

Data

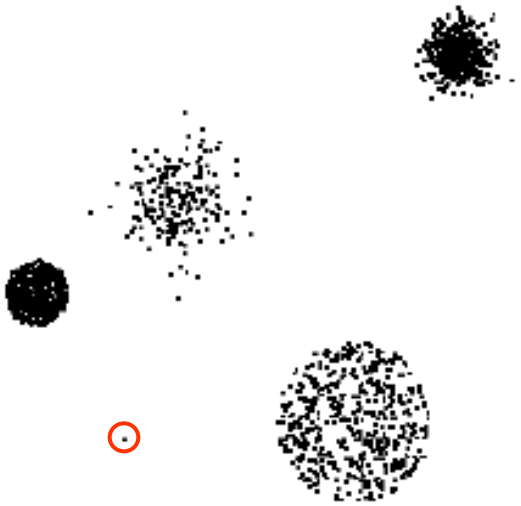
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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

Predictive Modeling



Association Rules

Anomaly Detection



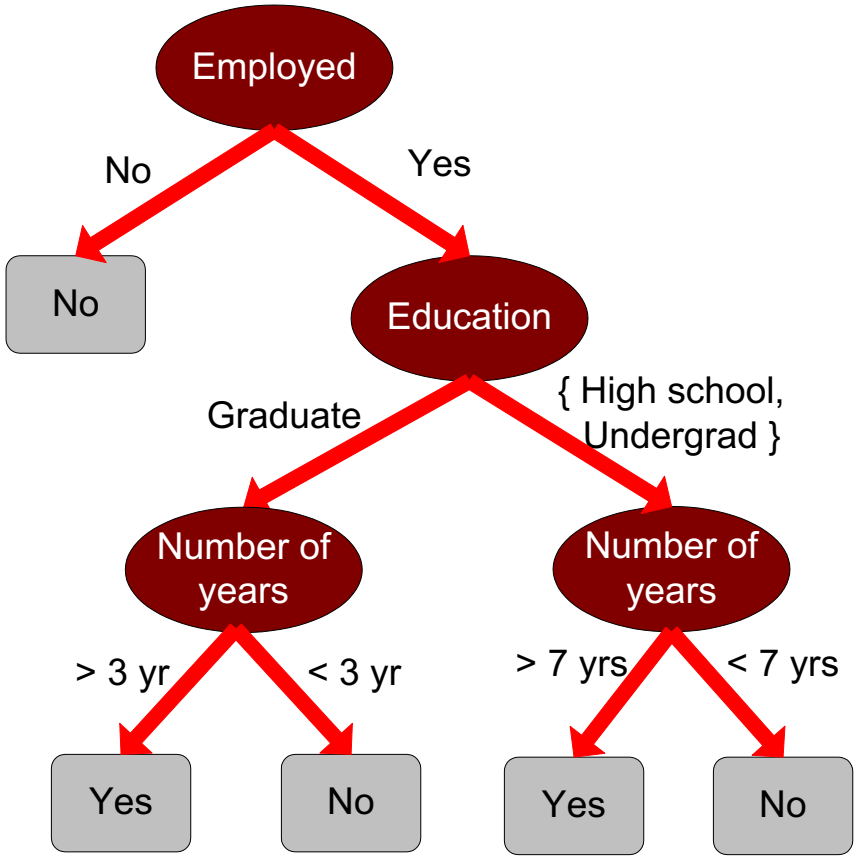
Predictive Modeling: Classification

- 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

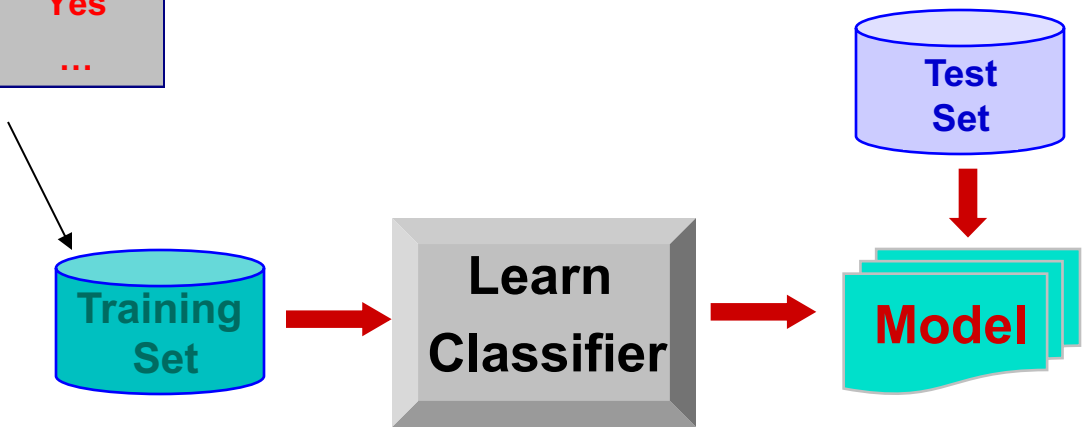


Classification Example

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	?
…	…	…	…	…

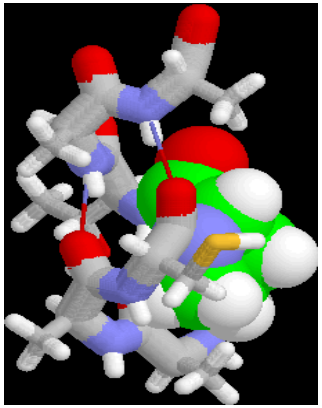


Examples of Classification Task

- 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

Classification: Application 1

- 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 2

- 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.

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.

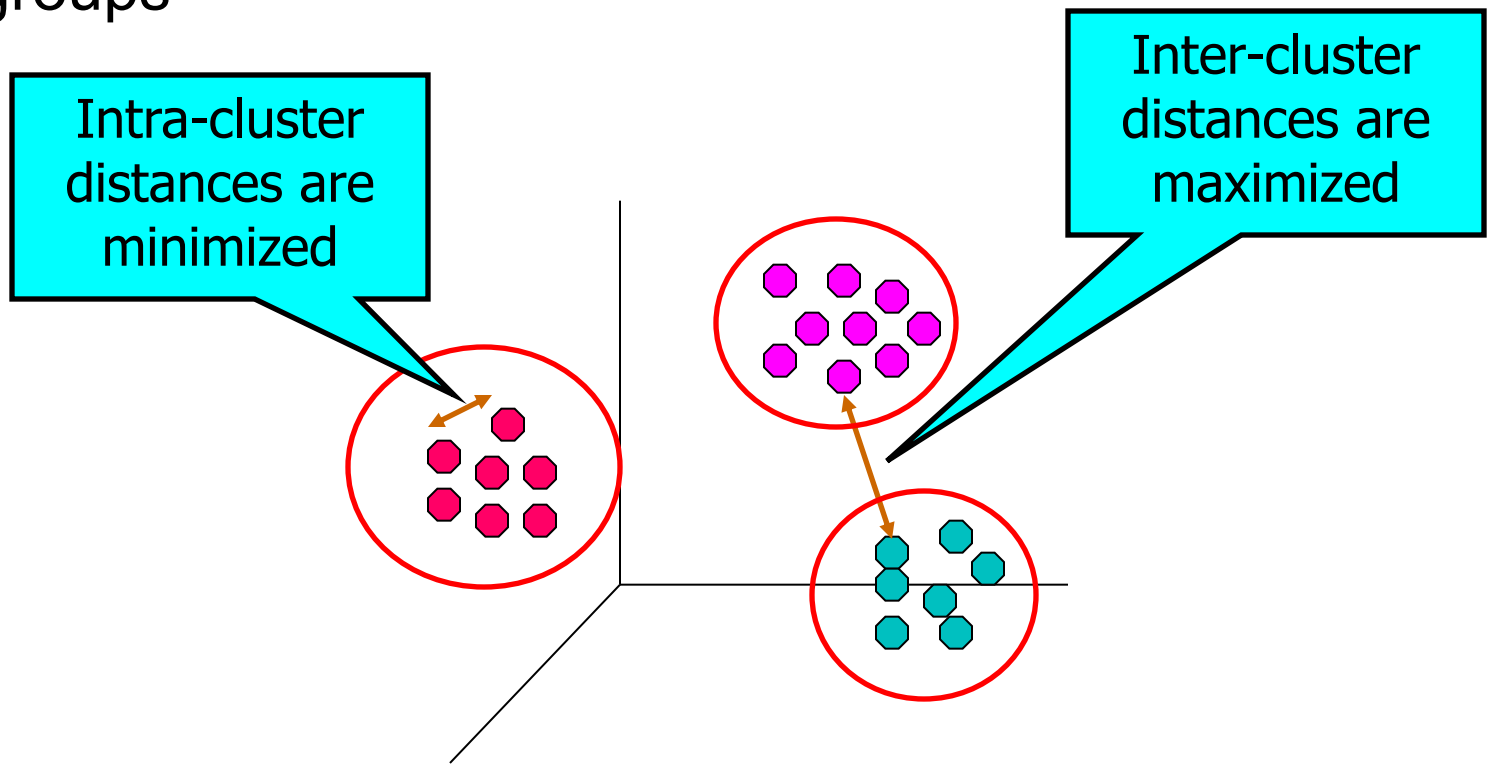
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

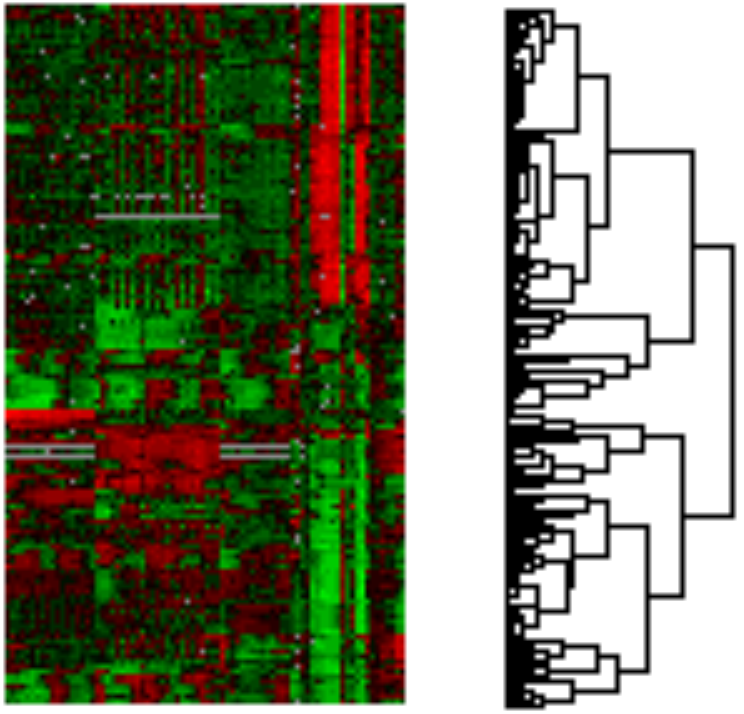
Clustering

- 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

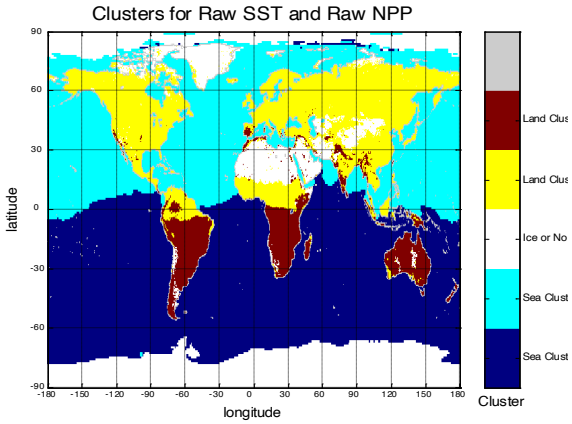


Applications of Cluster Analysis

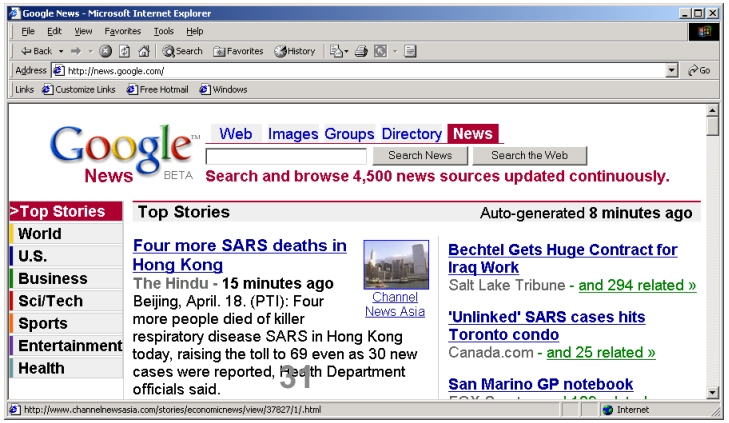
- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



Courtesy: Michael Eisen



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



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: 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 Rule Discovery: Application

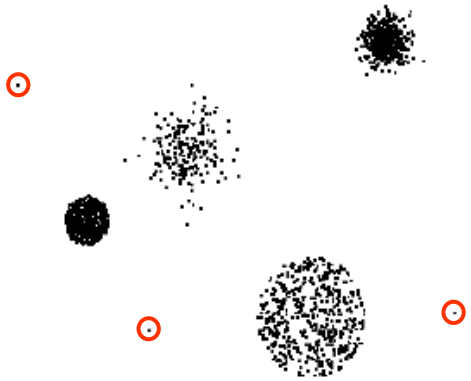
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is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

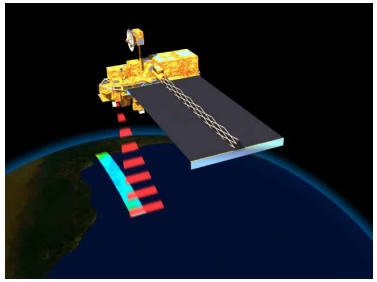
Association Analysis: Applications

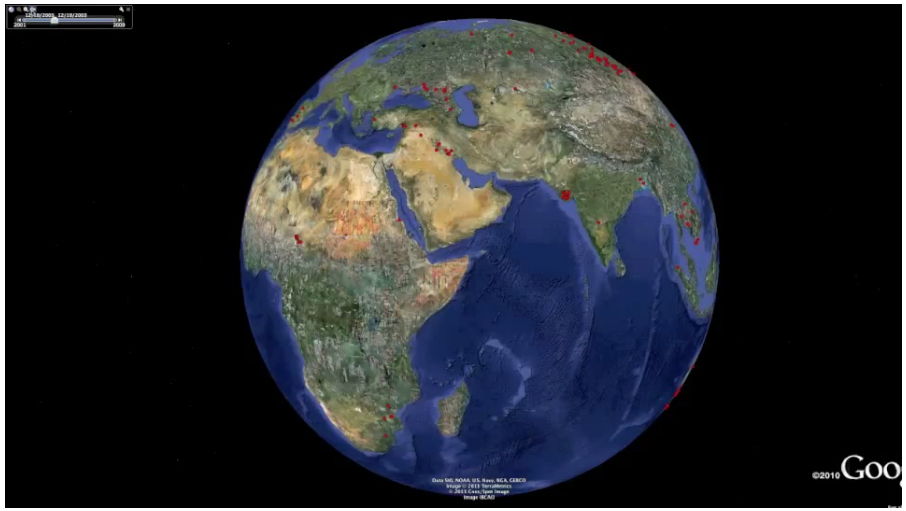
- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination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

Deviation/Anomaly/Change Detection

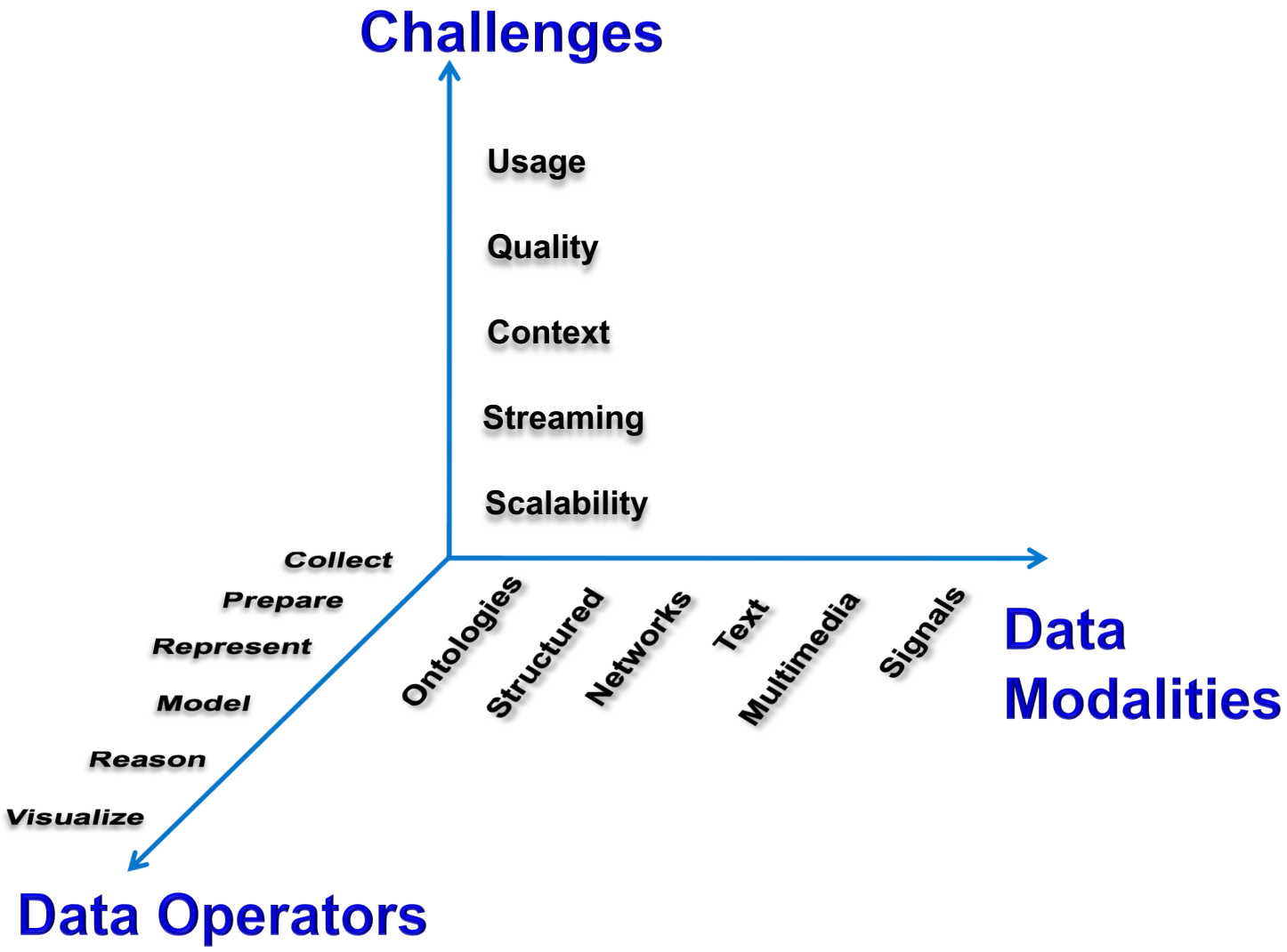
- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.







What matters when dealing with data?



What will we learn?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is labeled
- **We will learn to solve real-world problems:**
  - Market Basket Analysis
  - Fraud Detection
- **We will learn various “tools”**