SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A+' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

COURSE NAME : 19CSO302 & Introduction to Data Analytics

III YEAR/ VI SEMESTER

UNIT – I INTRODUCTION

Topic: Data Science Process

P.Poonkodi

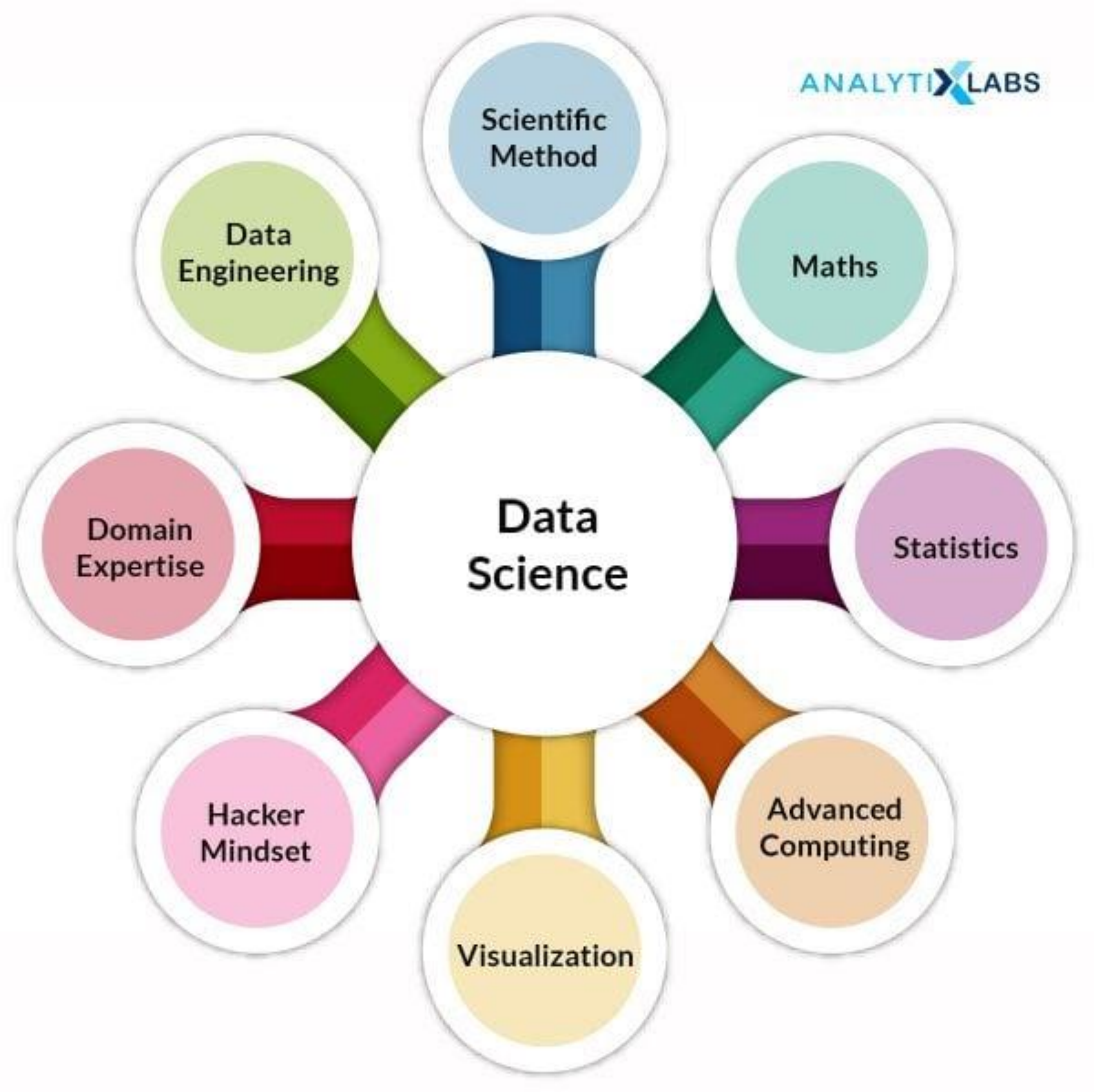
Assistant Professor

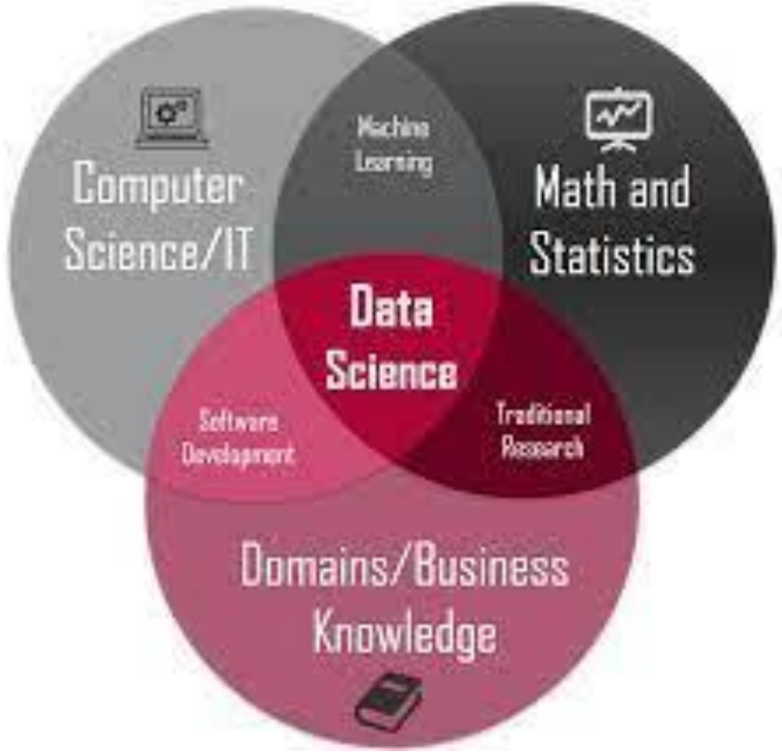
Department of Computer Science and Engineering

Data Science Process

- Real time example
- Roles and Stages
- Working with files and databases
- Exploring and managing data

Real time example





Introduction

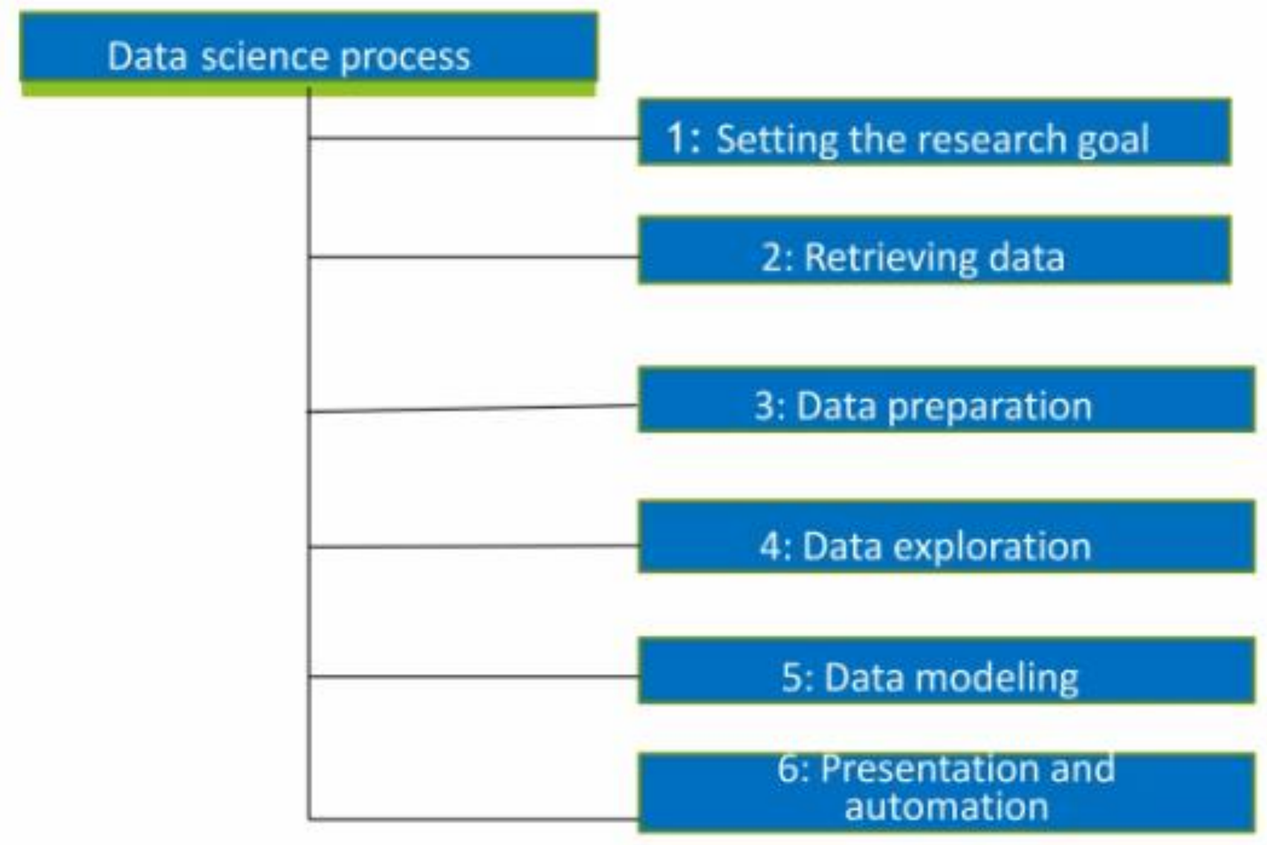
- Data science is the study of data to extract meaningful insights for business
- It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data

Uses

- Data science is used for a wide range of applications, including predictive analytics, machine learning, data visualization, recommendation systems, fraud detection, sentiment analysis, and decision-making in various industries like healthcare, finance, marketing, and technology

Stages

The data science process typically consists of six steps, as you can see in the mind map



Data science process

1: Setting the research goal

2: Retrieving data

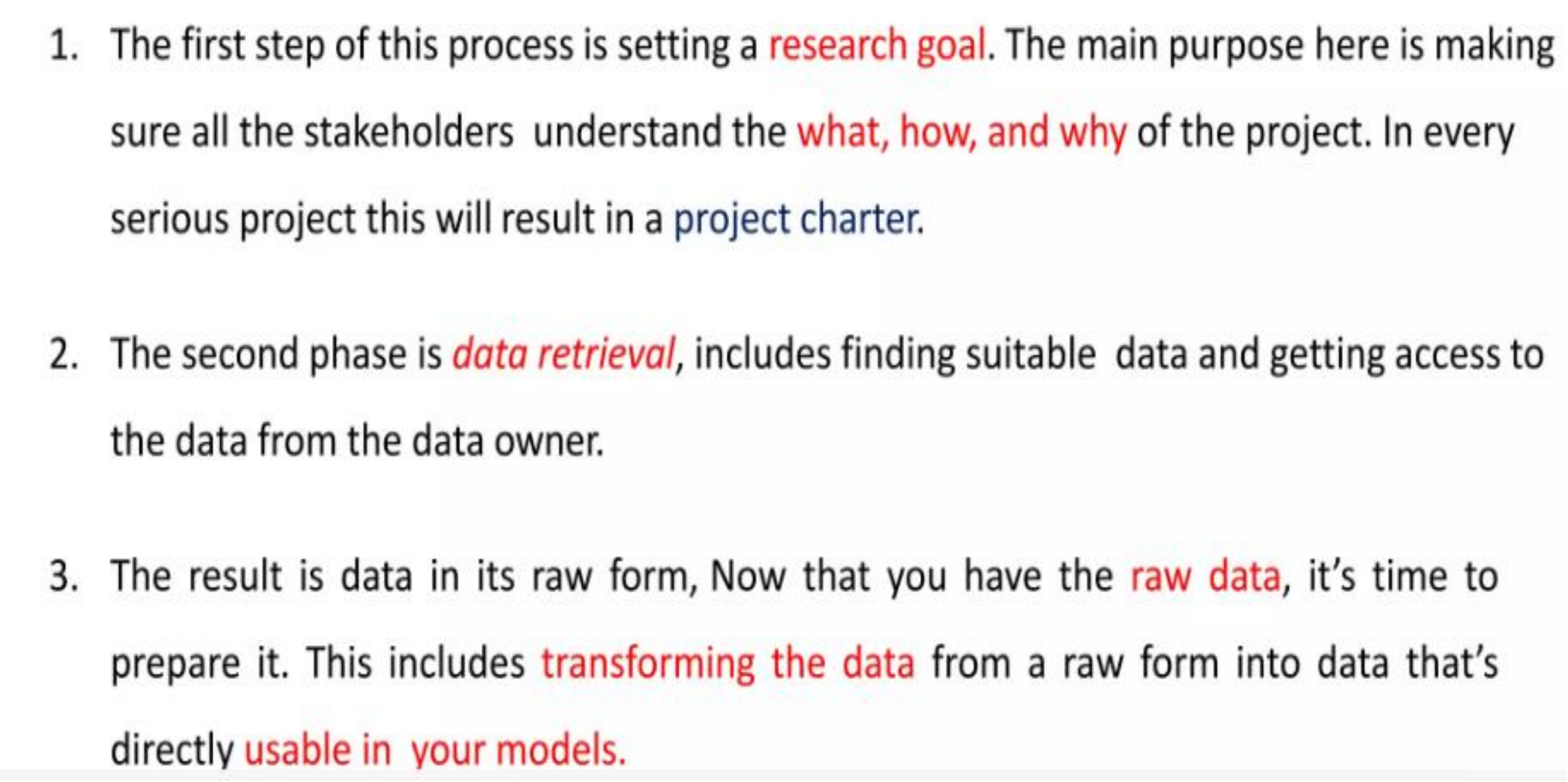
3: Data preparation

4: Data exploration

5: Data modeling

6: Presentation and automation

Data Science Process



1. The first step of this process is setting a research goal. The main purpose here is making sure all the stakeholders understand the what, how, and why of the project. In every serious project this will result in a project charter.

2. The second phase is *data retrieval*, includes finding suitable data and getting access to the data from the data owner.

3. The result is data in its raw form, Now that you have the raw data, it's time to prepare it. This includes transforming the data from a raw form into data that's directly usable in your models.

Contd.,

4. The fourth step is *data exploration*. The goal of this step is to gain a deep understanding of the data, look for patterns, correlations, and deviations based on visual and descriptive techniques.

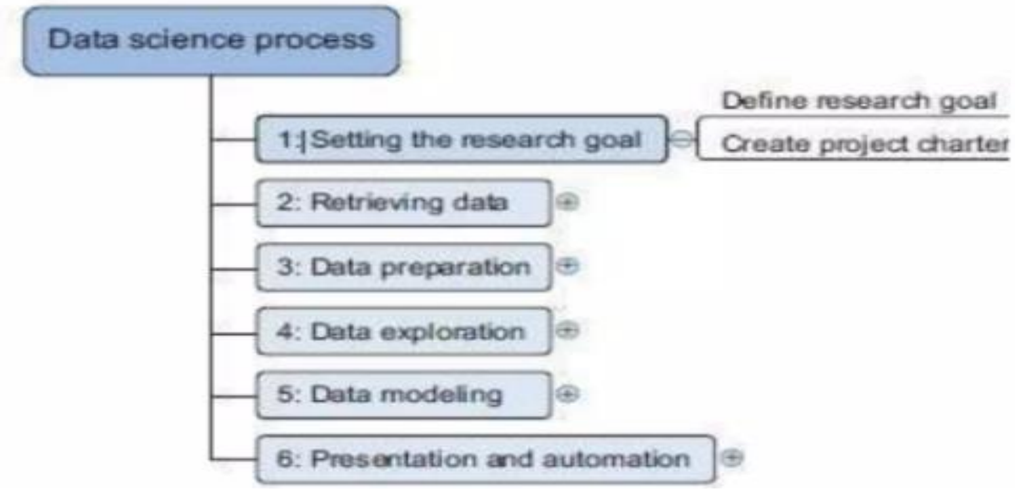
5. Finally: *model building* (often referred to as “data modeling”)- present the results to your business..

6. The last step of the data science model is *presenting your results and automating the analysis*, if needed. One goal of a project is to change a

Data Preparation.,

- Data collection is an error-prone process; in this phase you enhance the quality of the data and prepare it for use in subsequent steps. This phase consists of three subphases:
- Data cleansing removes false values from a data source and inconsistencies across data sources, Data Transformations, Data integration enriches data sources by combining information.

STEP 1: Defining Research Goals and Creating A Project Charter



- A project starts by understanding the *what*, the *why*, and the *how* of your project. The outcome should be a clear research goal, a good understanding of the context, well-defined deliverables, and a plan of action with a timetable. This information is then best placed in a project charter.

Understanding the goals and context of your research

- Understanding the business goals and context is critical for project success.

Create a project charter

A project charter requires teamwork,

- A clear research goal
- The project mission and context
- How you're going to perform your analysis
- What resources you expect to use
- Proof that it's an achievable project, or proof of concepts-idea turned to reality.
- Deliverables and a measure of success
- A timeline

to make an estimation of the project costs and the data and people required for your project to become a success.

Understanding the goals and context of your research

- Understanding the business goals and context is critical for project success.

Create a project charter

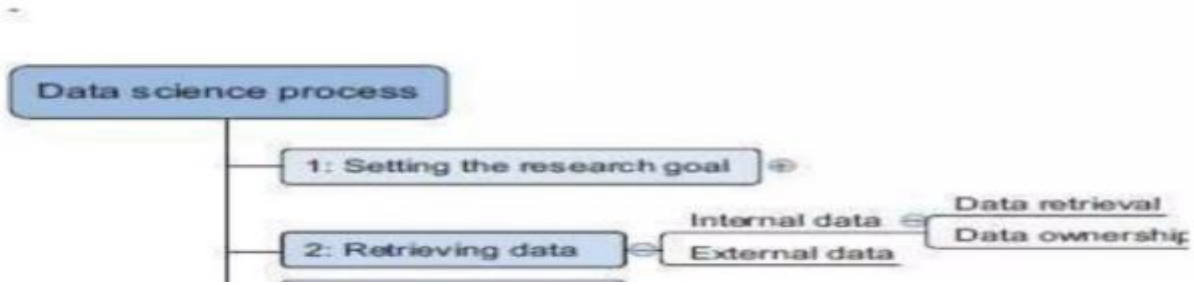
A project charter requires teamwork,

- A clear research goal
- The project mission and context
- How you’re going to perform your analysis
- What resources you expect to use
- Proof that it’s an achievable project, or proof of concepts-idea turned to reality.
- Deliverables and a measure of success
- A timeline

to make an estimation of the project costs and the data and people required for your project to become a success.

STEP 2: Retrieving Data

- The next step in data science is to retrieve the required data. Some times we need to go into the field and design a data collection process ourselves.



Data science process

1: Setting the research goal

2: Retrieving data

Internal data

Data retrieval

Data ownershi

External data

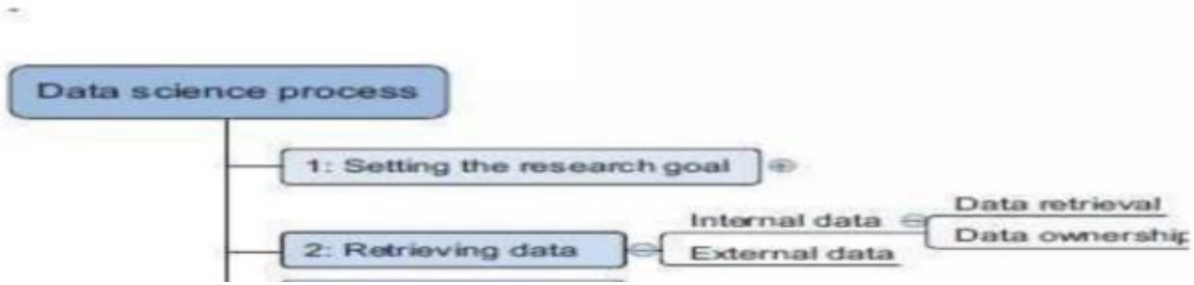
- Data can be stored in many forms, ranging from simple text files to tables in a database.

- The objective now is acquiring all the data you need.

- **Example:** Data is often like a diamond in the rough: it needs polishing to be of any use to you.

STEP 2: Retrieving Data

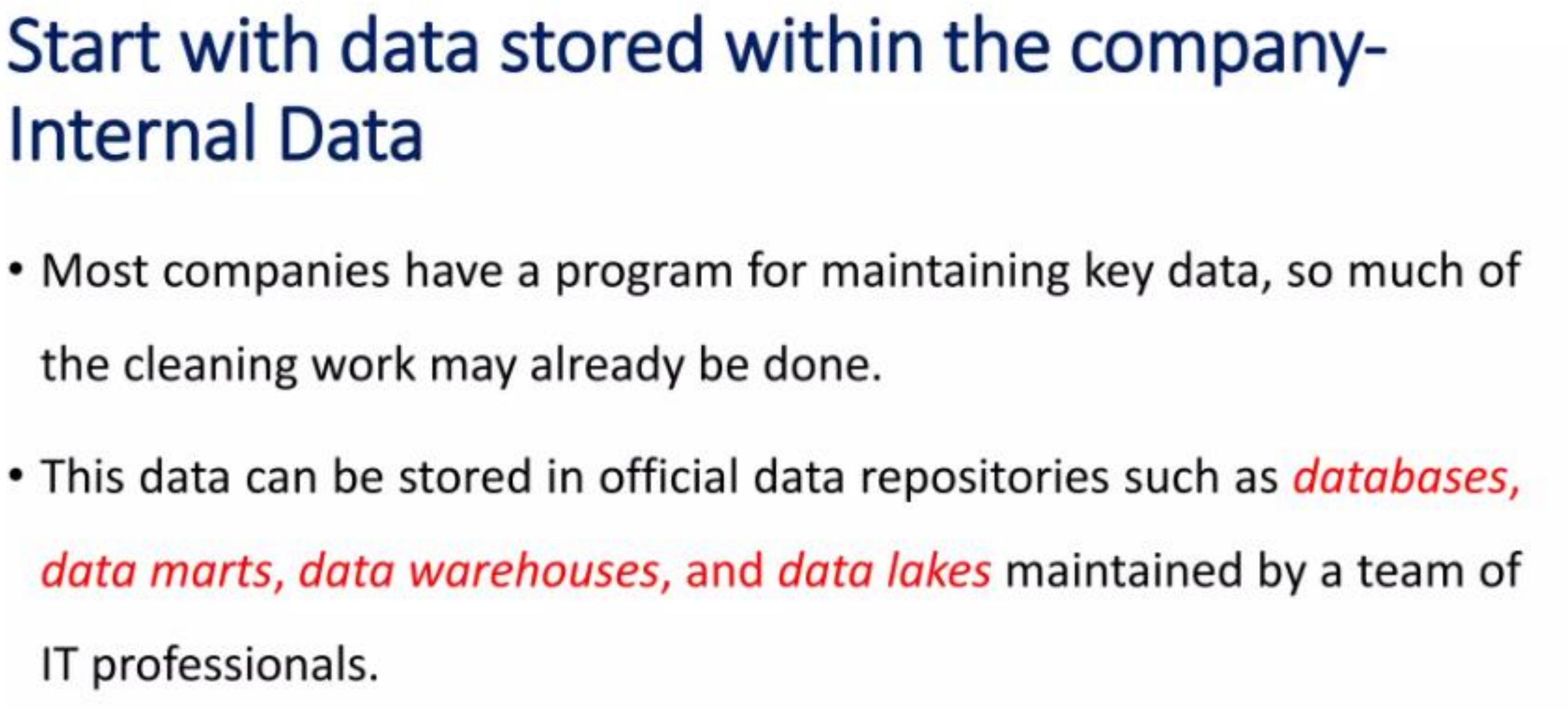
- The next step in data science is to retrieve the required data. Some times we need to go into the field and design a data collection process ourselves.



- Data can be stored in many forms, ranging from simple text files to tables in a database.

- The objective now is acquiring all the data you need.

- **Example:** Data is often like a diamond in the rough: it needs polishing to be of any use to you.



Start with data stored within the company-Internal Data

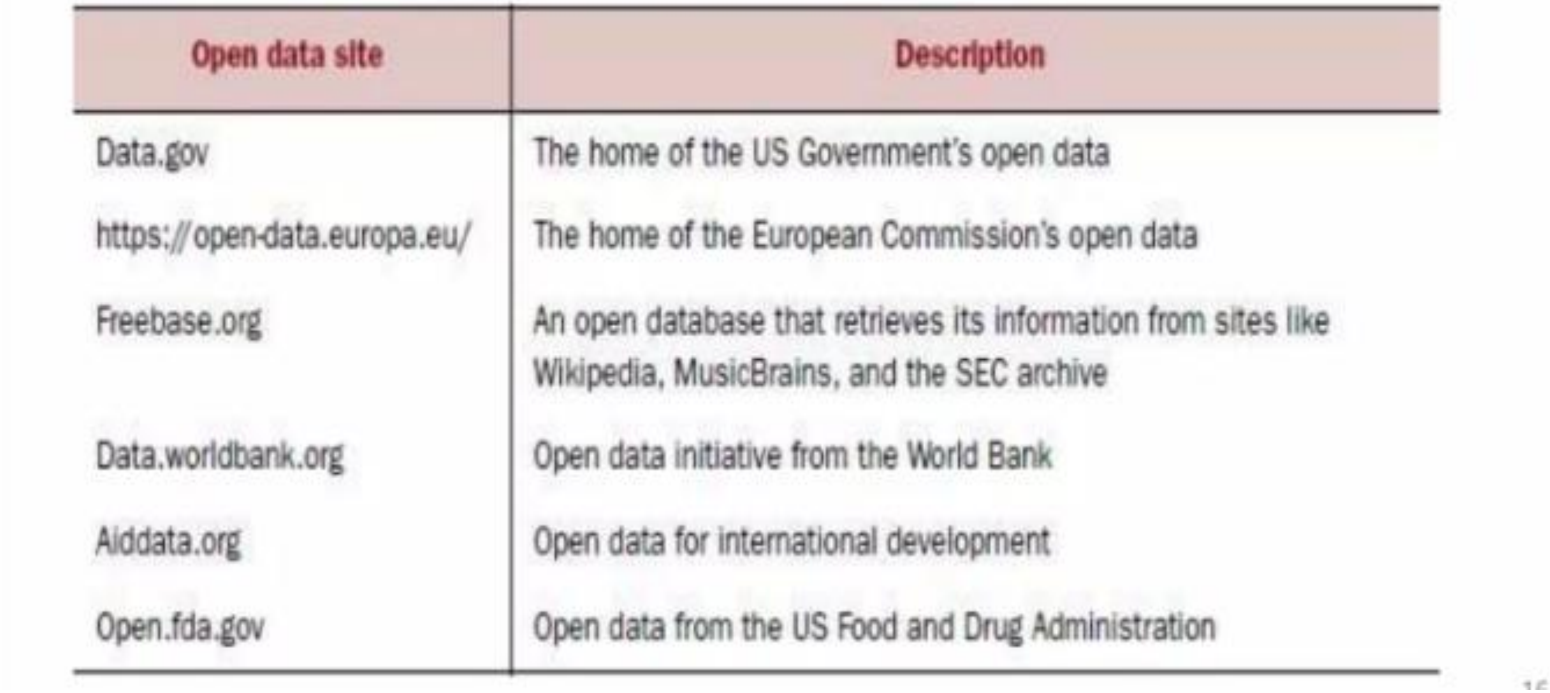
- Most companies have a program for maintaining key data, so much of the cleaning work may already be done.
- This data can be stored in official data repositories such as *databases, data marts, data warehouses,* and *data lakes* maintained by a team of IT professionals.

- The primary goal of a database is data storage, while a data warehouse is designed for reading and analyzing that data.
- A data mart is a subset of the data warehouse and geared toward serving a specific business unit.
- While data warehouses and data marts are home to preprocessed data, data lakes contains data in its natural or raw format.
- But the possibility exists that your data still resides in Excel files on the desktop of a domain expert.

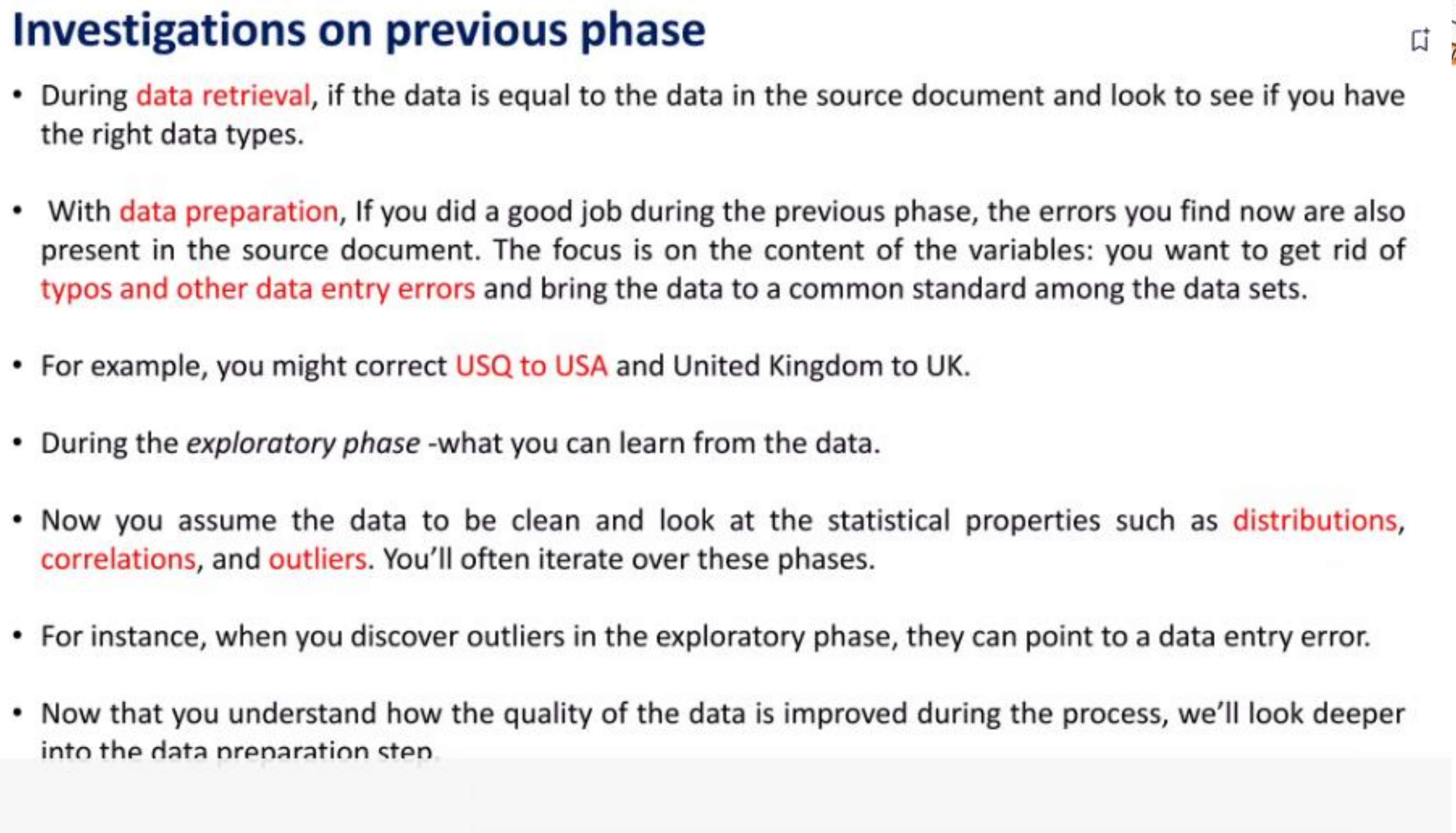
- Finding data even within your own company can sometimes be a challenge. As companies grow, their data becomes scattered around many places. Knowledge of the data may be dispersed as people change positions and leave the company.
- Getting access to data is another difficult task. Organizations understand the value and sensitivity of data and often have policies in place so everyone has access to what they need and nothing more. These policies translate into physical and digital barriers called Chinese walls. These “walls” are mandatory and well-regulated for customer data in most countries.

Don’t be afraid to shop around-External Data

- If data isn’t available inside your organization, look outside your organizations. Companies provide data so that you, in turn, can enrich their services and ecosystem. Such is the case with Twitter, LinkedIn, and Facebook.
- More and more governments and organizations share their data for free with the world.
- A list of open data providers that should get you started.



Open data site	Description
Data.gov	The home of the US Government’s open data
https://open-data.europa.eu/	The home of the European Commission’s open data
Freebase.org	An open database that retrieves its information from sites like Wikipedia, MusicBrains, and the SEC archive
Data.worldbank.org	Open data initiative from the World Bank
Aiddata.org	Open data for international development
Open.fda.gov	Open data from the US Food and Drug Administration

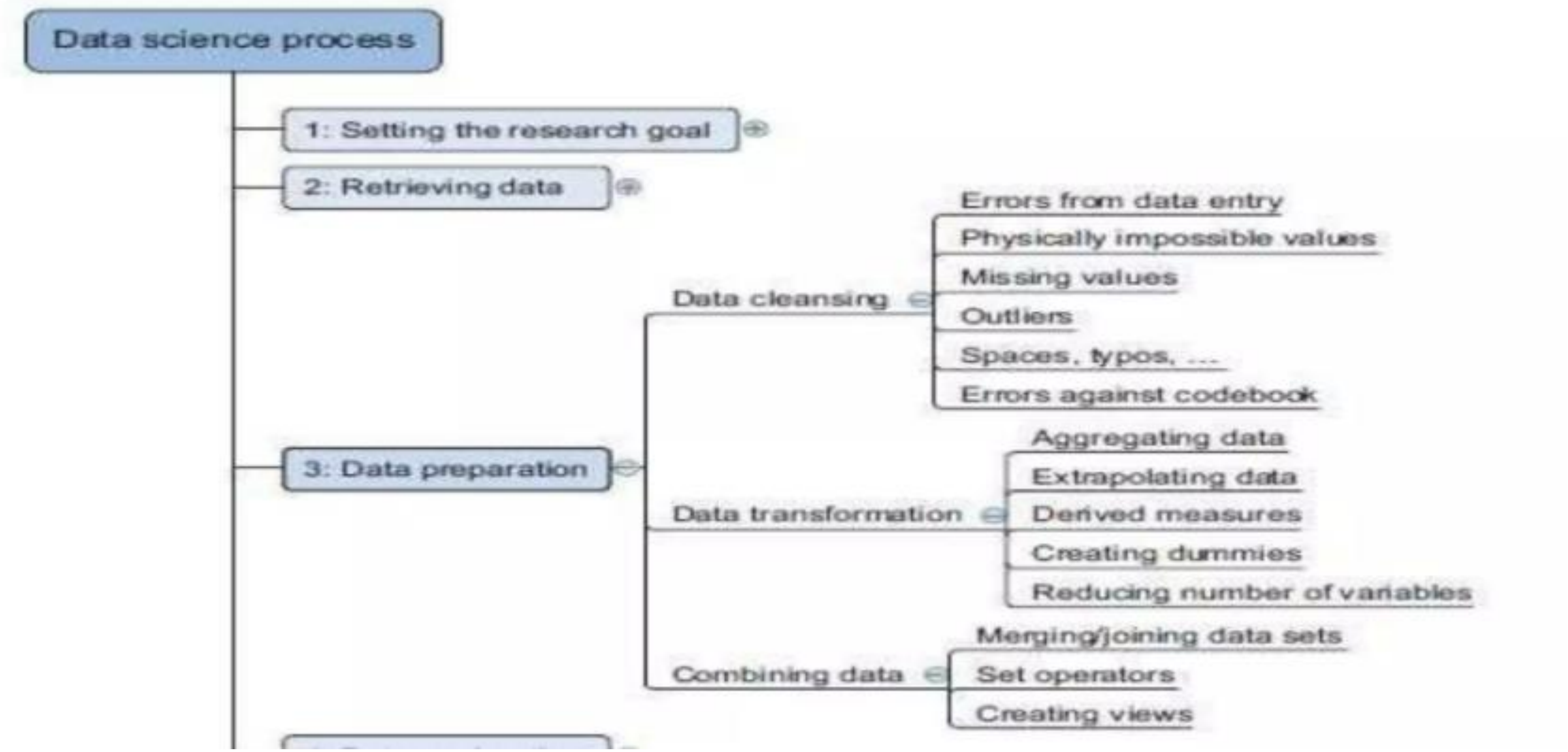


Investigations on previous phase

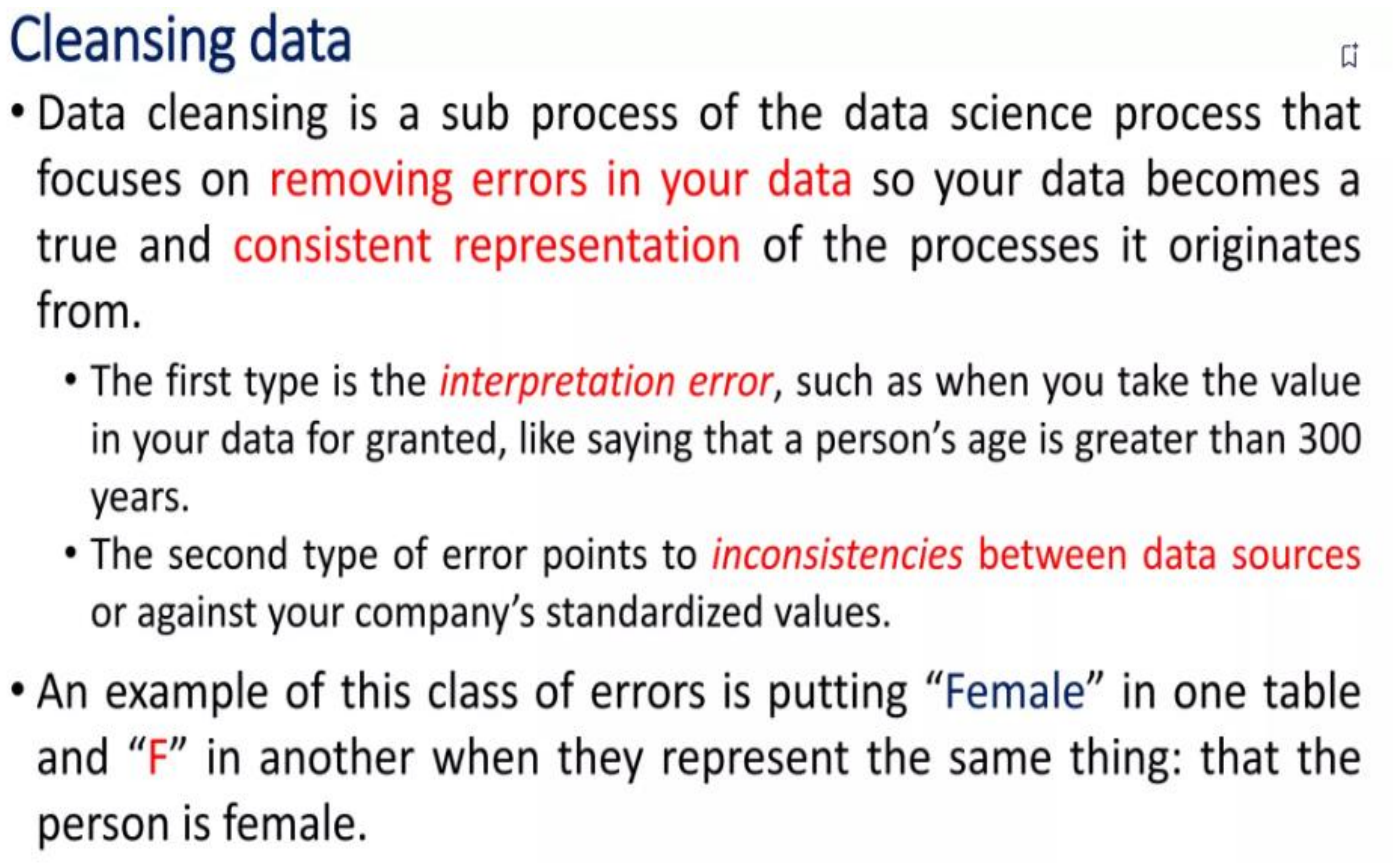
- During data retrieval, if the data is equal to the data in the source document and look to see if you have the right data types.
- With data preparation, If you did a good job during the previous phase, the errors you find now are also present in the source document. The focus is on the content of the variables: you want to get rid of typos and other data entry errors and bring the data to a common standard among the data sets.
- For example, you might correct USQ to USA and United Kingdom to UK.
- During the *exploratory phase* -what you can learn from the data.
- Now you assume the data to be clean and look at the statistical properties such as distributions, correlations, and outliers. You'll often iterate over these phases.
- For instance, when you discover outliers in the exploratory phase, they can point to a data entry error.
- Now that you understand how the quality of the data is improved during the process, we'll look deeper into the data preparation step.

Step 3: Cleansing, Integrating, And Transforming Data

- The data received from the data retrieval phase is likely to be “a diamond in the rough.” Task now is to sanitize and prepare it for use in the modeling and reporting phase.



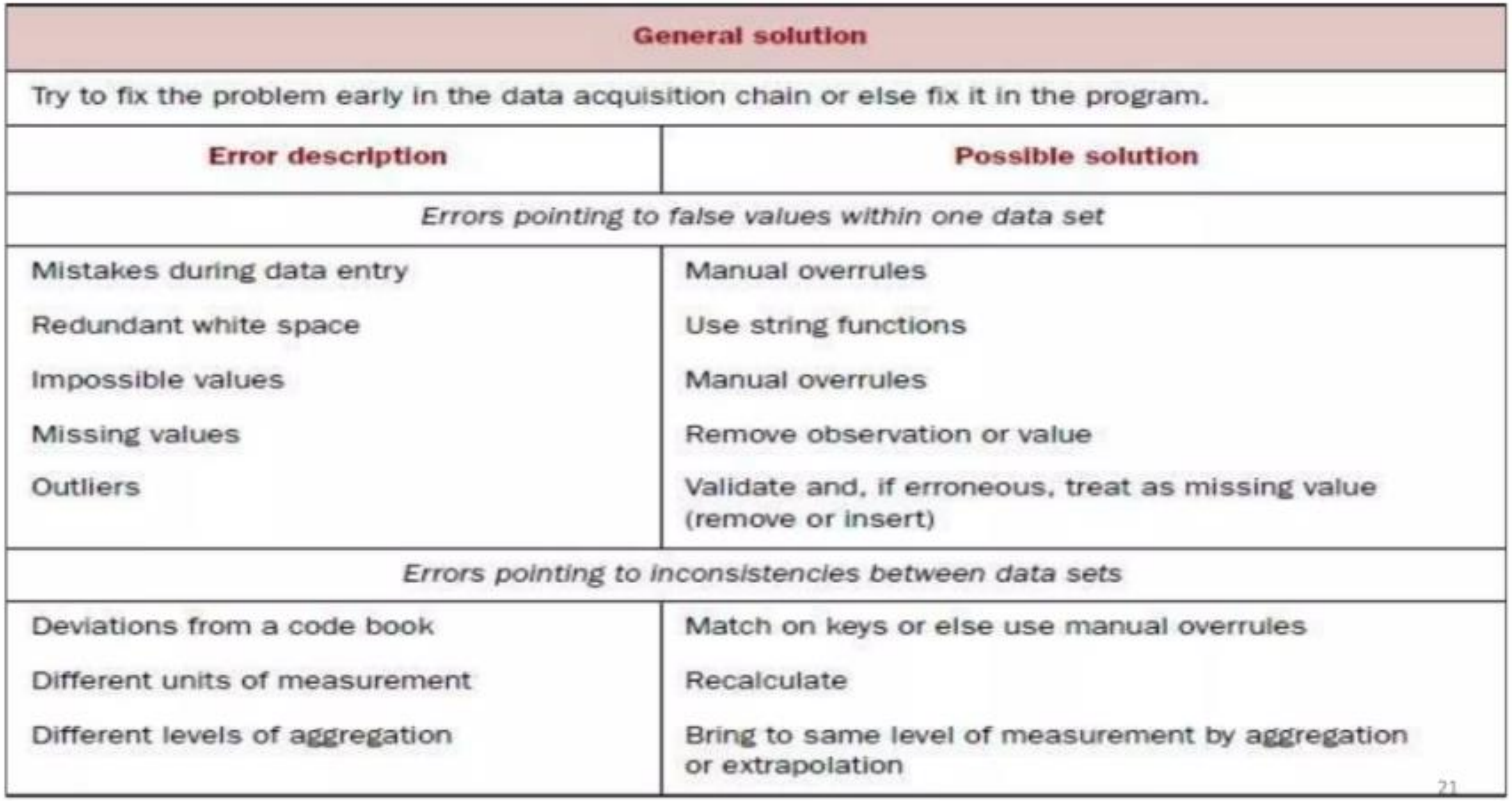
- Data science process
  - 1: Setting the research goal
  - 2: Retrieving data
  - 3: Data preparation
    - Data cleansing
      - Errors from data entry
      - Physically impossible values
      - Missing values
      - Outliers
      - Spaces, typos, ...
      - Errors against codebook
    - Data transformation
      - Aggregating data
      - Extrapolating data
      - Derived measures
      - Creating dummies
      - Reducing number of variables
    - Combining data
      - Merging/joining data sets
      - Set operators
      - Creating views



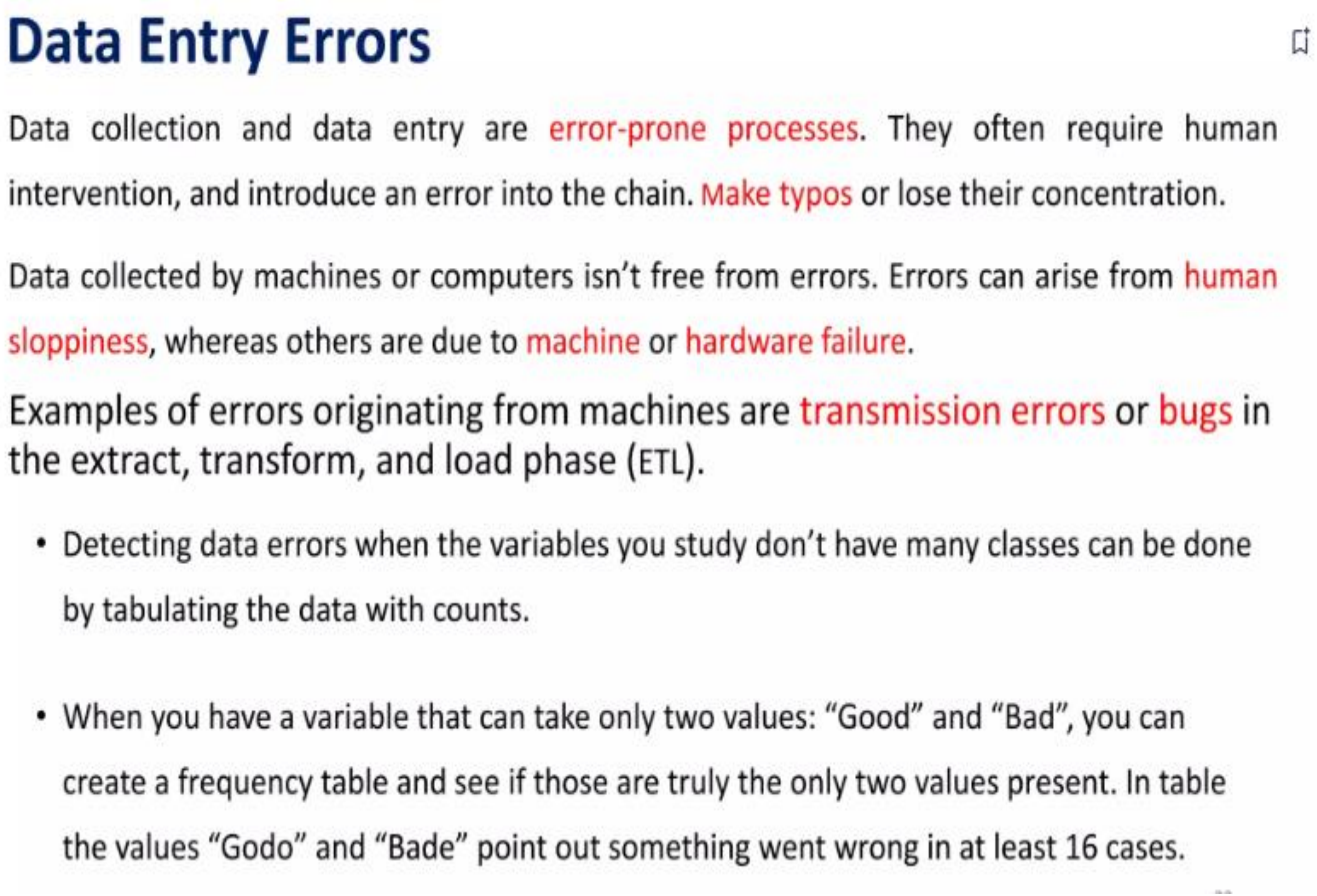
Cleansing data

- Data cleansing is a sub process of the data science process that focuses on removing errors in your data so your data becomes a true and consistent representation of the processes it originates from.
  - The first type is the *interpretation error*, such as when you take the value in your data for granted, like saying that a person’s age is greater than 300 years.
  - The second type of error points to *inconsistencies* between data sources or against your company’s standardized values.
- An example of this class of errors is putting “Female” in one table and “F” in another when they represent the same thing: that the person is female.

Common Errors



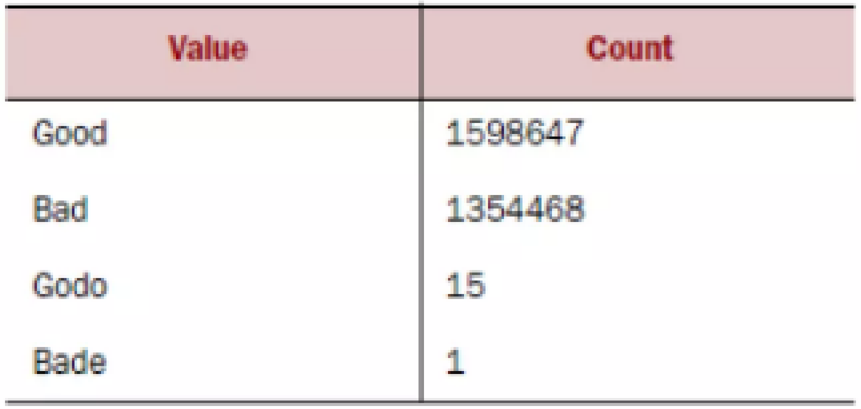
General solution	
Try to fix the problem early in the data acquisition chain or else fix it in the program.	
Error description	**Possible solution**
Errors pointing to false values within one data set	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
Errors pointing to inconsistencies between data sets	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation



Data Entry Errors

- Data collection and data entry are error-prone processes. They often require human intervention, and introduce an error into the chain. Make typos or lose their concentration.
- Data collected by machines or computers isn't free from errors. Errors can arise from human sloppiness, whereas others are due to machine or hardware failure.
- Examples of errors originating from machines are transmission errors or bugs in the extract, transform, and load phase (ETL).
  - Detecting data errors when the variables you study don't have many classes can be done by tabulating the data with counts.
  - When you have a variable that can take only two values: "Good" and "Bad", you can create a frequency table and see if those are truly the only two values present. In table the values "Godo" and "Bade" point out something went wrong in at least 16 cases.

Contd.,

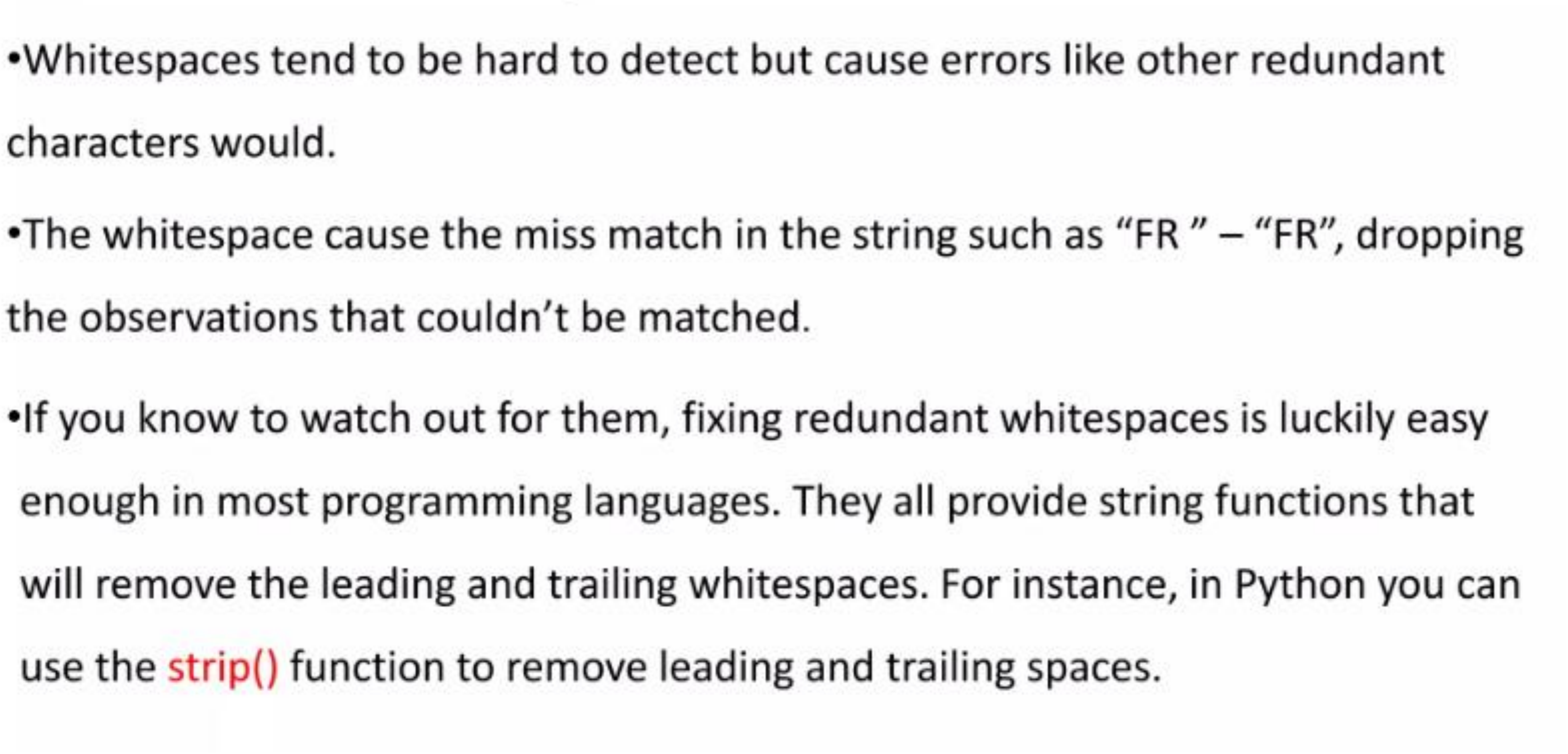


Value	Count
Good	1598647
Bad	1354468
Godo	15
Bade	1

Most errors of this type are easy to fix with simple assignment statements and if-then else rules:

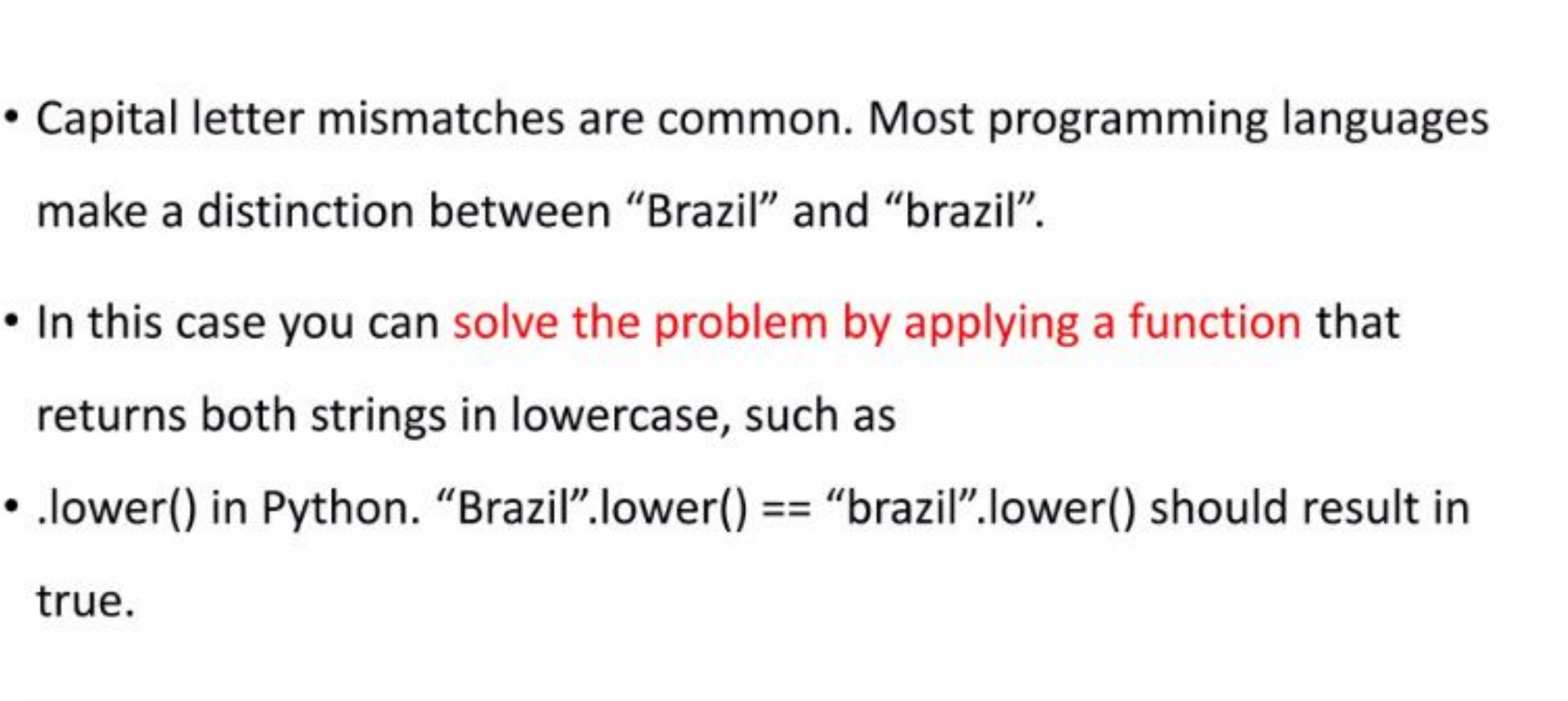
```
if x == “Godo”:
x = “Good”
if x == “Bade”:
x = “Bad”
```

Redundant Whitespace

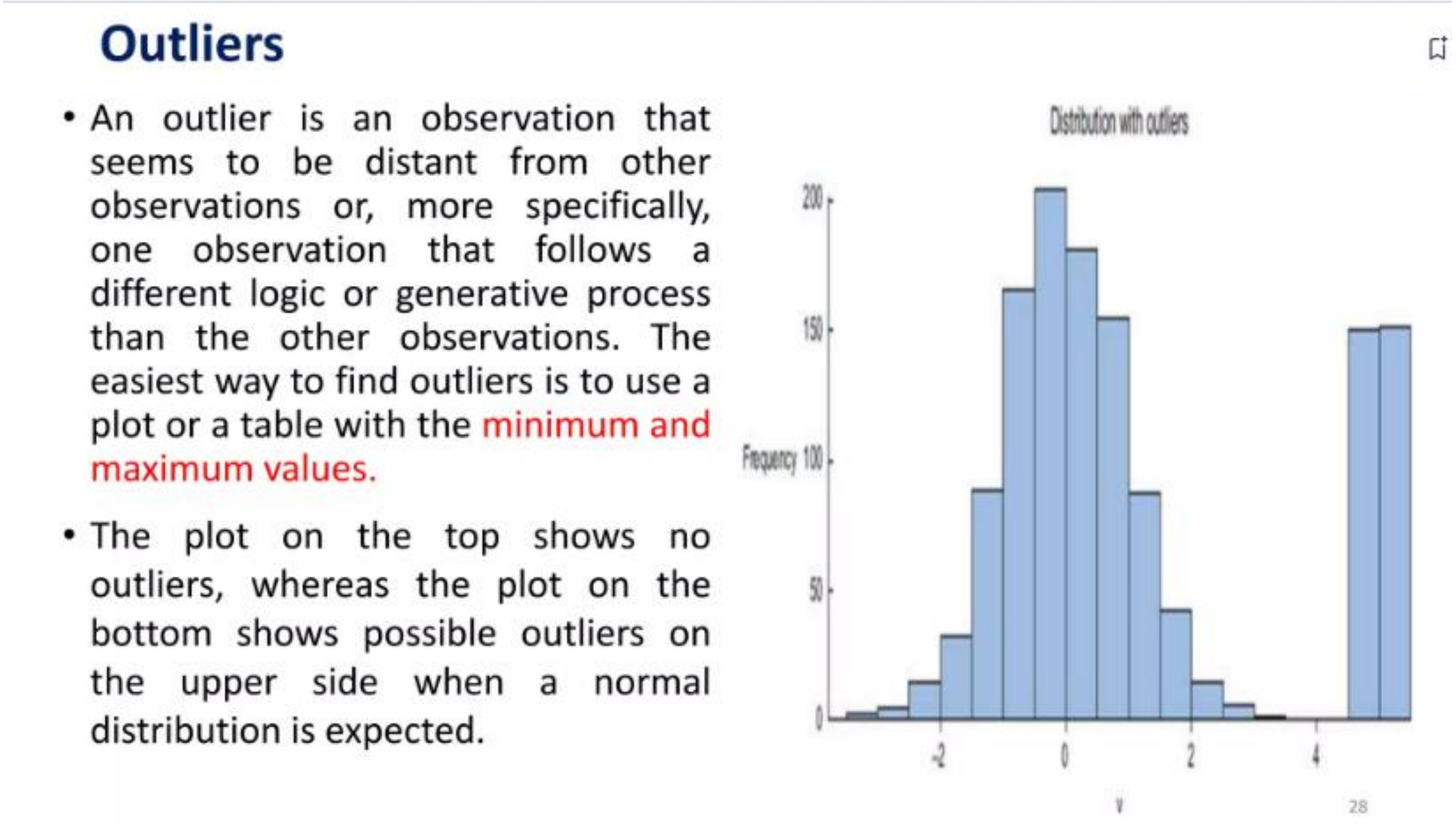


- Whitespaces tend to be hard to detect but cause errors like other redundant characters would.
- The whitespace cause the miss match in the string such as “FR ” – “FR”, dropping the observations that couldn’t be matched.
- If you know to watch out for them, fixing redundant whitespaces is luckily easy enough in most programming languages. They all provide string functions that will remove the leading and trailing whitespaces. For instance, in Python you can use the strip() function to remove leading and trailing spaces.

Fixing Capital Letter Mismatches



- Capital letter mismatches are common. Most programming languages make a distinction between “Brazil” and “brazil”.
- In this case you can solve the problem by applying a function that returns both strings in lowercase, such as
- .lower() in Python. “Brazil”.lower() == “brazil”.lower() should result in true.



Outliers

- An outlier is an observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations. The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.
- The plot on the top shows no outliers, whereas the plot on the bottom shows possible outliers on the upper side when a normal distribution is expected.

Dealing with Missing Values

- Missing values aren’t necessarily wrong, but you still need to handle them separately; certain modeling techniques can’t handle missing values. They might be an indicator that something went wrong in your data collection or that an error happened in the ETL process. Common techniques data scientists use are listed in table.

Step 4: Exploratory Data Analysis

- During exploratory data analysis you take a deep dive into the data.
- Information becomes much easier to grasp when shown in a picture, therefore you mainly use graphical techniques to gain an understanding of your data and the interactions between variables.
- Bar Plot, Line Plot, Scatter Plot ,Multiple Plots , Pareto Diagram , Link and Brush Diagram ,Histogram , Box and Whisker Plot .

Step 5: Build the Models

- Build the models are the next step, with the goal of making better predictions, classifying objects, or gaining an understanding of the system that are required for modeling.

Step 6: Presenting findings and building applications on top of them –

- The last stage of the data science process is where your soft skills will be most useful, and yes, they’re extremely important.
- Presenting your results to the stakeholders and industrializing your analysis process for repetitive reuse and integration with other tools.

Usage of Data Science Process

- The Data Science Process is a systematic approach to solving data-related problems and consists of the following steps:
- **Problem Definition:** Clearly defining the problem and identifying the goal of the analysis.
- **Data Collection:** Gathering and acquiring data from various sources, including data cleaning and preparation.
- **Data Exploration:** Exploring the data to gain insights and identify trends, patterns, and relationships.
- **Data Modeling:** Building mathematical models and algorithms to solve problems and make predictions.
- **Evaluation:** Evaluating the model’s performance and accuracy using appropriate metrics.
- **Deployment:** Deploying the model in a production environment to make predictions or automate decision-making processes.
- **Monitoring and Maintenance:** Monitoring the model’s performance over time and making updates as needed to improve accuracy.

Issues of Data Science Process

- **Data Quality and Availability**: Data quality can affect the accuracy of the models developed and therefore, it is important to ensure that the data is accurate, complete, and consistent. Data availability can also be an issue, as the data required for analysis may not be readily available or accessible.
- **Bias in Data and Algorithms**: Bias can exist in data due to sampling techniques, measurement errors, or imbalanced datasets, which can affect the accuracy of models. Algorithms can also perpetuate existing societal biases, leading to unfair or discriminatory outcomes.
- **Model Overfitting and Underfitting**: Overfitting occurs when a model is too complex and fits the training data too well, but fails to generalize to new data. On the other hand, underfitting occurs when a model is too simple and is not able to capture the underlying relationships in the data.
- **Model Interpretability**: Complex models can be difficult to interpret and understand, making it challenging to explain the model’s decisions and decisions. This can be an issue when it comes to making business decisions or gaining stakeholder buy-in.
- **Privacy and Ethical Considerations**: Data science often involves the collection and analysis of sensitive personal information, leading to privacy and ethical concerns. It is important to consider privacy implications and ensure that data is used in a responsible and ethical manner.
- **Technical Challenges**: Technical challenges can arise during the data science process such as data storage and processing, algorithm selection, and computational scalability.

References

- Runkler TA, “Data Analytics: Models and algorithms for intelligent data analysis”, Springer, Third Edition 2020

Thank You!