ACADEMIC CITY
UNIVERSITY COLLEGE

INTRODUCTION TO DATA SCIENCE

Ishaya, Jeremiah Ayock

Lecture 01

January 20, 2025

Academic City University College, Agbogba Haatso, Ghana.

OUTLINE OF PRESENTATION

Introduction

DATA SCIENCE MOTIVATION

- We are living in the age of data. Not only the amount of data we currently have is **gigantic**, but the pace of data generation is also increasing day by day.
- There are more than **4.5 billion active** internet users which comprises about **60%** of the global population.
- These internet users create large volumes of data using social networks such as **Facebook, Twitter, Instagram, and YouTube.**

DATA SCIENCE MOTIVATION CON'T

- The real-time data produced by **IoT devices** is seeing an unprecedented growth rate. These mind-boggling statistics are growing at an exponential rate because of the digitization of the world.

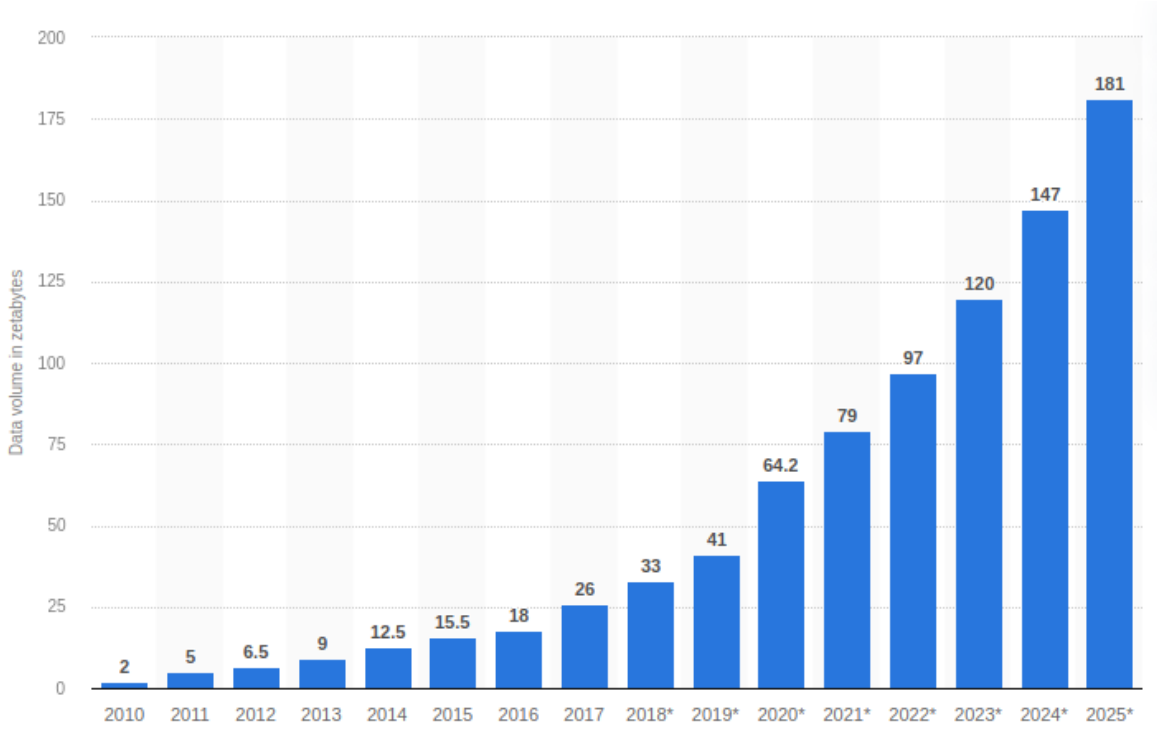

Figure 1: Image of global growth trend of data volume, 2010-2020

DATA SCIENCE MOTIVATION CONT

Figure 1 shows the global growth trend of data volume from 2010–2025. The graph above is measured in zettabytes (ZB) or 1021 bytes - 1 ZB represents 1 trillion gigabytes (GB). This is a huge amount of data. The value associated with this data diminishes because the data is not processed and analyzed at the same rate. Thus, it is of utmost importance to extract knowledge from the data.

DATA SCIENCE MOTIVATION CONT

Data science applies scientific principles, methods, algorithms, and processes to extract knowledge, information, and insights by collecting, processing, and analyzing structured and unstructured data, where the former type of data is obtained from a database management system such as MySQL, Oracle, or MongoDB, and the latter type of data comprises text, audio, video, and documents.

APPLICATIONS OF DATA SCIENCE

Data science has a whole lot of applications in a diverse set of industries.

- Financial companies use the data collected from their customers for fraud and risk detection.
- The healthcare sector receives great benefit from data science by analyzing medical images to detect tumors, and artery stenosis.
- Google use data science algorithms to provide us the best result for our searched query.
- Digital marketing strategies use data science algorithms to get insight into the preferences and needs of customers. Knowing the spending habits of people can help identify the customer base of a company.

APPLICATIONS OF DATA SCIENCE CON'T

- Internet giants like **Amazon, LinkedIn, Twitter, Google, Netflix, and IMDb** use recommendation systems to suggest relevant products to the users based upon their previous searches, thereby improving the user experience.
- Social media websites use face recognition to suggest tags to your uploaded images.
- Speech recognition products such as **Google Voice, Siri, Cortana,** etc. convert speech to text and commands.
- Airline companies employ data science to identify the strategic areas of improvements such as flight delay prediction and route planning

Tools For Data Science

PYTHON FOR DATA SCIENCE

Python is an open-source, versatile, and flexible programming language with a simple syntax that makes data science tasks quite easy.

Python offers numerous useful libraries that do all the tedious tasks for you in the background. Python offers libraries for data processing, analysis, modeling, and visualization that include:

- Numpy
- Pandas
- Scikit-Learn
- Matplotlib

DATA SCIENCE PIPELINE

The overall step by step process to collect, store, clean, preprocess, analyze, model, interpret, and visualize the data is known as a data science pipeline. The processes in the pipeline are followed in a particular order to make things work. The main steps of this pipeline are as follows:

- Data Acquisition,
- Data Preparation,
- Exploratory Data Analysis,
- Data Modeling and Evaluation, and
- Interpretation and Reporting of Findings.

DATA PIPELINE CONT

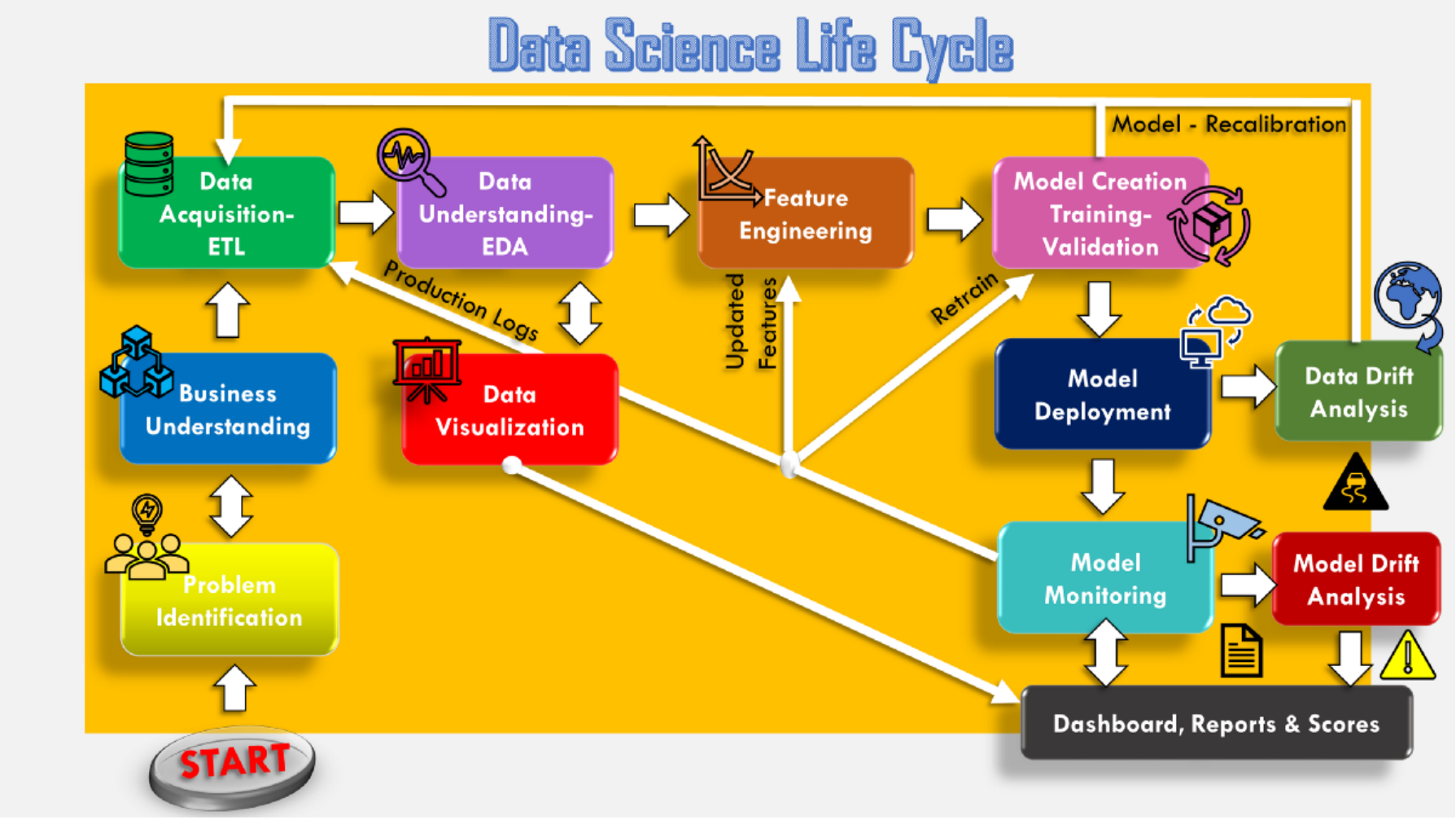

Figure 2: Image of the data science pipeline

DATA PIPELINE CONT

Understanding and following this pipeline enables us to:

- Recognize patterns in the given data;
- Extract useful information from the data;
- Determine suitable models to describe the data;
- Decide the best algorithms to be applied to the data;
- Make appropriate decisions at different levels of a data science project.

DATA ACQUISITION:

- We cannot perform any data science task without having data. The first step is to obtain data from either a database or from the internet. This data should be available in a usable format, for example,(CSV) and (TSV).
- The data can be structured as obtained from a database management system: MySQL, Oracle, and MongoDB. Alternatively, it can be unstructured, for example, text, audio, video, and documents.

DATA PREPARATION/CLEANING/SCRUBBING

The data acquired from different sources is in a raw form, which is usually not used directly. The data has to be cleaned and prepared for further stages of the data science pipeline. The results from data science and machine learning projects greatly depend upon what input we give them, essentially garbage in, garbage out. Therefore, cleaning or scrubbing of the acquired data has to be performed to amend or remove incorrect, incomplete, improperly formatted, or duplicated data.

The clean data is sometimes transformed and mapped into a format more suitable for further processing than the original data. This process is called data wrangling or data munging.

EXPLORATORY DATA ANALYSIS

In this phase, we apply different statistical tools to realize the range of values, important data variables and features, and data trends. We also extracted significant features from the data by analyzing the cleaned data.

DATA MODELING AND EVALUATION USING MACHINE LEARNING

A model or a machine learning model is a set of assumptions about the underlying data. For example, to increase its sales, a company spends money on advertising its products. The company keeps a record of the dollars spent on advertisements at each of its stores and sales in dollars from the same store. It discovers that the relationship between the aforementioned variables is almost linear. Therefore, a model for this situation can be a linear relationship or a line between advertisement and sales. A good model, which makes accurate assumptions about the data, is necessary for the machine learning algorithm to give good results.

DATA MODELING AND EVALUATION USING MACHINE LEARNING: CON'T

Machine learning algorithms typically build a mathematical model based on the given data, also known as the training data. Once a model is generated, it is used to make predictions or decisions on future test data. Often, when the model does not explain the underlying data, we revisit/update our model. Thus, it is a continuous process of making a model, assessing its performance, and updating the model if necessary, until a model of reasonable performance is obtained.

DATA MODELING AND EVALUATION USING MACHINE LEARNING: CON'T

A real-life example of predicting the future sale of products using machine learning models is by Walmart. The company records every purchase by the customer for future analysis. Data analysis by Walmart noticed a rise in the sales of toaster pastries, namely Pop-Tarts, whenever the National Weather Service (NWS) warned of a hurricane. Thus, store managers were instructed to put Pop-Tarts near store entrances during hurricane season. This move by the company saw a surge in the sale of Pop-Tarts. This story highlights the importance of machine learning models and their predictive powers.

INTERPRETATION AND REPORTING OF FINDINGS:

The next step in the data science pipeline is to interpret and explain our findings to others through communication. This can be achieved by connecting with and persuading people through interactive visualization and reporting to summarize the findings of a data science project.

QUESTION 1

What is the global growth trend of data volume in recent years?

- A.Linearly increasing
- B.Linearly decreasing
- C.Exponentially increasing
- D.Exponentially decreasing

QUESTION 2

Data science applies scientific principles, methods, algorithms, and processes to extract knowledge, information, and insights by:

- A.collecting data
- B.processing data
- C.analyzing data
- D.collecting, processing and analyzing data

QUESTION 3

Numpy, Pandas, Scikit-Learn and Matplotlib are popular

- A.Programming languages
- B.Inventors of Python
- C.IT companies
- D.Python libraries

QUESTION 4

In the list of following operations:

1.Exploratory Data Analysis

2.Data Preparation

3.Data Acquisition

4.Interpretation and Reporting of Findings

5.Data Modeling and Evaluation

What is the correct order of operations in a DS pipeline?

- A.12345
- B.54321
- C.32154
- D.13245

NEXT TASK

- Review of python for Data Science
  - Working with Numbers and Logic
  - Working with Strings
  - Dealing with Conditional Statements and iterations
  - Data Storage (Lists, Tuples, Sets, and Dictionaries)
- Data Acquisition

END OF PRESENTATION

THANK YOU