Unit II:

Data types: Quantitative, Qualitative, relationships: Ranking, Deviation, Nominal comparisons, Correlation, Partial and total relationships, Series over time.

Data Science Process: Research Goals- Retrieving data- Cleansing, integrating, and transforming data- Exploratory data analysis- Build the models

Data types, relationships, and visualization formats

There are a number of methods and approaches to creating visuals based on the nature and complexity of the data and the information. **Different kinds of graphics are used in data visualizations, including representations of statistics, maps, and diagrams.** These schematic, visual representations of content vary in their degree of abstraction.

In order to communicate effectively, it is important to understand different kinds of data and to establish visual relationships through the proper use of graphics. Enrique Rodríguez (2012), a data analyst at DataNauta, once explained in an interview that...

> **A good graphic is one that synthesizes and contextualizes all of the information that's necessary to understand a situation and decide how to move forward.”**

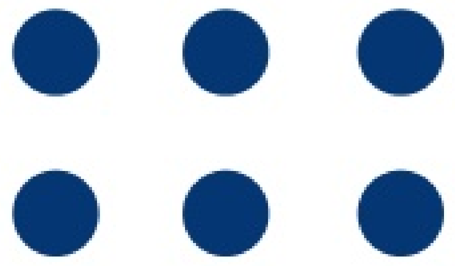
2 kinds of data

Before we talk about visuals themselves, we must first understand the different kinds of data that can be visualized and how they relate to one another. The most common kinds of data are[4]:

1) Quantitative (numeric)

Data that can be quantified and measured. This kind of data explains a trend or the results of research through numeric values. This category of data can be further subdivided into:

- **Discrete:** Data that consists of whole numbers (0, 1, 2, 3...). For example, the number of children in a family.
- **Continuous:** Data that can take any value within an interval. For example, people's height (between 60 - 70 inches) or weight (between 90 and 110 pounds).

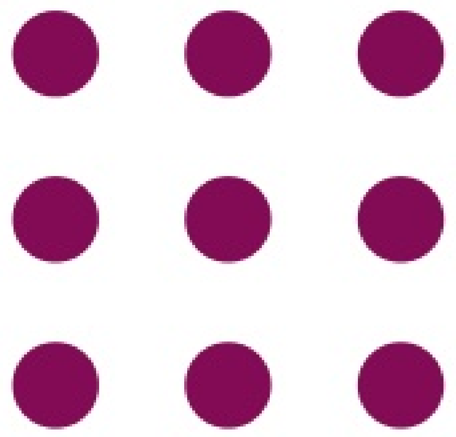


Quantitative

2) Qualitative (categoric)

This kind of data is divided into categories based on non-numeric characteristics. It may or may not have a logical order, and it measures qualities and generates categorical answers. It can be:

- **Ordinal:** Meaning it follows an order or sequence. That might be the alphabet or the months of the year.
- **Categorical:** Meaning it follows no fixed order. For example, varieties of products sold.



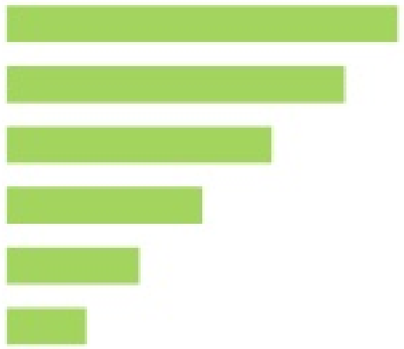
Qualitative

[5] Source: Hubspot, Prezy, and Infogram (2018). Presenting Data People Can't Ignore: How to Communicate Effectively Using Data. | p.10 of 16 | Available at: https://offers.hubspot.com/presenting-data-people-cant-ignore.

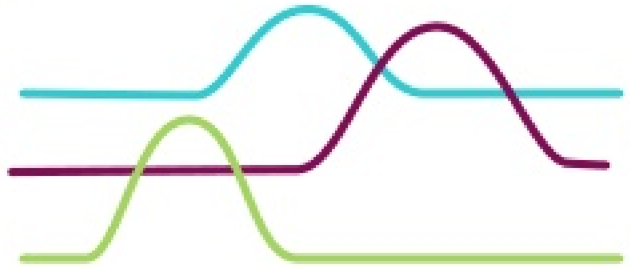
7 data relationships

Data relationships can be simple, like the progress of a single metric over time (such as visits to a blog over the course of 30 days or the number of users on a social network), or they can be complex, precisely comparing relationships, revealing structure, and extracting patterns from data. There are **seven data relationships** to consider:

Ranking: A visualization that relates two or more values with respect to a relative magnitude. For example: a company's most sold products.



Deviation: Examines how each data point relates to the others and, particularly, to what point its value differs from the average. For example: the line of deviation for tickets to an amusement park sold on a rainy versus a normal day.



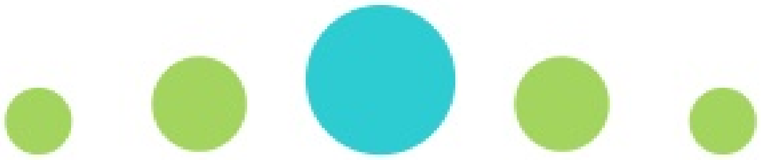
Nominal comparisons: Visualizations that compare quantitative values from different subcategories. For example: product prices in various supermarkets.

Correlation: Data with two or more variables that can demonstrate a positive or negative correlation with one another. For example: salaries based on level of education.

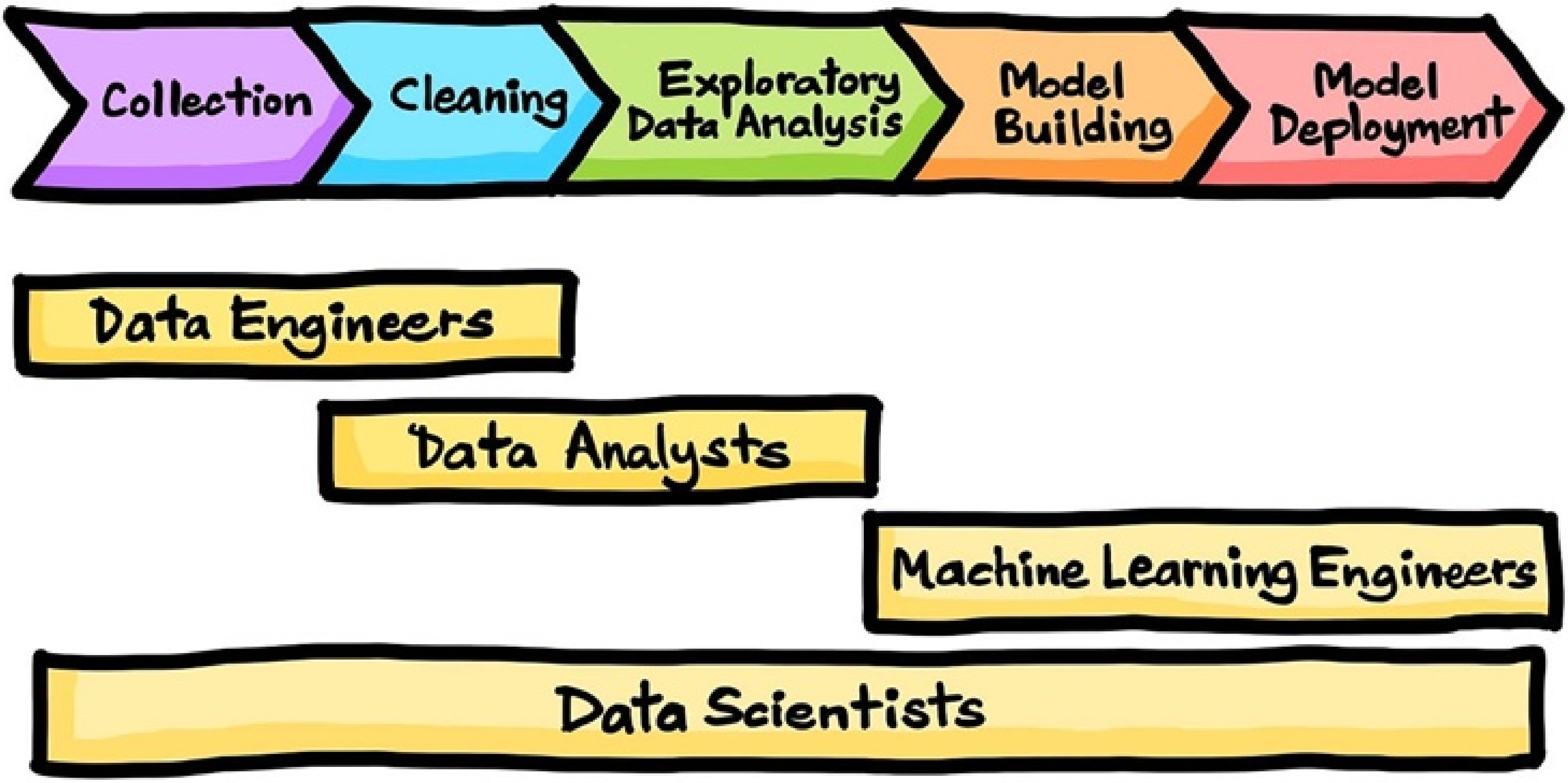
Partial and total relationships: Show a subset of data as compared with a larger total. For example: the percentage of clients that buy specific products.

Series over time: Here we can trace the changes in the values of a constant metric over the course of time. For example: monthly sales of a product over the course of two years.

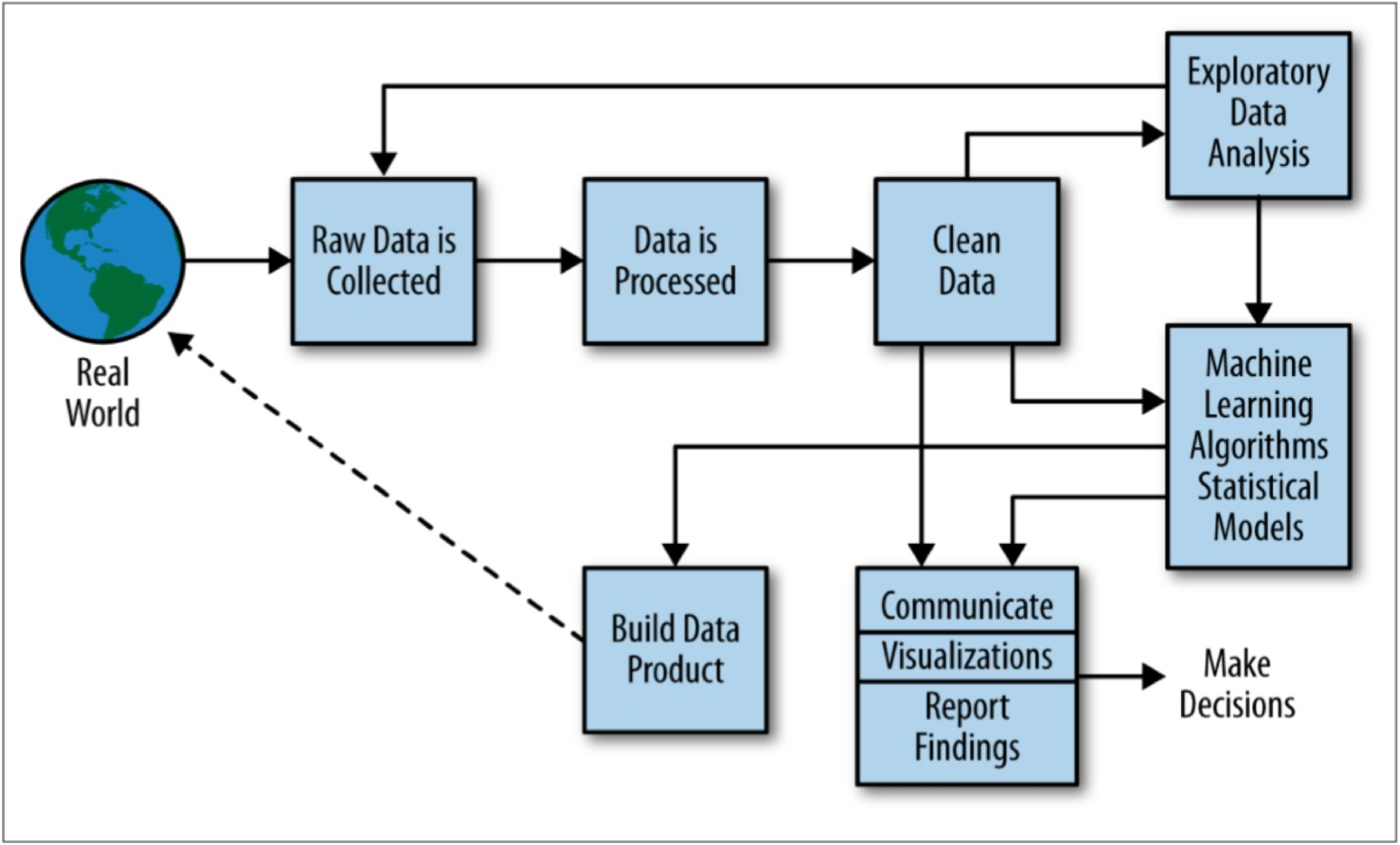
Distribution: Visualization that shows the distribution of data spatially, often around a central value. For example: the heights of players on a basketball team.



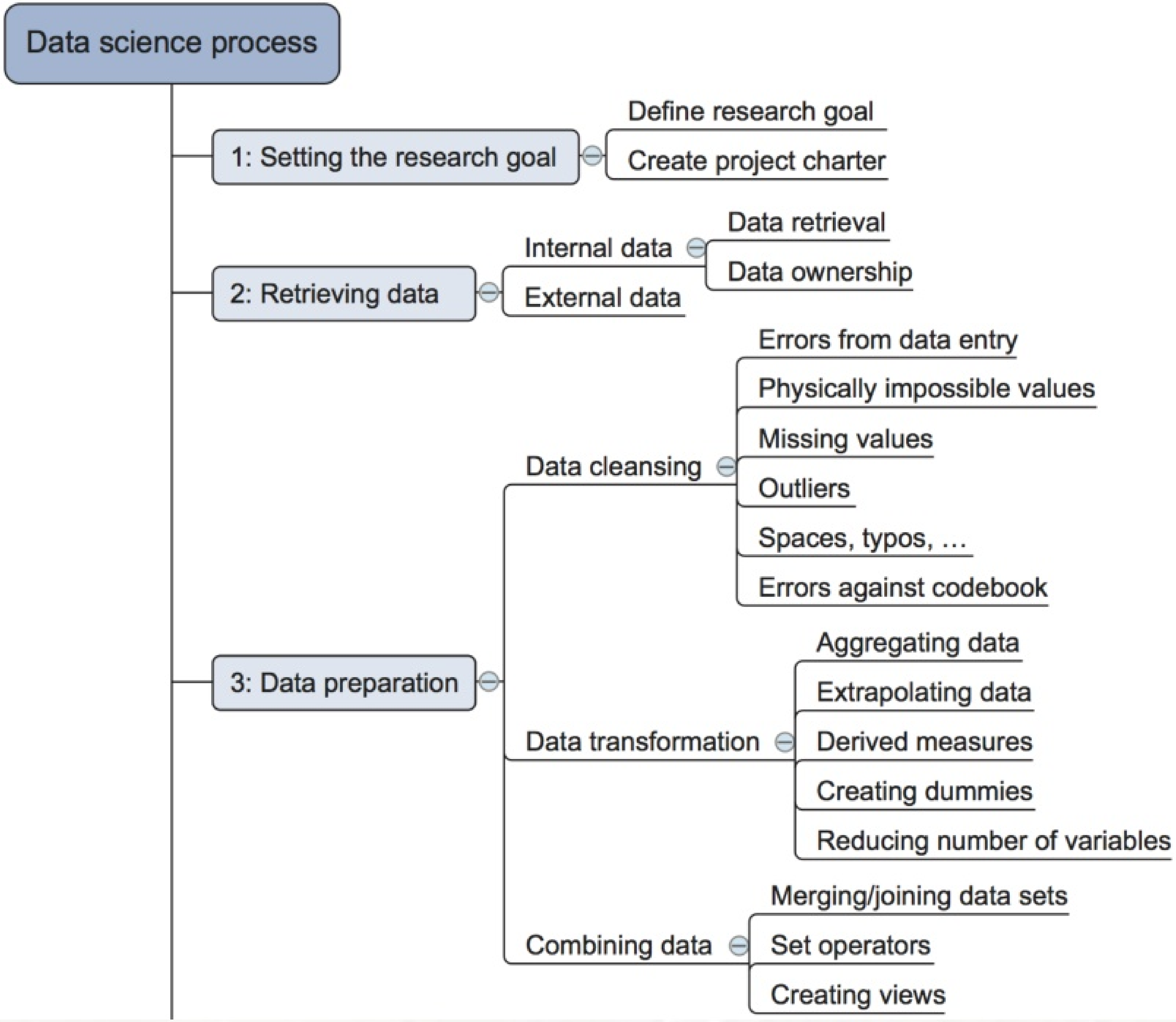
THE DATA SCIENCE PROCESS



It is a Process

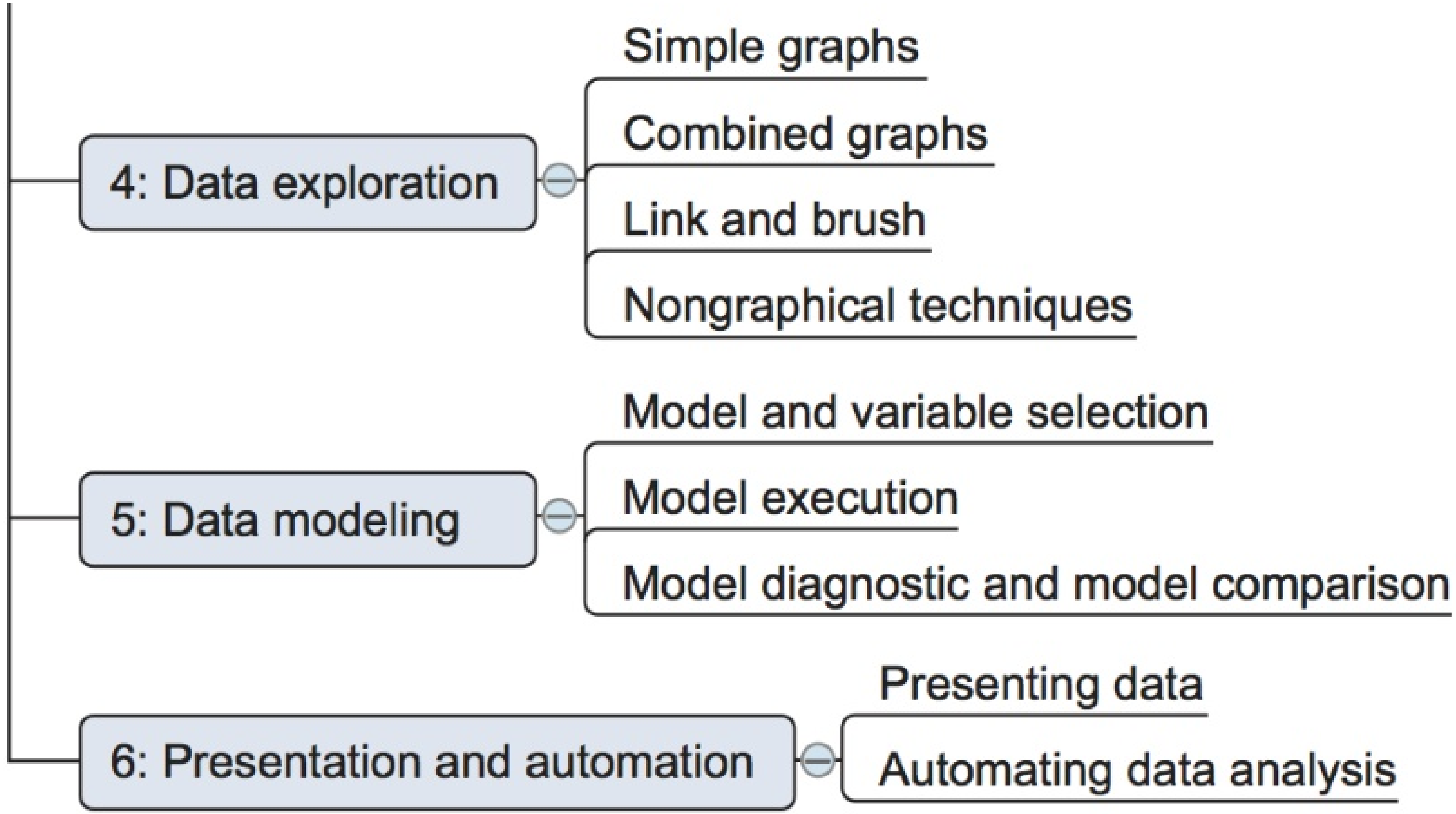


It is a Process (2)



- Data science process
  - 1: Setting the research goal
    - Define research goal
    - Create project charter
  - 2: Retrieving data
    - Internal data
      - Data retrieval
      - Data ownership
    - External 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

It is a Process (2)



4: Data exploration

- Simple graphs
- Combined graphs
- Link and brush
- Nongraphical techniques

5: Data modeling

- Model and variable selection
- Model execution
- Model diagnostic and model comparison

6: Presentation and automation

- Presenting data
- Automating data analysis

The Data Science Process is a systematic approach to solving data-related problems and consists of the following steps:

1. **Problem Definition:** Clearly defining the problem and identifying the goal of the analysis.
2. **Data Collection:** Gathering and acquiring data from various sources, including data cleaning and preparation.
3. **Data Exploration:** Exploring the data to gain insights and identify trends, patterns, and relationships.
4. **Data Modeling:** Building mathematical models and algorithms to solve problems and make predictions.
5. **Evaluation:** Evaluating the model's performance and accuracy using appropriate metrics.
6. **Deployment:** Deploying the model in a production environment to make predictions or automate decision-making processes.
7. **Monitoring and Maintenance:** Monitoring the model's performance over time and making updates as needed to improve accuracy.