BTW- Data Science Task 1

Bisma Fajar

June 2024

1 Data

Data is facts, information, observations and measurements that are used to make discoveries and to support informed decisions. A data point is a single unit of data with in a dataset, which is collection of data points.

1.1 How Data is Described

1.1.1 Raw Data

Raw data is data that has come from its source in its initial state and has not been analyzed or organized.

1.1.2 Quantitative Data

Quantitative data is numerical observations within a dataset and can typically be analyzed, measured and used mathematically. Some examples of quantitative data are: a country's population, a person's height or a company's quarterly earnings.

1.1.3 Qualitative Data

Qualitative data, also known as categorical data is data that cannot be measured objectively like observations of quantitative data. It's generally various formats of subjective data that captures the quality of something, such as a product or process. Sometimes, qualitative data is numerical and wouldn't be typically used mathematically, like phone numbers or timestamps. Some examples of qualitative data are: video comments, the make and model of a car or your closest friends' favorite color.

1.1.4 Structured Data

Structured data is data that is organized into rows and columns, where each row will have the same set of columns. Columns represent a value of a particular type and will be identified with a name describing what the value represents, while rows contain the actual values.

1.1.5 Unstructured data

typically cannot be categorized into rows or columns and doesn't contain a format or set of rules to follow. Because unstructured data has less restrictions on its structure it's easier to add new information in comparison to a structured dataset.

1.1.6 Semi-structured data

Semi-structured data has features that make it a combination of structured and unstructured data. It doesn't typically conform to a format of rows and columns but is organized in a way that is considered structured and may follow a fixed format or set of rules. The structure will vary between sources, such as a well defined hierarchy to something more flexible that allows for easy integration of new information. Metadata are indicators that help decide how the data is organized and stored and will have various names, based on the type of data. Some common names for metadata are tags, elements, entities and attributes. For example, a typical email message will have a subject, body and a set of recipients and can be organized by whom or when it was sent.

Examples of semi-structured data: HTML, CSV files, JavaScript Object Notation (JSON)

1.2 Sources of data

A data source is the origin of data, which can vary based on its collection method and timing. Primary data is generated directly by its users, while secondary data is collected by one party and shared for general use. For instance, scientists collecting data in a rainforest are generating primary data, but it becomes secondary data if shared with other researchers.

Databases are common data sources that use database management systems (DBMS) to store and manage data, accessed via queries. Files can also serve as data sources, including audio, image, video files, and spreadsheets like Excel. The Internet hosts various data sources, including databases and files. APIs (Application Programming Interfaces) allow programmers to share data over the internet, and web scraping extracts data directly from web pages. The "Working with Data" lessons cover how to utilize these various data sources.

2 What is Data Science?

Data Science is defined as a scientific field that uses scientific methods to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

2.1 Data Science a paradigm of science

Empirical: in which we rely mostly on observations and results of experiments
Theoretical: where new concepts emerge from existing scientific knowledge
Computational: where we discover new principles based on some computational experiments
Data-Driven: based on discovering relationships and patterns in the data

2.2 Other Related Fields

Since data is pervasive, data science itself is also a broad field, touching many other disciplines.

2.2.1 Databases

A critical consideration is **how to store** the data, i.e., how to structure it in a way that allows faster processing. There are different types of databases that store structured and unstructured data, which we will consider in our course.

2.2.2 Big Data

Often we need to store and process very large quantities of data with a relatively simple structure. There are special approaches and tools to store that data in a distributed manner on a computer cluster and process it efficiently.

2.2.3 Machine Learning

One way to understand data is to **build a model** that will be able to predict a desired outcome. Developing models from data is called **machine learning**. You may want to have a look at our Machine Learning for Beginners Curriculum to learn more about it.

2.2.4 Artificial Intelligence

An area of machine learning known as artificial intelligence (AI) also relies on data and involves building high complexity models that mimic human thought processes. AI methods often allow us to turn unstructured data (e.g., natural language) into structured insights.

2.2.5 Visualization

Vast amounts of data are incomprehensible for a human being, but once we create useful visualizations using that data, we can make more sense of it and draw conclusions. Thus, it is important to know many ways to visualize information. Related fields also include **Infographics** and **Human-Computer Interaction** in general.

3 Steps in the Data Journey

In Data Science, we focus on the following steps of the data journey:

3.1 1. Data Acquisition

The first step is to collect the data. While this can be straightforward, such as data from a web application, special techniques may be needed for more complex sources. For example, IoT sensors generate large volumes of data, and using buffering endpoints like IoT Hub can help manage this data before further processing.

3.2 2. Data Storage

Storing data, especially big data, presents challenges. The storage method should align with future querying needs. There are several ways to store data:

- **Relational Databases**: Use SQL to query collections of tables organized into schemas. Data often needs conversion to fit these schemas.
- **NoSQL Databases**: Such as CosmosDB, allow storage of complex data (e.g., hierarchical JSON documents) without enforcing schemas but lack rich querying capabilities and referential integrity.
- **Data Lakes**: Used for storing large collections of raw, unstructured data, often across clusters of servers. The Parquet format is commonly used for big data.

3.3 3. Data Processing

This involves converting data into a usable form for visualization or model training. Unstructured data (e.g., text, images) may require AI techniques to extract features and convert it to structured form.

3.4 4. Visualization / Human Insights

To understand data, visualization is crucial. Various techniques can reveal insights, and data scientists often explore different visualizations to identify relationships. Statistical methods can test hypotheses or prove correlations.

3.5 5. Training a Predictive Model

The ultimate goal of data science is decision-making based on data. Machine learning techniques can build predictive models to make predictions on new data sets with similar structures.

4 Digitalization

Digitalization refers to the process of converting information into a digital format. It involves the use of digital technologies to transform manual or non-digital processes into digital ones. Key aspects of digitalization include:

- Conversion : Converting analog information or processes into digital form.
- Efficiency : Improving efficiency through automation and streamlined digital workflows.
- Access : Enhancing access to information and services through digital platforms.
- Data Utilization : Leveraging digital data for analysis, decision-making, and innovation.
- Examples : Examples include digitizing paper records, implementing digital payment systems, and automating manufacturing processes.

5 Digital Transformation

Digital transformation goes beyond digitalization by fundamentally changing how organizations operate and deliver value to customers. It involves integrating digital technologies into all areas of a business, fundamentally altering business processes, customer experiences, and business models. Key aspects of digital transformation include:

- Strategy : Developing a digital strategy aligned with business goals and customer needs.
- Innovation : Fostering innovation through digital technologies such as AI, IoT, and big data analytics.
- Customer Experience : Enhancing customer experiences through digital channels and personalized interactions.
- Agility : Improving agility and responsiveness to market changes and customer demands.
- Cultural Change : Encouraging a digital-first mindset and culture throughout the organization.

6 Digital Ethics: Concepts and Applications

Digital ethics encompasses principles and practices governing moral behavior in data science, AI, and technology. It addresses critical issues such as data privacy, algorithmic fairness, and ethical frameworks for responsible innovation.

6.1 Ethics Principles

Ethical AI principles include accountability, transparency, fairness, reliability, safety, privacy, security, and inclusiveness. For instance, Microsoft's Responsible AI framework emphasizes these principles to ensure ethical AI development.

6.2 Ethics Challenges

Ethical challenges in data science involve issues like data ownership, informed consent, dataset bias, algorithmic fairness, and data privacy breaches. Case studies like the Netflix data privacy incident highlight these challenges in real-world contexts.

6.3 Applied Ethics

Practical approaches include professional codes of conduct, ethics checklists (e.g., Deon checklist), and compliance with data protection regulations (e.g., GDPR, CCPA). Establishing an ethics culture within organizations fosters accountability and ethical decision-making in AI and data-driven projects.