# Introduction to Data Mining

Tran Giang Son, tran-giang.son@usth.edu.vn

ICT Department, USTH

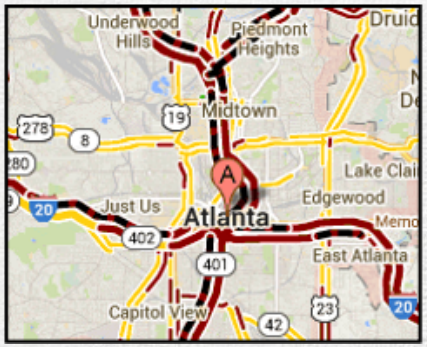
# Data

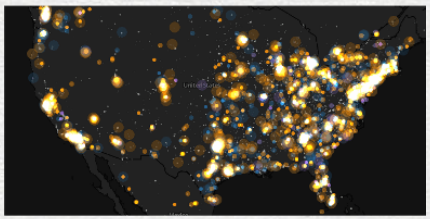
# Data

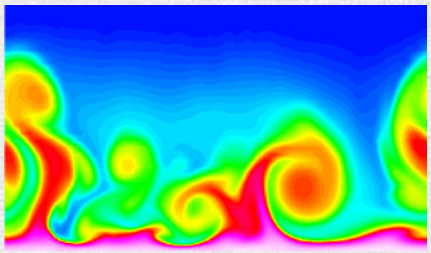
- Large-scale Data is Everywhere!
  - Commercial and scientific databases
  - Data generation and collection technologies
- Gather whatever data you can whenever and wherever possible.
  - Value either for the purpose collected or for a purpose not envisioned.

# Data


[NEWSFACTOR NETWORK]


Underwood Hills
Piedmont Heights
Midtown
Atlanta
Edgewood
Just Us
East Atlanta
Capitol View
278
8
19
20
402
401
42
23





# Data

- Data collected and warehoused
  - Web data
    - Google: Petabytes of web data
    - Facebook: billions of active users
  - Purchases at department/grocery stores, e-commerce
    - Amazon: millions of visits/day
  - Bank/Credit Card transactions

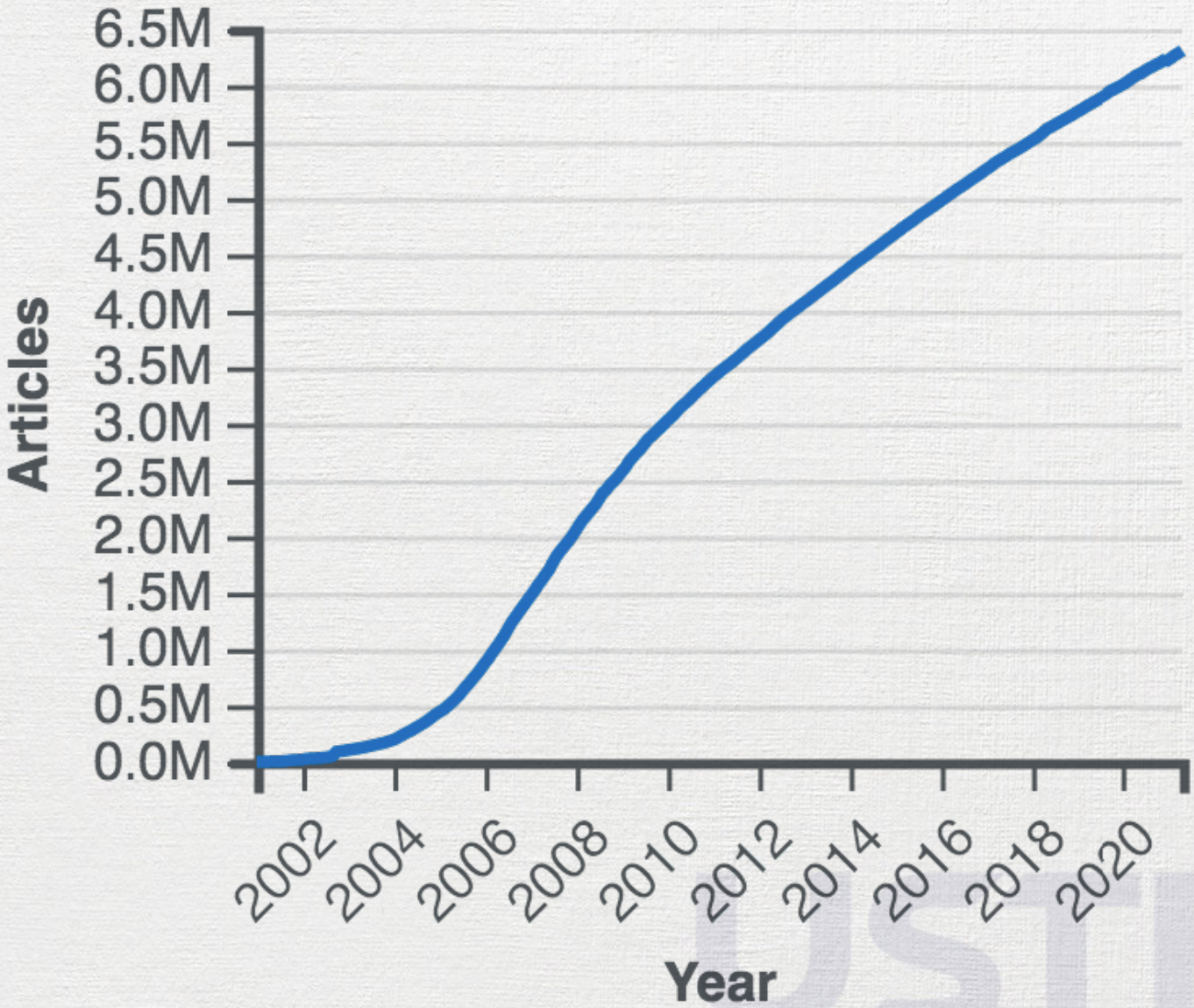
# Data

- Multiple types of data: tables, time series, images, graphs, etc
- Spatial and temporal aspects
- Interconnected data of different types:
  - mobile phone: location of the user, friendship information, check-ins to venues, opinions through twitter, images though cameras, queries to search engines

# Data

- Example: document data
  - Web as a document repository: estimated 50 billions of web pages
  - Wikipedia: 6.4 million articles (and counting)
  - Online news portals: steady stream of 100's of new articles every day
  - Twitter: 300 million tweets every day

# Data: Wikipedia




Wikipedia articles

# Data

- Example: network data
  - Web: 50 billion pages linked via hyperlinks
  - Facebook: 2.9 billion users
  - Twitter: 186 million users
  - Instant messenger: 2 billion users
  - Blogs: 250 million blogs worldwide, presidential candidates run blogs

# Data

- Example: genome data
  - https://www.internationalgenome.org/data
  - Full sequence of 1000 individuals
  - $3 \times 10^9$ nucleotides per person, $3 \times 10^{12}$ nucleotides
  - Lots more data in fact: medical history of the persons, gene expression data

# Data

- Example: Climate data
  - http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php
  - a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center
  - 6000 temperature stations, 7500 precipitation stations, 2000 pressure stations
  - Spatiotemporal data

# Data

- Example: Behavioral data
  - Mobile phones today record a large amount of information about the user behavior
    - GPS records position
    - Camera produces images
    - Communication via phone and SMS
    - Text via facebook updates
    - Association with entities via check-ins
  - Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.
  - Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.

# What

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
- Object is also known as record, point, case, sample, entity, or instance

# What: Attributes

- Categorical
  - Eye color, zip codes, words, rankings (e.g, good, fair, bad), height in {tall, medium, short}
  - Nominal (no order or comparison) vs Ordinal (order but not comparable)
- Numeric
  - Dates, temperature, time, length, value, count.
  - Discrete (counts) vs Continuous (temperature)
  - Special case: Binary attributes (yes/no, exists/not exists)

# What: Record data

- Multi-dimensional space: Same fixed set of numeric attributes
- n-by-d data matrix, where there are n rows, one for each object, and d columns, one for each attribute

| Projection of x Load | Projection of y load | Distance | Load | Thickness |
|---|---|---|---|---|
| 10.23 | 5.27 | 15.22 | 2.7 | 1.2 |
| 12.65 | 6.25 | 16.22 | 2.2 | 1.1 |

Record data

# What: Categorical data

- Data that consists of a collection of records, each of which consists of a fixed set of categorical attributes

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | High | No |
| 2 | No | Married | Medium | No |
| 3 | No | Single | Low | No |
| 4 | Yes | Married | High | No |
| 5 | No | Divorced | Medium | Yes |
| 6 | No | Married | Low | No |
| 7 | Yes | Divorced | High | No |

# What: Transactional data

- Each record (transaction) is a set of items.
- A set of items can also be represented as a binary vector, where each attribute is an item.

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

Transactional data

# What: Ordered data

- Genomic sequence data
- Data is a long ordered string

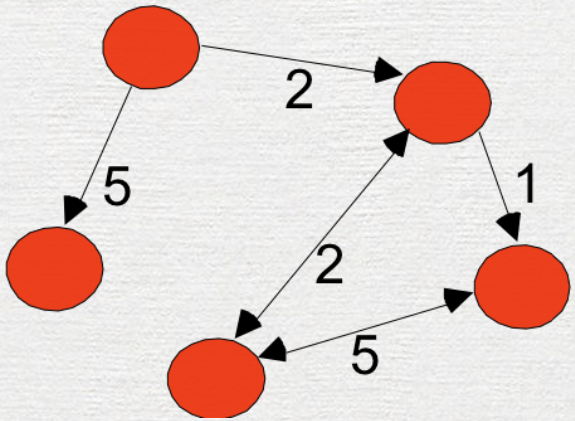
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC

# What: Ordered data

- Time series
  - Sequence of ordered (over "time") numeric values.

# What: Graph data

- Examples: Web graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Graph data

# Data Mining

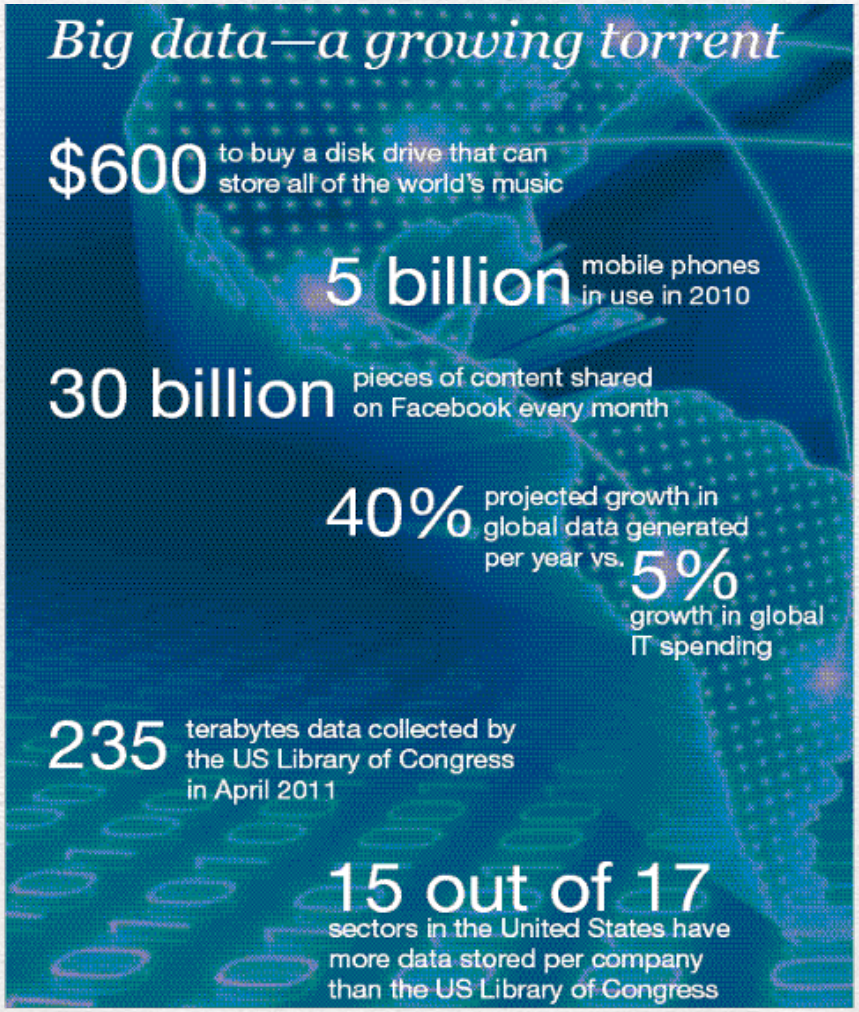
# Why

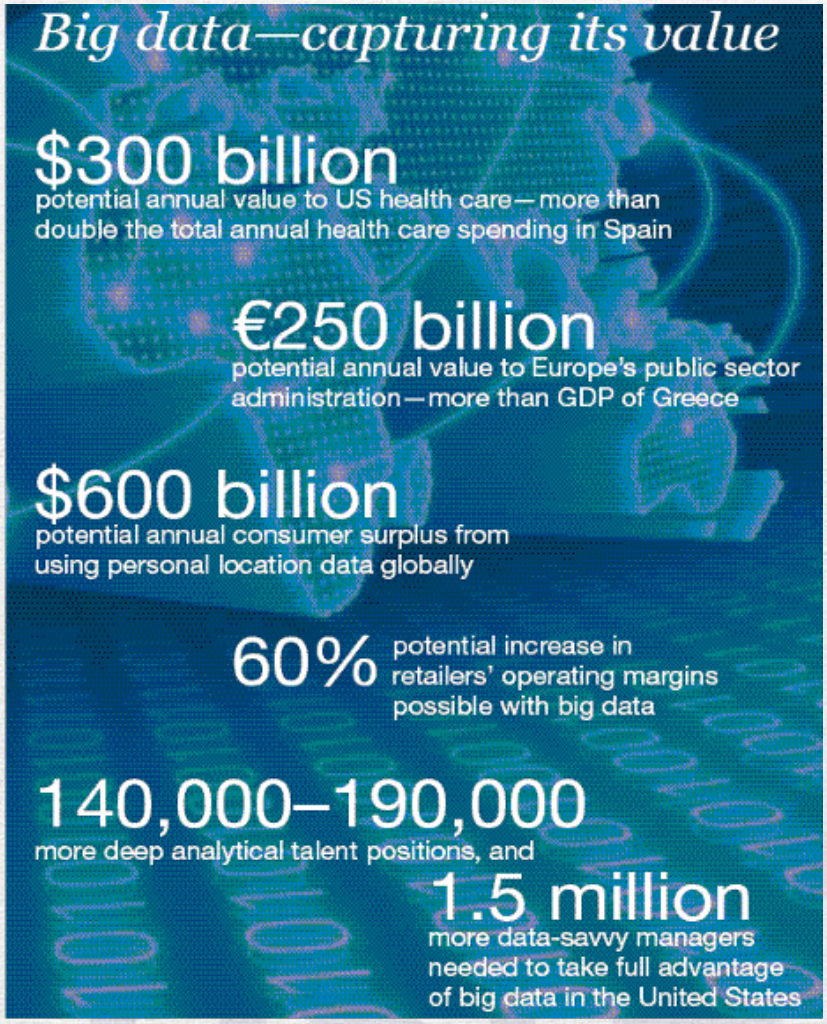
- Cheaper and more powerful computers
- Strong competitive pressure
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

# Why

- Research data collected and stored at enormous speeds
  - Remote sensors on a satellite
    - NASA EOSDIS: petabytes of earth science data / year
  - Telescopes scanning the skies
  - 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

# Why

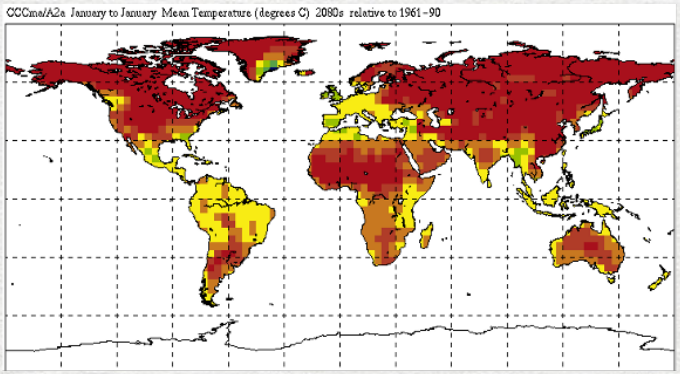

Big data—a growing torrent
$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
235 terabytes data collected by the US Library of Congress in April 2011
15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress


Big data—capturing its value
$300 billion potential annual value to US health care—more than double the total annual health care spending in Spain
€250 billion potential annual value to Europe's public sector administration—more than GDP of Greece
$600 billion potential annual consumer surplus from using personal location data globally
60% potential increase in retailers' operating margins possible with big data
140,000–190,000 more deep analytical talent positions, and
1.5 million more data-savvy managers needed to take full advantage of big data in the United States

# Why

Healthcare

Green Energy



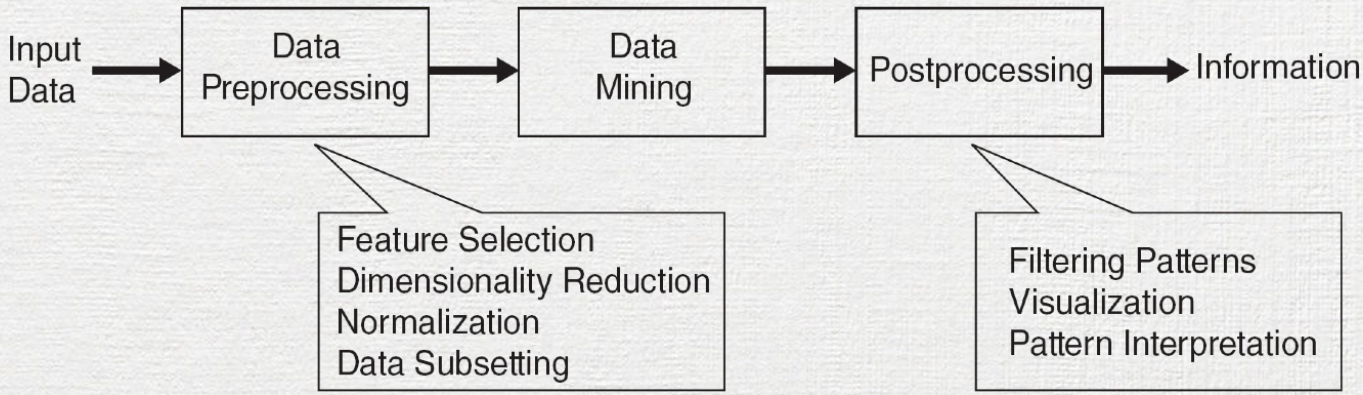

Climate change

Argiculture production

# What

- 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
  - The use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data.
  - The discovery of models for data

# What

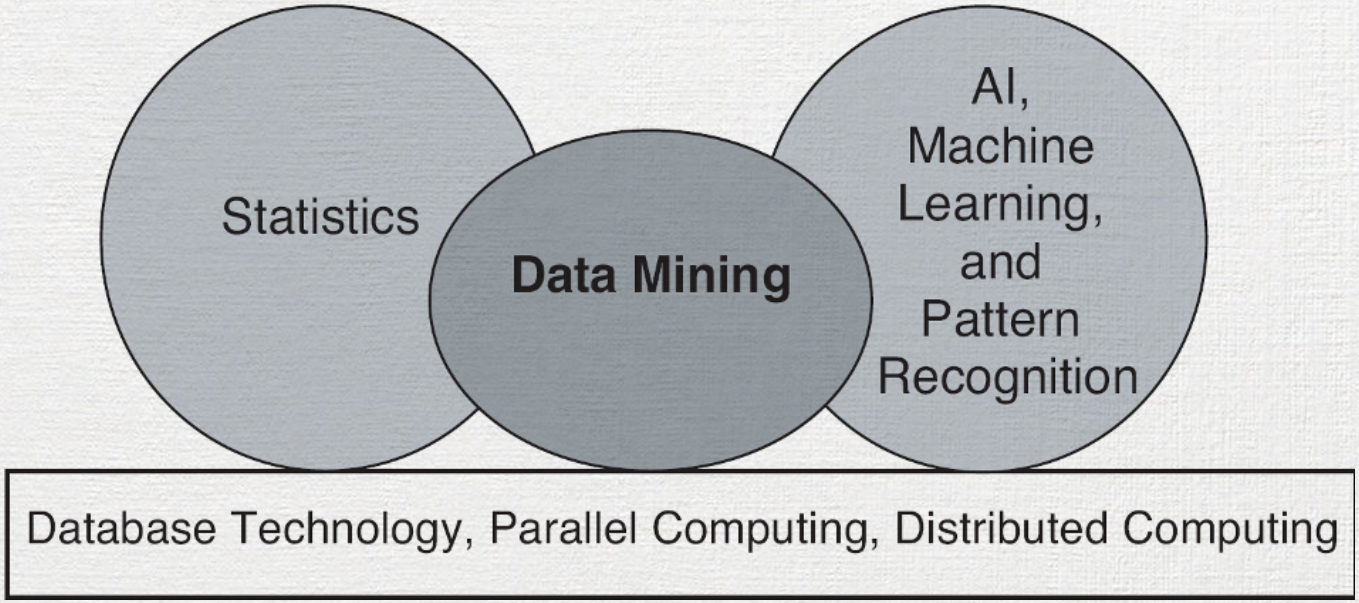



Pipeline

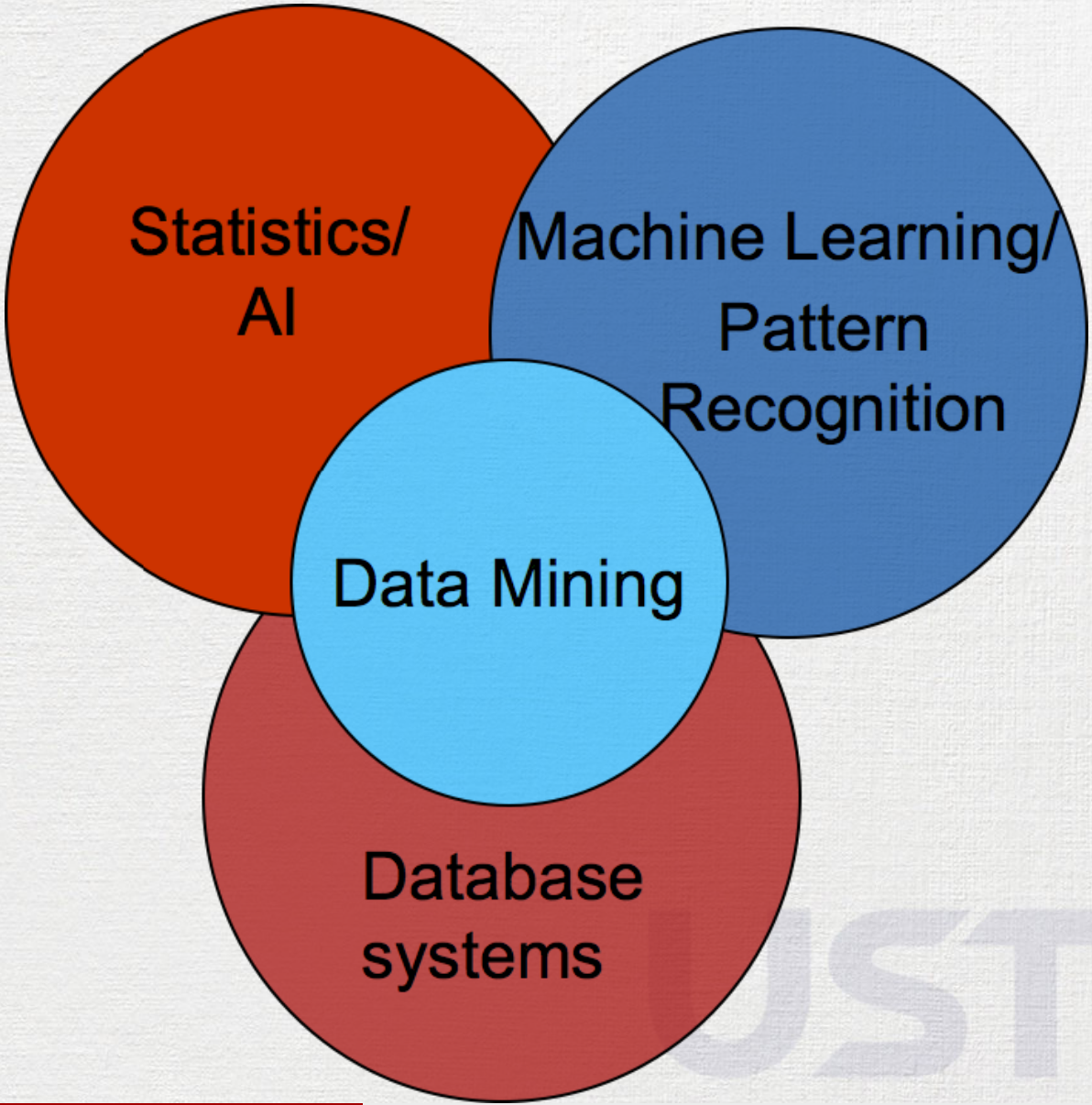
# What

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Difficult data
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed
- A key component of the emerging field of data science and data-driven discovery

# What


Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

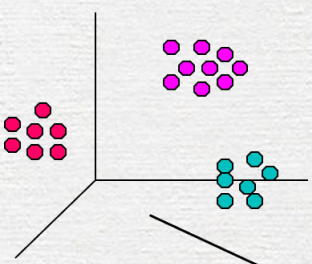
# What


Statistics/
AI
Machine Learning/
Pattern
Recognition
Data Mining
Database
systems

# What: 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.

# What: Tasks

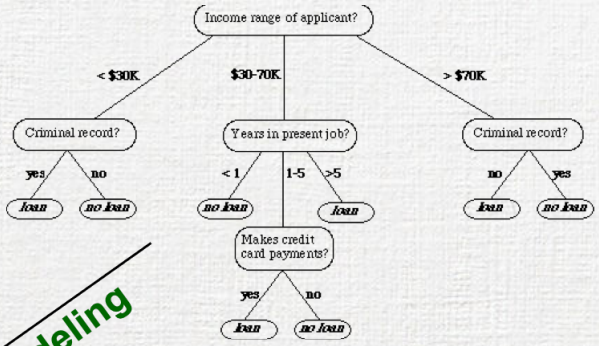


Clustering

## 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

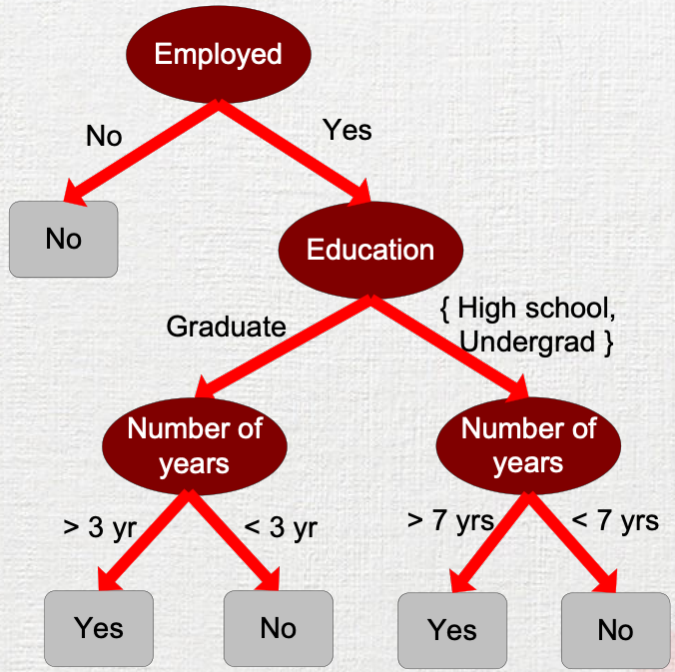
# Classification

# What: Prediction Methods - Classification

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

| 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 |
| ... | ... | ... | ... | ... |

Input



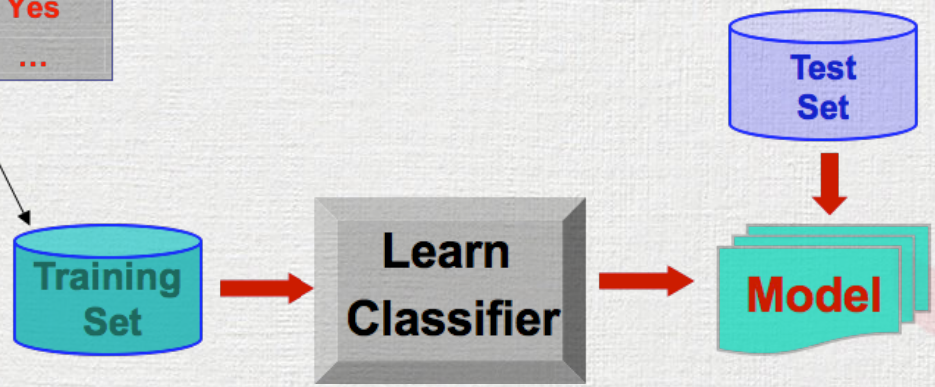

Rules

# What: Prediction Methods - Classification

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 | ? |
| … | … | … | … | … |



# Example: Classification

- Credit card transactions: legitimate or fraudulent
- Land covers: water bodies, urban areas, forests, etc. using satellite data
- News stories: finance, weather, entertainment, sports, etc.
- Identifying intruders in the cyberspace
- Tumor cells: benign or malignant
- Secondary structures of protein: alpha-helix, beta-sheet, or random coil

# Example: Classification

- 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.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

# Example: Classification

- Customer loyalty
  - 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.

# Example: Classification

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images. 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.

# Regression

# What: Regression

- Predict continuous valued variable
  - Based on the values of other variables
  - Linear or nonlinear model of dependency
- Extensively studied
  - Statistics
  - Neural network

# Example: Regression

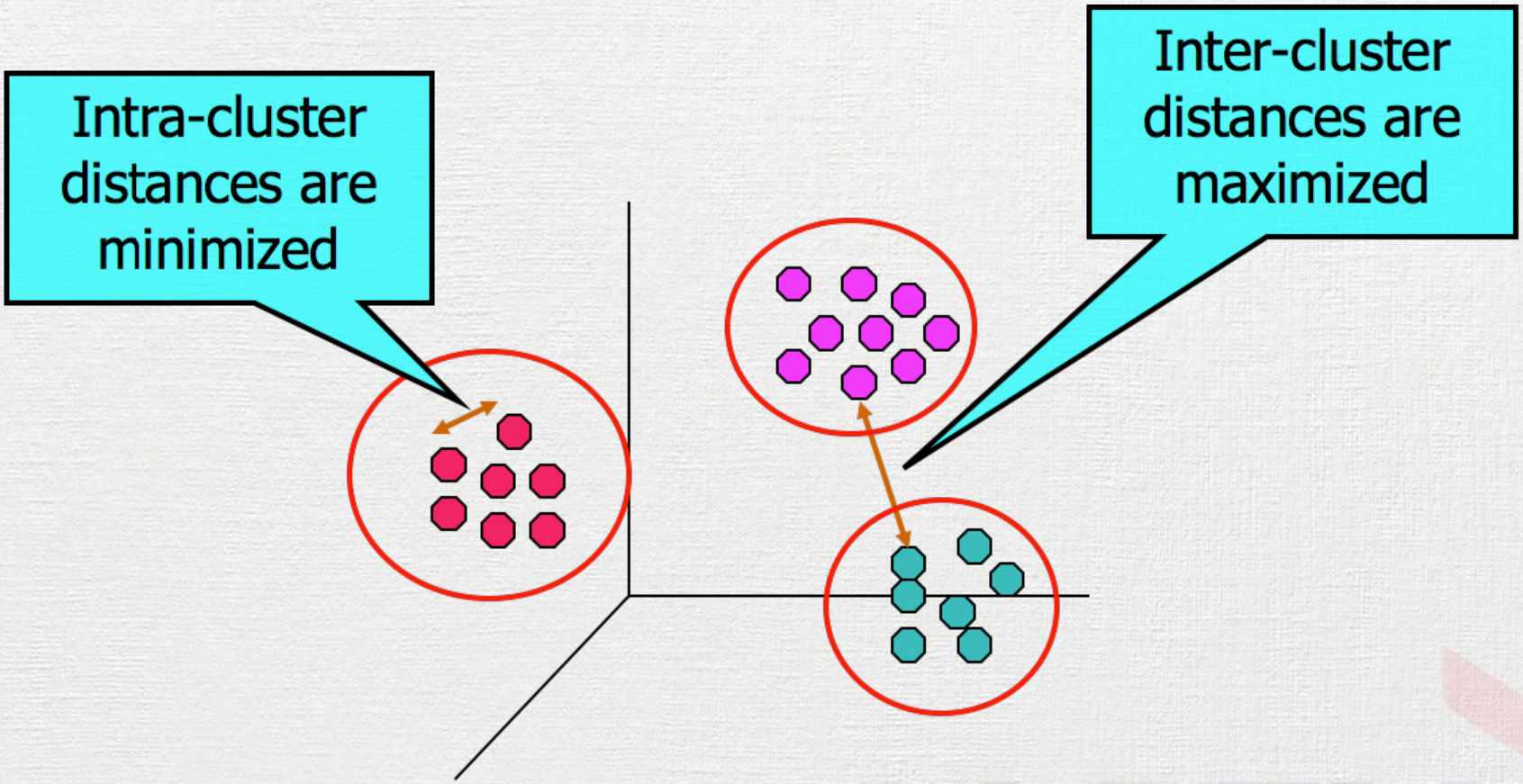
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc
  - Time series prediction of stock market indices

# Clustering

# What

- Finding groups of objects
  - Similar objects in the same group
  - Different from objects in other groups

# What


Intra-cluster distances are minimized
Inter-cluster distances are maximized

# Why

- 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

# Example: Cluster Analysis

- 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.

# Example: Cluster Analysis

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms
  - Approach:
    - Identify frequently occurring terms in each document
    - Form a similarity measure based on the frequencies of different terms
    - Use similarity to cluster

# Association Rules

# What

- Input: set of records
  - Some number of items from a given collection
- Ouput: dependency rules
  - predict occurrence of an item based on occurrences of other items

# What

| 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}

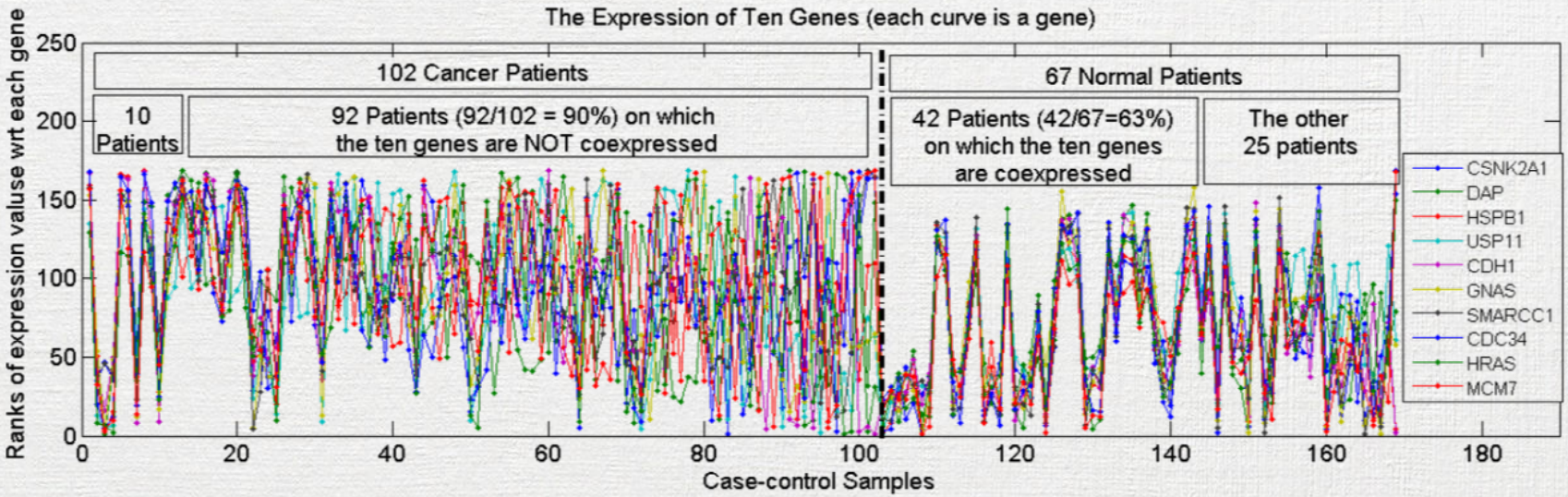
# Why

- Market-basket analysis
  - Sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Find combination of patient symptoms and test results associated with certain diseases

# Example: Association Rules

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset
  - Enriched with the TNF/NFB signaling pathway
    - Well-known to be related to lung cancer
  - P-value: $1.4 \times 10^{-5}$ (6/10 overlap with the pathway)

# Example: Association Rules



# Example: Association Rules

- 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.

# Disadvantages

# Why not?

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