Subject : Data Warehouse and Data Mining

Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

B Tech CSE-AI-DS
SY Sem IV

Disclaimer:

1. Information included in this slides came from multiple sources. We have tried our best to cite the sources. Please refer to the References to learn about the sources, when applicable.
2. The slides should be used only for academic purposes (e.g., in teaching a class), and should not be used for commercial purposes.

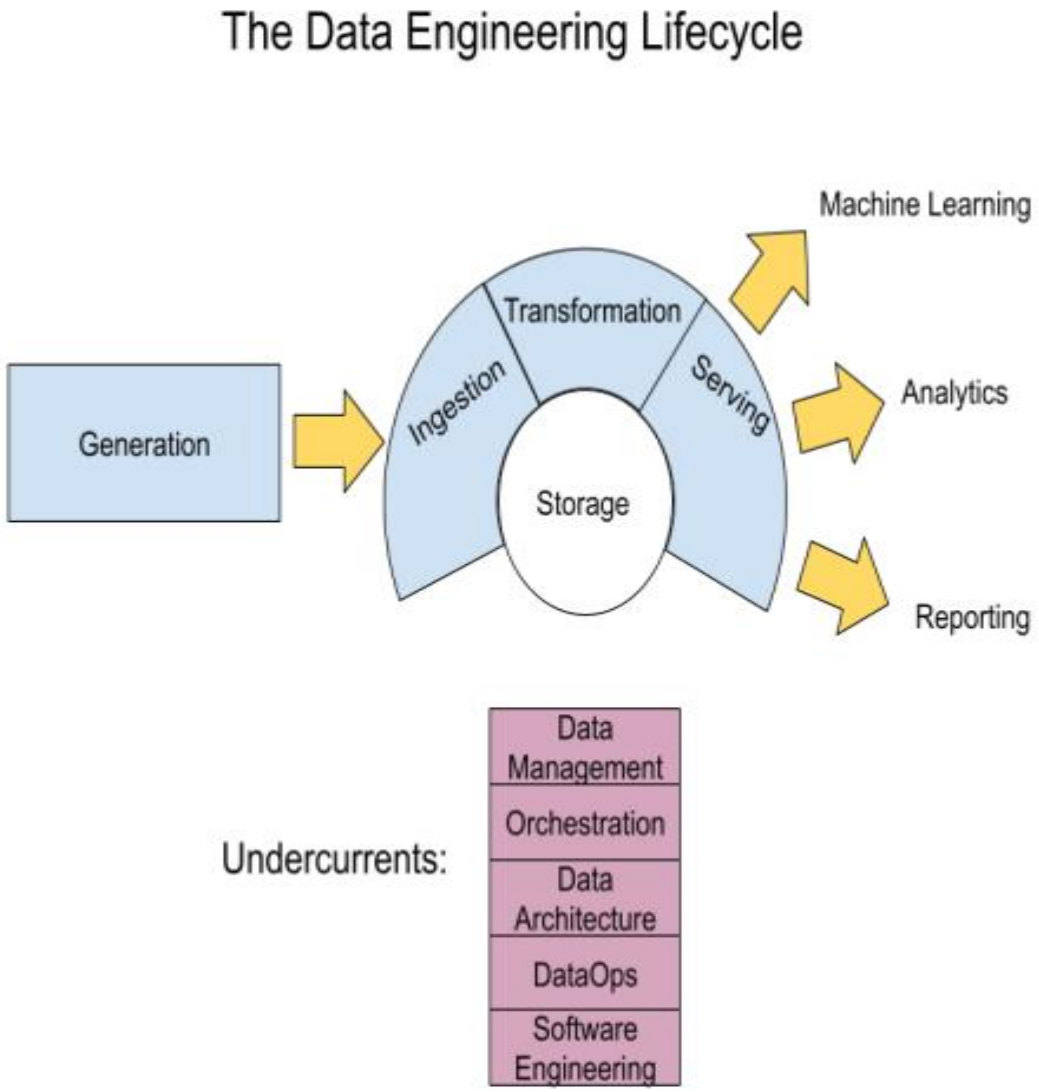
UNIT I - DATA PREPROCESSING: AN OVERVIEW

- Introduction to Data Engineering
- Defining Data Engineering, Overview of the Data Engineering Ecosystem, Data Engineering Lifecycle
- Data objects and attribute types, Data Characteristics, Types of Data, Structured, Unstructured, Semi-structured, Discrete, Continuous, Ordinal, Nominal, Qualitative, Quantative, Time series data, Geographical data.
- Measures of Central Tendency: Mean, Median, Mode
- Measures of Dispersion: Range, Variance, Standard Deviation.

Introduction to Data Engineering

- Data engineering: The development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information that supports downstream use cases, such as analysis and machine learning.
- Data engineering is the intersection of data management, data architecture ,software engineering.
- Data engineers Manage the data engineering lifecycle beginning with ingestion and ending with serving data for use cases, such as analysis or machine learning.

Data Engineering Lifecycle



Evolution of Database Technology

- 1960s: Data collection, database creation, IMS and network DBMS
- 1970s: Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s: Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and data mining
  - Data mining, applications and Web technology

What is Data Mining?

- Data mining is also called *knowledge discovery and data mining* (KDD)
- Data mining is
  - Extraction of useful patterns from data sources, e.g.- Databases, texts, web, image
- Patterns must be:
  - Valid, novel, potentially useful, understandable

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems

Why is Data Mining important?

- Rapid computerization of businesses produce huge amount of data
- How to make best use of data?
- It is used to discover patterns and relationships in the data in order to help make better business decisions
- Data mining technology can generate new business opportunities by:
  - *Automated prediction of trends and behaviors*
  - *Automated discovery of previously unknown patterns*

Why Data Mining?

The Explosive Growth of Data: from terabytes to petabytes

- **Data and storage predictions for the year 2025(zettabytes)**
  - The storage industry will ship 42ZB of capacity over the next seven years.
  - 90ZB of data will be created on IoT devices by 2025.
  - By 2025, 49 percent of data will be stored in public cloud environments.
  - Nearly 30 percent of the data generated will be consumed in real-time by 2025.

Why Data Mining?

The Explosive Growth of Data: from terabytes to zettabytes

- Data collection and data availability
  - Automated data collection tools, database systems, Web, computerized society
- Major sources of abundant data
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bioinformatics, scientific simulation, …
  - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Example of Discovered Patterns

- Association rules:

  "80% of customers who buy *cheese* and *milk* also buy *bread*, and 5% of customers buy all of them together"

  Cheese, Milk→ Bread [sup =5%, confid=80%]

Data mining Algorithms: can be characterized as consisting of three parts

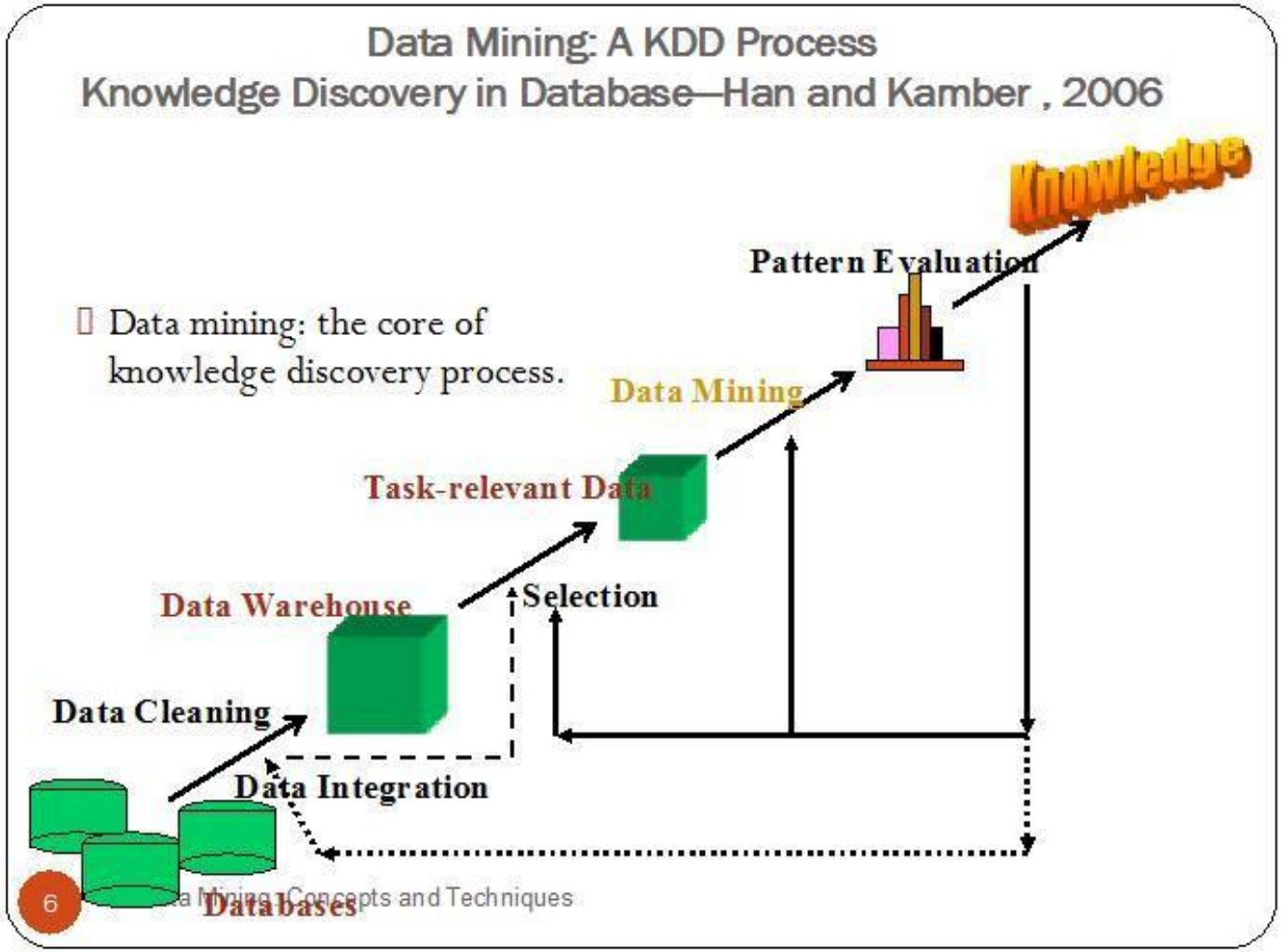
- Model: is to be fit on data.
- Preference: criteria to select one model over other
- Search: techniques to evaluate data point or searching of data.

Goal of Data Mining

- Goal of Data Mining to provide efficient tools and techniques for KDD

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discover process



Knowledge Discovery (KDD) Process

1. **Developing an understanding of:**
   - The application domain
   - The relevant prior knowledge
   - The goals of the end-user
   - Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

2. **Data cleaning and preprocessing.**
   - Removal of noise or outliers.
   - Collecting necessary information to model or account for noise.
   - Strategies for handling missing data fields.
   - Accounting for time sequence information and known changes.

3. **Data reduction and projection**
   - Finding useful features to represent the data depending on the goal of the task.
   - Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

Knowledge Discovery (KDD) Process

4. Choosing the data mining task.
   - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
5. Choosing the data mining algorithm(s)
   - Selecting method(s) to be used for searching for patterns in the data.
   - Deciding which models and parameters may be appropriate.
   - Matching a particular data mining method with the overall criteria of the KDD process.
6. Data mining.
   - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
7. Interpreting mined patterns.
8. Consolidating discovered knowledge.

Sequence of the steps

- **Data Cleaning:** To remove noise and inconsistent data.
- **Data Integration:** where multiple data sources may be combined.
- **Data Selection:** Where data relevant to the analysis task are retrieved from the database.
- **Data Transformation:** where data transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining:** An essential process where intelligent methods are applied to extract data patterns.
- **Pattern evaluation:** To identify the truly interesting pattern representing knowledge based on interesting measures.
- **Knowledge Presentation:** where visualization and knowledge representation tech. are used to present mined knowledge to users.

Applications of Data Mining

- Data mining applications are widely used in
  - Direct marketing
  - Health industry
  - E-commerce
  - Customer relationship management (CRM)
  - Telecommunication industry and financial sector, etc..
- Data mining is available in various forms
  - Text mining
  - Web mining
  - Audio & video data mining
  - Pictorial data mining
  - Relational databases data mining
  - Social networks data mining

Data Warehouse and Data Mining

- Introduction to KDD
- **Data Preprocessing: An Overview**
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Why Is Data Preprocessing Important?

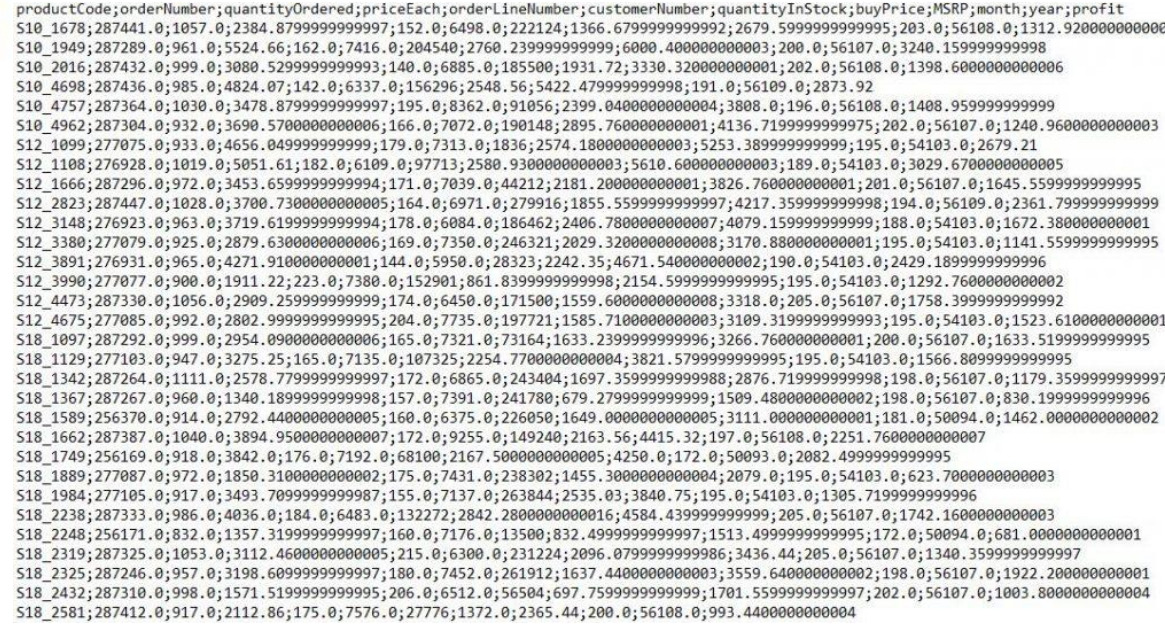
- No quality data, no quality mining results!
  - **Quality** decisions must be based on quality data
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application (90%).

Types of Data :Structured

- Structured data is data with a high degree of organization, usually stored in some sort of spreadsheet. Excel sheet, which is a prime example of structured data.

CUSTOMER

CUSTOMER_ID	LAST_NAME	FIRST_NAME	STREET	CITY	ZIP_CODE	COUNTRY
10302	Boucher	Leo	54, rue Royale	Nantes	44000	France
11244	Smith	Laurent	8489 Strong St	Las Vegas	83030	USA
11405	Han	James	636 St Kilda Road	Sydney	3004	Australia
11993	Mueller	Tomas	Berliner Weg 15	Tamm	71732	Germany
12111	Carter	Nataly	5 Tomahawk	Los Angeles	90006	USA
14121	Cortez	Nola	Av. Grande, 86	Madrid	28034	Spain
14400	Brown	Frank	165 S 7th St	Chester	33134	USA
14578	Wilson	Sarah	Seestreet #6101	Emory	1734	USA
14622	Jones	John	71 San Diego Ave	Arlington	69004	USA



```
productCode;orderNumber;quantityOrdered;priceEach;orderLineNumber;customerNumber;quantityInStock;buyPrice;MSRP;month;year;profit
S10_1678;287441.0;1057.0;2384.8799999999997;152.0;6498.0;222124;1366.6799999999992;2679.5999999999995;203.0;56108.0;1312.920000000000
S10_1949;287289.0;961.0;5524.66;162.0;7416.0;204540;2760.239999999999;6000.400000000003;200.0;56107.0;3240.159999999998
S10_2016;287432.0;999.0;3080.5299999999993;140.0;6885.0;185500;1931.72;3330.320000000001;202.0;56108.0;1398.6000000000006
S10_4698;287436.0;985.0;4824.07;142.0;6337.0;156296;2548.56;5422.479999999998;191.0;56109.0;2873.92
S10_4757;287364.0;1030.0;3478.8799999999997;195.0;8362.0;91056;2399.0400000000004;3808.0;196.0;56108.0;1408.959999999999
S10_4962;287304.0;932.0;3690.5700000000006;166.0;7072.0;190148;2895.760000000001;4136.7199999999975;202.0;56107.0;1240.9600000000003
S12_1099;277075.0;933.0;4656.049999999999;179.0;7313.0;1836;2574.1800000000003;5253.389999999999;195.0;54103.0;2679.21
S12_1108;276928.0;1019.0;5051.61;182.0;6109.0;97713;2580.9300000000003;5610.600000000003;189.0;54103.0;3029.6700000000005
S12_1666;287296.0;972.0;3453.6599999999994;171.0;7039.0;44212;2181.200000000001;3826.760000000001;201.0;56107.0;1645.5599999999995
S12_2823;287447.0;1028.0;3700.7300000000005;164.0;6971.0;279916;1855.5599999999997;4217.359999999998;194.0;56109.0;2361.799999999999
S12_3148;276923.0;963.0;3719.6199999999994;178.0;6084.0;186462;2406.7800000000007;4079.159999999999;188.0;54103.0;1672.380000000001
S12_3380;277079.0;925.0;2879.6300000000006;169.0;7350.0;246321;2029.3200000000008;3170.880000000001;195.0;54103.0;1141.5599999999995
S12_3891;276931.0;965.0;4271.910000000001;144.0;5950.0;28323;2242.35;4671.540000000002;190.0;54103.0;2429.1899999999996
S12_3990;277077.0;900.0;1911.22;223.0;7380.0;152901;861.8399999999998;2154.5999999999995;195.0;54103.0;1292.7600000000002
S12_4473;287330.0;1056.0;2909.259999999999;174.0;6450.0;171500;1559.6000000000008;3318.0;205.0;56107.0;1758.3999999999992
S12_4675;277085.0;992.0;2802.9999999999995;204.0;7735.0;197721;1585.7100000000003;3109.3199999999993;195.0;54103.0;1523.6100000000001
S18_1097;287292.0;999.0;2954.0900000000006;165.0;7321.0;73164;1633.2399999999996;3266.760000000001;200.0;56107.0;1633.5199999999995
S18_1129;277103.0;947.0;3275.25;165.0;7135.0;107325;2254.7700000000004;3821.5799999999995;195.0;54103.0;1566.8099999999995
S18_1342;287264.0;1111.0;2578.7799999999997;172.0;6865.0;243404;1697.3599999999988;2876.719999999998;198.0;56107.0;1179.3599999999997
S18_1367;287267.0;960.0;1340.1899999999998;157.0;7391.0;241780;679.2799999999999;1509.4800000000002;198.0;56107.0;830.1999999999996
S18_1589;256370.0;914.0;2792.4400000000005;160.0;6375.0;226050;1649.0000000000005;3111.000000000001;181.0;50094.0;1462.0000000000002
S18_1662;287387.0;1040.0;3894.9500000000007;172.0;9255.0;149240;2163.56;4415.32;197.0;56108.0;2251.7600000000007
S18_1749;256169.0;918.0;3842.0;176.0;7192.0;68100;2167.5000000000005;4250.0;172.0;50093.0;2082.4999999999995
S18_1889;277087.0;972.0;1850.3100000000002;175.0;7431.0;238302;1455.3000000000004;2079.0;195.0;54103.0;623.7000000000003
S18_1984;277105.0;917.0;3493.7099999999987;155.0;7137.0;263844;2535.03;3840.75;195.0;54103.0;1305.7199999999996
S18_2238;287333.0;986.0;4036.0;184.0;6483.0;132272;2842.2800000000016;4584.439999999999;205.0;56107.0;1742.1600000000003
S18_2248;256171.0;832.0;1357.3199999999997;160.0;7176.0;13500;832.4999999999997;1513.4999999999995;172.0;50094.0;681.0000000000001
S18_2319;287325.0;1053.0;3112.4600000000005;215.0;6300.0;231224;2096.0799999999986;3436.44;205.0;56107.0;1340.3599999999997
S18_2325;287246.0;957.0;3198.6099999999997;180.0;7452.0;261912;1637.4400000000003;3559.640000000002;198.0;56107.0;1922.200000000001
S18_2432;287310.0;998.0;1571.5199999999995;206.0;6512.0;56504;697.7599999999999;1701.5599999999997;202.0;56107.0;1003.8000000000004
S18_2581;287412.0;917.0;2112.86;175.0;7576.0;27776;1372.0;2365.44;200.0;56108.0;993.4400000000004
```

Data Types of Structured Data

Data types in structed data:

A structured data can have multiple data types in it. Let us look at a few of them.

- **Integers** – The data can be of integer or numeric format, which is a whole number. Example – Age, Number of runs scored, Number of cars owned.
- **Decimals** – The data can be of decimal type which is also known as fractions. Example – CGPA, Price.
- **Text** – The data can be of text format. Examples are Name, Location, Company Name.
- **Date** – The data can be of data format. It can of any data format depending on the problem statement. Example DOB, Order data

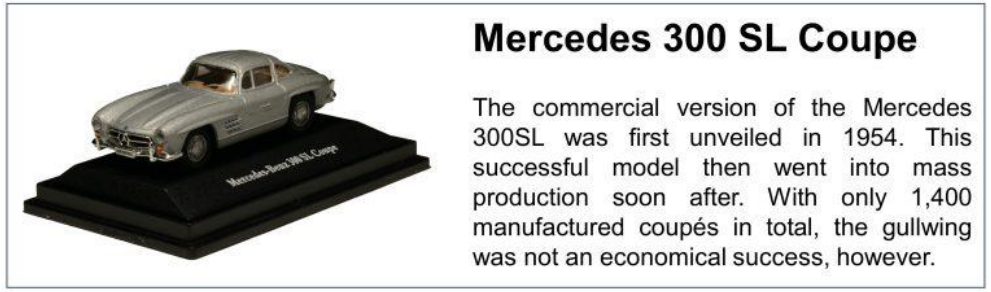
Semi structured Data

- Semi-structured is data which has some degree of organization in it.
- It is not as rigorously structured as structured data, but also not as messy as unstructured data.
- This degree of organization is typically achieved with some sort of tags or other elements with defined properties which introduce a hierarchy and system into a file.
- However, the order and amount of such structuring tags and elements may vary.
- Therefore, the structure imposed on a dataset it not as rigorous as in structured datasets where all data has to conform to the structure of the data table (spreadsheet).

```
{
    "EMPLOYEES": {
        "SALES": {
            "648229": {
              "NAME" : "Olivia Johnson"
              "DOB" : "1989-08-08"
            },
            "648666": {
              "NAME" : "Frank Mueller"
              "DOB" : "1985-05-11"
              "MISC" : "On paternal leave from 2019-01-01 until 2020-01-01"
            }
        }
    }
}
```

Unstructured data

- Unstructured data is data with no pre-defined organizational form or specific format. Or in other words, unstructured data is any data which is not structured or semi-structured.



Difference

	Structured data	Semi-structured data	Unstructured data
What is it?	Data with a high degree of organization, typically stored in a spreadsheet-like manner	Data with some degree of organization	Data with no predefined organizational form and no specific format
To put it simply	Think of a spreadsheet (e.g. Excel) or data in a tabular format	Think of a TXT file with text that has some structure (headers, paragraphs, etc.)	Essentially anything that is not structured or semi-structured data (which is a lot)
Example formats	•Excel spreadsheets •Comma-separated value file (.csv) •Relational database tables	•Hypertext Markup Language (HTML) files •JavaScript Object Notation (JSON) files •Extensible Markup Language (XML) files	•Images such as .jpeg or .png files •Videos such as .mp4 or m4a files •Sound files such as .mp3 or .wav files •Plain text files •Word files •PDF files
Characte- ristics	•Data is structured in a spreadsheet-like manner (e.g. in a table) •Within that table, entries have the same format and a predefined length and follow the same order •Is easily machine-readable and can therefore be analysed without major pre-processing of the data •It is commonly said that around 20% of the world's data is structured	•Data is stored in files that have some degree of organization and structure •Tags or other markers separate elements and enforce hierarchies, but the size of elements can vary and their order is not important •Needs some pre-processing before it can be analysed by a computer •Has gained importance with the emergence of the World Wide Web	•Data that can take any form and thus be stored as any kind of file (formless) •Within that file, there is no structure of content •Typically needs major pre-processing before it can be analysed by a computer, but often easily consumable for humans (e.g. pictures, videos, plain texts) •Most of the data that is created today is unstructured

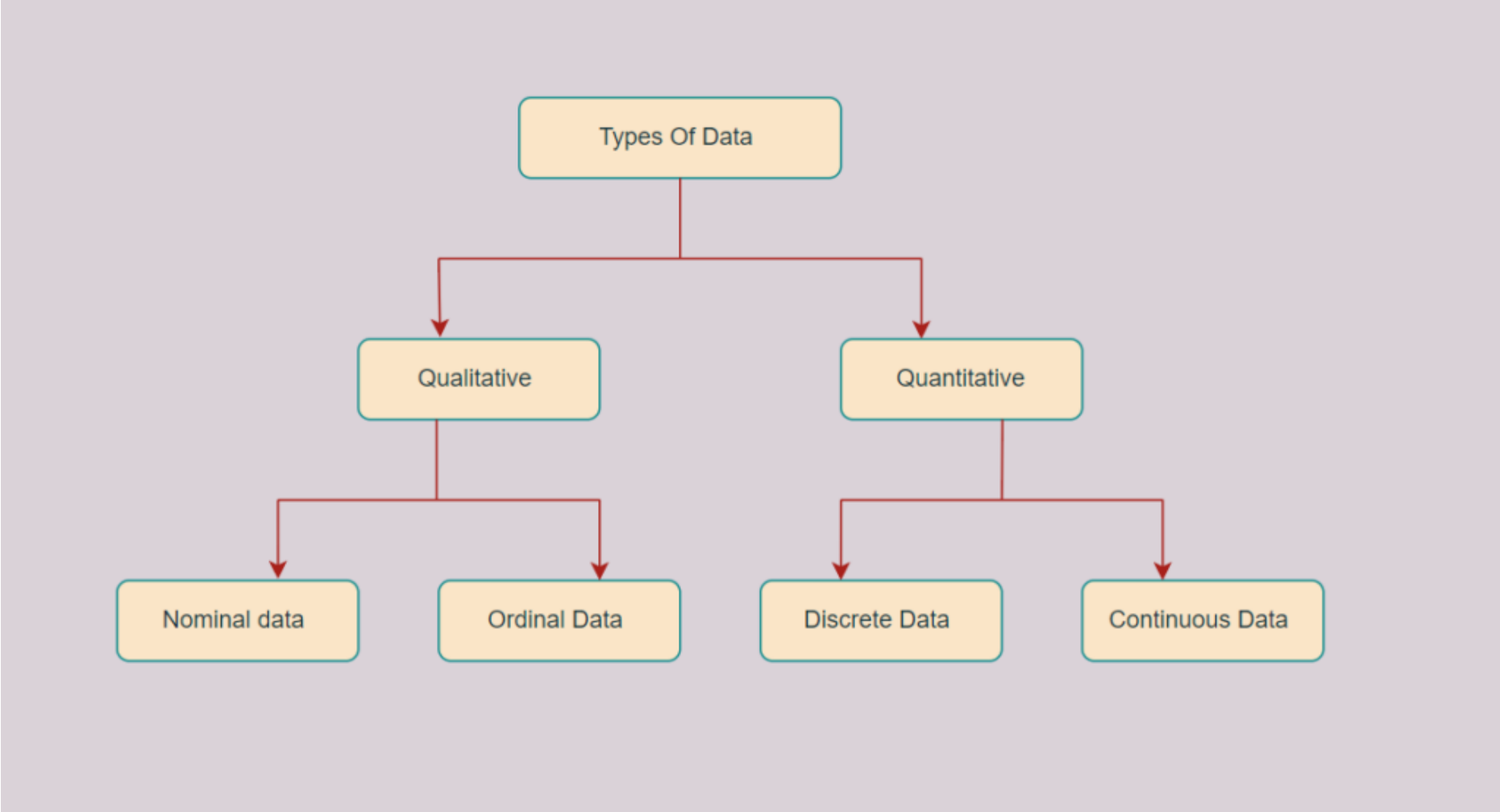
Structured Vs Unstructured Data

Structured vs Unstructured data:

Let us look at a few major differences between structured and unstructured data.

Structured data	Unstructured data
It is a quantitative data	It is a qualitative data
Easy to perform data analysis	Should be processed into proper form before performing analysis
Less than 20% of the total data in the world are in structured form	More than 80% of the total data in the world are in unstructured form
Can be either numeric or text	Can be text, numeric, images, videos, audio, and many more

Discrete, Continuous, Ordinal, Nominal, Qualitative, Quantative , Time series data, Geographical data.



Discrete, Continuous, Ordinal, Nominal, Qualitative, Quantative , Time series data, Geographical data.

Qualitative or Categorical Data

Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers.

These types of data are sorted by category, not by number.

These data consist of audio, images, symbols, or text. The gender of a person, i.e., male, female, or others, is qualitative data.

Qualitative data tells about the perception of people. This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly.

Nominal Data

Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can't be compared with another color.

The name "nominal" comes from the Latin name "nomen," which means "name." With the help of nominal data, we can't do any numerical tasks or can't give any order to sort the data. These data don't have any meaningful order; their values are distributed into distinct categories.

Ordinal Data

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.

Quantitative Data

Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like "how much," "how many," and "how often."

Quantitative data can be used for statistical manipulation. These data can be represented on a wide variety of graphs and charts, such as bar graphs, histograms, scatter plots, boxplots, pie charts, line graphs, etc.

Q1: Discrete Data

The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values.

The discrete data are countable and have finite values; their subdivision is not possible. These data are represented mainly by a bar graph, number line, or frequency table.

Examples of Discrete Data :

Total numbers of students present in a class

Cost of a cell phone

Q2:**Continuous Data**

Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range. **Examples of Continuous Data :** Height of a person

Discrete Vs Continuous Data

Discrete Data	Continuous Data
Discrete data are countable and finite; they are whole numbers or integers	Continuous data are measurable; they are in the form of fractions or decimal
Discrete data are represented mainly by bar graphs	Continuous data are represented in the form of a histogram
The values cannot be divided into subdivisions into smaller pieces	The values can be divided into subdivisions into smaller pieces
Discrete data have spaces between the values	Continuous data are in the form of a continuous sequence
Examples: Total students in a class, number of days in a week, size of a shoe, etc	**Example:** Temperature of room

Time Series Data

Time period	Sales
1/1/2021	124
1/2/2021	153
1/3/2021	130
1/4/2021	158
1/5/2021	120
1/6/2021	174
1/7/2021	155
1/8/2021	146
1/9/2021	126
1/10/2021	145
1/11/2021	155
1/12/2021	172
1/13/2021	198

- When the data is collected over equal intervals of time, it is called as time series data.
- The time intervals can be as small as seconds to as large as years. The only condition is the time-period should be of equal intervals and it should be clearly defined.
- A few examples of time series data are the stock prices of a company throughout the year, the sales data of a company for the past three years, the daily expenses a company has incurred over the past 6 months.
  - Characteristics of a time series data:
    Time series data has 4 characteristics, let us look at them one by one.
    1) Equal interval – As mentioned above, the time series data should be collected in equal interval of time and should not be uneven.
    2) Seasonality – When the time series data is influenced by seasonal factors such as rainfall, temperature, etc.
    3) Cyclicity - When the time series data is rising and falling in intervals those are not at fixed periods such as business cycles.
    4) Trend - When there is a long tern increase or decrease in the data. Based on the movement of data it can be called either an uptrend or a downtrend.

Understanding Data Attribute Types with

- For a customer object, attributes can be customer-id, address, etc
- First step before Data pre-processing - differentiate between different types of attributes and then pre-process the data.

- **Data attribute Types**
  - **Qualitative**
    - **Nominal**
    - **Ordinal**
    - **Binary**
  - **Quantitative**
    - **Numeric**

Qualitative Attributes

1. **Nominal Attributes –**
   - Values are name of things or some kind of symbols
   - Values of Nominal attributes represents some category or state and that's why nominal attribute also referred as **categorical attributes** and
   - There is no order (rank, position) among values of nominal attribute

Attribute	Values
Colours	Black, Brown, White

Qualitative Attributes

2. **Binary Attributes :** Binary data has only 2 values/states.

   For Example yes or no, true or false.

   i. **Symmetric :** Both values are equally important (Gender). No preference on which should be coded as 0 or 1

   ii. **Asymmetric :** Both values are not equally important. Most important outcome is coded as 1

Attribute	Values
Gender	Male , Female

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

Qualitative Attributes

3. **Ordinal Attributes :**
    - Values that have a meaningful sequence or ranking(order) between them
    - But magnitude of values is not actually known

Attribute	Value
Grade	A,B,C,D,E,F

Quantitative Attributes

Numeric :

- A numeric attribute is quantitative because, it is a measurable quantity, represented as integer or real values.
- Numerical attributes are of 2 types, **interval** and **ratio**.

i) Interval-scaled attributes

- Have order
- Values can be added and subtracted but cannot be multiplied or divided
- Eg. Temperature of 10 degree Celsius should not be considered as twice hot as 5 degree Celsius since 10 degree Celsius is 50 degree Fahrenheit and 5 degree Celsius is 41 degree Fahrenheit which isn't twice

Quantitative Attributes(Contd..)

- Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal
- Eg. difference between 100 degrees Fahrenheit and 90 degrees Fahrenheit is the same as 60 degrees Fahrenheit and 70 degrees Fahrenheit.
- Don’t have a true zero

ii) **Ratio-scaled attributes**

- Has all properties of interval-scaled
- Have a true zero
- Values can be added , subtracted , multiplied & divided
- Eg. Weight, height,etc

Geographical data

- Geographical data is a type of information that captures the physical features of a place.
- This can include things like the location of a city or town, the layout of its streets and highways, and its natural features such as rivers and mountains
- Geographical data can be used for a variety of purposes, including mapping out areas for planning purposes.
- Geographical data is often used by social media networks to provide information about users' location.
- Construction companies may use geographical data to map out the locations of roads and highways
- Geographical data is often presented in a two-dimensional format on a paper map or globe. However, geographical data can also be represented in three dimensions through digital means such as virtual globes.
- Geographic information systems (GIS) are software applications that allow users to view, edit, and analyze geographical data in three dimensions.

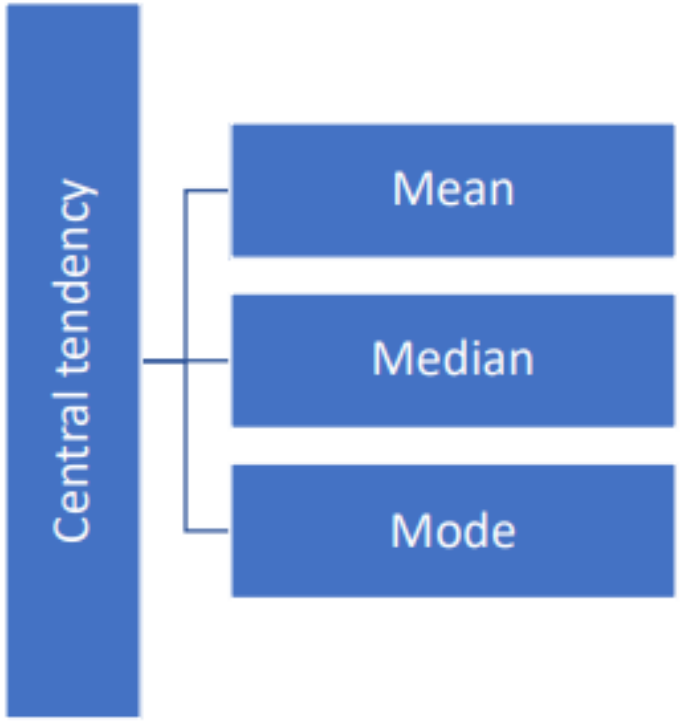
Exercise

- Classify the following attributes as discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

1. Number of telephones in your house
2. Size of French Fries (Medium or Large or X-Large)
3. Ownership of a cell phone
4. Number of local phone calls you made in a month
5. Length of longest phone call
6. Length of your foot
7. Price of your textbook
8. Zip code
9. Temperature in degrees Fahrenheit
10. Temperature in degrees Celsius
11. Temperature in Kelvin

Measures of Central Tendency: Mean, Median, Mode, Mid-range

- Measures of central tendency are used to numerically summarize data by identifying the central position within that set of data. They are also called as summary statistic.



Mean

- Mean: The mean represents the average value of the dataset.
- Mean is calculated by taking the sum of the values and dividing with the numbers of values in a dataset
- The mean (or average) is the most popular and well-known measure of central tendency.
- It can be used with both discrete and continuous data, although its use is most often with continuous data
- An important property of the mean is that it includes every value in your data set as part of the calculation.
- Limitations of mean: The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value

Example

- Ungrouped data: - 0 1 2 3 4 5 6 7 8 9 10 Mean = 5
- For grouped data

Grouped data:

Masses	Frequency(f)	X	FX
40-49	6	40+49/2 = 44.5	267
50-59	8	54.5	436
60-69	12	64.5	774
70-70	14	74.5	1043
80-89	7	84.5	591.5
90-99	3	94.5	283.5
Total	50		3395

Mean= 3395/50
67.9

Median

- Median is the middle value of the dataset in which the dataset is arranged in the ascending order or in descending order.
- When the dataset contains an even number of values, then the median value of the dataset can be found by taking the mean of the middle two values
- Consider the given dataset with the odd number of observations arranged in ascending order 2, 5, 6, 7, 9, 10, 12, 13, 15, 16, 18, 21, 23
- Here 12 is the middle or median number that has 6 values above it and 6 values below it
- Advantages of median: The median is less affected by outliners and skewed data than the mean and usually the preferred measure of central tendency when the distribution is not symmetrical
- Limitations of median: Unsuitable for fractions and percentage , it is not suitable for further algebraic calculation

Mode

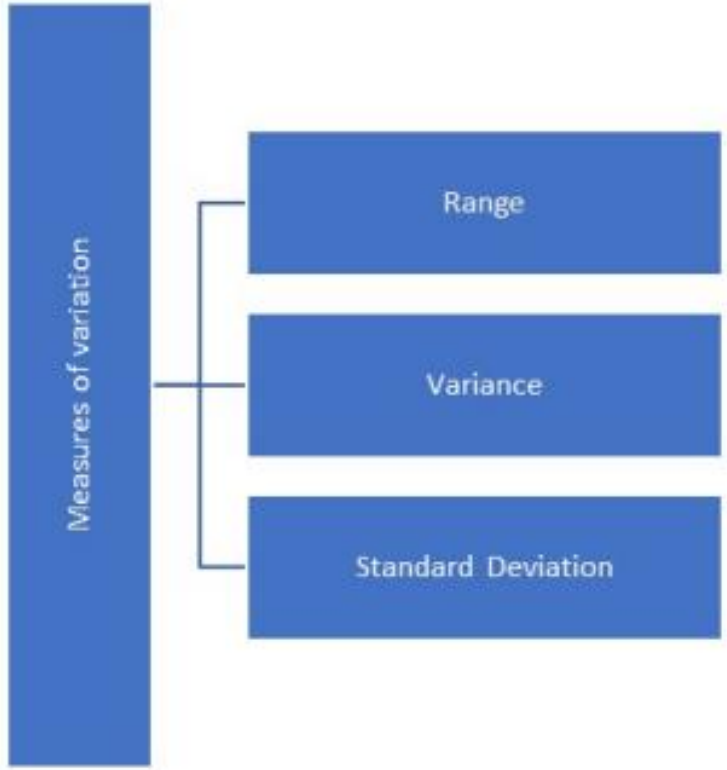
- The mode represents the frequently occurring value in the dataset.
- Sometimes the dataset may contain multiple modes, and in some cases, it does not contain any mode at all.
- Consider the given dataset 5, 4, 2, 3, 2, 1, 5, 4, 5 Since the mode represents the most common value. Hence, the most frequently repeated value in the given dataset is 5.
- Advantages of mode: The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.
- Limitations of mode: In some distributions, the mode may not reflect the center of the distribution very well.

How to select measure

- Based on the properties of the data, the measures of central tendency are selected.
- If you have a symmetrical distribution of continuous data, all the three measures of central tendency hold good. But most of the times, the analyst uses the mean because it involves all the values in the distribution or dataset.
- If you have skewed distribution, the best measure of finding the central tendency is the median.
- If you have the original data, then both the median and mode are the best choice of measuring the central tendency.
- If you have categorical data, the mode is the best choice to find the central tendency

Measures of Dispersion: Range, Variance, Standard Deviation

- Measures of variation give information on the spread or variability or dispersion of the data values
- Finding the central value of the observations is important in this descriptive analysis.
- It is also important to find out the extent of variation near the center.
- There maybe two sets of data that may exhibit same position of the center, but that could differ with respect to their variability.



Range

- Range is referred to as the difference between the highest value and the lowest value in a data set.
- The range is extremely sensitive to outliers and can be a very coarse measurement of the spread of data.
- This means a single data value can greatly affect the analysis of the problem when using range and this can be a limitation to the statisticians who work using range.
- The below image is the pictorial representation of range

Range = Max - Min

Range Example

Example 1:

- Participants in a race had the following finishing times in seconds: 16, 21, 28, 12, 30, 21, and 24. Compute the range.
- The maximum value is 30 and minimum value is 12,
- Hence the range is 30 – 12 = 18.

Example 2:

- The following is the data of measurements taken in meters. Find the range for the
- data. {10.2, 9.5, 12.6, 13, 7.9, 11.7, 8.1, 9, 11.8}.
- The range for the above data will be 13 – 7.9 = 5.1

Standard Deviation:

$$SD = \sqrt{\frac{\sum((x - mean)^2)}{N}}$$

- Standard Deviation: It gives us how spread out the observed values are from the mean. We can calculate the standard deviation using the below formula

  Standard Deviation is always a positive value. Standard deviation is the square root of the variance. Variance is the sum of the squared difference of the observed value from the mean.
  Standard deviation is a measure of the dispersion of a set of data from its mean.
  If the data points are farther from the mean, there is higher deviation within the data set.
  Standard deviation is calculated as the square root of variance by determining the variation between each data point relative to the mean.

Example

Example: Find the standard deviation for the below data;

{2, 4, 5, 7, 9, 11, 12, 14}

Sl	X	$X-\bar{X}$	$(X-\bar{X})^2$
1	2	$2-8=-6$	36
2	4	$4-8=-4$	16
3	5	$5-8=-3$	9
4	7	$7-8=-1$	1
5	9	$9-8=1$	1
6	11	$11-8=3$	9
7	12	$12-8=4$	16
8	14	$14-8=6$	36
			$\sum(X-\bar{X})^2=124$

$$\bar{X}=\frac{2+4+5+7+9+11+12+14}{8}=\frac{64}{8}=8$$

$$SD=\sqrt{\frac{\sum_{i=1}^{8}(Xi-8)^2}{8}}$$

$$SD=\sqrt{\frac{124}{8}}$$

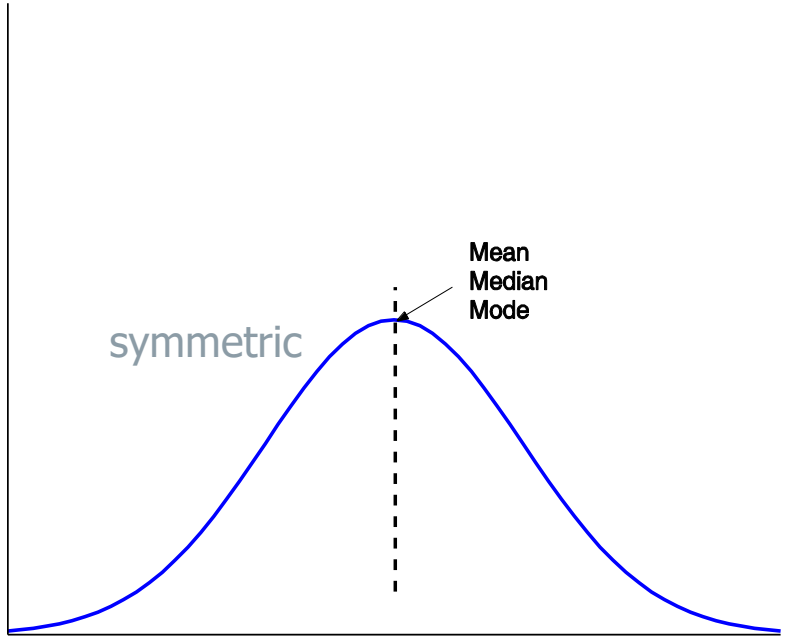
$$SD=3.93$$

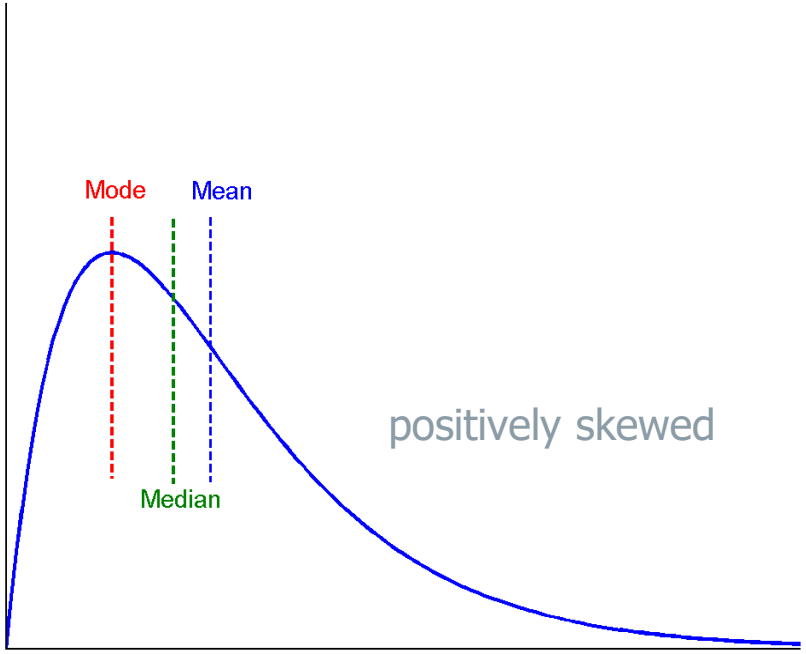
Variance

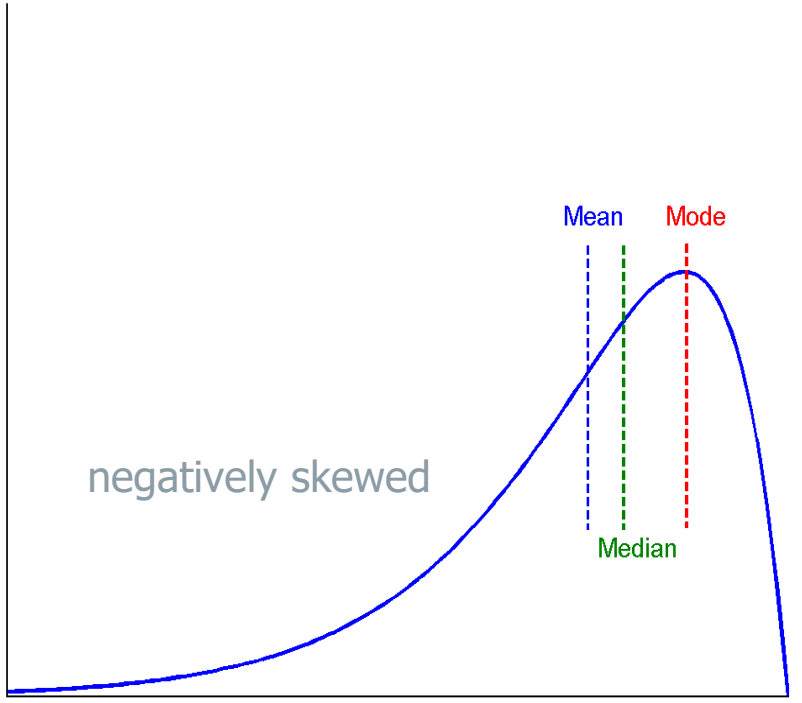
- Variance: The average deviation of all data values from the mean.
- The square of standard deviation is called variance.
- The difference between standard deviation and variance is only in terms of unit of measurement.

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Box Plot : Measuring Dispersion

Quartiles, outliers and boxplots

Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)

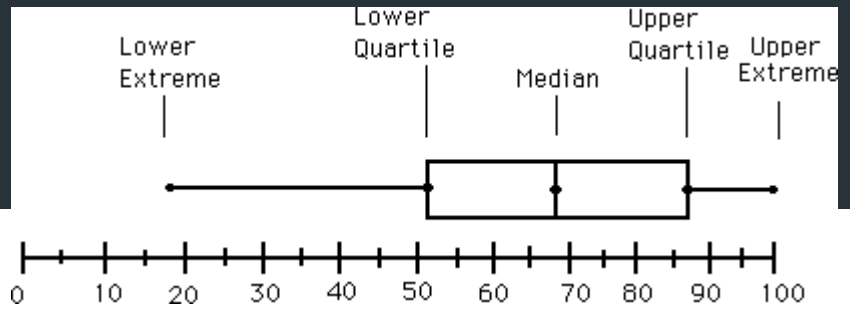
Inter-quartile range: $IQR = Q_3 - Q_1$

Five number summary: min, Q_1, median, Q_3, max

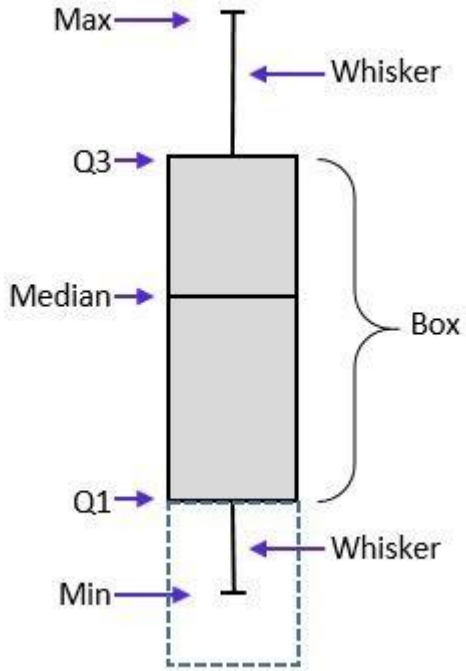
Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually

Outlier: usually, a value higher/lower than 1.5 x IQR

Boxplot Analysis



- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



Box Plot Example

- **Example: Finding the five-number summary**
- A sample of 10 boxes of raisins has these weights (in grams):
- 25, 28, 29, 29, 30, 34, 35, 35, 37, 38
- **Make a box plot of the data.**
- **Step 1:** Order the data from smallest to largest.
- Our data is already in order.
- 25, 28, 29, 29, 30, 34, 35, 35, 37, 38

Step 2

Step 2: Find the median.

The median is the mean of the middle two numbers:

$$25, 28, 29, 29, 30, 34, 35, 35, 37, 38$$

$$\frac{30 + 34}{2} = 32$$

The median is 32.

Step 3

Step 3: Find the quartiles.

The first quartile is the median of the data points to the *left* of the median.

$25, 28, \mathbf{29}, 29, 30$

$Q_1 = 29$

The third quartile is the median of the data points to the *right* of the median.

$34, 35, \mathbf{35}, 37, 38$

$Q_3 = 35$

Step 4

Step 4: Complete the five-number summary by finding the min and the max.

The min is the smallest data point, which is 25.

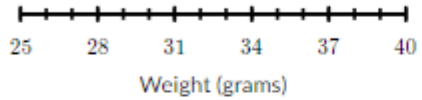
The max is the largest data point, which is 38.

The five-number summary is $25, 29, 32, 35, 38$.

Step 5

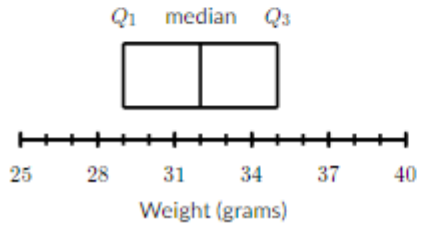
Let's make a box plot for the same dataset from above.

Step 1: Scale and label an axis that fits the five-number summary.



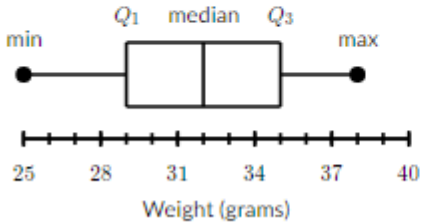
Step 2: Draw a box from Q_1 to Q_3 with a vertical line through the median.

Recall that $Q_1 = 29$, the median is 32, and $Q_3 = 35$.



Step 3: Draw a whisker from Q_1 to the min and from Q_3 to the max.

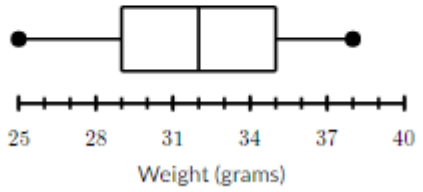
Recall that the min is 25 and the max is 38.



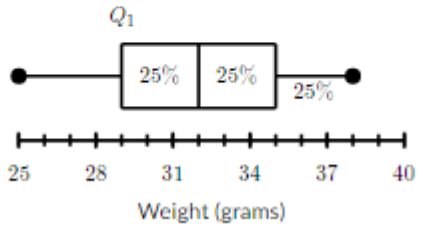
Interpreting Boxplot

Example: Interpreting quartiles

About what percent of the boxes of raisins weighed more than 29 grams?



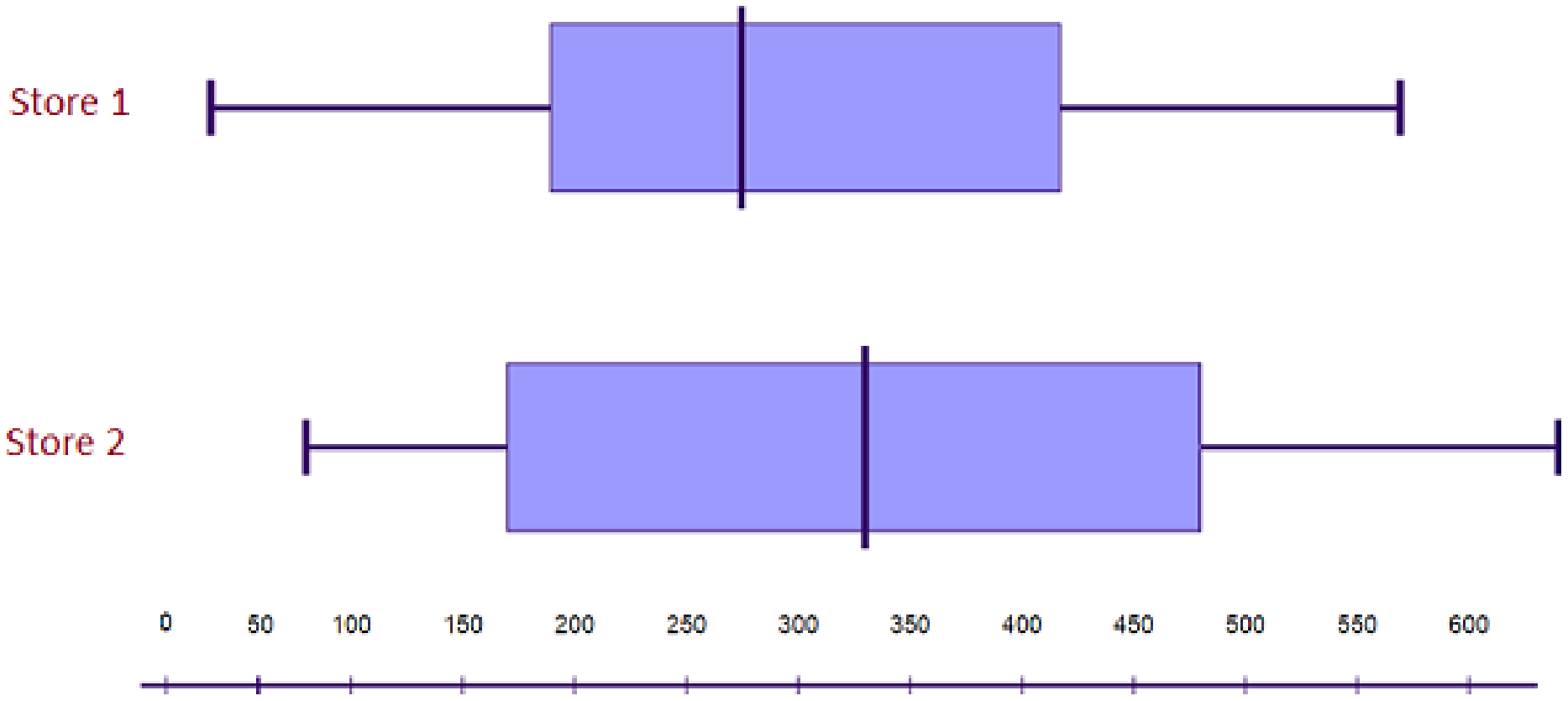
Since $Q_1 = 29$, about 25% of data is lower than 29 and about 75% is above is 29.



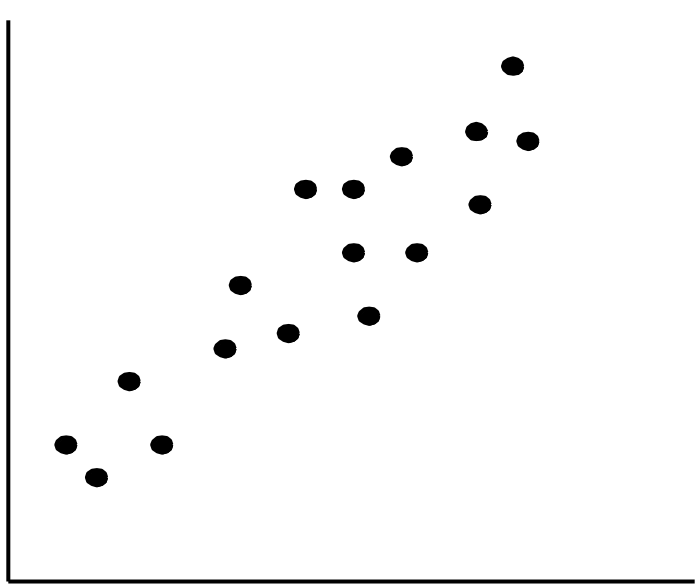
Solve :Draw Box Plot with 5 no summary. Compare Performance of Two Store using Box

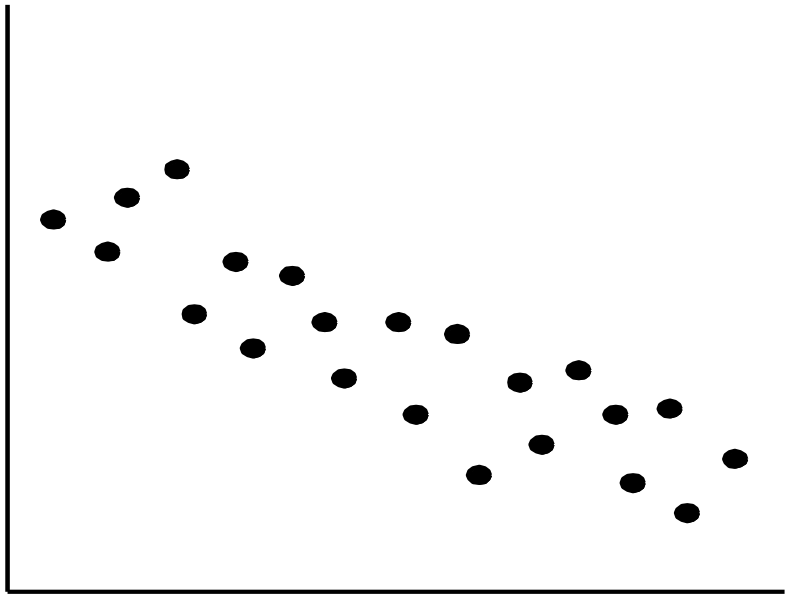
- **Store 1**:
- 350, 460, 20, 160, 580, 250, 210, 120, 200, 510, 290, 380.
- **Store 2:**
- 520, 180, 260, 380, 80, 500, 630, 420, 210, 70, 440, 140.

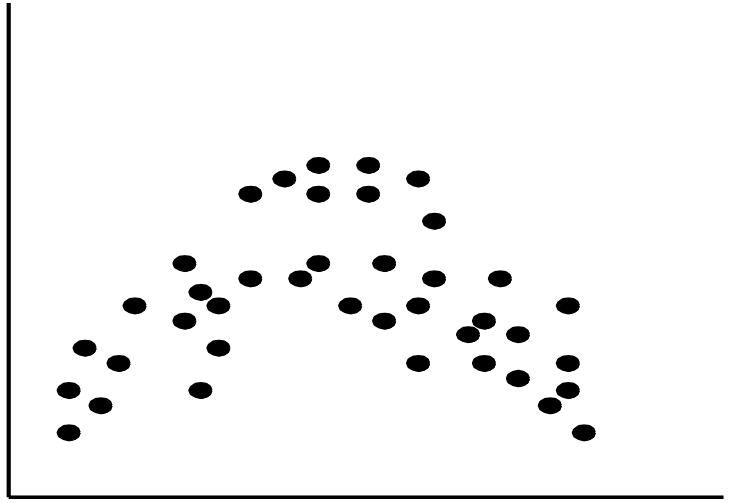
Comment on performance by interpreting box-plot



Positively and Negatively Correlated Data

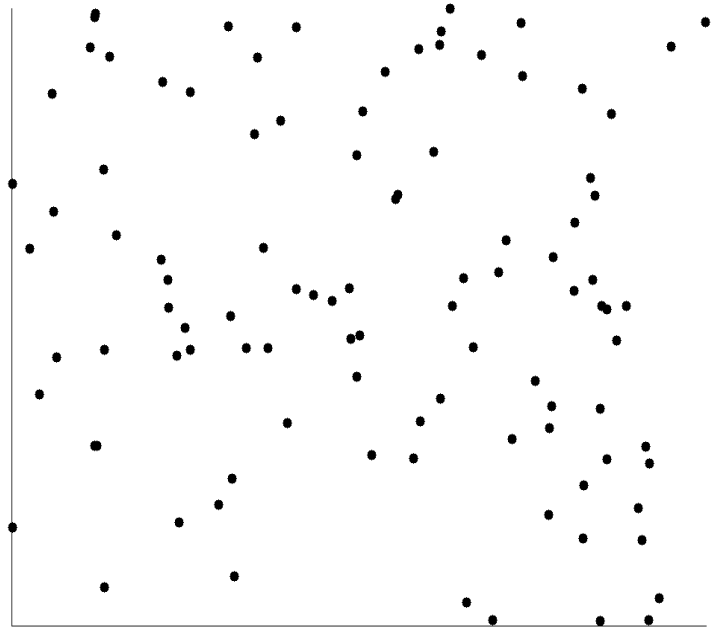


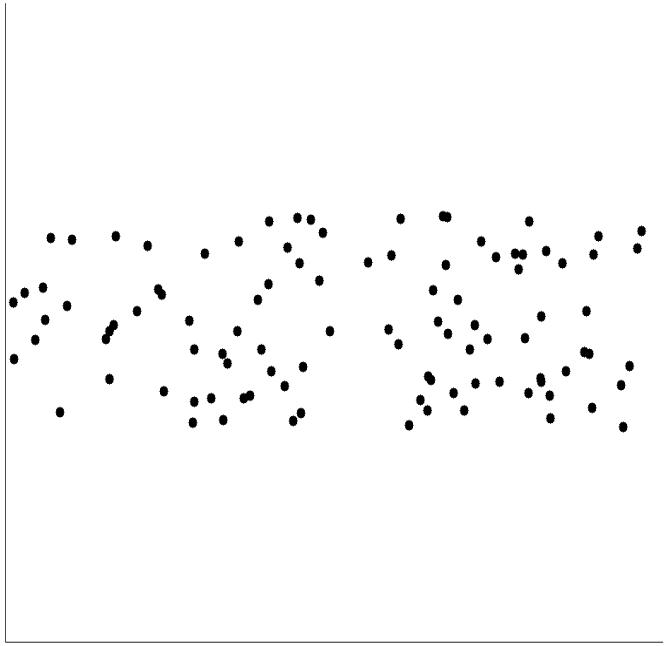


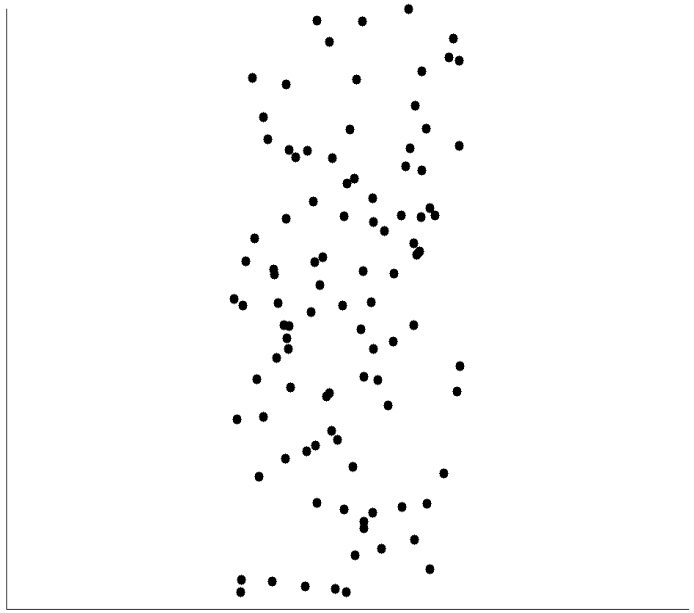


- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data







Reference

- Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, and Jian Pei, 3rd edition