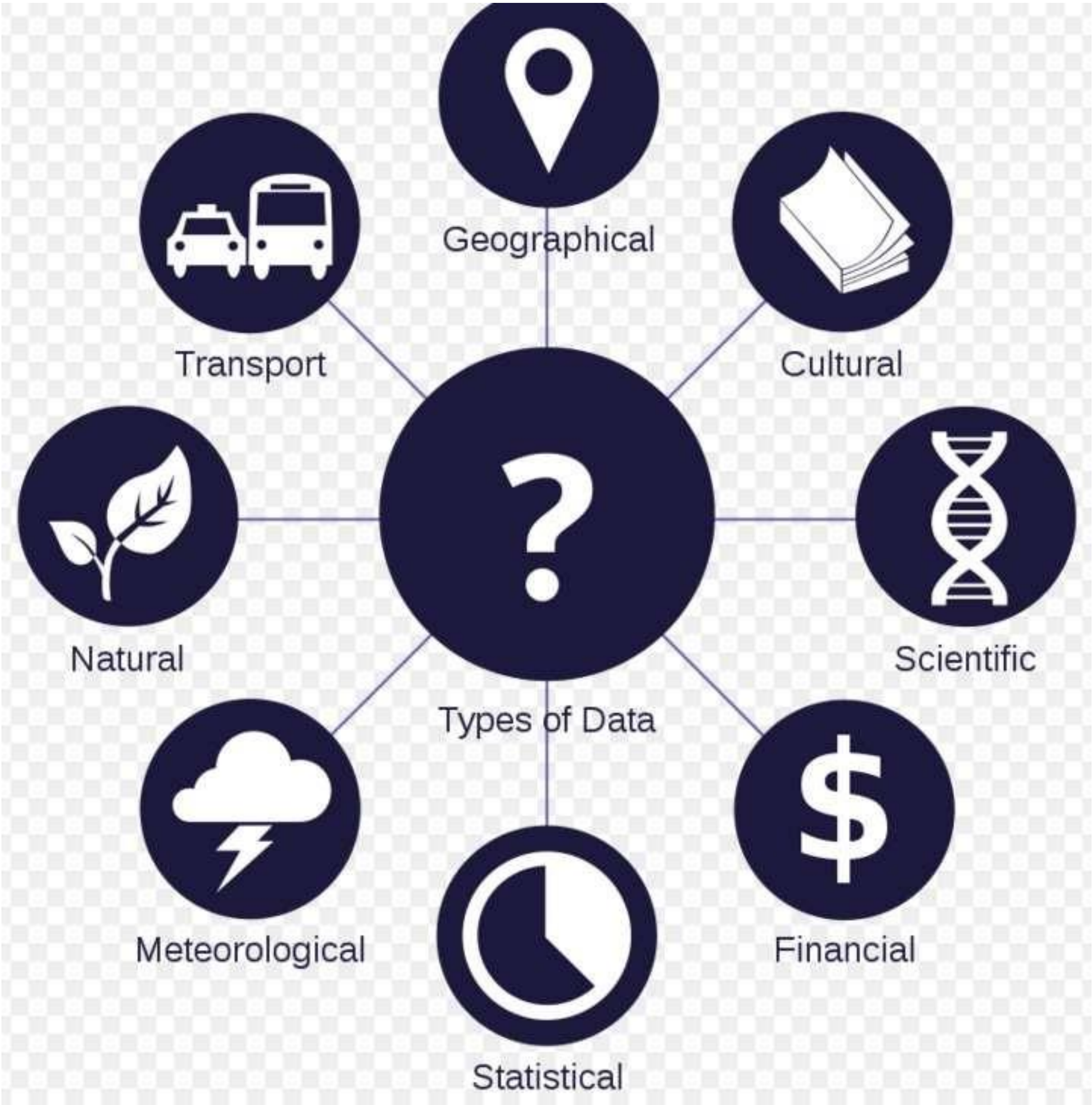
Data and Types of Data in Data Science

Data



Geographical
Transport
Cultural
Natural
Scientific
Types of Data
Meteorological
Financial
Statistical

What is Data?

- Data are the actual pieces of information that you collect through your study and used for the purpose of analysis.
- Data refers to a set of values, which are usually organized by variables
- It is the raw information from which statistics are created.
- For example, if you ask five of your friends how many pets they own, they might give you the following data: 0, 2, 1, 4, 18.
- Not all data are numbers; let’s say you also record the gender of each of your friends, getting the following data: male, male, female, male, female.
- Most data fall into one of two groups OR Variables are of different types and can be classified in many ways: numerical or categorical

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Numerical
  - Categorical

Numerical Data

- Numerical data is information that is measurable, and it is, of course, data represented as numbers and not words or text.
- These data have meaning as a measurement, such as a person's height, weight, IQ, or blood pressure; or they're a count, such as the number of stock shares a person owns, how many teeth a dog has, or how many pages you can read of your favourite book before you fall a sleep.
- Numerical data can be further broken into two types:

➢ Discrete

➢ Continuous.

Numerical Data

➢ **Discrete Data:** Discrete data is a numerical type of data that includes whole, concrete numbers with specific and fixed data values determined by counting.

- They have a logical end to them.
- Discrete data represent items that can be counted
- Finite number of possible values
- They take on possible values that can be listed out. The list of possible values may be fixed (also called finite); or it may go from 0, 1, 2, on to infinity.
- Examples: number of people in a room, number of items in a basket, number of hours in a day, money

Numerical Data

- **Continuous Data:** Continuous data includes complex numbers and varying data values that are measured over a specific time interval.
- Numbers that don’t have a logical end to them.
- Continuous data represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line.
- Infinite number of possible values.
- For example, the exact amount of gas purchased at the pump for cars with 20-gallon tanks would be continuous data from 0 gallons to 20 gallons, represented by the interval [0, 20], inclusive. You might pump 8.40 gallons, or 8.41, or 8.414863 gallons, or any possible number from 0 to 20.

Categorical Data

- Categorical data, this is any data that isn’t a number, which can mean a string of text or date.
- It describes an event using a string of words rather than numbers.
- Categorical data represent characteristics such as a person’s gender, marital status, hometown, or the types of movies they like. Categorical data can take on numerical values (such as “1” indicating male and “2” indicating female), but those numbers don’t have mathematical meaning.
- These variables can be broken down into nominal, ordinal values.

Nominal Data

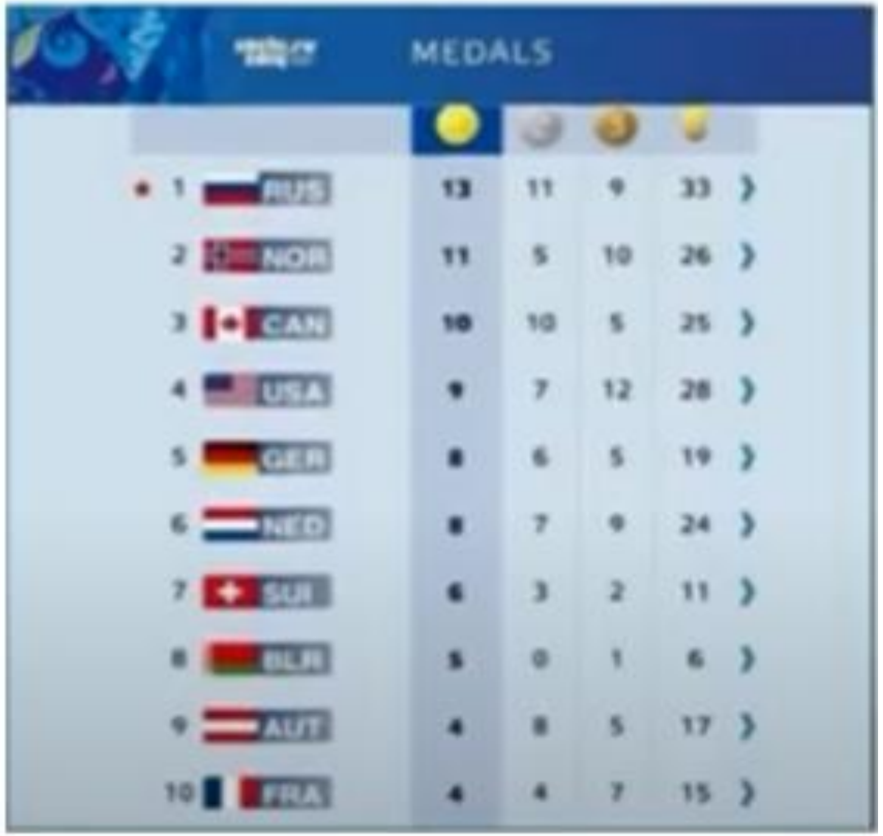
- There is no order at all: each category has its unique meaning
- We can only count but can not order or measure nominal data
- A nominal scale describes a variable with categories that do not have a natural order or ranking
- Examples: Names of cars, book titles in a library, marital status

Ordinal Data

- If there is a sense of order there, the variables are called ordinal.
- An ordinal scale is one where the **order matters** but **not the difference between values**.
- Examples of ordinal variables include:
  - socio economic status (“low income”,”middle income”,”high income”),
  - education level (“high school”,”BS”,”MS”,”PhD”),
  - income level (“less than 50K”, “50K-100K”, “over 100K”),
  - satisfaction rating (“extremely dislike”, “dislike”, “neutral”, “like”, “extremely like”).

Ordinal Data

- Note the differences between adjacent categories do not necessarily have the same meaning. For example, the difference between the two income levels “less than 50K” and “50K-100K” does not have the same meaning as the difference between the two income levels “50K-100K” and “over 100K”.



MEDALS

1	RUS	13	11	9	33
2	NOR	11	5	10	26
3	CAN	10	10	5	25
4	USA	9	7	12	28
5	GER	8	6	5	19
6	NED	8	7	9	24
7	SUI	6	3	2	11
8	BLR	5	0	1	6
9	AUT	4	8	5	17
10	FRA	4	4	7	15

Binary Data

- Binary data has only two possible states:

➢ Yes or No

➢ 0 or 1

➢ True or False

- Binary data is used heavily for classification machine learning models.
- Examples of binary variables can include whether a person has stopped their subscription service or not, or if a person bought a car or not, Toss of a coin, Switch On or Off

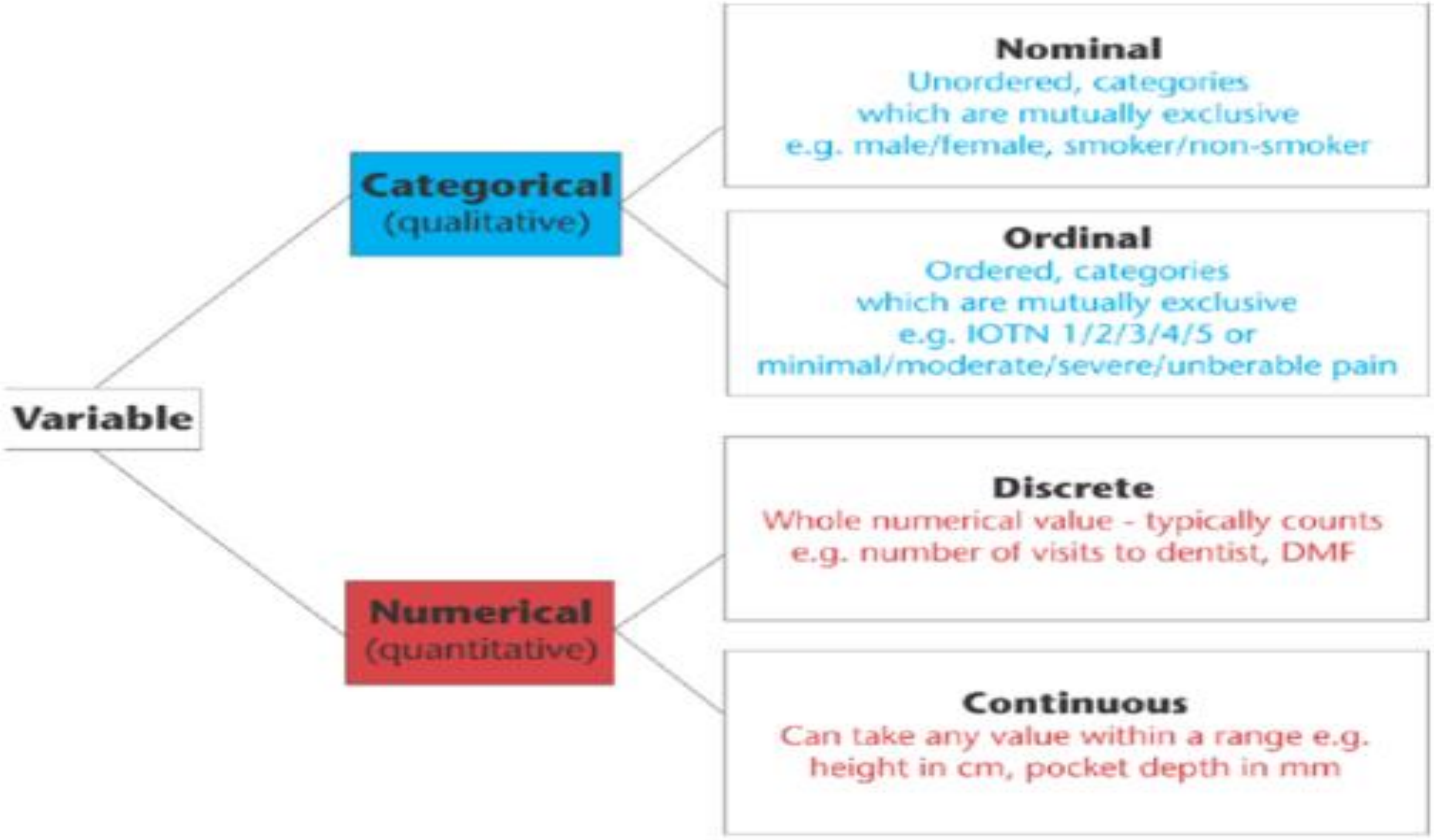
Interval Data

- An interval scale is one where there is **order and the difference between two values is meaningful**.
- Examples of interval variables include:
  - The temperature measured in Celsius, the difference in temperature between 50-60 degrees is the same as the difference in temperature between 80-90 degrees.
  - pH, SAT score (200-800), credit score (300-850), year.
- Interval data doesn’t have ‘absolute zero’ value. The zero points, in this case, are arbitrary as there can be a temperature below zero degrees Celsius.
- The mathematical operations such as addition, subtraction, mean, median, mode and standard deviation can be calculated.

Ratio Variable

- ➢ A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable.
- ➢ It has an absolute zero value.
- ➢ For example, if you measure temperature in degrees Kelvin then it is considered ratio data. This is because zero points are absolute as there can’t be temperatures below zero degrees Kelvin. Ratio data doesn’t have any negative numerical value. For example, height can’t be negative.
- Other Examples: enzyme activity, dose amount, reaction rate, flow rate, concentration, pulse, weight, length, survival time.

Qualitative (Categorical) vs Quantitative (Numerical)



Variable

Categorical (qualitative)

Nominal
Unordered, categories which are mutually exclusive e.g. male/female, smoker/non-smoker

Ordinal
Ordered, categories which are mutually exclusive e.g. IOTN 1/2/3/4/5 or minimal/moderate/severe/unberable pain

Numerical (quantitative)

Discrete
Whole numerical value - typically counts e.g. number of visits to dentist, DMF

Continuous
Can take any value within a range e.g. height in cm, pocket depth in mm

Qualitative (Categorical) vs Quantitative (Numerical)

- There are other ways of classifying variables that are common in statistics. One is qualitative vs. quantitative.
- Qualitative variables are descriptive/categorical. Qualitative data is data concerned with descriptions, which can be observed but cannot be computed. Many statistics, such as mean and standard deviation, do not make sense to compute with qualitative variables.
- Quantitative variables have numeric meaning, so statistics like means and standard deviations make sense.

Qualitative Data

- Qualitative data is defined as the data that approximates and characterizes.
- Qualitative data can be observed and recorded.
- This data type is non-numerical in nature.
- This type of data is collected through methods of observations, one-to-one interviews, conducting focus groups, and similar methods.
- Qualitative data in statistics is also known as categorical data – data that can be arranged categorically based on the attributes and properties of a thing or a phenomenon.

Qualitative Data

- For example, think of a student reading a paragraph from a book during one of the class sessions. A teacher who is listening to the reading gives feedback on how the child read that paragraph. If the teacher gives feedback based on fluency, intonation, throw of words, clarity in pronunciation without giving a grade to the child, this is considered as an example of qualitative data.

➢ The cake is orange, blue, and black in colour

➢ Females have brown, black, and red hair

- Qualitative data does not include numbers in its definition of traits

Quantitative Data

- Quantitative data is all about numbers.
- Quantitative data is any quantifiable information that can be used for mathematical calculation or statistical analysis.
- This form of data helps in making real-life decisions based on mathematical derivations.
- Quantitative data is used to answer questions like how many? How often? How much? This data can be validated and verified.

➢ Example: There are four cakes and three muffins kept in the basket

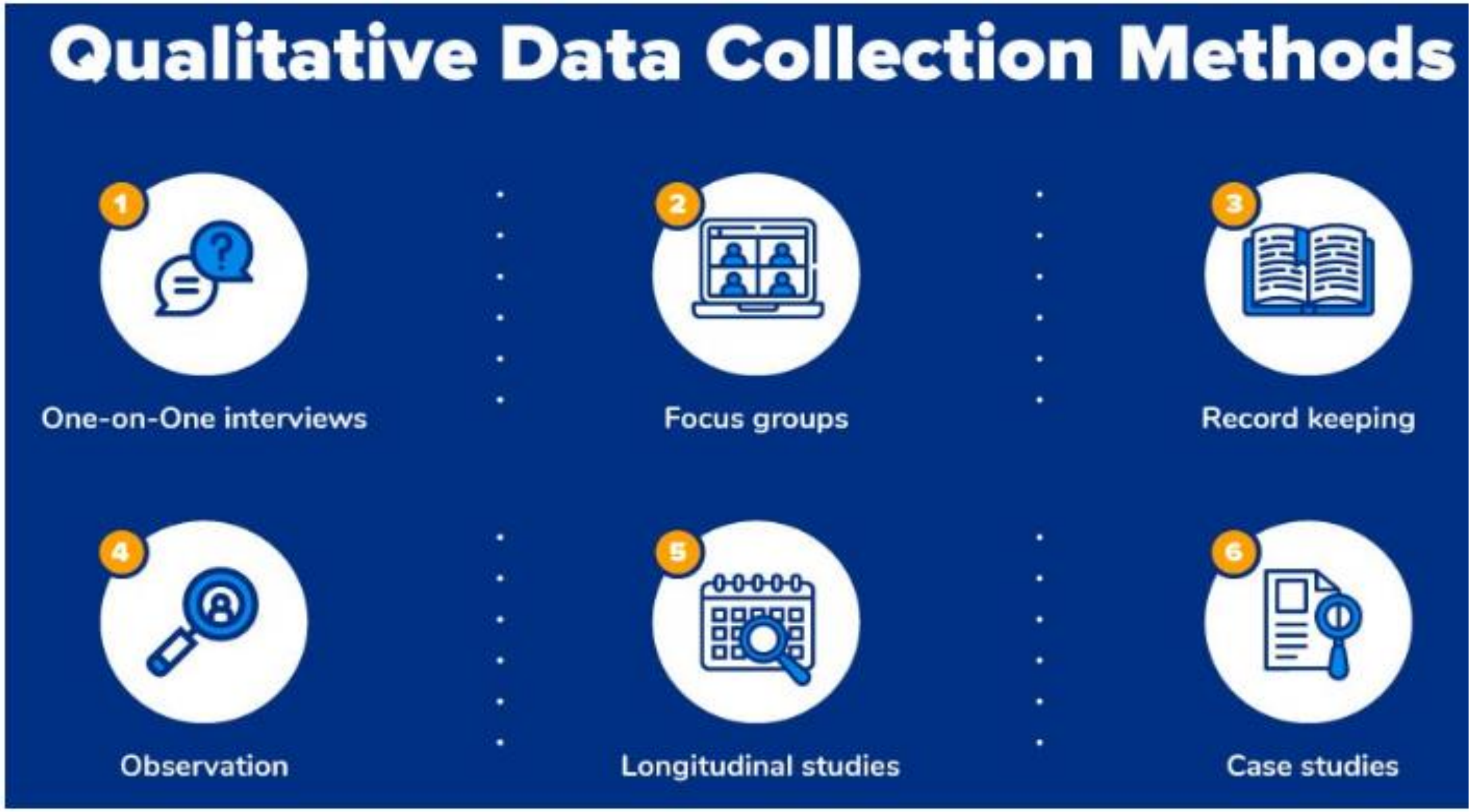
➢ Example: One glass of fizzy drink has 97.5 calories

Key Differences (Quantitative vs Qualitative Data)

Quantitative Data	Qualitative Data
These are data that deal with quantities, values, or numbers.	These data, on the other hand, deals with quality.
Measurable.	They are generally not measurable.
Expressed in numerical form.	They are descriptive rather than numerical in nature.
Conclusive	Exploratory
Measures quantities such as length, size, amount, price, and even duration.	Narratives often make use of adjectives and other descriptive words to refer to data on appearance, color, texture, and other qualities.
Approach is Objective	Approach is Subjective
Determines Level of occurrence	Determines Depth of understanding
Data Collection Techniques: Quantitative surveys, Interviews, Experiments	Data Collection Techniques: Qualitative Survey, Focus group méthode, Documental revision, etc.
Data Structure: Structured	Data Structure: Unstructured

Qualitative Data Collection Methods

- Qualitative Data Collection Methods



Qualitative Data Collection Methods

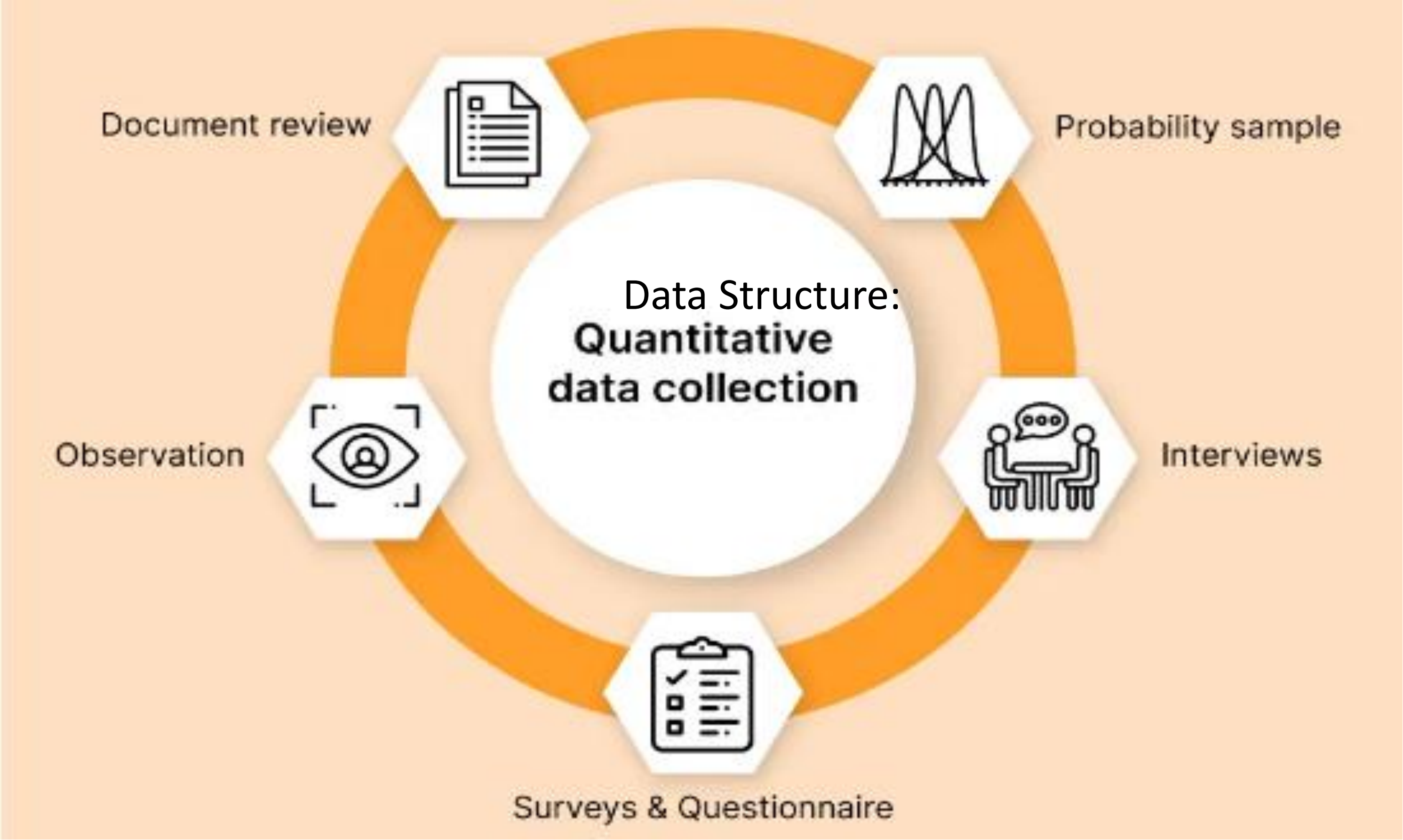
1. One-on-One interviews
2. Focus groups
3. Record keeping
4. Observation
5. Longitudinal studies
6. Case studies

Qualitative Data Collection Methods

- Exploratory in nature, these methods are mainly concerned at gaining insights and understanding of underlying reasons and motivations, so they tend to dig deeper.
- Since they cannot be quantified, measurability becomes an issue.
- This lack of measurability leads to the preference for methods or tools that are largely unstructured or, in some cases, maybe structured but only to a very small, limited extent.
- Generally, qualitative methods are **time-consuming and expensive** to conduct, and so researchers try to lower the costs incurred by decreasing the sample size or number of respondents.

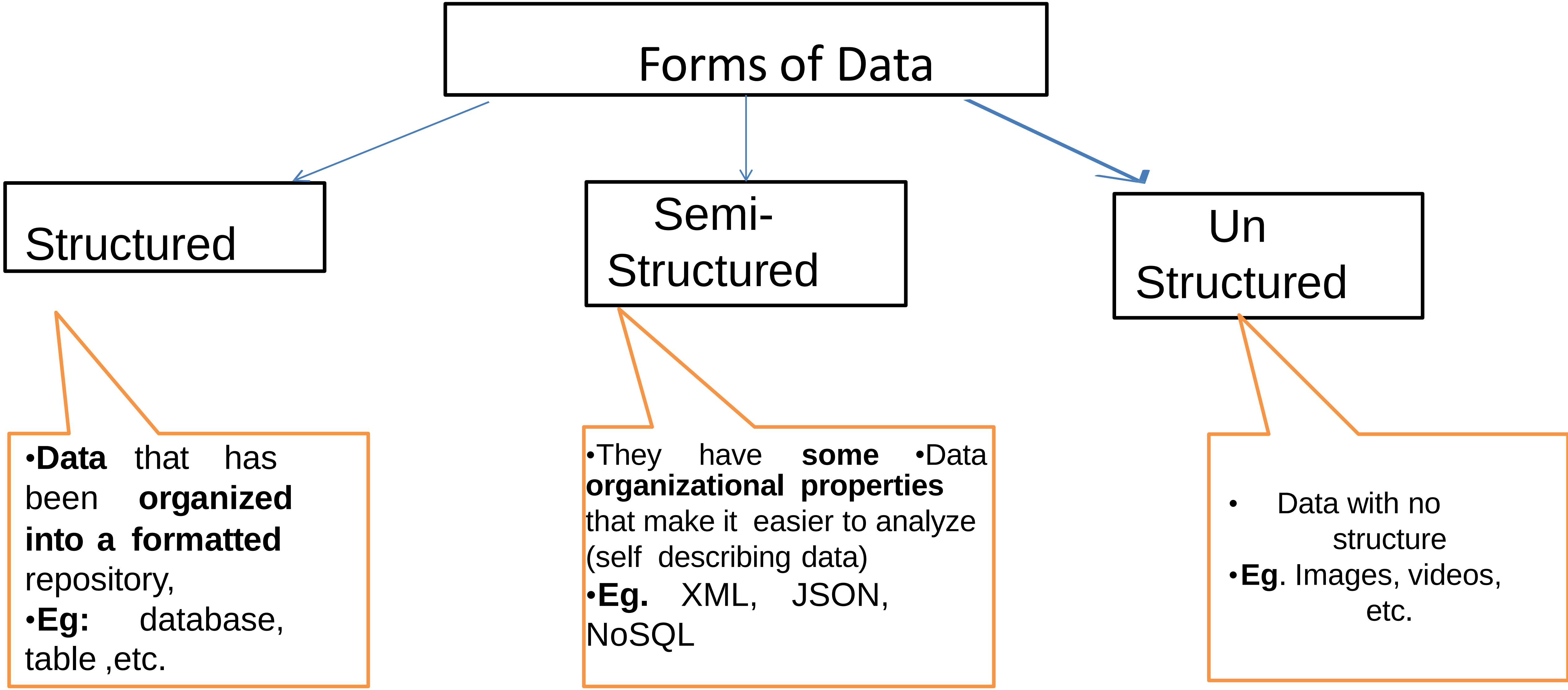
Quantitative Data Collection Methods

- Qualitative Data Collection Methods



Quantitative Data Collection Methods

- Data can be readily quantified and generated into numerical form, which will then be converted and processed into useful information mathematically.
- The result is often in the form of statistics that is meaningful and, therefore, useful.
- Unlike qualitative methods, these quantitative techniques usually make use of larger sample sizes because its measurable nature makes that possible and easier.
- Example: Administrative data collection->financial data, performance data, resource allocation etc.



Forms of Data

Structured

- **Data** that has been **organized into a formatted** repository,
- **Eg:** database, table ,etc.

Semi-Structured

•They have **some** •Data **organizational properties** that make it easier to analyze (self describing data)

•**Eg.** XML, JSON, NoSQL

Un Structured

- Data with no structure
- **Eg**. Images, videos, etc.

Structured Data

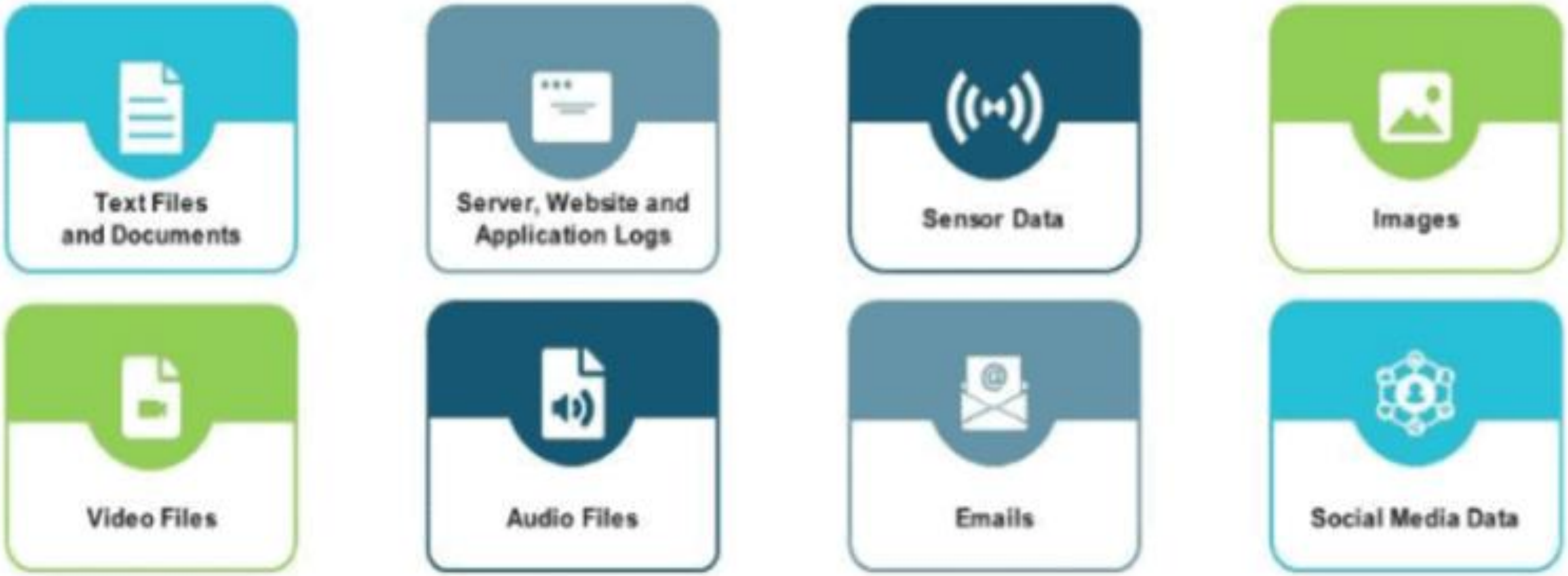
- **Structured data** is generally tabular data that is represented by columns and rows in a database.
- Databases that hold tables in this form are called ***relational databases****.*
- The mathematical term “*relation*” specify to a formed set of data held as a table.
- In structured data, all row in a table has the same set of columns.
- SQL (Structured Query Language) programming language used for structured data.

Semi-structured Data

- Semi-structured data is information that doesn’t consist of Structured data (relational database) but still has some structure to it.
- Semi-structured data consist of documents held in JavaScript Object Notation (JSON) format. It also includes key-value stores and graph databases.

Unstructured Data

- **Unstructured data** is information that either does not organize in a pre-defined manner or not have a pre-defined data model.
- Unstructured information is a set of text-heavy but may contain data such as numbers, dates, and facts as well.
- **Videos, audio, and binary** data files might not have a specific structure. They're assigned to as **unstructured** data.



Acquire Data

Data Acquisition

- What is data acquisition or data collection?
- What kind of data do we need?
- Where you can find these data sets?
- What are some common methods of accessing this data?

Data Acquisition

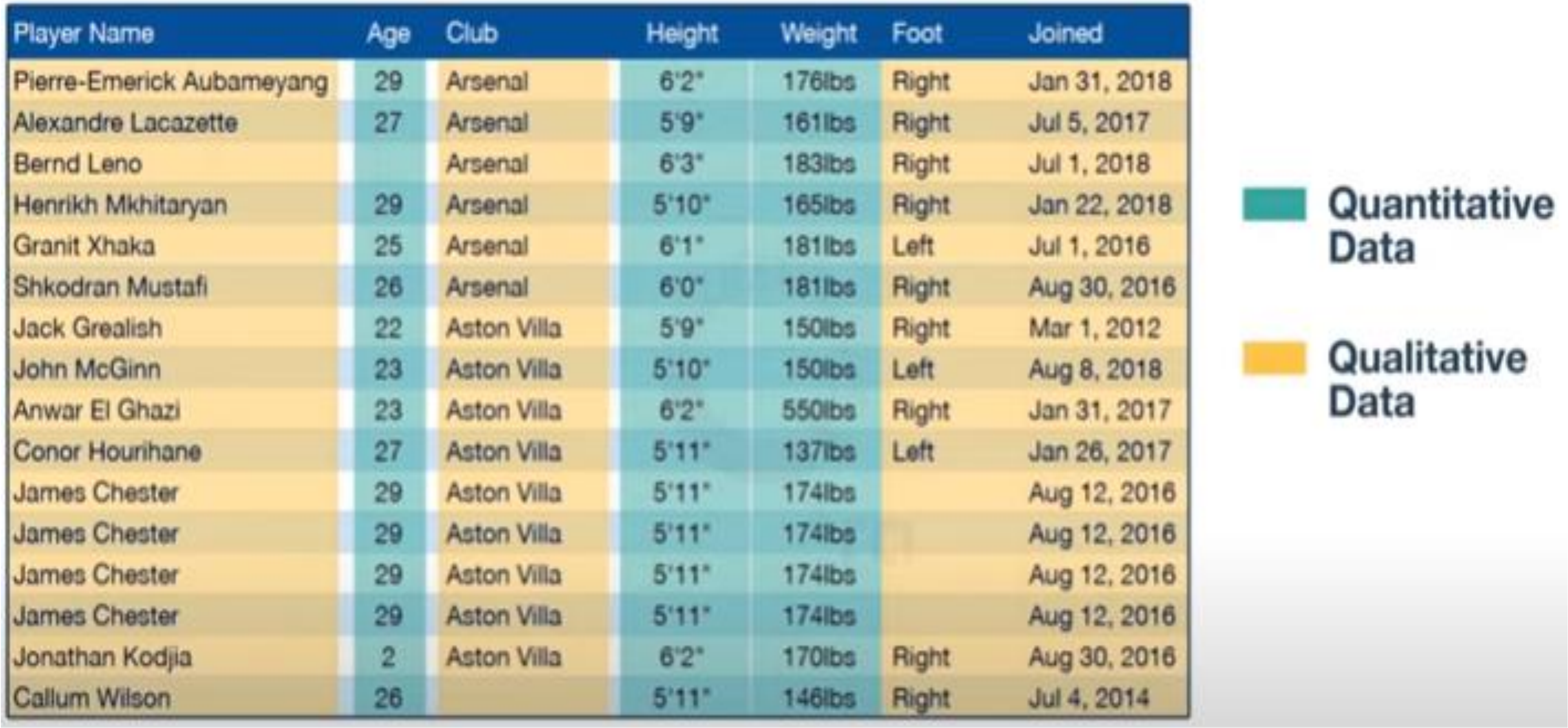
- Data collection is a systematic approach to gather relevant information from a variety of sources.
- Depending on the problem statements, the data collection method is broadly classified into two categories.

➢ Primary data collection

➢ Secondary data collection

Sample Data

- For Example, we have collected and aggregated the data from various open-source websites such as Github, Kaggle, and datahub. A snapshot of the data collected is shown on the screen.

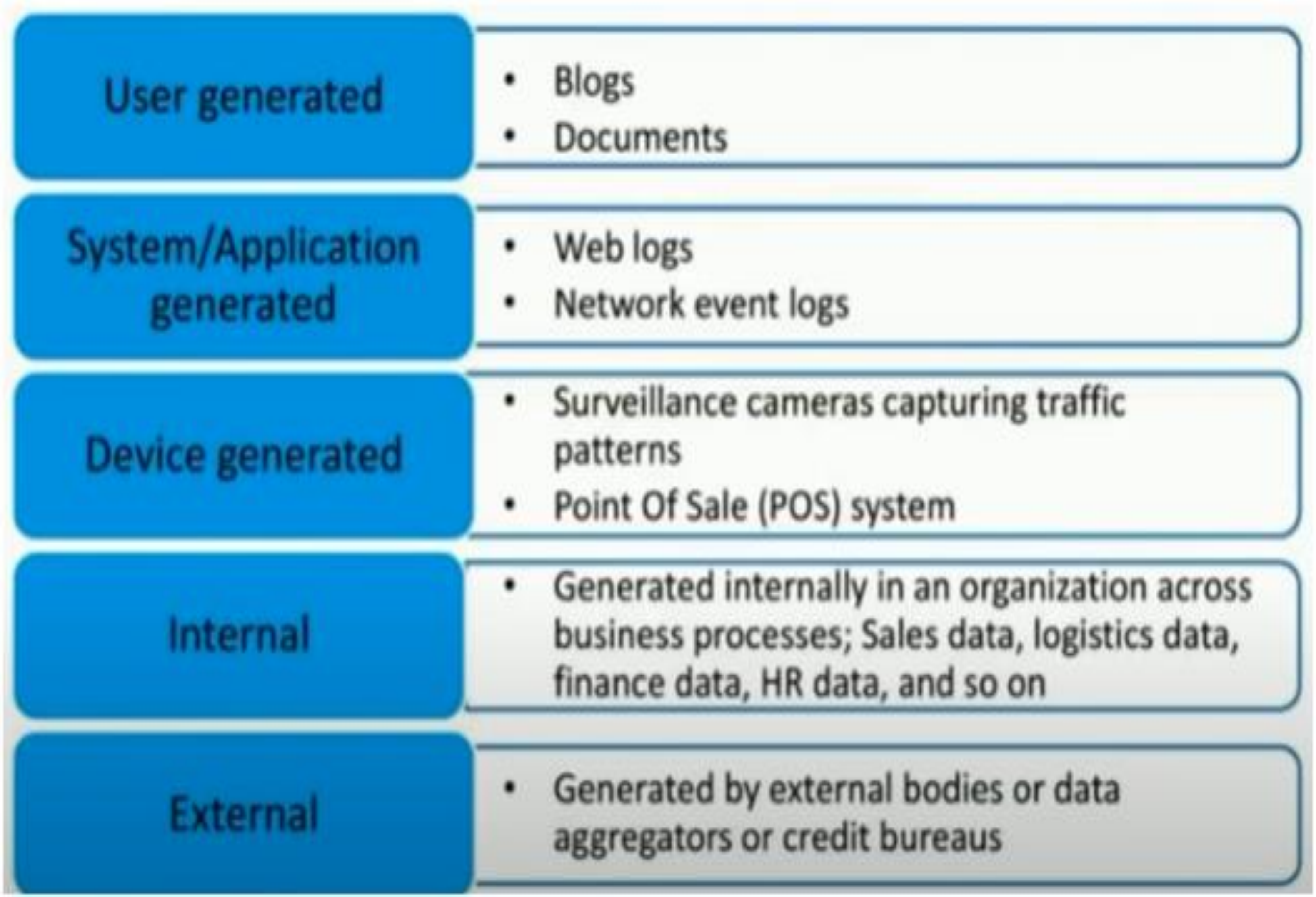


Player Name	Age	Club	Height	Weight	Foot	Joined
Pierre-Emerick Aubameyang	29	Arsenal	6'2"	176lbs	Right	Jan 31, 2018
Alexandre Lacazette	27	Arsenal	5'9"	161lbs	Right	Jul 5, 2017
Bernd Leno		Arsenal	6'3"	183lbs	Right	Jul 1, 2018
Henrikh Mkhitaryan	29	Arsenal	5'10"	165lbs	Right	Jan 22, 2018
Granit Xhaka	25	Arsenal	6'1"	181lbs	Left	Jul 1, 2016
Shkodran Mustafi	26	Arsenal	6'0"	181lbs	Right	Aug 30, 2016
Jack Grealish	22	Aston Villa	5'9"	150lbs	Right	Mar 1, 2012
John McGinn	23	Aston Villa	5'10"	150lbs	Left	Aug 8, 2018
Anwar El Ghazi	23	Aston Villa	6'2"	550lbs	Right	Jan 31, 2017
Conor Hourihane	27	Aston Villa	5'11"	137lbs	Left	Jan 26, 2017
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
James Chester	29	Aston Villa	5'11"	174lbs		Aug 12, 2016
Jonathan Kodjia	2	Aston Villa	6'2"	170lbs	Right	Aug 30, 2016
Callum Wilson	26		5'11"	146lbs	Right	Jul 4, 2014

🟩 Quantitative Data

🟨 Qualitative Data

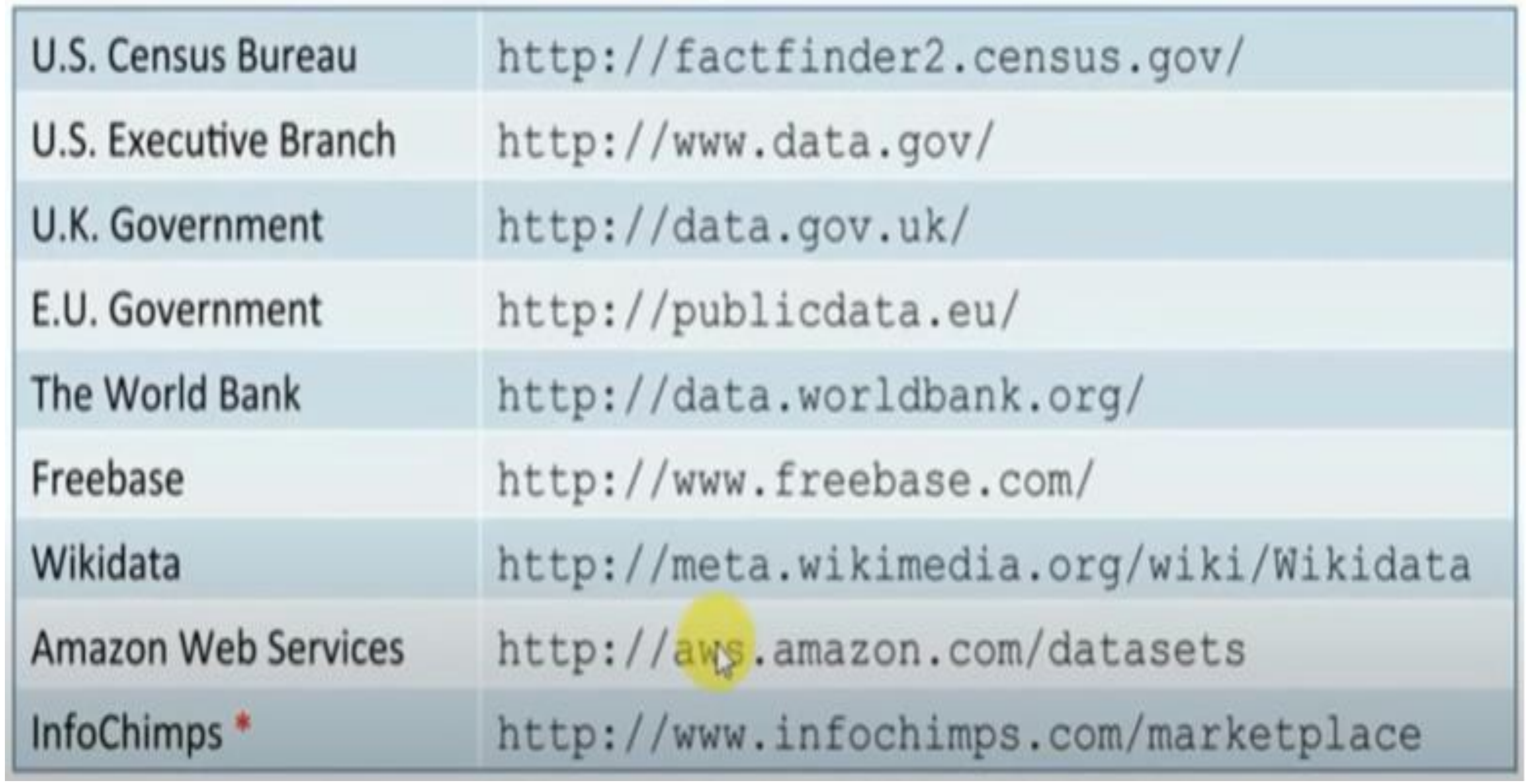
Sources of Data



User generated	• Blogs • Documents
System/Application generated	• Web logs • Network event logs
Device generated	• Surveillance cameras capturing traffic patterns • Point Of Sale (POS) system
Internal	• Generated internally in an organization across business processes; Sales data, logistics data, finance data, HR data, and so on
External	• Generated by external bodies or data aggregators or credit bureaus

Freely available data sources

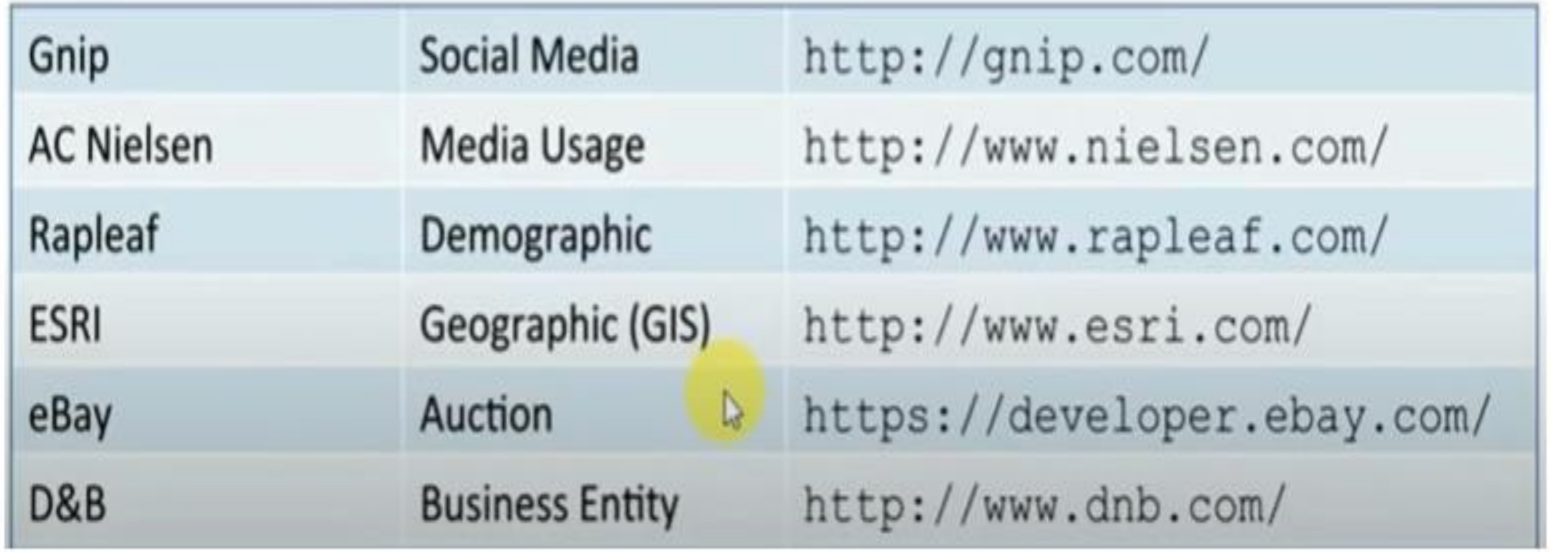
- There are many sources of data available at no cost
  - ➢ Some are public domain and some are copyrighted
  - ➢ Be sure to check the license to verify that your use is allowed



U.S. Census Bureau	http://factfinder2.census.gov/
U.S. Executive Branch	http://www.data.gov/
U.K. Government	http://data.gov.uk/
E.U. Government	http://publicdata.eu/
The World Bank	http://data.worldbank.org/
Freebase	http://www.freebase.com/
Wikidata	http://meta.wikimedia.org/wiki/Wikidata
Amazon Web Services	http://aws.amazon.com/datasets
InfoChimps *	http://www.infochimps.com/marketplace

Commercial data sources

- Many companies also offer data
  - ➢ Usually for a fee, but sometimes available at no cost
  - ➢ Always be sure to check the license terms



Gnip	Social Media	http://gnip.com/
AC Nielsen	Media Usage	http://www.nielsen.com/
Rapleaf	Demographic	http://www.rapleaf.com/
ESRI	Geographic (GIS)	http://www.esri.com/
eBay	Auction	https://developer.ebay.com/
D&B	Business Entity	http://www.dnb.com/

Acquition techniques

- Data, comes from, many places, local and remote, in many varieties, structured and un-structured. And, with different velocities.
- There are many techniques and technologies to access these different types of data.
- For examples: A lot of data exists in conventional **relational databases**, like structure big data from organizations.
- The **tool** of choice to access data from databases is structured query language or **SQL**, which is supported by all relational databases management systems.
- Data can also exist in files such as text files and Excel spread sheets. **Scripting languages** are generally used to get data from files.

Acquition techniques

- For extracting data from websites; we can use web scraping tools like Scrapy, Scraper API, ParseHub, Webhouse.io, Content Grabber, Common crawl etc.
- NoSQL storage systems are increasingly used to manage a variety of data types in big data. These data stores are databases that do not represent data in a table format with columns and rows. NoSQL data stores provide APIs to allow users to access data. These APIs can be used directly or in an application that needs to access the data

Data Acquisition

- Example: WIFIRE case study as a real project that acquires data using several different mechanisms. The WIFIRE project stores sensor data from weather stations in a relational database. We use SQL to retrieve this data from the database to create models to identify weather patterns associated with Santa Ana conditions.
- To determine whether a particular weather station is currently experiencing Santa Ana conditions, we access real time data using a **web socket** service. Once we start listening to this service, we **receive weather station measurements** as they occur. This data is then processed and compared to patterns found by our models to determine if a weather station is experiencing Santa Ana conditions.

Data Acquisition

- At the same time Tweets are retrieved using hashtags related to any fire that is occurring in the region. The **Tweet messages** are retrieves using the **Twitter REST service**. The idea is to **determine the sentiment of these tweets** to see if people are expressing fear, anger or are simply nonchalant about the nearby fire.
- The combination of **sensor data and tweet sentiments** helps to give us a sense of the **urgency of the fire situation**. As a summary, big data and data science comes from many places. Finding and evaluating data useful to your big data analytics is important before you start acquiring data. Depending on the source and structure of data, there are alternative ways to access it.

Data Pre-processing

Data Pre processing

- Data Pre processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.
- Therefore, certain steps are executed to convert the data into a small clean data set. This technique is performed before the execution of the Iterative Analysis. The set of steps is known as Data Pre processing. It includes –

➢ Data Cleaning

➢ Data Integration

➢ Data Transformation

➢ Data Reduction

Data Pre processing

- **Tasks of Data Preparation**

1. **Data Cleaning**

➢ This is the first step which is implemented in Data Preprocessing. In this step, the primary focus is on handling missing data, noisy data, detection, and removal of outliers minimizing duplication, and computed biases within the data.

2. **Data Integration**

➢ This process is used when data is gathered from various data sources and data are combined to form consistent data. This consistent data after performing data cleaning is used for Data Preparation and analysis.

Data Pre processing

- **Tasks of Data Preparation**

3. Data Transformation

➢ This step is used to convert the raw data into a specified format according to the need of the model. The options for the transformation of data are given below -

- **Normalization -** In this method, numerical data is converted into the specified range, i.e., between 0 and one so that scaling of data can be performed.
- **Aggregation -** The concept can be derived from the word itself, this method is used to combine the features into one. For example, combining two categories can be used to form a new group.

Data Pre processing

- **Tasks of Data Preparation**
- **Generalization -** In this case, lower level attributes are converted to a higher standard.

4. Data Reduction

- After the transformation and scaling of data duplication, i.e., redundancy within the data is removed and efficiently organize the data during Data Preparation.

Data Quality

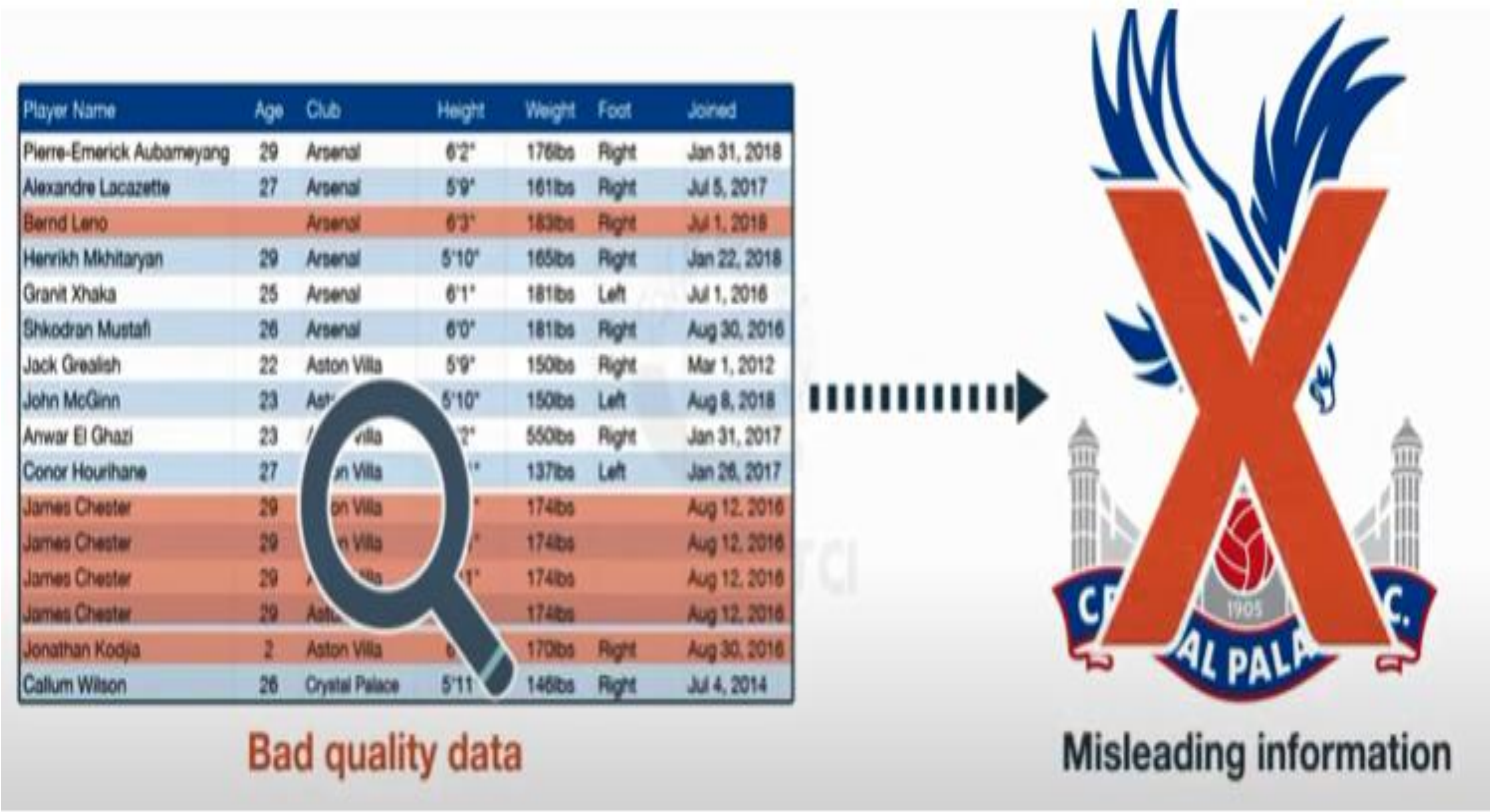
- Importance of good quality data
- Factors that cause data quality issues
- Data Quality Remediation

Importance of good quality data

- After collecting the data, most people start the analysis on it. Often, they forget to do a sanity check on the data. If the data is of bad quality, it can give misleading information.
- For example, if you start the analysis without ensuring data quality then you might get unexpected results such as the Crystal Palace club will win the next EPL. However, your domain knowledge on EPL says that the result looks inaccurate as Crystal Palace has never even finished in the top 4.
- A surprising fact is that a professional data scientist spends approximately 60% of his time ensuring that data is of high quality.

Importance of good quality data

- Example



Player Name	Age	Club	Height	Weight	Foot	Joined
Pierre-Emerick Aubameyang	29	Arsenal	6'2"	176lbs	Right	Jan 31, 2018
Alexandre Lacazette	27	Arsenal	5'9"	161lbs	Right	Jul 5, 2017
Bernd Leno		Arsenal	6'3"	183lbs	Right	Jul 1, 2018
Henrikh Mkhitaryan	29	Arsenal	5'10"	165lbs	Right	Jan 22, 2018
Granit Xhaka	25	Arsenal	6'1"	181lbs	Left	Jul 1, 2016
Shkodran Mustafi	26	Arsenal	6'0"	181lbs	Right	Aug 30, 2016
Jack Grealish	22	Aston Villa	5'9"	150lbs	Right	Mar 1, 2012
John McGinn	23	Ast[illegible]	5'10"	150lbs	Left	Aug 8, 2018
Anwar El Ghazi	23	[illegible] Villa	[illegible]2"	550lbs	Right	Jan 31, 2017
Conor Hourihane	27	[illegible]n Villa	[illegible]1"	137lbs	Left	Jan 26, 2017
James Chester	29	[illegible]on Villa	[illegible]"	174lbs		Aug 12, 2016
James Chester	29	[illegible]n Villa	[illegible]"	174lbs		Aug 12, 2016
James Chester	29	A[illegible]lla	[illegible]1"	174lbs		Aug 12, 2016
James Chester	29	Ast[illegible]	[illegible]	174lbs		Aug 12, 2016
Jonathan Kodjia	2	Aston Villa	6[illegible]	170lbs	Right	Aug 30, 2016
Callum Wilson	26	Crystal Palace	5'11	146lbs	Right	Jul 4, 2014

Bad quality data

Misleading information

Raw Data

- We collect data from the source like twitter, blogs, websites etc.
- Raw data may:
  - Have errors
  - Not validated
  - Multiple forms
  - Unformatted
  - Requiring confirmation or citation
- Input to the data processing process
- Raw data Example: If the correct format is not specified in an application form, the date of birth data can take many forms, such as “ 31st July 2021”, “31/07/2021”, “31/07/21” and “31 July 21”,
- This raw data needs to be processed to a common format for further use by systems/ humans.

Why does the data quality issue occur?

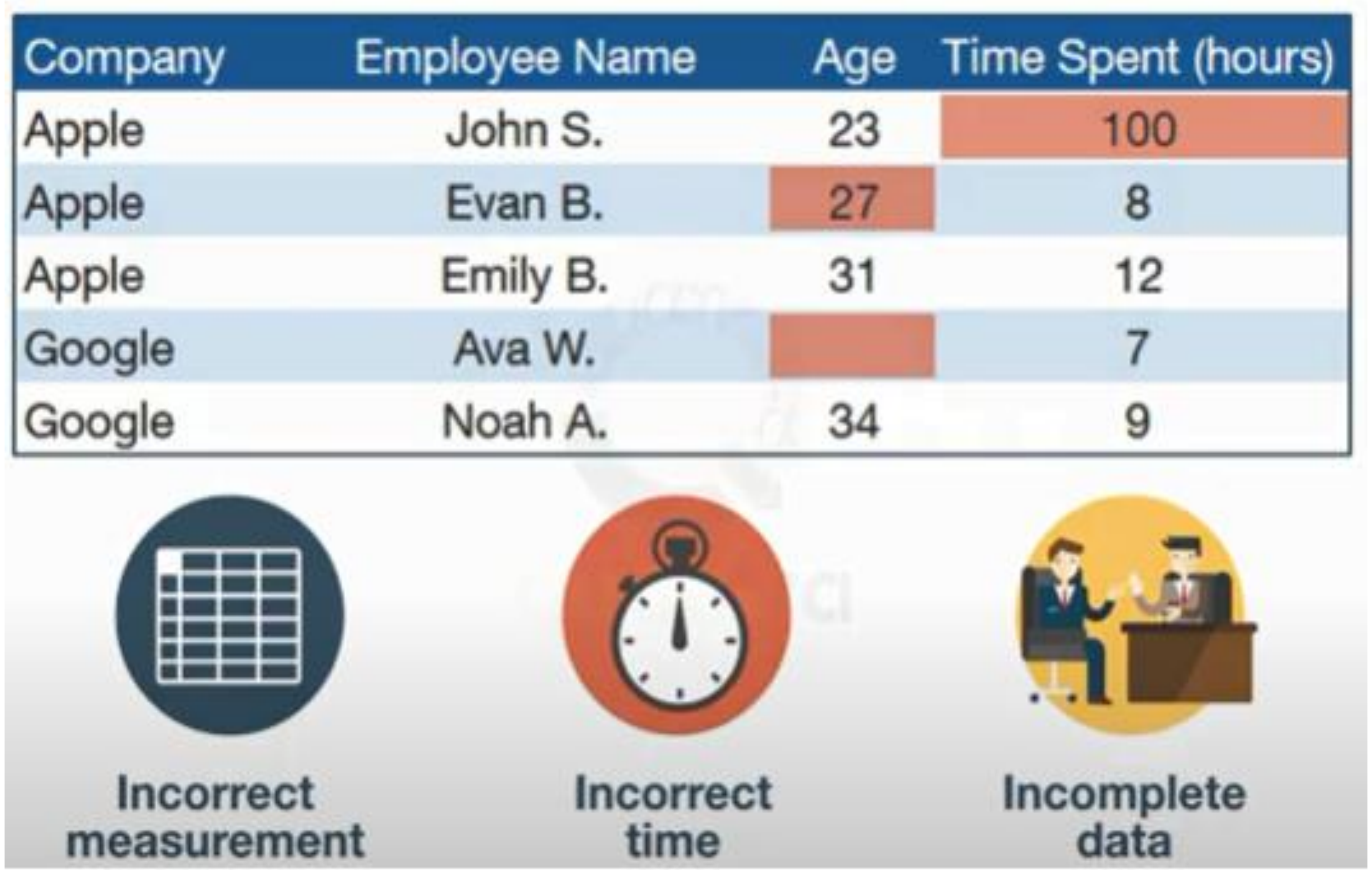
- **Need of Data Preparation Process and Pre processing**

➢ Some specified Machine Learning and Deep Learning model need information in a specified format, for example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values has to be managed from the original raw data set.

➢ Another aspect of Data Preparation and analysis is that the data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and the best out of them is chosen.

Why does the data quality issue occur?

- Improper Data Collection

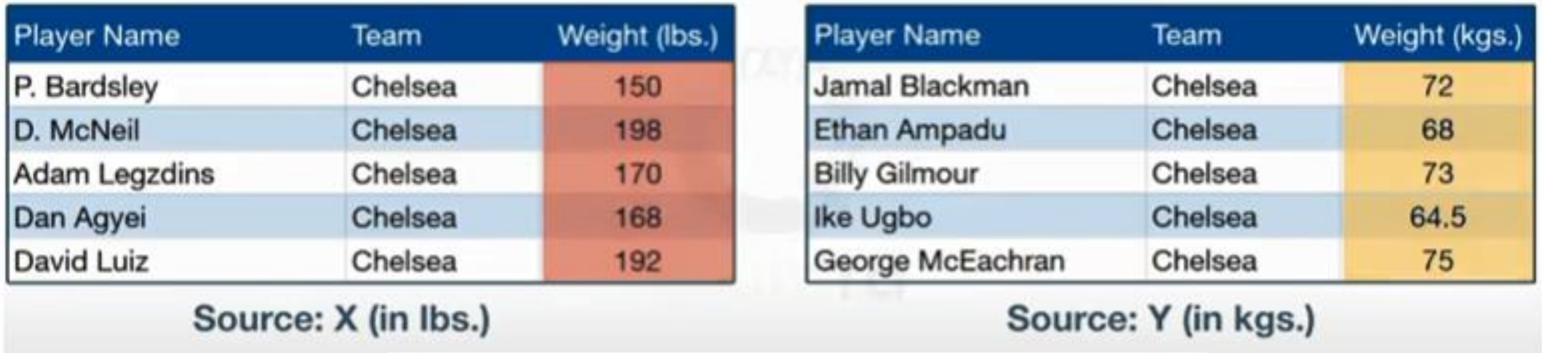


Company	Employee Name	Age	Time Spent (hours)
Apple	John S.	23	100
Apple	Evan B.	27	8
Apple	Emily B.	31	12
Google	Ava W.		7
Google	Noah A.	34	9

Incorrect measurement
Incorrect time
Incomplete data

Why does the data quality issue occur?

- Improper Data Integration

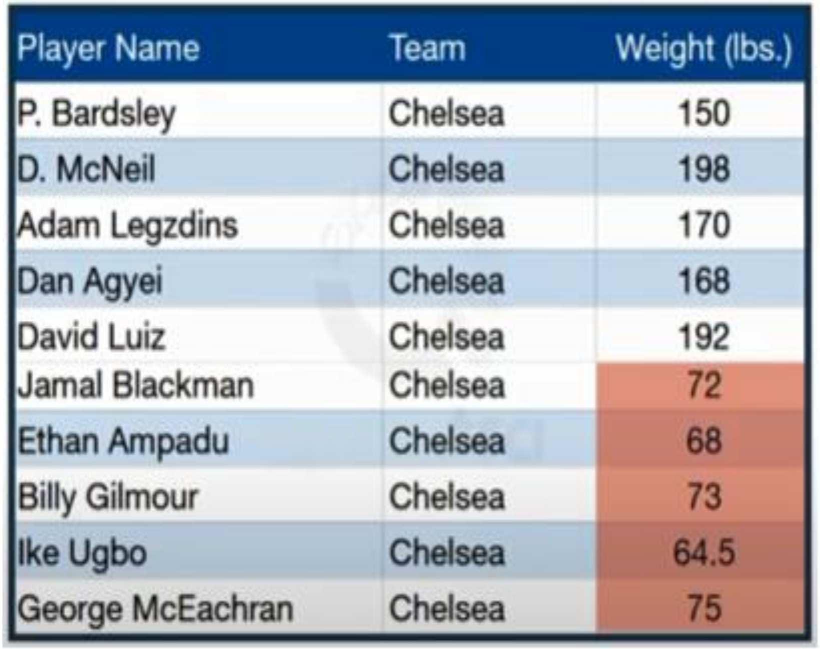


Player Name	Team	Weight (lbs.)
P. Bardsley	Chelsea	150
D. McNeil	Chelsea	198
Adam Legzdins	Chelsea	170
Dan Agyei	Chelsea	168
David Luiz	Chelsea	192

Source: X (in lbs.)

Player Name	Team	Weight (kgs.)
Jamal Blackman	Chelsea	72
Ethan Ampadu	Chelsea	68
Billy Gilmour	Chelsea	73
Ike Ugbo	Chelsea	64.5
George McEachran	Chelsea	75

Source: Y (in kgs.)



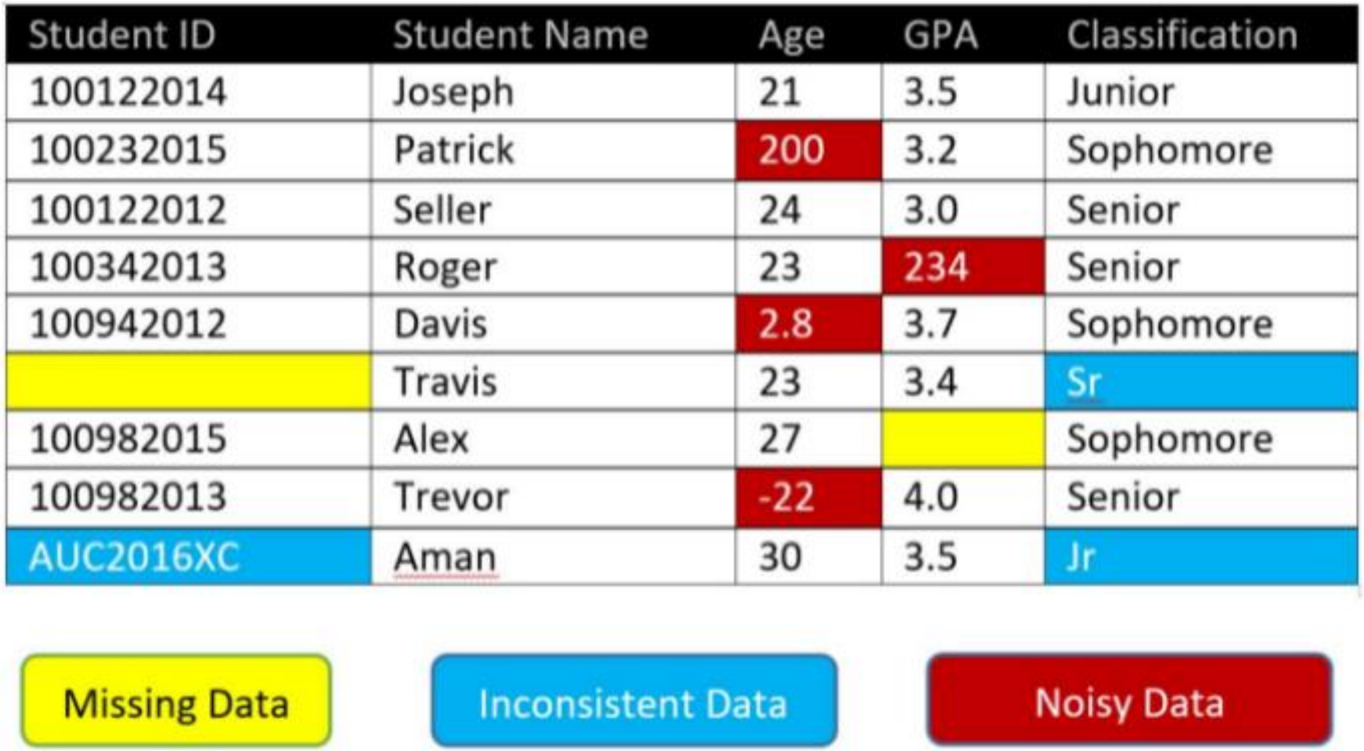
Player Name	Team	Weight (lbs.)
P. Bardsley	Chelsea	150
D. McNeil	Chelsea	198
Adam Legzdins	Chelsea	170
Dan Agyei	Chelsea	168
David Luiz	Chelsea	192
Jamal Blackman	Chelsea	72
Ethan Ampadu	Chelsea	68
Billy Gilmour	Chelsea	73
Ike Ugbo	Chelsea	64.5
George McEachran	Chelsea	75

Factors that cause data quality issues

- It can occur during data collection or data integration. Consider that you are recording the average time that employees spend in a cafeteria weekly across companies. You recorded 100 hours instead of 10 hours, or the unit of **measurement recorded incorrectly**. Also, if you are interviewing and someone may choose not to respond to certain questions which leads to **missing values**
- Another cause is when data is **collected from different sources and merged**. For example, you require the weight of all players of EPL in a single file. You extract the player's weight data from source X. But the weight data of some new players is not available, so you get it from source Y. The unit of weight measurement in source X is in pounds, and source y is in kilograms. If data collected from both sources are combined as it is, then there will be inconsistency in the data, and it will result in **inaccurate data**.

Factors that cause data quality issues

- Inconsistent data is when data fails to match. Let’s say, the user entered birthday to be May 07, 1993 and the age attribute displays 50. Or over time the ratings of a movie have changed from the numeric rating 1, 2, 3 to alphabets — A, B, C. Thus, in the same column data is not consistent.



Student ID	Student Name	Age	GPA	Classification
100122014	Joseph	21	3.5	Junior
100232015	Patrick	200	3.2	Sophomore
100122012	Seller	24	3.0	Senior
100342013	Roger	23	234	Senior
100942012	Davis	2.8	3.7	Sophomore
	Travis	23	3.4	Sr
100982015	Alex	27		Sophomore
100982013	Trevor	-22	4.0	Senior
AUC2016XC	Aman	30	3.5	Jr

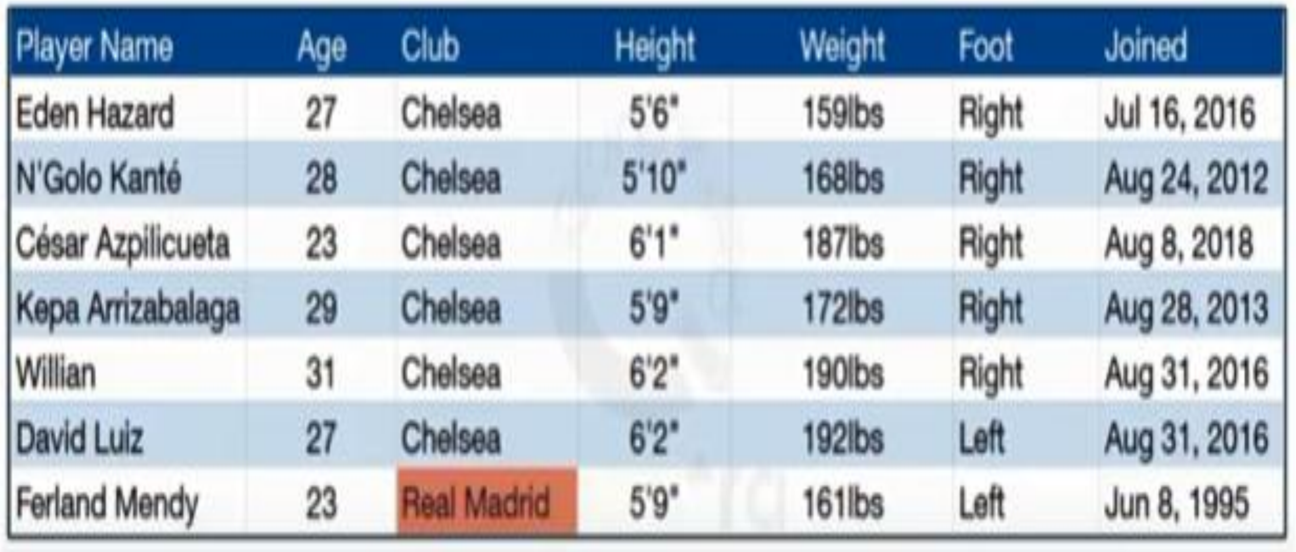
Missing Data

Inconsistent Data

Noisy Data

Types of data quality issues

- The common data quality issues that are easy to spot are **missing values, duplicate values, and inconsistent data**.
- However, some issues are difficult to spot. For example, If you follow EPL, then there is no club with the name of Real Madrid in EPL. From this example, you can infer that, detecting and handling such problems would require domain knowledge.



Player Name	Age	Club	Height	Weight	Foot	Joined
Eden Hazard	27	Chelsea	5'6"	159lbs	Right	Jul 16, 2016
N'Golo Kanté	28	Chelsea	5'10"	168lbs	Right	Aug 24, 2012
César Azpilicueta	23	Chelsea	6'1"	187lbs	Right	Aug 8, 2018
Kepa Arrizabalaga	29	Chelsea	5'9"	172lbs	Right	Aug 28, 2013
Willian	31	Chelsea	6'2"	190lbs	Right	Aug 31, 2016
David Luiz	27	Chelsea	6'2"	192lbs	Left	Aug 31, 2016
Ferland Mendy	23	Real Madrid	5'9"	161lbs	Left	Jun 8, 1995

Types of data quality issues

- **Summary**

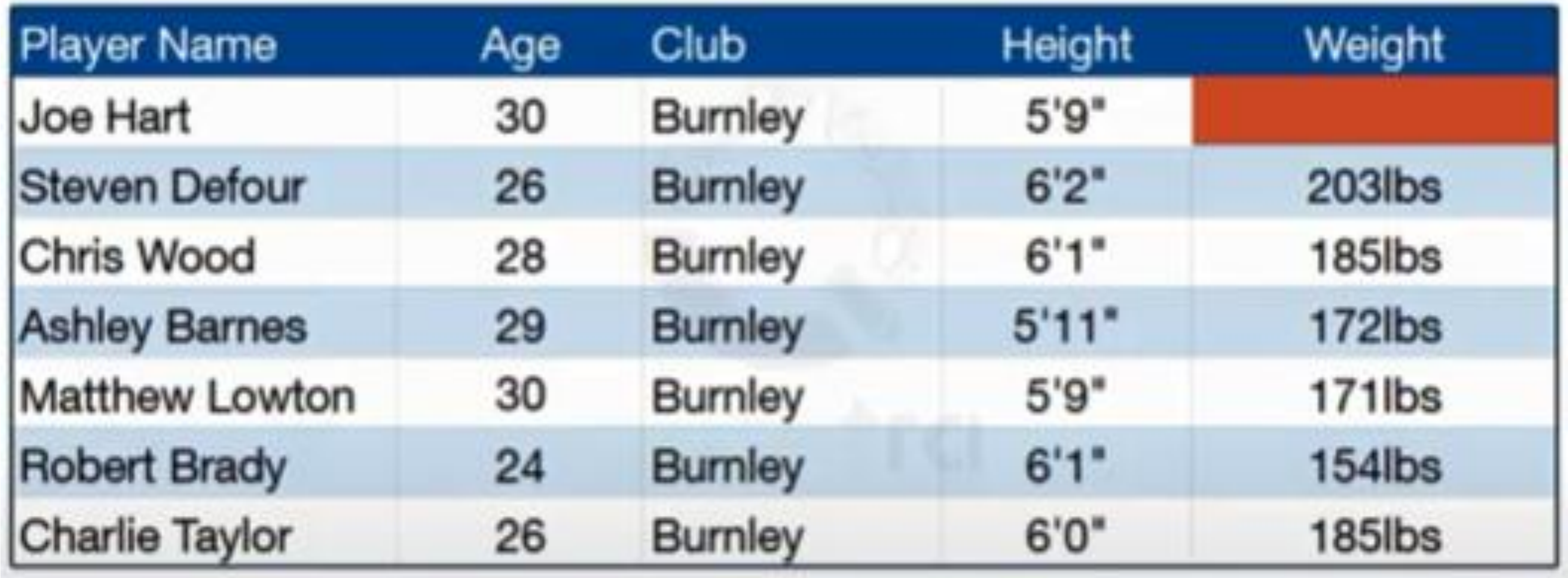
- Data Pre processing is necessary because of the presence of unformatted real-world data. Mostly real-world data is composed of –
  - **Inaccurate data (missing data) -** There are many reasons for missing data such as data is not continuously collected, a mistake in data entry, technical problems with biometrics
  - **The presence of noisy data (erroneous data and outliers) -** The reasons for the existence of noisy data could be a technological problem of gadget that gathers data, a human mistake during data entry and much more.
  - **Inconsistent data -** The presence of inconsistencies are due to the reasons such that existence of duplication within data, human data entry, containing mistakes in codes or names, i.e., violation of data constraints and much more necessitate Data Preparation and analysis.

Data Quality Remediation

- Therefore as a data scientist, you should develop a good **understanding** of the **domain**, and the **problem** you are solving.
- **How to fix these data quality issues?**
- Once you identify the inaccurate and missing data, you can use the alternate source of data, if available.
  - For example, if Bernd Leno’s age is missing, then you can get it from the **alternative source**, such as Wikipedia. However, this approach is not always possible, as you can’t find an alternative source for every data set, or data point.
  - In that case, a simple approach is to **remove the inaccurate data**. This can work well if you have a few inaccurate data points.
  - But, if there are many records with data quality problems, then this approach can reduce the data size, resulting in a poor analysis.

Data Quality Remediation

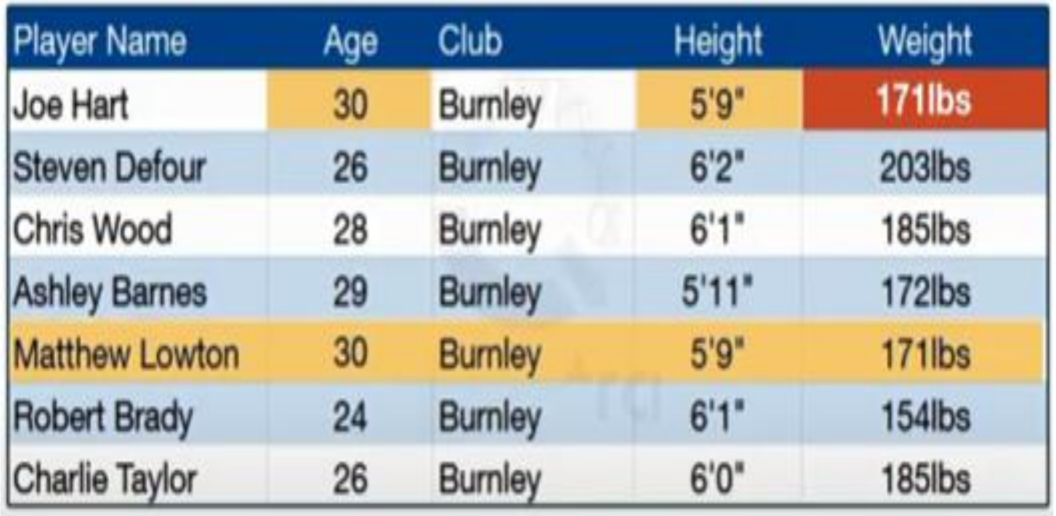
- A better approach, would be to **replace incorrect or missing values by mean, mode, or median**.
- For example, in this dataset, you can impute the missing weight of Joe Hart, by the **mode of 185**, or **mean of 178.3**, or **median of 178.5**. But, there are chances that the **imputed values are inaccurate**.



Player Name	Age	Club	Height	Weight
Joe Hart	30	Burnley	5'9"	🟥
Steven Defour	26	Burnley	6'2"	203lbs
Chris Wood	28	Burnley	6'1"	185lbs
Ashley Barnes	29	Burnley	5'11"	172lbs
Matthew Lowton	30	Burnley	5'9"	171lbs
Robert Brady	24	Burnley	6'1"	154lbs
Charlie Taylor	26	Burnley	6'0"	185lbs

Data Quality Remediation

- Another approach, is to **estimate the missing** weight **value**, based on the player whose height and age is similar to Joe Hart.
- For example, Matthew Lowton has the same age and height as Joe Hart, so you can assign 171.



Player Name	Age	Club	Height	Weight
Joe Hart	30	Burnley	5'9"	171lbs
Steven Defour	26	Burnley	6'2"	203lbs
Chris Wood	28	Burnley	6'1"	185lbs
Ashley Barnes	29	Burnley	5'11"	172lbs
Matthew Lowton	30	Burnley	5'9"	171lbs
Robert Brady	24	Burnley	6'1"	154lbs
Charlie Taylor	26	Burnley	6'0"	185lbs

- However, not all values can be estimated from the values of other attributes. Thus, the **approach to remediate the data quality issues depends**, on the **type of data** you are dealing with, and the **domain understanding** of the data.

How is Data Pre processing performed?

- Data Pre processing is carried out to remove the cause of unformatted real-world data which we discussed. Therefore by using these three different steps we can **handled missing data** during Data Preparation–

➢ **Ignoring the missing record -** It is the simplest and efficient method for handling the missing data. But, this method should not be performed at the time when the number of missing values is immense or when the pattern of data is related to the unrecognized primary root of the cause of the statement problem.

➢ **Filling the missing values manually -** This is one of the best-chosen methods of Data Preparation process. But there is one limitation that when there are large data set, and missing values are significant then, this approach is not efficient as it becomes a time-consuming task.

How is Data Pre processing performed?

- **Filling using computed values -** The missing values can also be occupied by **computing mean, mode or median** of the observed given values. Another method could be the predictive values in Data Preprocessing is that are computed by using any **Machine Learning or Deep learning tools and algorithms**. But one drawback of this approach is that it can generate bias within the data as the **calculated values are not accurate** concerning the observed values.
- **Constant**: You may replace the missing values of a column by using a constant such as “Unknown” or “ ∞”.

Imputation

- Imputation is the process of replacing unknown values with the best guess.
- Usually, we use the statistical parameters such as mean, median or mode to impute the missing values.
- When to use which parameter is decided by the parameters ability to provide the best central location for the data.

Imputation

- **Mean**

- Mean is defined as the average value of all the data points.
- The mean can be used to impute any missing data if the distribution of data is even. If you impute the missing data with the mean, the overall mean of the data remains the same but the standard deviation will reduce.
- For example, if the data is skewed. By imputing with mean, the distribution of the actual dataset will change completely. This may result in erroneous conclusions.
- The mean has one more disadvantage. It is very susceptible to the influence of outliers. Outliers are unusually large or small value compared to the rest of the dataset. These values shift the mean towards one side and make it unrepresentative of the dataset. In such cases, it is better to use the median to impute values

Imputation

- **Median**

➢ Median can be defined as the central value of a dataset. Unlike the mean, the median value is less affected by outliers or skewed data. In such case a median imputation would be more appropriate than mean imputation.

- **Mode**

➢ The mode can be defined as the most frequent value in a dataset. Unlike the mean, the mode is less affected by outliers or skewed data. We can use the mode value to impute the missing data.

➢ Since it is the most likely value in a dataset, the distribution pattern doesn’t affect the mode value much. The mode is the natural choice for imputation in most cases.

Imputation

- **Mode**

- ➢ In case where there are multiple modes, you have to decide which mode value should be used for imputation.
- ➢ In situations where the data contains a lot of outliers in a particular range then the mode of the data is not close to the most of the data and cannot be treated a representative of the data. In such cases using the median would be a better option.

How is Data Pre processing performed?

- Meaning of **noisy data**
- Noisy Data: Noisy data is a **meaningless data** that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc.
- For example, out of range values like a person filling out the numeric value -679 in the salary field or some negative four digit random number in the age field. Impossible data combinations like — Gender: Male, Pregnant: Yes adds to the noise in the data.
- Noisy data is used interchangeably with the term **corrupt data**.
- The process of removing noise from a data set is termed as **data smoothing**.

How is Data Pre processing performed?

- How we can deal with **noisy data**?
- **Data Binning**

➢ Binning is a technique where we sort the data and then partition the data into equal frequency bins. Then you may either replace the noisy data with the bin mean, bin median or the bin boundary.

➢ This is a simple example of data binning.

Sorted data for Age: 3, 7, 8, 13, 22, 22, 22, 26, 26, 28, 30, 37

Smoothing the data by equal frequency bins

- **Bin 1:** 3, 7, 8, 13
- **Bin 2:** 22, 22, 22, 26
- **Bin 3:** 26, 28, 30, 37

How is Data Pre processing performed?

- ➢ Smoothing by bin means
- **Bin 1:** 8, 8, 8, 8
- **Bin 2:** 23, 23, 23, 23
- **Bin 3:** 30, 30, 30, 30
- In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
- ➢ Smooth the data by bin boundaries: pick the minimum and maximum value, Put the minimum on the left side and maximum on the right side and Middle values in bin boundaries move to its closest neighbour value with less distance.
- **Bin 1:** 3, 3, 3, 13
- **Bin 2:** 22, 22, 22, 26
- **Bin 3:** 26, 26, 26, 37
- In smoothing by bin boundaries, each bin value is replaced by the closest boundary value.

How is Data Pre processing performed?

- How we can deal with noisy data?
- **Pre processing in Clustering**

➢ In the approach, the outliers may be detected by grouping similar data in the same group, i.e., in the same cluster. Thus, values that fall far apart from the cluster may be considered noise or outliers.

- **Machine Learning**

➢ A Machine Learning algorithm can be executed for the smoothing of data during Data Preprocessing . For example, **Regression** Algorithm can be used for the smoothing of data using a specified linear function.

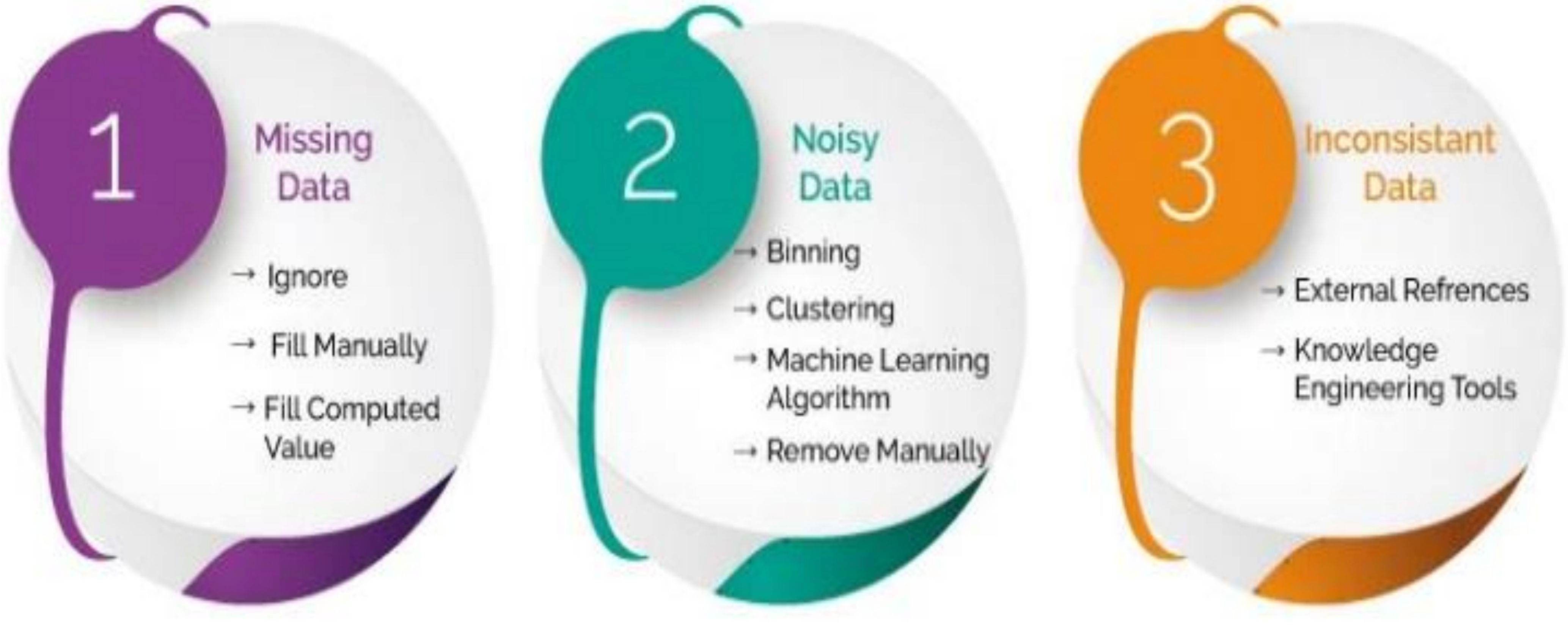
How is Data Pre processing performed?

- How we can deal with noisy data?
- **Removing manually**

➢ The noisy data can be deleted manually by the human being, but it is a time-consuming Data Preparation process, so mostly this method is not given priority. To deal with the inconsistent data manually and perform Data Preparation and analysis properly, the data is managed using external references and knowledge engineering tools like the knowledge engineering process.

How is Data Pre processing performed?

- Data Cleaning steps for data pre-processing



1 Missing Data

→ Ignore

→ Fill Manually

→ Fill Computed Value

2 Noisy Data

→ Binning

→ Clustering

→ Machine Learning Algorithm

→ Remove Manually

3 Inconsistant Data

→ External Refrences

→ Knowledge Engineering Tools

Data Integration

- There are numerous tools available in the market that would help us query the data effectively since our data will not integrate itself.
- We have some **Open Source** Data Integration Tools, **Cloud-based** Tools, and also the **On-premises Data** Integration tools.
- The best tool to choose depends on the **requirements, platform, and type of data** that particular business organizations are likely to use.
- List of Common Open-source Tools

➢ Apache Airflow

➢ CloverETL

➢ Talend Open Studio

➢ Karma

➢ Pentaho

➢ Dell Bhoomi AtomSphere

Data Transformation

- Turning the data into an appropriate format for the computer to learn from.
- Example: For research about smog around the globe, you have data about **wind speeds**. However, the data got mixed, and we have three **variants** of figures: **meters per second, miles per second, and kilometers per hour.** We need to transform these data to the same scale for ML modeling.
- Here are the techniques for data transformation or data scaling:

➢ Aggregation

➢ Normalization

➢ Discretization

➢ Concept hierarchy generation

➢ Generalization

Data Transformation

- **Aggregation**

➢ In the case of data aggregation, the data is pooled together and presented in a unified format for data analysis.

➢ Working with a large amount of high-quality data allows for getting more reliable results from the ML model.

➢ Example: If we want to build a neural network algorithm that simulates the style of Vincent Van Gogh, we need to provide as many paintings by this famous artist as we can to provide enough material for training. The **images** need to have the **same digital format**, and we will use data transformation techniques to achieve that.

Data Transformation

- **Normalization**

➢ It helps you to scale the data within a range to avoid building incorrect ML models while training and/or executing data analysis. If the data range is very wide, it will be hard to compare the figures. With various normalization techniques, you can transform the original data linearly, perform decimal scaling or Z-score normalization.

➢ For example, to compare the population growth of city X (1+ million citizens) to 1 thousand new citizens in city Y, we need to normalize.

Data Transformation

- **Discretization**

➢ During discretization, a programmer transforms the data into sets of small intervals. For example, putting people in categories “young”, “middle age”, “senior” rather than working with continuous age values. Discretization helps to improve efficiency.

- **Concept hierarchy generation**

➢ If you use the concept hierarchy generation method, you can generate a hierarchy between the attributes where it was not specified. For example, if you have the location information that includes a street, city, province, and country but they have no hierarchical order, this method can help you transform the data.

Data Transformation

- **Generalization**

➢ With the help of generalization, it is possible to convert low-level data features to high-level data features. For example, house addresses can be generalized to higher-level definitions, such as town or country.

Data Reduction

- When we work with large amounts of data, it becomes harder to come up with reliable solutions. Data reduction can be used to **reduce the amount of data and decrease the costs** of analysis.
- Researchers really need data reduction when working with **verbal speech** datasets. Massive arrays contain individual features of the speakers, for example, interjections and filling words. In this case, huge databases can be decreased to a representative sampling for the analysis.

➢ Attribute feature selection

➢ Dimensionality reduction

➢ Numerosity reduction

➢ Data compression

Data Reduction

- **Attribute feature selection**

- ➢ If we construct a **new feature combining the given features** in order to make the data mining process more efficient, it is called an attribute selection.
- ➢ For example, the features male/female and student can be constructed into male student/female student. This can be useful if we conduct research about how many men and/or women are students but their study field doesn’t interest us.

Data Reduction

- **Dimensionality reduction**

➢ Datasets that are used to solve real-life tasks have a huge number of features. Computer vision, speech generation, translation, and many other tasks cannot sacrifice the speed of operation for the sake of quality. It’s possible to use dimensionality reduction to cut the number of features used.

- **Numerosity reduction**

➢ Numerosity reduction is a method of data reduction that replaces the original data by a smaller form of data representation. There are two types of numerosity reduction methods –

➢ Parametric

➢ Non-Parametric.

Data Reduction

- **Parametric Methods**

➢ Parametric methods use models to represent data. Commonly, regression is used to build such models.

- **Non-parametric methods**

➢ These techniques allow for storing reduced representations of the data through histograms, data sampling, and data cube aggregation.

- **Data compression**

➢ Example: Audio/video compression

Data Wrangling

Data Wrangling

- Data Wrangling is a technique that is executed at the time of making an interactive model. In other words, it is used to convert the raw data into the format that is convenient for the consumption of data.
- This technique is also known as Data Munging.
- This method also follows certain steps such as after extracting the data from different data sources, **sorting** of data using the certain algorithms are performed, **decompose** the data into a different structured format and finally **store** the data into another database.
- Transforming data into a form that is most appropriate for learning algorithms.

Data Wrangling

- **The need of Data Wrangling**

- ➢ Data Wrangling is an important aspect of implementing the model. Therefore, data is **converted to the proper feasible format** before applying any model to it. By performing filtering, grouping, and selecting appropriate data; accuracy and performance of the model could be increased.
- ➢ Another concept is that when time-series data has to be handled every algorithm is executed with different aspects. Therefore Data Wrangling is used to **convert the time series data** into the **required format** of the applied model.
- ➢ In simple words, the **complex data is transformed into a usable format** for performing analysis on it.

Data Wrangling

- **Why is Data Wrangling Important?**

➢ Data Wrangling is used to handle the issue of **Data Leakage** while implementing Machine Learning and Deep Learning.

Data Leakage

- Data Leakage is responsible for the cause of an invalid Machine Learning/Deep Learning model due to the over-optimization of the applied model.
- Data Leakage is the term used when the data from outside, i.e., not part of the training dataset is used for the learning process of the model. This additional learning of information by the applied model will disapprove of the computed estimated performance of the model.
- For example when we want to use the particular feature for performing Predictive Analysis, but that **specific feature is not present** at the time of **training** of dataset then data leakage will be introduced within the model.

Data Leakage

- Data Leakage can be demonstrated in many ways that are given below -

➢ The Leakage of data from test dataset to the training data set.

➢ Leakage of computed correct prediction to the training dataset.

➢ Usage of data outside the scope of the applied algorithm

- In general, the leakage of data is observed from two primary sources of Machine Learning/Deep Learning algorithms such as **feature attributes** (variables) and **training data set**.

Data Leakage

- **Checking the presence of Data Leakage within the applied model**
- Data Leakage is observed at the time of usage of complex datasets. They are described below –

➢ At the time of dividing the time series dataset into training and test, the dataset is a complex problem.

➢ The implementation of sampling in a graphical problem is a complex task.

➢ Storage of analog observations in the form of audios and images in separate files having a defined size and timestamp.

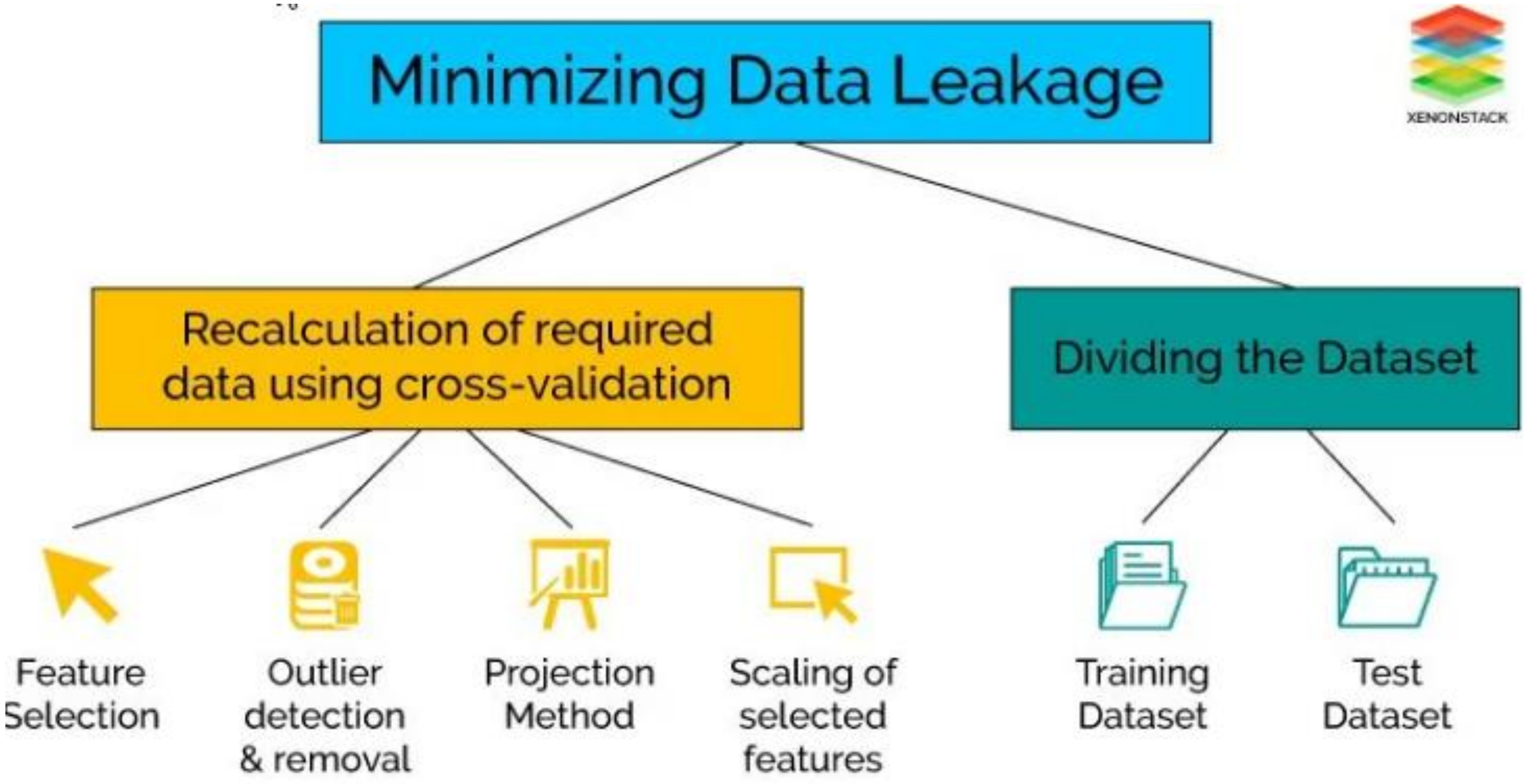
How is Data Wrangling performed?

- Data Wrangling is conducted to minimize the effect of Data Leakage while executing the model.
- The effect of Data Leakage could be minimized by recalculating for the required Data Preparation during the **cross-validation process** that includes **feature selection, outliers detection, and removal, projection methods, scaling of selected features** and much more.
- Another solution is that dividing the complete dataset into **training dataset** that is used to train the model and **validation dataset** which is used to evaluate the performance and accuracy of the applied model.

How is Data Wrangling performed?

- The selection of the model is made by looking at the results of the test data set in the cross-validation process. This conclusion will not always be valid as the sample of the test data set could vary.
- The performance of different models is evaluated for the particular type of test dataset. Therefore, while selecting the best model, test error is overfitting. The test error variance is determined by using different samples of the test dataset . Choosing of suitable model happens in this way.

How is Data Wrangling performed?



Minimizing Data Leakage

Recalculation of required data using cross-validation

Dividing the Dataset

Feature Selection

Outlier detection & removal

Projection Method

Scaling of selected features

Training Dataset

Test Dataset

XENONSTACK

Data Preparation vs Data Wrangling

- Data Preprocessing steps are performed before the Data Wrangling.
- In this case, Data Preprocessing data is prepared **exactly after receiving the data from the data source**. In this initial transformations, Data Cleaning or any aggregation of data is performed. It is executed once.
- For example, we have data where one attribute has three variables, and we have to convert them into three attributes and delete the special characters from them. The concept of Data Preparation steps **performed before applying any iterative model** and will be executed once in the project.

Data Preparation vs Data Wrangling

- On the other hand, Data Wrangling is **performed during the iterative analysis** and **model building**. This concept at the time of feature engineering. The conceptual view of the dataset changes as different models are applied to achieve a good analytic model.
- For example, we have data containing 30 attributes where two attributes are used to compute another attribute, and that computed feature is used for further analysis. In this way, the **data** could be **changed according to the requirement of the applied model**, and Data Preparation can be effective.

Data Preparation vs Data Wrangling

- Summary
- Data Preprocessing: Preparation of data directly after accessing it from a data source. Typically realized by a developer or data scientist for initial transformations, aggregations and data cleansing. This step is done before the interactive analysis of data begins. It is executed once.
- Data Wrangling: Preparation of data during the interactive data analysis and model building. Typically done by a data scientist or business analyst to change views on a dataset and for features engineering. This step iteratively changes the shape of a dataset until it works well for finding insights or building a good analytic model.

Tasks of Data Wrangling

- **Discovering**
- ➢ Firstly, data should be understood thoroughly and examine which approach will best suit. For example: You can liken it to looking in your refrigerator before cooking a meal to see what ingredients you have at your disposal.
- **Structuring**
- ➢ As the data is gathered from different sources, the data will be present in various shapes and sizes. Therefore, there is a need for structuring the data in a proper format.
- **Cleaning**
- ➢ Cleaning or removing of data should be performed that can degrade the performance of the analysis.

Tasks of Data Wrangling

- **Enrichment**

➢ Extract new features or data from the given data set to optimize the performance of the applied model.

- **Validating**

➢ This approach is used for improving the quality of data and consistency rules so that transformations that are applied to the data could be verified.

- **Publishing**

➢ After completing the steps of Data Wrangling, the steps can be documented so that similar steps can be performed for the same kind of data to save time.

Data Wrangling vs ETL

- Data Wrangling is used to analyze the data that was gathered from different data sources. It is **designed specially to handle diverse and complex data of any scale**. But in the case of ETL, it can **handle structured data** that was originated from different databases or operating systems.
- Data Wrangling technology is used by **business analysts**, users engaged in business, and managers. On the other hand, ETL (Extract, Transform, and Load) is employed by **IT Professionals**. They receive the requirements from business people and then they use ETL tools to deliver the data in a required format.
- The primary task of the Data Wrangling method is to **manage the newly generated data** from various sources **for the analysis process** whereas the goal of ETL is to extract, transform and load the data into the central enterprise Data Warehouse for performing **analysis process using business applications.**

Tools

- **Data Preprocessing Tools**

➢ Data Preprocessing in R

➢ Data Preprocessing in Python

➢ Data Preprocessing in Weka

- **Data Wrangling Tools**

➢ Data Wrangling in Tabula

➢ Data Wrangling in R

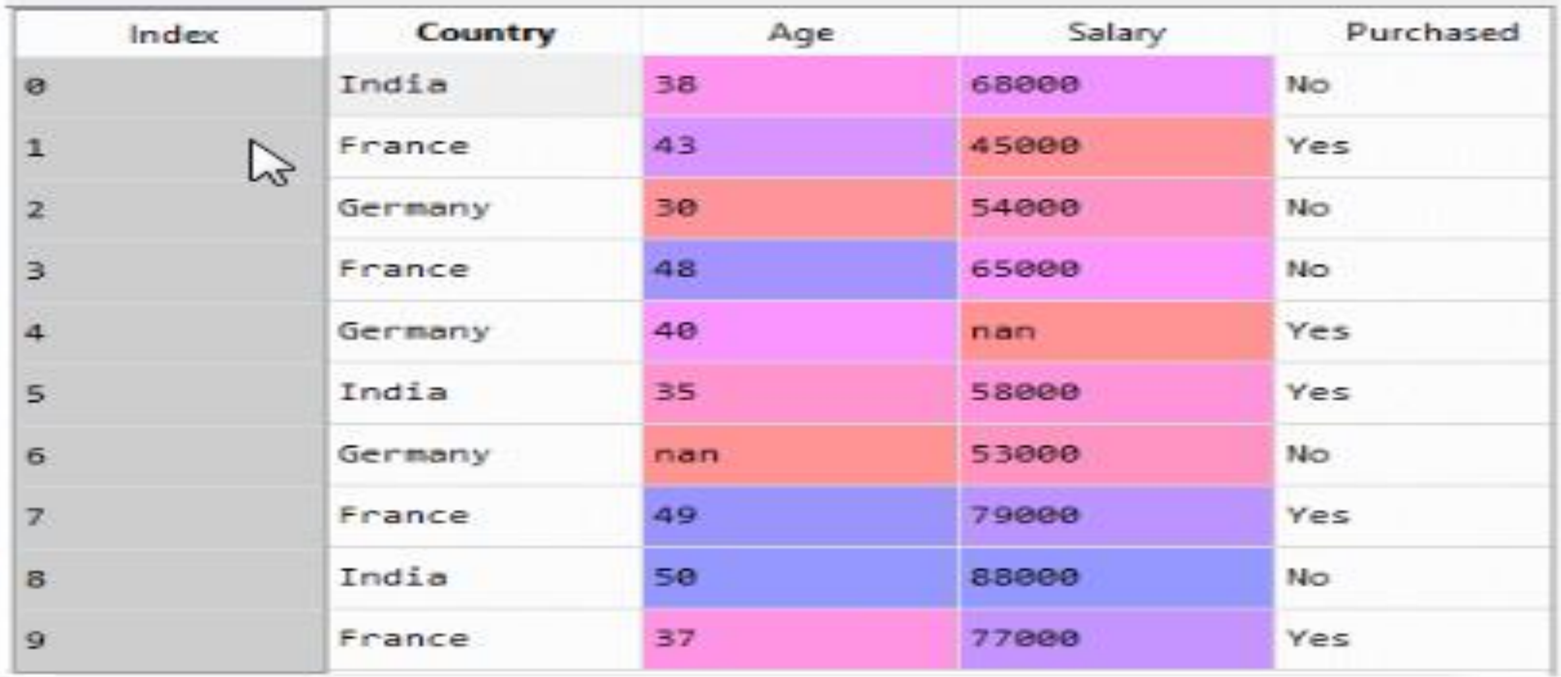
➢ Data Wrangling in CSVKit

➢ Data Wrangling using Python with Pandas

➢ Data Wrangling using Mr. Data Converter

Encoding the categorical data

- Categorical data refers to the information that has specific categories within the dataset. Machine Learning models are primarily based on mathematical equations. Thus, we can intuitively understand that keeping the categorical data in the equation will cause certain issues since we would only need numbers in the equations.



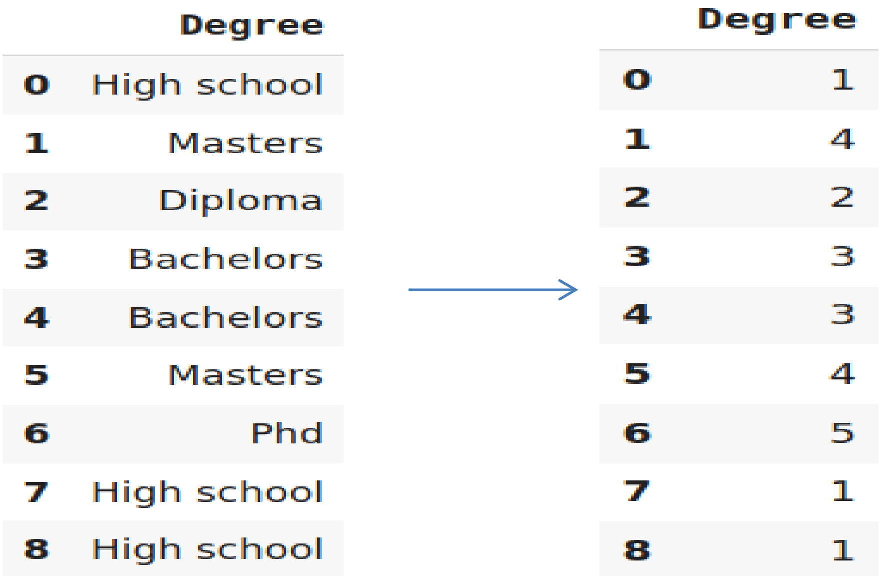
Index	Country	Age	Salary	Purchased
0	India	38	68000	No
1	France	43	45000	Yes
2	Germany	30	54000	No
3	France	48	65000	No
4	Germany	40	nan	Yes
5	India	35	58000	Yes
6	Germany	nan	53000	No
7	France	49	79000	Yes
8	India	50	88000	No
9	France	37	77000	Yes

Encoding the categorical data

- **Label Encoding or Ordinal Encoding**

➢ We use this categorical data encoding technique when the categorical feature is ordinal. In this case, retaining the order is important. Hence encoding should reflect the sequence.

➢ In Label encoding, each **label is converted into an integer** value. We will create a variable that contains the categories representing the education qualification of a person.



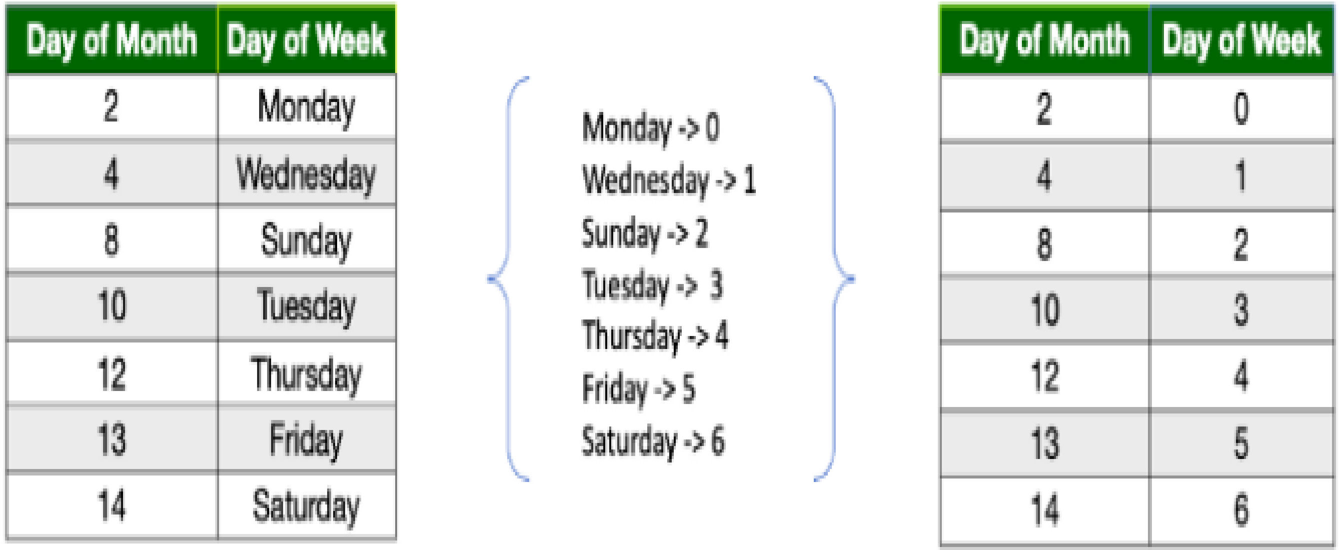
	Degree
0	High school
1	Masters
2	Diploma
3	Bachelors
4	Bachelors
5	Masters
6	Phd
7	High school
8	High school

	Degree
0	1
1	4
2	2
3	3
4	3
5	4
6	5
7	1
8	1

Encoding the categorical data

- **Label Encoding or Ordinal Encoding**

➢ In this process, we assign a discrete number to each unique category using some defined process. For example, we can sort the variables in order of the **number of occurrences** and **number them in increasing order**.



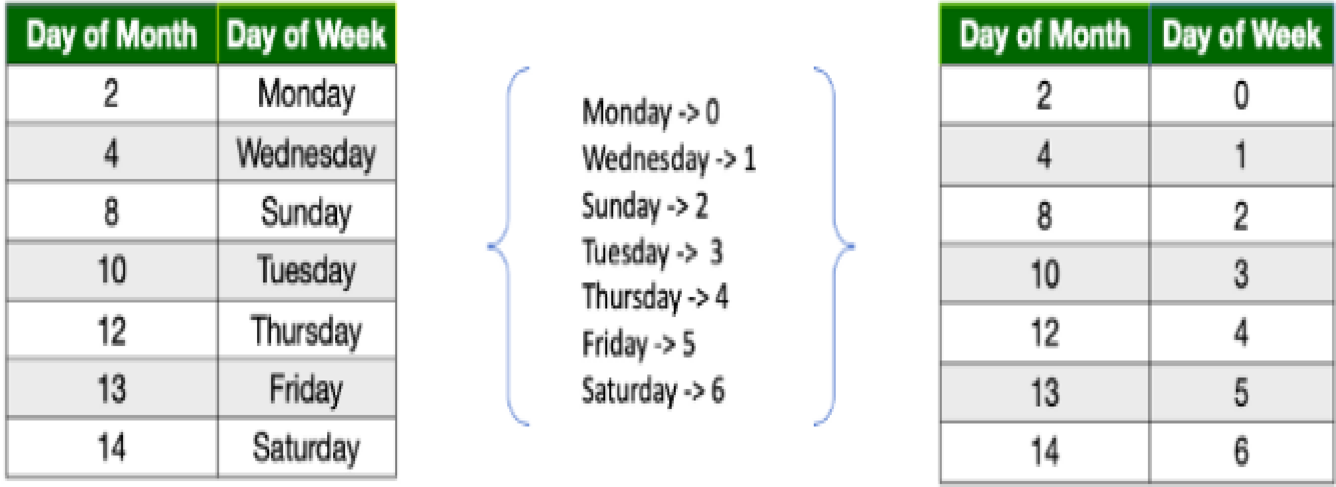
Day of Month	Day of Week
2	Monday
4	Wednesday
8	Sunday
10	Tuesday
12	Thursday
13	Friday
14	Saturday

Monday -> 0
Wednesday -> 1
Sunday -> 2
Tuesday -> 3
Thursday -> 4
Friday -> 5
Saturday -> 6

Day of Month	Day of Week
2	0
4	1
8	2
10	3
12	4
13	5
14	6

Encoding the categorical data

- **Label Encoding or Ordinal Encoding**
- In this example, the days are labelled in the order of their appearance in the data. The major problems here are:-
- **Natural ordering is lost**
- Common relationships between categories are not captured. (For example, Saturday and Sunday together make a weekend and hence should be closer to each other)



Day of Month	Day of Week
2	Monday
4	Wednesday
8	Sunday
10	Tuesday
12	Thursday
13	Friday
14	Saturday

Monday -> 0
Wednesday -> 1
Sunday -> 2
Tuesday -> 3
Thursday -> 4
Friday -> 5
Saturday -> 6

Day of Month	Day of Week
2	0
4	1
8	2
10	3
12	4
13	5
14	6

Encoding the categorical data

- **One Hot Encoding**

➢ We use this categorical data encoding technique when the features are nominal(do not have any order). In one hot encoding, for each level of a categorical feature, we create a new variable. Each category is mapped with a binary variable containing either 0 or 1. Here, **0 represents** the **absence**, and **1 represents** the **presence** of that category.

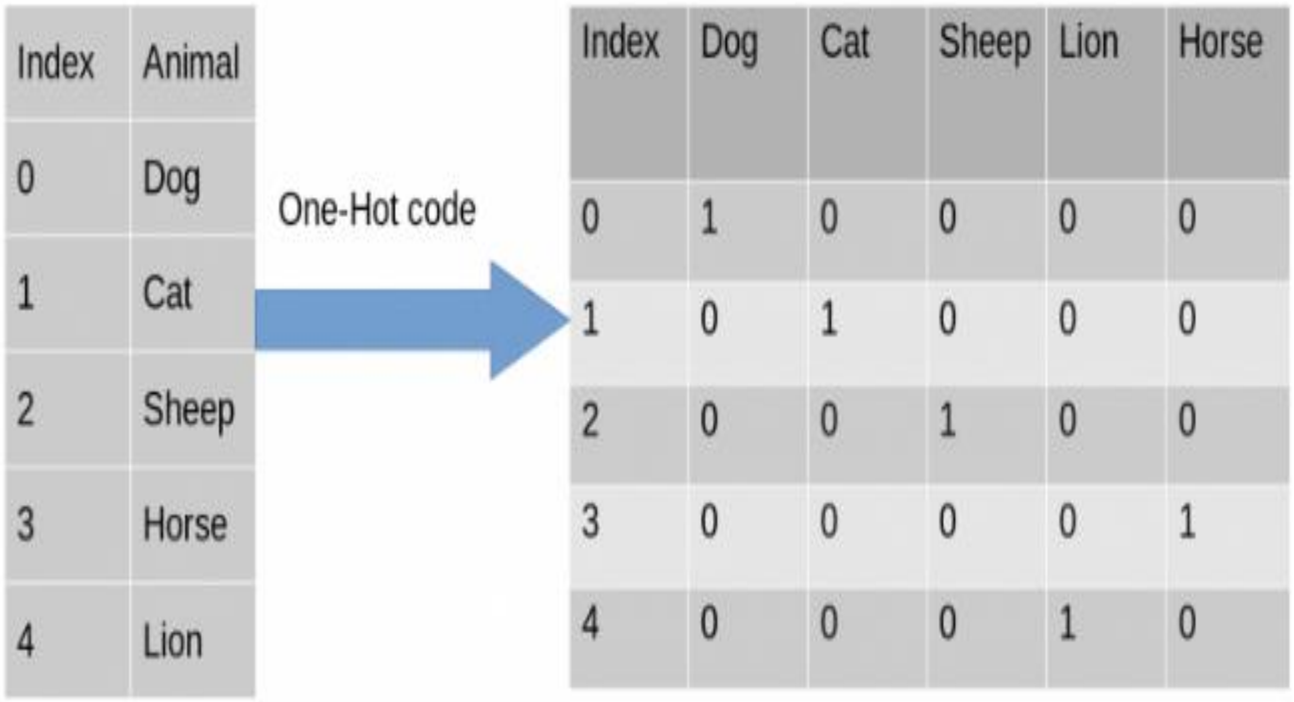
➢ These newly created binary features are known as **Dummy variables**. The number of dummy variables depends on the levels present in the categorical variable. This might sound complicated. Let us take an example to understand this better.

➢ Suppose we have a dataset with a category animal, having different animals like Dog, Cat, Sheep, Cow, Lion. Now we have to one-hot encode this data.

Encoding the categorical data

- **One Hot Encoding**

➢ After encoding, we have dummy variables each representing a category in the feature Animal. Now for each category that is present, we have 1 in the column of that category and 0 for the others.



Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

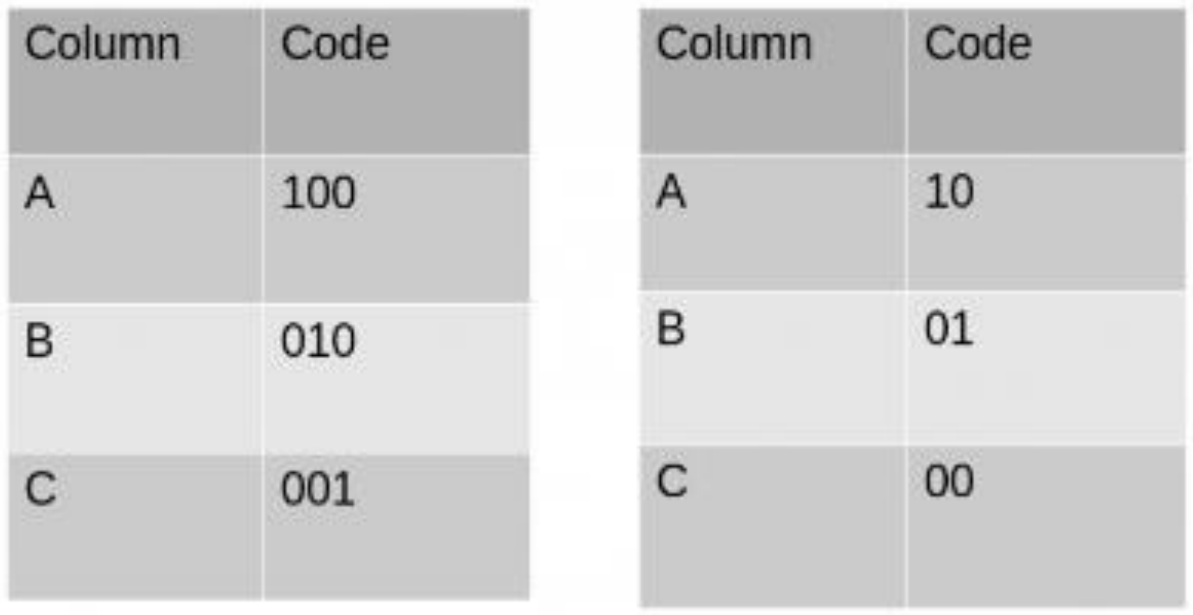
Encoding the categorical data

- **Dummy Encoding**

- ➢ Dummy coding scheme is similar to one-hot encoding. This categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables). In the case of one-hot encoding, for N categories in a variable, it uses N binary variables. The dummy encoding is a small improvement over one-hot-encoding. Dummy encoding uses **N-1 features** to represent N labels/categories.
- ➢ To understand this better here we are coding the same data using both one-hot encoding and dummy encoding techniques. While one-hot uses 3 variables to represent the data whereas dummy encoding uses 2 variables to code 3 categories.

Encoding the categorical data

- **Dummy Encoding**



Column	Code
A	100
B	010
C	001

One- Hot Coding

Column	Code
A	10
B	01
C	00

Dummy Code

Encoding the categorical data

- **Drawbacks of One-Hot and Dummy Encoding**

➢ One hot encoder and dummy encoder are two powerful and effective encoding schemes.

➢ They are also very popular among the data scientists, But may not be as effective when- If there are multiple categories in a feature variable in such a case we need a similar number of dummy variables to encode the data. For example, a column with 30 different values will require 30 new variables for coding.

➢ Due to the massive increase in the dataset, coding slows down the learning of the model along with deteriorating the overall performance that ultimately makes the model computationally expensive.

Encoding the categorical data

- **Count or frequency encoding**

➢ Replace the categories by the count of the observations that show that category in the dataset. Similarly, we can replace the category by the frequency -or percentage- of observations in the dataset. That is, if 10 of our 100 observations show the colour blue, we would replace blue by 10 if doing count encoding, or by 0.1 if replacing by the frequency.

➢ Limitation: If two different categories appear the same amount of times in the dataset, that is, they appear in the same number of observations, they will be replaced by the same number, hence, may lose valuable information.

Encoding the categorical data

- **Summary**

➢ As handling categorical variables in any dataset is crucial step in feature engineering, any of the above techniques can be applied depending upon type of model.

➢ If there are **lesser categories** and it is **nominal** categorical data, then **one-hot encoding** works just fine. If the relationship between any categorical column as independent variable and dependent variable (Target Variable) is important, then **Ordered Integer Encoding** can be applied. For **ordinal** categorical data, simply **Label Encoding** can be used.

➢ Traditional techniques for handling categorical variables kind of works but limits the capabilities of algorithms.

Encoding the categorical data

- **One hot vectors vs Word embedding(word2vec, GloVe)**

➢ One-hot vectors are high-dimensional and sparse, while word embedding’s are low-dimensional and dense. When we use one-hot vectors as a feature in a classifier, your feature vector grows with the vocabulary size; word embedding’s are more computationally efficient.

➢ Word embedding’s have the ability to generalize, due to semantically similar words having similar vectors, which is not the case in one-hot vectors (each pair of such vectors wi,wj has cosine similarity cos(wi,wj)=0).

➢ If feature vector contains one-hot vectors of the documents’ words, we will only be able to consider features we’ve seen during training; when we use embedding, semantically similar words will create similar features, and will lead to similar classification.

Encoding the categorical data

- **One hot vectors vs Word embedding(word2vec, GloVe)**

➢ Example. Let’s say that your classifier works with the bag-of-words approach, i.e. the feature vector is the sum of all the document’s word vectors (which is equivalent to a vector that counts the number of occurrences of each word in the vocabulary in the one-hot representation).

➢ Suppose that in your training data you have a document with the single word ***school***, and your test data contains a document with the single word ***education***, that wasn’t previously observed during training. In the one-hot vector representation the feature vectors of these two instances will be completely different, possibly leading to different class prediction, while in word embedding representation, they will be similar, hopefully leading to the same classification.

References

- Data Integration tool- Talend Open Studio

➢ https://www.talend.com/resources/introduction-talend-open-studio-data-integration/

➢ https://info.talend.com/rs/talend/images/WP_EN_DI_Talend_Definitive_Guide_DataIntegration.pdf

- Data wrangling tool

➢ http://vis.stanford.edu/wrangler/

- One hot encoding

➢ https://colab.research.google.com/github/alzayats/Google_Colab/blob/master/6_1_one_hot_encoding_of_words_or_characters.ipynb#scrollTo=xaiRJYibT-u3

Thank you.