Data Science

•Data Analysis is focused on answering specific, known questions about the past and present. It's about summarizing data, finding patterns, and creating reports to support immediate business decisions.

•Example: "What were our sales figures for Q2?" "Which marketing campaign generated the most leads last month?"

•

•Data Science is focused on discovering new questions to ask and predicting future outcomes. It uses advanced techniques to model data and prescribe actions.

•Example: "Based on past sales and customer behavior, what will our sales be next quarter, and which customers are most likely to churn?" "Can we build a system to recommend products to users in real-time?"

•Datafication refers to the process of turning various forms of information into structured digital data that can be analyzed and used to gain insights or inform decision-making.

This can involve collecting and processing data from a variety of sources, such as social media, IoT devices, and sensors, and storing it in databases for analysis

•The benefits of datafication are many:

•By analyzing data, organizations can make more informed decisions, optimize operations, and improve customer experiences.

For example, data analysis can help businesses identify patterns and trends in customer behavior, allowing them to tailor their products and services to meet the needs of their customers more effectively

•The 3 V's (volume, velocity and variety) are three defining properties or dimensions of big data . Volume refers to the amount of data, velocity refers to the speed of data processing, and variety refers to the number of types of data.

• Structured data is traditional data, ordered and conforming to a formal structure. It is data stored in relational database systems. For example, a bank statement includes the date, time, and amount.

• Semi-structured data is incompletely sorted data that deviates فرحني from the standard data structure: log files, JSON files, CSV files, etc.

• Unstructured data is unorganized data that cannot fit into relational databases: text files, emails, photos, movies, voice messages, audio files

Benefits of Data Science

- extracting insights and knowledge from data
- gain a competitive advantage
- identifying areas where costs can be reduced.
- uncover new insights and opportunities
- in the healthcare industry to improve patient diagnosis
- help companies identify and mitigate risksتخفيف المخاطر by analyzing data

Data •

Data are raw facts and figures with no fundamental meaning.
• Data plainly reports part of a situation without providing an interpretation. transactions.” • Data can be:
It is an unprocessed form of knowledge that does not convey value نقل القيمة or significance
• For data to have some useful meaning, it has to be organized, analyzed and interpreted تم تفسيره. In the context of an organization, “data are most usefully described as structured records of
• Quantitative كمي: when data can be counted or measured like cost, weight, and volume.
• Qualitative نوعي : when data describes things like name, color, and shape

Information is processed data.

• It is organized, classified, structured and provides meaningful and useful context.
• In contrast to data, information has meaning. “Data becomes information when its creator adds meaning. We transform data into information by adding value in various ways.”
• Here are several important processes that convert data to information:
• Calculation: data are mathematically or statistically scrutinised

- Categorisation: data are sorted into groups or classes
- Condensing: summarizing data to be more concise
- Contextualising: gathering data for a purpose
- Correcting: editing errors out from the data

Knowledge

- Knowledge is a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information.
- It originates and is applied in the minds of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices, and norms.
- Knowledge is the fact or condition of knowing something with familiarity gained through experience.
- It is also defined as the theoretical or practical understanding of a subject, gained through experience or education.

The Data Science Process

Key steps include

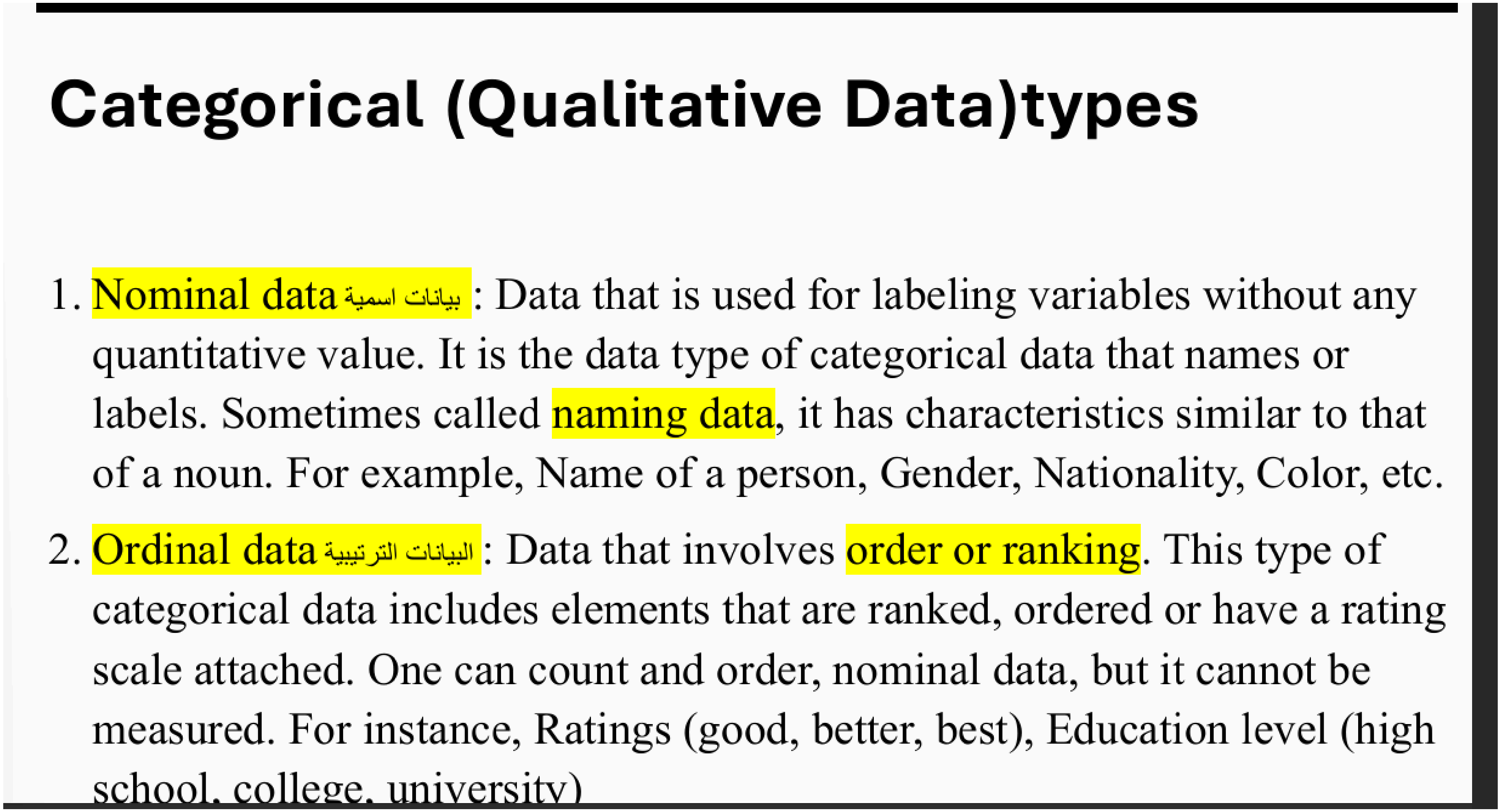
- Defining the problem
- Collecting data
- Cleaning data
- Exploring
- Building models
- Testing
- Putting solutions to work

The Pillars of Data Science

- Domain Knowledge
- Math and Statistics Expertise
- Computer Science
- Communication and Visualization Skills

categorical or qualitative data

- numerical or quantitative data



Categorical (Qualitative Data)types

1. Nominal data بيانات اسمية : Data that is used for labeling variables without any quantitative value. It is the data type of categorical data that names or labels. Sometimes called naming data, it has characteristics similar to that of a noun. For example, Name of a person, Gender, Nationality, Color, etc.
2. Ordinal data البيانات الترتيبية : Data that involves order or ranking. This type of categorical data includes elements that are ranked, ordered or have a rating scale attached. One can count and order, nominal data, but it cannot be measured. For instance, Ratings (good, better, best), Education level (high school, college, university)

Numerical(Quantitative)Data Types

- There are two types of numerical data, namely, discrete data and continuous data.

1. Discrete Data: Discrete data can be defined as the data that can only take certain values and cannot be made more precise. It is a type of numerical data with countable elements (i.e., they have a one-to-one mapping with natural numbers). Discrete data can either be countably finite or countably infinite. Some general examples of discrete data are Number of cars in a parking lot, number of students in a class, number of candidates in an election, etc.

Numerical(Quantitative)Data Types

- Continuous Data: Continuous Data can be defined as the data that can take any value within a given range. It is a numerical data type with uncountable elements. They are represented as a set of intervals on a real number line. Some examples of continuous data are Height, Weight, Temperature, etc.
- Similar to discrete data, continuous data can also be either finite or infinite. An uncountable finite data set has an end, while an uncountable infinite data set tends to be infinite.

Continuous Data Types

1. Interval Data: This is when numbers have units that are of equal magnitude as well as rank order on a scale without absolute zero. Scales of this type can have an arbitrarily assigned “zero”, but it will not correspond to an absence of the measured variable. For example, the temperature in Fahrenheit scale
2. Ratio Data: When numbers have units that are of equal magnitude as well as rank order on a scale with an absolute zero. An example is blood pressure.

Data Sources and Collection Methods

- Data collection is the process of collecting data aiming to gain insights.

Primary Data البيانات الأولية

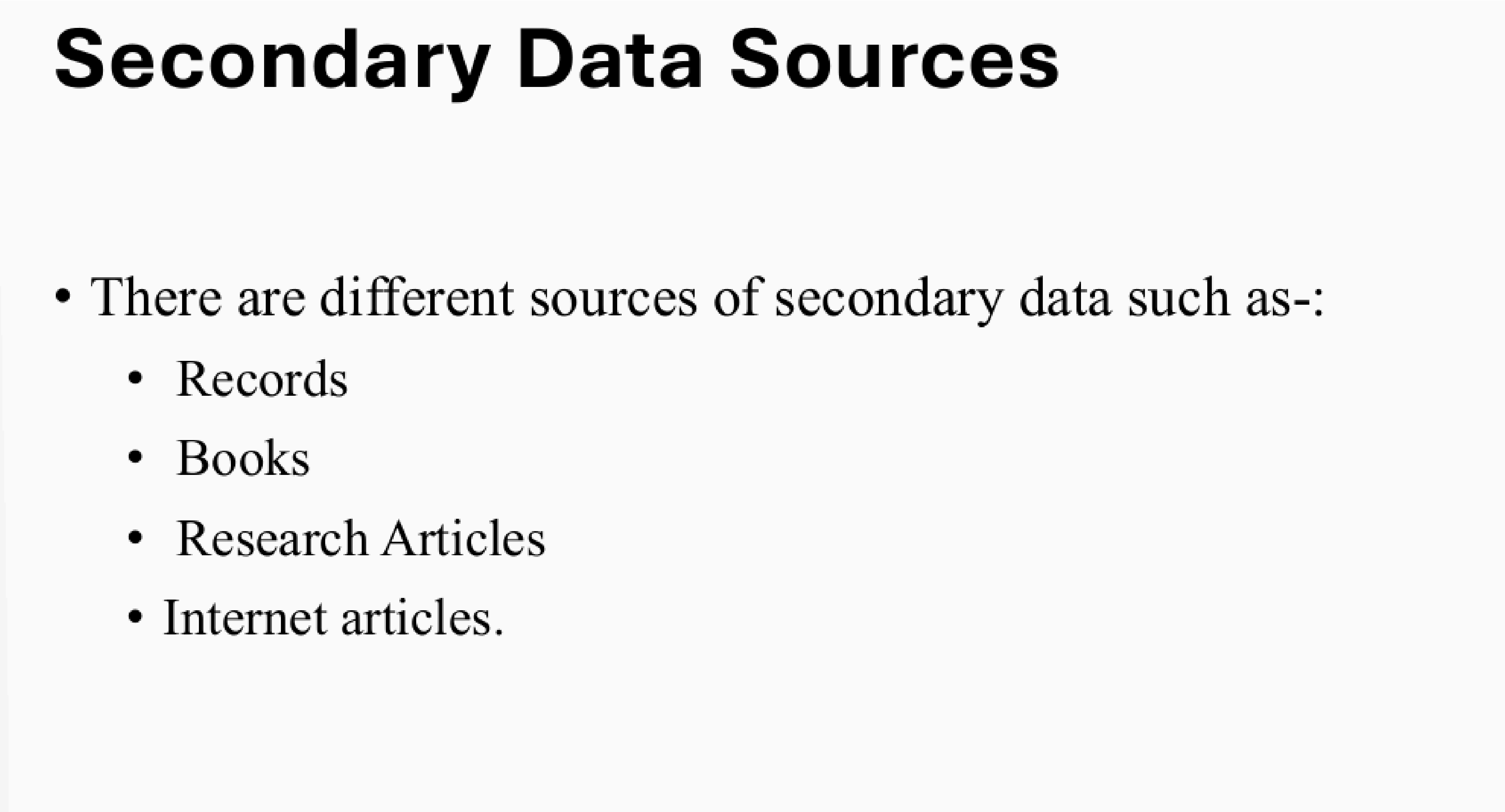
- Data that is not published yet and is the first-hand information معلومات مباشرةwhich is not changed by any individual is known as primary data.
- In other words, researchers use different approaches to gather and collect primary data for a specific purpose.
- Thus, the validity, reliability, objectivity, and authenticity of data are more in primary data in comparison with the secondary data types.
- These qualities are important in some types of research methods such as statistical surveys as the use of the information is specific to a problem and cannot be provided from published references.

Primary Data Sources

- To achieve primary data, different sources can be used such as:
  - Experiments
  - Surveys
  - Interviews
  - Questionnaires.

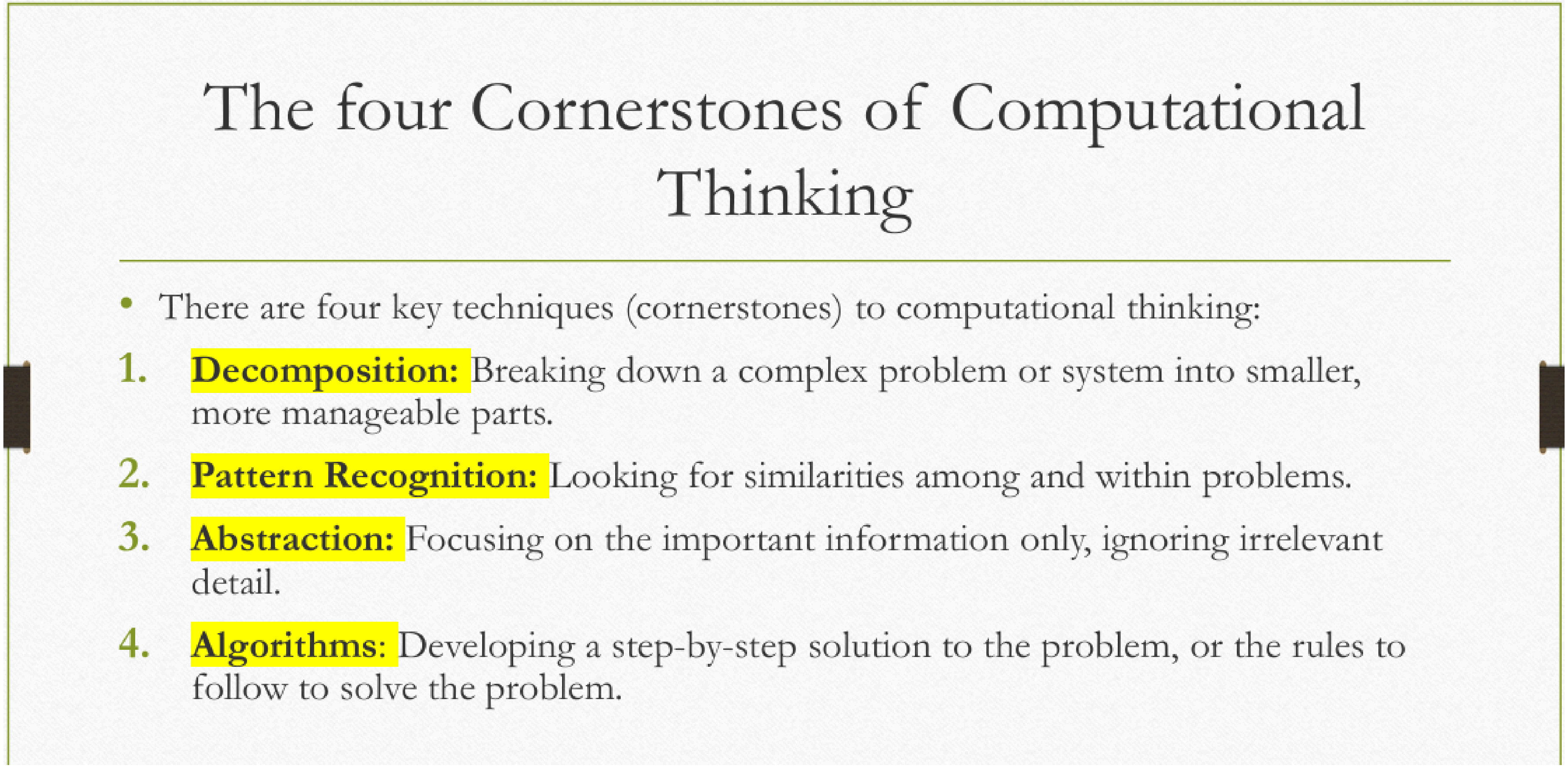
Secondary Data

- Secondary data is the data gathered from published sources meaning that the data is already gathered by someone else for another reason and can be used for other purposes in research as well.
- In all papers, the literature review section is based on secondary data sources.



Secondary Data Sources

- There are different sources of secondary data such as-:
  - Records
  - Books
  - Research Articles
  - Internet articles.



The four Cornerstones of Computational Thinking

- There are four key techniques (cornerstones) to computational thinking:

1. **Decomposition:** Breaking down a complex problem or system into smaller, more manageable parts.
2. **Pattern Recognition:** Looking for similarities among and within problems.
3. **Abstraction:** Focusing on the important information only, ignoring irrelevant detail.
4. **Algorithms**: Developing a step-by-step solution to the problem, or the rules to follow to solve the problem.

Data Preprocessing

- Algorithms play a vital role in the data preprocessing stage, which involves transforming raw data into a format suitable for analysis. Some common data preprocessing algorithms include:-
- **Data Cleaning**: Algorithms that handle missing values, remove outliers, and correct inconsistencies in the data.
- **Feature Engineering**: Algorithms that create new features from the existing data, improving the predictive power of machine learning models.
- **Data Transformation**: Algorithms that convert data between different formats or scales, such as normalization and standardization.

Model Evaluation

- Algorithms are also essential for evaluating the performance of machine learning models, ensuring their reliability and effectiveness. Some common evaluation algorithms include:
- **Cross-Validation:** An algorithm that assesses a model's performance by partitioning the data into training and validation sets.
- **Confusion Matrix:** An algorithm that provides a detailed breakdown of a model's performance, including accuracy, precision, recall, and F1-score.

Optimization

- Algorithms can also be used to optimize various aspects of the data science workflow, such as hyperparameter tuning and model selection. Some optimization algorithms include:
- **Gradient Descent:** An algorithm that iteratively adjusts the parameters of a model to minimize a cost function.
- **Genetic Algorithms:** Algorithms inspired by the process of natural selection, used for optimization and problem-solving.