Introduction to Data Mining

by

Eng. Asma’a Hassan

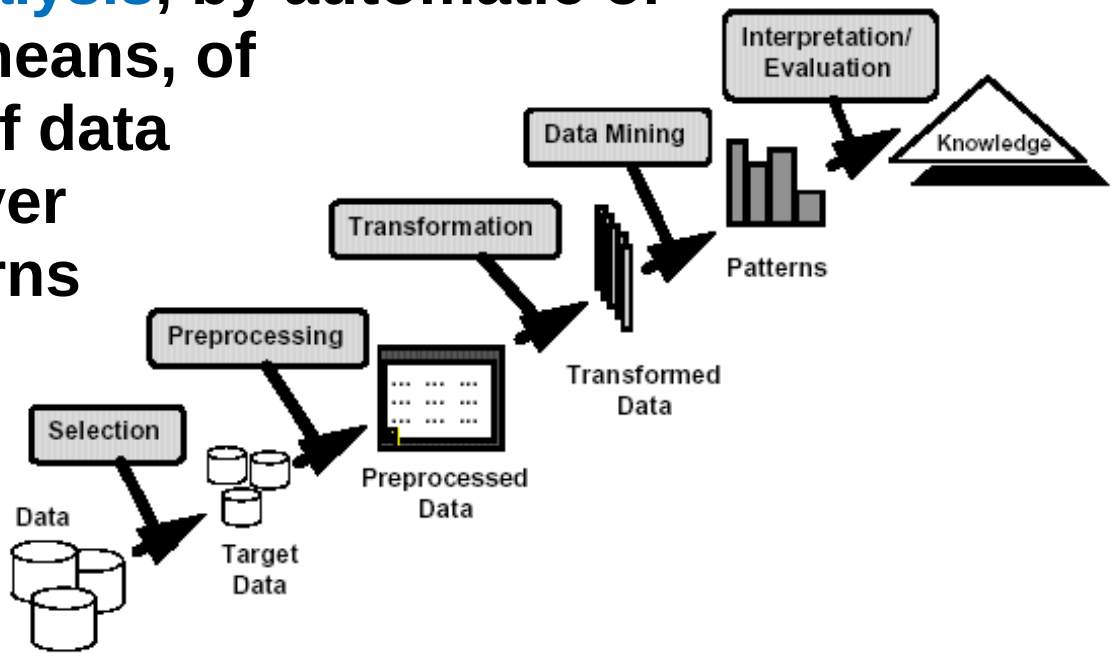
Learning Objectives

After completing this lesson, students should be able to:

- Define data mining
- Justify usage of data mining
- Give examples for data mining applications
- Recognize the origins of data mining
- Classify data mining tasks
- Summarize challenges of data mining

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 (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'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Data Explosion

We are drowning in data, but starving for knowledge

"The amount of data stored in various media has doubled in three years, from 1999 to 2002. The amount of data put into storage in 2002, five exabytes (one quintillion bytes), was equal to the contents of a half a million new libraries, each containing a digitised version of the print collection of the entire US Library of Congress" (Lyman and Varian, UC Berkeley, 2003)

Scale of Data

Organization	Scale of Data
Walmart	~ 20 million transactions/day
Google	> 4.2 billionWeb pages
Yahoo	~10 GBWeb data/hr
NASA	satellites ~ 1.2 TB/day
NCBI GenBank	~ 22 million genetic sequences
France Telecom	29.2 TB
UK Land Registry	18.3 TB
AT&T Corp	26.2 T

"The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don't have a clue as to what any of that data actually means" (Stephen Cass, "A Fountain of Knowledge," IEEE Spectrum, January 2004)

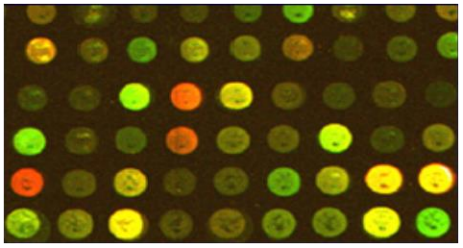
Why Mine Data? Commercial Viewpoint

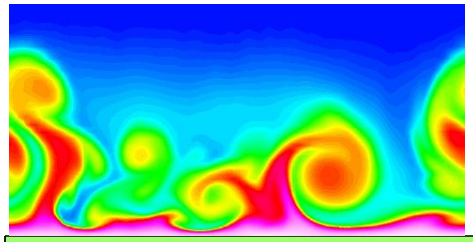
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

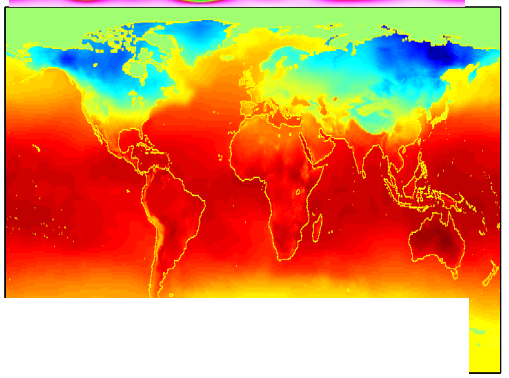
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data

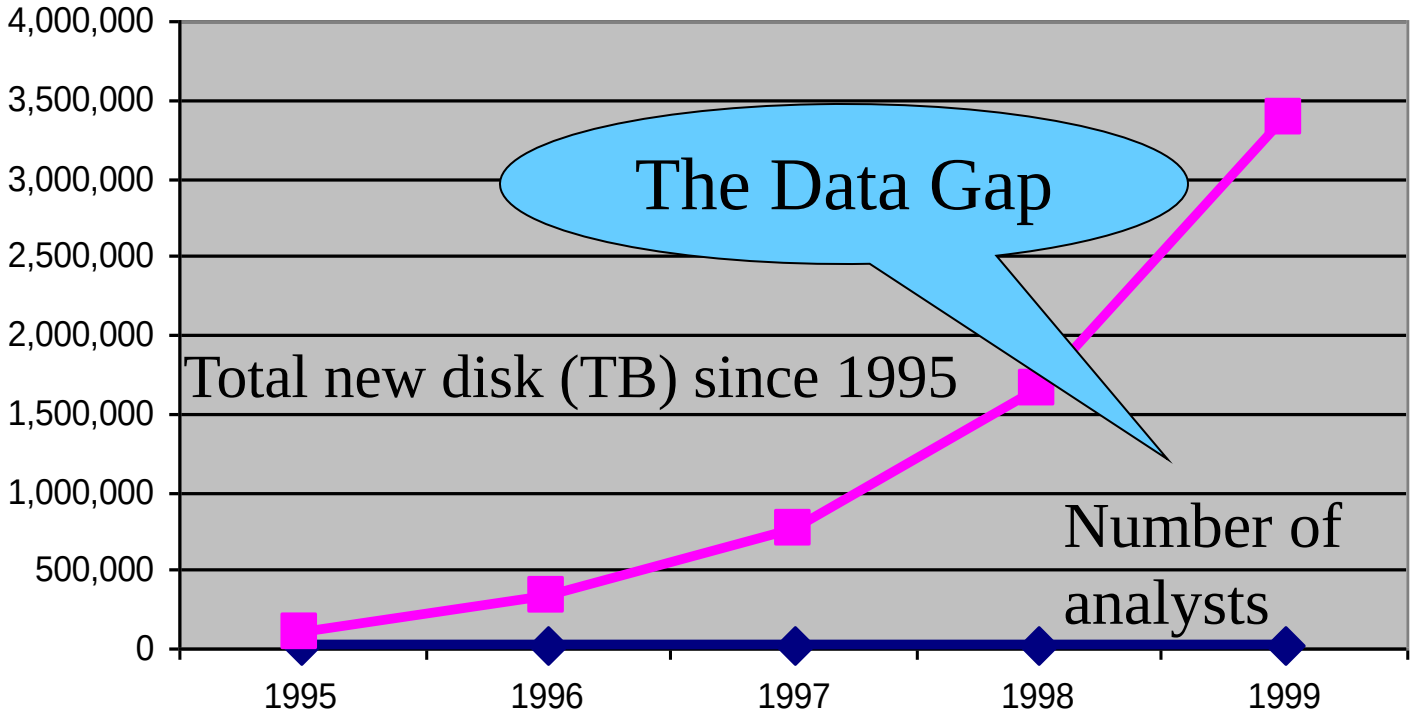






Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



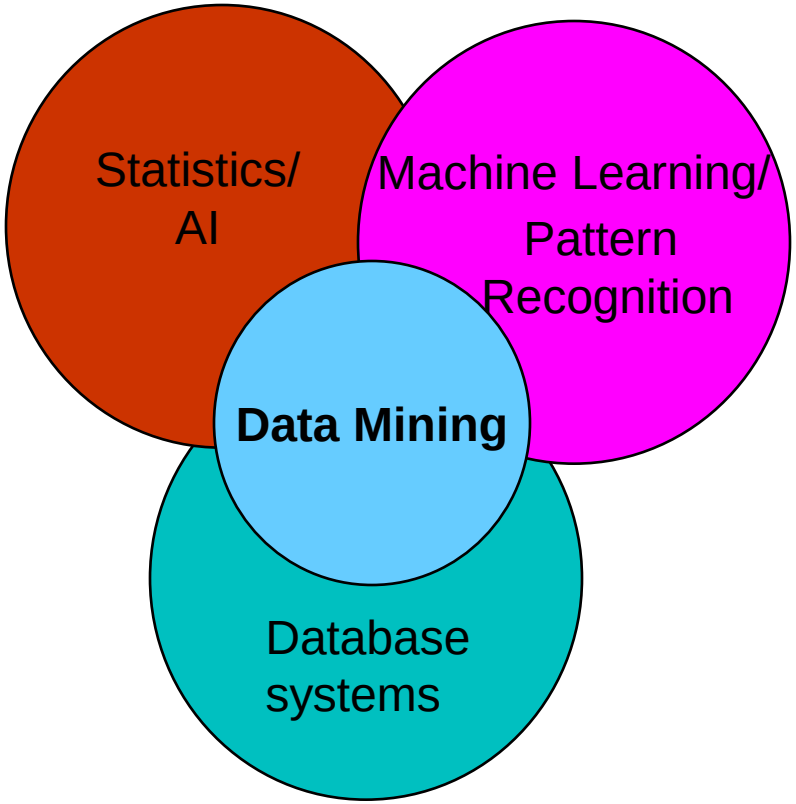
From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Data Mining Applications

Application	Input	Output
Business Intelligence	Customer purchase history, credit card information	What products are frequently bought together by customers
Collaborative Filtering	User-provided ratings for movies ,or other products	Recommended movies or other products
Network Intrusion Detection	TCPdump trace or Cisco NetFlow logs	Anomaly score assigned to each network connection
Web search	Query provided by user	Documents ranked based on their relevance to user input
Medical Diagnosis	Patient history,physiological, and demographic data	Diagnosis of patient as sick or healthy
Climate Research	Measurements from sensors aboard NASA Earth observing satellites	Relationships among Earth Science events, trends in time series, etc
Process Mining	Event-based data from workflow logs	Discrepancies between prescribed models and actual process executions

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



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...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential pattern Discovery[Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification: Definition

- **Given**:
  - a collection of records (*training set*)
  - Each record contains a set of *attributes*
  - one of the attributes is the *class*.
- **Task**:

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

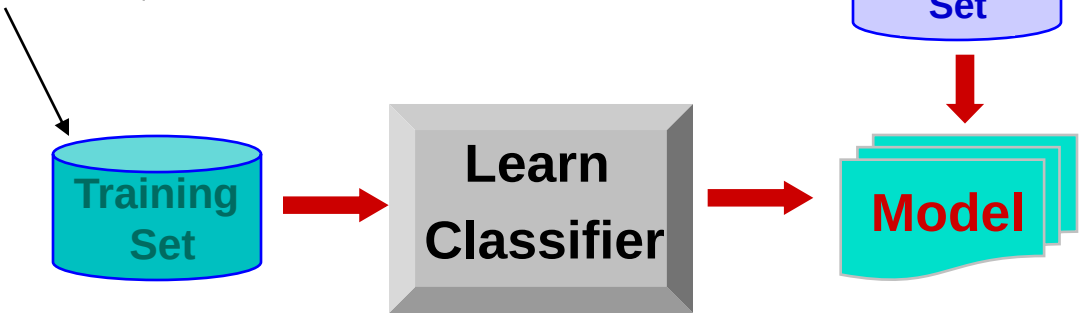
  - use the model to predict the class for previously unseen records
- **Goal**:

  Model should accurately predict the class for previously unseen records.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
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

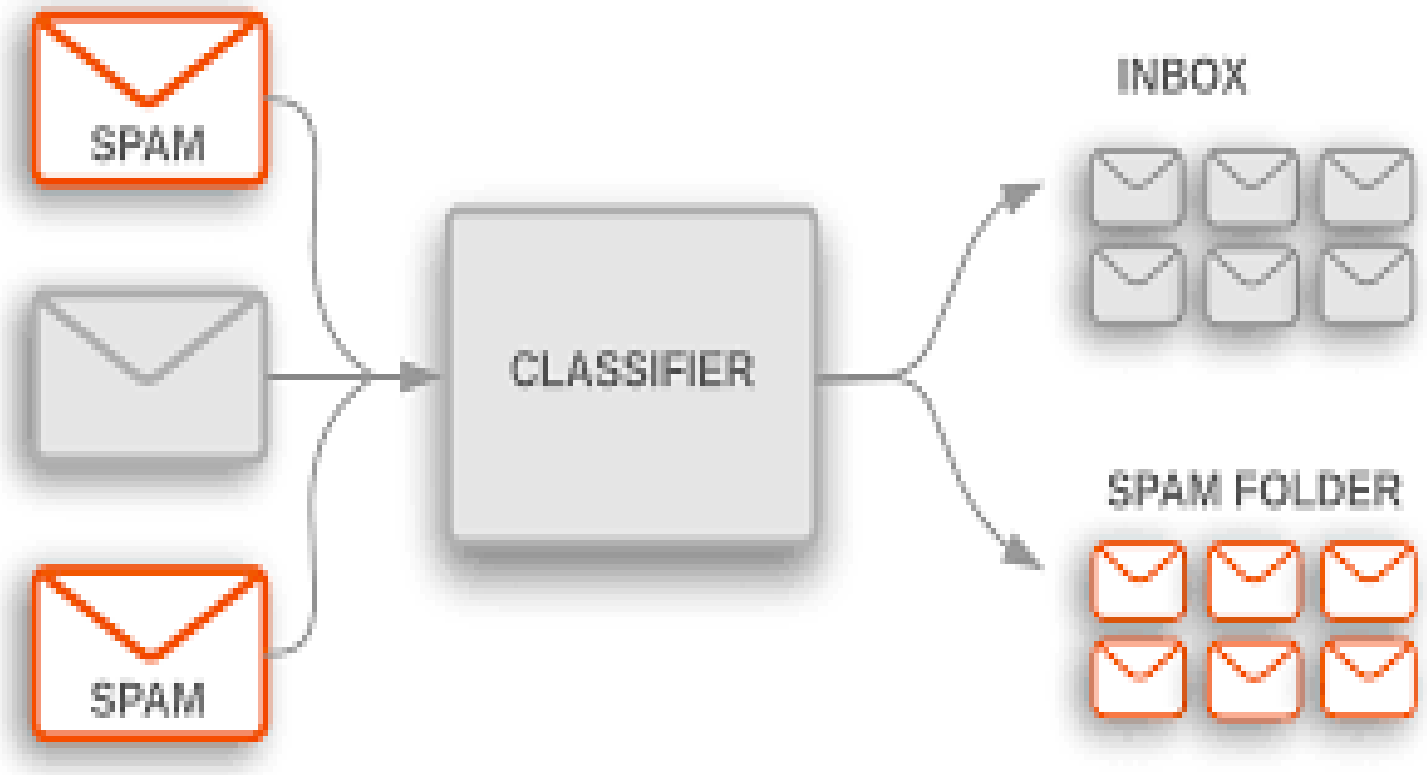
Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Applications

- Direct marketing
  - Predict consumers who will most likely buy a new product base on their demographic, lifestyle, and previous buying behavior
- Spam detection
  - Categorize email messages as spam or non-spam based on message header and content
- Functional classification of proteins
  - Assign sequences of unknown proteins to their respective functional classes
- Galaxy classification
  - Classify galaxies based on their image features
- Automated target recognition
  - Identify target objects (enemy tanks, trucks, etc) based on signals gathered from sensor arrays

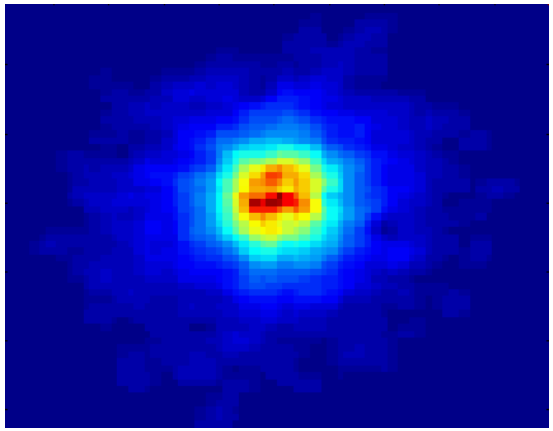
Classification Example



Classifying Galaxies

Courtesy: http://aps.umn.edu

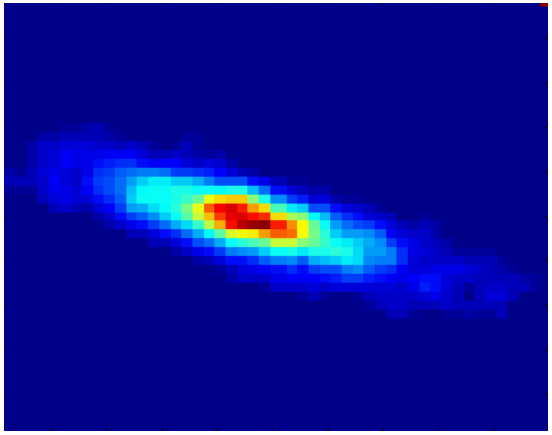
Early



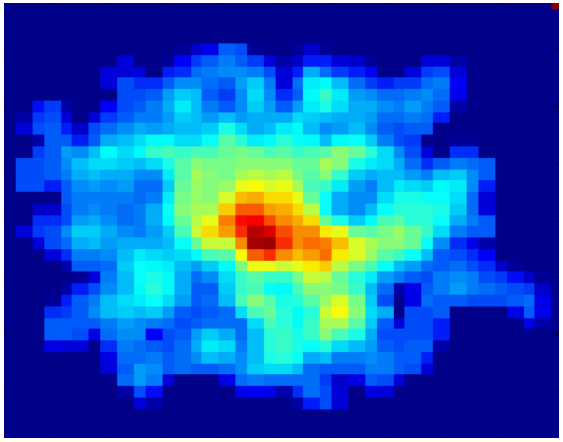
Class:
- **Stages of Formation**

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**

Clustering Definition

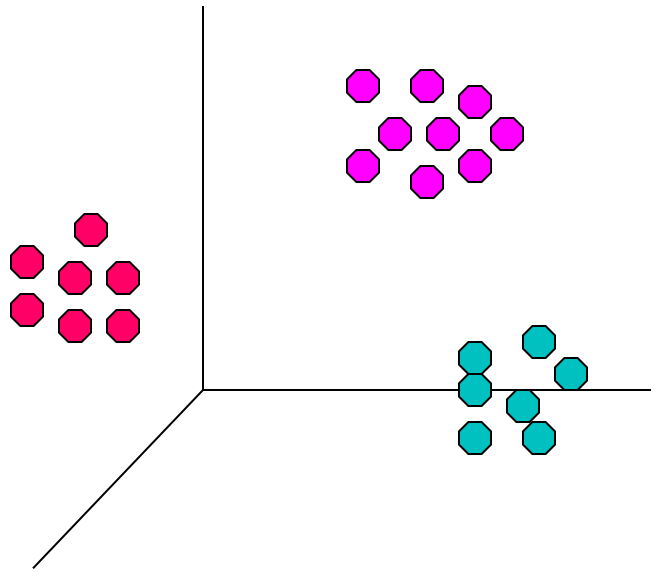
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance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

Illustrating Clustering

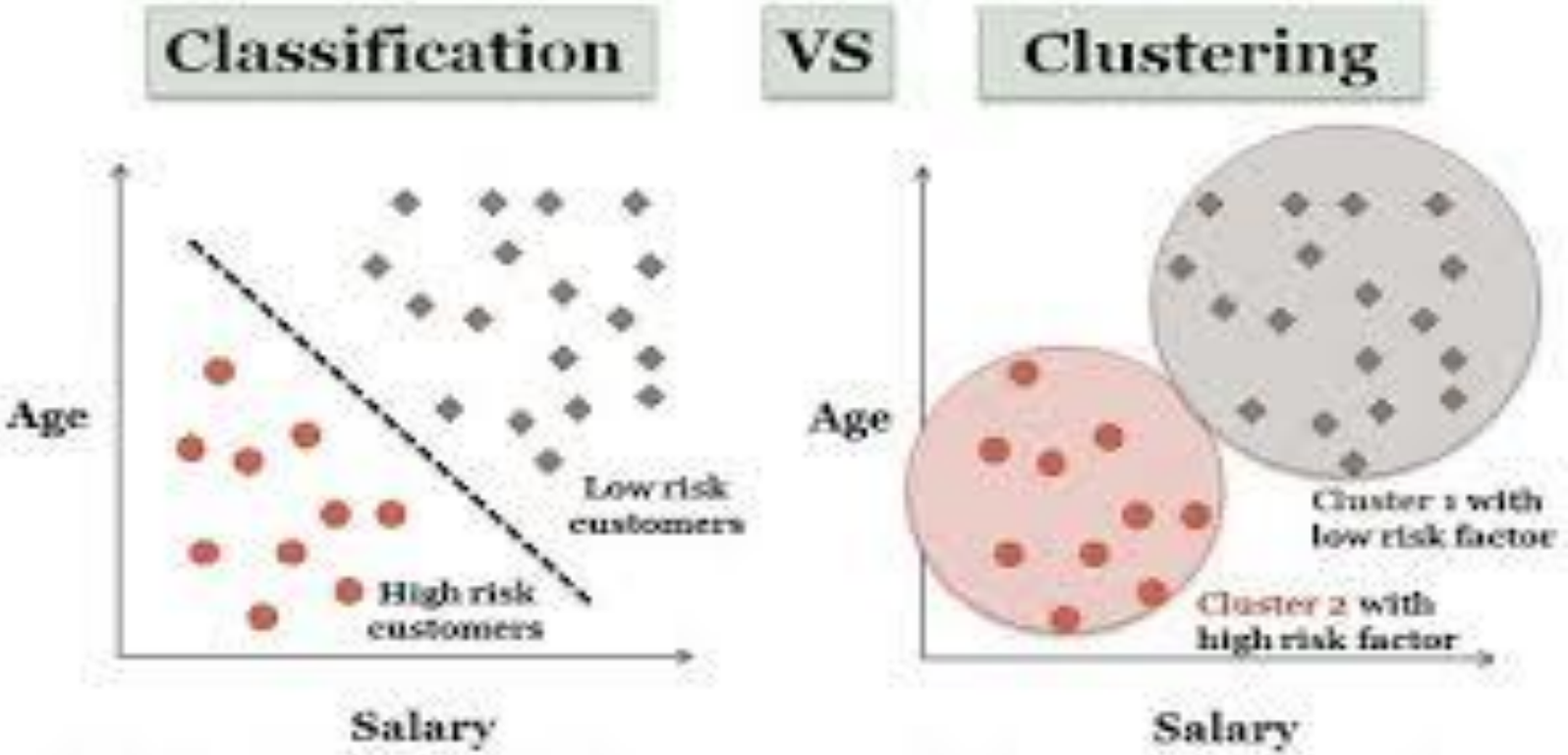
- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Comparing Clustering and Classification



Risk classification for the loan payees on the basis of customer salary

Clustering: Applications

- Market Segmentation
  - Subdivide customers based on their geographical and lifestyle related information
- Document clustering
  - Find groups of documents that are similar to each other based on the important terms appearing in them
- Time series clustering
  - Find groups of similar time series (e.g., stock prices, ECG, seismic waves) based on their shapes
- Sequence clustering
  - Find groups of sequences (e.g., Web or protein sequences) with similar features

Association Rule Discovery: Definition

Given:

- A collection of transactions
- Each transaction contains a set of items

Task:

- Discover dependency rules that will predict the presence of an item in a record based on the presence of other items

Goal:

- Rules must have high support, i.e., applicable to sufficiently large number of records
- Rules must have high confidence, i.e., make accurate prediction

Illustratin Association Rule Mining

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}

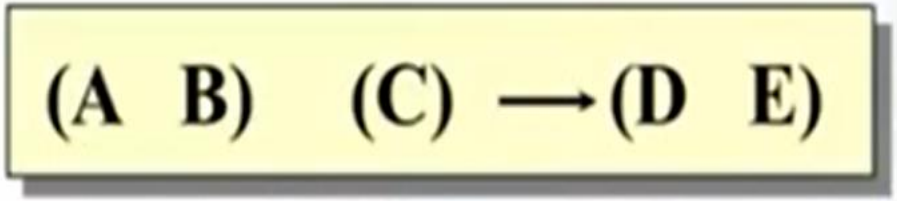
{Diaper, Milk} --> {Beer}

Association Rule Mining: Applications

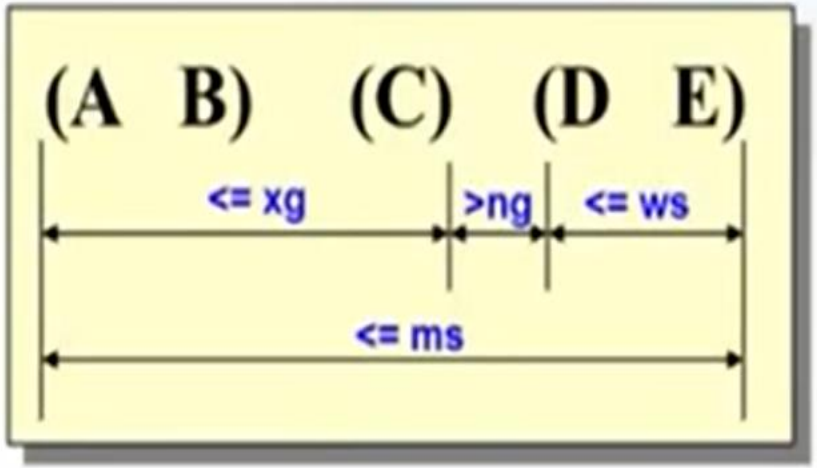
- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- World-Wide Web
  - Rules are used to develop Web caching and prefetching techniques

Sequential Pattern Discovery: Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.



- Rules are formed by first disovering patterns. Event occurrences in the patterns are governed by timing constraints.



Sequential Pattern Discovery: Definition

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)

    (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:

    (Intro_To_Visual_C) (C++_Primer) -->

    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:

    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

Regression: Definition

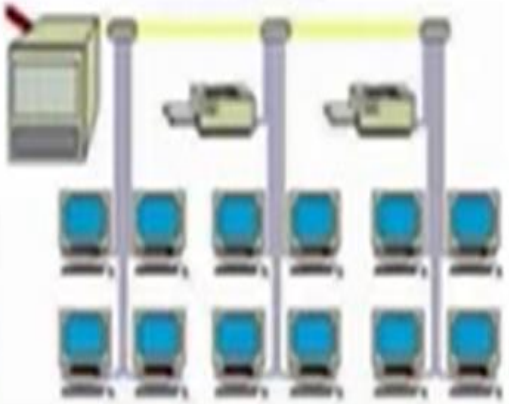
- Given:
  - A collection of records (training set) Each record contains a set of attributes
  - One of the continuous-valued attributes is designated as the target variable
- Task:
  - Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency
- Greatly studied in statistics, neural network fields

Regression: Applications

- Marketing
  - Predicting sales amounts of new product based on advertising expenditure
- Earth Science
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc
- Finance
  - Time series prediction of stock market indices
- Agriculture
  - Predicting crop yield based on soil fertility and weather information
- Socio-economy
  - Predicting electricity consumption in single family homes based on outdoor temperatures

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection



Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data