Data Mining: Introduction

Lecture Notes for Chapter 1

Introduction to Data Mining

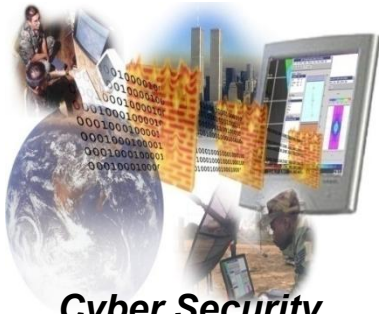
Outline

Unit 1: Introduction to Data Mining [3 Hrs.]

- Data Mining
- Data Mining Goals
- Data Mining Tools and Technologies
- Data Mining Methodologies: CRISP, DM, KDD, SEMMA

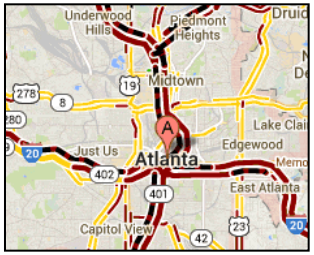
Large-scale Data is Everywhere!

- 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.

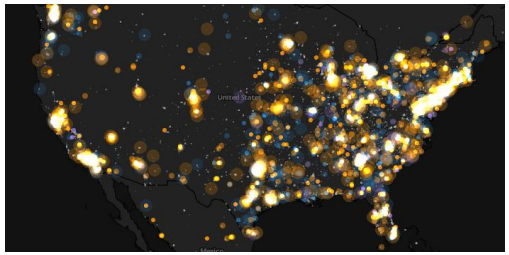


Cyber Security

E-Commerce

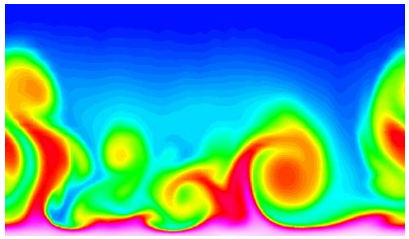


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

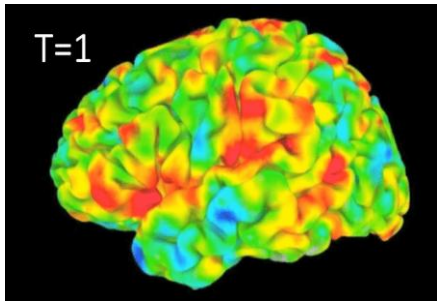
Why Data Mining? Commercial Viewpoint

- 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
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

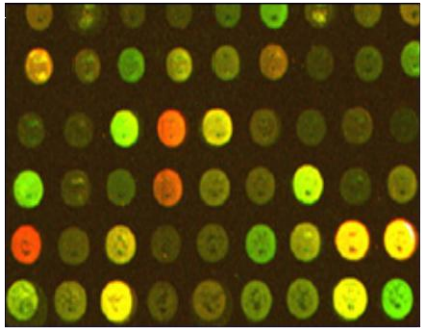
Why Data Mining? Scientific Viewpoint

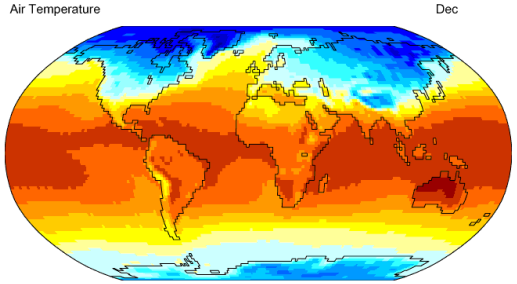
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated 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

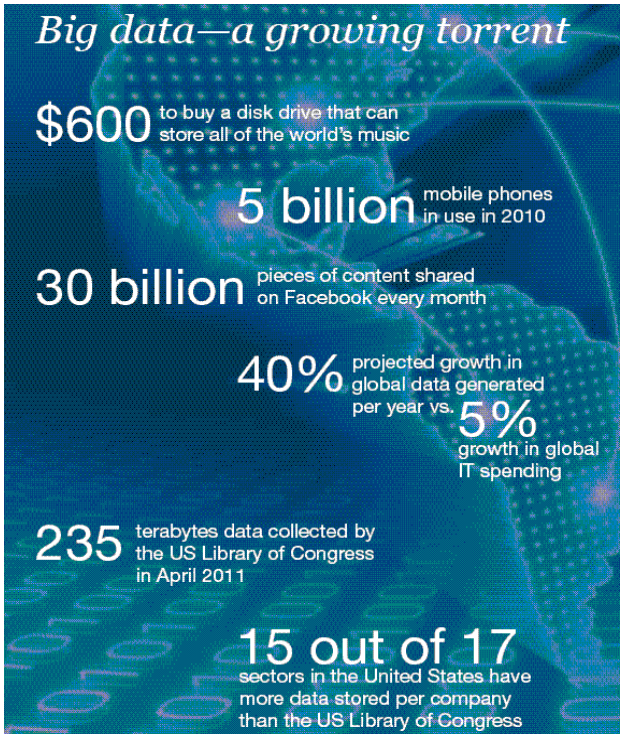


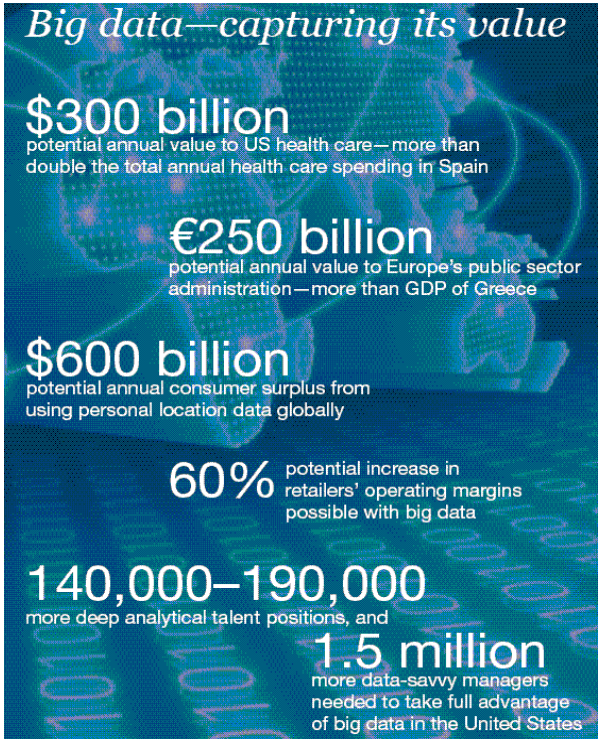
Surface Temperature of Earth

Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

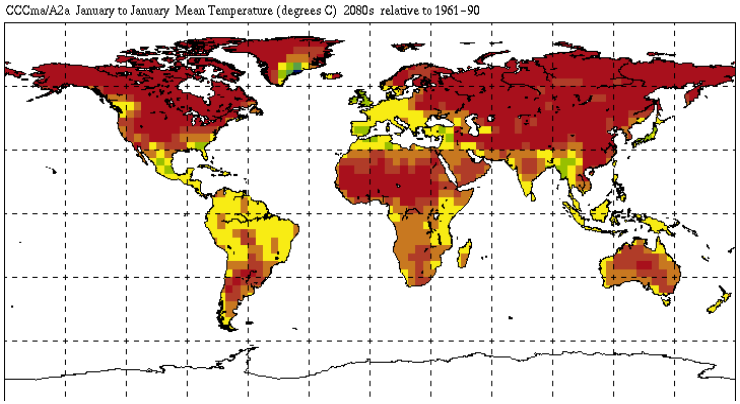
Big data: The next frontier for innovation, competition, and productivity





Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



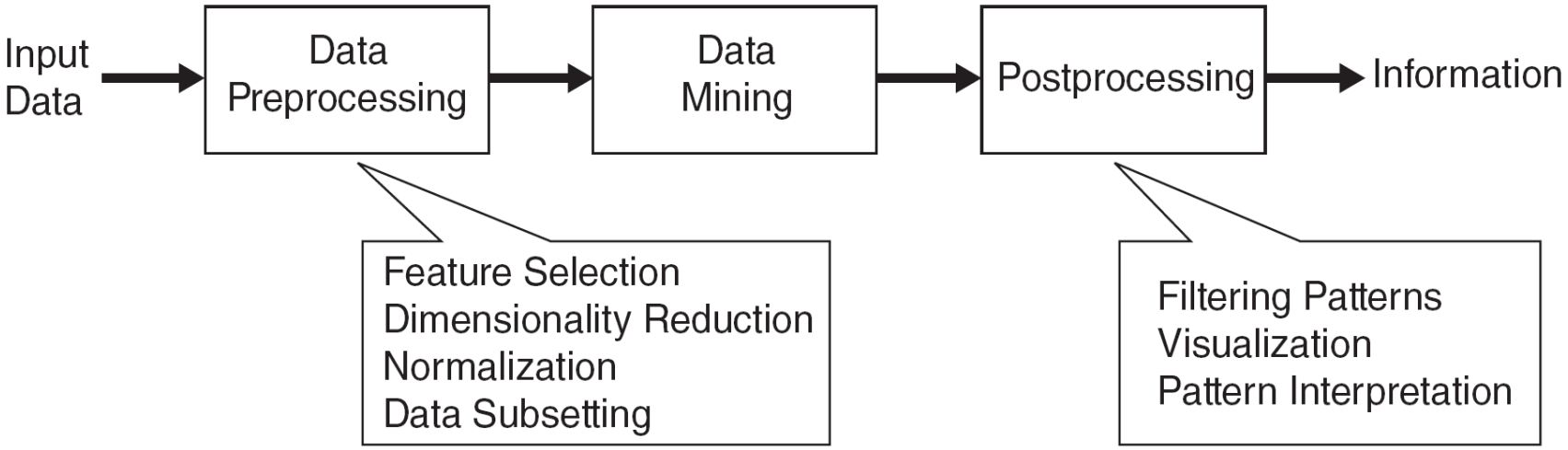
Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

What is Data Mining?

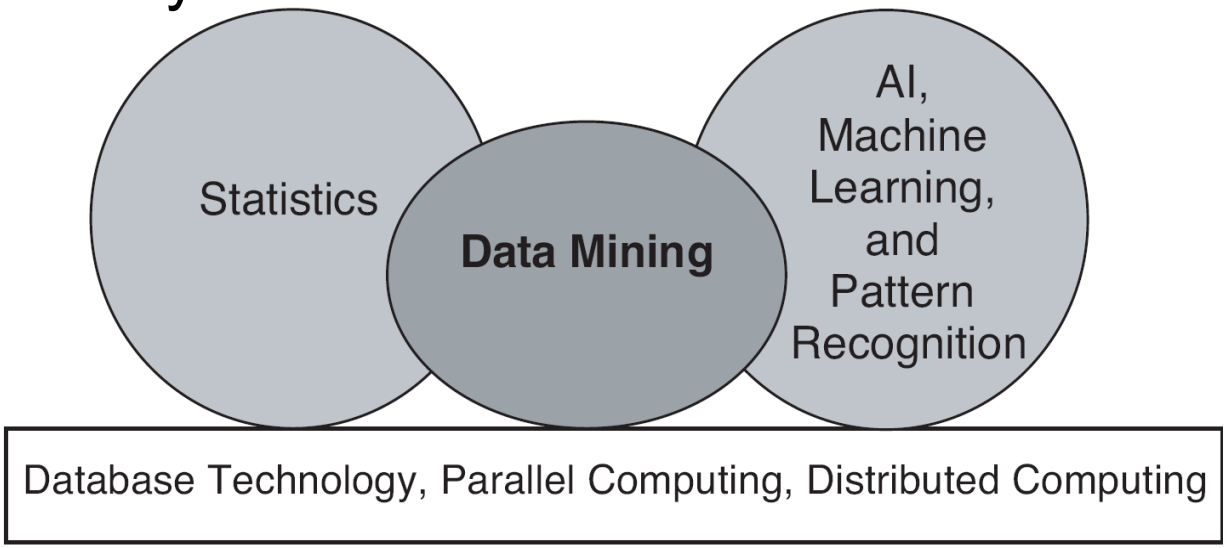
- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



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

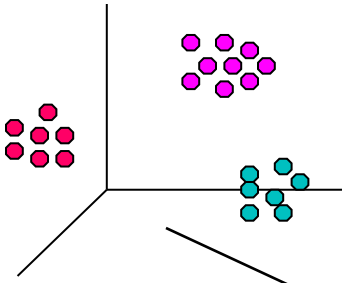
Data Mining Tasks

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

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

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

Data Mining Tasks ...

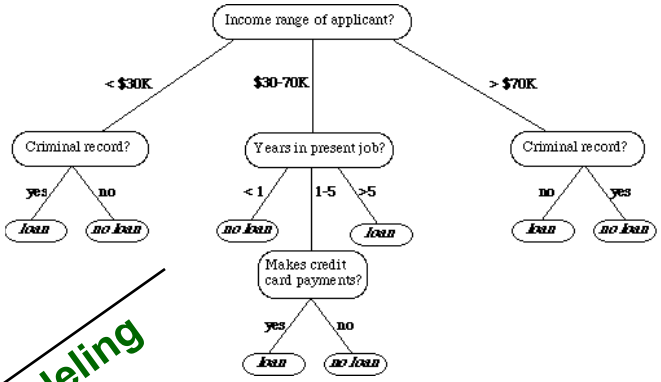


Clustering

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

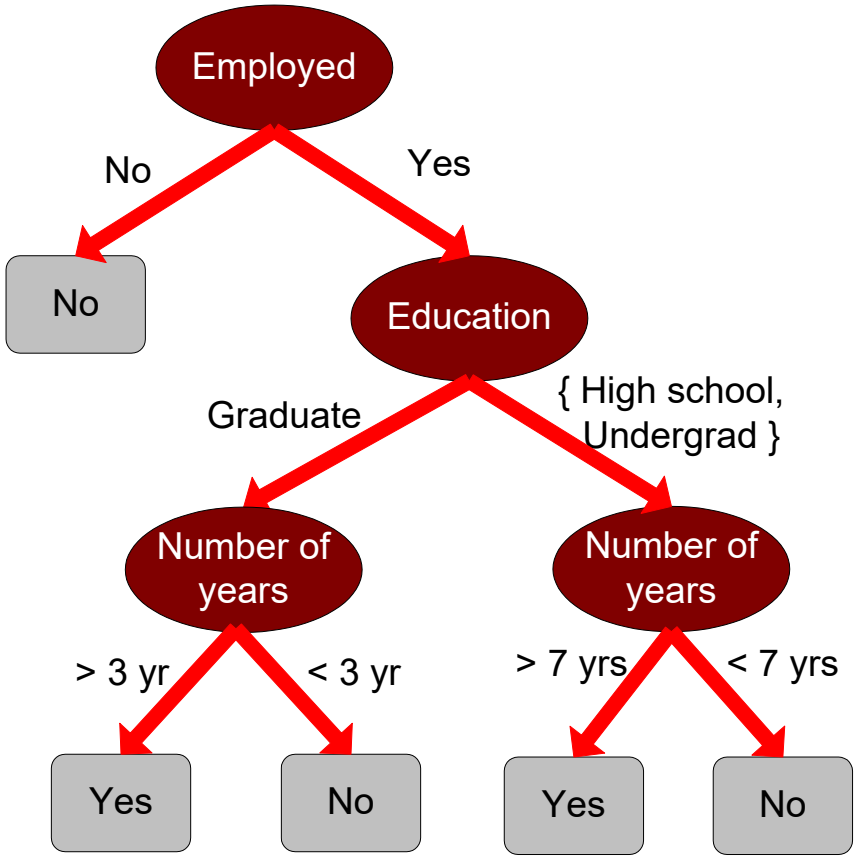
Predictive Modeling: Classification

- 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

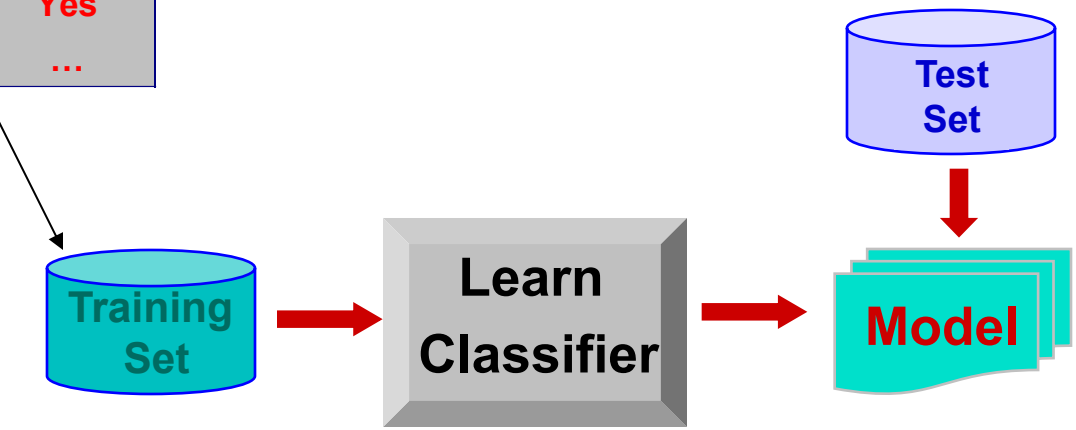


Classification Example

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	?
…	…	…	…	…

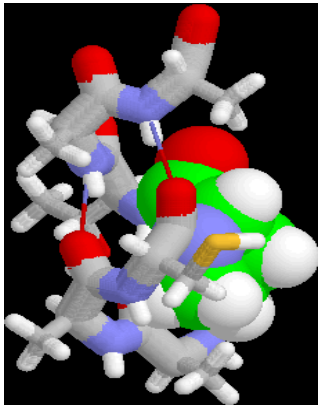


Examples of Classification Task

- 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

- Churn prediction for telephone customers
  - **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.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 3

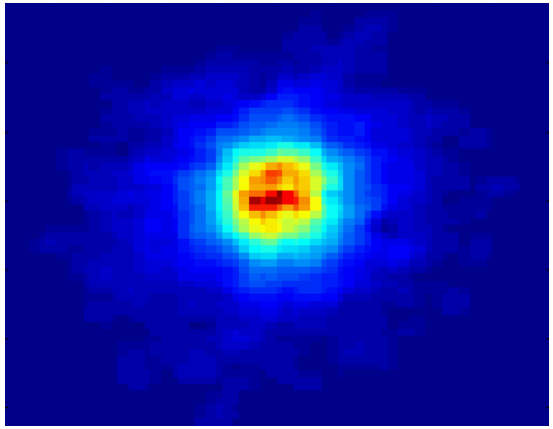
- Sky Survey Cataloging
  - **Goal:** To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - **Approach:**
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

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

Classifying Galaxies

Courtesy: http://aps.umn.edu

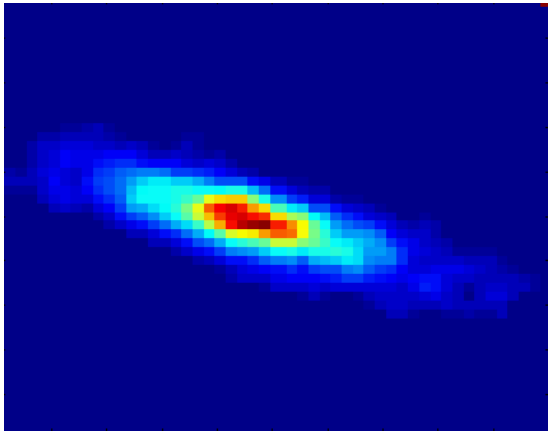
Early



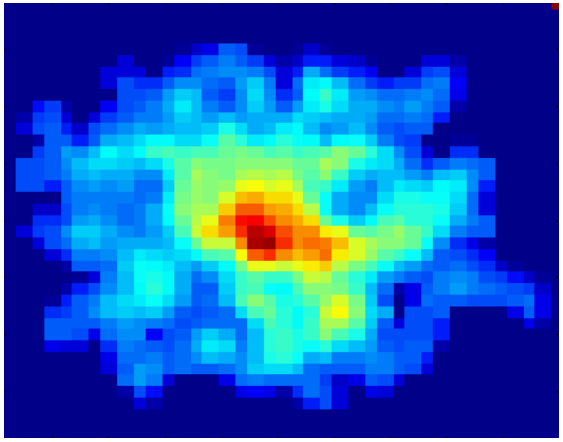
Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



Late



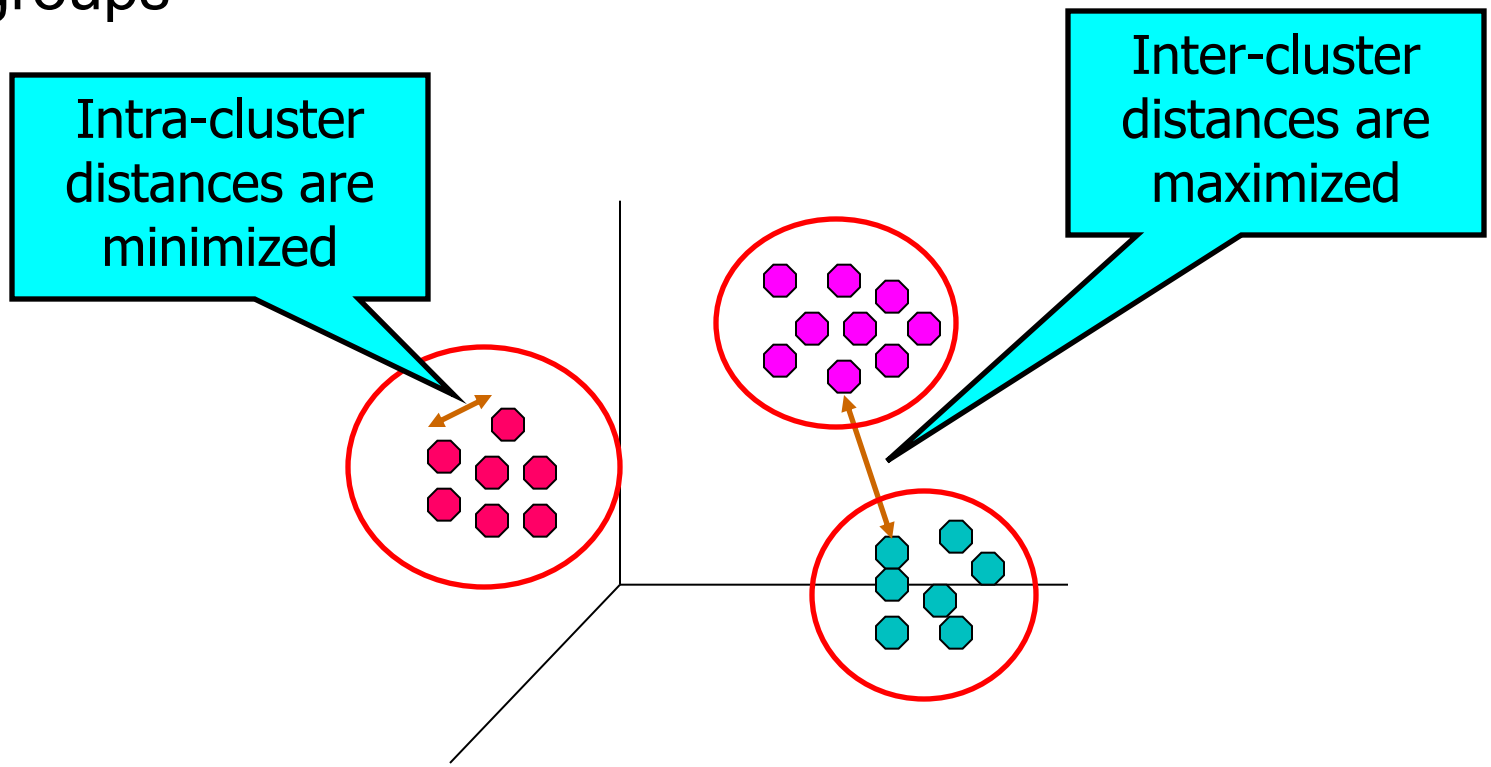
Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

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

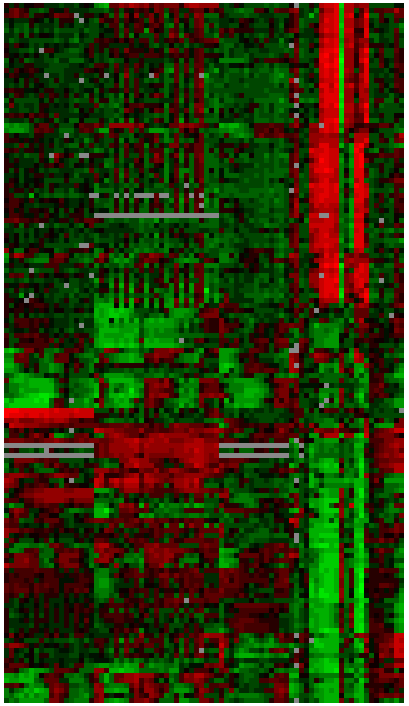
Clustering

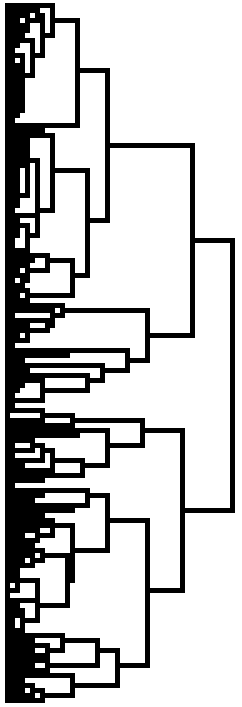
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



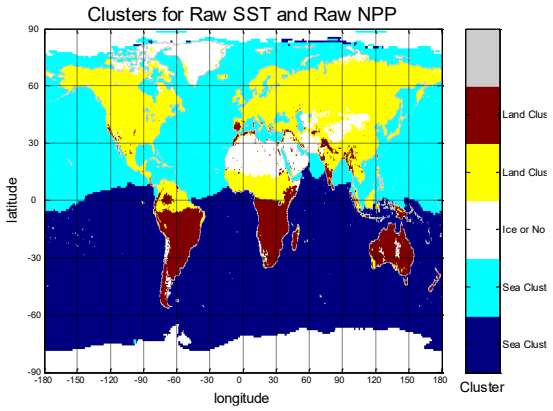
Applications of Cluster Analysis

- **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

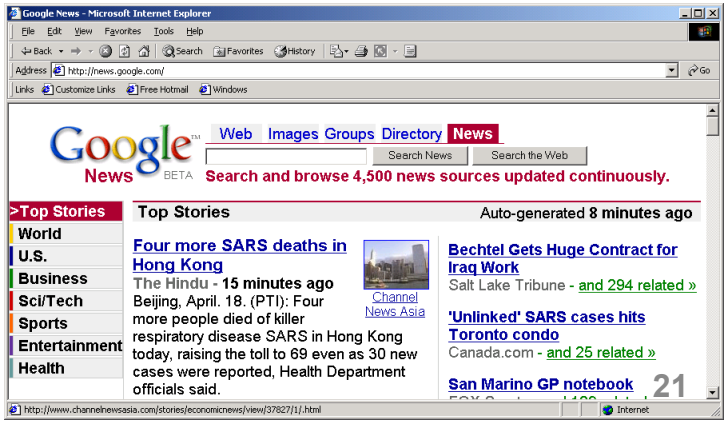




Courtesy: Michael Eisen



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



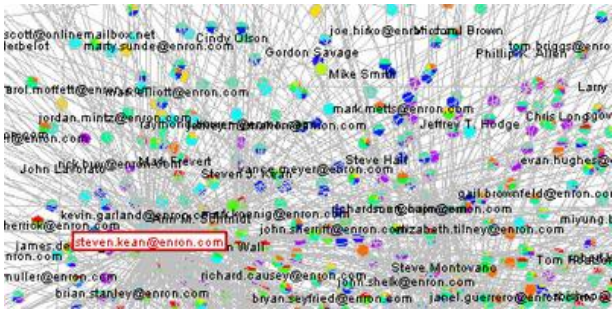
Clustering: Application 1

- 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.

Enron email dataset



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}

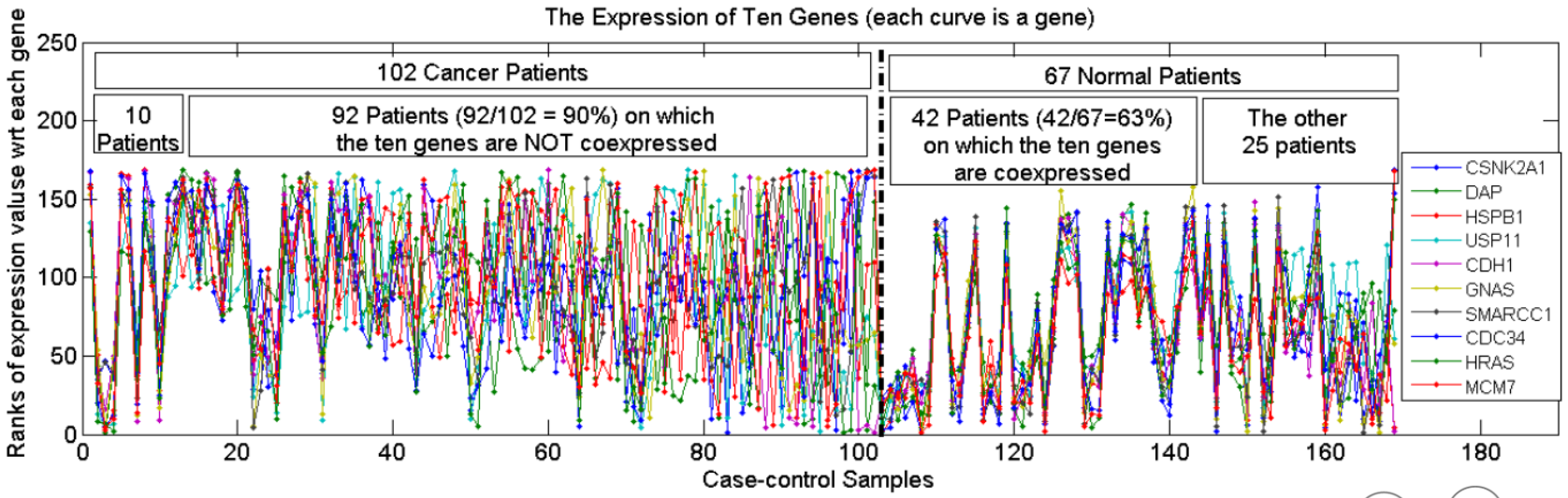
Association Analysis: Applications

- 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

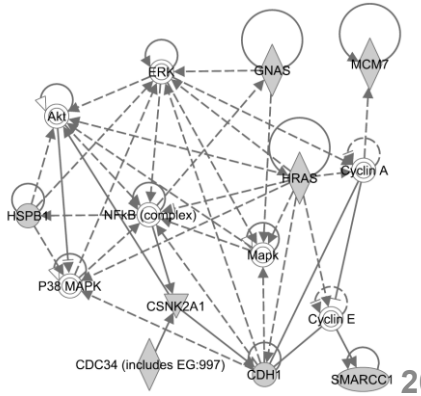
- An Example Subspace Differential Coexpression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjee et a 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

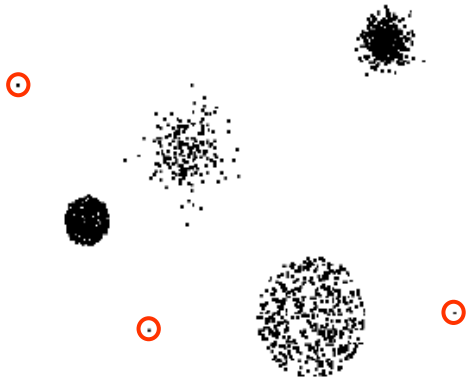
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

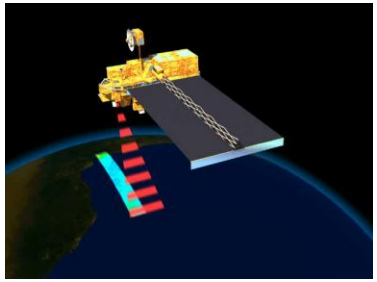


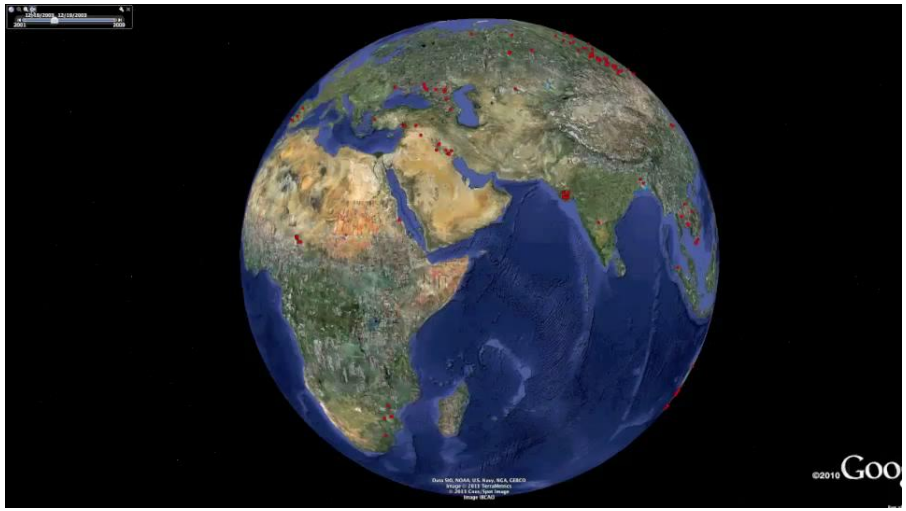
[Fang et al PSB 2010]

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.







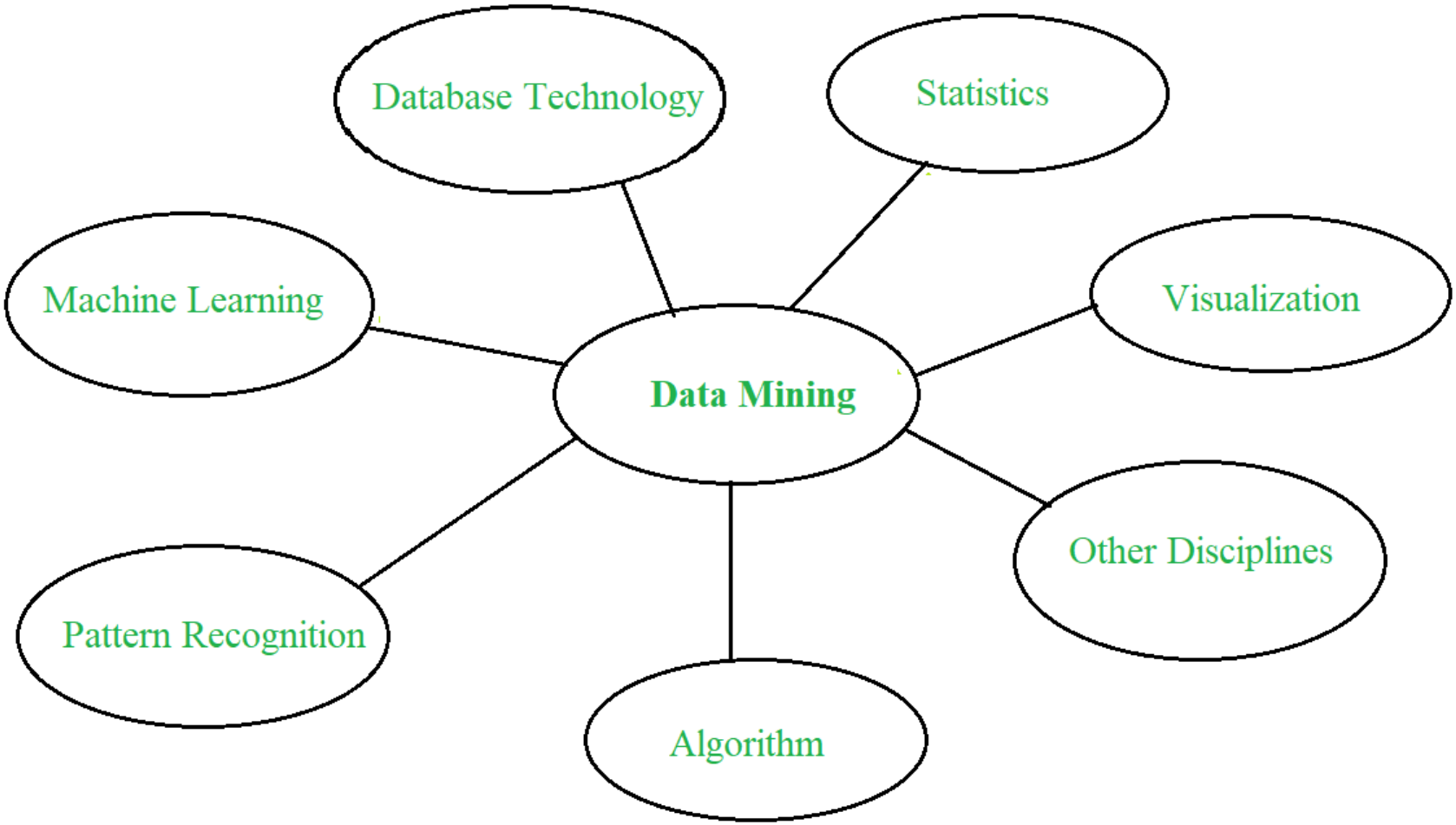
Motivating Challenges

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

Data Mining Tools and Technologies

Tools	Description
RapidMiner	Drag-and-drop interface for data analysis and predictive modeling.
WEKA	Open-source software with machine learning algorithms.
KNIME	Open-source platform for data analytics and reporting.
Orange	Visual programming tool with interactive data visualization.
R & Python	Programming languages widely used for statistical computing and data mining.
SAS Enterprise Miner	Commercial tool for predictive analytics and data mining.

Technologies Used in Data Mining



Data Mining Methodologies

- **CRISP-DM**, KDD, SEMMA
- CRISP-DM
  - Cross-Industry Standard Process for Data Mining

...

Business Understanding:

- Focuses on understanding the project objectives and requirements from a business perspective.
- Determines the business objectives, assesses the situation, defines the data mining goals, and produces a project plan.

Data Understanding:

- Involves collecting initial data, describing the data, exploring the data, and verifying data quality.
- Helps in understanding the data's characteristics and identifying any data quality issues.

...

Data Preparation:

- Covers all activities needed to construct the final dataset from the initial raw data.
- Includes tasks such as selecting data, cleaning data, constructing data, integrating data, and formatting data.

Modeling:

- Involves selecting modeling techniques, generating test design, building models, and assessing models.
- Different modeling techniques may require different data formats and assumptions.

...

Evaluation:

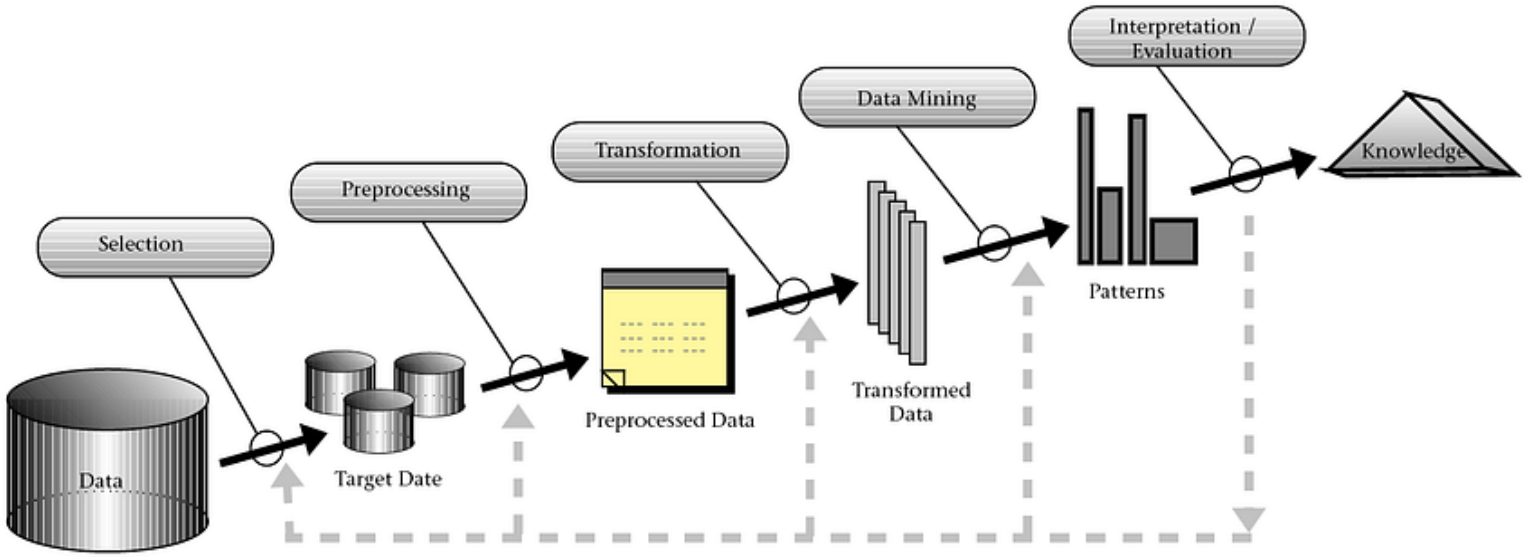
- Assesses the model thoroughly to ensure it meets the business objectives.
- Involves reviewing the steps executed, ensuring that the model is achieving the intended goals, and deciding on the next steps.

Deployment:

- Involves deploying the model into the operational environment for use.
- Can include generating reports, implementing the model within an application, or creating a repeatable data mining process for ongoing use.

KDD

- Knowledge Discovery in Databases (KDD) is a comprehensive process used in data mining and machine learning to extract useful knowledge from large datasets.



...

Selection:

- **Objective**: In the selection stage, the focus is on identifying and retrieving data from various sources that are relevant to the analysis and decision-making process.

Activities:

- Define the criteria for selecting data based on the problem domain and objectives.
- Gather data from databases, data warehouses, or other sources that meet the defined criteria.
- Ensure the data collected is comprehensive and representative of the problem at hand.

...

Pre-processing:

- **Objective**: Pre-processing involves cleaning and transforming the raw data to prepare it for further analysis.

Activities:

- Clean the data by handling missing values, outliers, and noise.
- Normalize or standardize data to ensure consistency and comparability across different variables.
- Feature selection or extraction to identify relevant attributes that contribute most to the analysis.

...

Transformation:

- **Objective**: Transformation aims to convert the pre-processed data into a format suitable for mining and modeling.

Activities:

- Aggregate or summarize data to reduce complexity and improve efficiency in subsequent analysis.
- Perform dimensionality reduction techniques such as PCA (Principal Component Analysis) to reduce the number of variables while retaining important information.
- Apply encoding techniques for categorical variables to convert them into numerical format if necessary.

...

Data Mining:

- **Objective**: Data mining involves applying algorithms and statistical methods to discover patterns, relationships, and insights from the transformed data.

Activities:

- Use various data mining techniques such as clustering, classification, association rule mining, and regression to extract patterns.
- Evaluate and compare different models to identify the most suitable one for the problem at hand.
- Iteratively refine models based on feedback and insights gained from the evaluation process.

...

Interpretation/Evaluation:

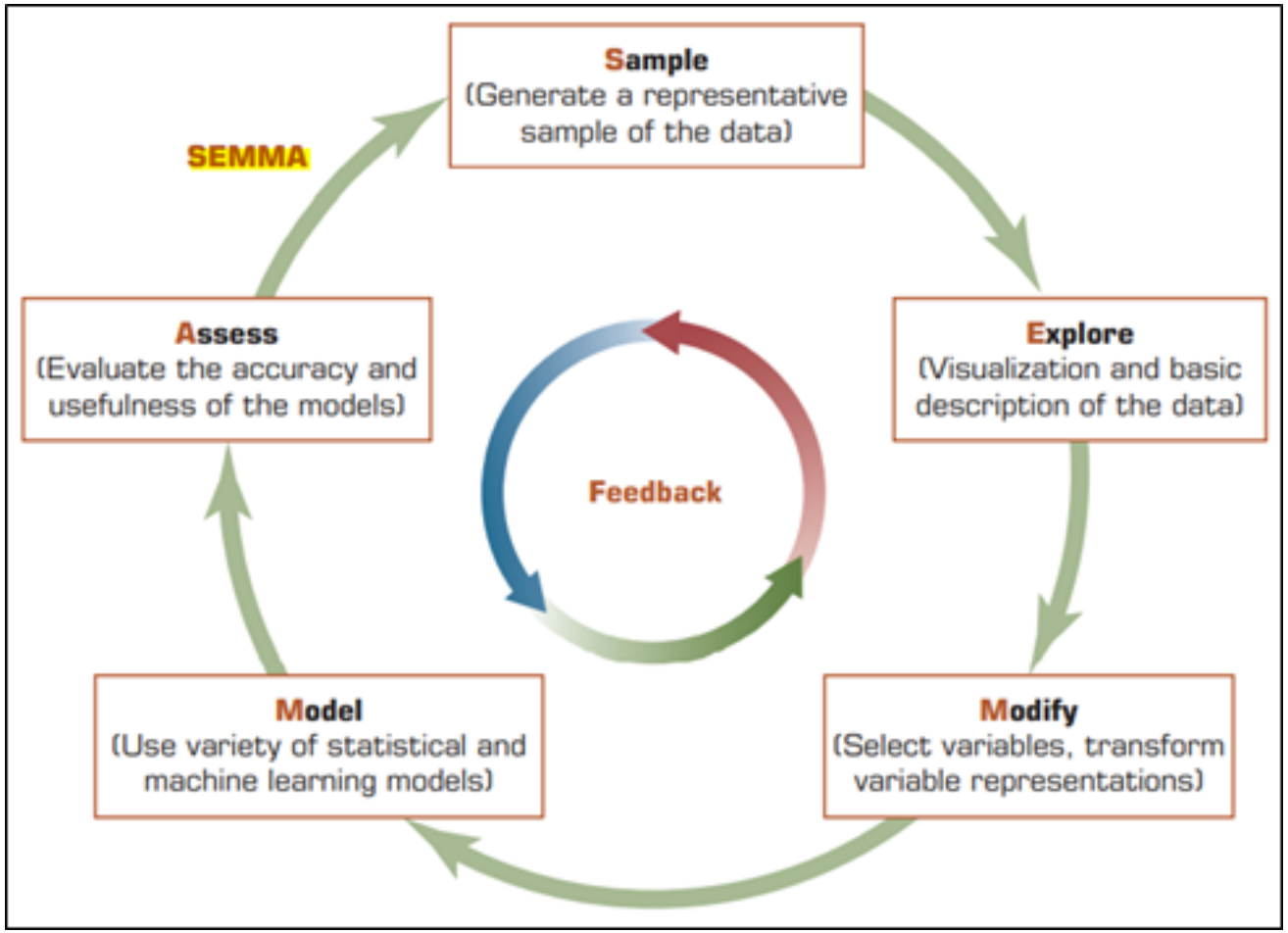
- **Objective**: The final stage focuses on interpreting the patterns discovered and evaluating the effectiveness and reliability of the models developed.

Activities:

- Interpret and visualize the patterns and relationships discovered during the data mining process.
- Assess the quality and relevance of the insights gained against the initial objectives and business goals.
- Communicate findings and recommendations to stakeholders in a clear and understandable manner.

SEMMA

- The SEMMA process is a methodology developed by SAS (Statistical Analysis System) for data mining and predictive modeling.



...

Sample:

- **Objective**: The first phase involves selecting a representative sample of data from the population for analysis. This sample should accurately reflect the characteristics of the entire dataset.

Activities:

- Define sampling criteria based on project goals and data characteristics.
- Randomly select samples from the dataset using appropriate sampling techniques.
- Ensure the sample size is sufficient for meaningful analysis.

...

Explore:

- **Objective**: In this phase, the selected data sample is explored to understand its characteristics, identify patterns, and gain insights into potential relationships between variables.

Activities:

- Perform descriptive statistics to summarize data distribution, central tendencies, and variability.
- Visualize data through charts, graphs, and plots to identify trends and outliers.
- Conduct preliminary data analysis to identify potential data quality issues or missing values.

...

Modify:

- **Objective**: The modify phase focuses on data preparation and preprocessing to ensure data quality and suitability for modeling.

Activities:

- Clean the data by handling missing values, outliers, and inconsistencies.
- Transform variables as needed, such as normalization, scaling, or encoding categorical variables.
- Select relevant features or variables that are most predictive for the modeling phase.

...

Model:

- **Objective**: The modeling phase involves building and validating predictive or descriptive models using statistical or machine learning techniques.

Activities:

- Select appropriate modeling techniques based on project objectives and data characteristics (e.g., regression, classification, clustering).
- Train models using the prepared dataset and evaluate model performance using appropriate metrics.
- Iteratively refine models by tuning parameters and assessing model robustness and generalization.

...

Assess:

- **Objective**: The assess phase evaluates the performance and effectiveness of the models developed in the previous phase.

Activities:

- Assess model performance using evaluation metrics (e.g., accuracy, precision, recall, F1-score for classification; RMSE, MAE for regression).
- Validate models by testing against new or unseen data to ensure they generalize well.
- Interpret and communicate results to stakeholders, including recommendations based on model findings.

Differences

Aspect	KDD	SEMMA	CRISP-DM
Origin	Academic and research communities	SAS Institute	Cross-industry initiative
Focus	Entire knowledge discovery process	Technical data mining processes	Structured data mining methodology
Phases	Selection, Preprocessing, Transformation, Data Mining, Interpretation/Evaluation	Sample, Explore, Modify, Model, Assess	Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment
Business Emphasis	Moderate	Low	High
Iterative Approach	Yes	Yes	Yes
Tool Dependency	None	Tied to SAS tools	Tool-agnostic
Flexibility	High	Moderate	High

The End