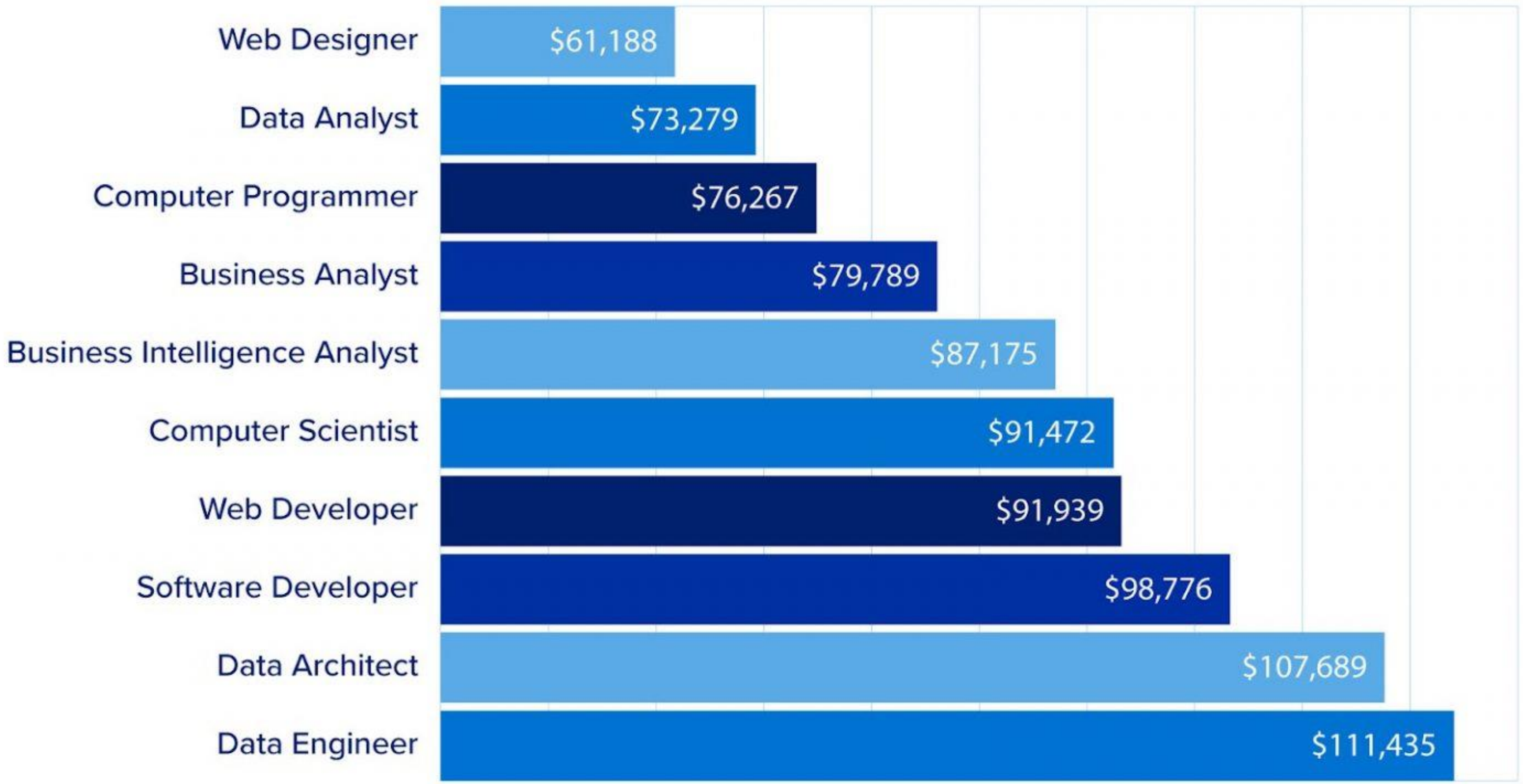
LECTURE 1 : INTRODUCTION

- DS3002 – Data Mining
- By Eesha Tur Razia Babar
- Content Taken from Multiple Sources including Introduction to Data Mining By PN TAN, Data Mining Concepts and Techniques by Han, google, Notes of Dr. Stephan Mandt, and selected papers

WHY BOTHER? JOB PERSPECTIVE

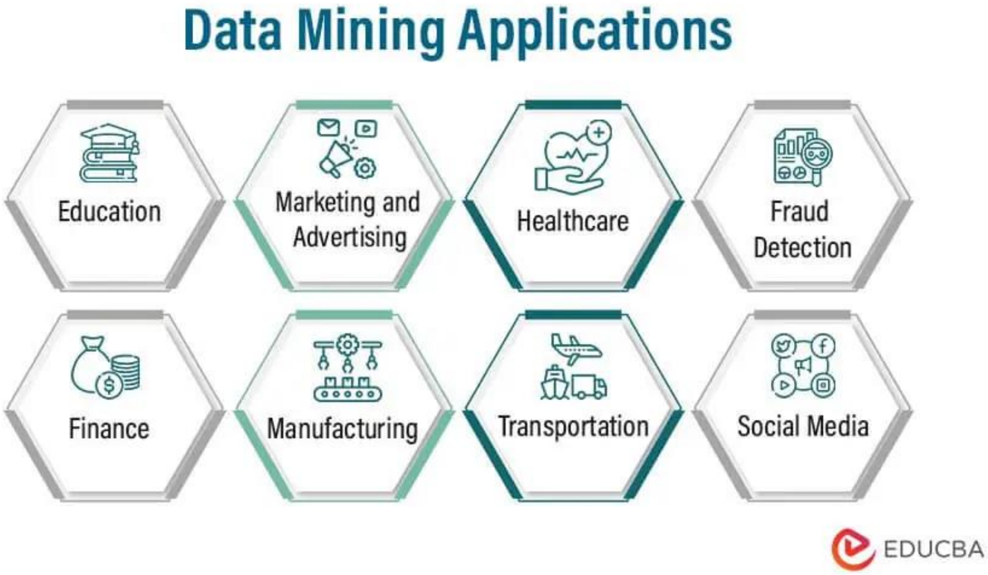
- Data Scientist
- Data Analyst
- Data Engineer
- Data Mining Specialist
- Data Mining Consultant

Computer Science Salaries



Burning Glass Technologies, "Labor Insight Real-Time Labor Market Information Tool."
(2022) http://www.burning-glass.com

DATA MINING APPLICATIONS



WHY DATA MINING?

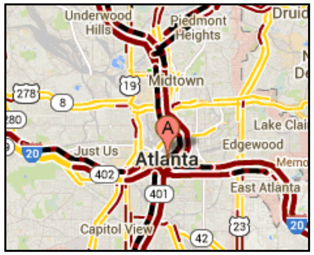
“We are living in the information age” is a popular saying; however, we are living in the data age.

LARGE-SCALE DATA IS EVERYWHERE!

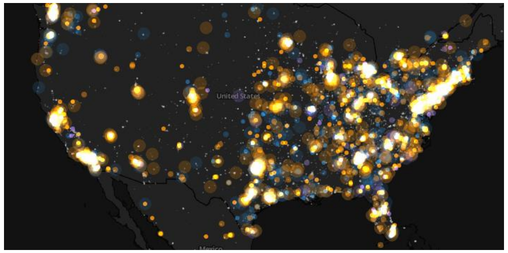
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies

- New mantra
  - Gather whatever data you can whenever and wherever possible.

E-Commerce

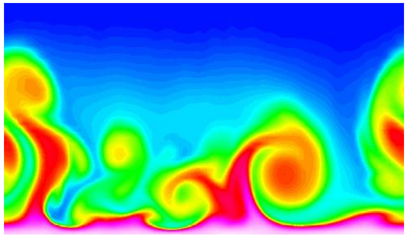


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

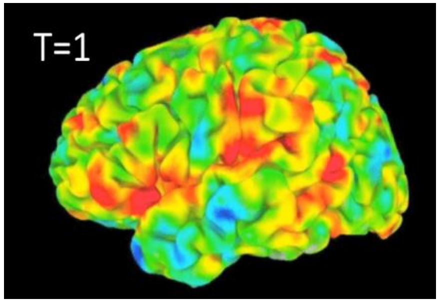
WHY DATA MINING? SCIENTIFIC VIEWPOINT

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

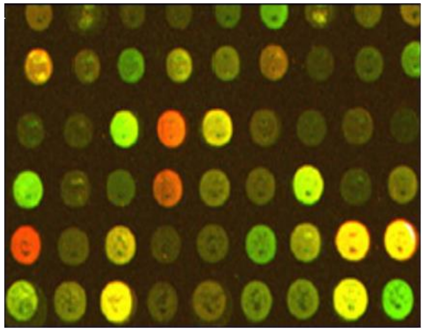
Data mining helps scientists

- in automated analysis of massive datasets

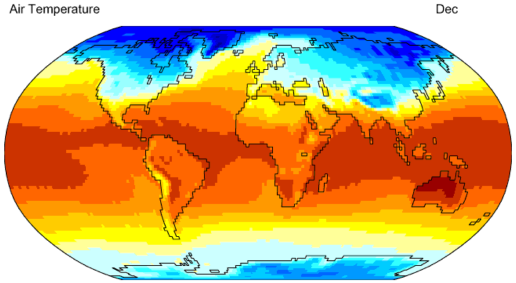


fMRI Data from Brain

Sky Survey Data



Gene Expression Data



Surface Temperature of Earth

WHY DATA MINING? COMMERCIAL VIEWPOINT

Lots of data is being collected and warehoused

- Web data
  - Google 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

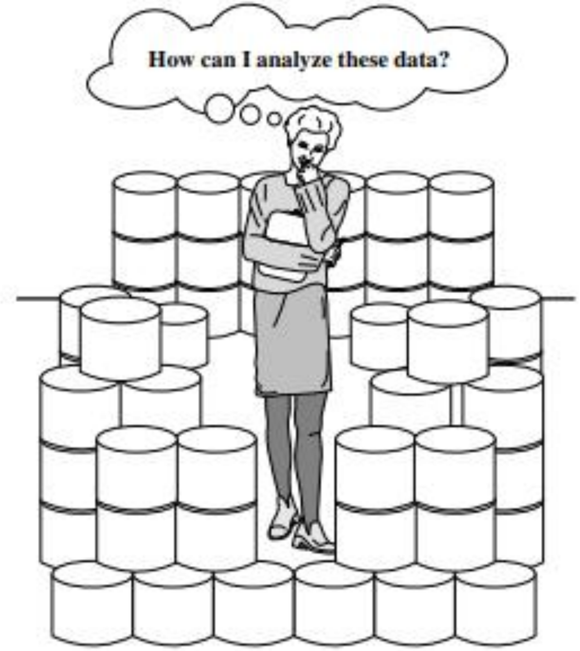
Competitive Pressure is Strong

WORLD IS DATA RICH BUT INFORMATION POOR

This explosively growing, widely available, and gigantic body of data makes our time truly the data age.

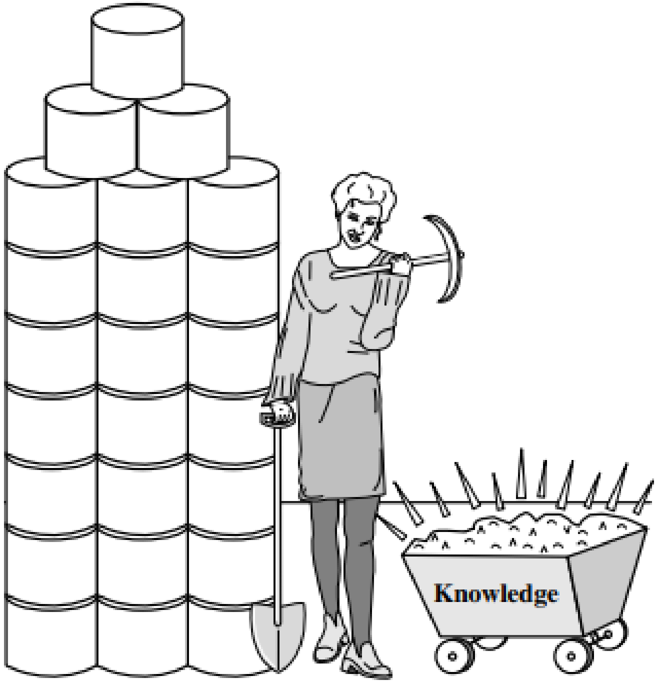
Powerful and versatile tools are badly needed to automatically uncover valuable information from the amounts of data and to transform such data into organized ones.

This necessity has led to the birth of data mining.



USE DATA MINING

Data mining turns a large collection of data into knowledge



WHAT IS DATA MINING

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery

KDD is the overall process of converting raw data into useful information

Data mining (an essential process where intelligent methods are applied to extract data patterns)

CHEF – DATA SCIENTIST ANALOGY

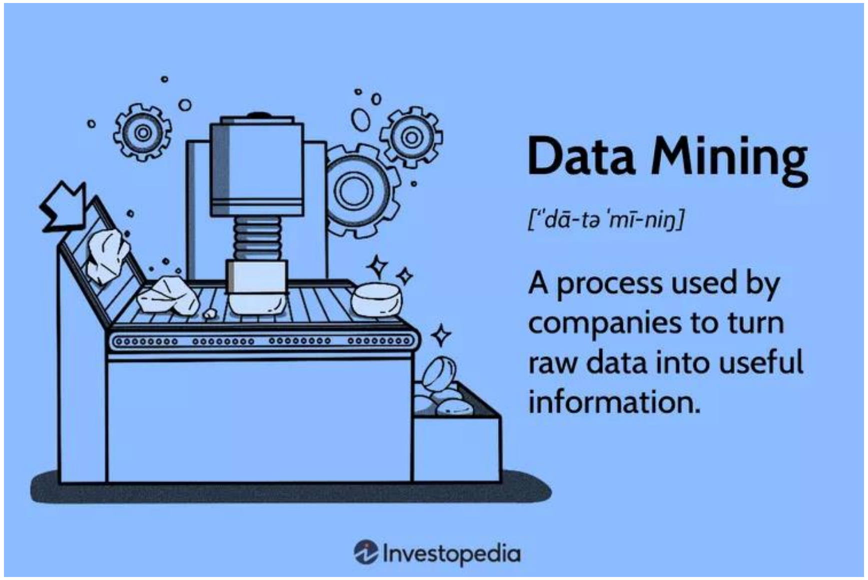
Chef	Data Scientist
Chef carefully selects, cleans, and prepares raw vegetables	A data scientist curates and processes raw data with precision.
Just as a chef combines various vegetables to create a harmonious flavor profile	A data scientist integrates different values in a data set to extract meaningful
The cooking process for vegetables involves techniques such as chopping, sautéing, and seasoning to enhance their taste and texture	Similarly, data preprocessing, transformation, and analysis refine raw data into meaningful information
A carefully crafted recipe results in a delicious meal	While meticulous data analysis leads to valuable insights and informed decision-making.

WHAT IS DATA MINING? EXPLAINED

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



Data Mining

['dā-tə 'mī-niŋ]

A process used by companies to turn raw data into useful information.

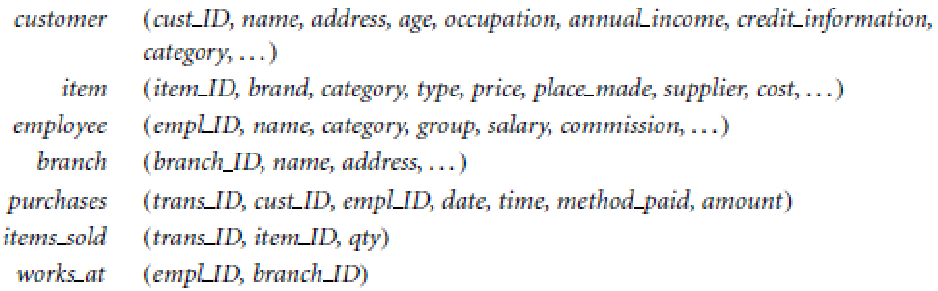
Investopedia

EXAMPLE :

A database for *All Electronics*

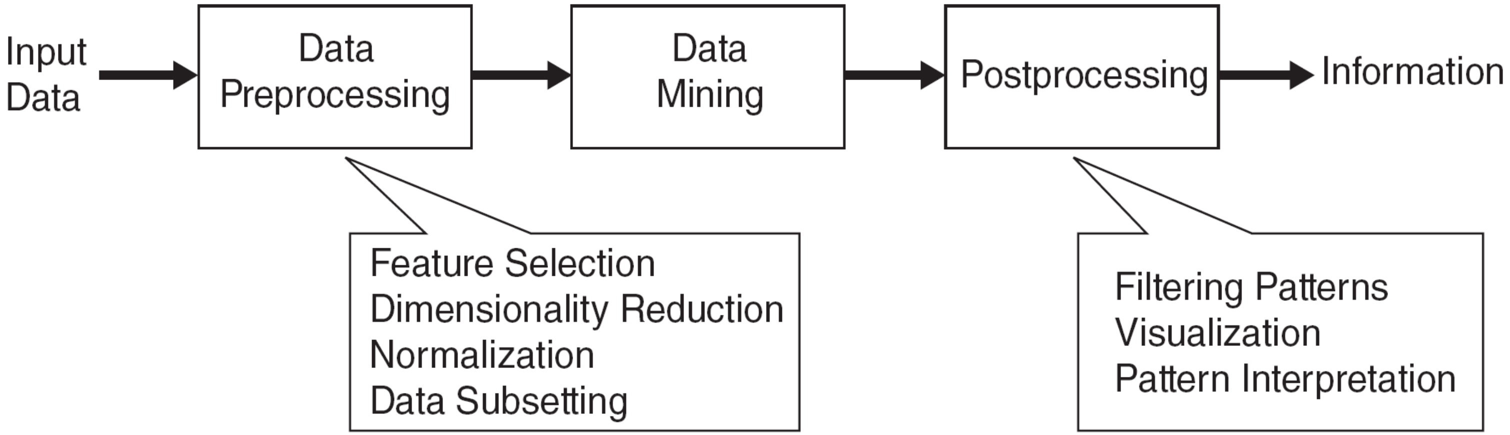
A customer relationship manager at All Electronics may order the following data mining task: Summarize the characteristics of customers who spend more than $5000 a year at All Electronics.

The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.



customer	*(cust_ID, name, address, age, occupation, annual_income, credit_information, category, . . .)*
item	*(item_ID, brand, category, type, price, place_made, supplier, cost, . . .)*
employee	*(empl_ID, name, category, group, salary, commission, . . .)*
branch	*(branch_ID, name, address, . . .)*
purchases	*(trans_ID, cust_ID, empl_ID, date, time, method_paid, amount)*
items_sold	*(trans_ID, item_ID, qty)*
works_at	*(empl_ID, branch_ID)*

KNOWLEDGE DISCOVER PROCESS



KNOWLEDGE DISCOVERY PROCESS

DEFINE THE PROBLEM

What goal are you seeking to achieve?

What type of data will you require to address the problem?

Why students are prone to drop out of college

DATA COLLECTION

Software tools (e.g., web crawlers) or specialized hardware (e.g., sensors).

Questionnaires and user surveys.

save data in Excel, CSV files, or a database. Whichever storage medium you choose, ensure you can easily retrieve the data when required.

DATA PROCESSING: WHY BOTHER?

The raw data you collect is usually not ready to be processed directly.

It may be in unsupported formats or structures or contain missing values, which can affect the overall data mining results.

The dataset may also contain mixed information. For example, you may find data on students who dropped out of college, but you some data points about students who graduated from college.

DATA PROCESSING

Data cleaning. In this stage, you need to deal with outliers, which are basically missing, abnormal, or incorrect entries. You may need to drop certain attributes or values and estimate missing entries for consistency.

Data integration. Data integration or aggregation involves combining data from different sources into a single file for easier analysis.

DATA PROCESSING

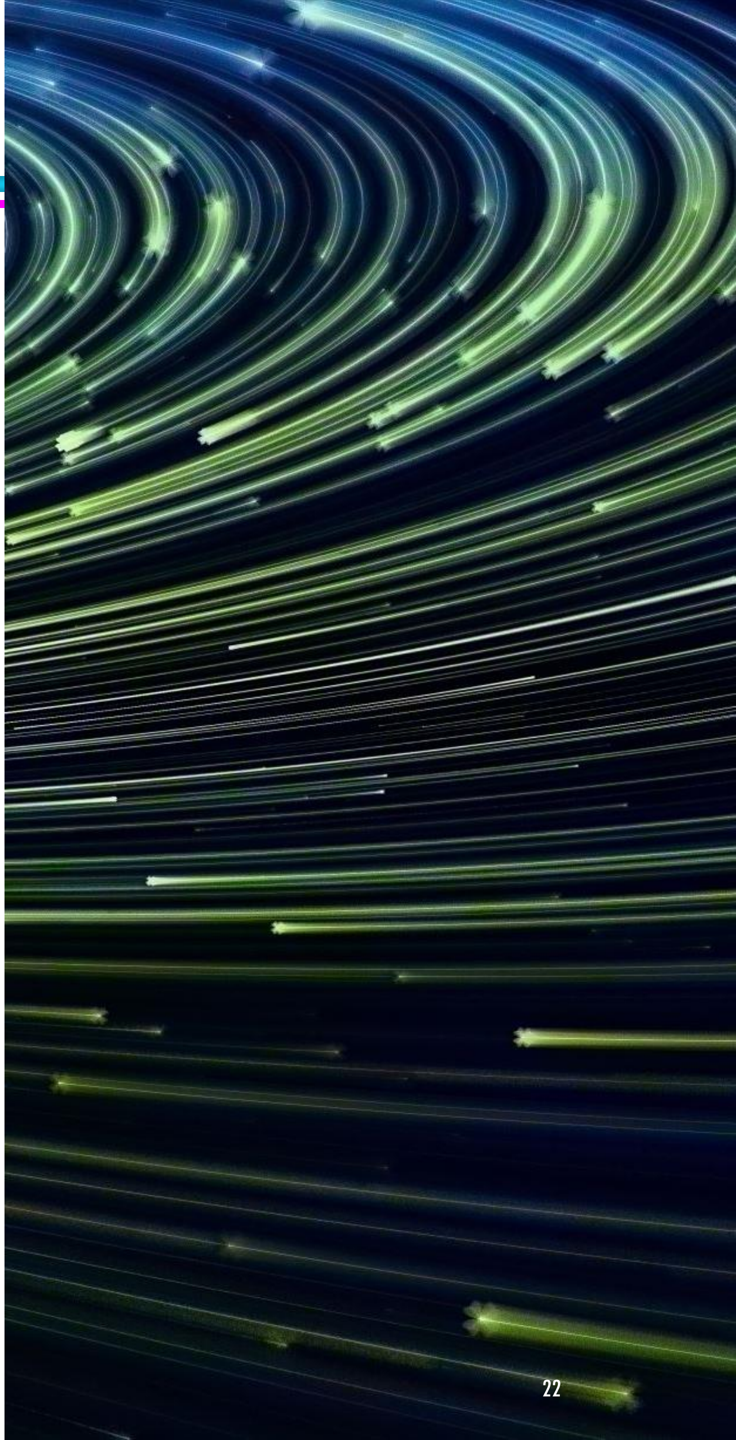
Feature extraction. In certain cases, you may collect and end up with large volumes of data. It's your responsibility to sort through this data to identify features that are relevant to your objective.

Data transformation. Often, the final result of the data collection stage is usually high-dimensional data, which may need to be clustered or otherwise transformed to make it better suited for data analysis.

MODEL ESTIMATION

A data mining model analyzes raw data, allowing the identification and understanding of different trends and patterns.

Each data mining process is unique; therefore, each requires a specialized model to be evaluated well.



RESULT ANALYSIS

One of the final stages of a data mining project involves interpreting the model's results. A good data mining model should allow you to make quality decisions quickly. Unless required, avoid complex models that are difficult to interpret; these often need a lot of time and expertise to understand.

DRAWING CONCLUSIONS

Drawing conclusions is an essential part of data mining. This phase is highly dependent on your ability to understand results from the use of data mining techniques. You also need to summarize what you've learned from the data mining process and assess the strengths of the models.

Use the results for personal decision-making or present them to the management team to determine how to use the information discovered from the data.

DATA MINING AND KDD

The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation. However, in industry, in media, and in the research milieu, the term *data mining* is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than *knowledge discovery from data*). Therefore, we adopt a broad view of data mining Functionality.

Data mining is the *process* of discovering interesting patterns and knowledge from *large* amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

DATA MINING FUNCTIONALITIES

There are several ***data mining functionalities***.

- ❑Characterization and discrimination
- ❑The mining of frequent patterns, associations, and correlations
- ❑Classification and Regression
- ❑Clustering analysis
- ❑Outlier analysis

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.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Characterization and Discrimination

CHARACTERIZATION

Data characterization is a summarization of the general characteristics or features of a target class of data.

Example:

Manager at All Electronics may order the following data mining task: Summarize the characteristics of customers who spend more than $5000 a year at All Electronics.

The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings.

DATA DISCRIMINATION

Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. The target and contrasting classes can be specified by a user.

Example :

All Electronics Manager may want to compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g., less than three times a year).

The resulting description provides a general comparative profile of these customers, such as that 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree

Classification and Regression

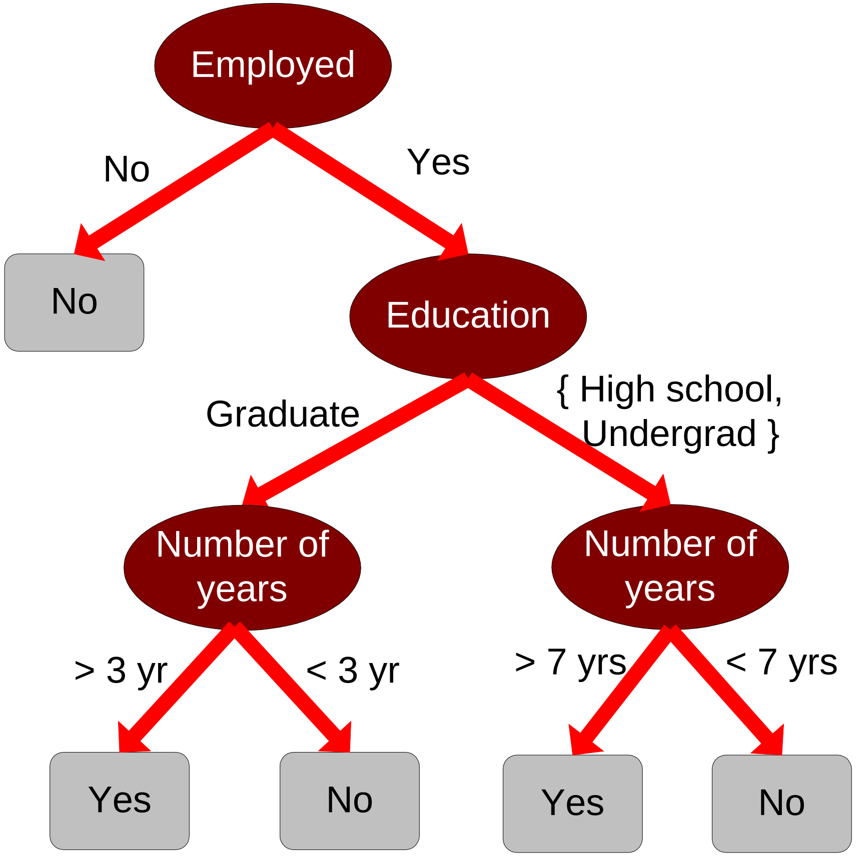
Classification

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

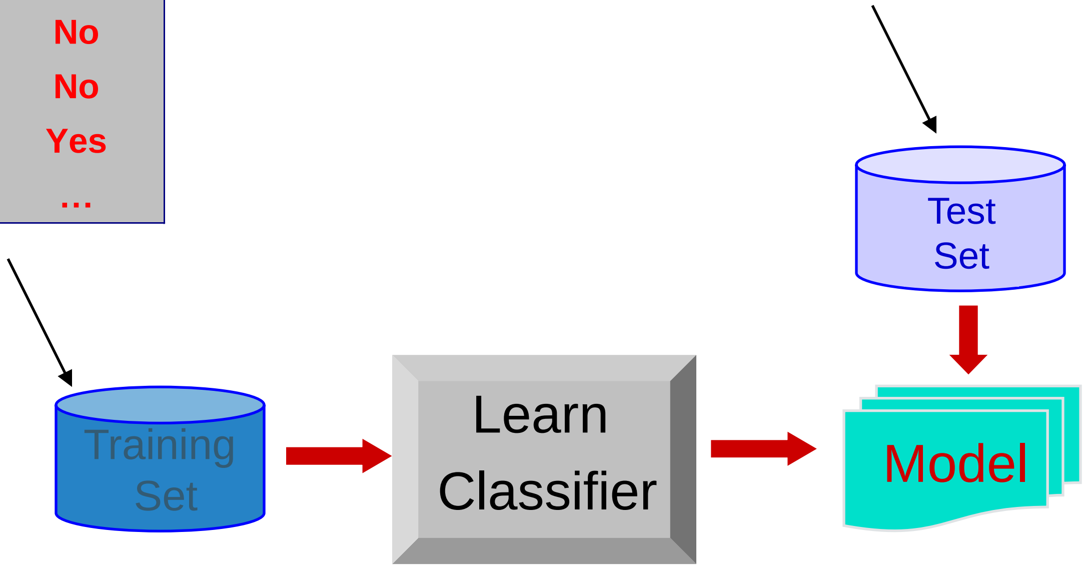


CLASSIFICATION EXAMPLE

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
…	…	…	…	…

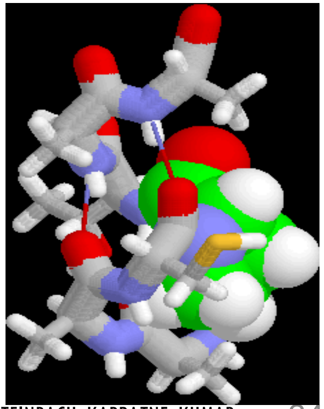
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



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

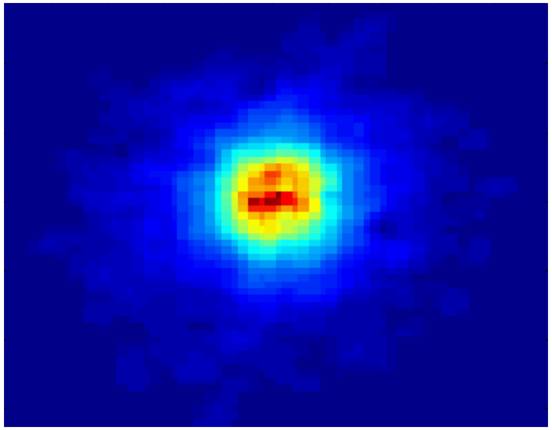


- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

CLASSIFYING GALAXIES

Courtesy: http://aps.umn.ed

Early



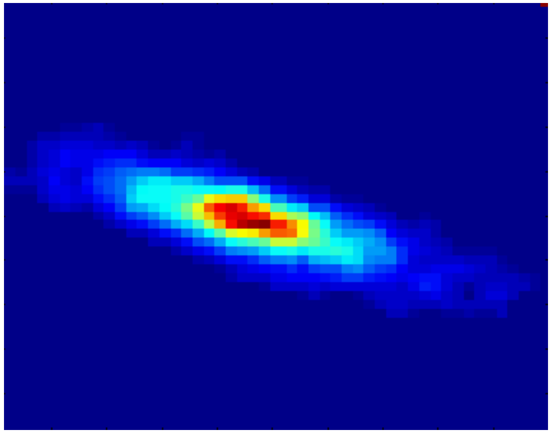
Class:

- Stages of Formation

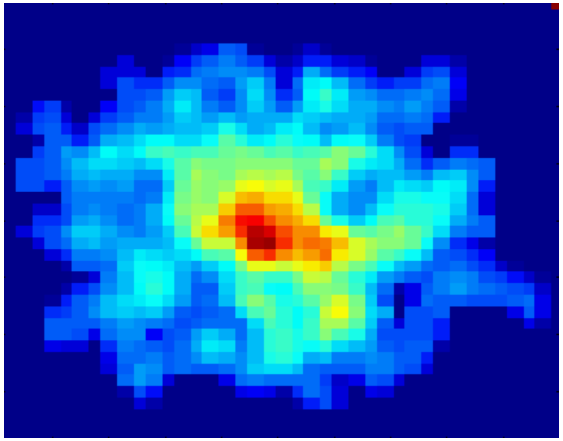
Attributes:

- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late



Data Size:

- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

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.

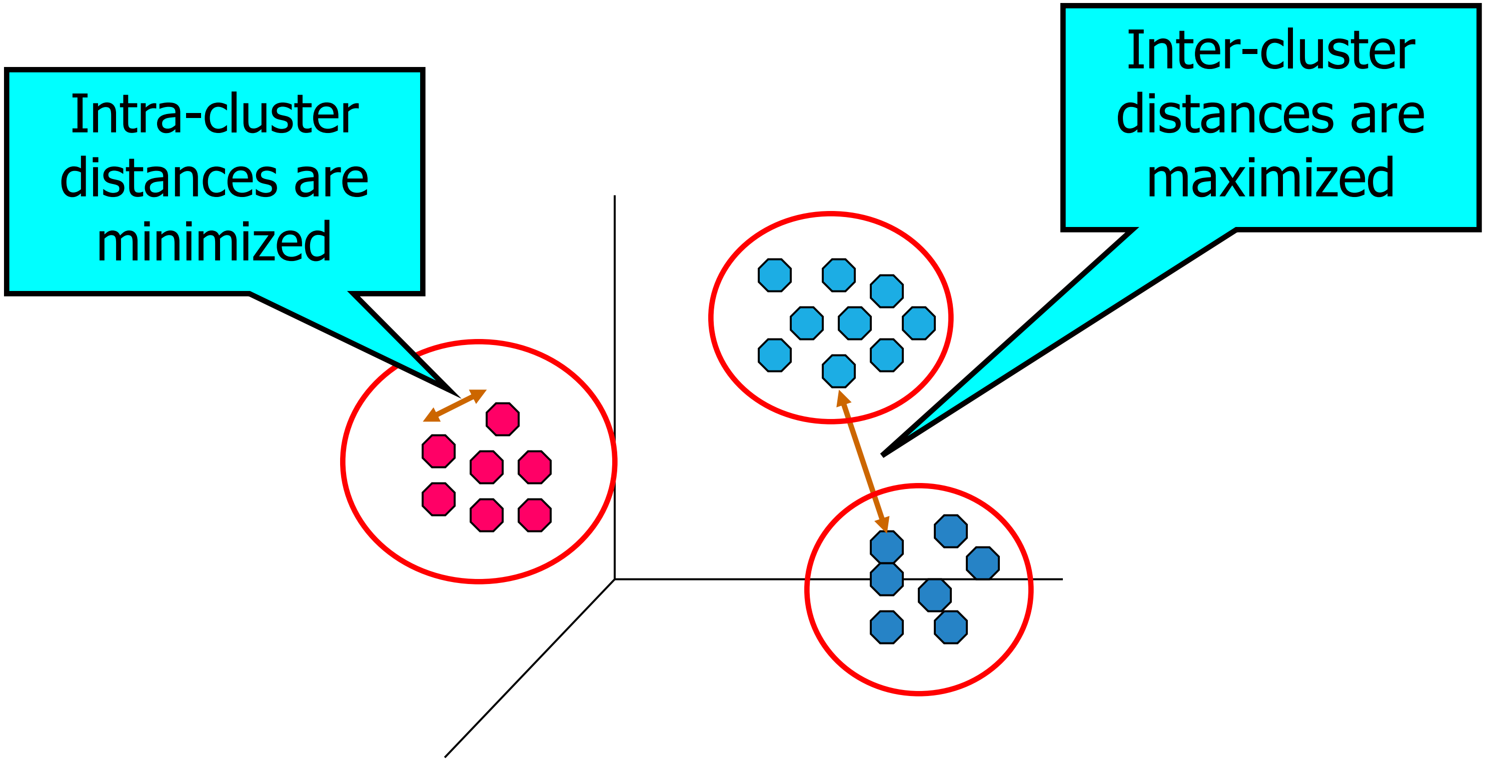
Extensively studied in statistics, neural network fields.

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


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

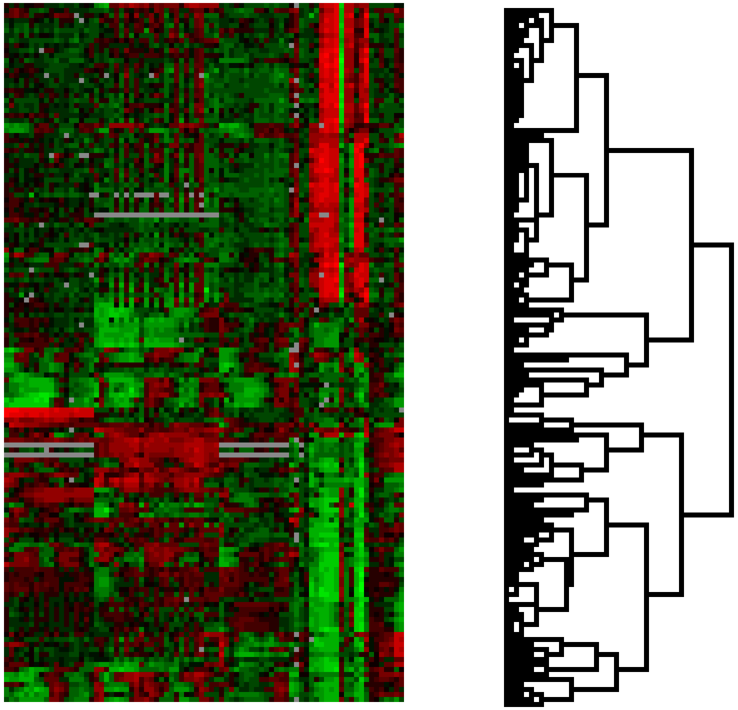
APPLICATIONS OF CLUSTER ANALYSIS

Understanding

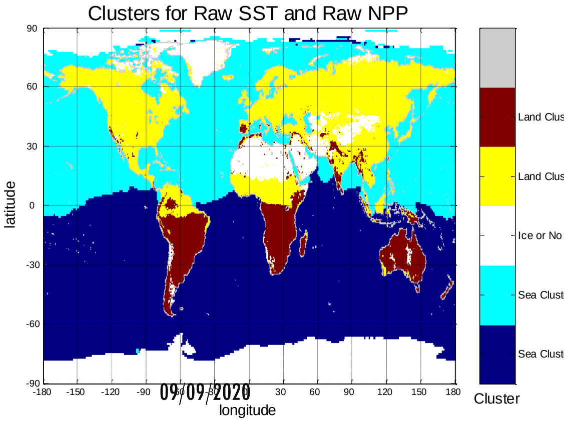
- Custom profiling for targeted marketing
- Group related documents for browsing
- Group genes and proteins that have similar functionality
- Group stocks with similar price fluctuations

Summarization

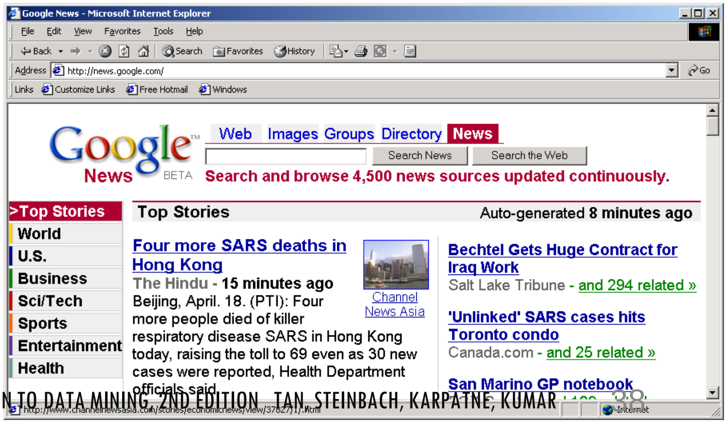
- Reduce the size of large data sets



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.



Search and browse 4,500 news sources updated continuously.

Top Stories — Auto-generated 8 minutes ago

Four more SARS deaths in Hong Kong
The Hindu - 15 minutes ago
Beijing, April. 18. (PTI): Four more people died of killer respiratory disease SARS in Hong Kong today, raising the toll to 69 even as 30 new cases were reported, Health Department officials said.
Channel News Asia

Bechtel Gets Huge Contract for Iraq Work
Salt Lake Tribune - and 294 related »

'Unlinked' SARS cases hits Toronto condo
Canada.com - and 25 related »

San Marino GP notebook

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

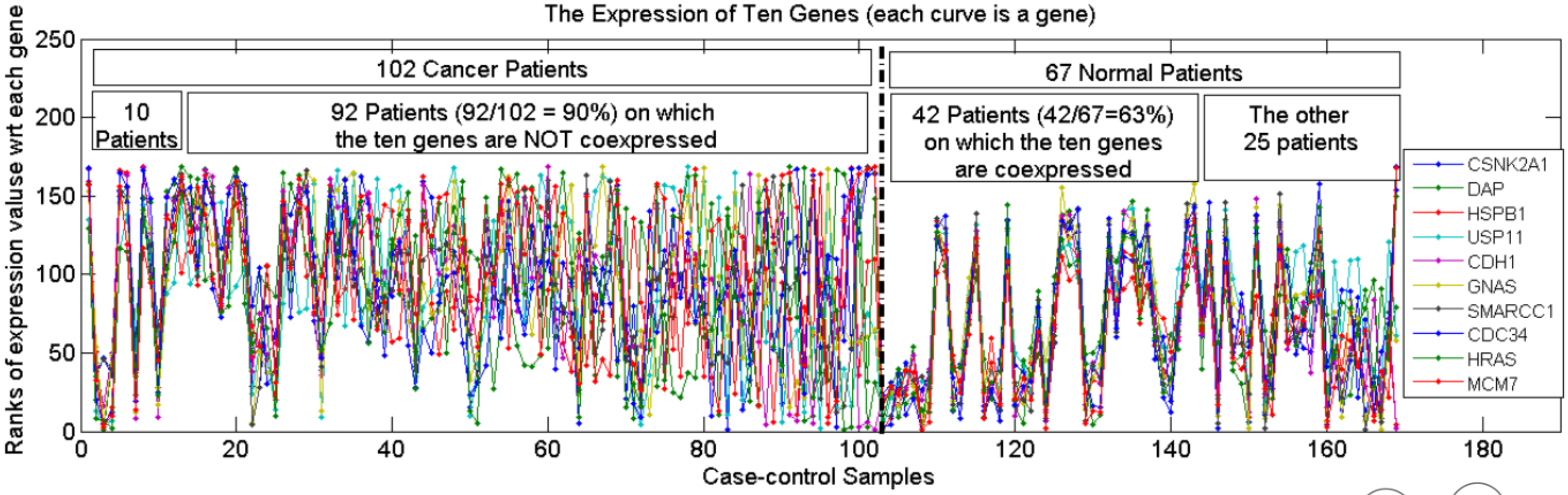
Medical Informatics

- Rules are used to find combination of patient symptoms and test results associated with certain diseases

Association Analysis: Applications

An Example Subspace Differential Coexpression Pattern from lung cancer dataset

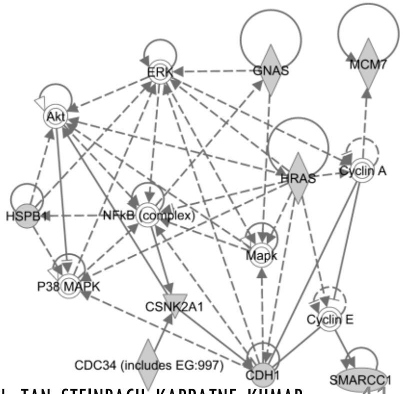
Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]



Enriched with the TNF/NFB signaling pathway

which is well-known to be related to lung cancer

P-value: 1.4*10^{-5} (6/10 overlap with the pathway)



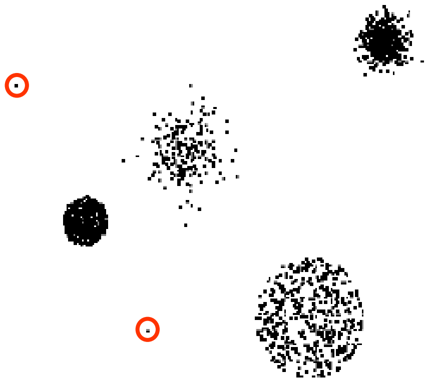
[Fang et al PSB 2010]

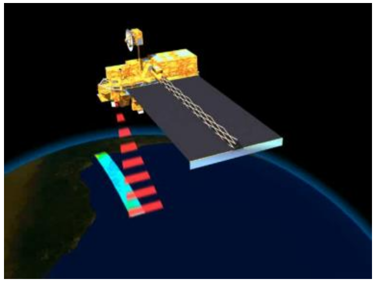
DEVIATION/ANOMALY/CHANGE DETECTION

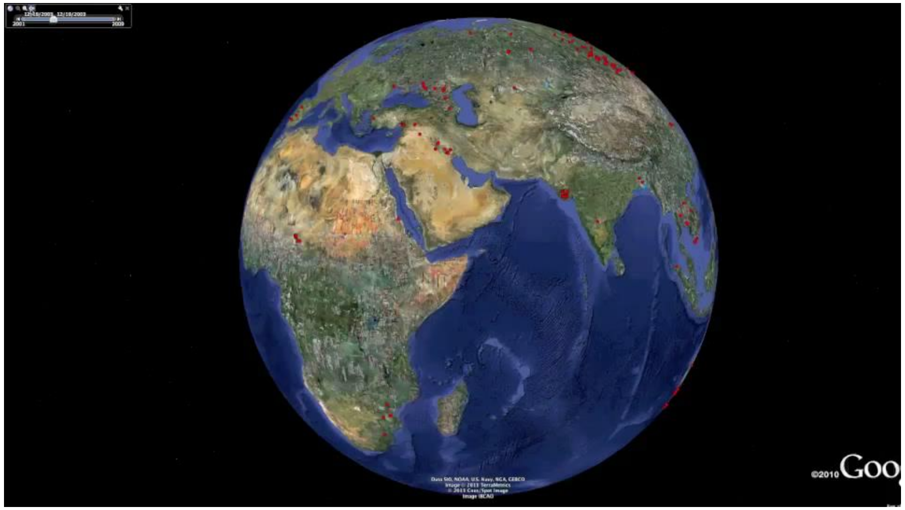
Detect significant deviations from normal behavior

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.







MOTIVATING CHALLENGES

Scalability

High Dimensionality

Heterogeneous and Complex Data

Data Ownership and Distribution

Non-traditional Analysis