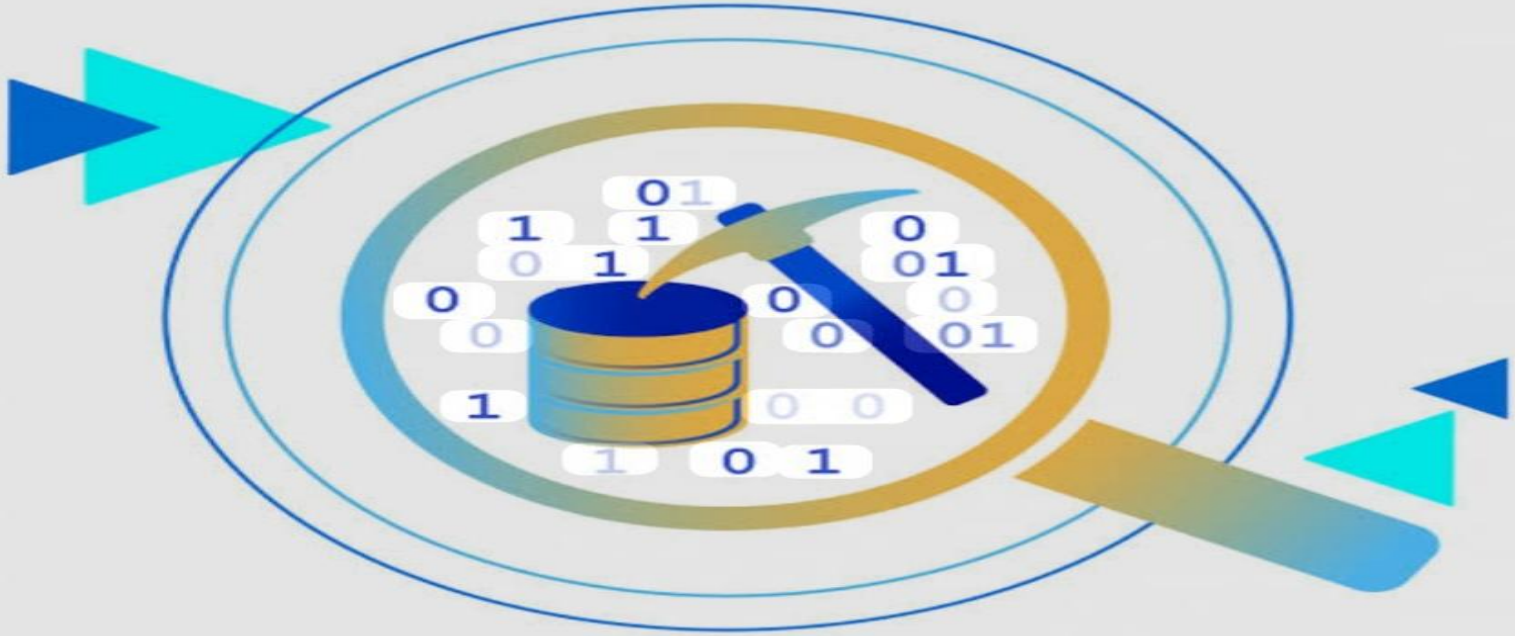
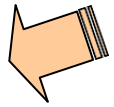
Data Science – Fall 2025/2026

INTRODUCTION TO DATA MINING

Lecture 1 - Introduction

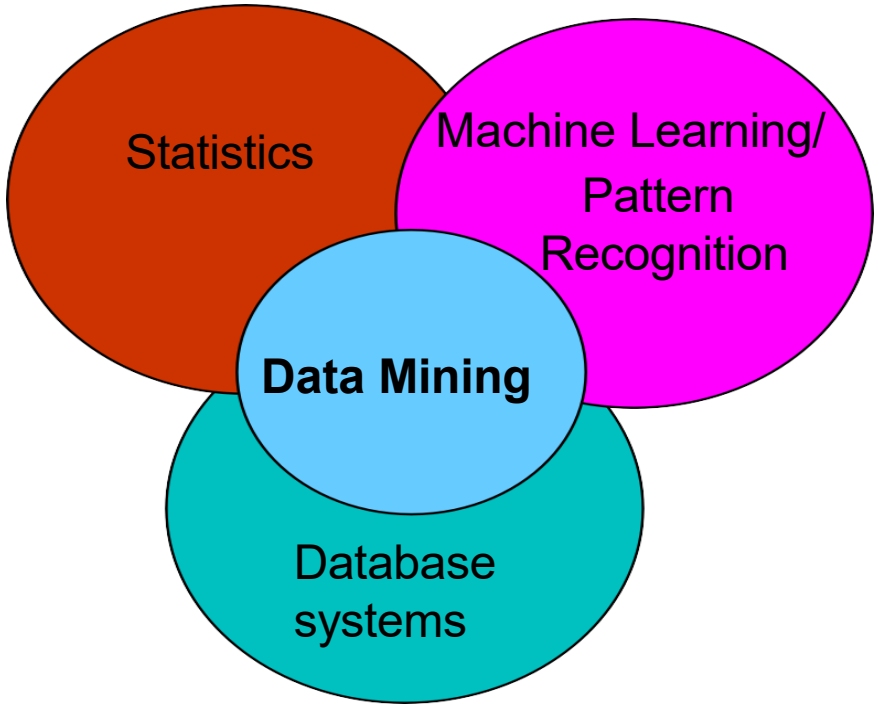


- **What is Data Mining ?**
- **What is Machine Learning ?**
- **KDD Process**
- **Database**
- **Basic Statistical Description of Data**
- **Data Mining Resources**

- **What is Data Mining ?** 
- **What is Machine Learning ?**
- **KDD Process**
- **Database**
- **Basic Statistical Description of Data**
- **Data Mining Resources**

What is DM ?

- **Data Mining** (DM), also known as **Knowledge Discovery in Databases** (KDD), is the process of extracting meaningful patterns, trends, and insights from large datasets.
- **DM** combines techniques from fields like machine learning, statistics, and database systems to enable data-driven decision-making



" The ultimate goal of Data Mining is to turn data into knowledge and knowledge into action. "

" It is estimated that to extract enough gold to make a single gold ring, you'd need to sort through around 26 tons of rock and other stuff. "

What is DM ?: Data Deluge

- Explosive Growth of Data: from terabytes (10^{12} bytes) to petabytes (10^{15} bytes) or even brontobytes (10^{27}bytes)
- Data collection and data availability: Automated data collection tools, database systems, Web, Computerized Society ...

http://www.internetlivestats.com/one-second/

" Every second, there is:

7,998 Tweets sent

839 Instagram photos uploaded

1,364 Tumblr posts

3,083 Skype calls

55,560GB of Internet traffic

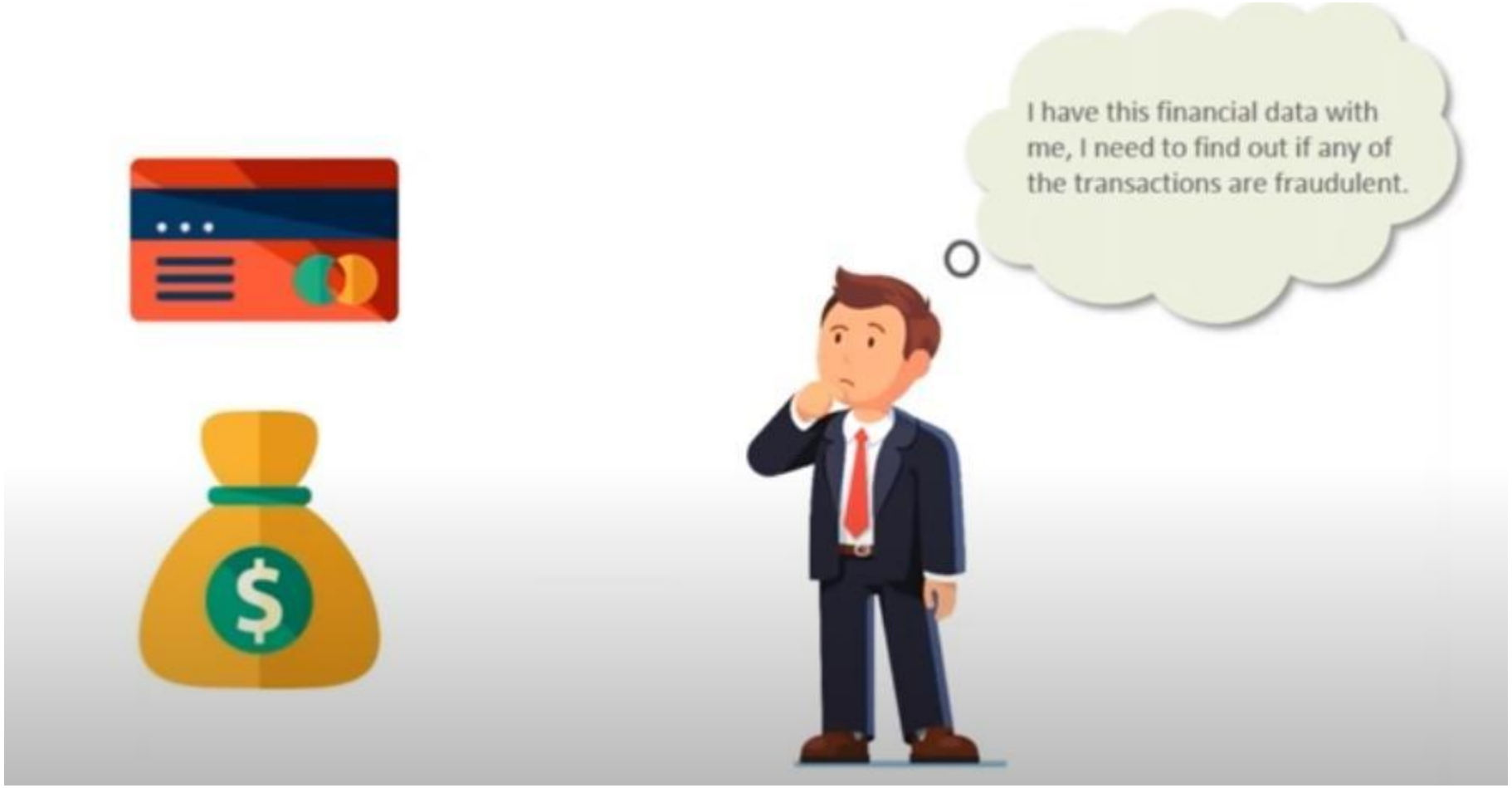
66,335 Google searches

73,391 YouTube videos viewed

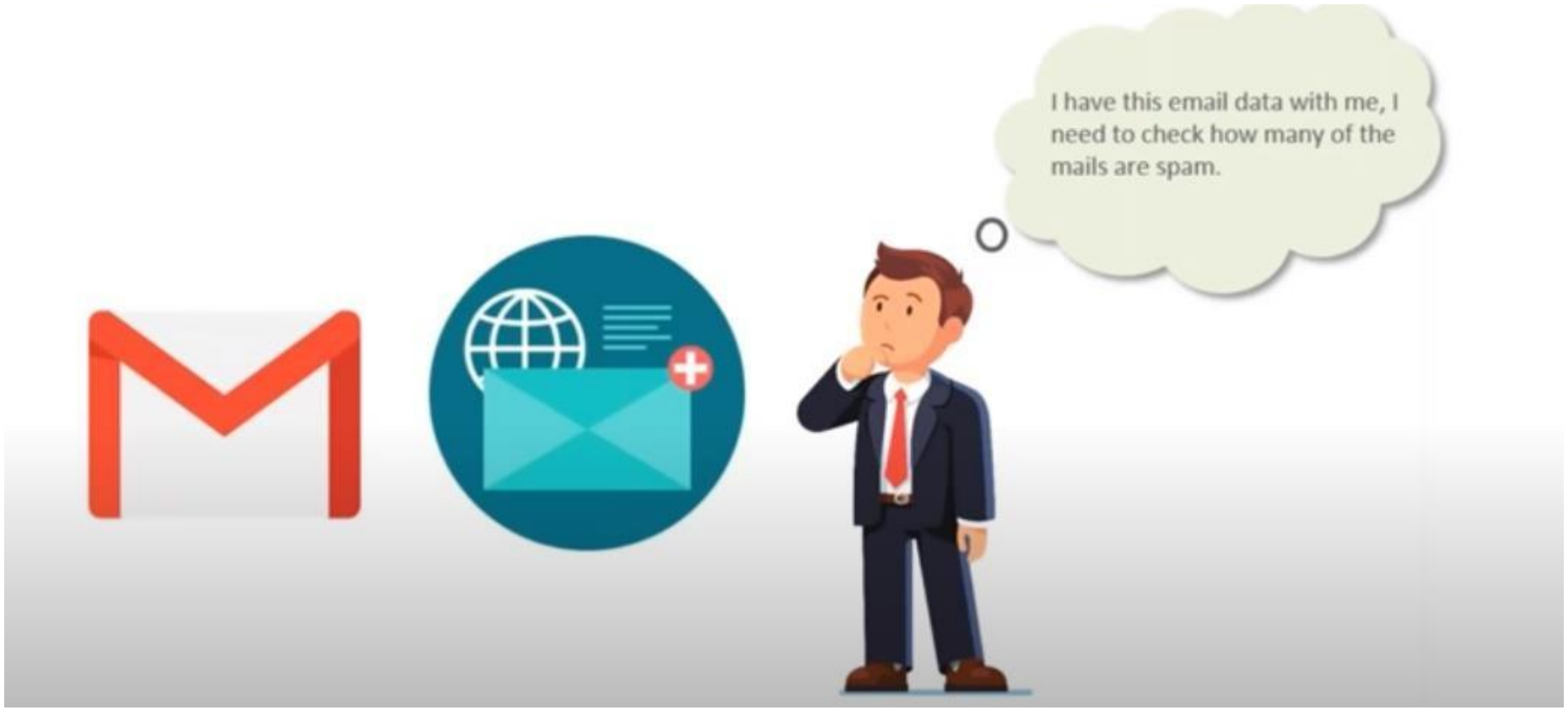
2,681,874 Emails sent "

" We are drowning in Data and straving for knowledge !!! "

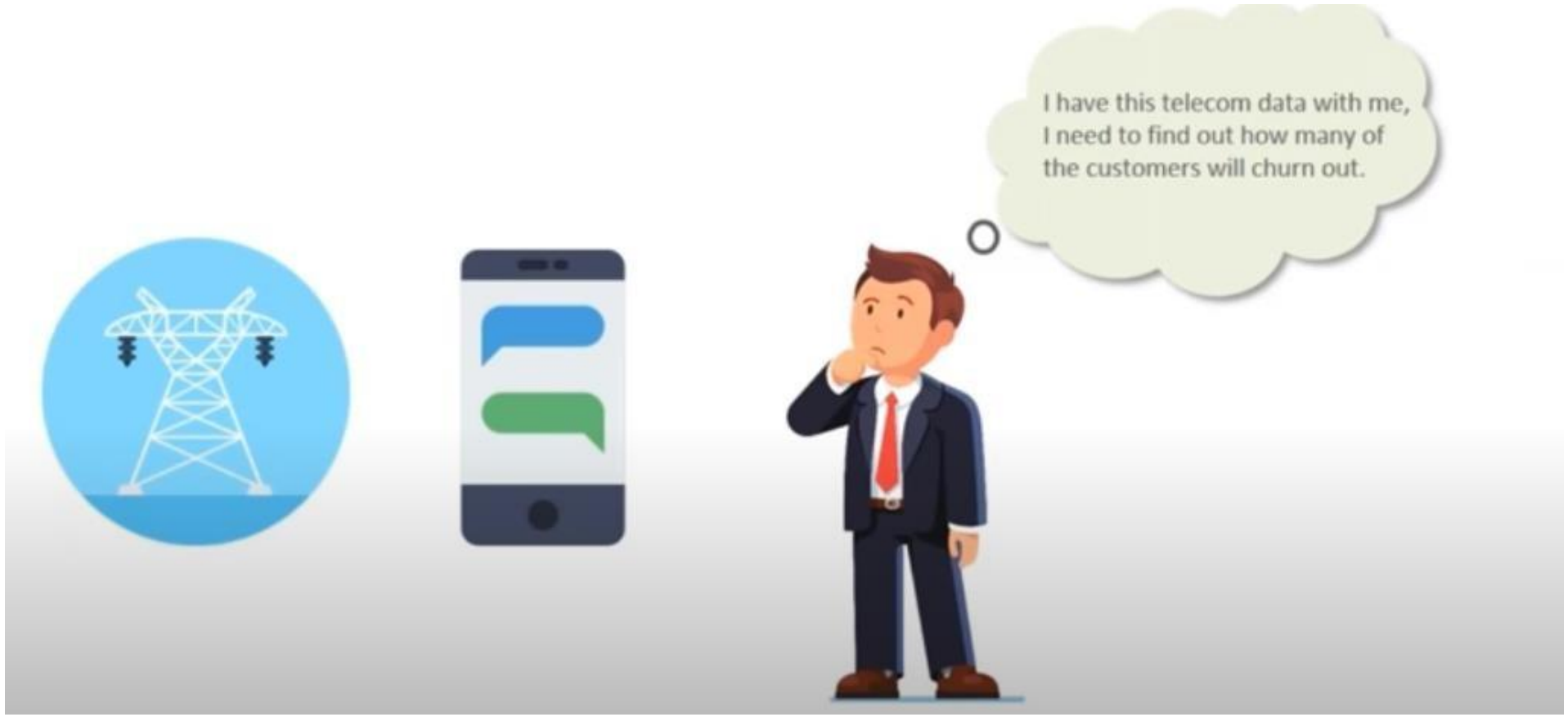
What is DM ? : DM Scenario 1



What is DM ? : DM Scenario 2



What is DM ? : DM Scenario 3



What is DM ? A RESCUE

What is DM ? Where DM is Used

- **Business & Marketing**
  - *Goal :* Understand customer behavior and improve sales.
  - *Example:* Market basket analysis (e.g., “People who buy bread often buy butter”

- **Finance & Banking**
  - *Goal :* Detect fraud, evaluate risk, and manage investments.
  - *Example:* Credit card fraud detection ; Predicting loan defaults.

- **Education**
  - *Goal :* Improve diagnosis and healthcare.
  - *Example:* Predicting diseases in healthcare ; finding patterns in patient records

- **Telecommunications**
  - *Goal :* Improve service and reduce churn (loss of customers)
  - *Example:* Identifying customers likely to switch providers; Detecting network problems early

What is DM ? Where DM is Used

black and white image colorization

Zhang, Isola, Efros.
Colorful Image Colorization.
In ECCV, 2016.

http://richzhang.github.io/colorization/

See also https://machinelearningmastery.com/inspirational-applications-deep-learning/

What is DM ? Where DM is Used

Image recognition using deep neural networks

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"girl in pink dress is jumping in air."

"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

Andrej Karpathy & Li Fei-Fei
"Deep Visual-Semantic Alignments for Generating Image Descriptions"
CVPR 2015

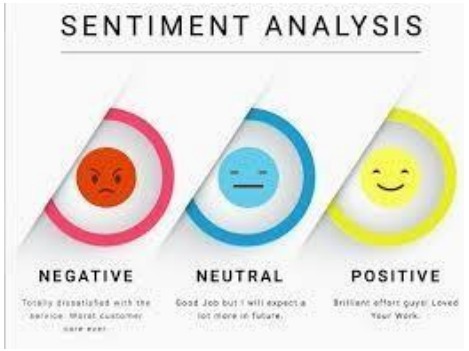
https://cs.stanford.edu/people/karpathy/deepimagesent/

What is DM ? Where DM is Used

Email spam filtering

Blanzieri, E. & A. Bryl. "A survey of learning-based techniques of email spam filtering" Artificial Intelligence Review March 2008, Vol. 29, Issue 1, pp 63–92

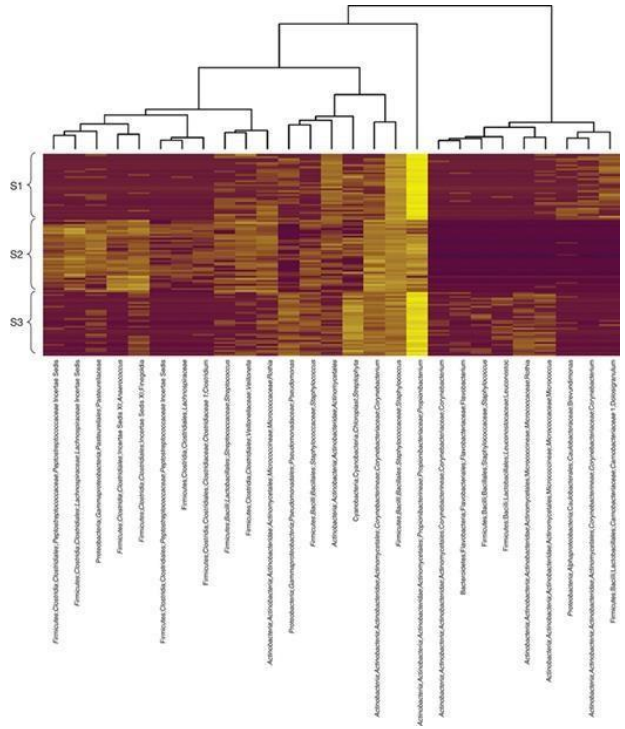
Document sentiment analysis



Liu B., Zhang L. "A Survey of Opinion Mining and Sentiment Analysis." In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA. 2012

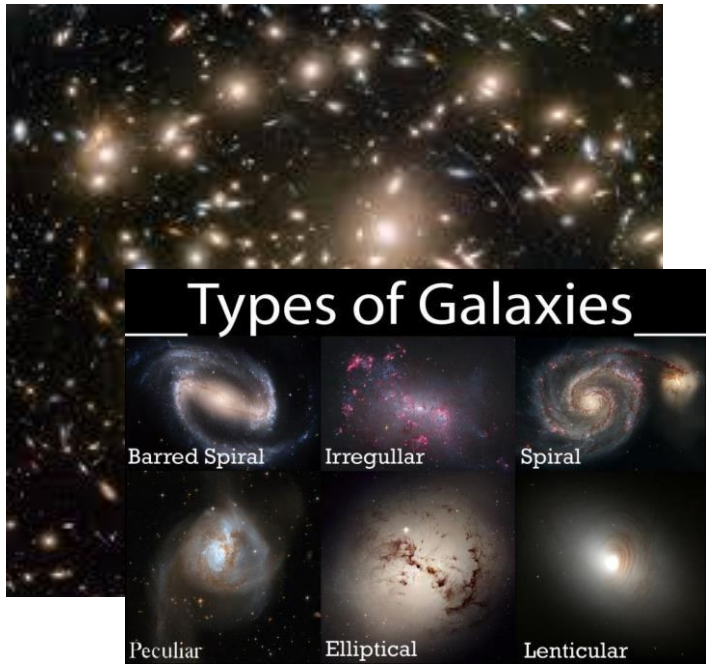
What is DM ? Where DM is Used

Identifying important groups of microorganisms in the human body



Dan Knights Elizabeth K. Costello Rob Knight
"Supervised classification of human microbiota"
FEMS Microbiology Reviews, Volume 35,
Issue 2, 1 March 2011, Pages 343–359

Classifying galaxies in the universe



Fowler, L., Schawinski, K., & Brandt, B.-E.
Galaxy Classification using Machine Learning.
Paper presented at the American Astronomical Society Meeting Abstracts. 2017

What is DM ?: Where DM is Used

Image and video processing

https://www.classaction.org/blog/facebook-sued-over-face-recognition-feature

recommender systems

NETFLIX

YouTube

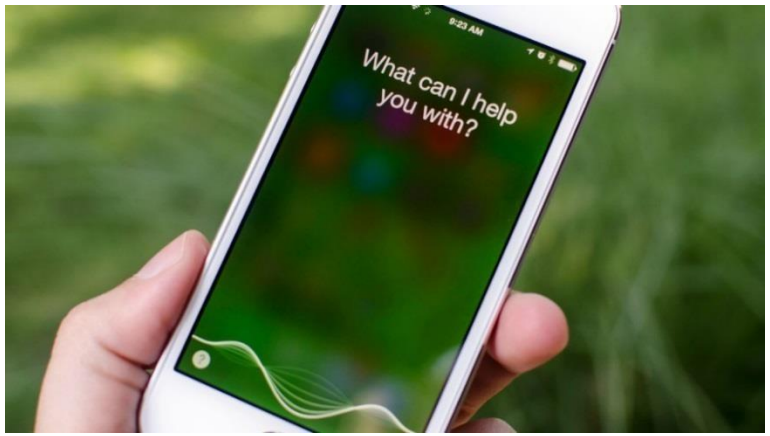
PANDORA

amazon

Audio and voice processing

Personal assistants

alexa

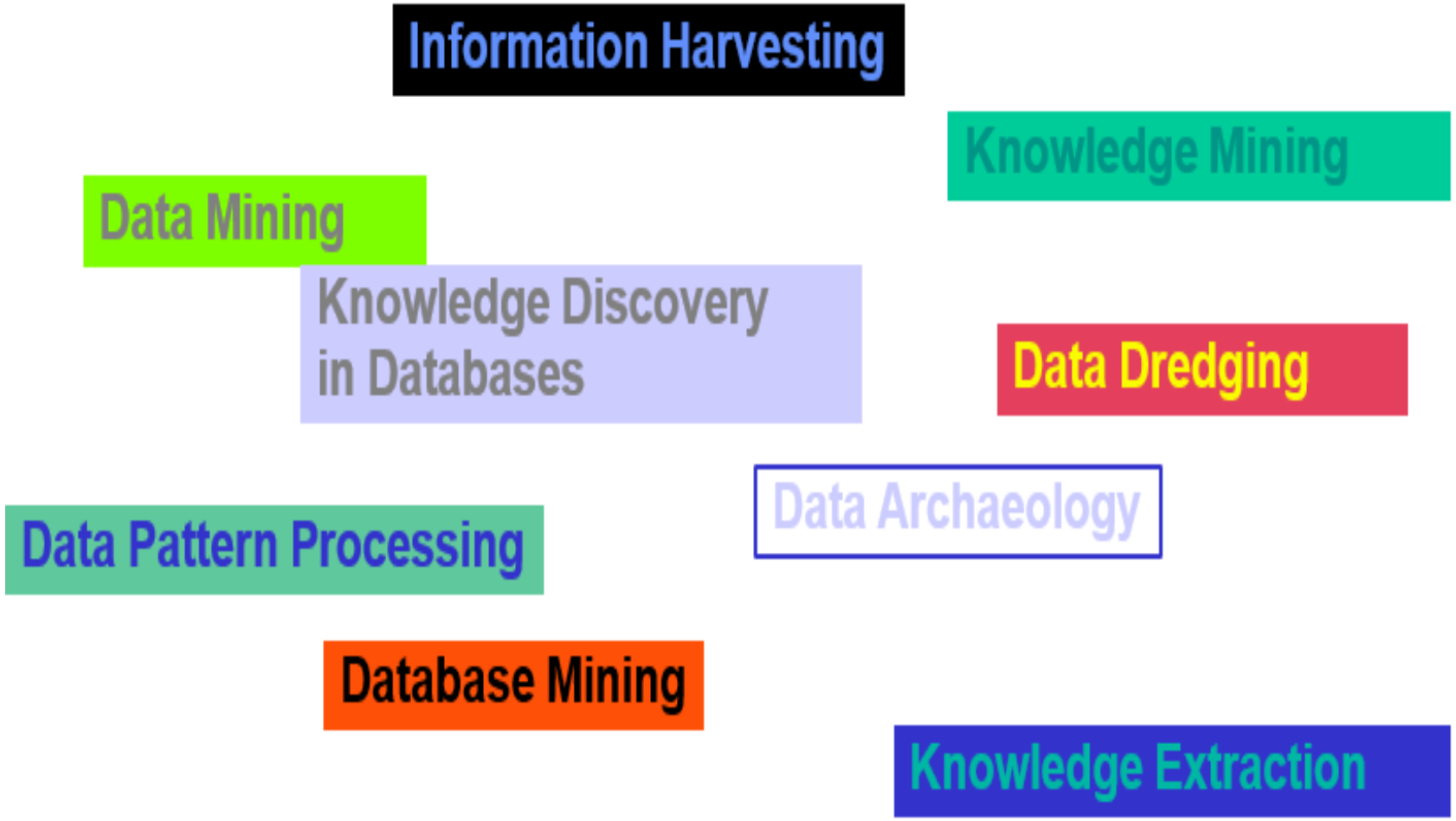


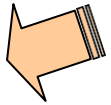
Bgr.com/tag/siri

What is DM ? : Evolution of DM

Evolutionary Step	Business Question	Enabling Technologies
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases

What is DM ? : Synonyms of DM



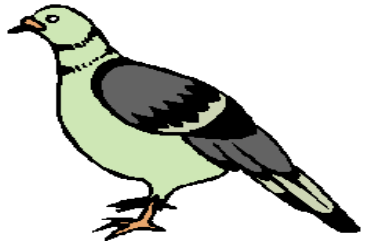
- **What is Data Mining ?**
- **What is Machine Learning ?** 
- **KDD Process**
- **Database**
- **Basic Statistical Description of Data**
- **Data Mining Resources**

What is Machine Learning ?

- **Machine Learning** (ML) is a field of study within AI (*Artificial Intelligence)* concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions (auto-learns).

- ML finds application in many fields, including NLP (*natural language processing*), image recognition, email filtering, agriculture, medecine and other applications

- *Pattern recognition* in machine learning is the automated discovery of regularities and structures within data, allowing a system to categorize new observations or make predictions based on what it has learned from past examples.



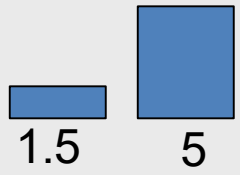
We will return to the actual topic in two minutes. In the meantime, we are going to play a quick game.

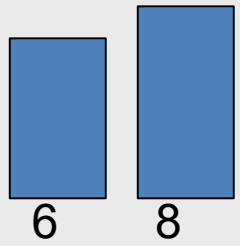
I am going to show you some problems which were shown to pigeons!

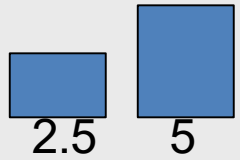
Let's see if you are as smart as a pigeon!

Pigeon Problem 1

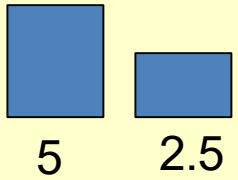
Examples of class A

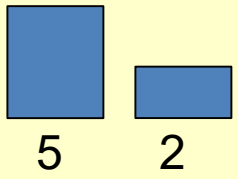


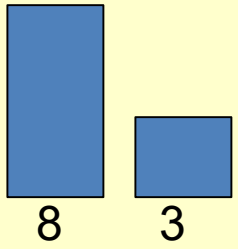


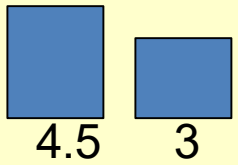


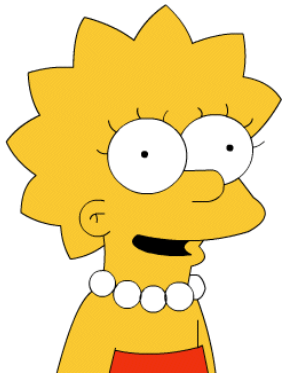
Examples of class B



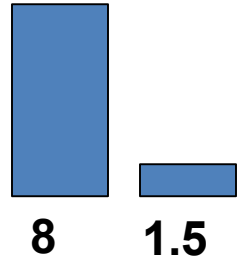




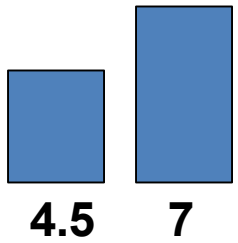




What class is this object?

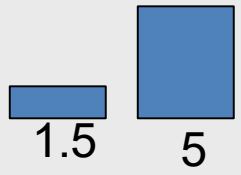


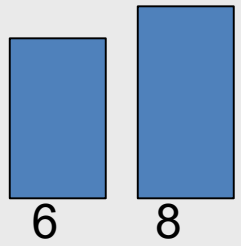
What about this one, A or B?

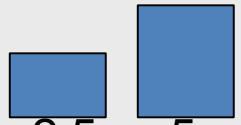


Pigeon Problem 1

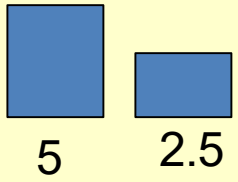
Examples of class A

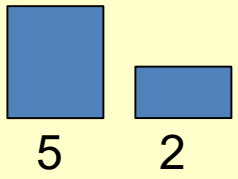


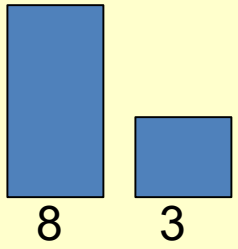


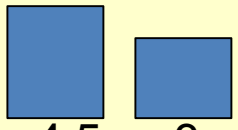


Examples of class B

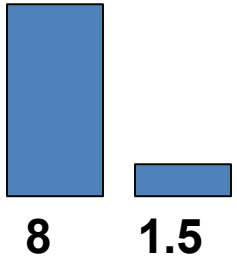








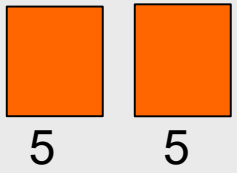
This is a B!

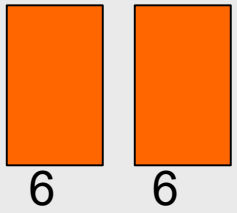


Here is the rule.
If the left bar is smaller than the right bar, it is an A, otherwise it is a B.

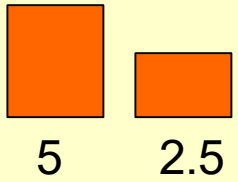
Pigeon Problem 2

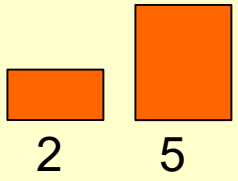
Examples of class A

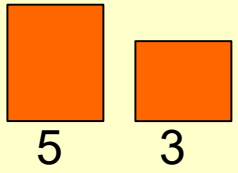


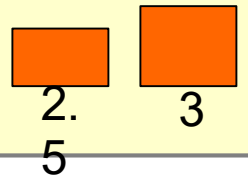


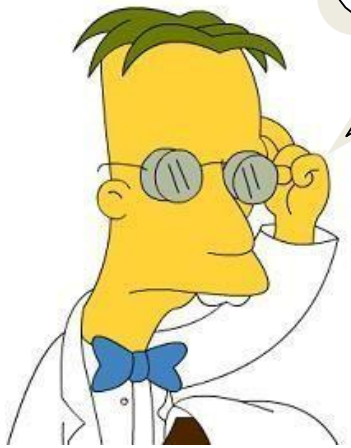
Examples of class B

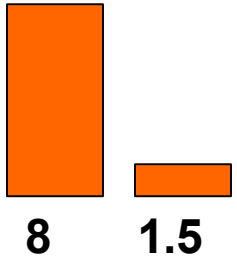


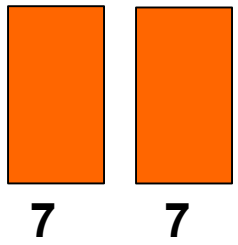






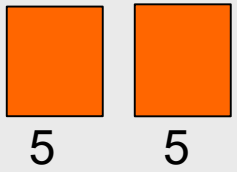


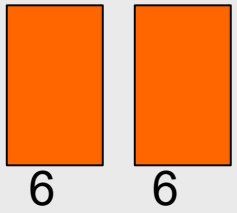




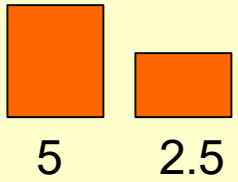
Pigeon Problem 2

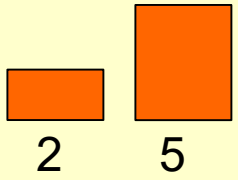
Examples of class A

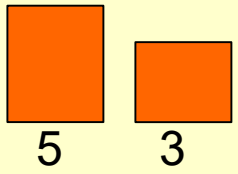


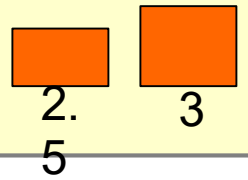


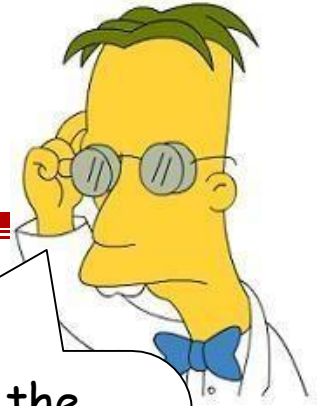
Examples of class B





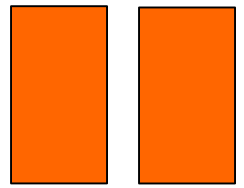






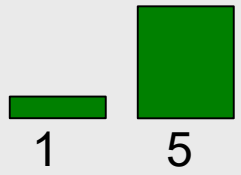
The rule is as follows, if the two bars are equal sizes, it is an A. Otherwise it is a B.

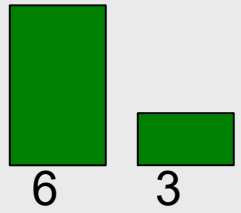
So this one is an A.

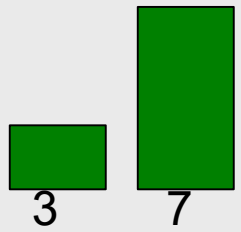


Pigeon Problem 3

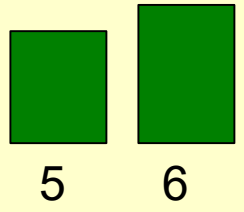
Examples of class A

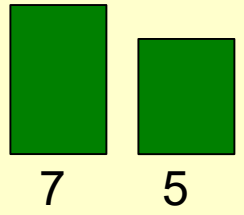


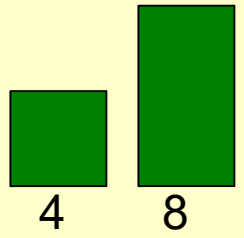


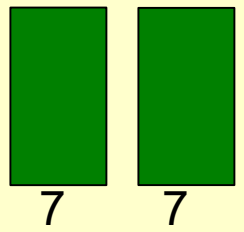


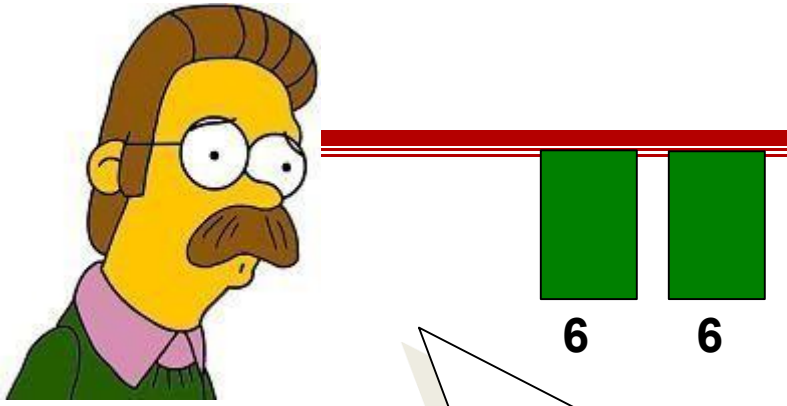
Examples of class B





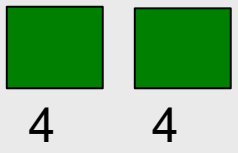


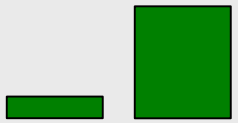


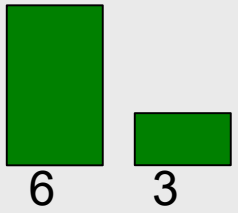


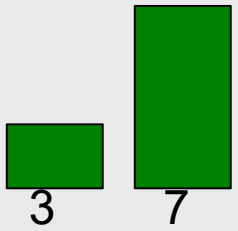
Pigeon Problem 3

Examples of class A

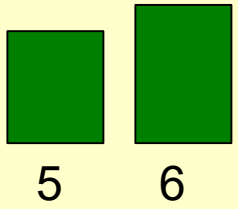


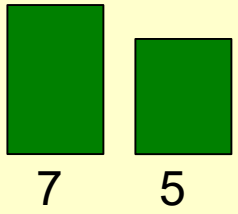


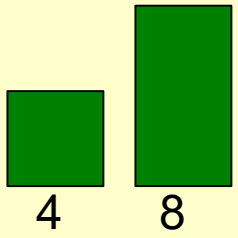


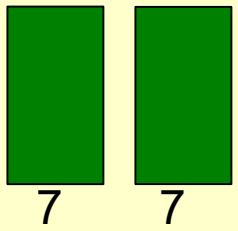


Examples of class B

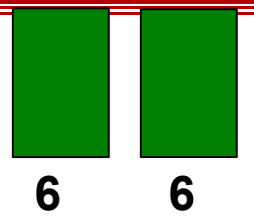








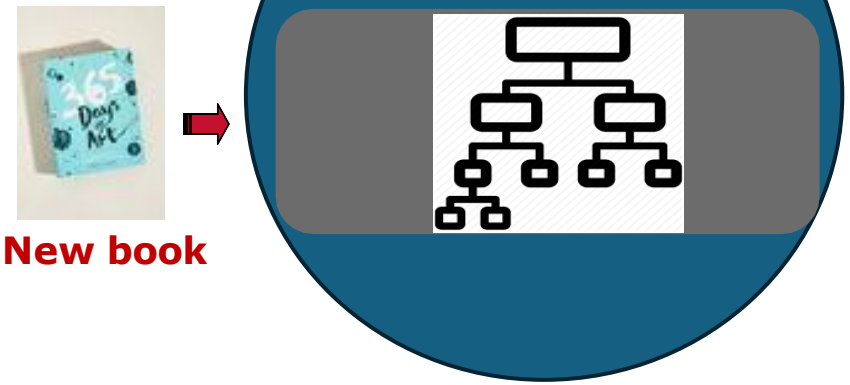
It is a B!



The rule is as follows, if the sum of the two bars is less than or equal to 10, it is an A. Otherwise it is a B.

ML: Classification

- Classification is a **supervised** learning process where labeled data is used to assign new data to predefine categories or classes



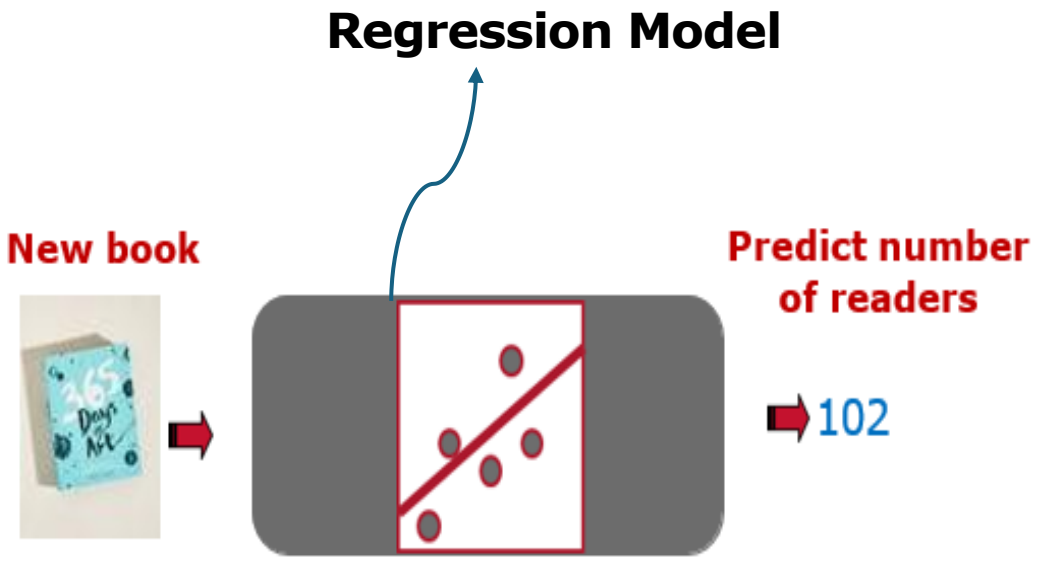
Classification Model

art
Predicted category

Data: Large collection of books. For each book: **title, info, full text and a category**

ML: Regression

- Regression is a **supervised** learning process where a model predicts a continuous numerical value based on input data

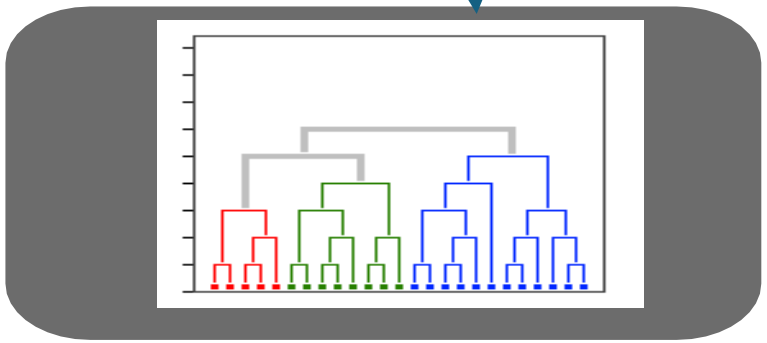


Data: Large collection of books. For each book: **title, info, full text and number of users that accessed the books in the past 12 months**

ML: Clustering

- Clustering is an **unsupervised** learning process that groups data points into clusters based on their similarities, without using labeled data.

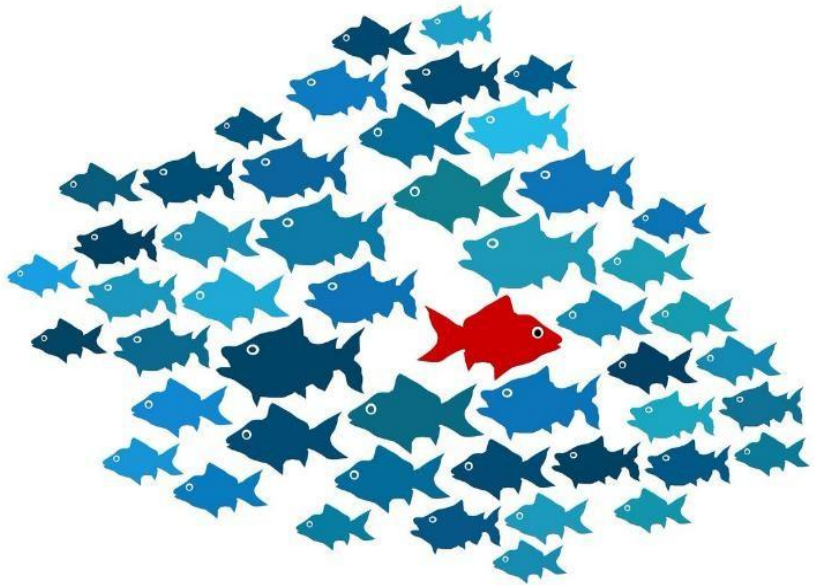
We can derive from these data a set of **clusters** that group books by similarity

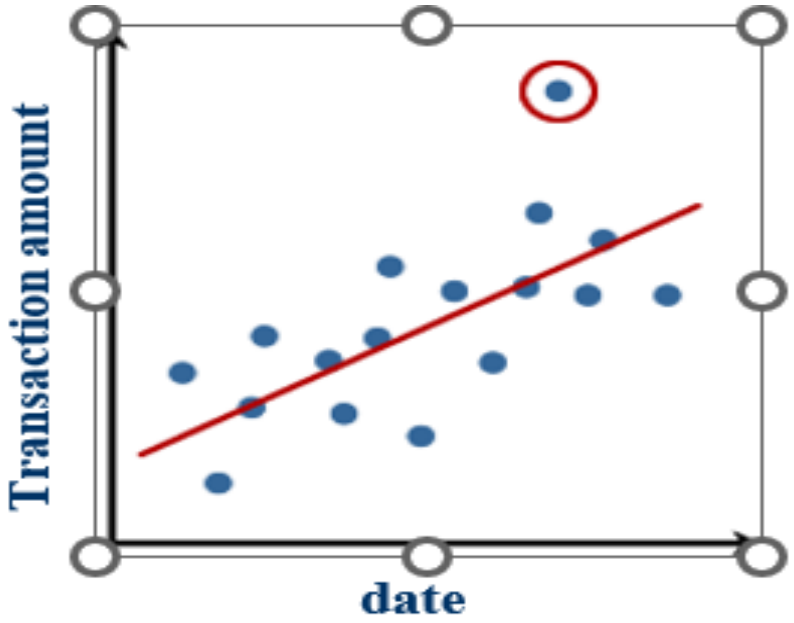


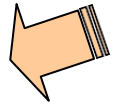
Data: Large collection of books. For each book: **title, author, info, full text …**

ML: Outlier Analysis

- Outlier analysis (or Anomaly Detection) is an **unsupervised learning** process that identifies data points that are significantly different from the rest of the dataset.

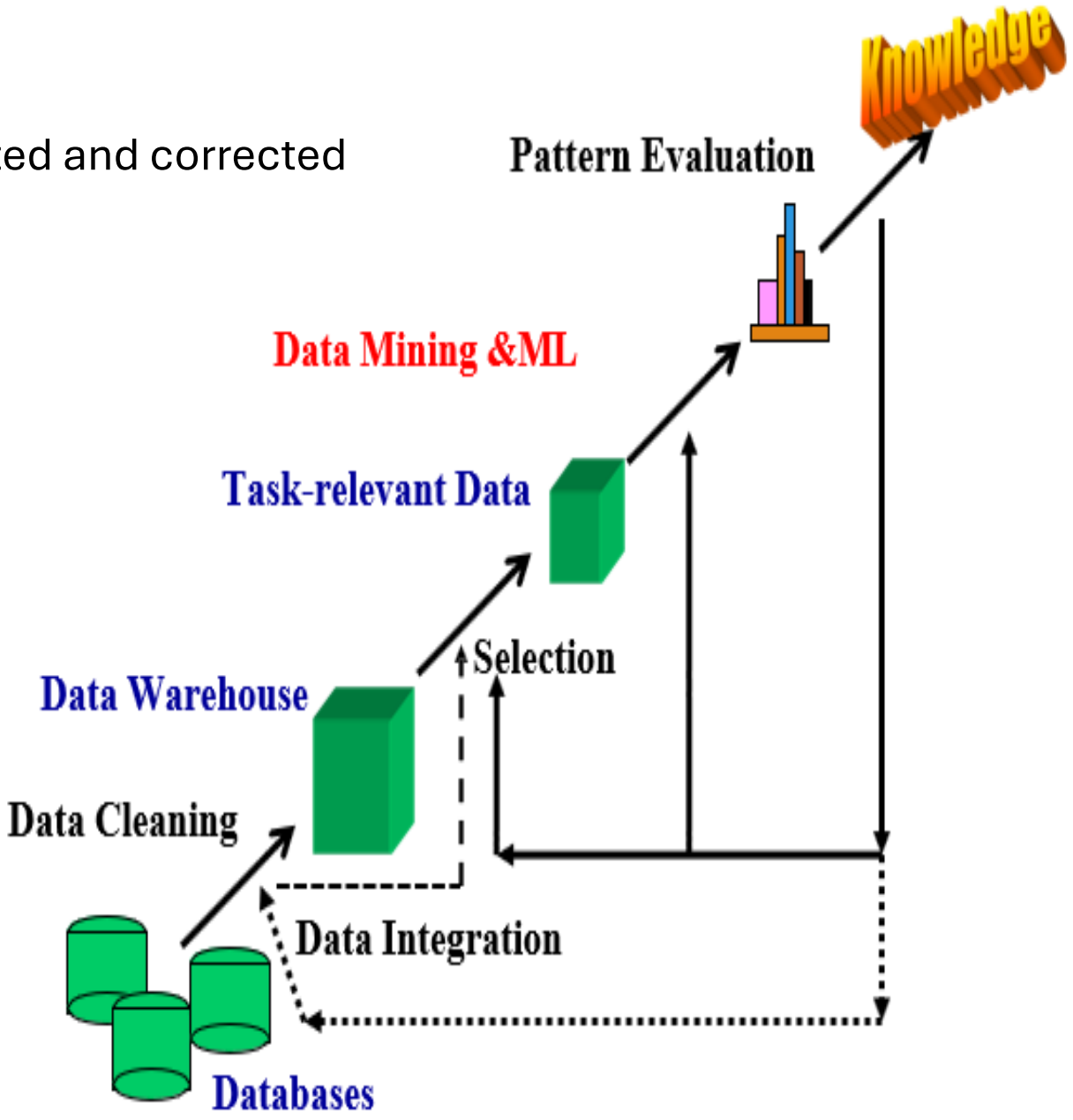




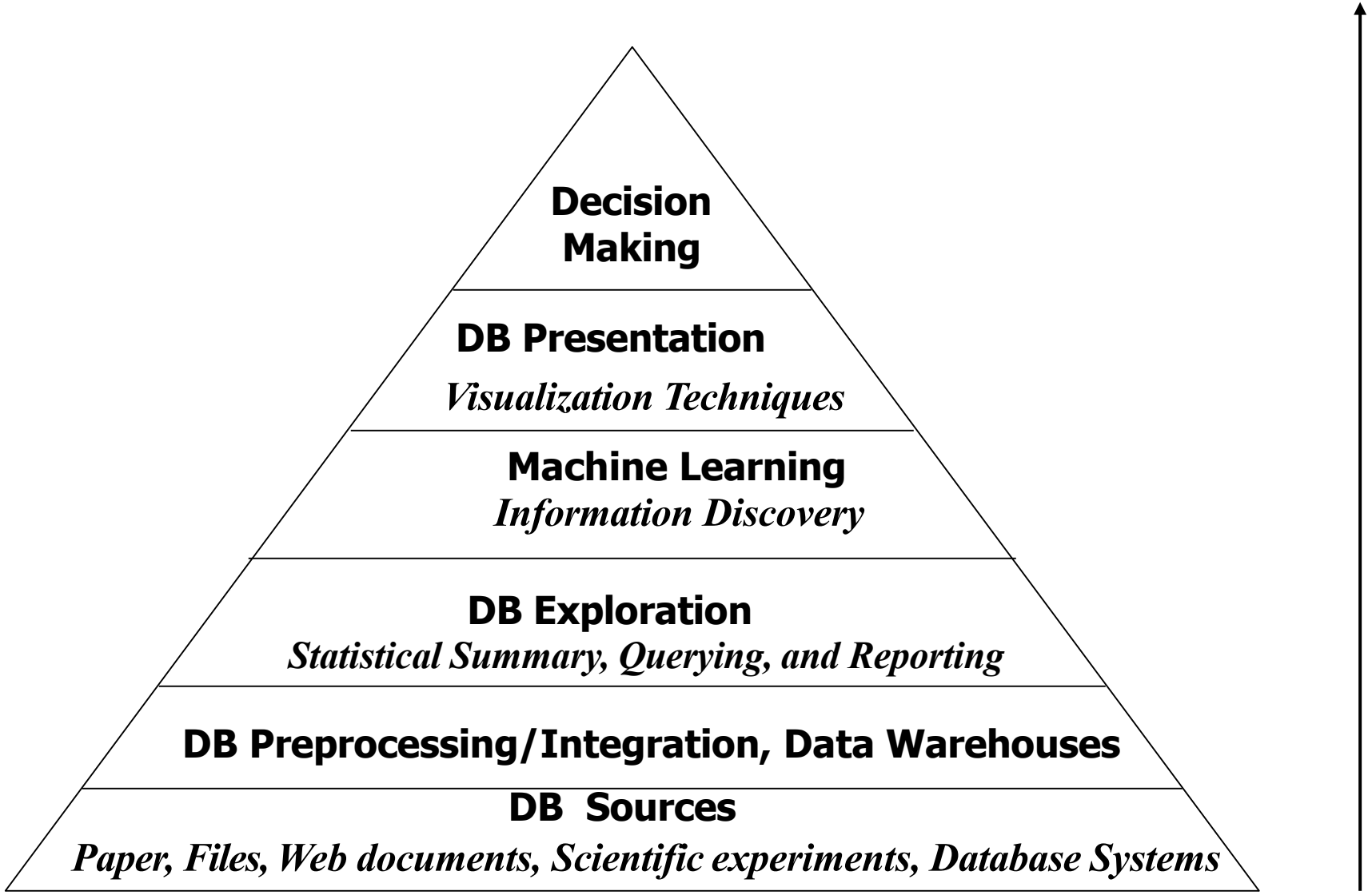
- **What is Data Mining ?**
- **What is Machine Learning ?**
- **KDD Process** 
- **Database**
- **Basic Statistical Description of Data**
- **Data Mining Resources**

KDD Process

- **Databases:** raw data provided from multiple sources
- **Data Cleaning:** errors and inconsistencies are detected and corrected
- **Data Integration:** data from different sources are combined into a single and unified view
- **Data Warehouse:** integrated data are stored in a centralized repository for analysis
- **Selection:** relevant data are retrieved from the data warehouse for specific analysis
- **DM & ML:** ML techniques are applied to discover patterns
- **Pattern Evaluation:** discovered patterns are assessed for their usefulness and validity
- **Knowledge:** valuable insights and actionable knowledge are extracted for decision-making



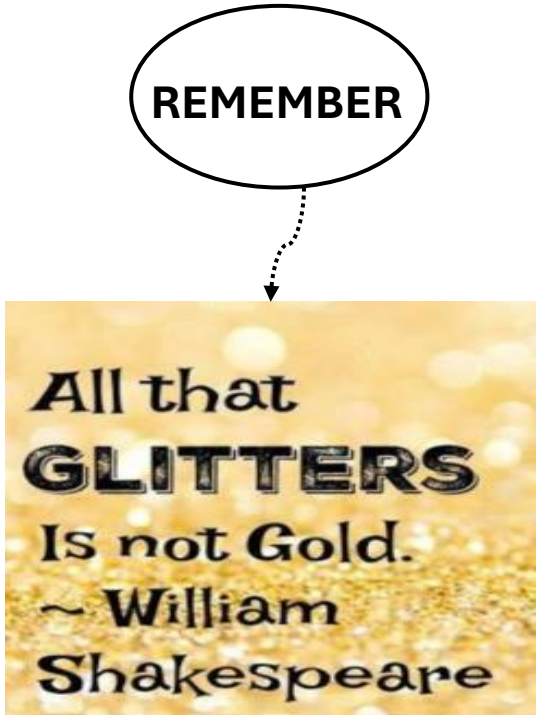
KDD Process : A hieracrhical process

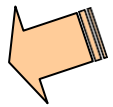


KDD Process : Evaluation of Discovered Patterns

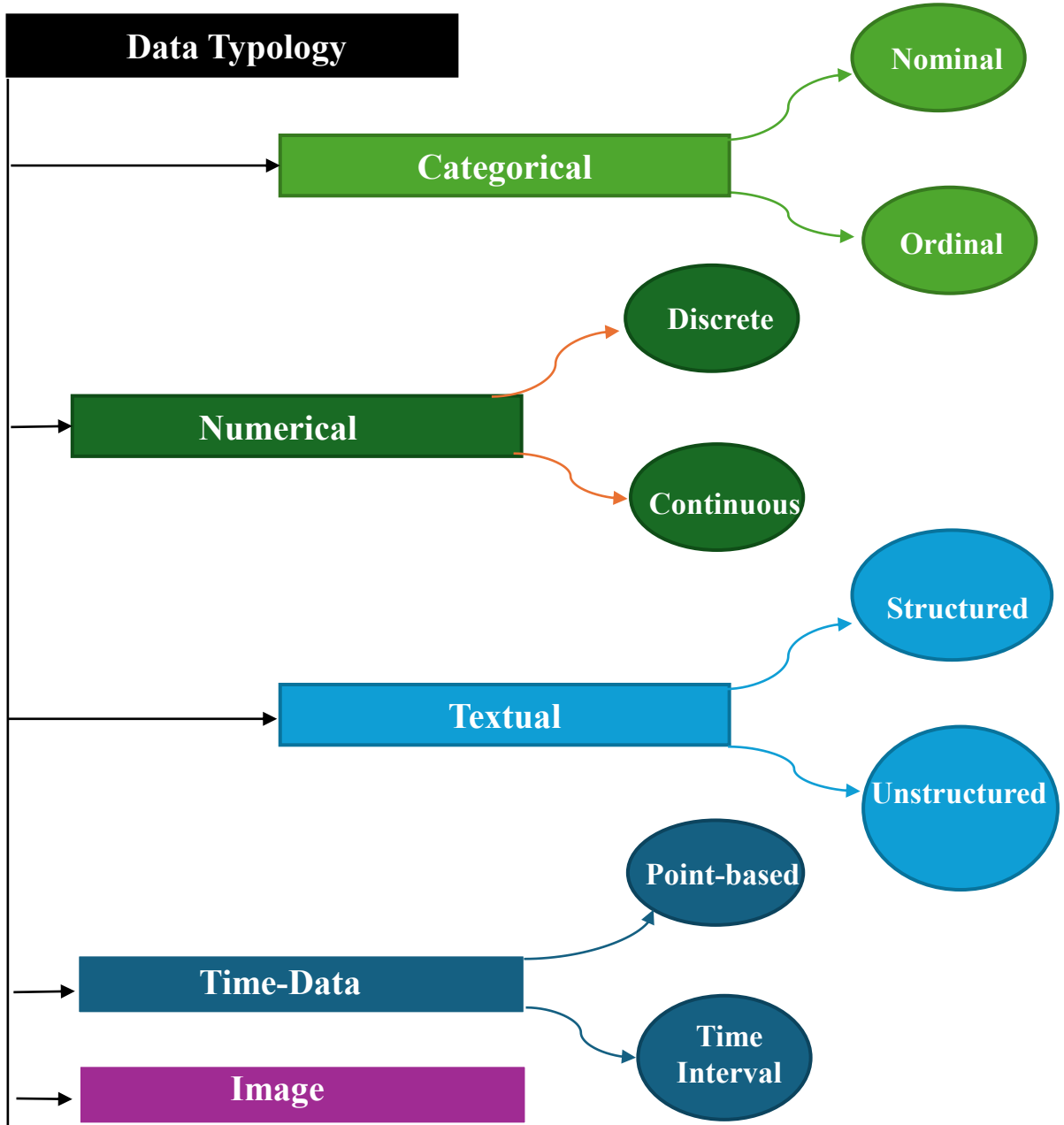
The **evaluation** step in the KDD process is crucial. It ensures that the discovered patterns are **valid**, **novel**, **useful** and **understandable** for effective knowledge extraction.

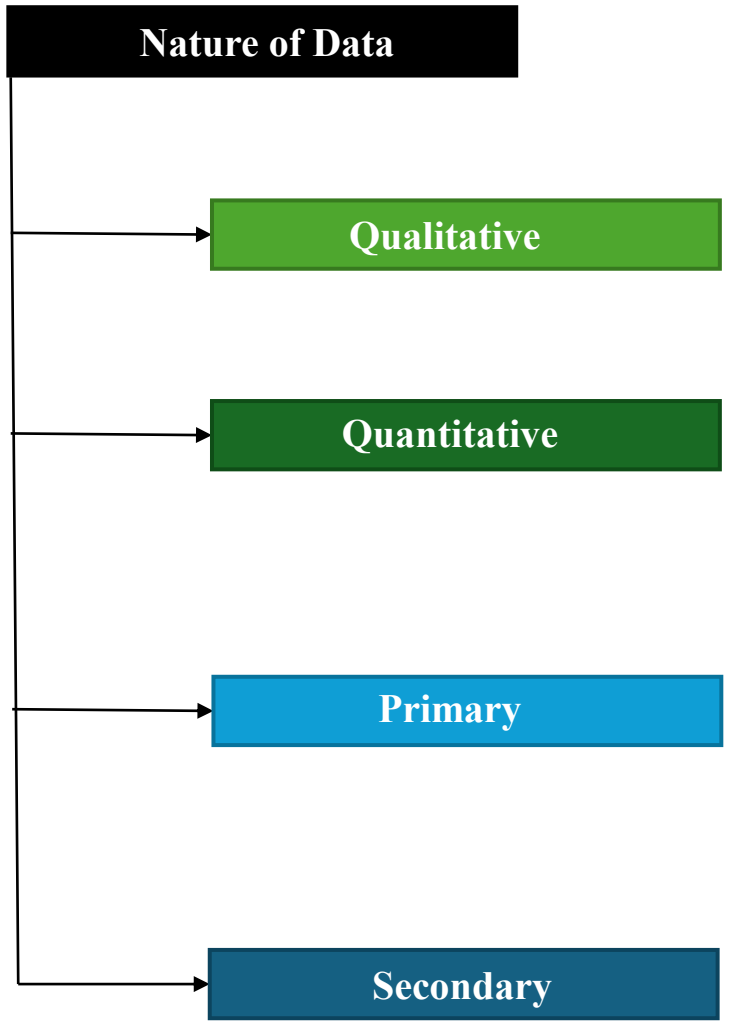
- **Valid:** The pattern accurately represents the data and holds true across different samples
- **Novel:** The pattern reveals new or unexpected information not previously known
- **Useful :** The pattern provides insights that can support decisions or actions
- **Understandable:** The pattern is simple, clear and easy for humans to interpret and apply



- **What is Data Mining ?**
- **What is Machine Learning ?**
- **KDD Process**
- **Database** 
- **Basic Statistical Description of Data**
- **Data Mining Resources**

Database : A Global vision



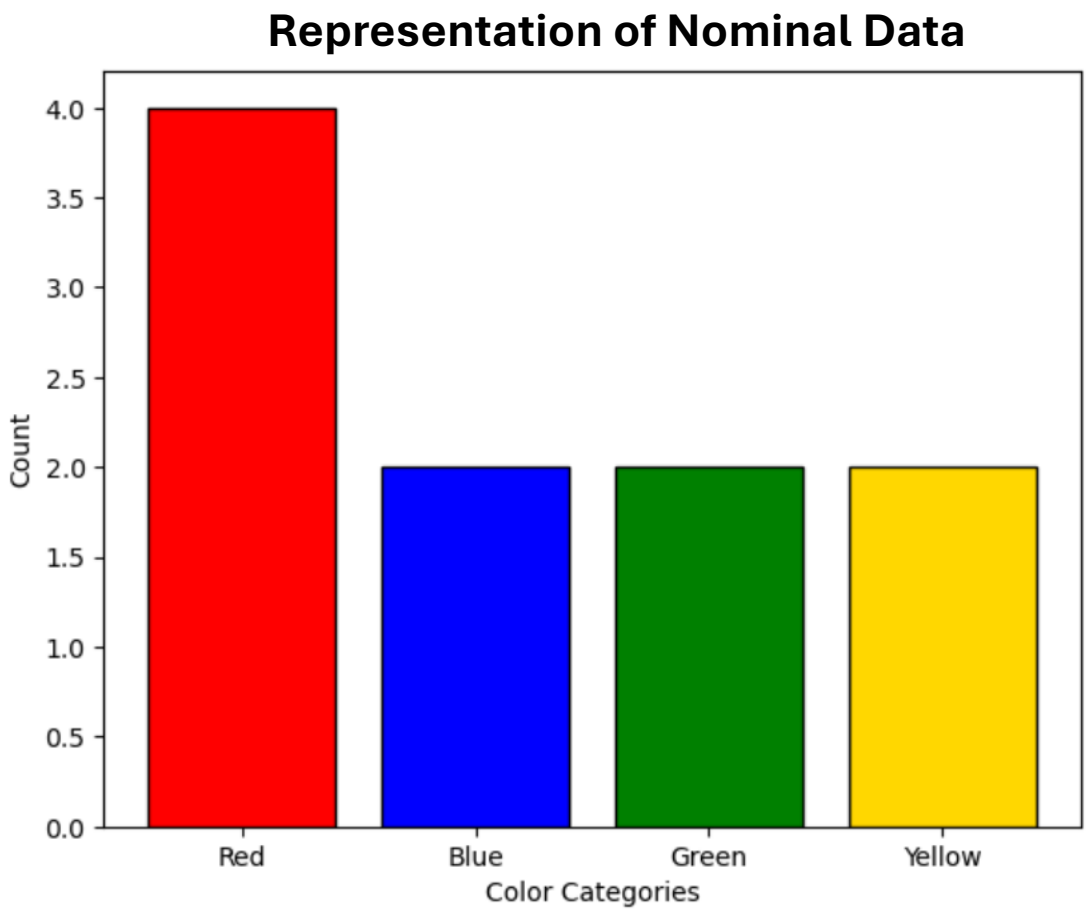


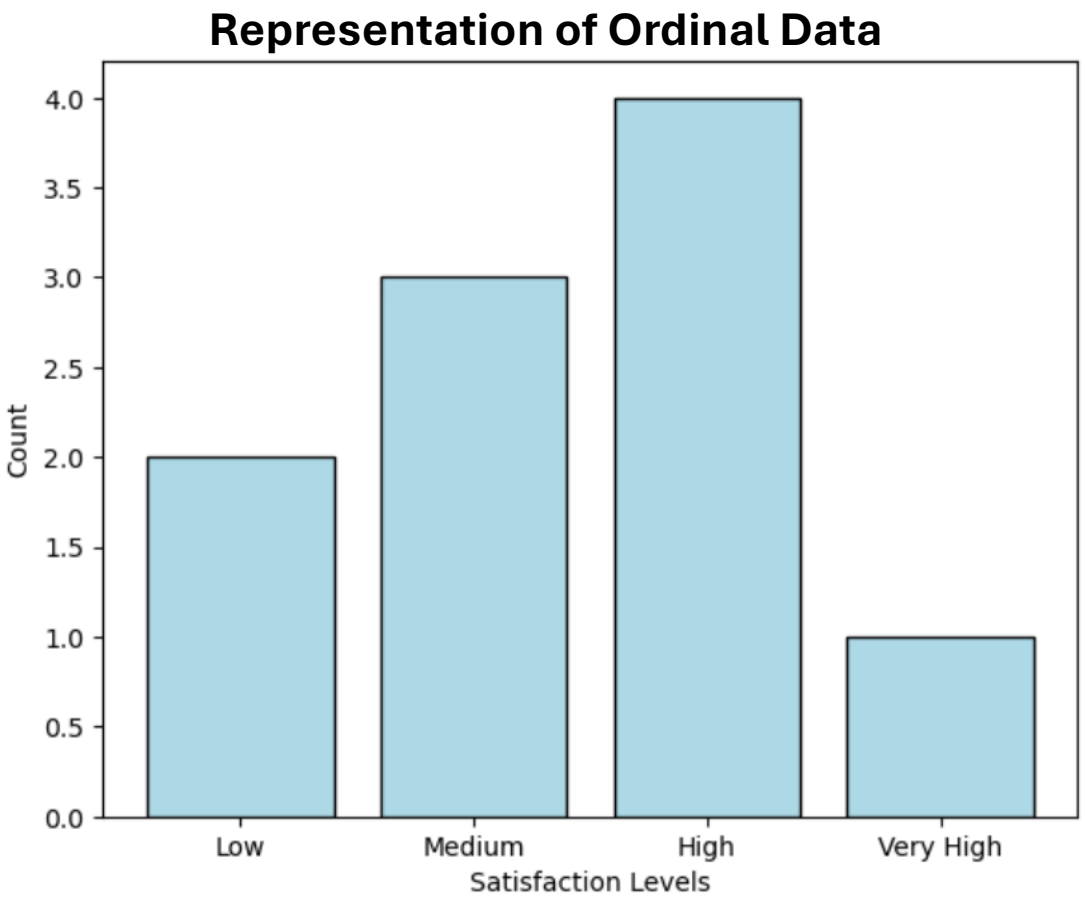
Database : Categorical Database

- Categorical data represent **groups or categories (not numbers)** that describe qualities or characteristics, such as colors, gender..
- Nominal Data : A type of categorical data where the **categories have no specific order**
- Ordered Data : A type of categorical data where the **categories follow a meaningful order or rank**

Feature	**Nominal Data**	**Ordinal Data**
Definition	Categories with **no natural order**	Categories with a **meaningful order**
Type of data	Qualitative (labels or names)	Qualitative (ordered labels)
Order or ranking	❌ No	✅ Yes
Examples	Colors (red, blue, green), gender (male, female)	Satisfaction (low, medium, high), education (high school, college, master)
Statistical measures	**Mode**	**Mode** and **Median**
Graph type	Bar chart (unordered)	Bar chart (ordered categories)

Database : Nominal VS Ordinal Data



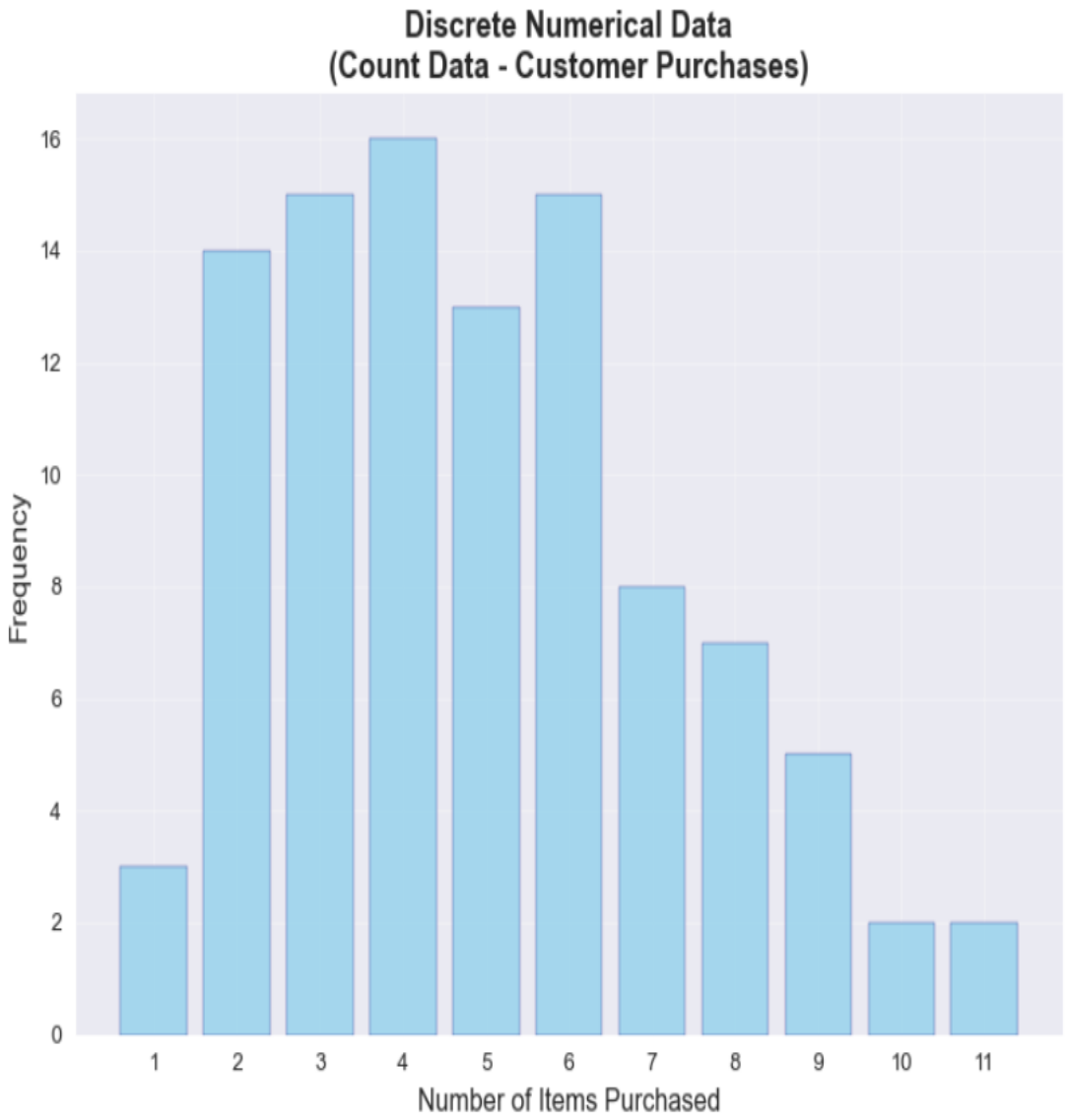


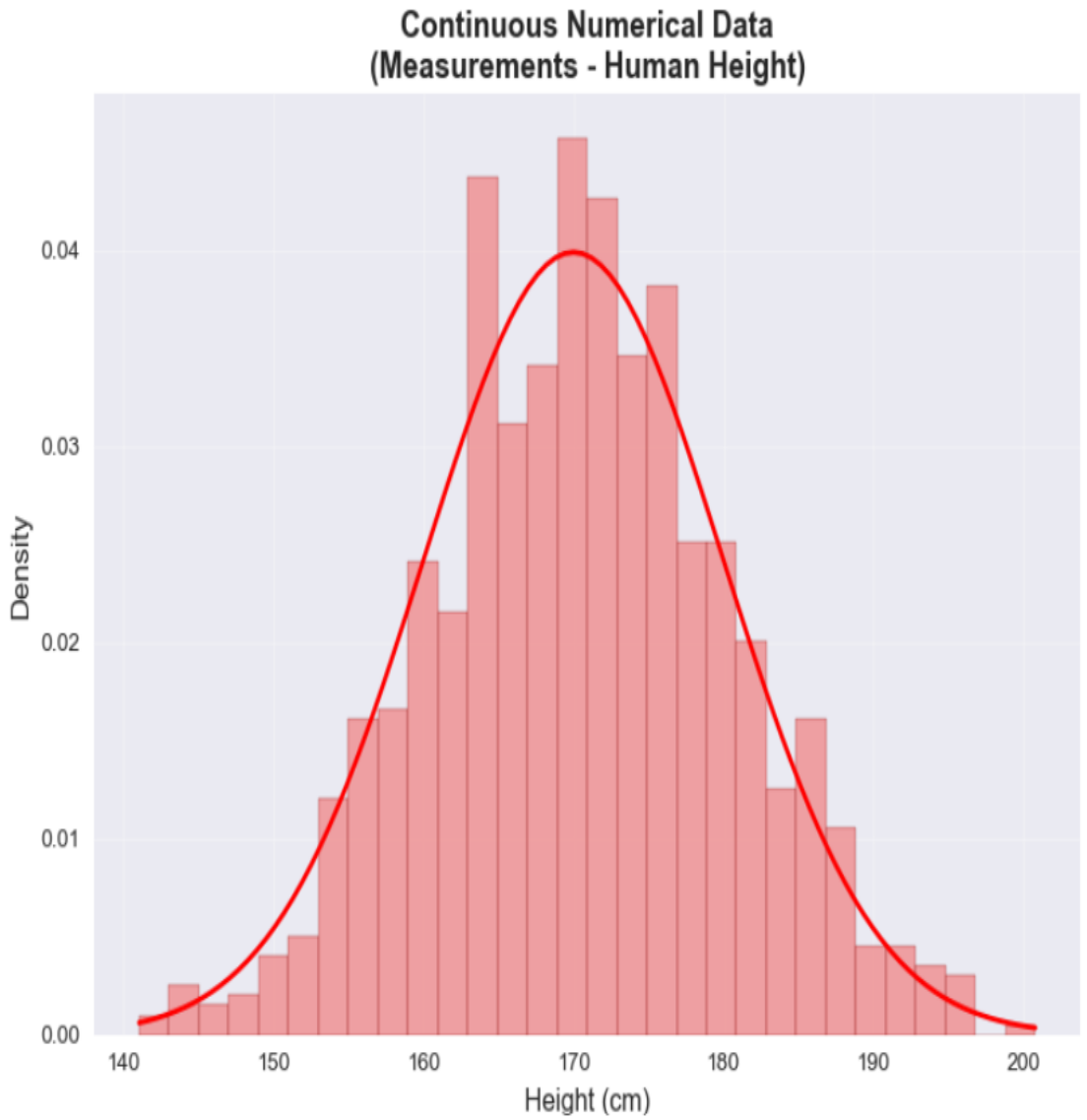
Database : Numerical Database

- Numerical data describe data that can be expressed as **numbers.**
- Allow for measurement, comparison, and calculation
- Show **how much** or **how many** of something
- Examples: 25°C, 170 cm, 60 kg, 5 hours

Type	Description	Example
Discrete Data	Countable numbers (no decimals)	Number of students, cars, books …
Continuous Data	Measurable values that can take **any value within a range** (including decimals)	Measures of height, weight, temperature, time…

Database : Discrete VS Continuous Data





Database : Textual Database

- Textual data base is made up of words, sentences, or paragraphs usually written in natural language (like Arabic, English, French ...).
- **Structured** data is organized **in a fixed format**, like rows and columns in a table. Example: an Excel sheet with “Name”, “Age”, “Salary” columns.
- **Unstructured** data has no fixed format or structure.

Contextual and nuanced richness: Unstructured data containing sentiments expressed in free-text comments

Name	Address	Mail	Date of birth	Date of comment	Comment
Smith	New York	b.smith@gmail.com	07/08/1987	01/02/2024 14:30:45	Very Good
Le grand	Paris	l.johnson@gmail.com	14/08/1997	01/02/2024 14:20:25	Parfait
William	London	e.willimam@yahoo.com	11/02/2001	01/02/2024 14:36:00	I did not like

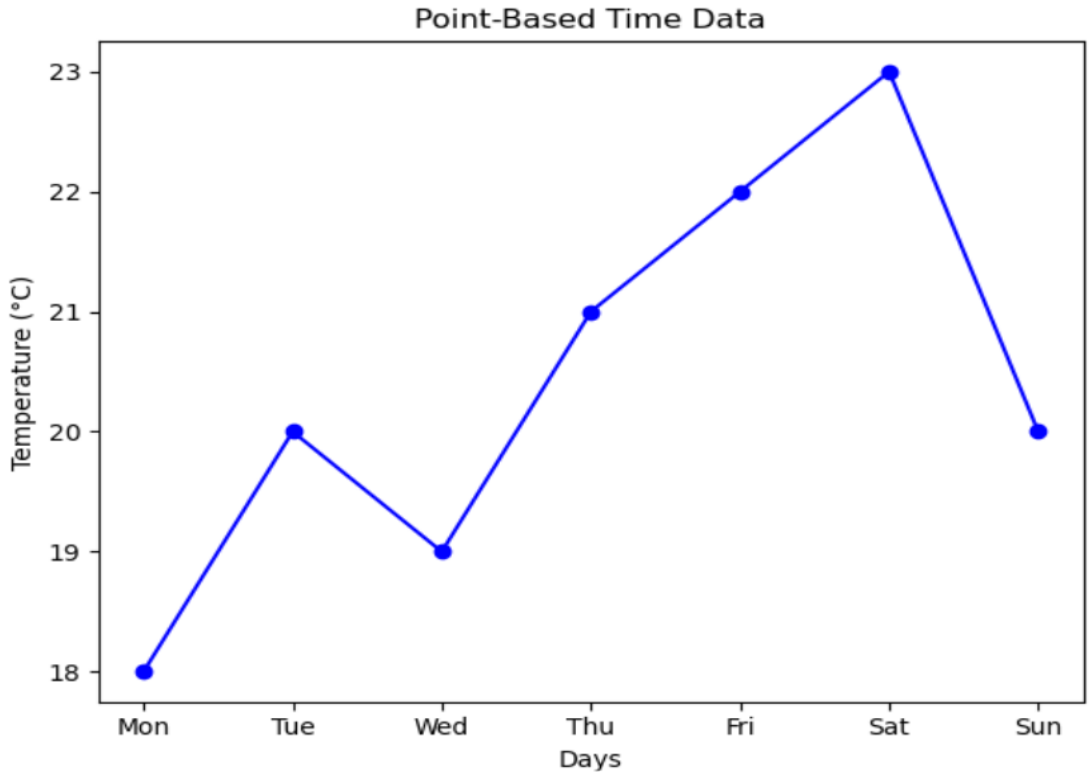
@Smith : Very good
@Le grand : Parfait
@William: I did not like

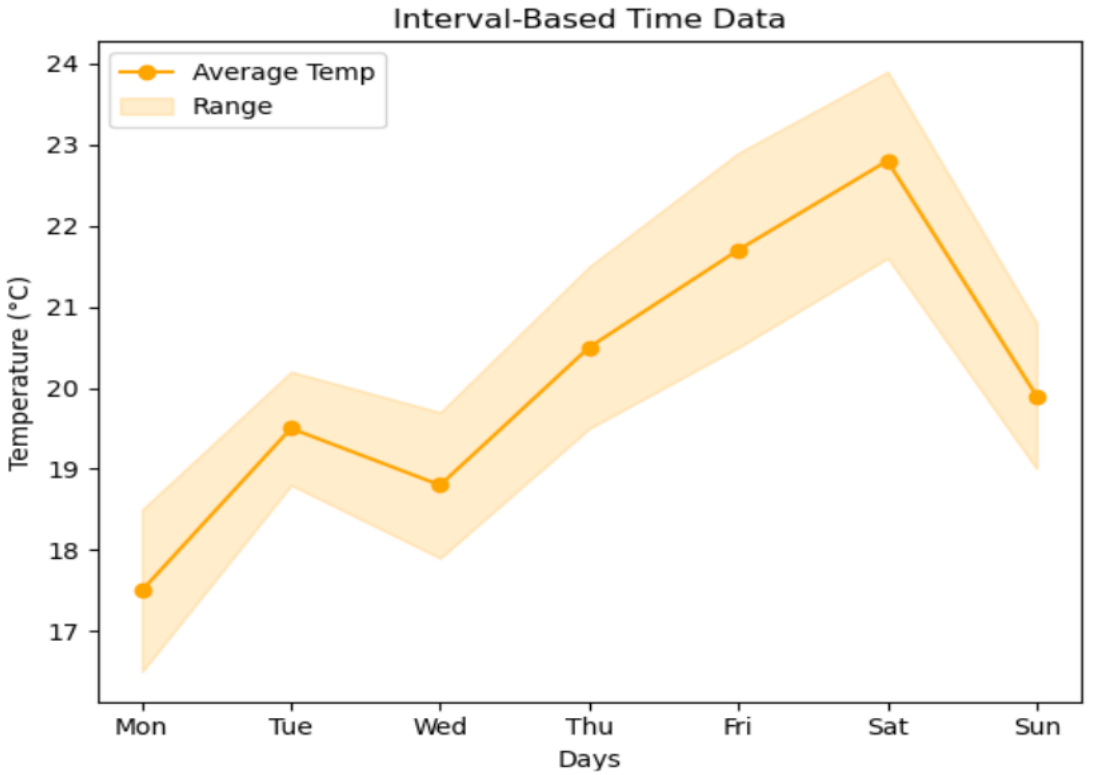
Database : Time Data

- Time-based data is associated with specific timestamps, showing when events occurred or measurements were taken.

- **Point-based time data** is associated with specific timestamps, showing when events occurred or measurements were taken.

- **Interval time data** is measured data over duration period (e.g., hourly sales, daily temperature averages, monthly revenue).





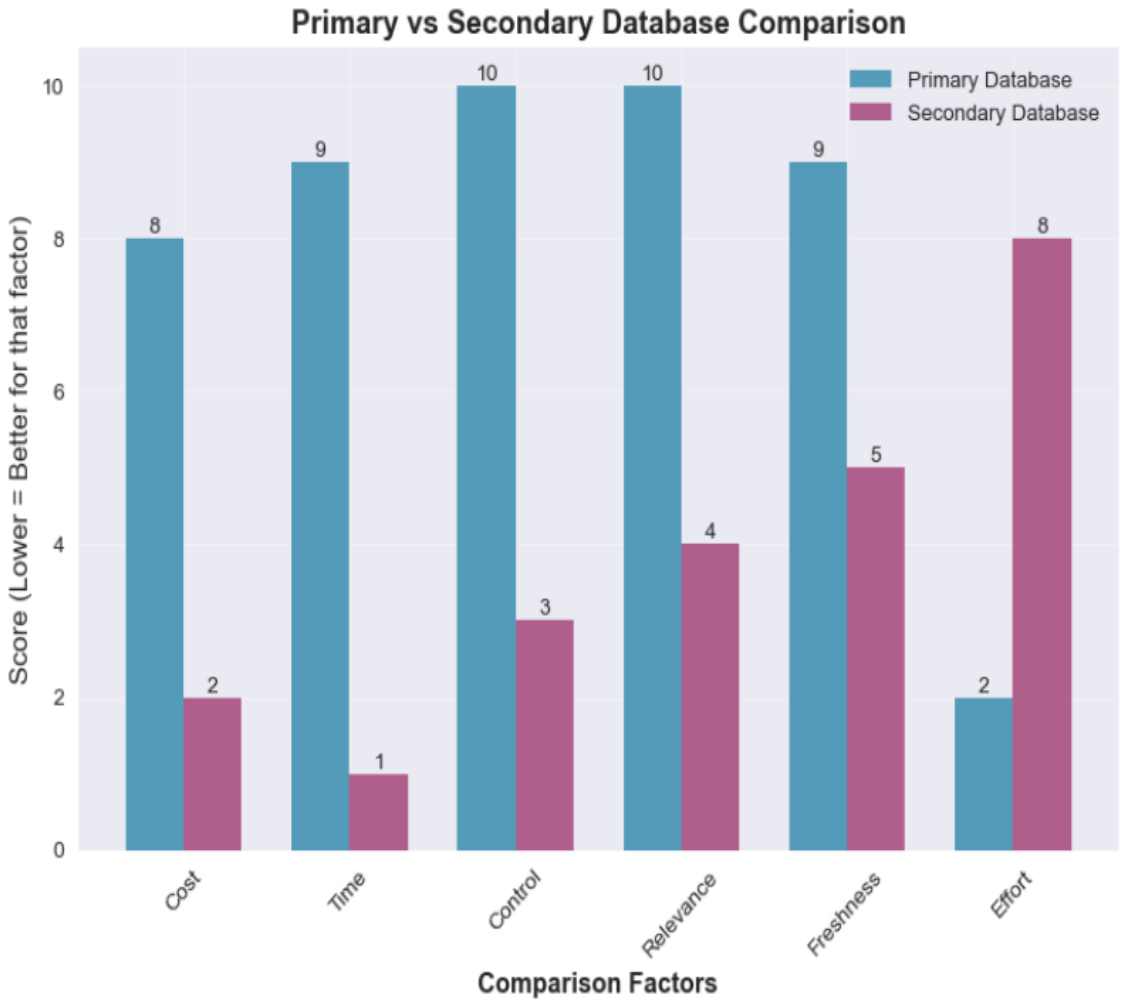
Database : Primary VS Secondary Database

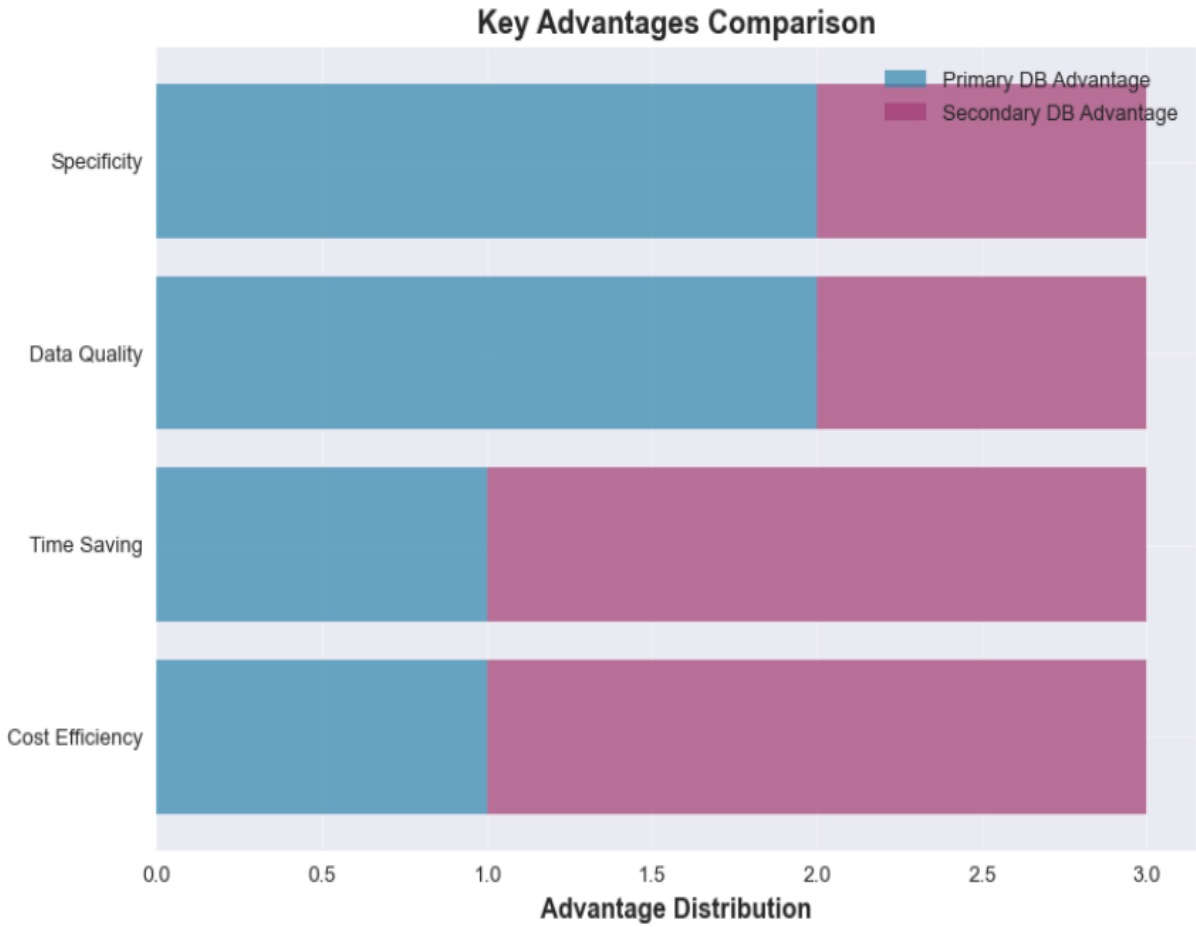
- **Primary Databse** : A main database that stores raw, original, and experimentally verified data

- **Secondary Databse** : Existing data originally collected by others for different purposes.

Feature	Primary Database	Secondary Database
Cost	High	Low
Data Type	Raw and unprocessed	Analyzed and interpreted
Bias	Known kollection Method	Unknown: potential biases
Reliability	High	Low

Database : Primary VS Secondary Database

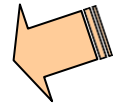




Database : Models with Examples

Different types of databases are designed to store and manage data in various formats, each optimized for specific **structures**, **relationships**, and **use cases**

- Relational DB: Employee list (ID, name, department, salary, ...)
- Transactional DB: Bank transfers (account number, debit, credit, date, ...)
- Document DB: Medical reports (patient ID, diagnosis, notes, ...)
- Key-Value DB: Shopping cart (cart ID, items)
- Graph DB (nodes & relationships): Social network (user, friend connections)
- Time series DB (data over time): Energy usage (meter ID, kilowatts, date)
- Spatial DB (geographic & geometric data) : City Locations (city ID, latitude, longitude, ...)
- Hierarchical DB: Airline reservation system (Univ. => Dep. => Course => Student)

- **What is Data Mining ?**
- **What is Machine Learning ?**
- **KDD Process**
- **Database**
- **Basic Statistical Description of Data** 
- **Data Mining Resources**

Basic Statistical Description of Data : Statistical Metrics

- Measures of central tendency describe the location of the middle or center (average) of a data distribution.
  - Given a Random Variable, they indicate where **most of its values fall**?
  - The main measures are: mean, median, mode

- Dispersion of the data describe how the data are spread out around the center.
  - It can be summarized using the range, quartiles, and interquartile range.
  - It can be represented visually with **the five-number summary and boxplots**
  - **The variance** also measure how much the data values deviate from the mean.

- Graphic displays provide visual representations of basic statistical descriptions to help us inspect and understand our data.
  - Common examples include bar charts, pie charts, and line graphs, as well as quantile plots, quantile–quantile (Q–Q) plots, histograms and scatter plots.

Basic Statistical Description of Data : The Mean

- Arithmetic mean: most common type of average. It shows the central value of the data. It tells what value each item would have if all were the same.

- Geometric mean: the average obtained by multiplying all the values together and then taking the *nth* root, (*n* is the number of values). It shows the typical value in a set of numbers that grow or change by percentages or ratios, such as growth rates.

- Trimmed mean: the average obtained after removing a predefined percentage of the smallest and largest values from the data set before calculating the mean. It shows the typical central value of the data while reducing the effect of extreme or unusual values.

- Weighted mean: the average of values that takes into account their relative importance or weights. It shows the central value when some numbers count more than others. For example, when certain scores or items have greater significance in the calculation

- **+**: Simple and intuitive
- **-**: can be skewed by outliers : it doesn't deal well with wildly varying samples. For example, the monthly income (in $) of 5 person:1000, 1200, 1100, 900, 10000

  The everage income is 2840 $ => It does not represent all people

Basic Statistical Description of Data : The Median

- Median: The middle value in a set of ordered data values. It is used for skewed asymmetric data
  - N values of observations, sorted in increasing order
  - If N is odd => median is the middle value **ELSE** the median is the average of the two middlemost values

Salaries of 10 staff

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

12 14 15 15 15 16 18 17 90 95

$$median = \frac{15 + 16}{2} = 15.5$$

- **+**: Handles outliers well. Splits data into two groups, each with the same number of items
- **-**: Not easy to calculate: you need to sort the list first => High Computing Cost

Basic Statistical Description of Data : The Mode

- Mode: value that occurs most frequently in the set.
  - Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**.
  - If each data value occurs only once, then there is no mode.

Salaries of 10 staff

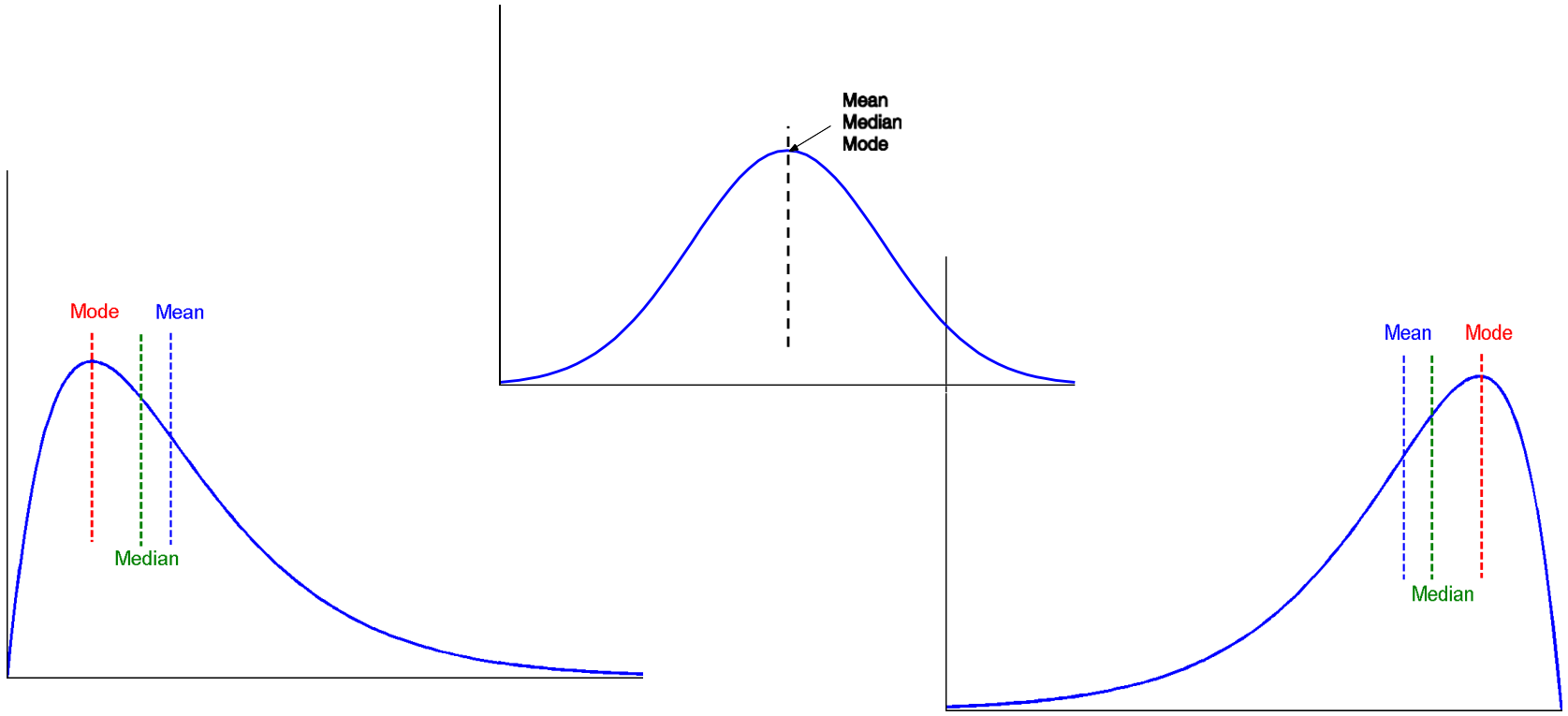
Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

$\text{mode} = 15$

- **+**:
  - Works well for exclusive voting situations (this choice or that one => no compromise)
  - Gives a choice that the most people wanted (whereas the average can give a choice that nobody wanted).
- **-**:
  - Requires more effort to compute (have to count the votes)

Basic Statistical Description of Data : Symetric VS Skewed Data

In **symmetric data**, values are evenly distributed around the center, while in **skewed data**, values stretch more to one side, either right (positive skew) or left (negative skew).



Positively skewed
Mean > median

Negatively skewed
Mean < median

Basic Statistical Description of Data : Dispersion of Data

- Dispersion metrics are used only for numerical data
- The dispersion of data answers the following question: "**How much does my data vary**?"
- A measure of spread gives us an idea of how well the mean represents the data. If the spread of values in the data set is large, the mean is not as representative of the data as if the spread of data is small.

Basic Statistical Description of Data : Percentiles and Quartiles

- Consider the Maximum value of a distribution. Think of it as the value in a set of data that has 100% of the observations at or below it. We call it 100th percentile

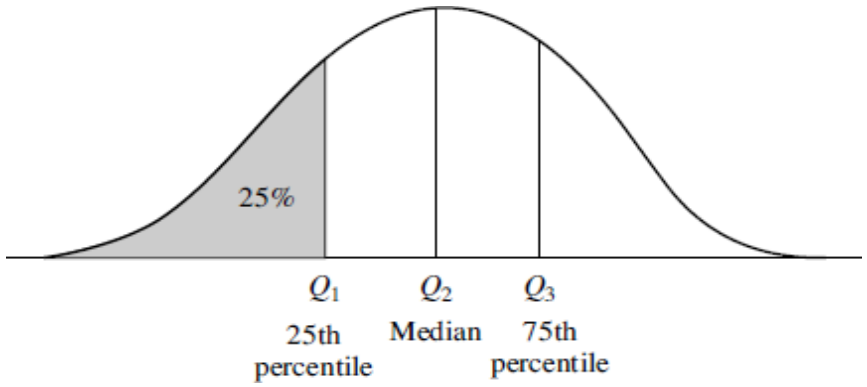
- From this perspective, the median, (which has 50% of the observations at or below it), is the 50th percentile (it is called second quartile)

- p^{th} percentile of a distribution is the value such that p percent of the observations fall at or below it

- The most commonly used percentiles other than the median are 25th percentile and the 75th percentile

Basic Statistical Description of Data : Percentiles and Quartiles

- 25^{th} percentile <=> first quartile
- 50^{th} percentile (median) <=> second quartile
- 75^{th} percentile <=> third quartile



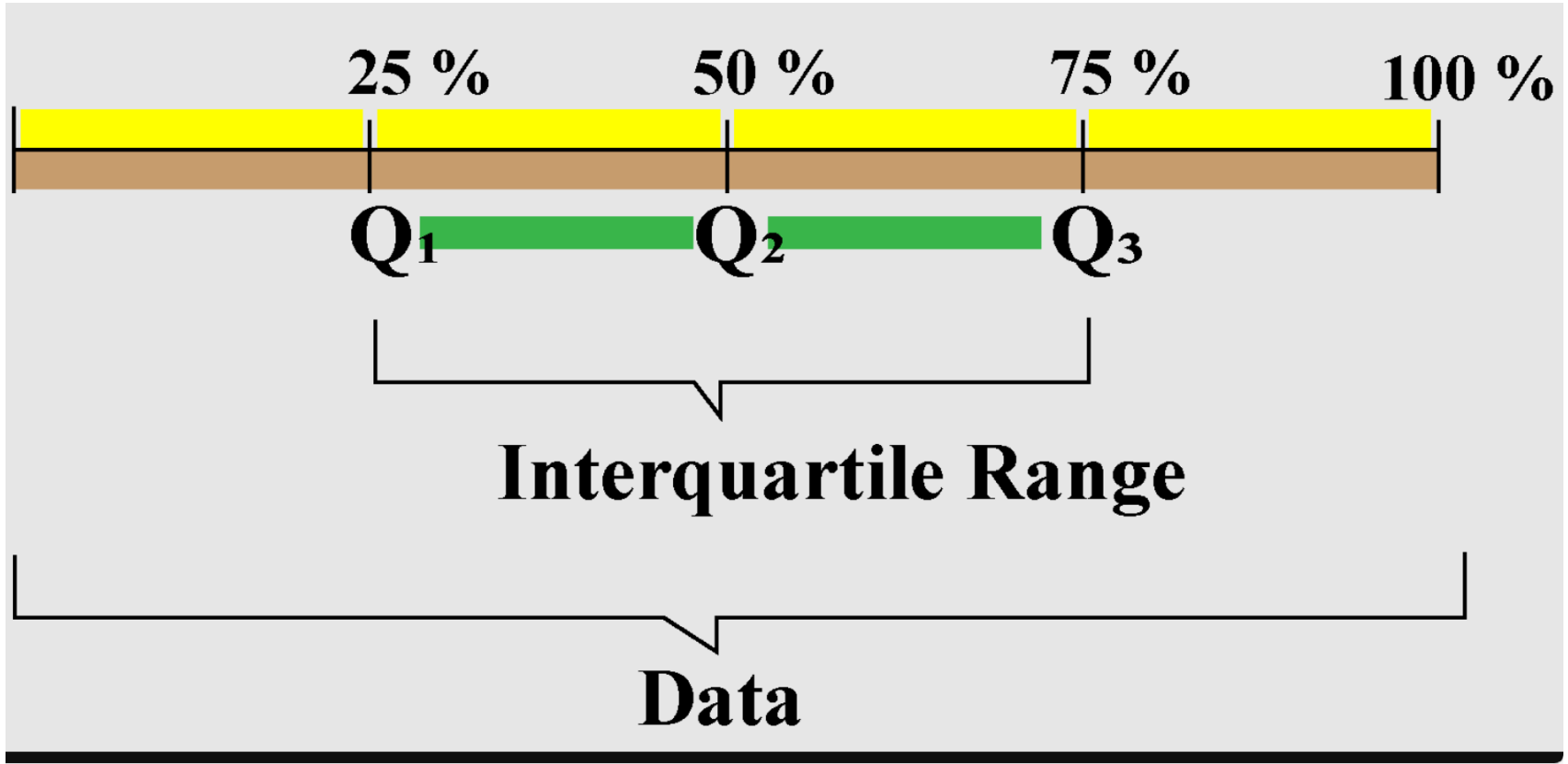
- Quartiles are a useful measure of spread because they are much less affected by outliers or a skewed data set than the equivalent measures of mean and standard deviation.

- Quartiles are often reported along with the median as the best choice of measure of spread and central tendency, respectively, when dealing with skewed and/or data with outliers

Basic Statistical Description of Data : The Interquartile Range

- The interquartile range (IQR) is the difference between the third quartile (Q3) and the first quartile (Q1)

- It measures the **spread of the middle 50% of the data**, showing how concentrated or dispersed the central values are.



Basic Statistical Description of Data : The five numbers

- There is no single numeric measure of spread that is very useful for describing skewed distributions.

- The five-number summary of a distribution consists of:
  - Minimum, Q1, Median, Q3, Maximum

- A common rule for identifying suspected outliers is to single out values falling:
  - Below Q1 – 1.5 × IQR
  - Above Q3 + 1.5 × IQR

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

  - In our example:
    - » Q1 – 1.5 × IQR = 15 - 1.5 × 3 = 10.5 => no outliers on the lower side
    - » Q3 + 1.5 × IQR = 18 + 1.5 × 3 = 22.5 => outliers on the higher side are 90k and 95k

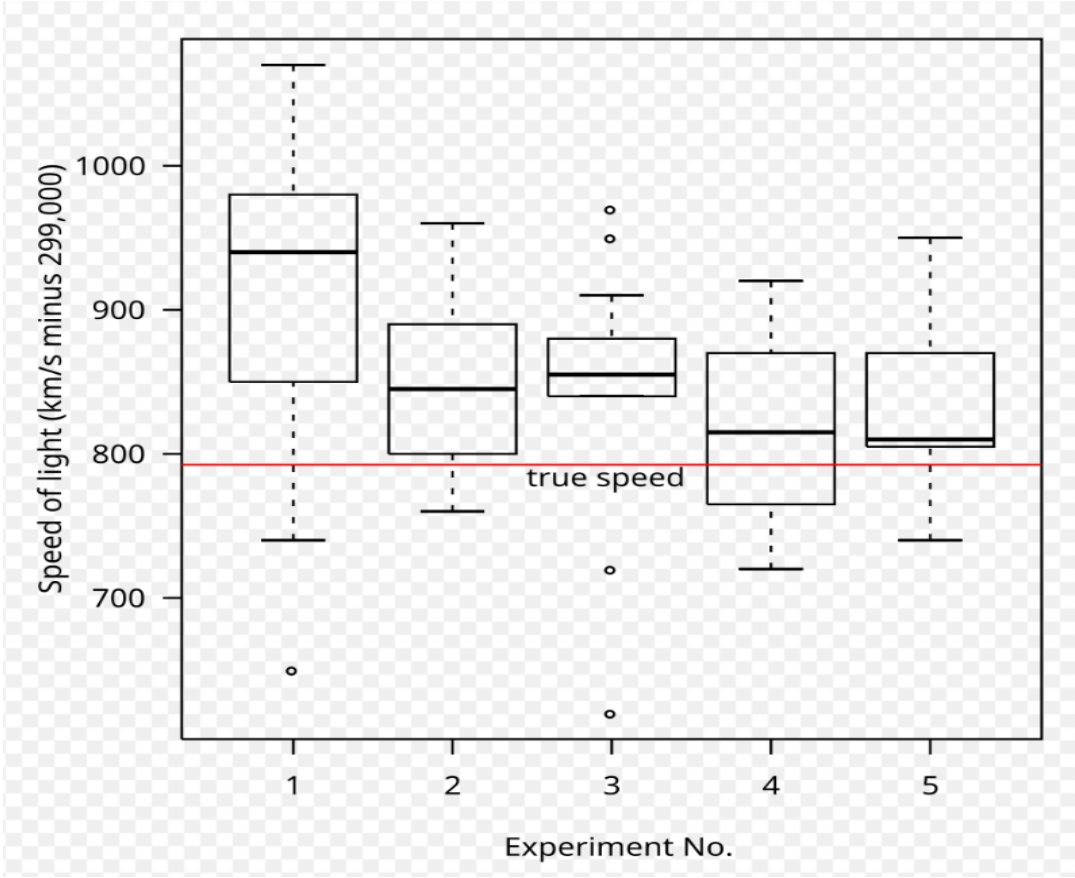
Basic Statistical Description of Data : Boxplots

- A **boxplot** is a simple picture that shows how observations of database are spread out. It uses a box and whiskers to give the 5-number summary of your data.

- The Box:
  - **Left edge of box** = 1st Quartile (Q1)
  - **Right edge of box** = 3rd Quartile (Q3)
  - **Line inside box** = Median (Q2)

- The Whiskers:
  - **Up whisker** (Lowest normal value) = Q1 - 1.5×IQR
  - **Down whisker** (Highest normal value) =Q3 + 1.5×IQR)

- **Outliers**:
  - Any dots outside the whiskers are unusual values



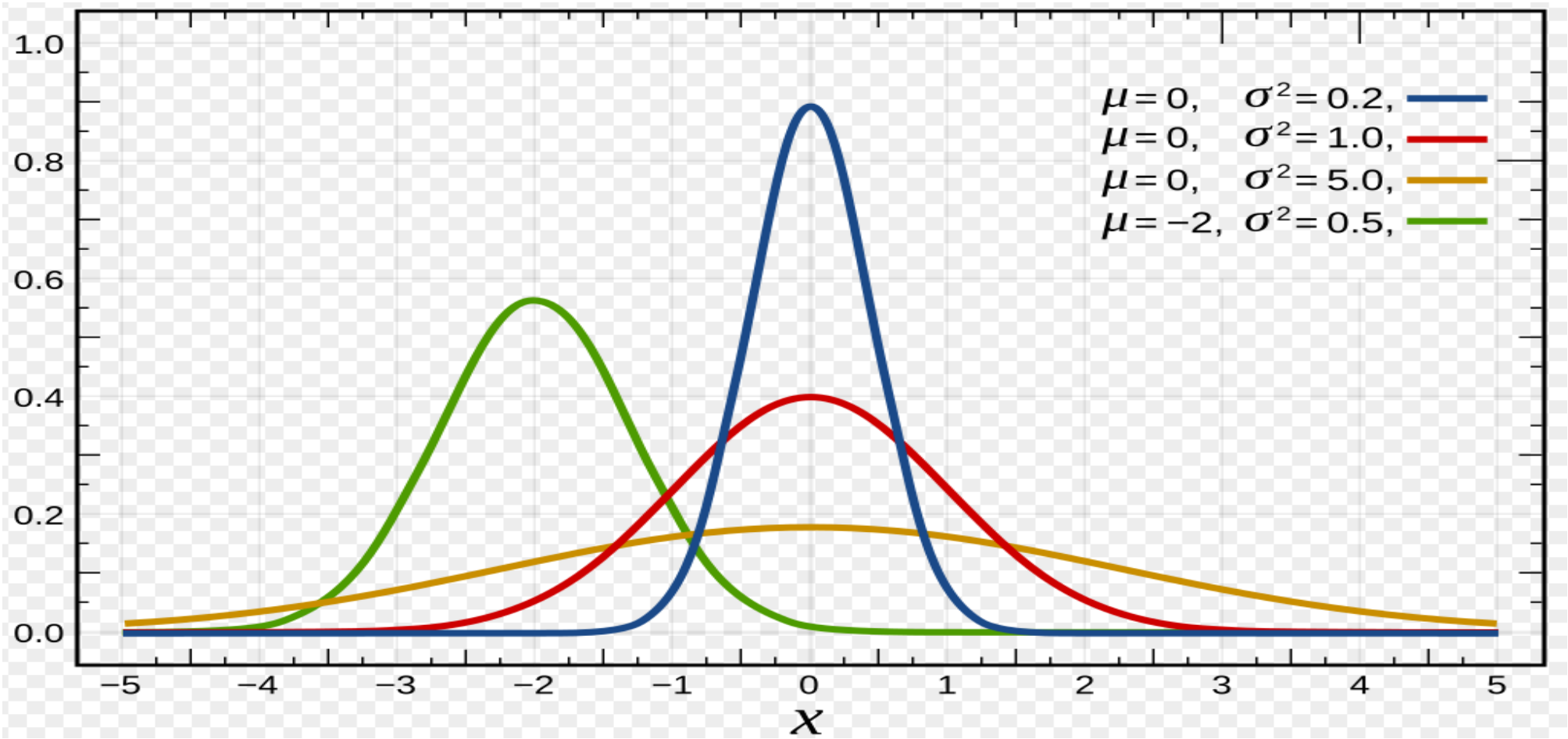
Speed of light Distribution.

Basic Statistical Description of Data : The Variance

- **Variance** measures how spread out a set of numbers is from their average value. It tells you how much your data points differ from the mean.
- Low Variance => Numbers are all close to the average (consistent)
- High Variance => Numbers are spread far from the average (widely dispersed)
- Variance measures spread about the mean and should be considered only when the mean is chosen as the measure of center
- (Var = 0) => there is no spread => all observations have the same value. Otherwise, var > 0.

$$\text{Variance} = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2$$

Basic Statistical Description of Data : The Variance

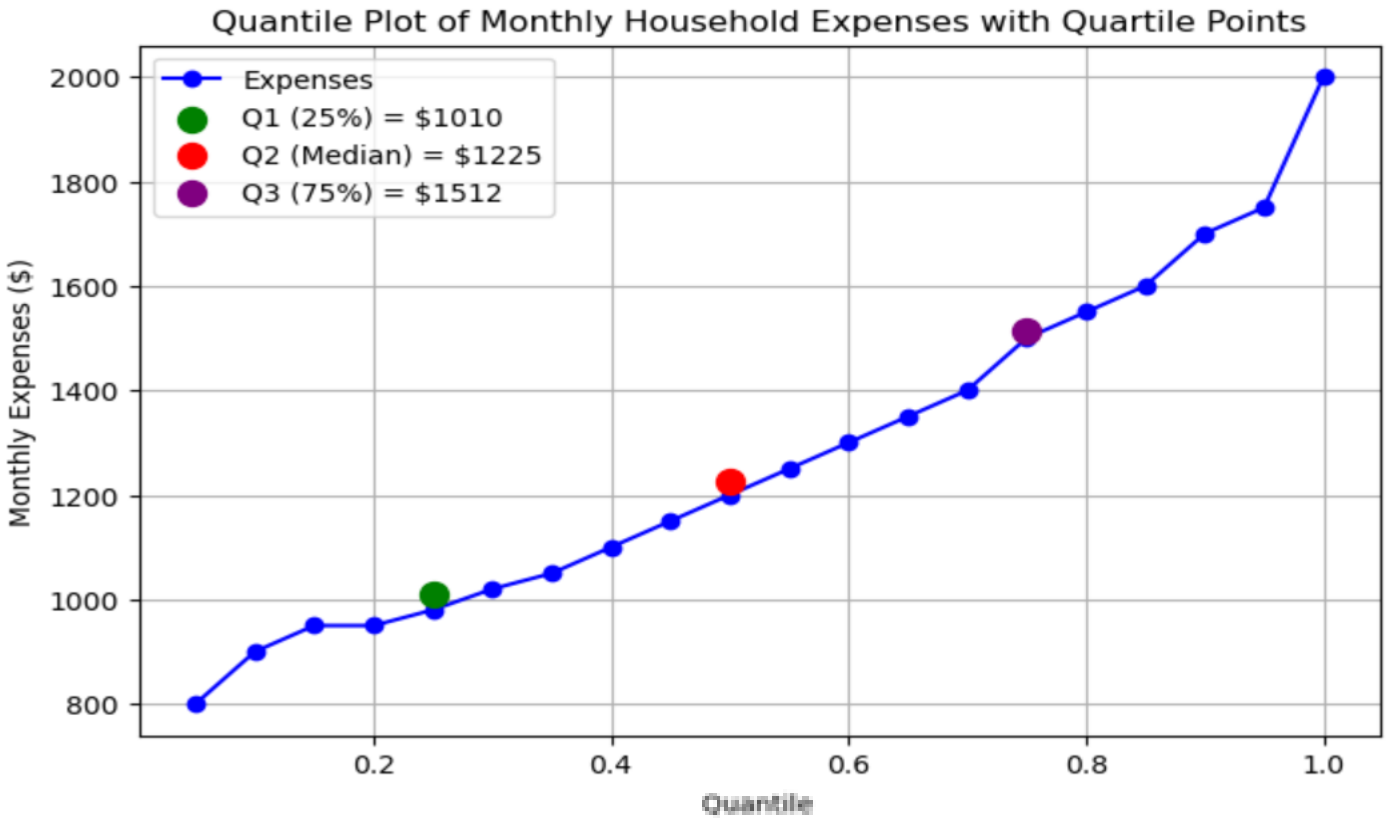


Normal Distribution with different Mean and Variance

Basic Statistical Description of Data : Quantile Plot

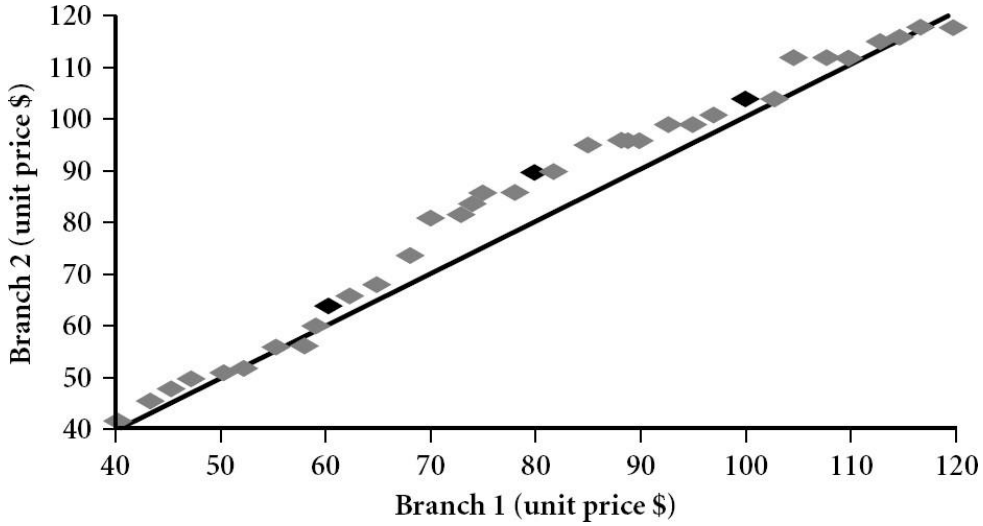
- A graph representing sorted data values versus their quantile levels to show how data are distributed across their range.

- A first step in exploratory data analysis to understand the distribution of the data.



Basic Statistical Description of Data : Quantile-Quantile Plot (Q-Q Plot)

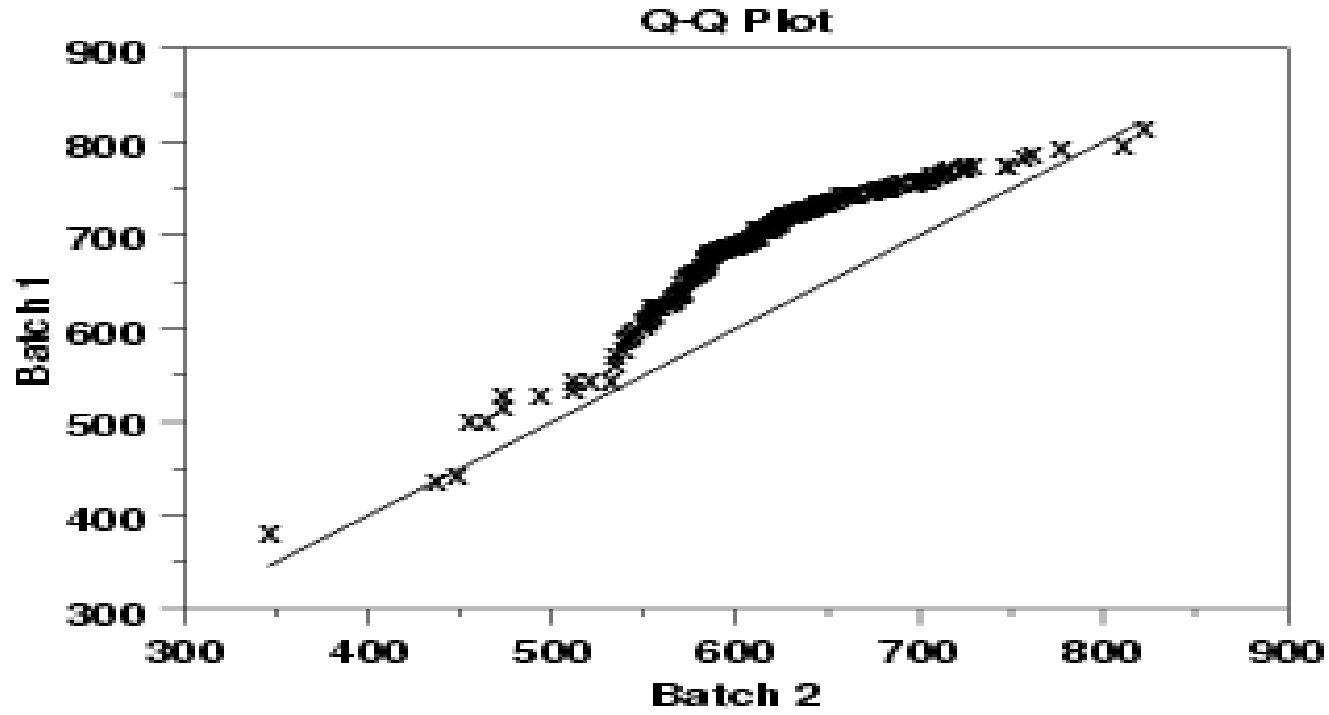
- Plots the quantiles of one univariate distribution against the corresponding quantiles of another
- Test if two data sets come from populations with a common distribution
- 45-degree reference line ($y=x$) is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
- The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions



Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

Basic Statistical Description of Data : Quantile to Quantile plot (Q-Q Plot)

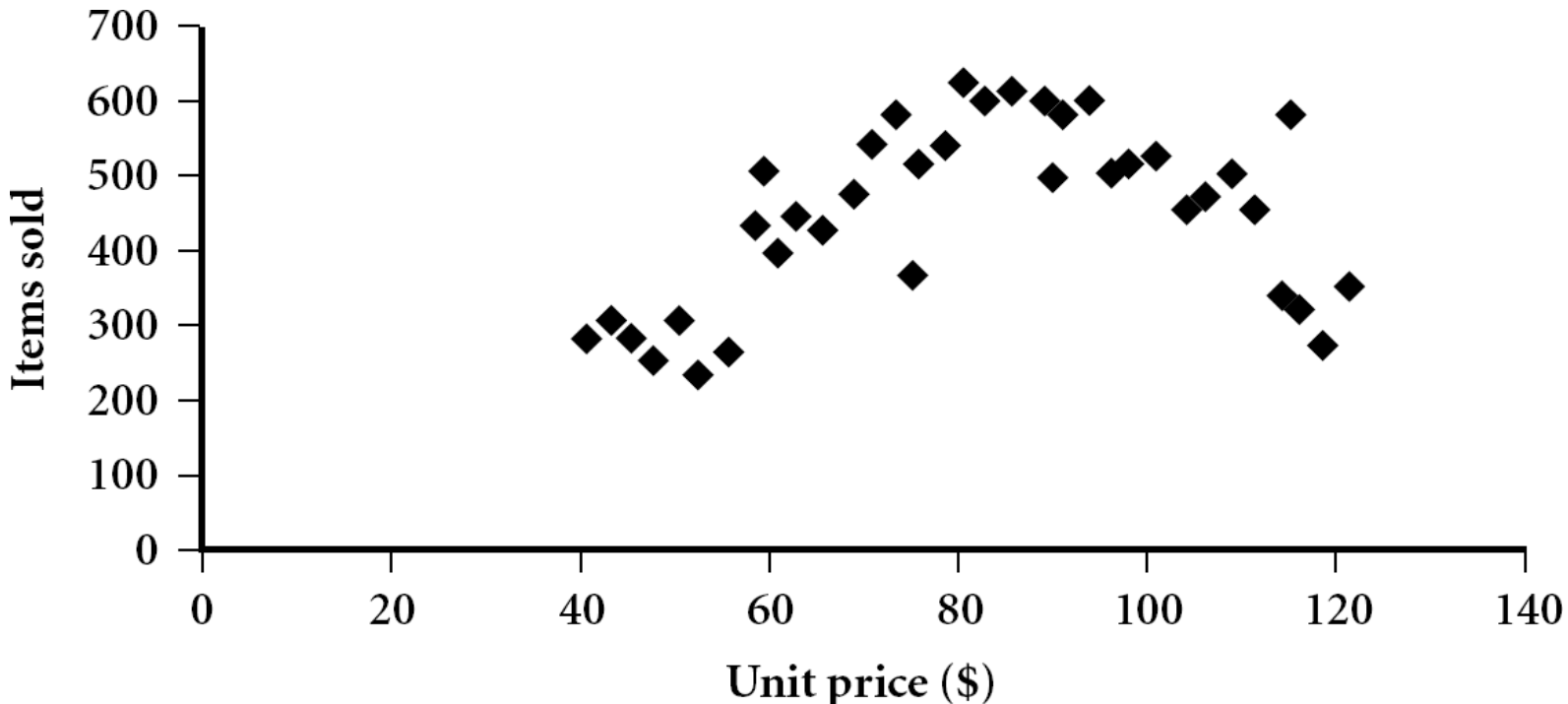
- What does this Q-Q plot indicates?



- These 2 batches do not appear to have come from populations with a common distribution.
- The batch 1 values are significantly higher than the corresponding batch 2 values.
- The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

Basic Statistical Description of Data : Scatter Plot

- Scatter plot: one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Basic Statistical Description of Data : Correlation

- **Linear correlation** is a statistical relationship between two numerical variables where changes in one variable are associated with proportional changes in the other, forming roughly a straight-line pattern on a scatter plot.

- **Covariance** measures the direction of the linear relationship between two numerical variables. It indicates whether the variables tend to increase or decrease together. A positive covariance means that as one variable increases, the other tends to increase as well, while a negative covariance means that as one variable increases, the other tends to decrease.

- **Correlation coefficient** (denoted as *r*) standardizes covariance to a value between **–1 and +1**, allowing for easier interpretation of the strength and direction of the linear relationship. A correlation close to +1 indicates a strong positive relationship, close to –1 indicates a strong negative relationship, and around 0 suggests little to no linear correlation. Unlike covariance, correlation is unit-free, making it useful for comparing relationships between different datasets.

- **Nonlinear (curvilinear) correlation:** the relationship follows a curved pattern rather than a straight line.

Basic Statistical Description of Data : Correlation

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

n = number of observations

x_i, y_i = individual data points

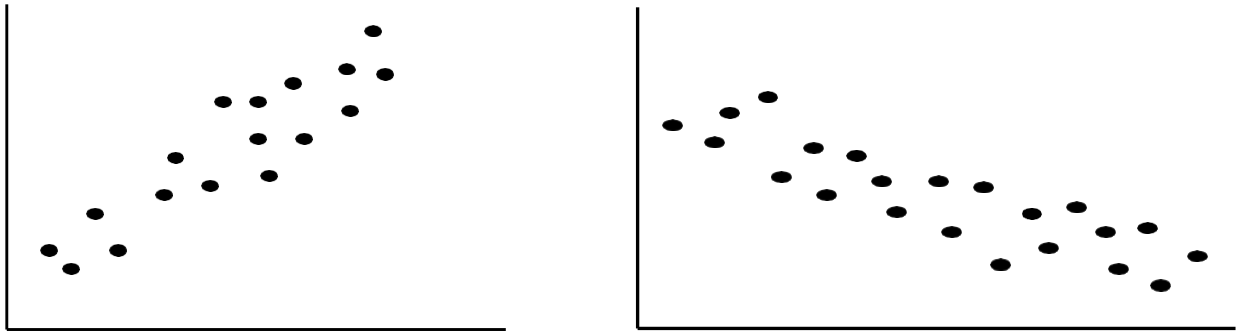
$\bar{x}, \bar{y}$ = mean of X and Y respectively

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X\, s_Y}$$

r = +1 → perfect positive linear relationship

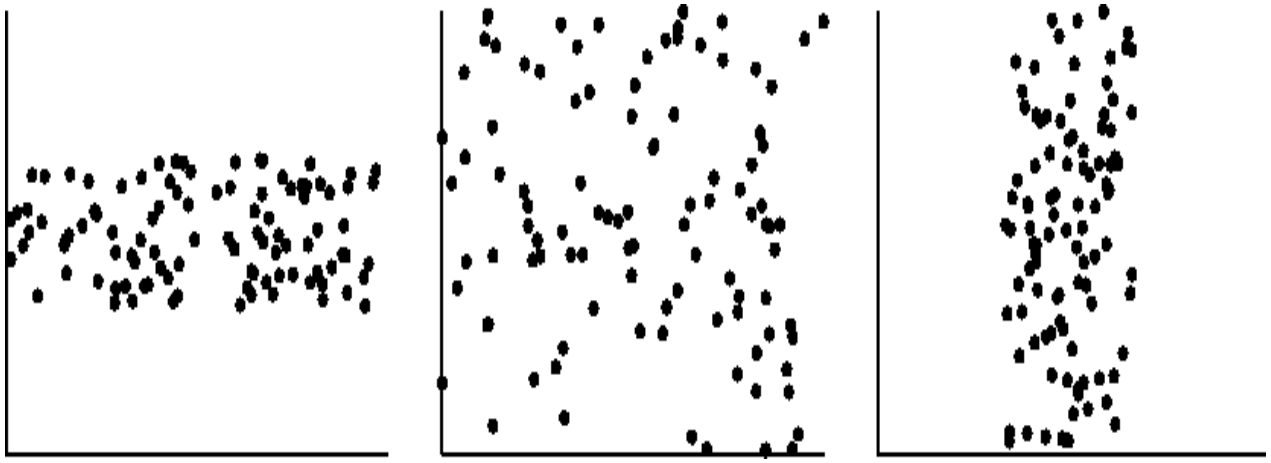
r = –1 → perfect negative linear relationship

r = 0 → no linear relationship

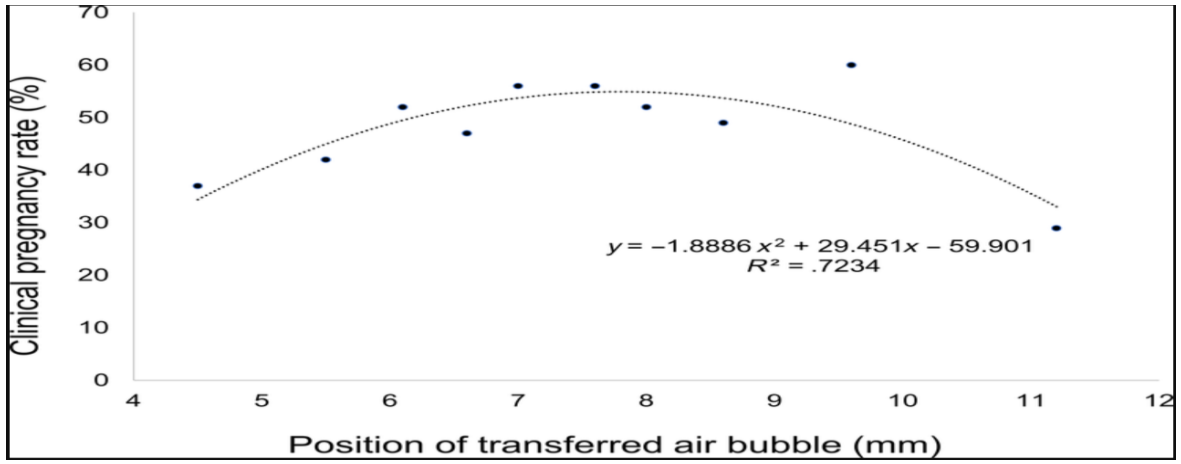


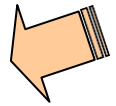
- The left half fragment is positively correlated
- The right half fragment is negative correlated

- Uncorrelated data



- Curvilinear Correlation



- **What is Data Mining ?**
- **What is Machine Learning ?**
- **KDD Process**
- **Database**
- **Basic Statistical Description of Data**
- **Data Mining Resources** 

Data Mining Books

- "Data Mining: Concepts and Techniques (3rd Edition)". J. Han and M. Kamber. Morgan Kaufmann Publishers. 2012. **(Textbook)**
- Introduction to Data Mining (2nd edition) P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar. Pearson, 2018.
- "Data Mining: Practical Machine Learning Tools and Techniques (4th Edition)" I.H. Witten, E. Frank, M. Hall, C. Pal. Morgan Kaufmann Publishers. 2017.
- "Advances in Knowledge Discovery and Data Mining". Eds.: Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy. The MIT Press, 1995.

Data Mining Journals

- Data Mining and Knowledge Discovery Journal
- ACM SIGKDD Explorations Newsletter
- TKDE: IEEE Transactions in Knowledge and Data Engineering
- TODS: ACM Transactions on Database Systems
- JACM: Journal of ACM
- Data and Knowledge Engineering
- JIIS: Intl. Journal of Intelligent Information Systems
- …

Data Mining Conferences

- KDD: ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining
- ICDM: IEEE International Conference on Data Mining,
- SIAM International Conference on Data Mining
- PKDD: European Conference on Principles and Practice of Knowledge Discovery in Databases
- PAKDD Pacific-Asia Conference on Knowledge Discovery and Data Mining
- DaWak: Intl. Conference on Data Warehousing and Knowledge Discovery

Other related Conferences:

- ICML: Intl. Conf. On Machine Learning
- IDEAL: Intl. Conf. On Intelligent Data Engineering and Automated Learning
- IJCAI: International Joint Conference on Artificial Intelligence
- AAAI: American Association for Artificial Intelligence Conference
- SIGMOD/PODS: ACM Intl. Conference on Data Management
- ICDE: International Conference on Data Engineering
- VLDB: International Conference on Very Large Data Bases

Data Mining Datasets

- Univ. of California Irvine Machine Learning Data Repository.
- Univ. of California Irvine KDD Data Repository.
- Datasets for Data Mining
- Datamob - Public data put to good use.
- Time Series Data Library
- CMU's StatLib-Datasets Archive
- Stanford Large Network Dataset Collection (SNAP)
- 100+ Interesting Data Sets for Statistics
- …

GITHUB REPOSITORY

https://github.com/rida87/DataMining