Data Engineering

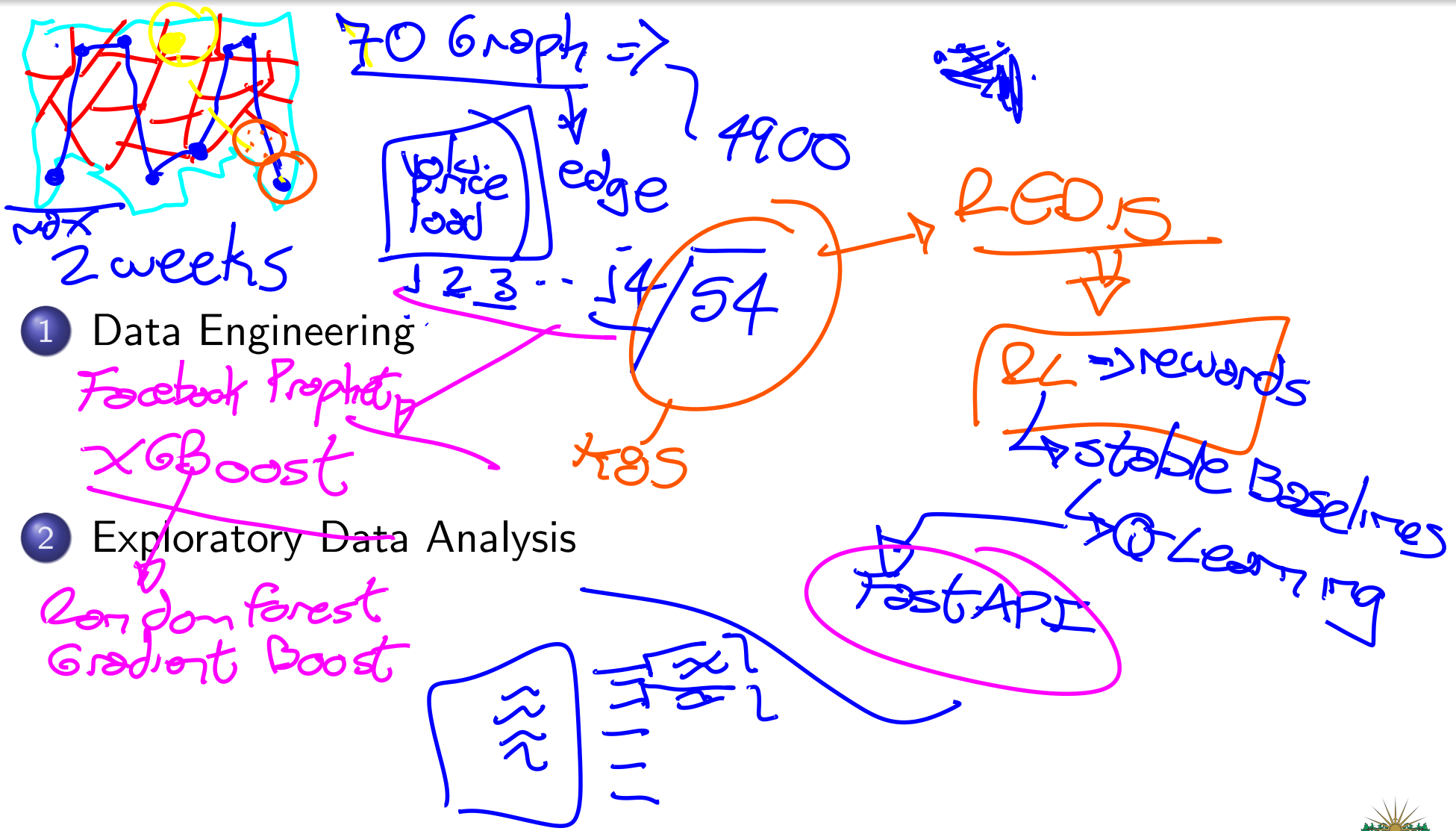
DataBase Foundations

Author: Eng. Carlos Andrés Sierra, M.Sc.
carlos.andres.sierra.v@gmail.com

Lecturer
Computer Engineer
School of Engineering
Universidad Distrital Francisco José de Caldas

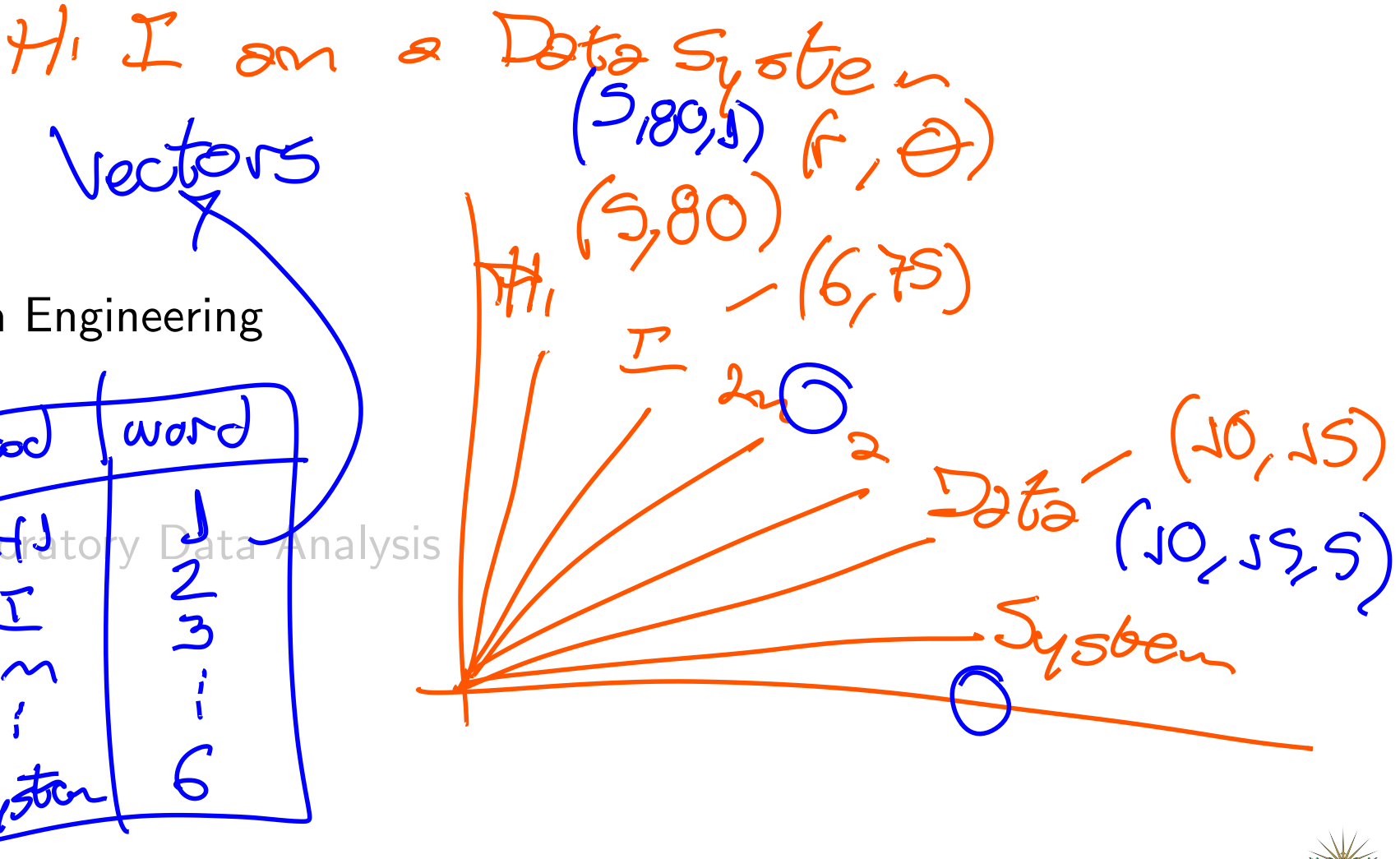
2024-I

Outline

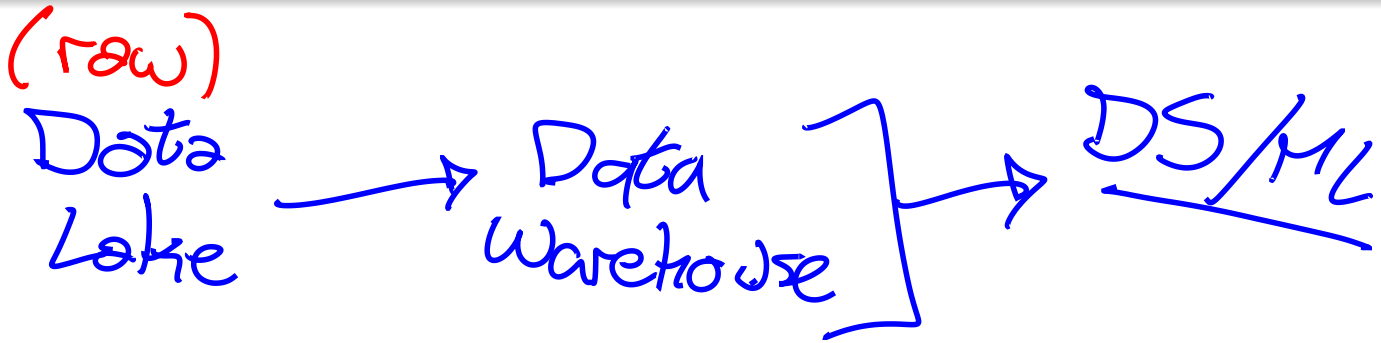


1. Data Engineering
2. Exploratory ~~Data~~ Analysis

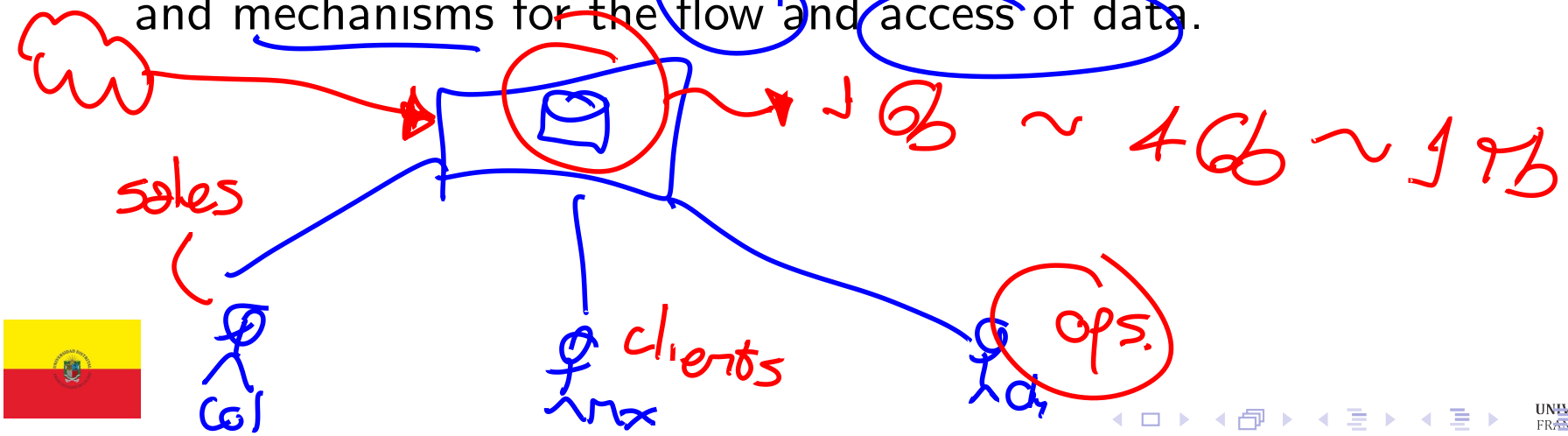
Outline



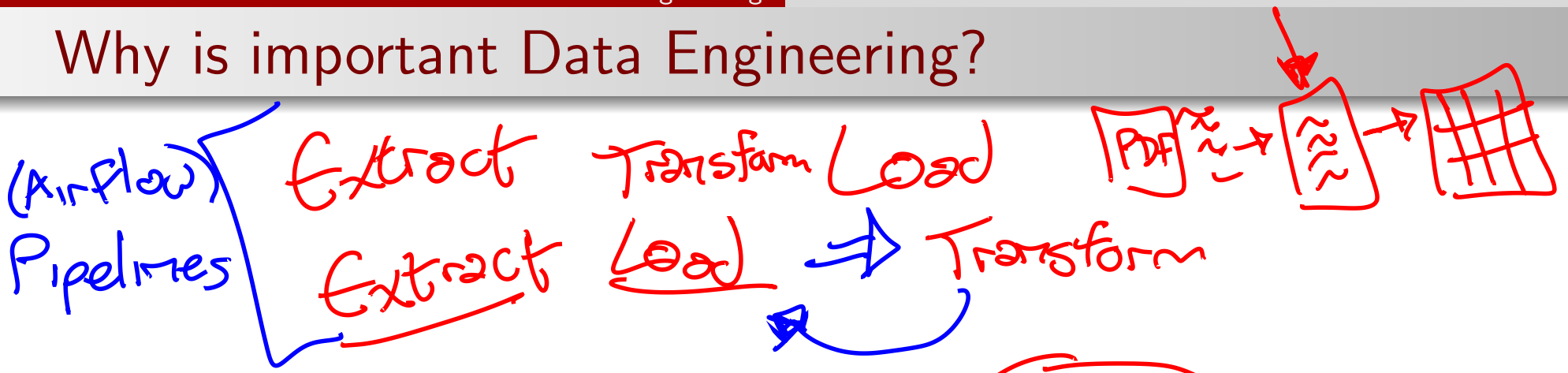
What is Data Engineering?



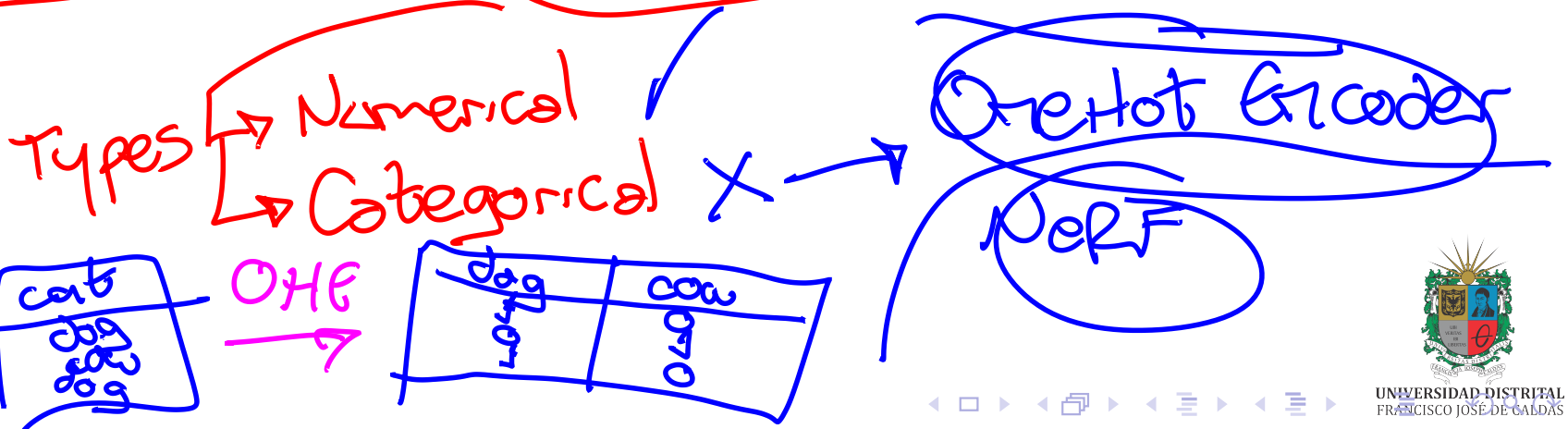
- Data Engineering is the aspect of data science that focuses on practical applications of **data collection and analysis**.
- Data Engineers are responsible for **building and maintaining the architecture** that allows data scientists to perform their work.
- Data Engineering is a set of operations aimed at creating interfaces and mechanisms for the flow and access of data.



Why is important Data Engineering?

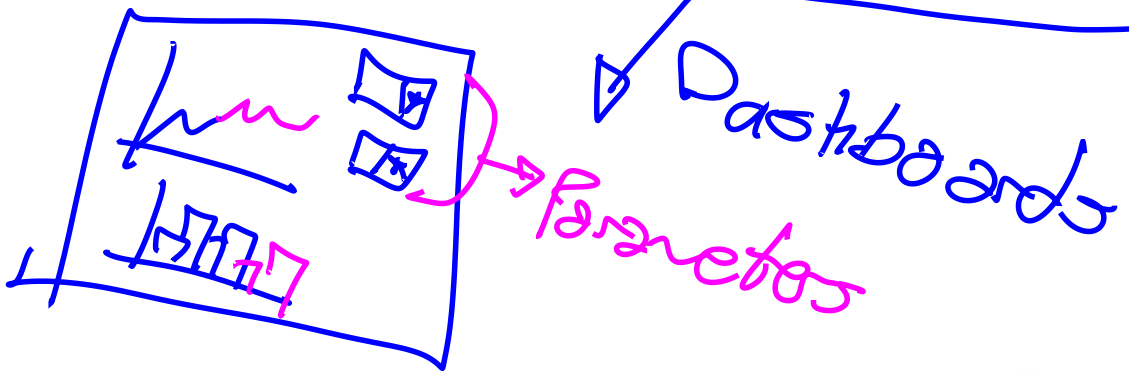


- Data Engineering is the foundation of the high-quality data that is necessary for effective data science.
- Data Engineering is the process of **collecting, transforming, and storing data** in a way that's accessible and easy to analyze.
- Data Engineering is the process of **building and maintaining the architecture** that allows data scientists to perform their work.



Data Science

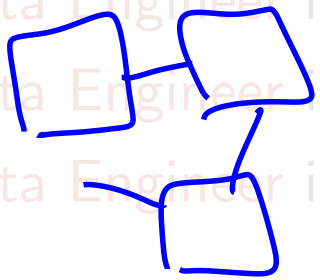
- Data Science is the process of **extracting knowledge** from data.
- Data Science is the process of **analyzing and interpreting complex digital data**.
- Data Science is the process of **creating models** that can predict future outcomes.
- Data Science is the process of **creating visualizations** to help understand data.



DBOps vs Data Engineer

DataOps

- DBOps is responsible for the **operation of the database**.
- DBOps is responsible for the **performance of the database**.
- DBOps is responsible for the **security of the database**.
- Data Engineer is responsible for the **data architecture**.
- Data Engineer is responsible for the **data quality**.
- Data Engineer is responsible for the **data flow**.



SQL → Docker
cloud

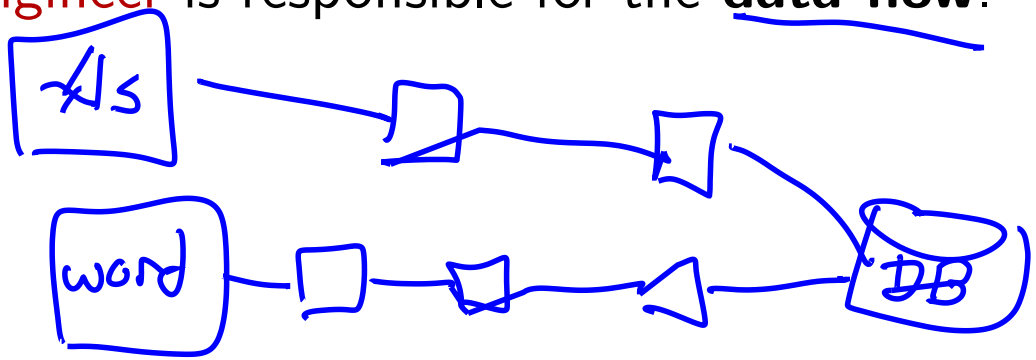
metrics

+ credentials
+ git ⇒ vulnerabilities

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

DBOps vs Data Engineer

- DBOps is responsible for the **operation of the database**.
- DBOps is responsible for the **performance of the database**.
- DBOps is responsible for the **security of the database**.
- Data Engineer is responsible for the **data architecture**.
- Data Engineer is responsible for the **data quality**.
- Data Engineer is responsible for the **data flow**.



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

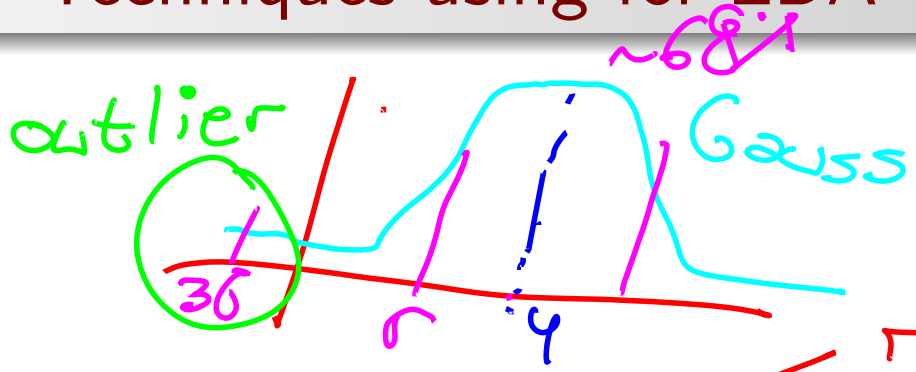
Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

What is Exploratory Data Analysis?

- Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics.
- Exploratory Data Analysis (EDA) is the process of **visualizing and analyzing data** to extract insights.
- Exploratory Data Analysis (EDA) is the process of **understanding the data** before building a model.
- Exploratory Data Analysis (EDA) is the process of **cleaning and preparing data** for analysis.
- Exploratory Data Analysis (EDA) is the process of **identifying patterns** in the data.

Techniques using for EDA



~ bias

mean – std

- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction

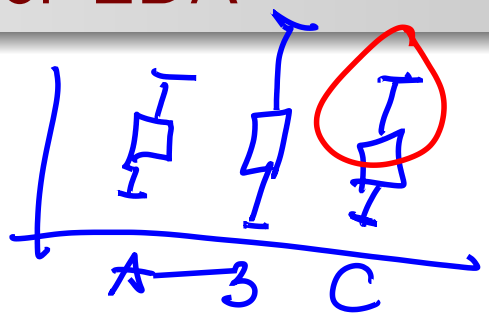
↳ Regression

→ sensibility ⇒ D – rain

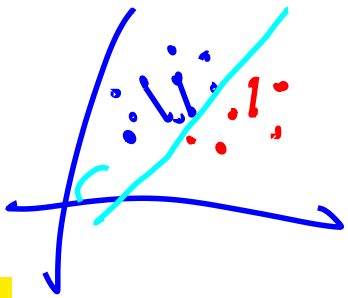
A – clouds 9
B – sun 8
C – cars 1

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Techniques using for EDA

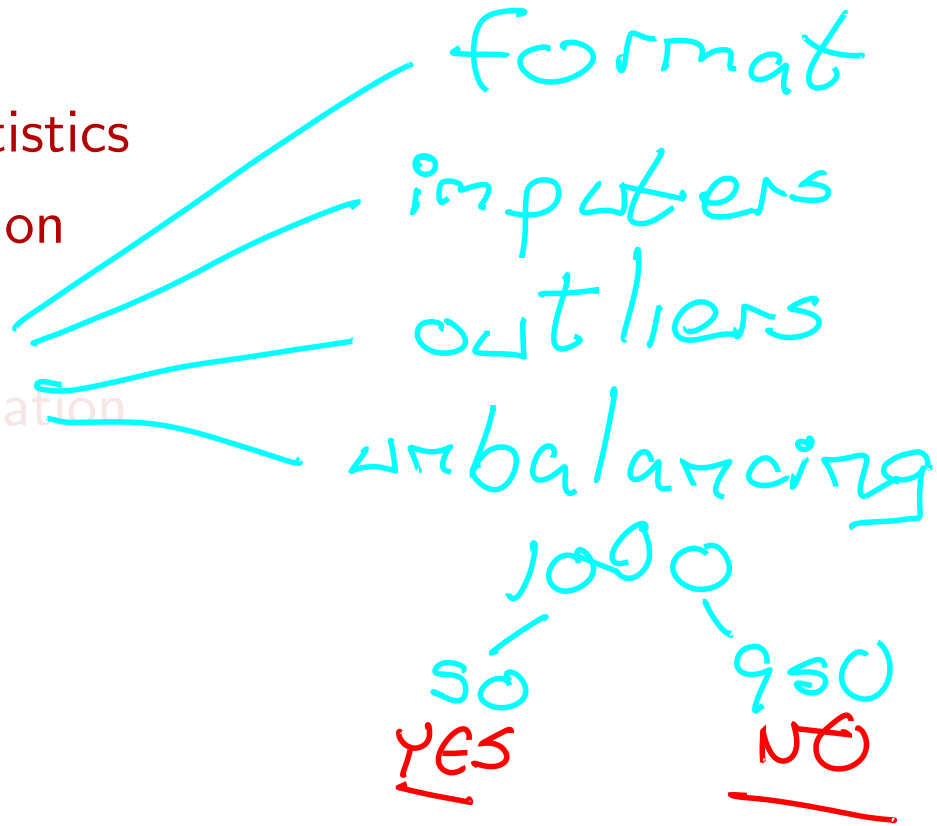


- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction



Techniques using for EDA

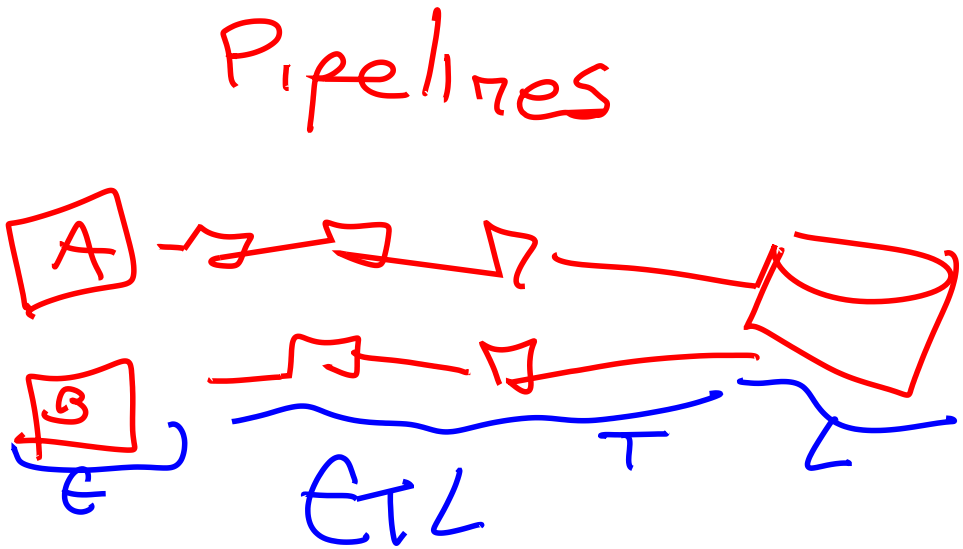
- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

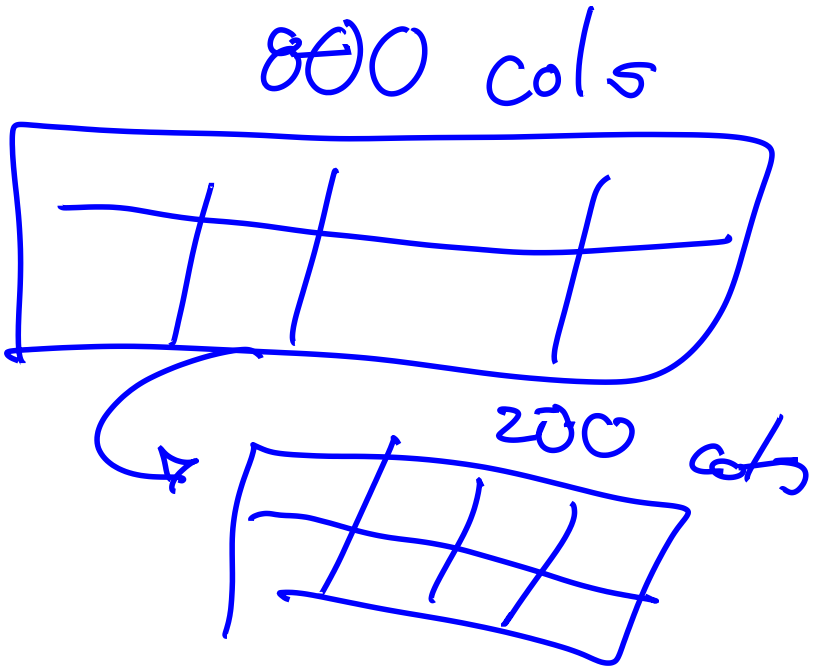
Techniques using for EDA

- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction



Techniques using for EDA

- Descriptive Statistics
- Data Visualization
- Data Cleaning
- Data Transformation
- Data Reduction



Big Data

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

How to improve data quality?

- Data Quality is the process of ensuring that data is accurate, complete, and reliable.
- Data Quality is the process of ensuring that data is consistent and up-to-date.
- Data Quality is the process of ensuring that data is free from errors and inconsistencies.
- Data Quality is the process of ensuring that data is of high quality and can be trusted.
- Data Quality is the process of ensuring that data is fit for purpose and can be used effectively.

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Thanks!

Questions?

Repo:
github.com/EngAndres/ud-public/tree/main/courses/databases-foundations

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS