# Welcome to Business Data Mining IS 4482 - 001

TUESDAY 6:00PM – 9:00PM

BUC 108

DA XU

# Course Description

This course introduce

- Basic concepts, fundamental information theories and methodologies in data mining
- Core techniques/algorithms for data analytics, including classification, numerical prediction, clustering, and association rule mining
- Hands on practices of major data mining techniques with real-world applications

# Materials

- Textbook (Optional): *Machine Learning with R*, Second Edition, by Brett Lantz, 2015.

- R and RStudio (RStudio Cloud)

# Hybrid Class

- This is NOT a computer science class, but you will be required to write code and likely learn new syntax
- This is NOT a master's level data mining class, so we will not take an in-depth dive into any one principle or method
- This is basically an introduction to data mining class where you will be exposed to a number of different concepts and learn to combine them to solve business problems

# Grading

| | Points | Percentage |
|---|---|---|
| 2 Exams (75 points each) | 150 | 30% |
| 2 Quizzes (25 points each) | 50 | 10% |
| 10 Lab Assignments (10 points each) | 100 | 20% |
| 4 Homework Assignments (50 points each) | 200 | 40% |
| | | |
| Total | 500 | 100% |

- Assignments must be complete individually
- All Quizzes and Exams are open book

# Grading Scale

| | | | |
|---|---|---|---|
| A = | 93-100 | C = | 73-76.99 |
| A- = | 90-92.99 | C- = | 70-72.99 |
| B+ = | 87-89.99 | D+ = | 67-69.99 |
| B = | 83-86.99 | D = | 63-66.99 |
| B- = | 80-82.99 | D- = | 60-62.99 |
| C+ = | 77-79.99 | E = | Less than 60 |

# Course Schedule

- Tentative course schedule and important dates are on Syllabus.

# Communication

- Primary communication method will be email
  - da.xu@eccles.utah.edu
- Canvas announcement will be made from time to time, please pay attention to them
- Office hours
  - Tuesday 4:20 – 5:50 pm (BUC 3)
  - Or by appointment

# Attendance

- The University expects regular attendance at all class meetings, however, no portion of your grade will be tied directly to attendance
- A large portion of your grade is based on in classroom activities and/or assignments
  - Labs, Quizzes, and Exams will be completed on class
- All material covered in class is expected to be used on the exams

# Canvas

- All class materials will be made available on Canvas including lab assignments, homework, data sets, and course slides
- All assignments must be submitted on Canvas
- Quiz and exam will be on Canvas as well

# Assignments

- All homework and lab assignments are due at 11.59 pm on the due date as posted in Canvas
- For late assignments, the final assignment score will be dropped by 30% (up to 3 days late)
- **Do not wait** until the last moment to submit on Canvas

# Quizzes and Exams

- All Quizzes and Exams are open book
- Quizzes and Exams include
  - Part1: True/False and multiple choice questions
  - Part2: Short answers
  - Part3: Code writing – you will be required to write and submit R code on Canvas.

# Question?

# Lecture 1: Introduction to Data Mining

# Overview

- What is Data Mining
- Why Data Mining
- Data Mining process
- Data Mining Tasks and Applications
- Data Mining Trends

# What is Data Mining

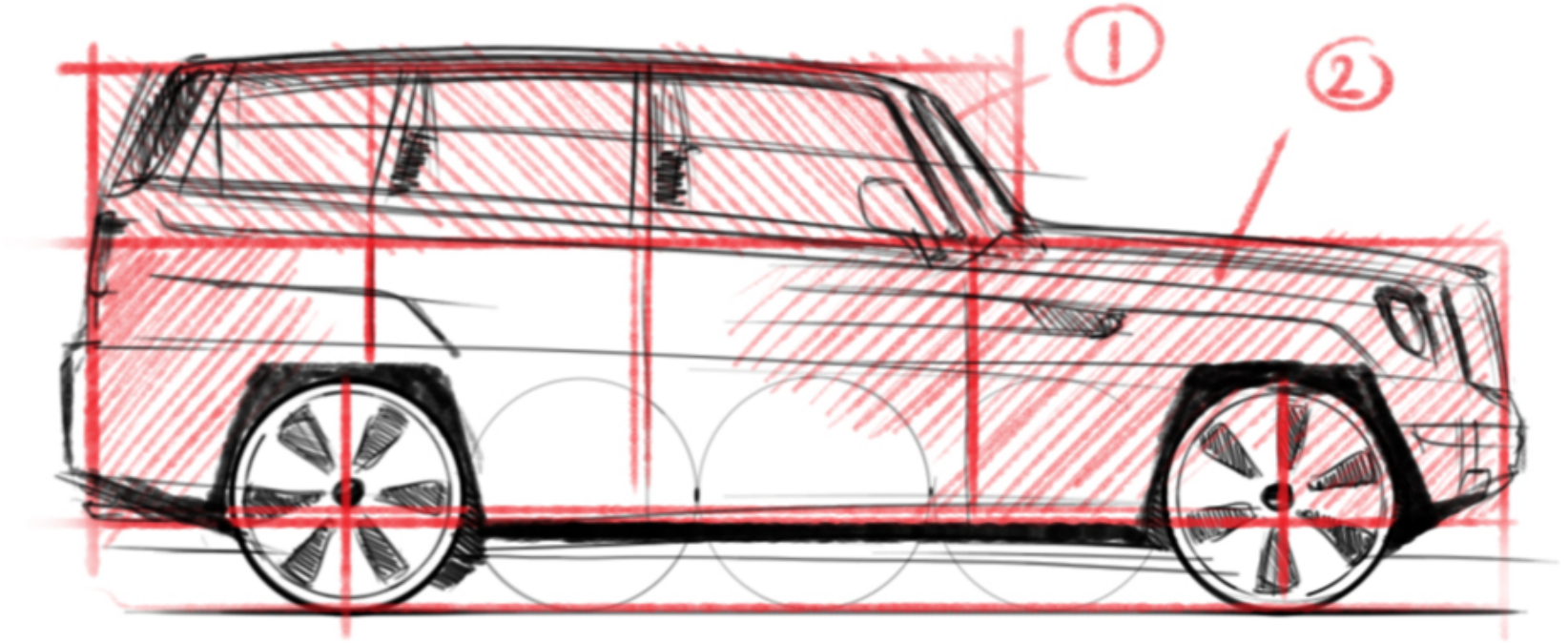
- **Data mining** (sometimes called data or knowledge discovery) is the process of efficiently discover useful **patterns** in large quantity of **data**.
  - **Big Data**- a term used to describe the large volume of data that we are learning to capture, store, and use in increasingly efficient ways
  - **Machine Learning**- the field of study interested in the development of computer algorithms to transform data into intelligent action
  - **Data Mining**- is concerned with the generation of novel insights from large databases

# What is Data Mining

- Data knows more about your music preference

# What is Data Mining

- Car manufacturer improved vehicle models via the social media platform



# What is Data Mining

- Use Big data to forecast crime
  - PredPol, working with Los Angeles and Santa Cruz police and a team of researchers, predicts the odds of a crime occurring. In Los Angeles, the distribution of theft and violent crime dropped by 33% and 21% correspondingly.

# Why Data Mining?

- Data: data are any facts, numbers, or text that can be processed by a computer.
  - Computerization of businesses produce huge amount of data
  - Online e-businesses are generating even larger data sets

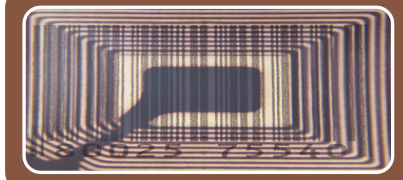
# What Is Data Mining?

- Data

Business data – e.g., demographic data, transactional databases, and customer behavior data

Web data – e.g., page content, page linkage and clickstream

System data – e.g., service logs, quality and availability of network, computer, and manufacturing devices



Sensor data – e.g. RFID devices for homeland security, supply-chain and environmental applications

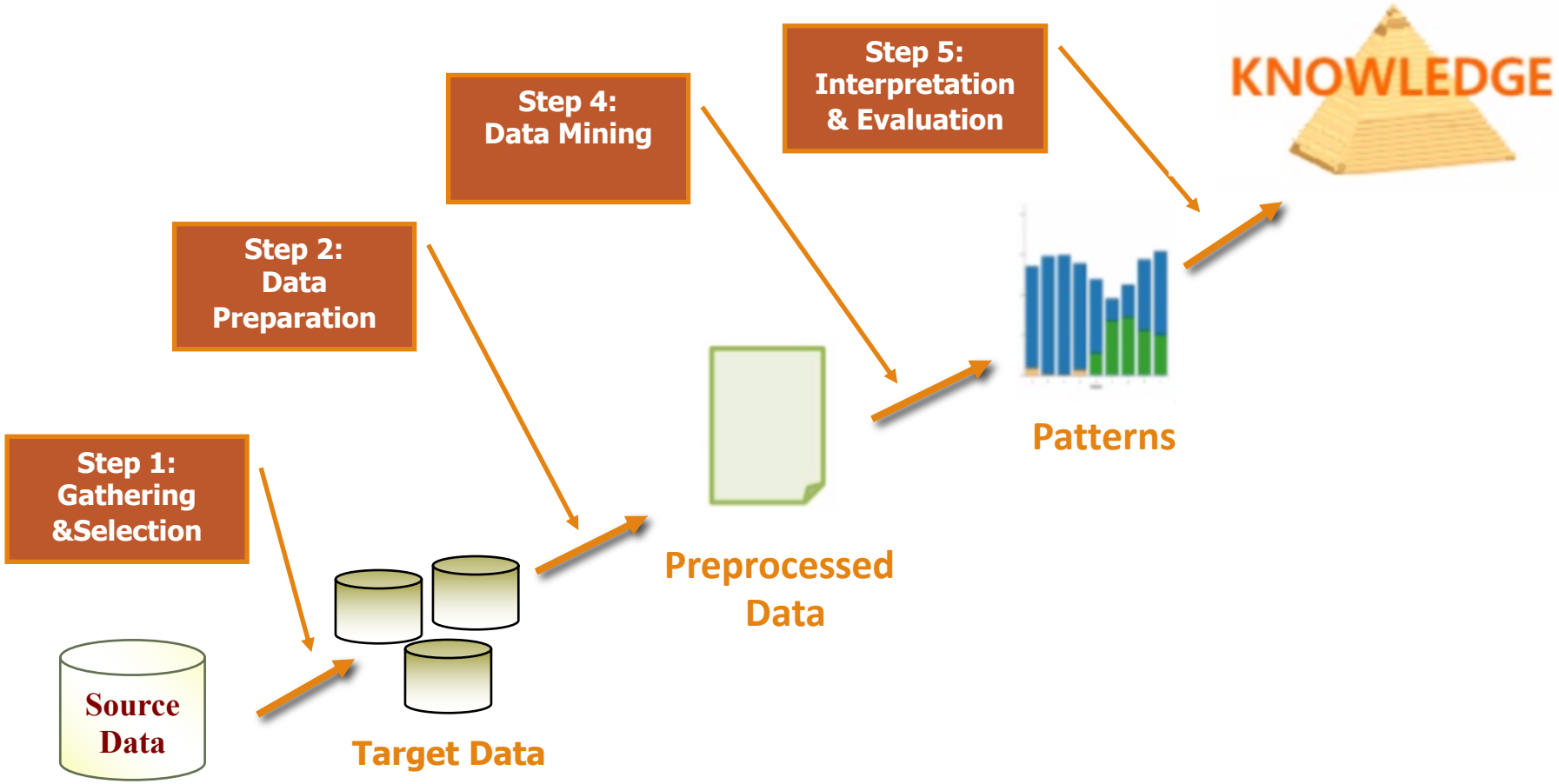
# Why Data Mining?

- There is a big gap from stored data to knowledge

# Why Data Mining?

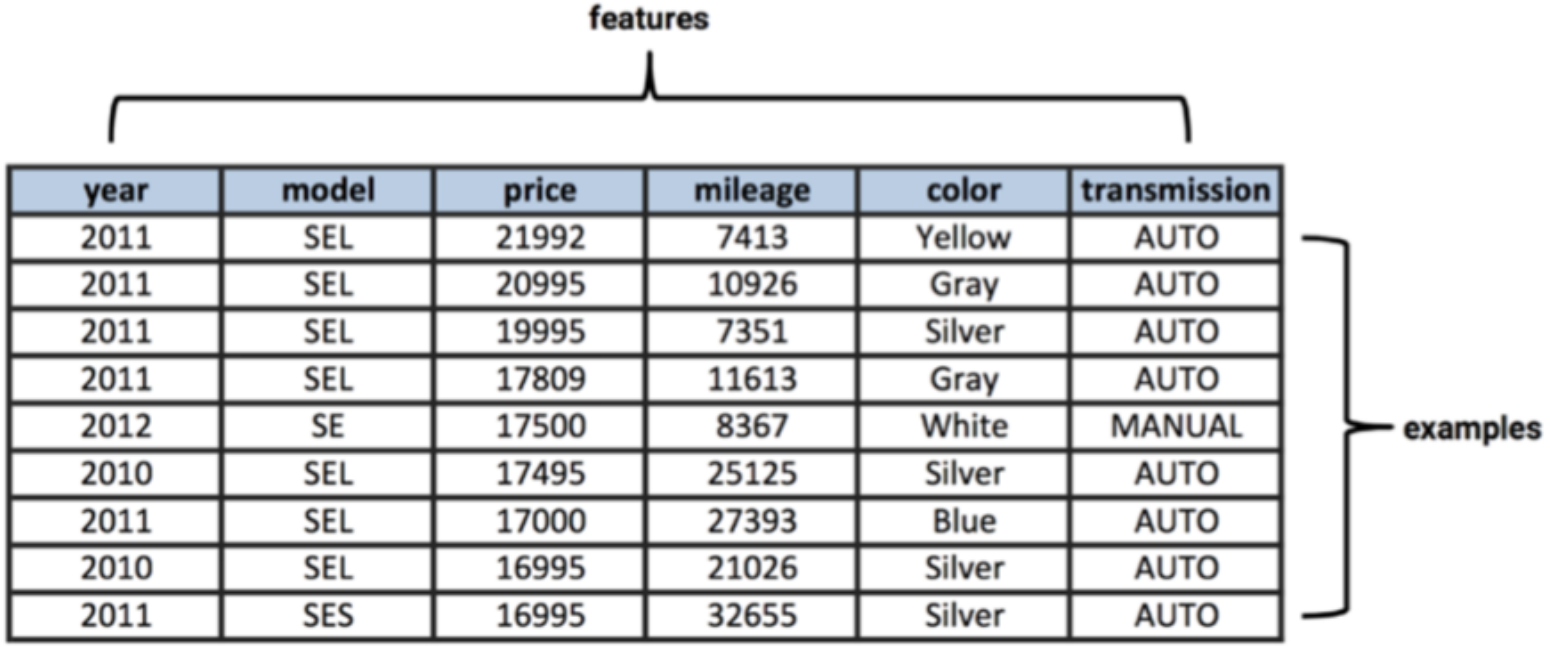
- Data mining helps generate insights from data to support business decisions and improve the profitability and efficiency

# Data Mining Process



# Data Mining Process

- **Data understanding:** With preliminary analysis, data exploration provides a high level overview of each attribute in the data set and interaction between the attributes.




| year | model | price | mileage | color | transmission |
|---|---|---|---|---|---|
| 2011 | SEL | 21992 | 7413 | Yellow | AUTO |
| 2011 | SEL | 20995 | 10926 | Gray | AUTO |
| 2011 | SEL | 19995 | 7351 | Silver | AUTO |
| 2011 | SEL | 17809 | 11613 | Gray | AUTO |
| 2012 | SE | 17500 | 8367 | White | MANUAL |
| 2010 | SEL | 17495 | 25125 | Silver | AUTO |
| 2011 | SEL | 17000 | 27393 | Blue | AUTO |
| 2010 | SEL | 16995 | 21026 | Silver | AUTO |
| 2011 | SES | 16995 | 32655 | Silver | AUTO |

If a feature represents a characteristic measured in numbers, it is unsurprisingly called **numeric.**

if a feature is an attribute that consists of a set of categories, the feature is called **categorical** or **nominal**.

# Data Mining Process

- **Data preparation:** Before applying the data mining algorithm, we need to prepare the data set for handling of any anomalies that may be present in the data.
  - Finding outliers; e.g., patients with age >120
  - Missing values; data sparsity
  - Duplicate
  - Highly correlated attributes; age/DOB
  - Data cleaning and transforming

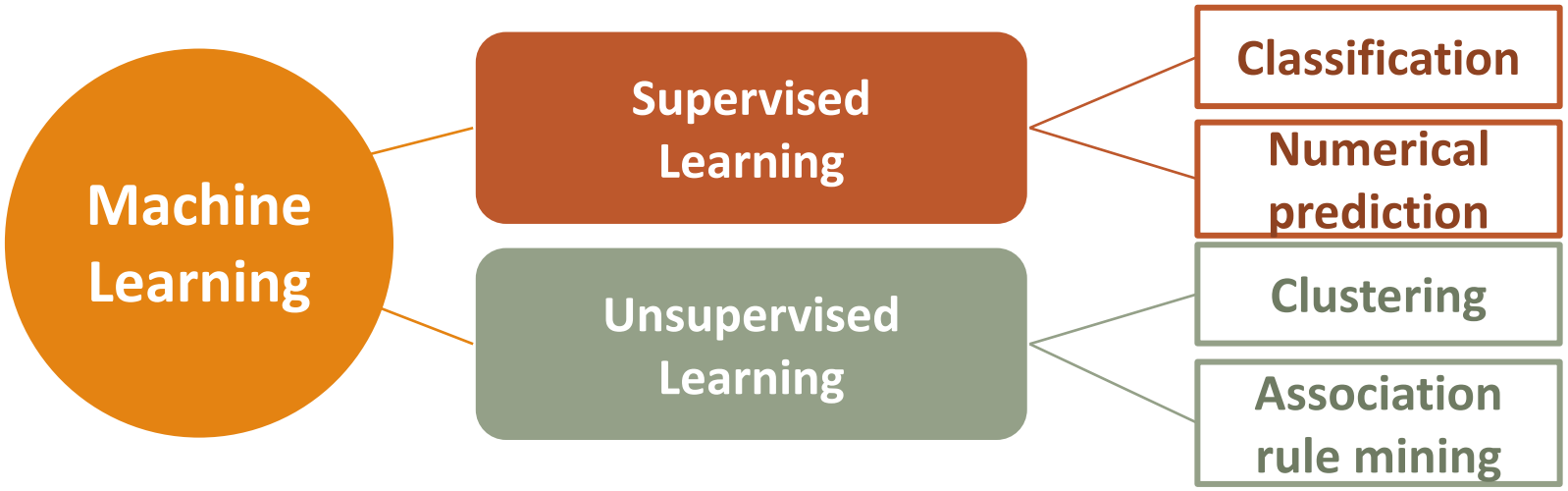
# Data Mining Process

- **Data mining tasks**
  - Classification
  - Numerical prediction
  - Clustering
  - Association rule mining

# Data Mining Process

- **Interpretation and evaluation:**
  - Understanding the prediction, classification, or clustering results of the data mining process.
  - Because each machine learning model results in a biased solution to the learning problem, it is important to evaluate how well the algorithm learns from its experience. Depending on the type of model used, you might be able to evaluate the accuracy of the model using a test dataset or you may need to develop measures of performance specific to the intended application.

# Data Mining Tasks



# Data Mining Tasks

- **Supervised learning** is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.
  - Y=f(X); X $\xrightarrow{\text{predict}}$ Y
  - The supervision does not refer to human involvement, but rather to the fact that the target values provide a way for the learner to know how well it has learned the desired task.
  - A **predictive model** is used for tasks that involve, as the name implies, the prediction of one value using other values in the dataset.

| age | sex | bmi | children | smoker | region | expenses | InsurancePlan |
|---|---|---|---|---|---|---|---|
| 18 | male | 23.2 | 0 | no | southeast | 1121.87 | Basic |
| 18 | male | 30.1 | 0 | no | southeast | 1131.51 | Basic |
| 18 | male | 33.3 | 0 | no | southeast | 1135.94 | Basic |
| 18 | male | 33.7 | 0 | no | southeast | 1136.4 | Basic |

# Data Mining Tasks

- **Unsupervised learning** is where you only have input data (X) and no corresponding output variables.
  - A **descriptive model** is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways.
  - There is no target to learn, the process of training a descriptive model is called unsupervised learning.

# Data Mining Tasks

- **Supervised learning-Classification**
  - A classification problem is to predict which category an example belongs to
    - Output/Target variable is a category
      - An e-mail message is spam
      - A person has cancer
      - A football team will win or lose
      - An applicant will default on a loan

# Data Mining Tasks

- **Supervised learning- Classification**
  - Predicting Pokemon battle winner with classification

Image Source : The Verge

# Data Mining Tasks

- **Supervised learning-Classification**

| # | Name | Type 1 | Type 2 | HP | Attack | Defense | Sp. Atk | Sp. Def | Speed | Generation | Legendary |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Bulbasaur | Grass | Poison | 45 | 49 | 49 | 65 | 65 | 45 | 1 | FALSE |
| 2 | Ivysaur | Grass | Poison | 60 | 62 | 63 | 80 | 80 | 60 | 1 | FALSE |
| 3 | Venusaur | Grass | Poison | 80 | 82 | 83 | 100 | 100 | 80 | 1 | FALSE |
| 4 | Mega Venus | Grass | Poison | 80 | 100 | 123 | 122 | 120 | 80 | 1 | FALSE |
| 5 | Charmander | Fire | | 39 | 52 | 43 | 60 | 50 | 65 | 1 | FALSE |
| 6 | Charmeleon | Fire | | 58 | 64 | 58 | 80 | 65 | 80 | 1 | FALSE |
| 7 | Charizard | Fire | Flying | 78 | 84 | 78 | 109 | 85 | 100 | 1 | FALSE |
| 8 | Mega Chariz | Fire | Dragon | 78 | 130 | 111 | 130 | 85 | 100 | 1 | FALSE |
| 9 | Mega Chariz | Fire | Flying | 78 | 104 | 78 | 159 | 115 | 100 | 1 | FALSE |

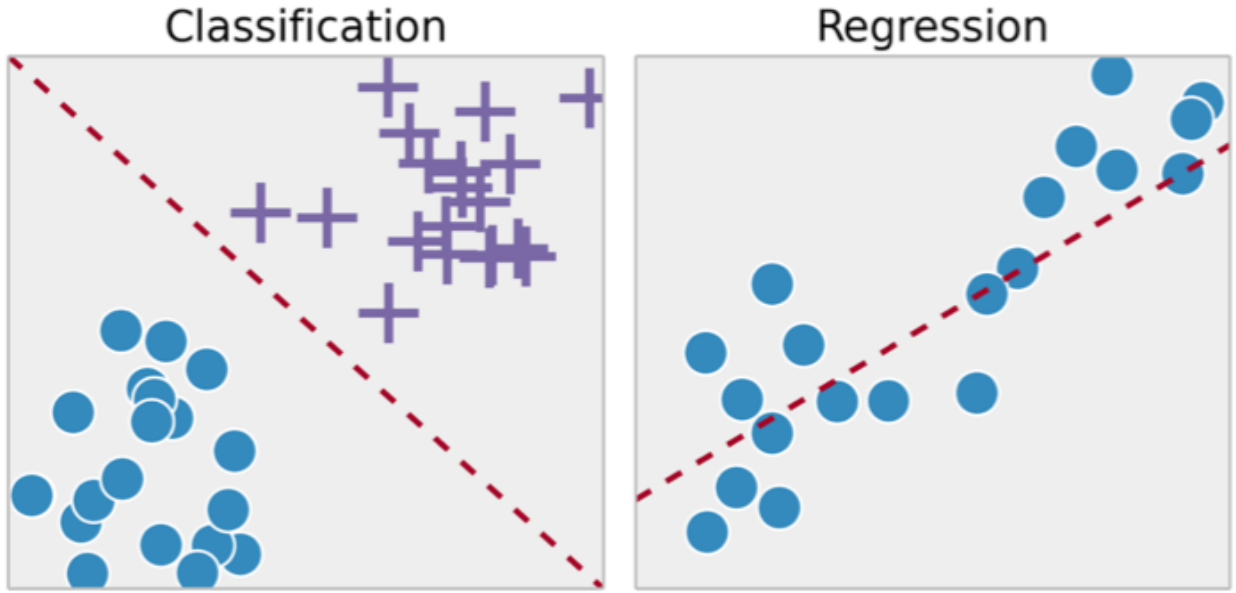
| First_pokem | Second_poke | Winner |
|---|---|---|
| 266 | 298 | 298 |
| 702 | 701 | 701 |
| 191 | 668 | 668 |
| 237 | 683 | 683 |
| 151 | 231 | 151 |
| 657 | 752 | 657 |

# Data Mining Tasks

- **Supervised learning-Classification**

# Data Mining Tasks

- **Supervised learning-Numerical Prediction**
  - A numerical prediction problem is when the output variable is a real value, such as “dollars” or “weight”.



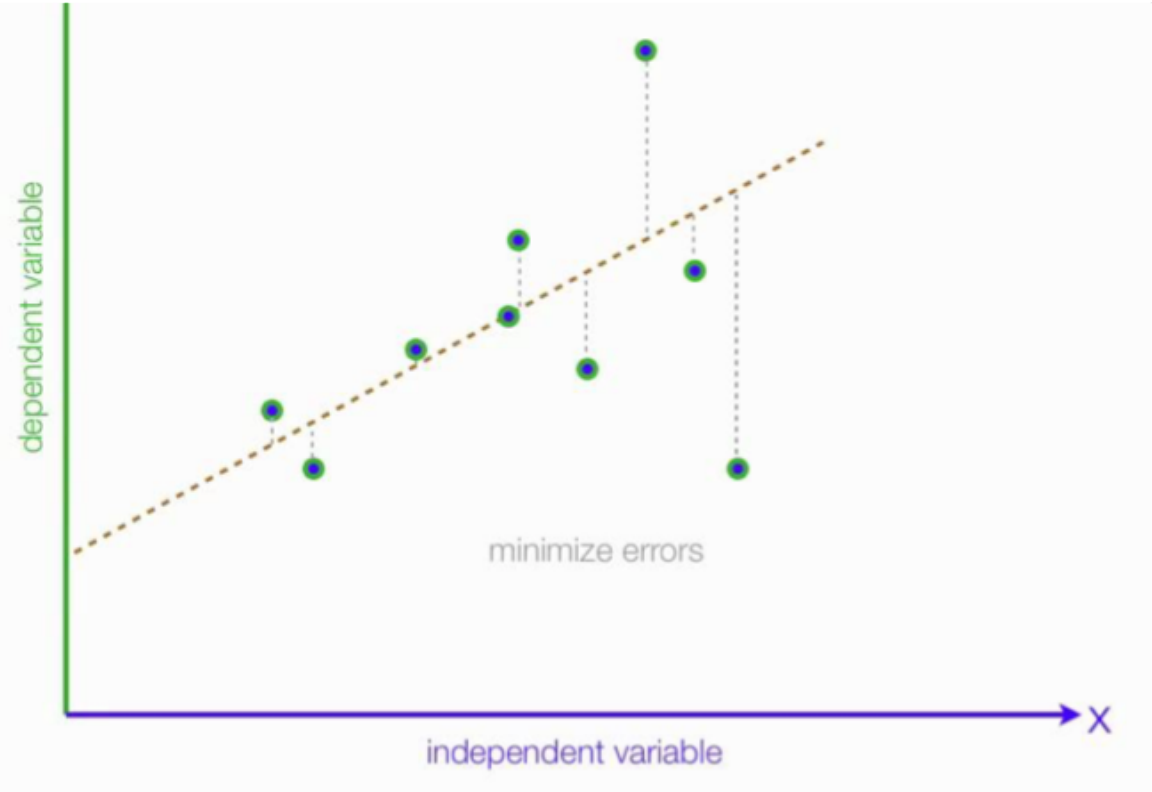
# Data Mining Tasks

- **Supervised learning-Numerical Prediction**
  - Predict medical expenses

| age | sex | bmi | children | smoker | region | expenses | InsurancePlan |
|---|---|---|---|---|---|---|---|
| 18 | male | 23.2 | 0 | no | southeast | 1121.87 | Basic |
| 18 | male | 30.1 | 0 | no | southeast | 1131.51 | Basic |
| 18 | male | 33.3 | 0 | no | southeast | 1135.94 | Basic |
| 18 | male | 33.7 | 0 | no | southeast | 1136.4 | Basic |

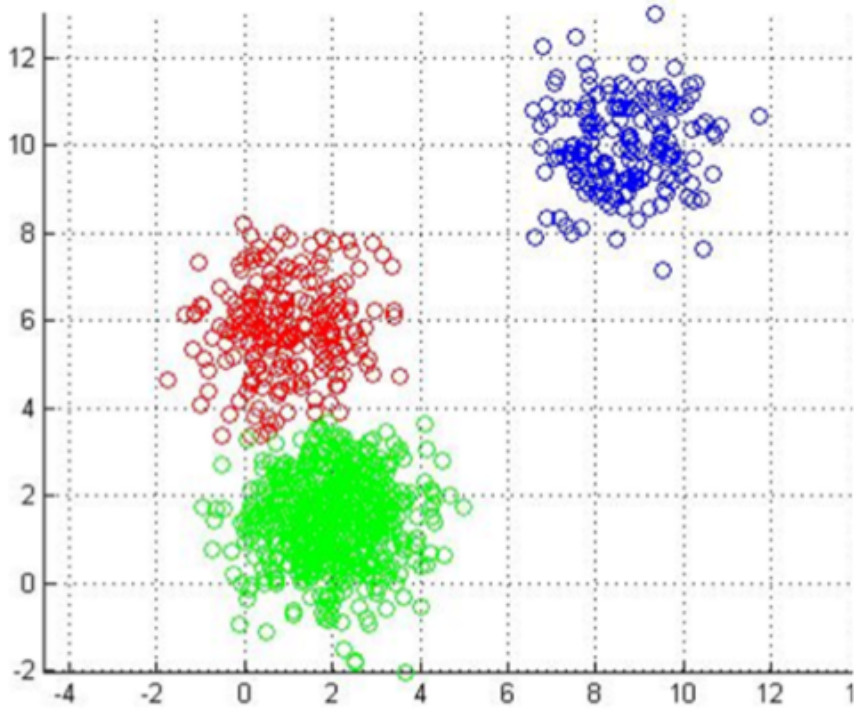
# Data Mining Tasks

- **Supervised learning-Numerical Prediction**
  - If a company's sales have increased steadily every month for the past few years, by conducting a linear analysis on the sales data with monthly sales, the company could forecast sales in future months.



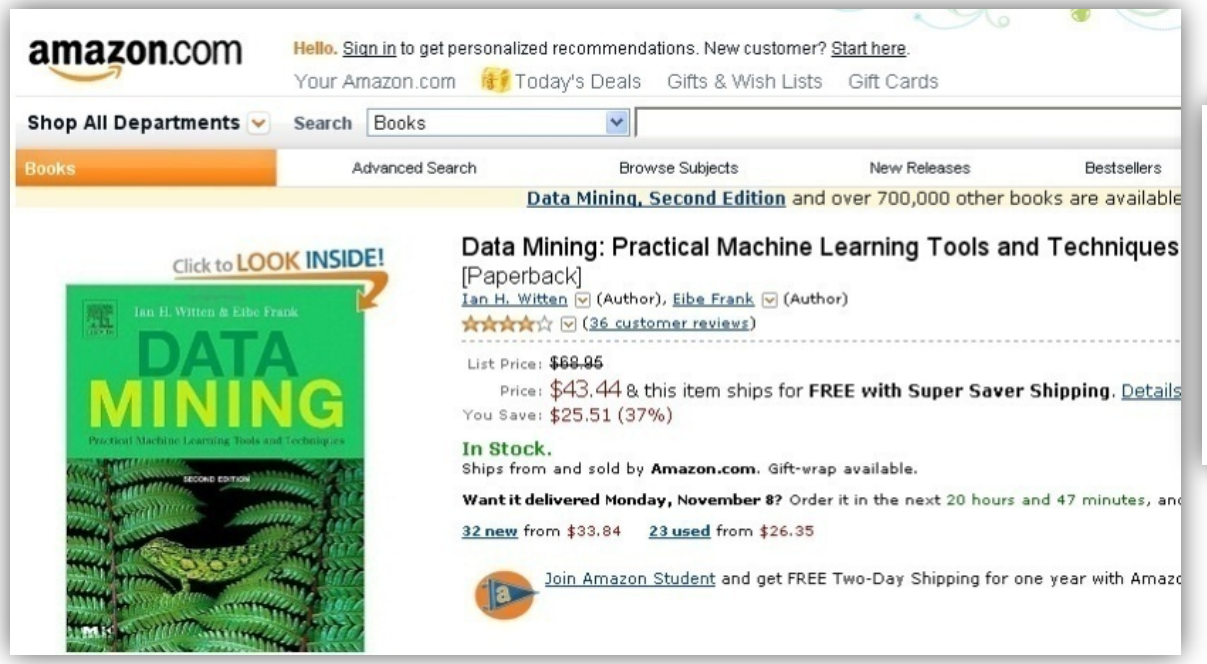
# Data Mining Tasks

- **Unsupervised learning-Clustering**
  - Customer segmentation
    - Retailers want to segment customers based on their spend patterns and understand price sensitivity of the customers. Some of the segmentation variables considered are – total spend, value of discounts, number of items bought on discounts.



# Data Mining Tasks

- **Unsupervised learning-Association rule mining**





Promotions to
retain customers
& increase sales

# Data Mining Tasks

- **Unsupervised learning-Association rule mining**
  - A classic case: Diaper and Beer

# More Data Mining Applications


Predicting the rise & fall of stocks

Weather Forecasting

Gene data

Medical Diagnosis

# More Data Mining Applications

- Target Marketing
- Credit Scoring
- Sales Forecasting
- Promotion Analysis
- Distribution Channel Analysis
- Customer Profiling
- Customer Profitability Analysis
- Cross Sell/Up Sell
- Help Desk Problem Resolution
- Customer Service Automation
- Network Forecasting
- Tariff Modeling
- E-government

- Fraud detection
- Security Management
- Product/Product Line Profitability
- Merchandise Planning
- Resource Management
- Operations Management
- Capacity Management
- Store/Branch Performance Analysis
- Store/Branch Site Selection
- Diagnosis decision support
- Gene and protein analysis
- Homeland Security

# Data Mining Trends

- Due to rapid growth of data, ***big data*** processing is becoming a major component of the knowledge discovering process.
  - Cloud computing technology has been emerged to manage large data sets efficiently.
  - Data analytical engines, such as Apache Spark, enable enterprises to mine their business data in the cloud.

# Data Mining Trends

- Machines That Learn More ***Effectively***
  - This will lead to developments in how algorithms are treated, such as AI deployments that can recognize, alter, and improve upon their own internal architecture with minimal human supervision.

# Data Mining Trends

- Deep Learning
  - On March 2016, Google DeepMind's deep learning program AlphaGo beat Lee Sedol, the Go World Champion
  - Recent advancements in automatic machine translation, object classification in photographs, image caption generation, and more.

# Data Mining Trends

- Deep Learning
  - Generative adversarial network (GAN) — a class of neural network — to generate some extremely realistic faces.

# Data Mining Trends

- Deep Learning
  - Generative adversarial network (GAN) — a class of neural network — to generate some extremely realistic faces.

# Data Mining Trends

- Deep Learning
  - Generative adversarial network (GAN) — a class of neural network — to generate some extremely realistic faces.