# CSEN1083: Data Mining

Seif Eldawlatly

Spring 2019

# CSEN1083: Data Mining

- Instructor
  - Associate Professor

    Faculty of Media Engineering and Technology, GUC

    Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University

    E-mail: seif.eldawlatly@guc.edu.eg

- Office Hours
  - Tuesdays– 12:00pm to 1:00pm (Office: TBD)

- Textbook
  - “Introduction to Data Mining" by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar, First edition (2006) or Second edition (2018), Pearson Education

# CSEN1083: Data Mining

- Course Evaluation
  - 3 Assignments (Programming): 30%
  - Mid-term exam: 20%
  - 3 Quizzes: 10%  (Best 2 out of 3)
  - Final exam: 40%

# Introduction

- Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data

- Other Definitions:
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data

  - Exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

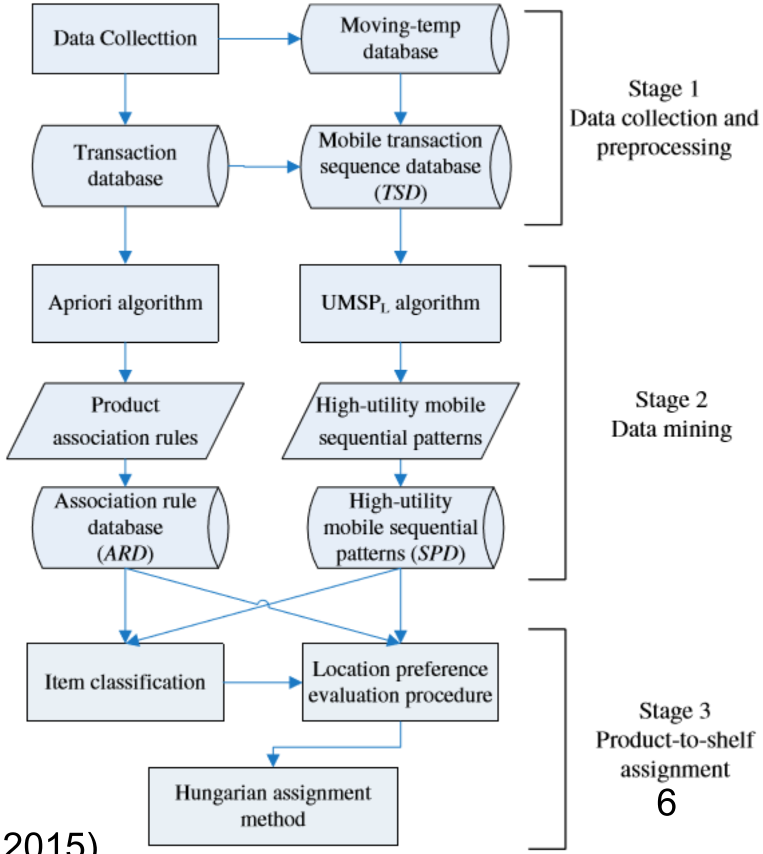
# Introduction

- Examples: Business
- Point-of-sale data collection (bar code scanners, radio frequency identification (RFID), and smart card technology) have allowed retailers to collect up-to-the-minute data

© © James Hardy/PhotoAlto/Corbis

# Introduction

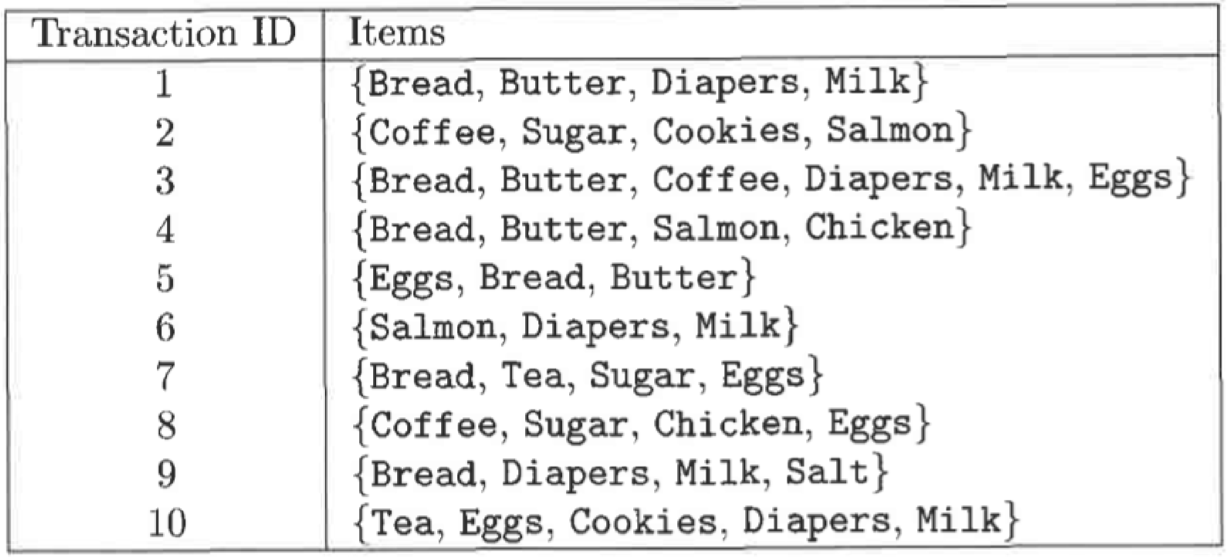
- Examples: Business
- Applications: Shelf Space Optimization
- Marketing the right merchandise, at the right place, at the right time, in the right quantities is key to retail revenues and profitability


Data Collecttion
Moving-temp database
Transaction database
Mobile transaction sequence database (TSD)
Stage 1
Data collection and preprocessing
Apriori algorithm
UMSP$_L$ algorithm
Product association rules
High-utility mobile sequential patterns
Association rule database (ARD)
High-utility mobile sequential patterns (SPD)
Stage 2
Data mining
Item classification
Location preference evaluation procedure
Hungarian assignment method
Stage 3
Product-to-shelf assignment

(Tsai and Huang, 2015)

# Introduction

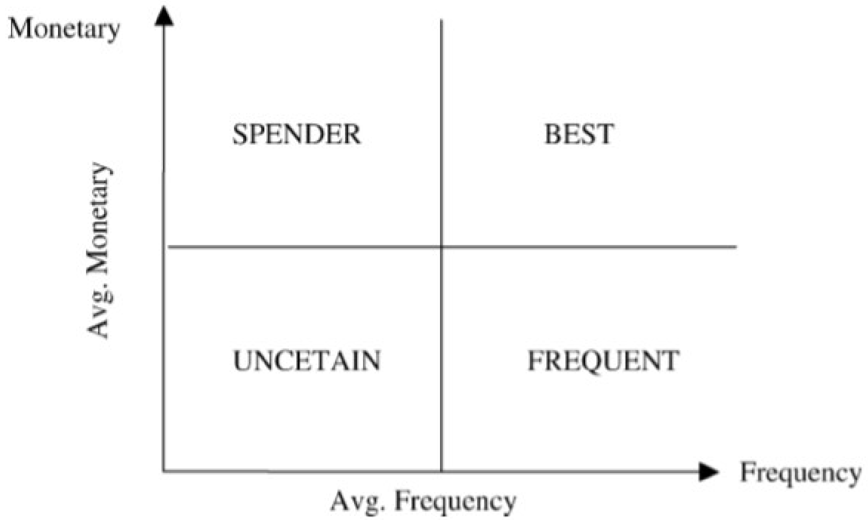
- Examples: Business
- Applications: Market Basket Analysis

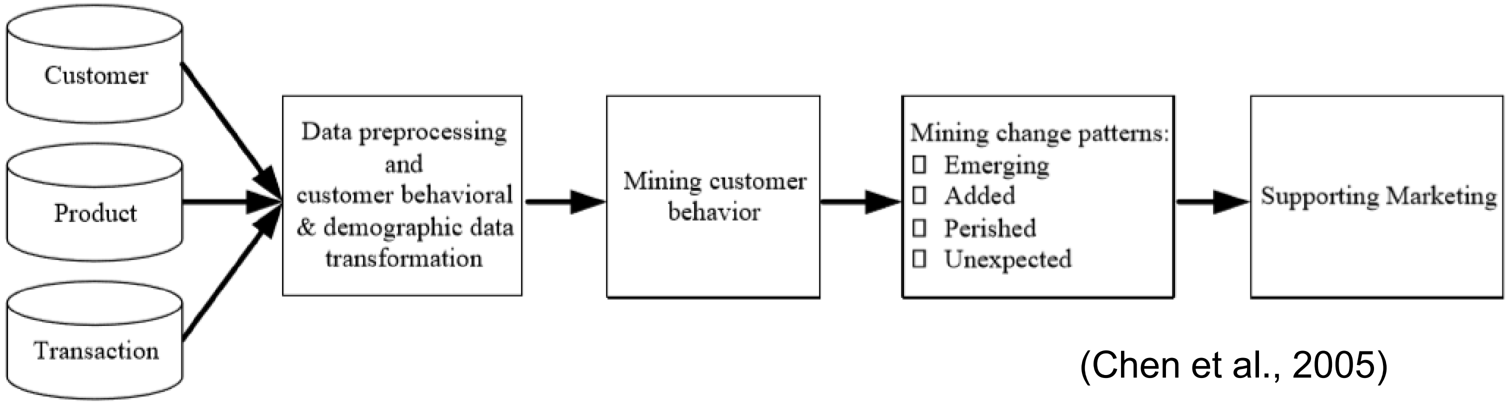


| Transaction ID | Items |
|---|---|
| 1 | {Bread, Butter, Diapers, Milk} |
| 2 | {Coffee, Sugar, Cookies, Salmon} |
| 3 | {Bread, Butter, Coffee, Diapers, Milk, Eggs} |
| 4 | {Bread, Butter, Salmon, Chicken} |
| 5 | {Eggs, Bread, Butter} |
| 6 | {Salmon, Diapers, Milk} |
| 7 | {Bread, Tea, Sugar, Eggs} |
| 8 | {Coffee, Sugar, Chicken, Eggs} |
| 9 | {Bread, Diapers, Milk, Salt} |
| 10 | {Tea, Eggs, Cookies, Diapers, Milk} |

# Introduction

- Examples: Business
- Applications: Mining changes in customer behavior



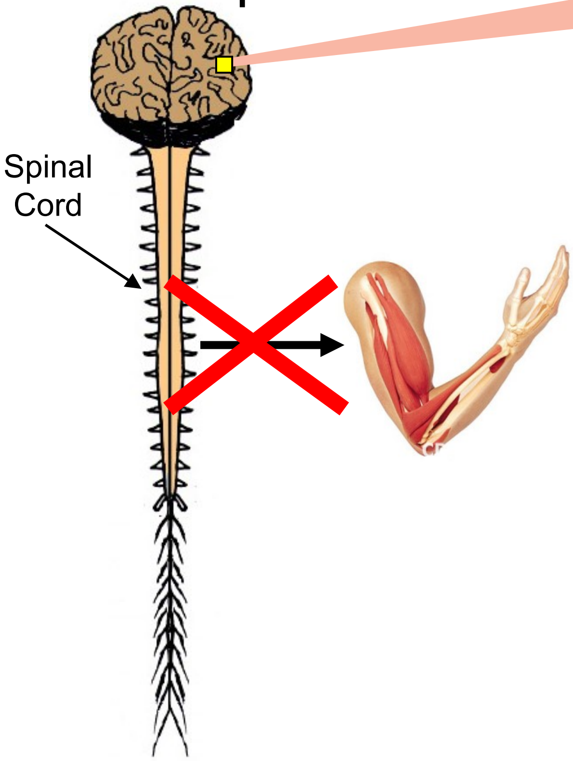


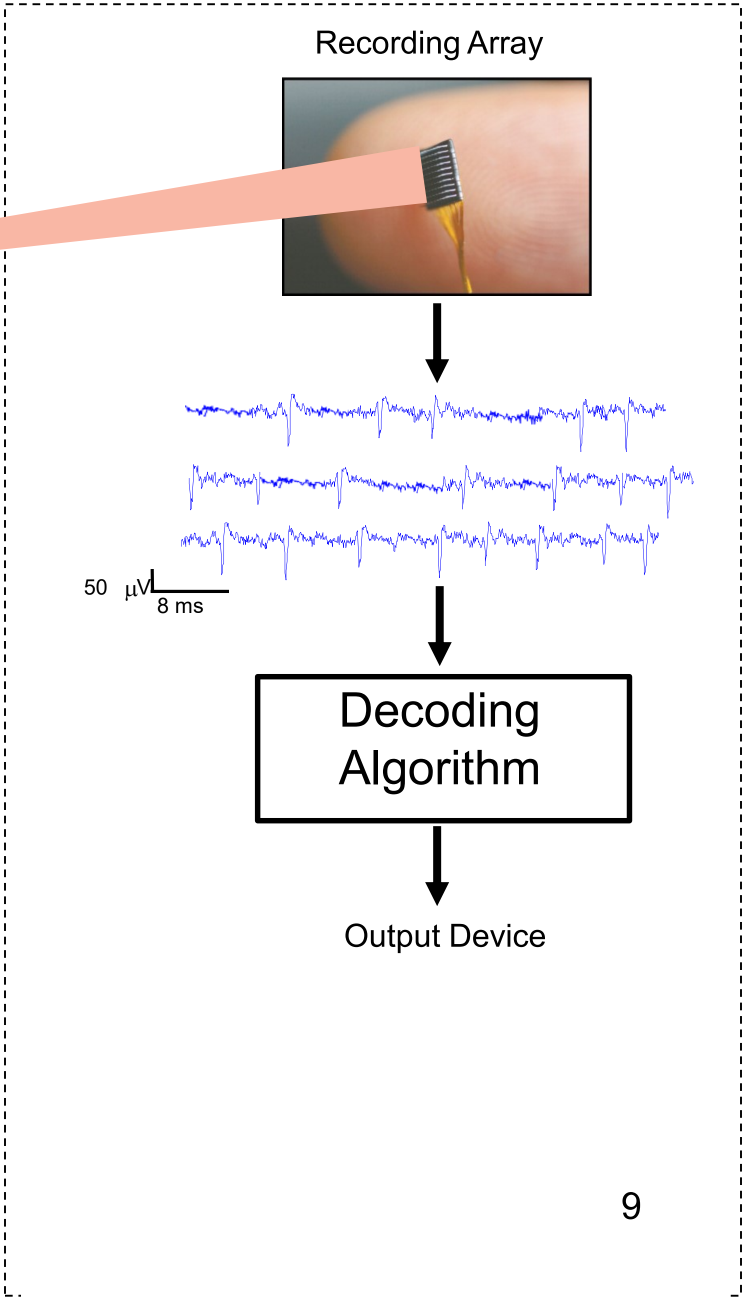
(Chen et al., 2005)

# Introduction

- Examples: Healthcare
- Applications: Brain-computer interface

(Hochberg et al., 2006)





Recording Array

Decoding Algorithm

Output Device

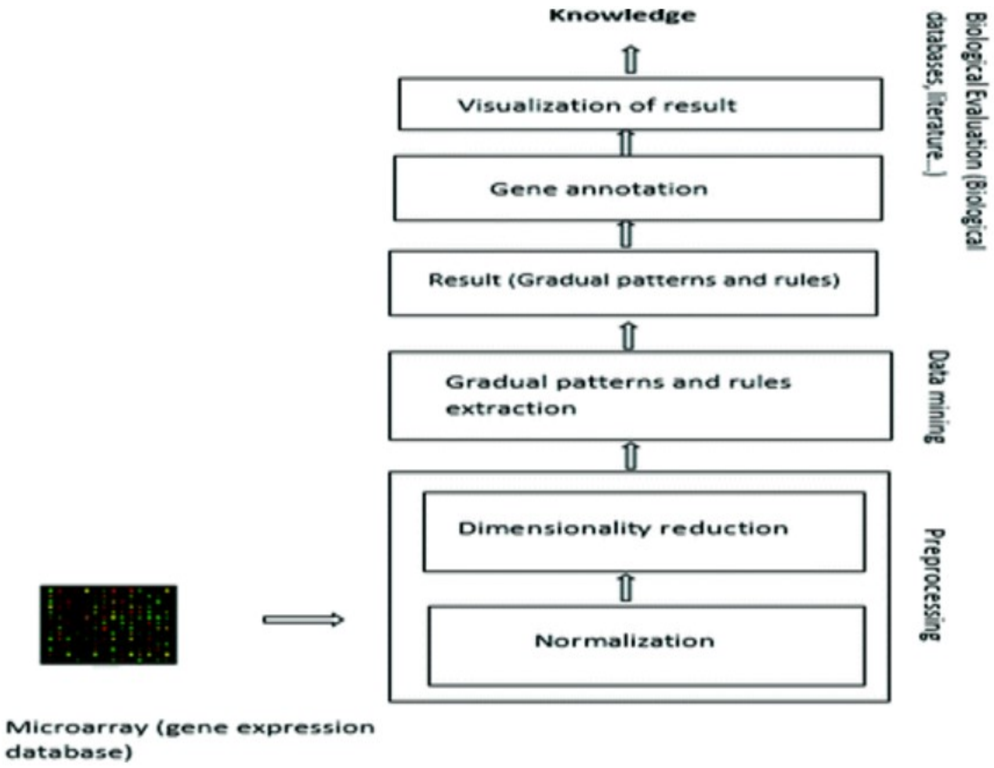
# Introduction

- In 2012, scientists at Brown University, USA, reported a BCI that a paralyzed subject can use to control a robotic arm to grab a bottle and drink from it (Hochberg et al., 2012)

http://www.youtube.com/watch?v=cg5RO8Qv6mc

# Introduction

- Examples: Healthcare
- Applications: Mining Gene Expression Data



Knowledge

Visualization of result

Gene annotation

Result (Gradual patterns and rules)

Gradual patterns and rules extraction

Dimensionality reduction

Normalization

Microarray (gene expression database)

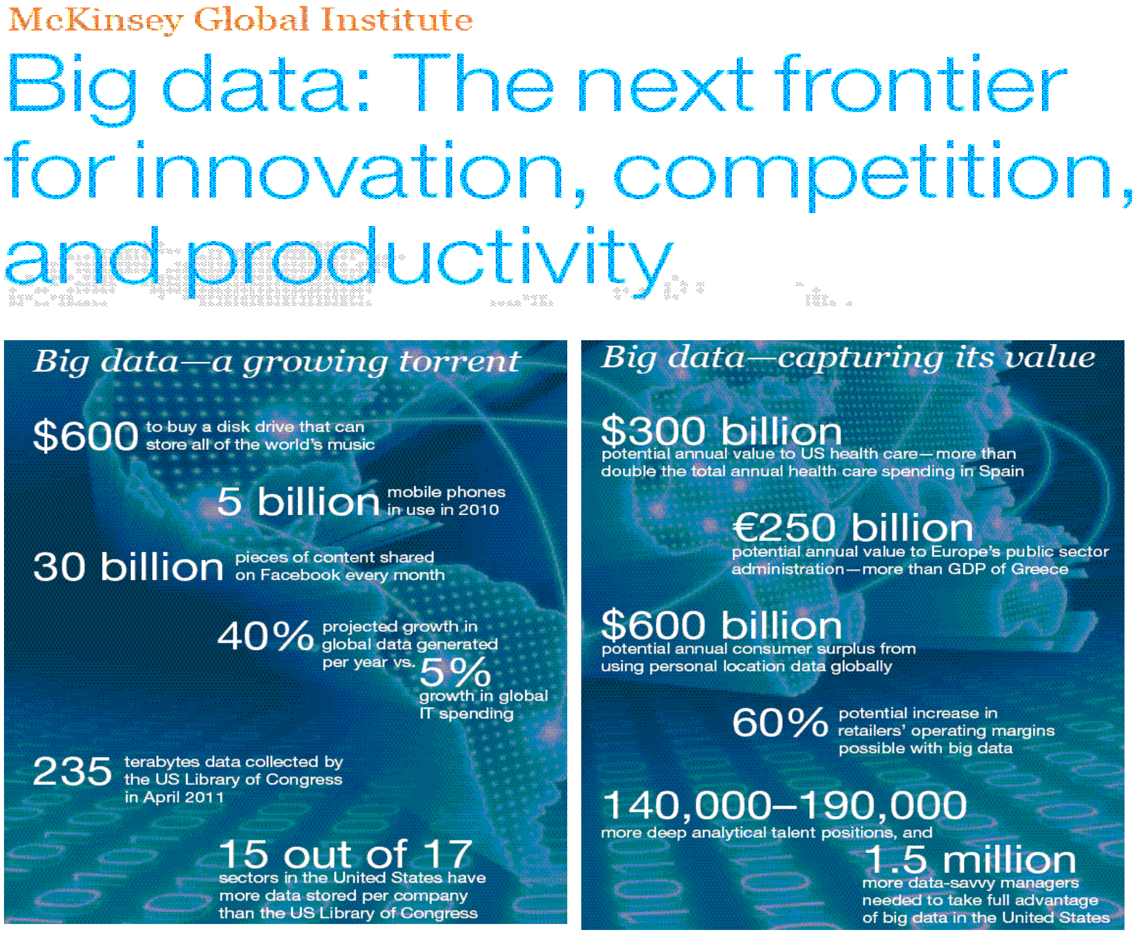
Biological Evaluation (Biological databases, literature...)

Data mining

Preprocessing

# Introduction

- Advances in Big Data technologies motivate more investment in Data Mining Techniques



McKinsey Global Institute

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

*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

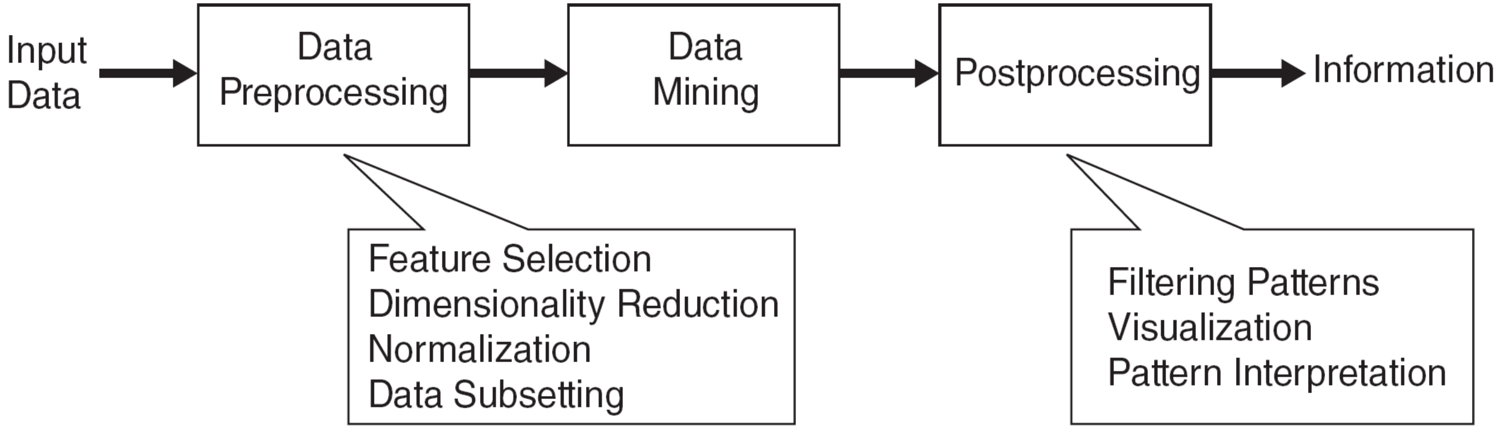
60% potential increase in retailers' operating margins possible with big data

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

# Data Mining and Knowledge Discovery

- Information retrieval is not Data Mining
- Examples of “not” Data Mining:
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”
- The process of Knowledge Discovery in Databases (KDD)

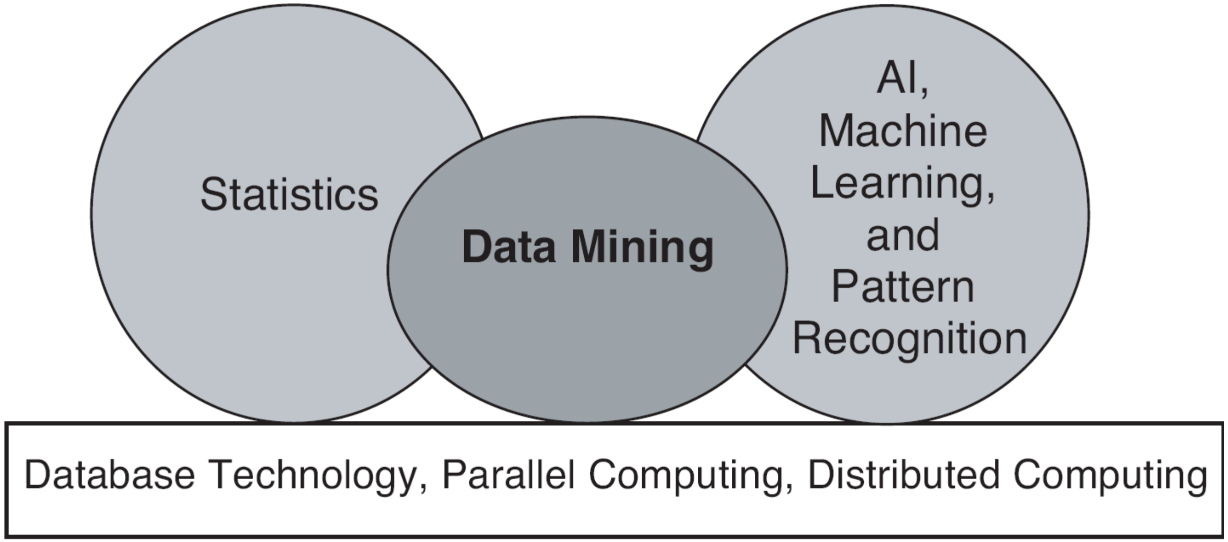


# Origins of Data Mining

- Data mining draws upon ideas, such as

  (1) sampling, estimation, and hypothesis testing from statistics

  (2) search algorithms, modeling techniques, and learning theories from artificial intelligence, pattern recognition, and machine learning



Statistics

Data Mining

AI, Machine Learning, and Pattern Recognition

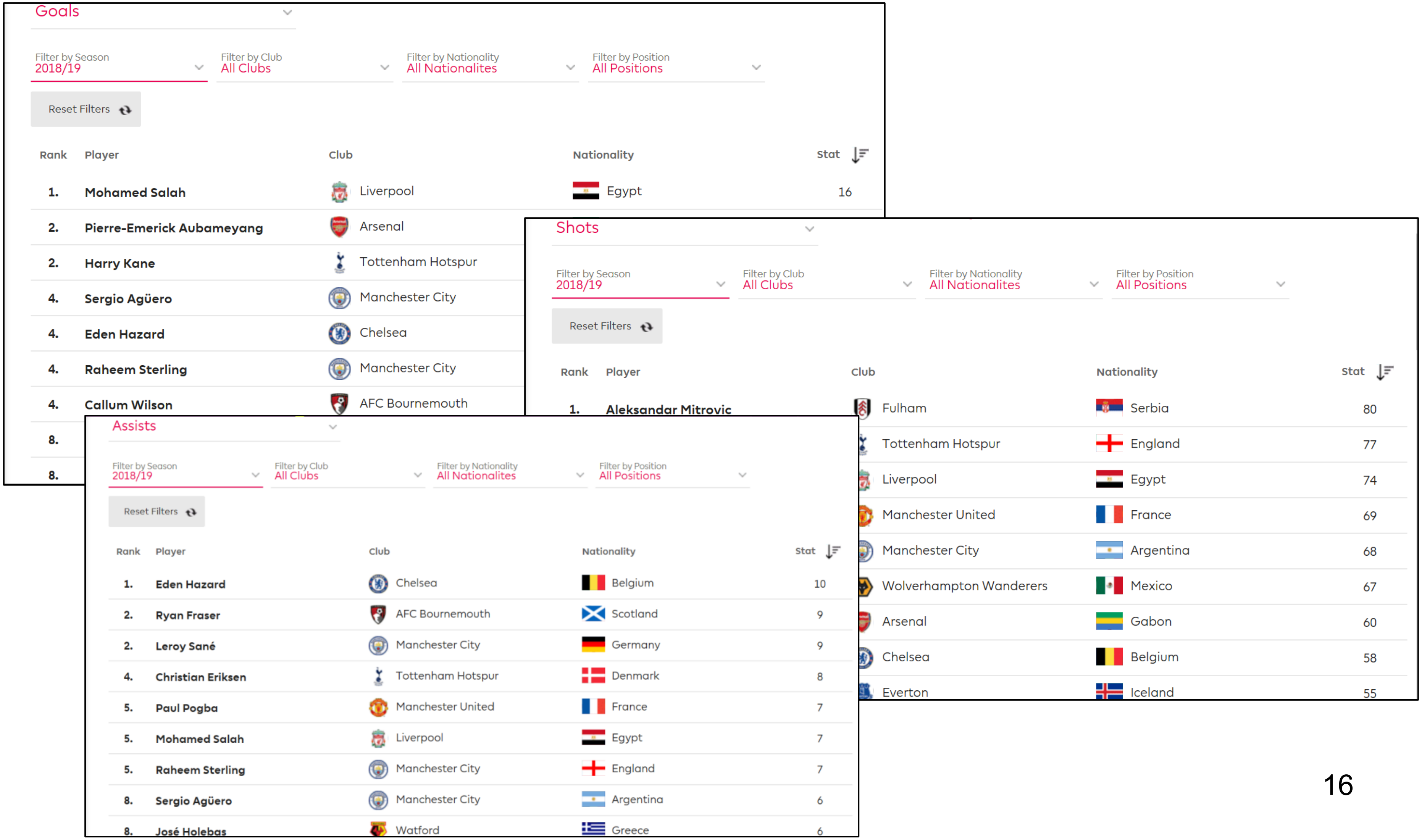
Database Technology, Parallel Computing, Distributed Computing

# Data Mining Tasks

- Data mining tasks are generally divided into two major categories:
  - Predictive tasks: To predict the value of a particular attribute based on the values of other attributes.
  - Descriptive tasks: To derive patterns that summarize the underlying relationships in data.

# Data Mining Tasks

- Core Data Mining Tasks: Premier League Player Statistics:



Goals

Filter by Season 2018/19 | Filter by Club All Clubs | Filter by Nationality All Nationalites | Filter by Position All Positions

Reset Filters

| Rank | Player | Club | Nationality | Stat |
|---|---|---|---|---|
| 1. | Mohamed Salah | Liverpool | Egypt | 16 |
| 2. | Pierre-Emerick Aubameyang | Arsenal | | |
| 2. | Harry Kane | Tottenham Hotspur | | |
| 4. | Sergio Agüero | Manchester City | | |
| 4. | Eden Hazard | Chelsea | | |
| 4. | Raheem Sterling | Manchester City | | |
| 4. | Callum Wilson | AFC Bournemouth | | |
| 8. | | | | |
| 8. | | | | |

Shots

Filter by Season 2018/19 | Filter by Club All Clubs | Filter by Nationality All Nationalites | Filter by Position All Positions

Reset Filters

| Rank | Player | Club | Nationality | Stat |
|---|---|---|---|---|
| 1. | Aleksandar Mitrovic | Fulham | Serbia | 80 |
| | | Tottenham Hotspur | England | 77 |
| | | Liverpool | Egypt | 74 |
| | | Manchester United | France | 69 |
| | | Manchester City | Argentina | 68 |
| | | Wolverhampton Wanderers | Mexico | 67 |
| | | Arsenal | Gabon | 60 |
| | | Chelsea | Belgium | 58 |
| | | Everton | Iceland | 55 |

Assists

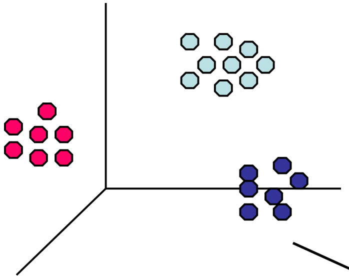
Filter by Season 2018/19 | Filter by Club All Clubs | Filter by Nationality All Nationalites | Filter by Position All Positions

Reset Filters

| Rank | Player | Club | Nationality | Stat |
|---|---|---|---|---|
| 1. | Eden Hazard | Chelsea | Belgium | 10 |
| 2. | Ryan Fraser | AFC Bournemouth | Scotland | 9 |
| 2. | Leroy Sané | Manchester City | Germany | 9 |
| 4. | Christian Eriksen | Tottenham Hotspur | Denmark | 8 |
| 5. | Paul Pogba | Manchester United | France | 7 |
| 5. | Mohamed Salah | Liverpool | Egypt | 7 |
| 5. | Raheem Sterling | Manchester City | England | 7 |
| 8. | Sergio Agüero | Manchester City | Argentina | 6 |
| 8. | José Holebas | Watford | Greece | 6 |

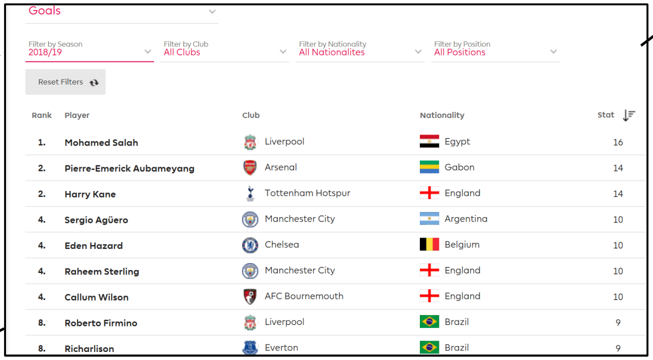
# Data Mining Tasks

- Core Data Mining Tasks: Premier League Player Statistics:



Clustering

Data



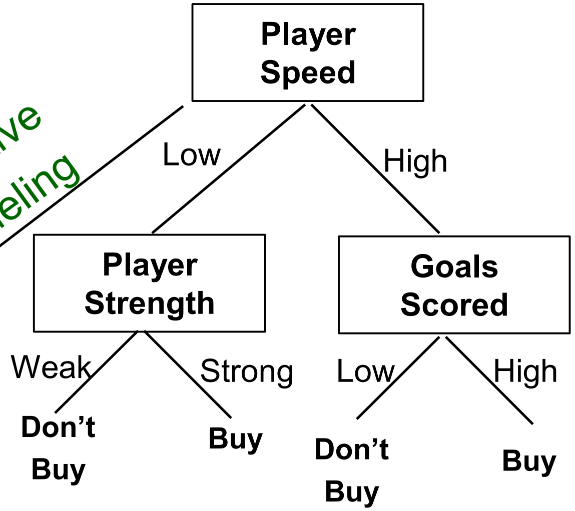
Goals

Filter by Season 2018/19 | Filter by Club All Clubs | Filter by Nationality All Nationalites | Filter by Position All Positions

Reset Filters

| Rank | Player | Club | Nationality | Stat |
|---|---|---|---|---|
| 1. | Mohamed Salah | Liverpool | Egypt | 16 |
| 2. | Pierre-Emerick Aubameyang | Arsenal | Gabon | 14 |
| 2. | Harry Kane | Tottenham Hotspur | England | 14 |
| 4. | Sergio Agüero | Manchester City | Argentina | 10 |
| 4. | Eden Hazard | Chelsea | Belgium | 10 |
| 4. | Raheem Sterling | Manchester City | England | 10 |
| 4. | Callum Wilson | AFC Bournemouth | England | 10 |
| 8. | Roberto Firmino | Liverpool | Brazil | 9 |
| 8. | Richarlison | Everton | Brazil | 9 |

Predictive Modeling



Player Speed

Low

High

Player Strength

Goals Scored

Weak

Strong

Low

High

Don't Buy

Buy

Don't Buy

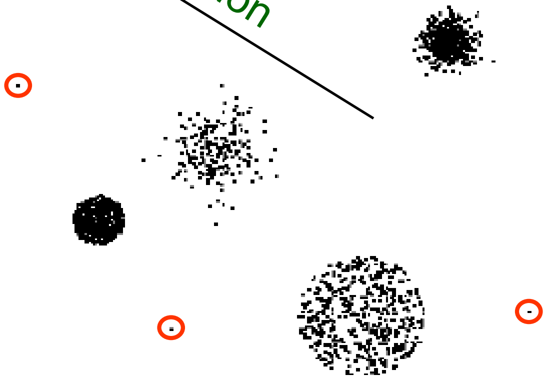
Buy

Association Rules

Player Speed

Number of Goals Scored

Anomaly Detection



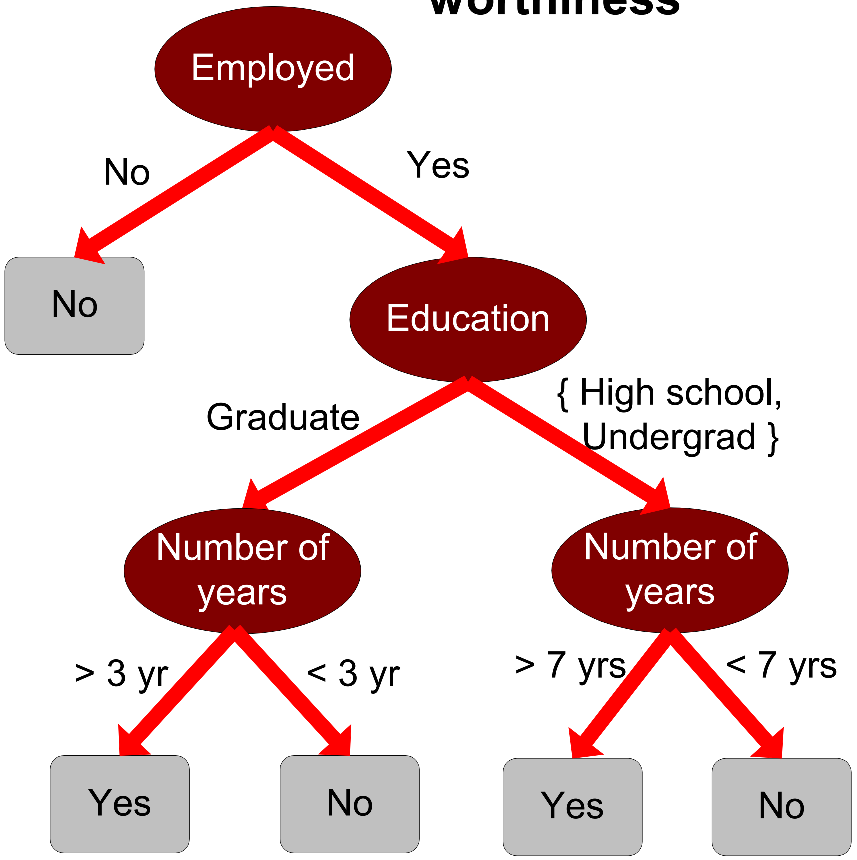
# Predictive Modeling: Classification

- 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**



# Predictive Modeling: Classification

categorical

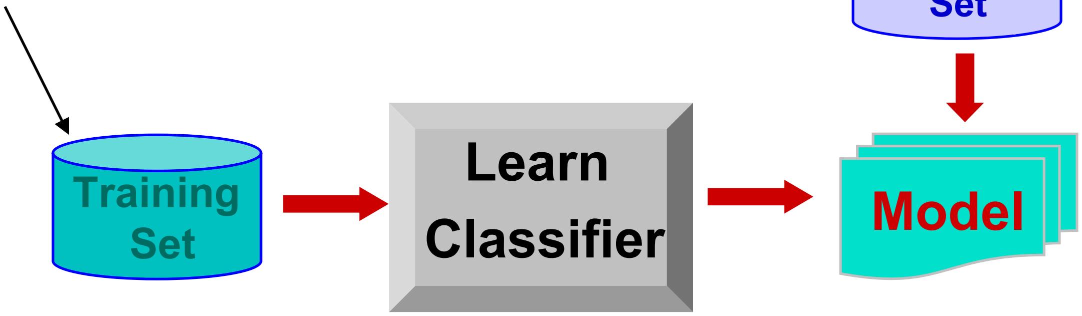
categorical

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 | ? |
| … | … | … | … | … |



Test Set

Training Set

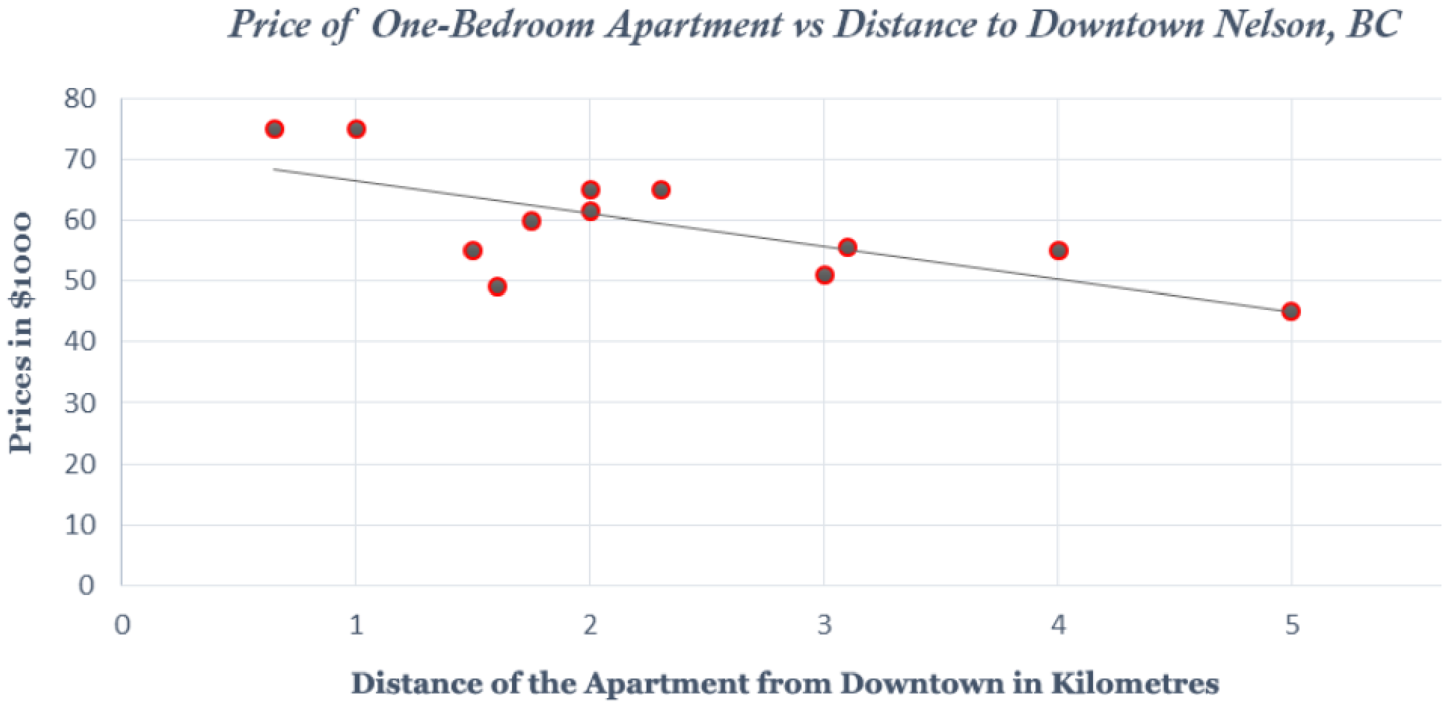
Learn Classifier

Model

# Predictive Modeling: Regression

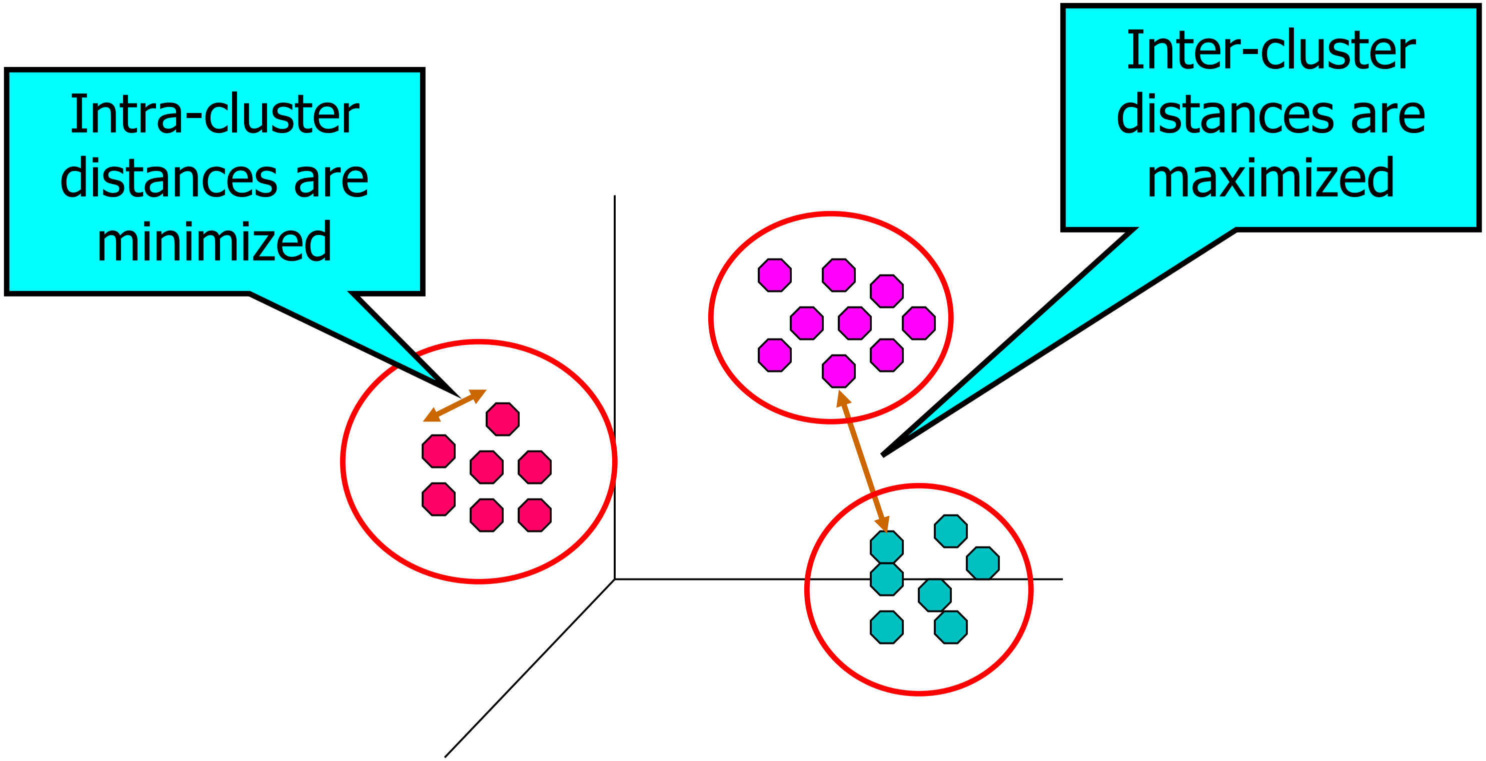
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency

- Example: Pricing



# 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

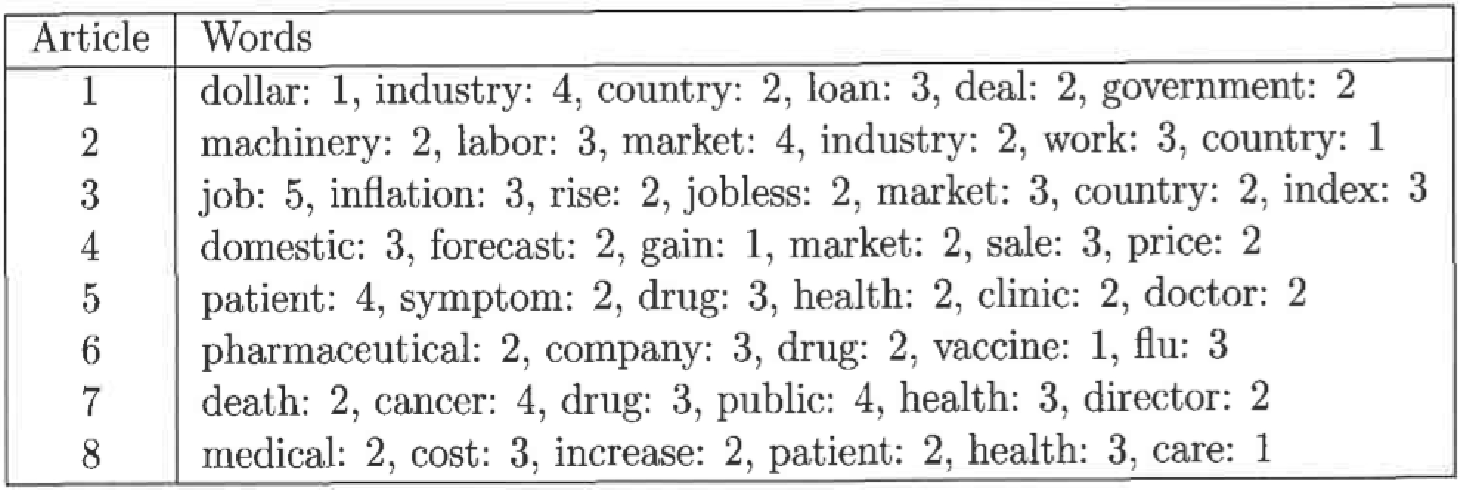


Intra-cluster distances are minimized

Inter-cluster distances are maximized

# Clustering

- Example: Document Clustering



| Article | Words |
|---|---|
| 1 | dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2 |
| 2 | machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1 |
| 3 | job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3 |
| 4 | domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2 |
| 5 | patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2 |
| 6 | pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3 |
| 7 | death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2 |
| 8 | medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1 |

Cluster 1: Economy

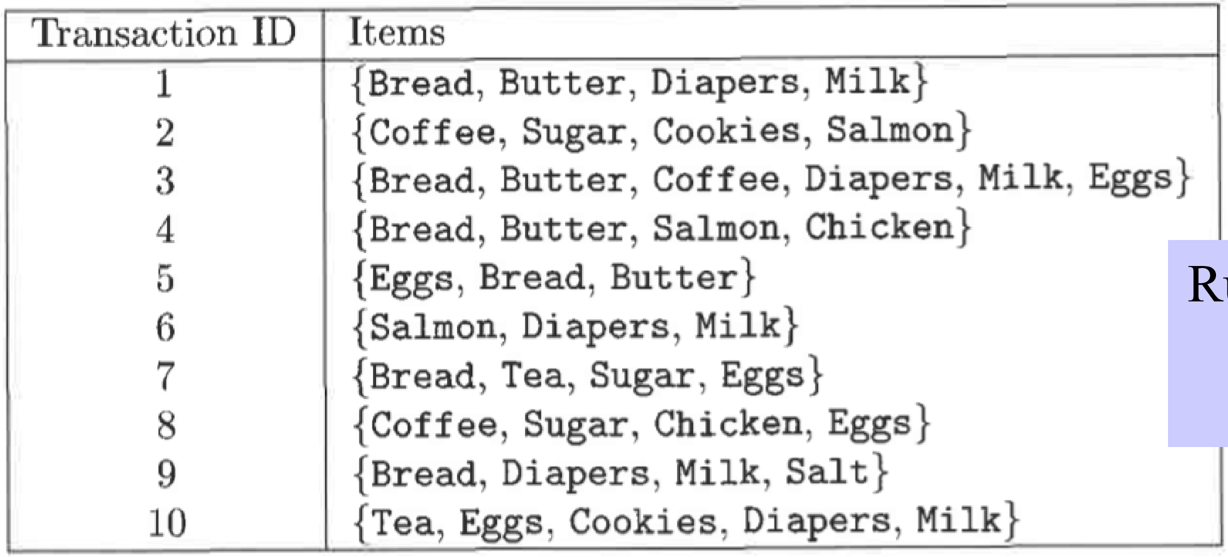
Cluster 2: Healthcare

# Association Rule Discovery

- Discover patterns that describe strongly associated features in the data

- The goal of association analysis is to extract the most interesting patterns in an efficient manner

- Example: Market Basket Analysis



| Transaction ID | Items |
|---|---|
| 1 | {Bread, Butter, Diapers, Milk} |
| 2 | {Coffee, Sugar, Cookies, Salmon} |
| 3 | {Bread, Butter, Coffee, Diapers, Milk, Eggs} |
| 4 | {Bread, Butter, Salmon, Chicken} |
| 5 | {Eggs, Bread, Butter} |
| 6 | {Salmon, Diapers, Milk} |
| 7 | {Bread, Tea, Sugar, Eggs} |
| 8 | {Coffee, Sugar, Chicken, Eggs} |
| 9 | {Bread, Diapers, Milk, Salt} |
| 10 | {Tea, Eggs, Cookies, Diapers, Milk} |

Rules Discovered:

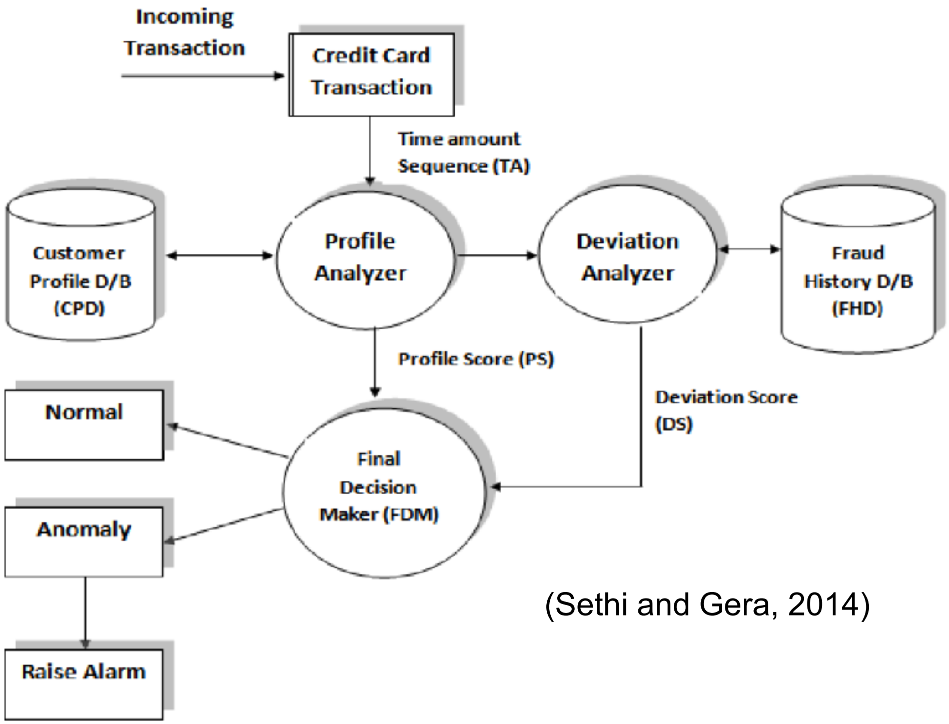
{Diapers} --> {Milk}

{Butter} --> {Bread}

# Anomaly Detection

- Identifying observations whose characteristics are significantly different from the rest of the data

- Example: Credit Card Fraud Detection
- The number of fraudulent cases is relatively small compared to the number of legitimate transactions



(Sethi and Gera, 2014)

# Course Outline

- Introduction
- Linear Algebra and Probability Theory Review
- Data
- Data Exploration
- Classification
- Association Analysis
- Cluster Analysis
- Anomaly Detection