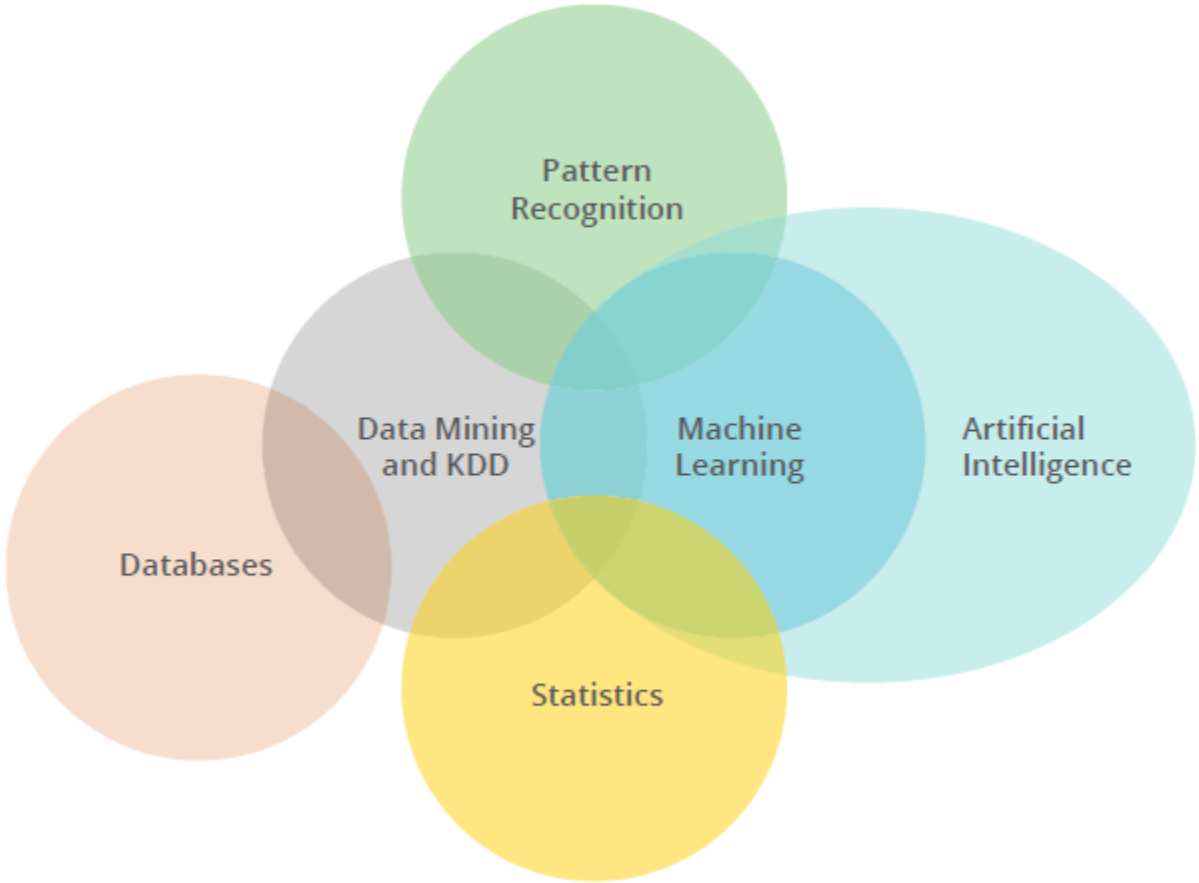
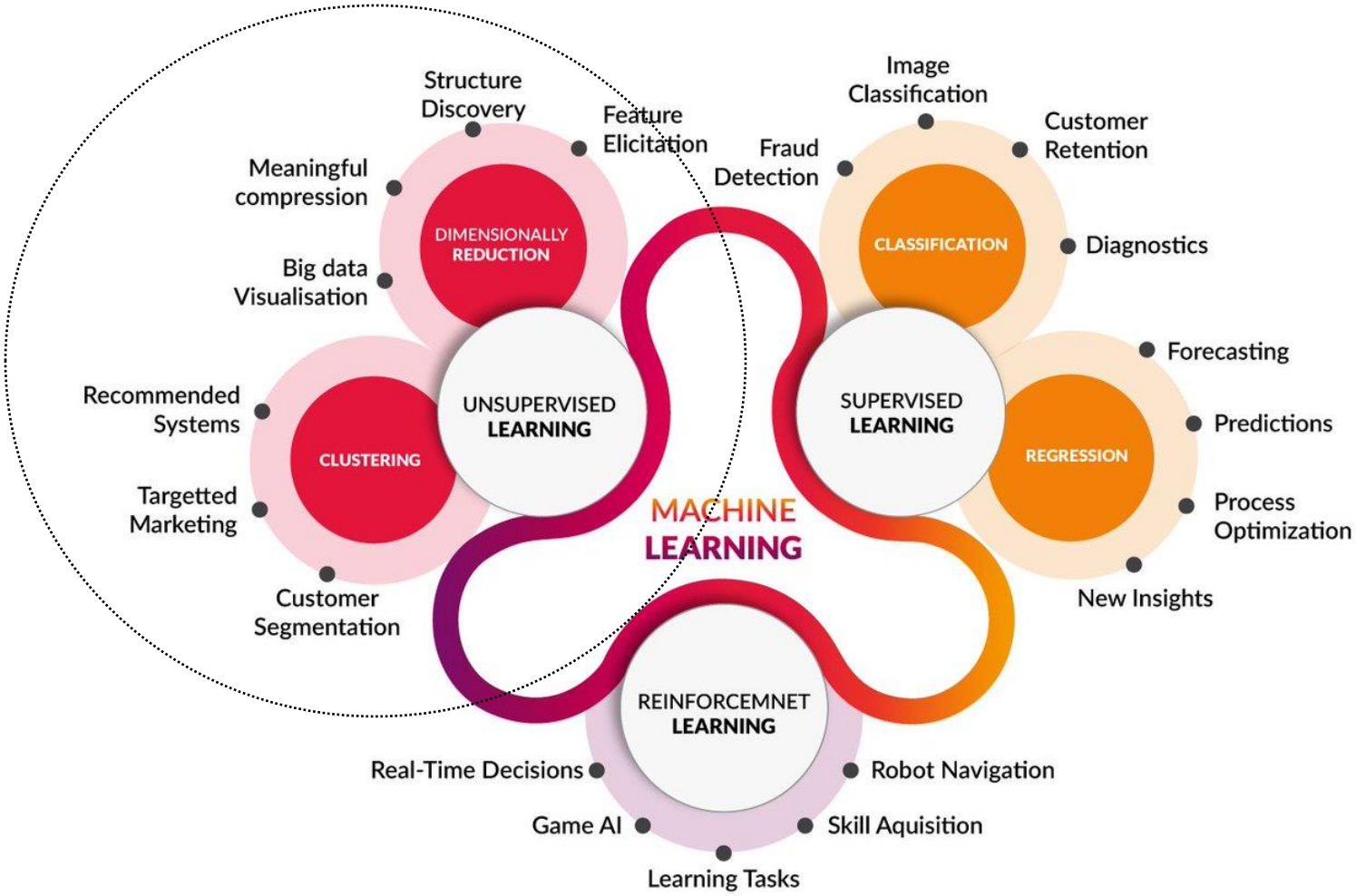
# Introduction to Data Mining

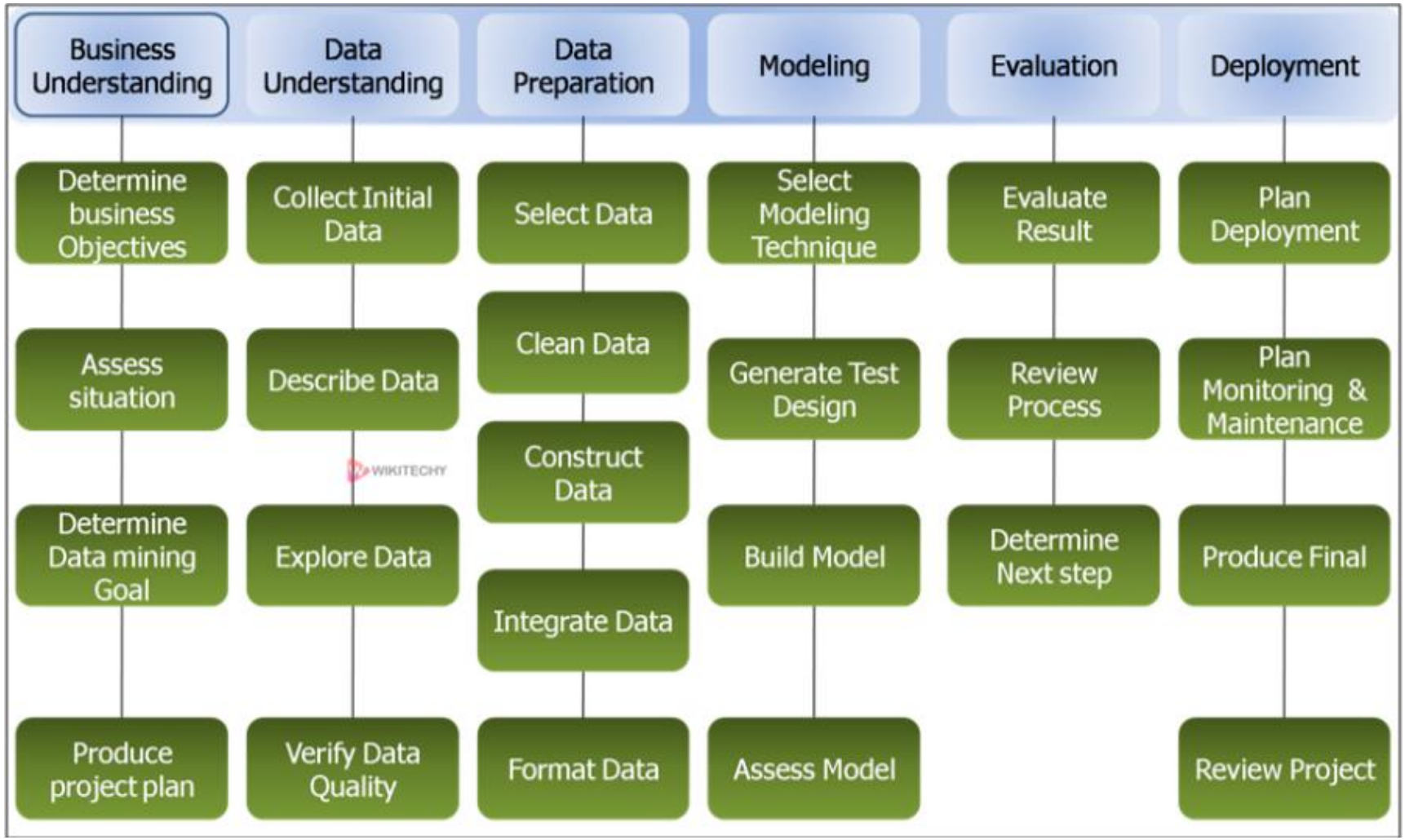
# Data Mining vs ML vs AI



# Machine Learning

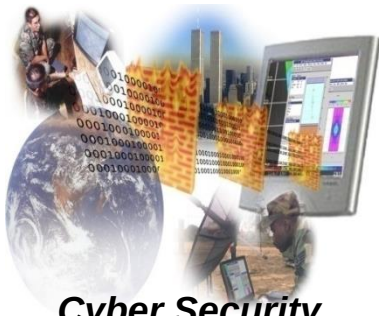


# Implementation Process of Data Mining



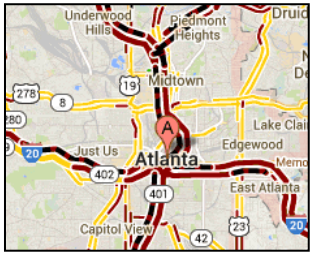
# Large-scale Data is Everywhere!

- Enormous data growth in commercial and scientific datasets 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.

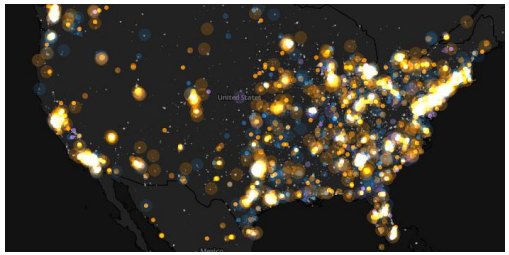


***Cyber Security***


***E-Commerce***

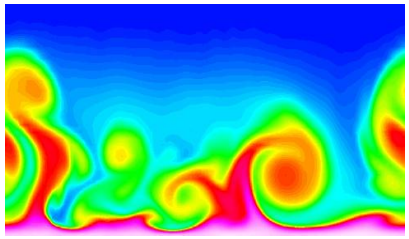



***Traffic Patterns***



***Social Networking: Twitter***

***Sensor Networks***



***Computational Simulations***

# Why Data Mining? Commercial Viewpoint

- Data Explosion: Lots of data is collected and warehoused
  - **Web Data**: Google stores Peta Bytes of web data.
  - **Social Media**: Facebook has billions of users.
  - **E-commerce**: Millions of daily transactions.
  - **Technological Advancements**: Cheaper and more powerful computers.

- Competitive Pressure is Strong
  - **Intense Pressure**: Stiff competition in data era.
  - **Strategic Advantage**: Offer better, customized services (e.g., Customer Relationship Management).

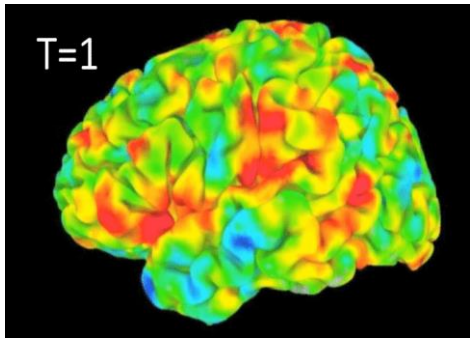



# Why Data Mining? Scientific Viewpoint

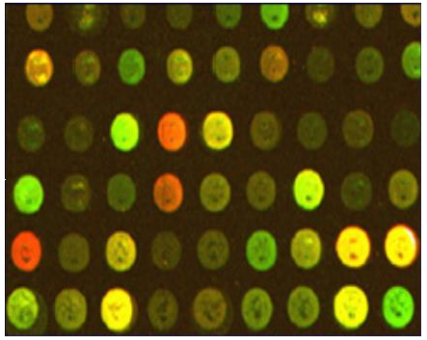
- Rapid Data Accumulation
  - Data collected and stored at incredible speeds.
  - Example: satellites spatial data collection.
  - NASA archives petabytes of earth science data annually.
  - Telescopes scanning the skies.
  - Sky survey data.

**Sky Survey Data**

- Biological Insights
  - High-throughput biological data.




**fMRI Data from Brain**

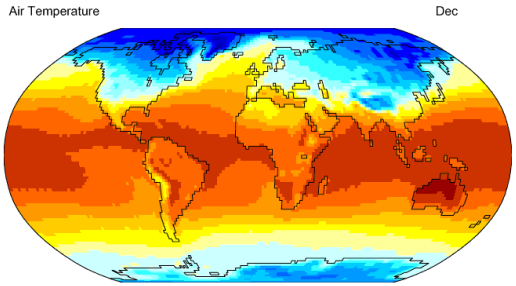


**Gene Expression Data**

- Simulating the Unseen
  - Scientific simulations generate terabytes

*Data Mining Empowers Scientists*

Automating analysis of massive datasets and aiding hypothesis formation.

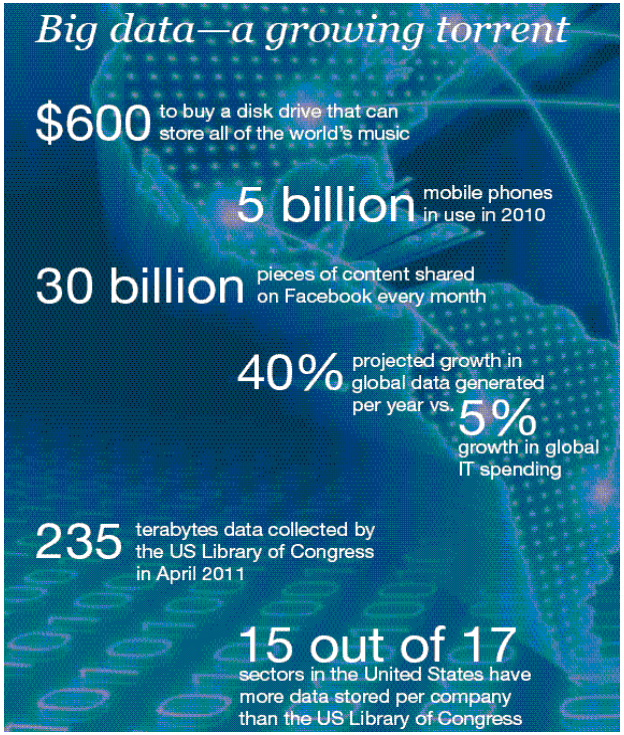


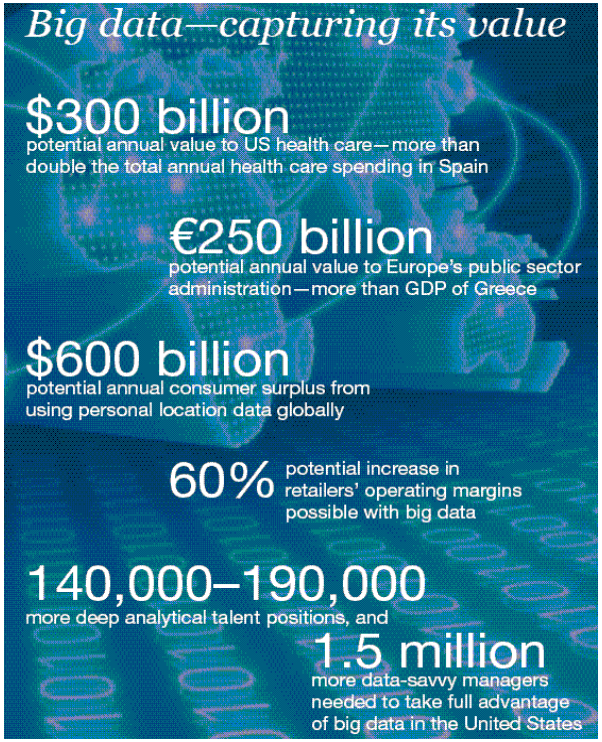

**Surface Temperature of Earth**

# Great opportunities to improve productivity

McKinsey Global Institute

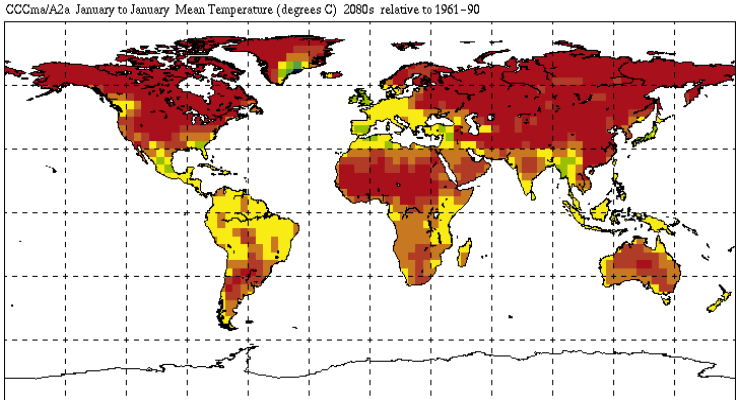
## Big data: The next frontier for innovation, competition, and productivity



# Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



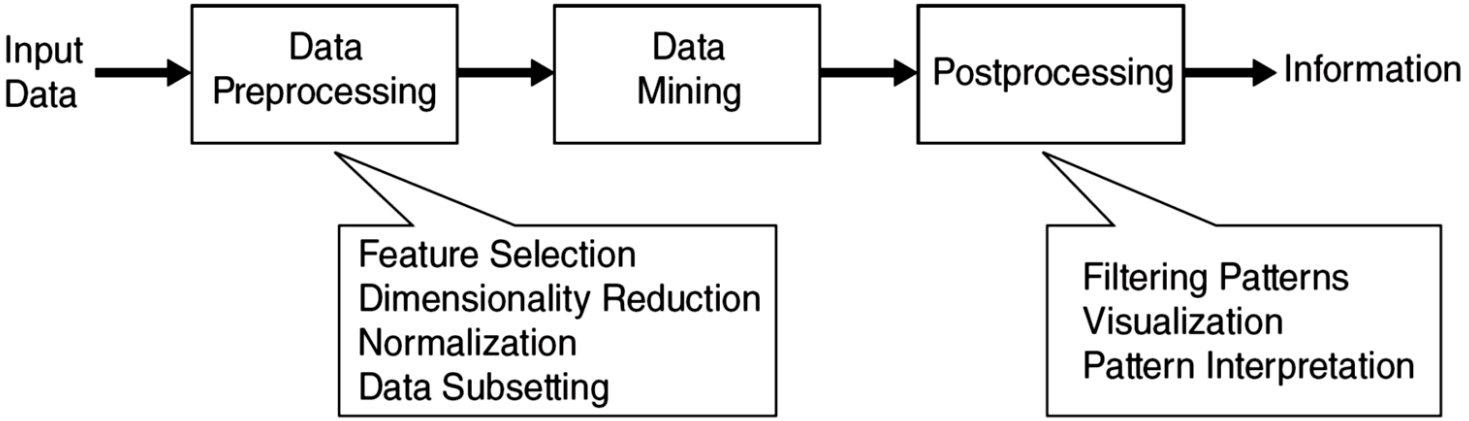

Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

# What is Data Mining?

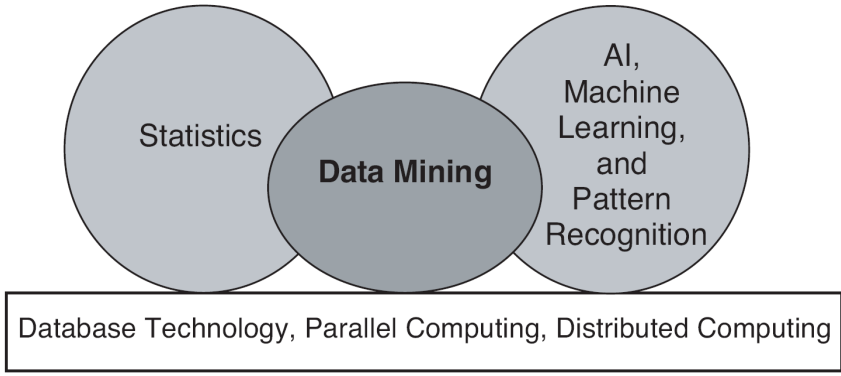
- Many Definitions
  - **Non-trivial extraction** of implicit, **previously unknown** and **potentially useful** information from data.
  - **Exploration & analysis**, by automatic or semi-automatic means, of **large quantities of data** in order to **discover meaningful patterns.**



# Origins of Data Mining

- Ideas from many fields
  - machine learning/AI, pattern recognition, statistics, and database systems

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



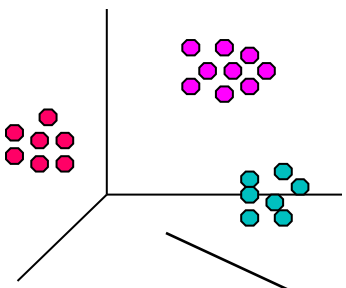

***A key component of the emerging field of data science and data-driven discovery***

# Data Mining Tasks

- **Prediction Methods**
  - Use some variables to predict unknown or future values of other variables.
  - **Example:**
    - Sales Forecasting in E-commerce

- **Description Methods**
  - Find human-interpretable patterns that describe the data.
  - **Example:**
    - Analyzing historical criminals data for profiling.

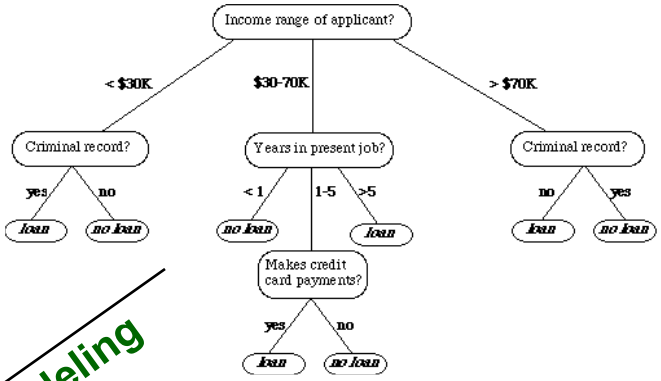
# Data Mining Tasks ...



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 |

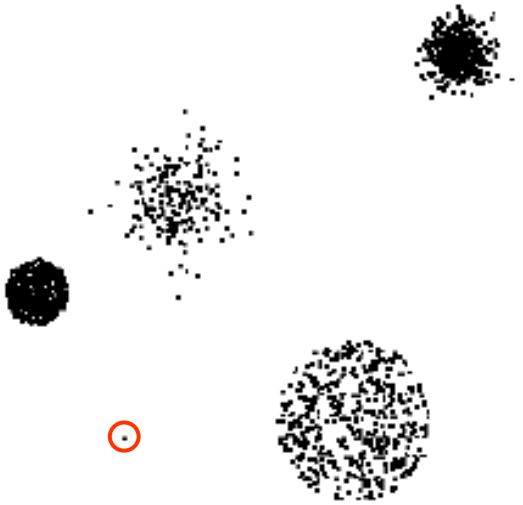



Predictive Modeling

Association Rules

Anomaly Detection



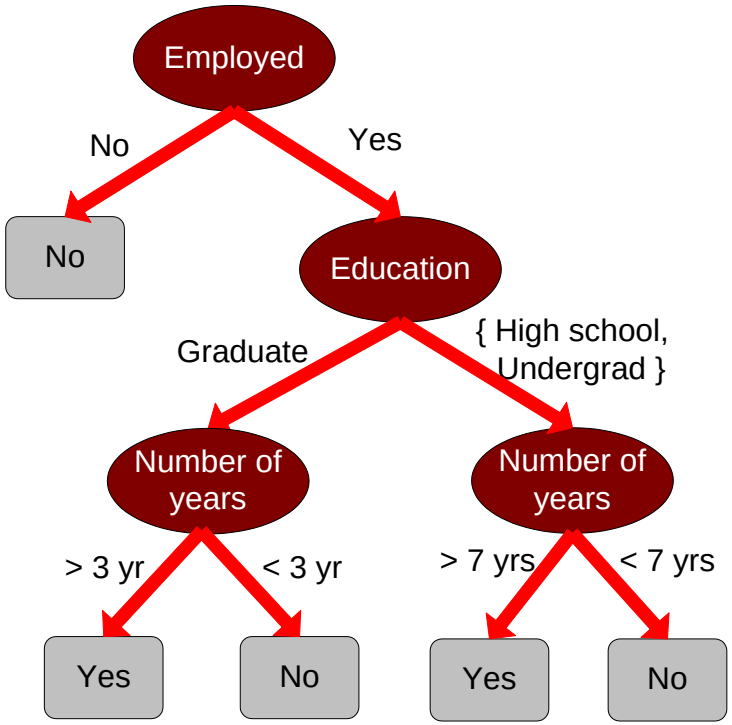


# Predictive Modeling: Classification

**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**

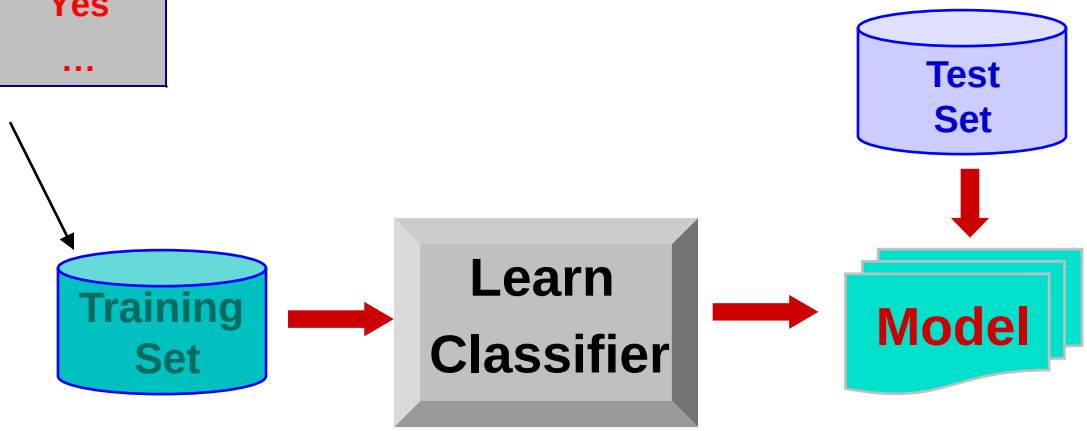



***Find a model for class attribute as a function of the values of attributes***

# Classification Example

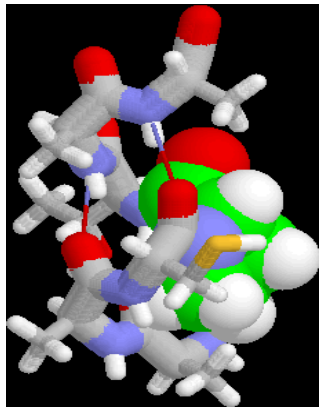
| Tid | Employed (categorical) | Level of Education (categorical) | # years at present address (quantitative) | Credit Worthy (class) |
|---|---|---|---|---|
| 1 | Yes | Graduate | 5 | Yes |
| 2 | Yes | High School | 2 | No |
| 3 | No | Undergrad | 1 | No |
| 4 | Yes | High School | 10 | Yes |
| … | … | … | … | … |

| 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

- Credit card transaction classification
  - Legitimate vs. fraudulent.
- Satellite data
  - Land cover classification (water bodies, urban areas, . etc.).
- News story categorization
  - Finance, weather, entertainment, sports, etc.
- Cyberspace security
  - Intruder identification.
- Medical diagnosis
  - Tumor cell classification (benign vs. malignant).
- Protein analysis
  - Classification of secondary structures (alpha-helix, beta-sheet, or random coil).

# Classification Application 1 : Fraud Detection

- **Goal**

  ***Predict and Prevent Fraud***

- **Approach**
  - **Data Collection:** Use transaction and account-holder data as attributes.
  - **Attribute Identification:** Timing, purchase details, payment history…etc.
  - **Transactions labeling:** Categorize as fraud or fair.
  - **Model Creation:** Develop a robust classification model.
  - **Real-time Monitoring:** Detect fraud in live credit card transactions.

# **Classification Application 2:** Churn prediction

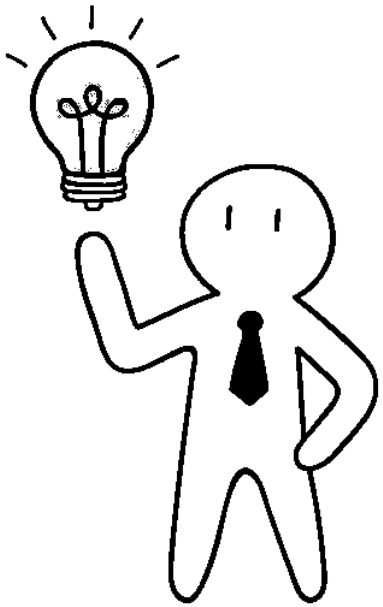
- **Goal**

  ***Predict whether a telephone customer is likely to be lost to a competitor***.

- **Approach**
  - **Data Collection:** Gather detailed transaction records for past and present customers.
  - **Attribute Identification:** Explore attributes like call frequency, call locations, peak call times, financial status, marital status, etc.
  - **Customer Labeling:** Categorize customers as loyal or disloyal.
  - **Model Development:** Create a robust model for predicting customer loyalty.
  - **Real-time Monitoring:** Detect potential churn among customers in real-time.

# Classification Application 3

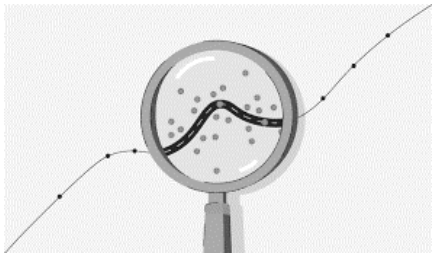
Imagine a use case and recommend a classification-based approach to resolve it.



# Regression

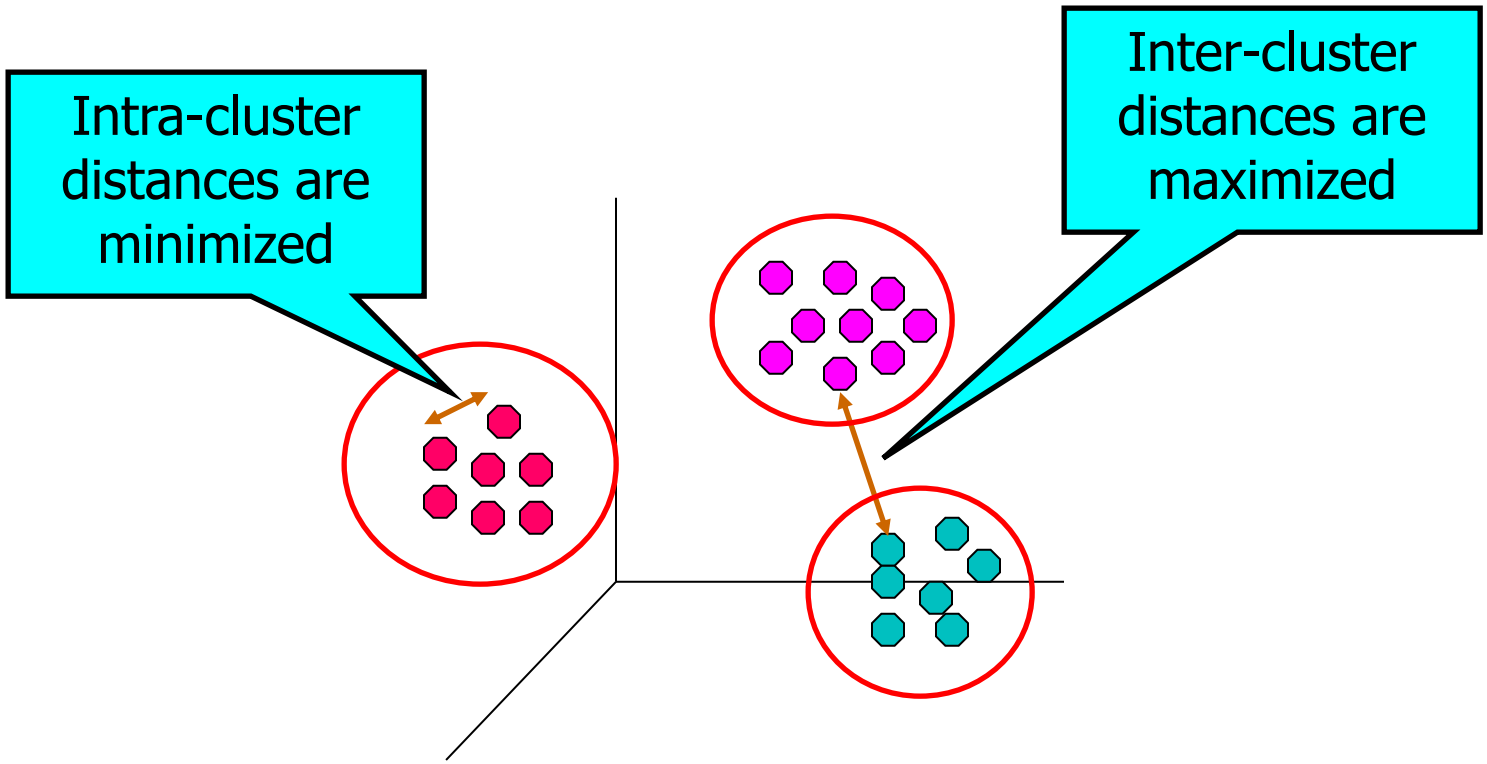
- **Predicting a continuous variable by considering the relationships with other variables, using linear or nonlinear models.**

- Widely explored in statistics and neural network domains.
- **Examples**
  - **Forecasting Sales:** Predicting new product sales by analyzing advertising expenditure.
  - **Wind Velocity Prediction**: Estimating wind velocities using variables like temperature, humidity, and air pressure.
  - **Stock Market Forecast**: Predicting stock market indices through time series analysis.



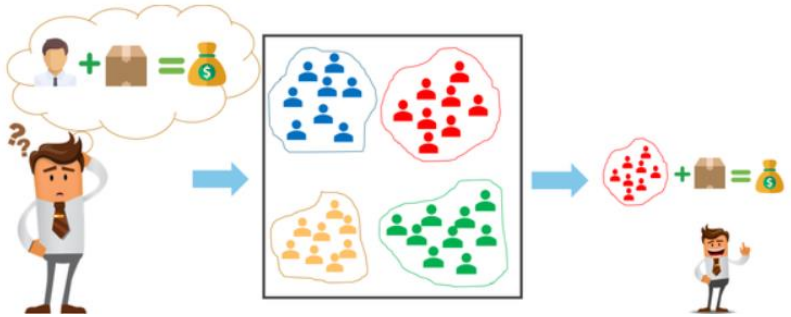
# Clustering

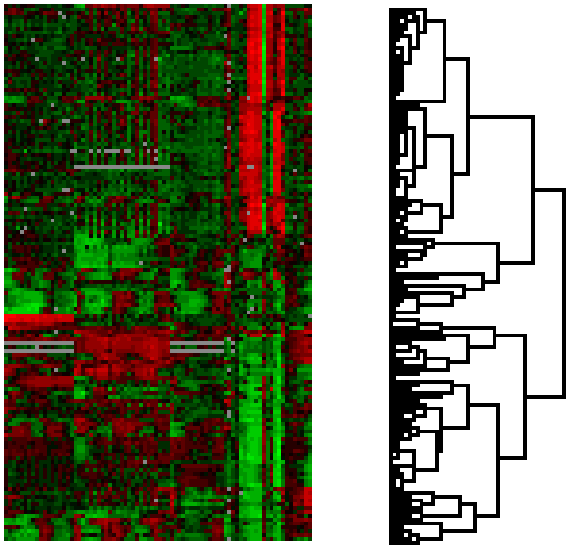
**Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups**

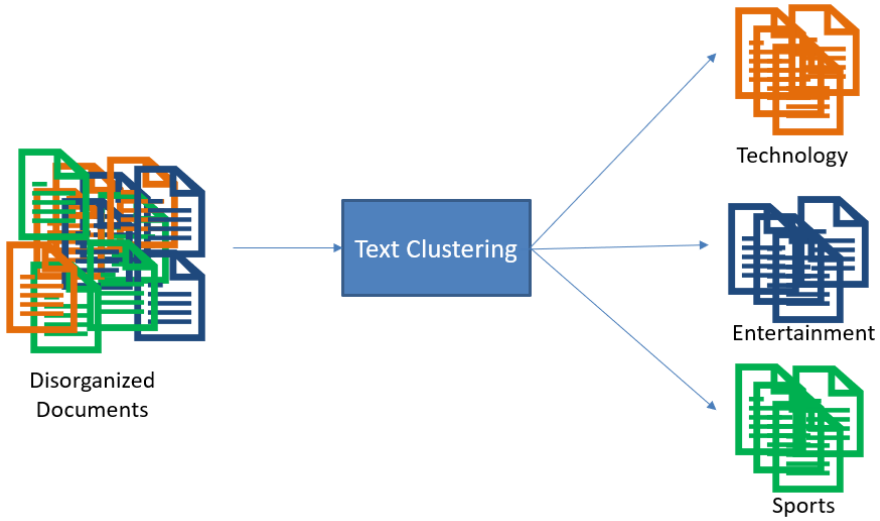


# Applications of Cluster Analysis

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







# **Clustering Application 1:** Market Segmentation

- **Goal**

  ***Segment the market into distinct customer subsets, allowing precise targeting with tailored marketing strategies.***

- **Approach:**
  - **Data Collection:** Gather diverse customer attributes, including geographical and lifestyle information.
  - **Customer Clustering:** Identify clusters of customers who share similar attributes and characteristics.
  - **Clustering Evaluation:** Assess the quality of clustering by analyzing buying patterns within and across clusters.

# **Clustering Application 2:** Document Clustering

- **Goal**

  ***Identify groups of documents that exhibit similarity based on their important words.***

- **Approach**
  - **Term Identification:** Recognize frequently occurring terms within each document.
  - **Similarity Measure:** Create a similarity metric by considering term frequencies across documents.
  - **Clustering Technique:** Utilize the similarity measure to cluster documents into cohesive groups.

# 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 | Juice, Coke, Milk |
| 2 | Juice, Bread |
| 3 | Juice, Coke, Diaper, Milk |
| 4 | Juice, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

Rules Discovered:
**{Milk} --> {Coke}**
**{Diaper, Milk} --> {Juice}**

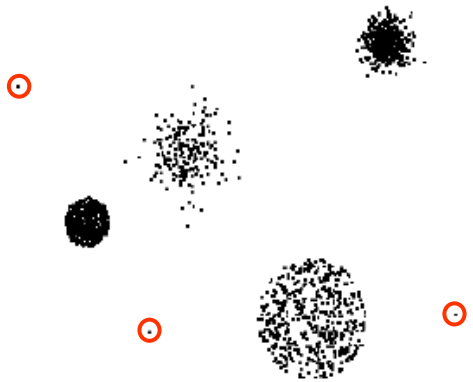
# Association Analysis: Applications

- **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 th



# Motivating Challenges

- **Scalability:** Handling Expansive Data
  - Efficiently managing and processing massive datasets.
- **High Dimensionality:** Dealing with Multidimensional Data
  - Addressing complex data structures with numerous attributes.
- **Heterogeneous and Complex Data:** Managing Data Variety
  - Handling diverse data types and intricate data structures.
- **Data Ownership and Distribution:** Navigating Data Access
  - Tackling geographically distributed data owned by multiple entities.
  - Challenges include minimizing communication, consolidating results, and ensuring data security and privacy.
- **Non-traditional Analysis:** Adapting to Advanced Techniques
  - Statical methodology is unable to deal with current data nature.