Introduction to Data Science

Data Fundamentals

Author: Eng. Carlos Andrés Sierra, M.Sc.
cavirguezs@udistrital.edu.co

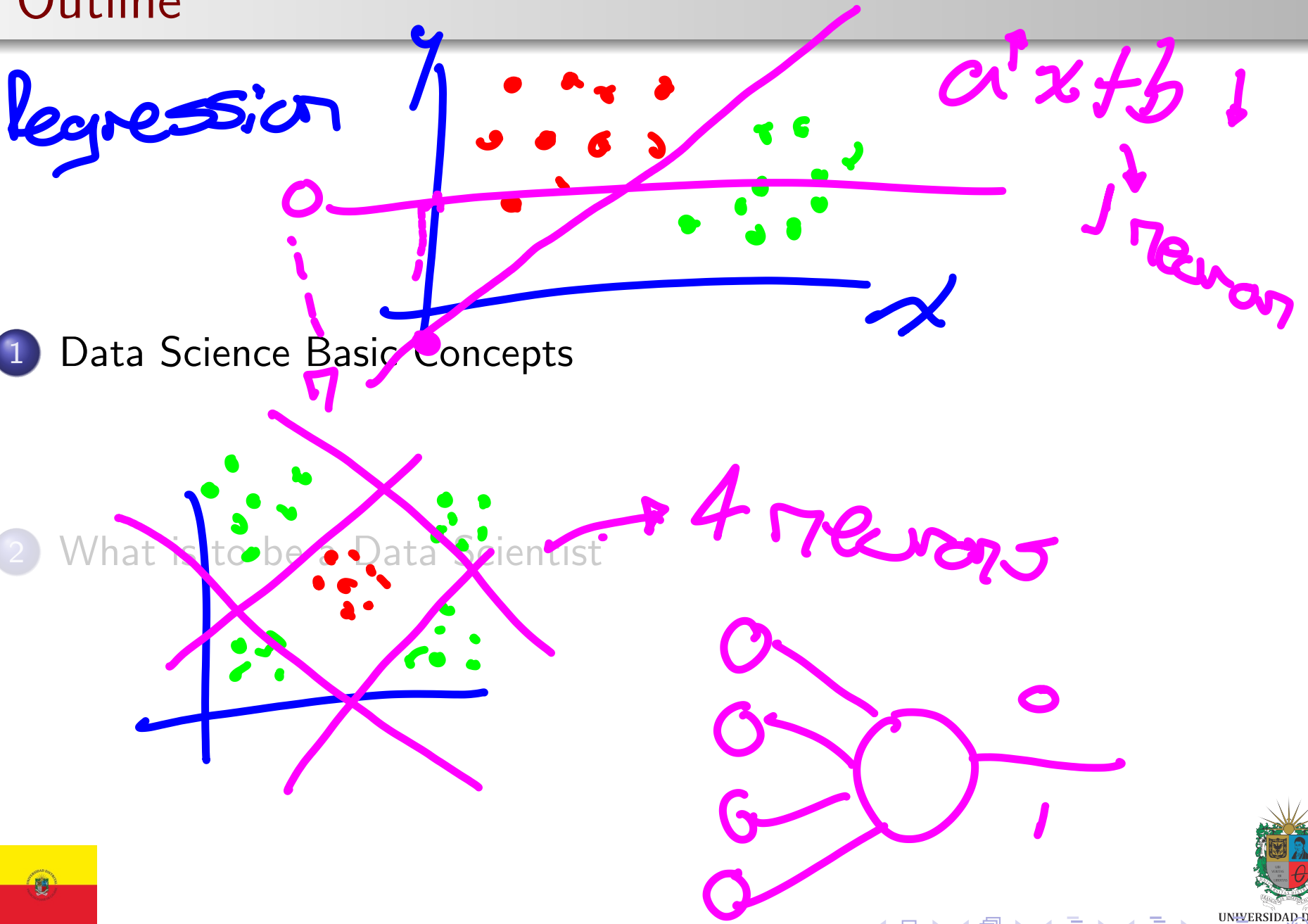
Full-time Adjunct Professor
Computer Engineering Program
School of Engineering
Universidad Distrital Francisco José de Caldas

2026-I

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

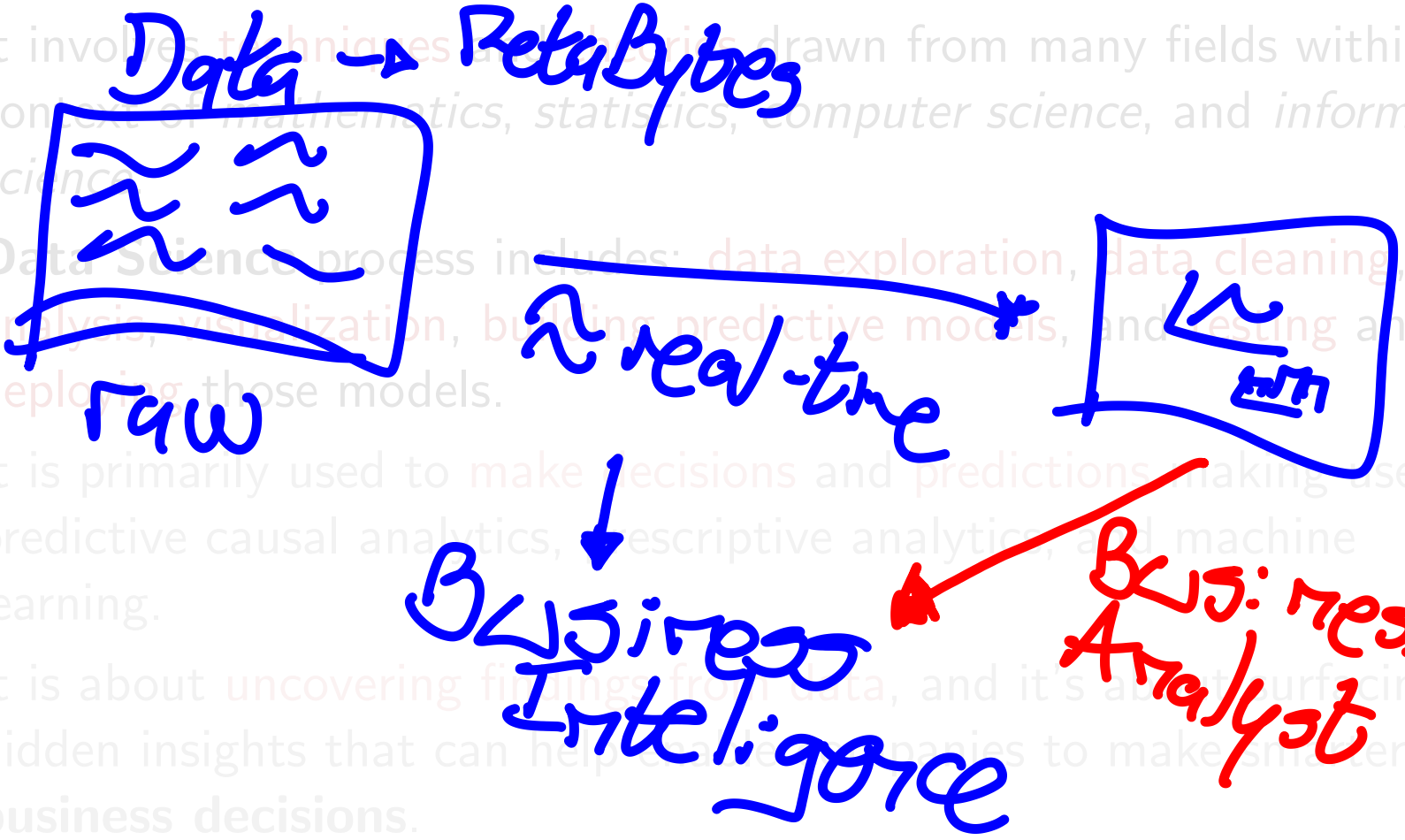
Outline

Outline



What is Data Science?

- **Data Science** is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- It involves techniques and theories drawn from many fields within the context of *mathematics*, *statistics*, *computer science*, and *information science*.
- **Data Science** process includes: data exploration, data cleaning, data analysis, visualization, building predictive models, and testing and deploying those models.
- It is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics, and machine learning.
- It is about uncovering findings from data, and it's about surfacing hidden insights that can help companies to make smarter **business decisions**.



What is Data Science?

- **Data Science** is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- It involves techniques and theories drawn from many fields within the context of *mathematics*, *statistics*, *computer science*, and *information science*.
- **Data Science** process includes: data exploration, data cleaning, data analysis visualization, building predictive models and testing and deploying those models.
- It is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics, and machine learning.
- It is about uncovering findings from data, and it's about surfacing hidden insights that can help enable companies to make smarter **business decisions**.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

What is Data Science?

- **Data Science** is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- It involves techniques and theories drawn from many fields within the context of *mathematics*, *statistics*, *computer science*, and *information science*.
- **Data Science** process includes data exploration, data cleaning, data analysis, visualization, building predictive models, and testing and deploying those models.
- It is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics, and machine learning.
- It is about uncovering findings from data, and it's about surfacing hidden insights that can help enable companies to make smarter **business decisions**.

What is Data Science?

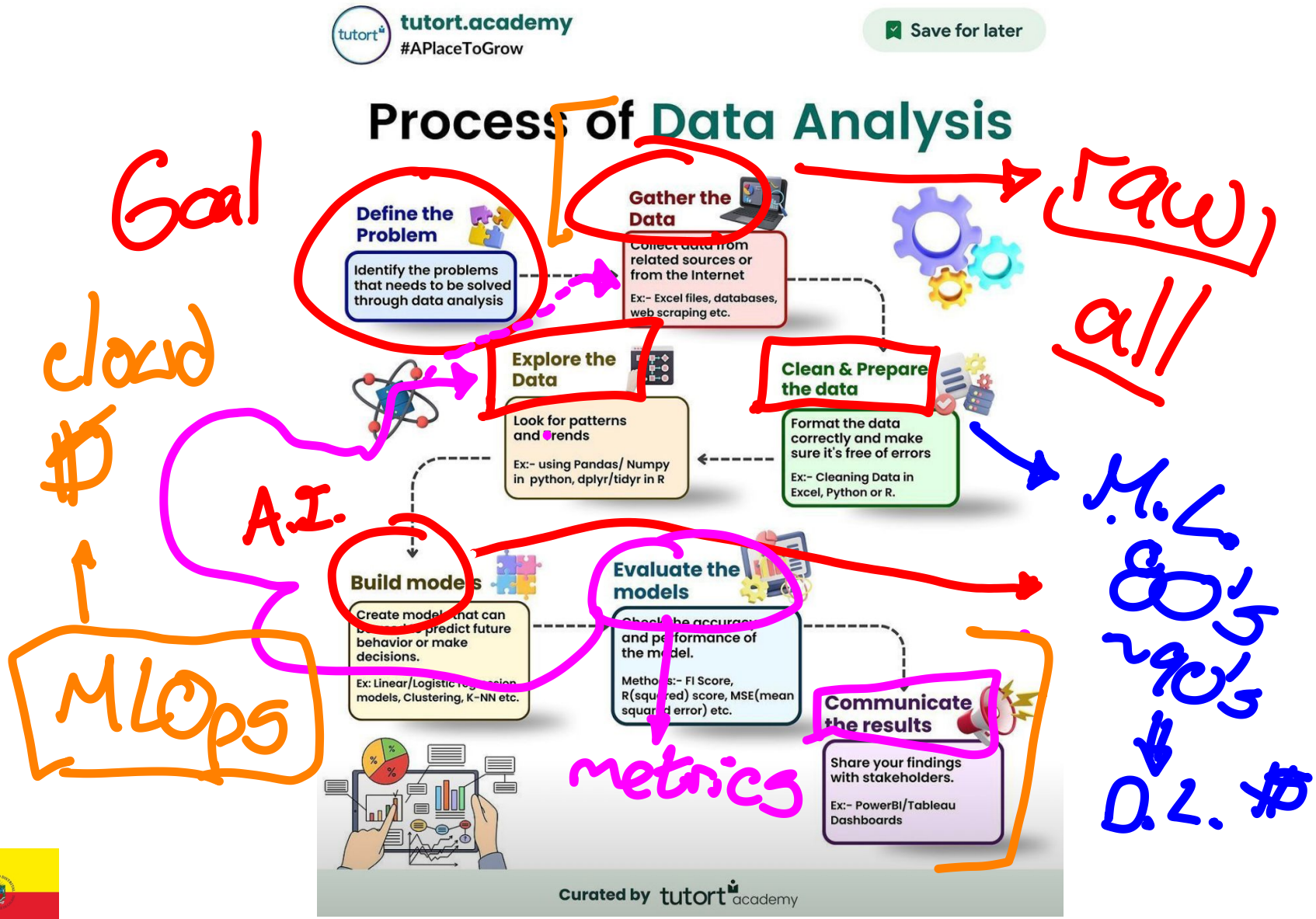
- **Data Science** is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- It involves techniques and theories drawn from many fields within the context of *mathematics*, *statistics*, *computer science*, and *information science*.
- **Data Science** process includes: data exploration, data cleaning, data analysis, visualization, building predictive models, and testing and deploying those models.
- It is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics, and machine learning.
- It is about uncovering findings from data, and it's about surfacing hidden insights that can help enable companies to make smarter **business decisions**.

What is Data Science?

- **Data Science** is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
- It involves techniques and theories drawn from many fields within the context of *mathematics*, *statistics*, *computer science*, and *information science*.
- **Data Science** process includes: data exploration, data cleaning, data analysis, visualization, building predictive models, and testing and deploying those models.
- It is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics, and machine learning.
- It is about uncovering findings from data, and it's about surfacing hidden insights that can help enable companies to make smarter **business decisions**.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

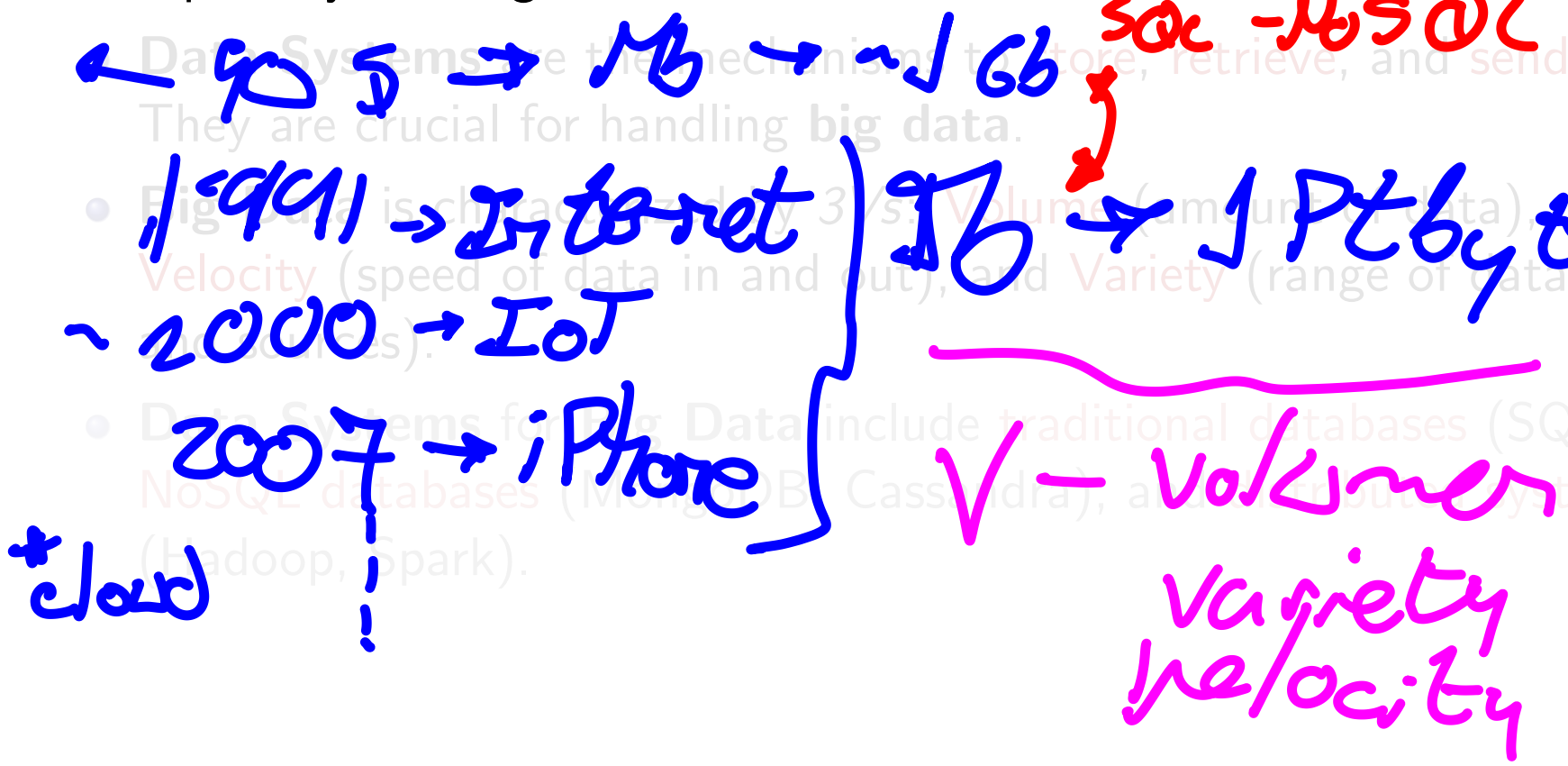
Process of Data Analysis



Types of Data Analysis

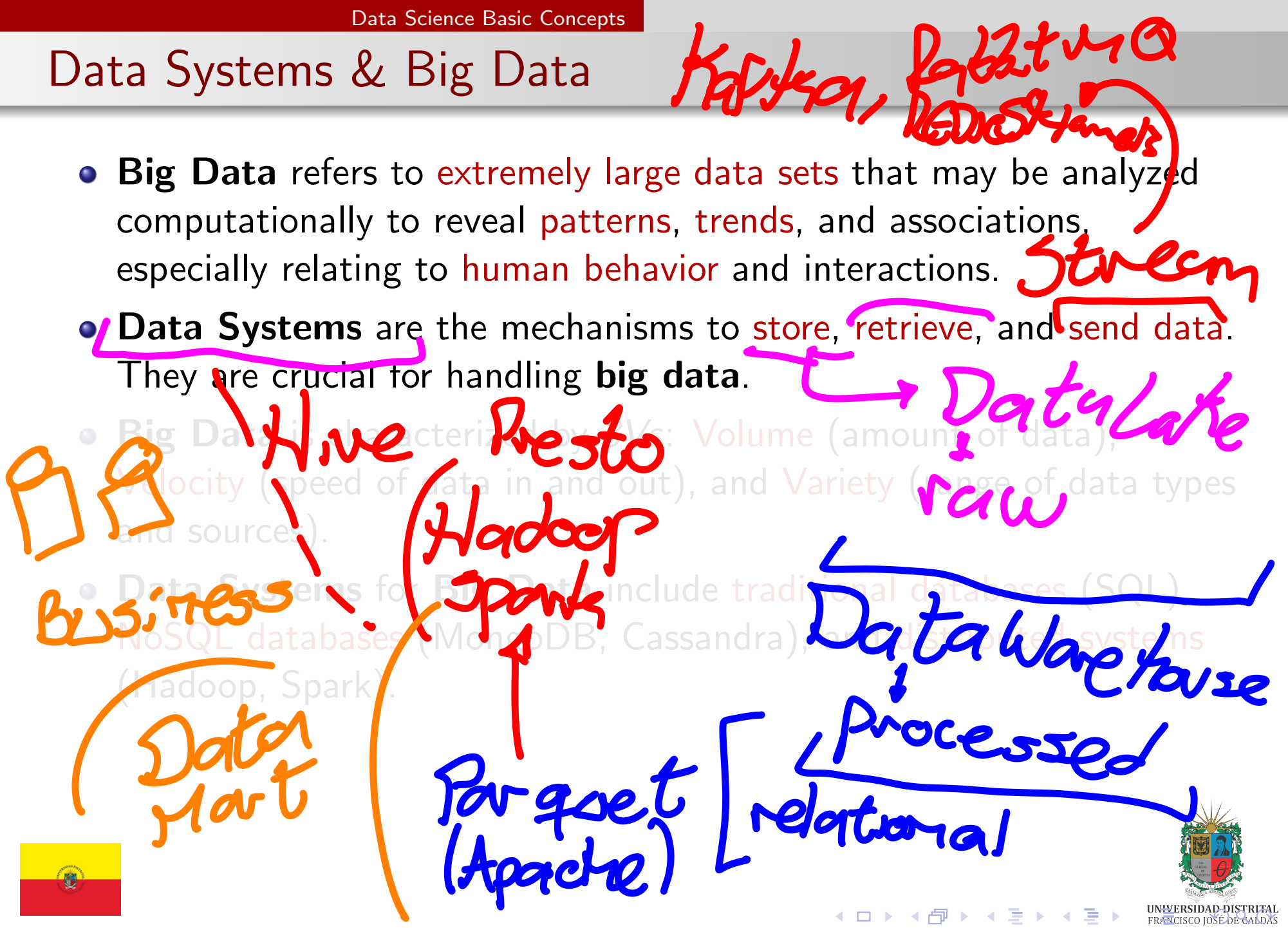
Data Systems & Big Data

- **Big Data** refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.



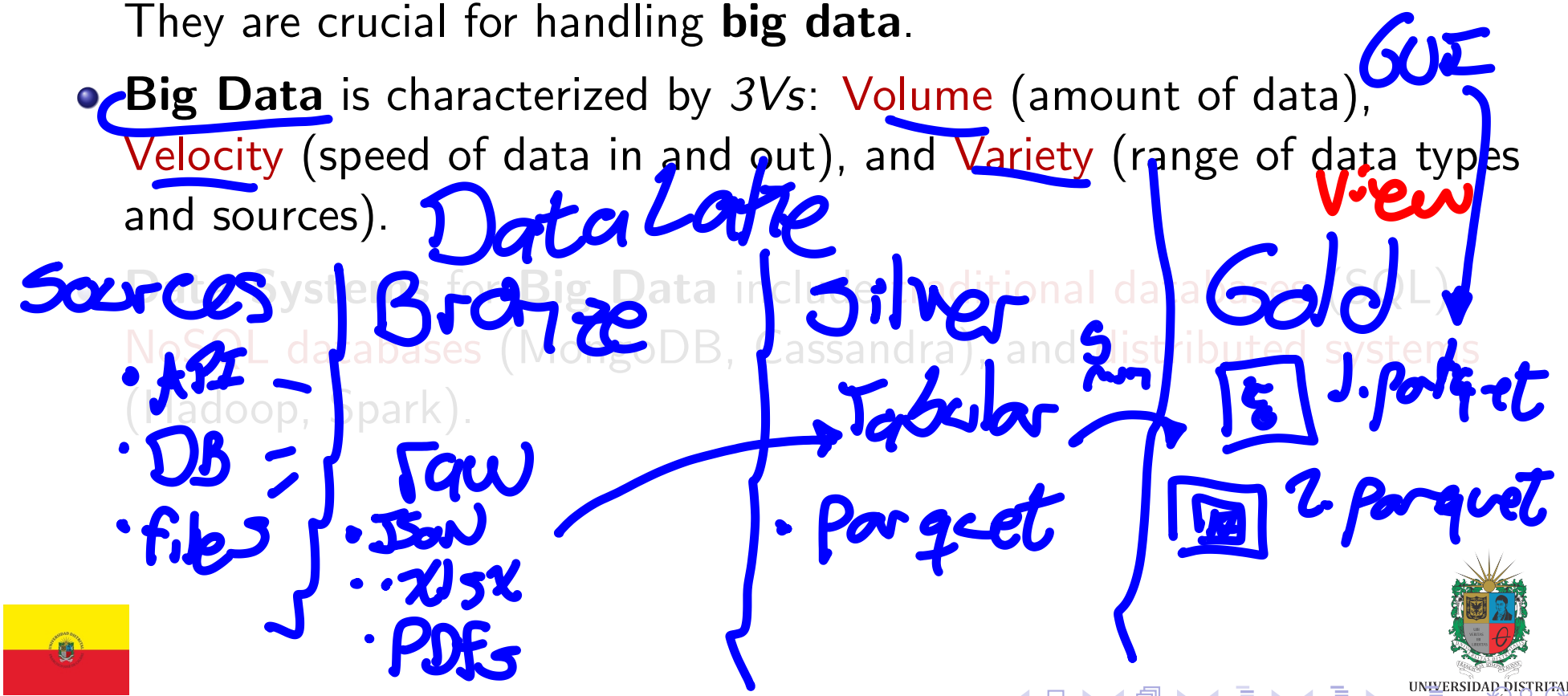
Data Systems & Big Data

- **Big Data** refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.
- **Data Systems** are the mechanisms to store, retrieve, and send data. They are crucial for handling **big data**.



Data Systems & Big Data

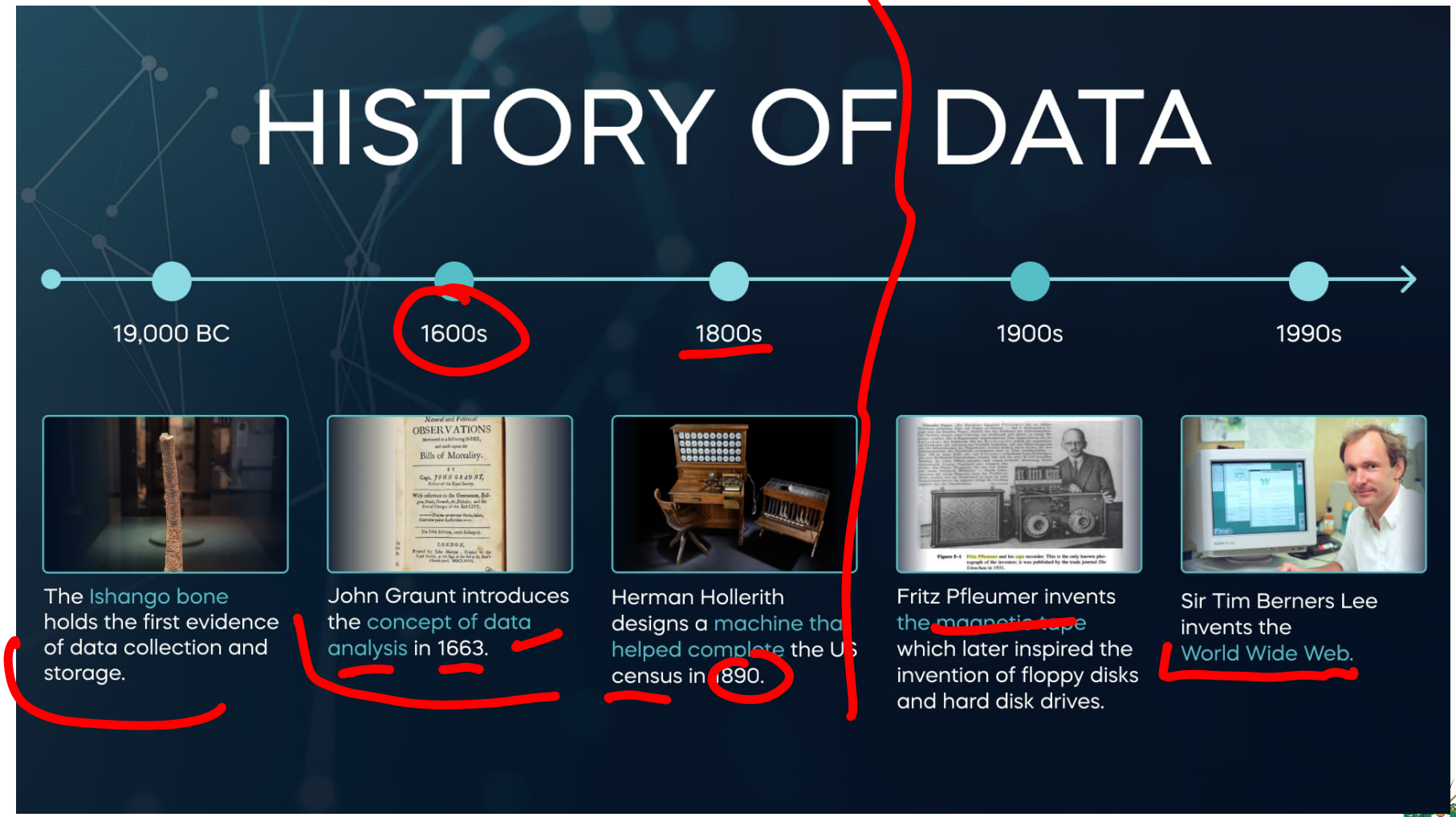
- **Big Data** refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.
- **Data Systems** are the mechanisms to store, retrieve, and send data. They are crucial for handling **big data**.
- **Big Data** is characterized by *3Vs*: Volume (amount of data), Velocity (speed of data in and out), and Variety (range of data types and sources).



Data Systems & Big Data

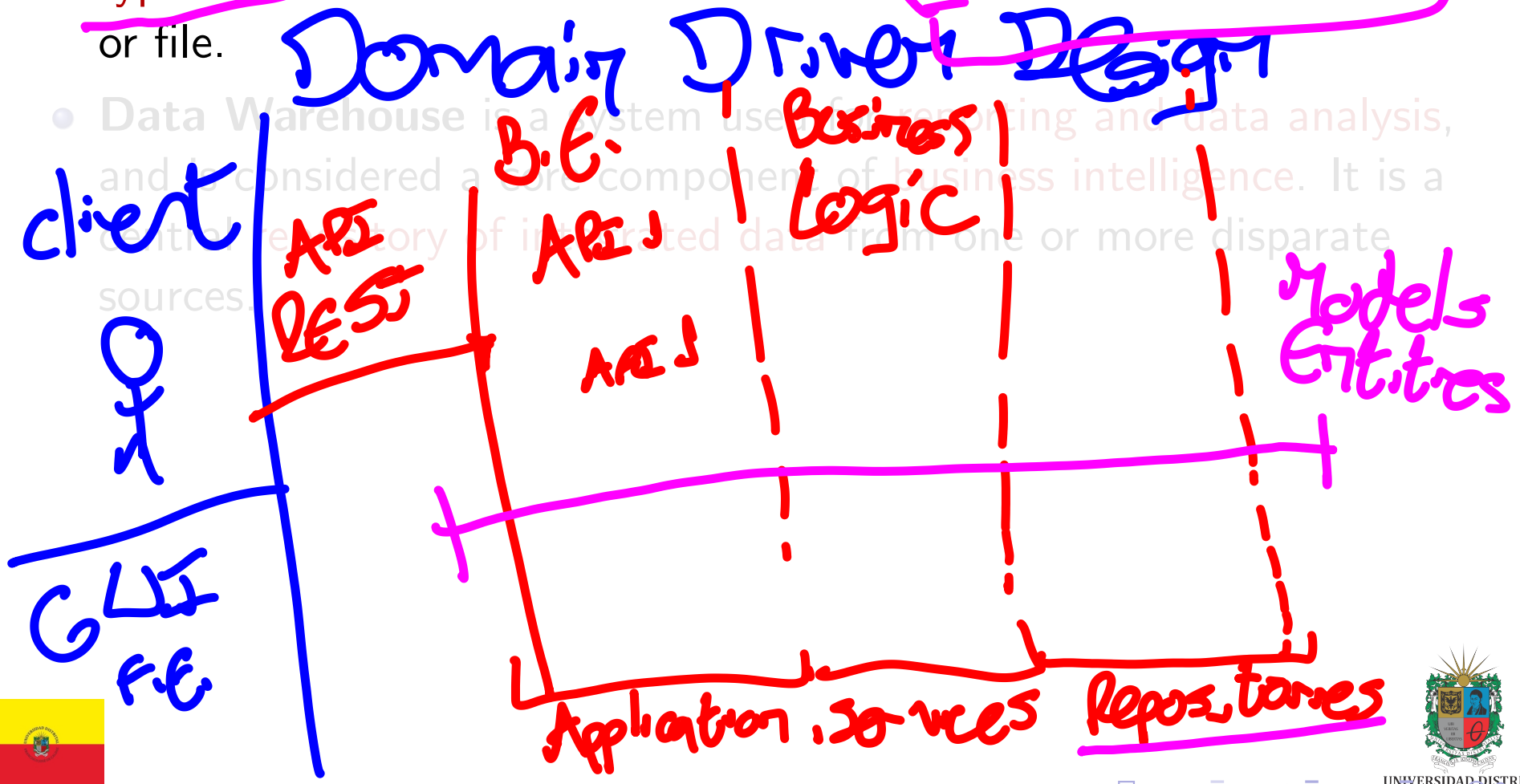
- **Big Data** refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.
- **Data Systems** are the mechanisms to store, retrieve, and send data. They are crucial for handling **big data**.
- **Big Data** is characterized by *3Vs*: Volume (amount of data), Velocity (speed of data in and out), and Variety (range of data types and sources).
- **Data Systems** for **Big Data** include traditional databases (SQL), NoSQL databases (MongoDB, Cassandra), and distributed systems (Hadoop, Spark).

History of Data



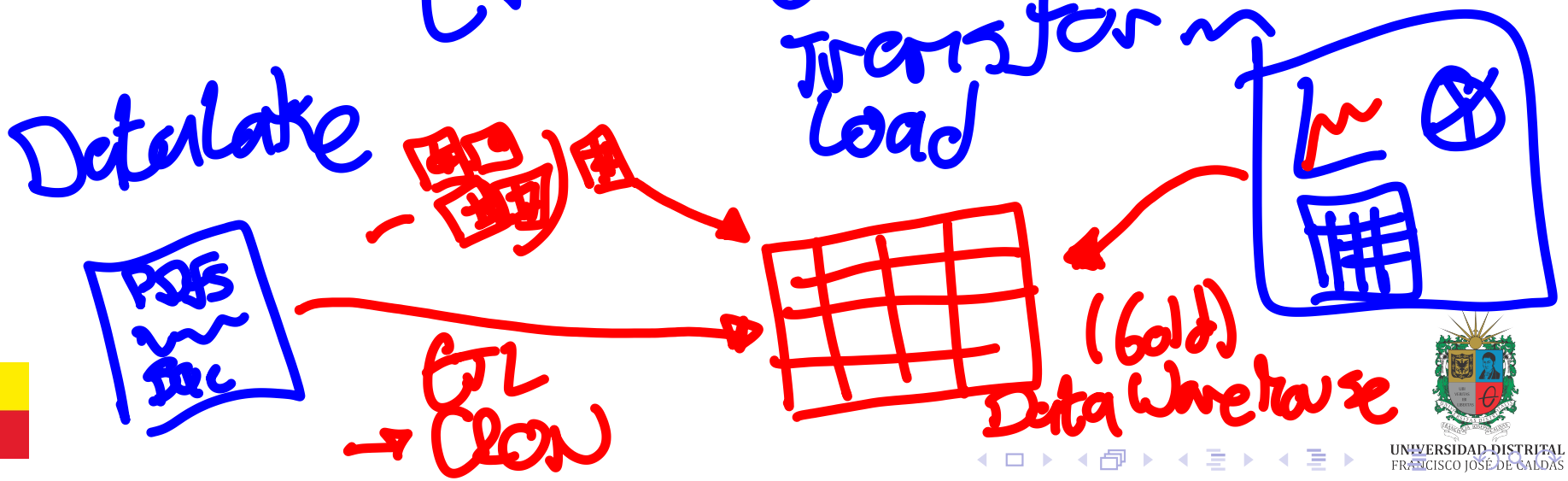
Data Lake Vs. Data WareHouse

- **Data Lake** is a storage repository that holds a vast amount of raw data in its native format until it is needed. It is a place to store every type of data in its native format with no fixed limits on account size or file.
- **Data Warehouse** is a system used for reporting and data analysis, and is considered a core component of business intelligence. It is a central repository of integrated data from one or more disparate sources.

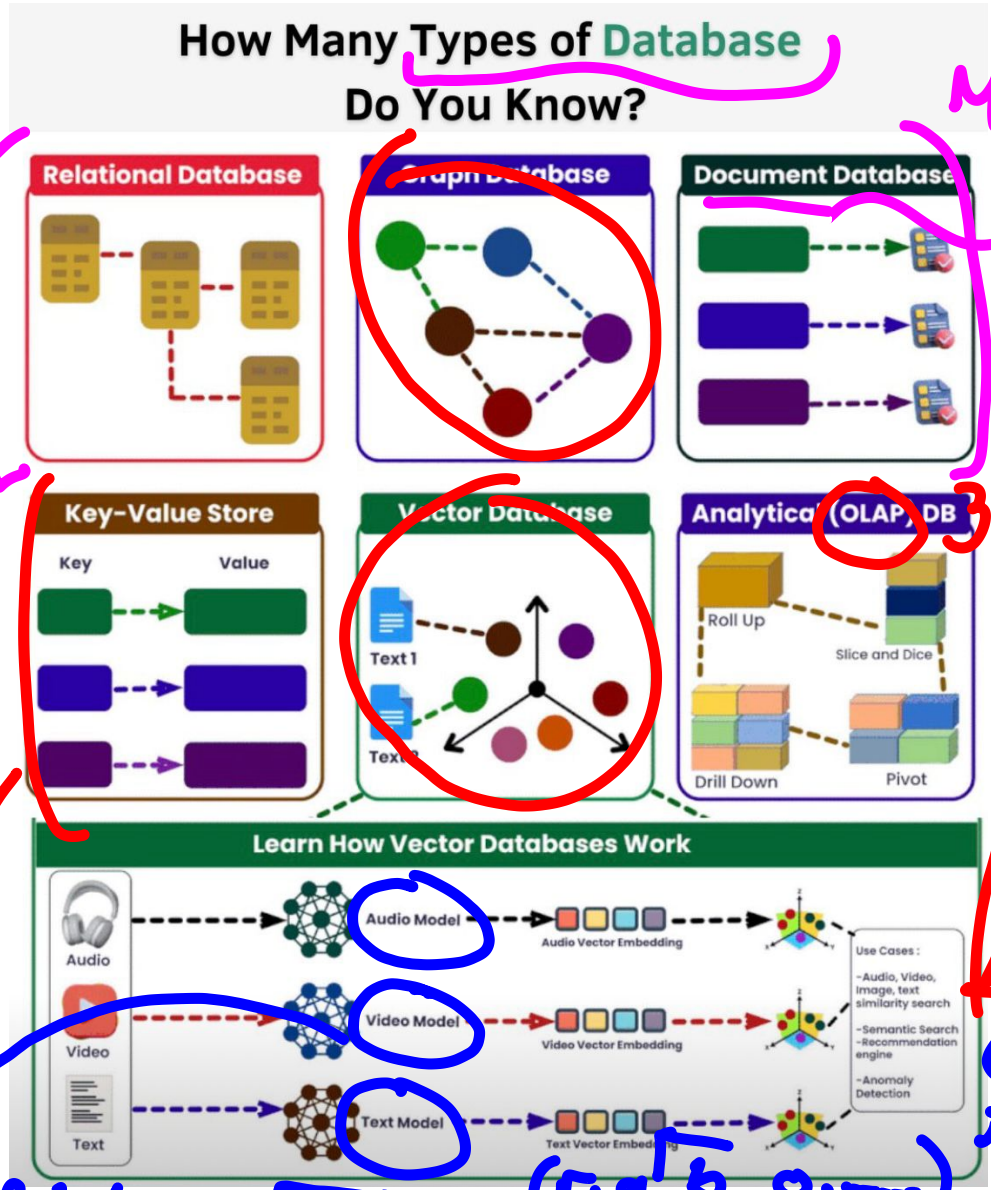


Data Lake Vs. Data WareHouse

- **Data Lake** is a storage repository that holds a vast amount of raw data in its native format until it is needed. It is a place to store every type of data in its native format with no fixed limits on account size or file.
- **Data Warehouse** is a system used for reporting and data analysis, and is considered a core component of business intelligence. It is a central repository of integrated data from one or more disparate sources.

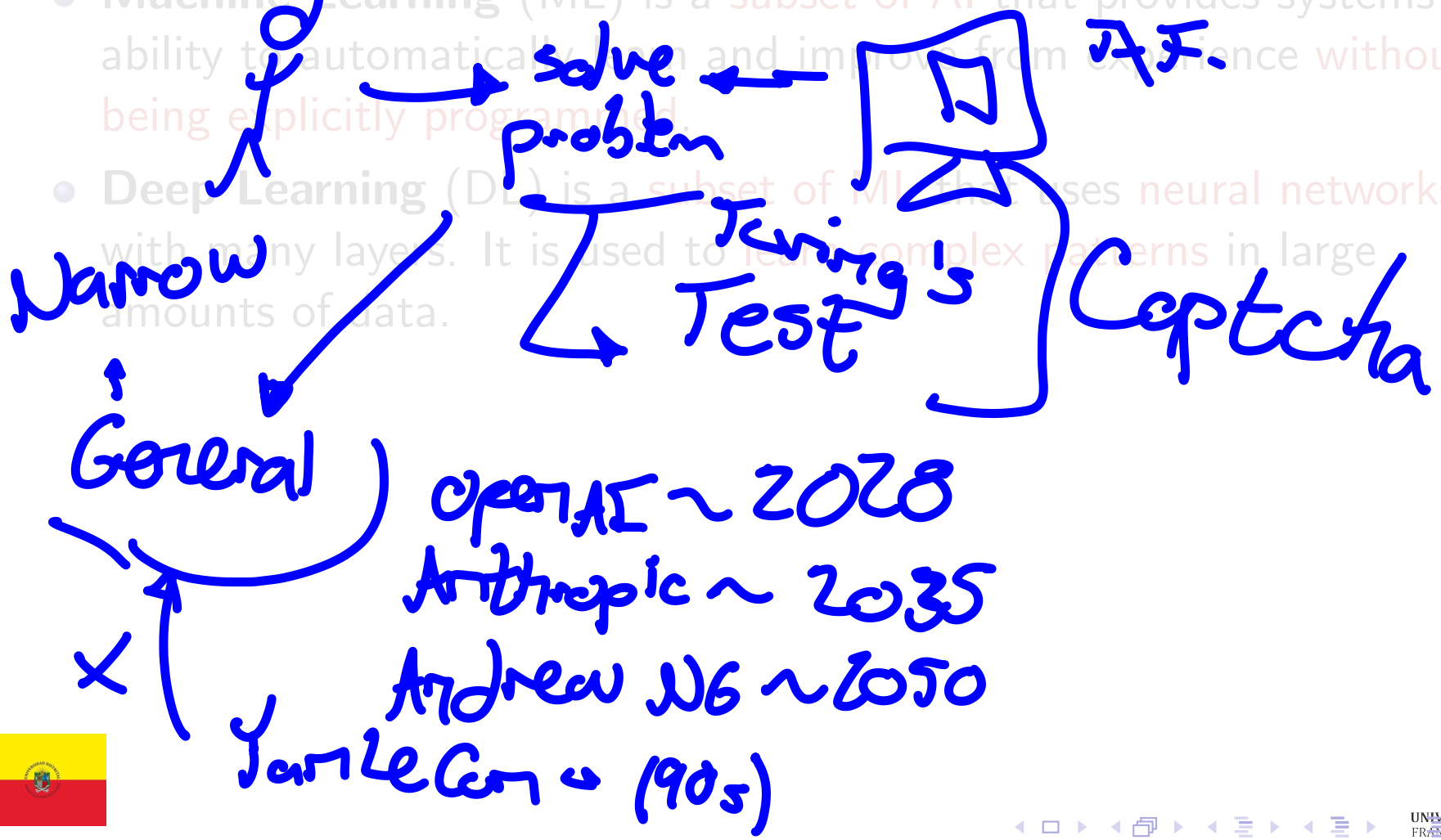


Types of Database



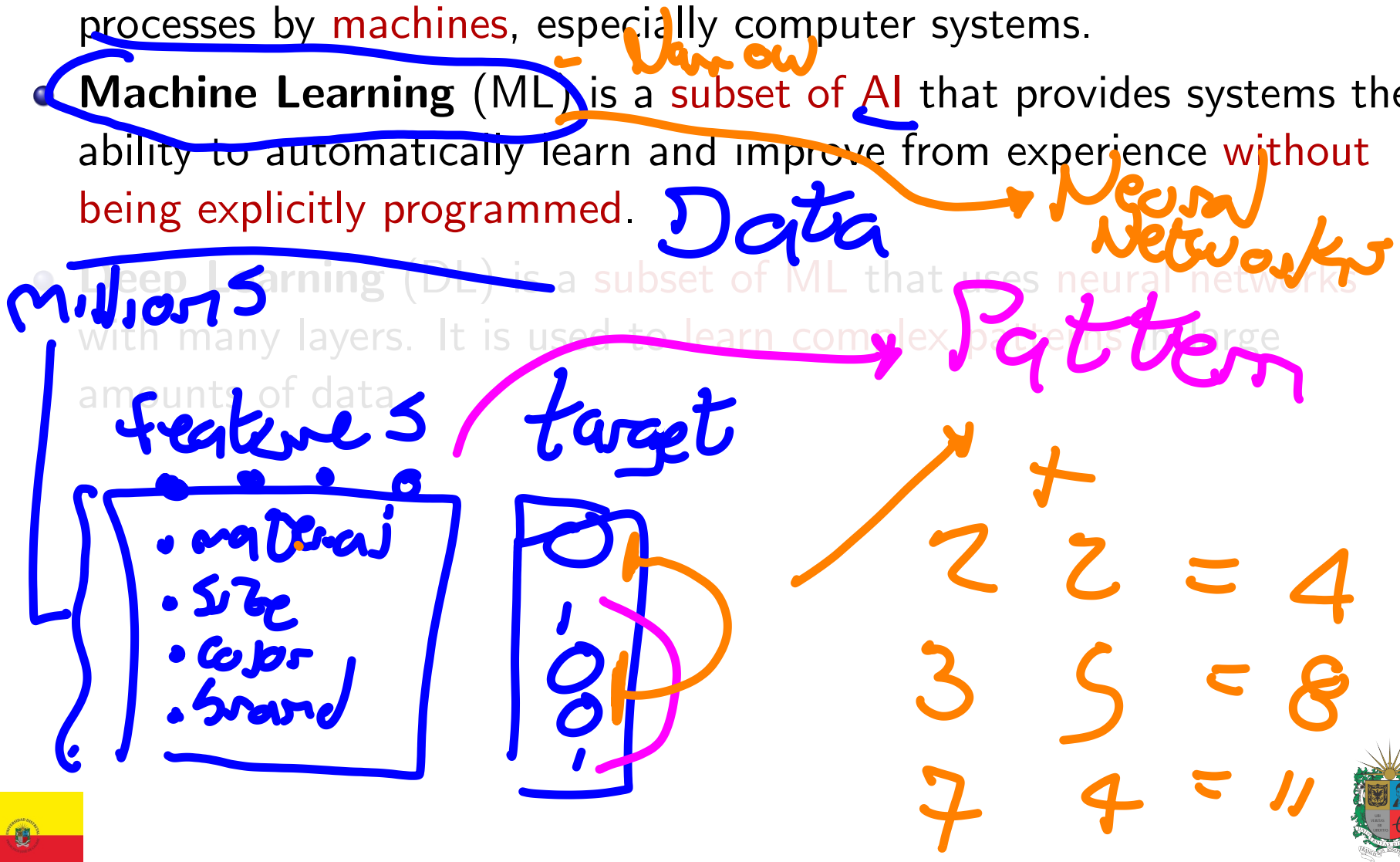
Artficial Inteligence, Machine Learning, Deep Learning

- **Artificial Intelligence** (AI) is the simulation of human intelligence processes by machines, especially computer systems.
- **Machine Learning** (ML) is a subset of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- **Deep Learning** (DL) is a subset of ML that uses neural networks with many layers. It is used to learn complex patterns in large amounts of data.



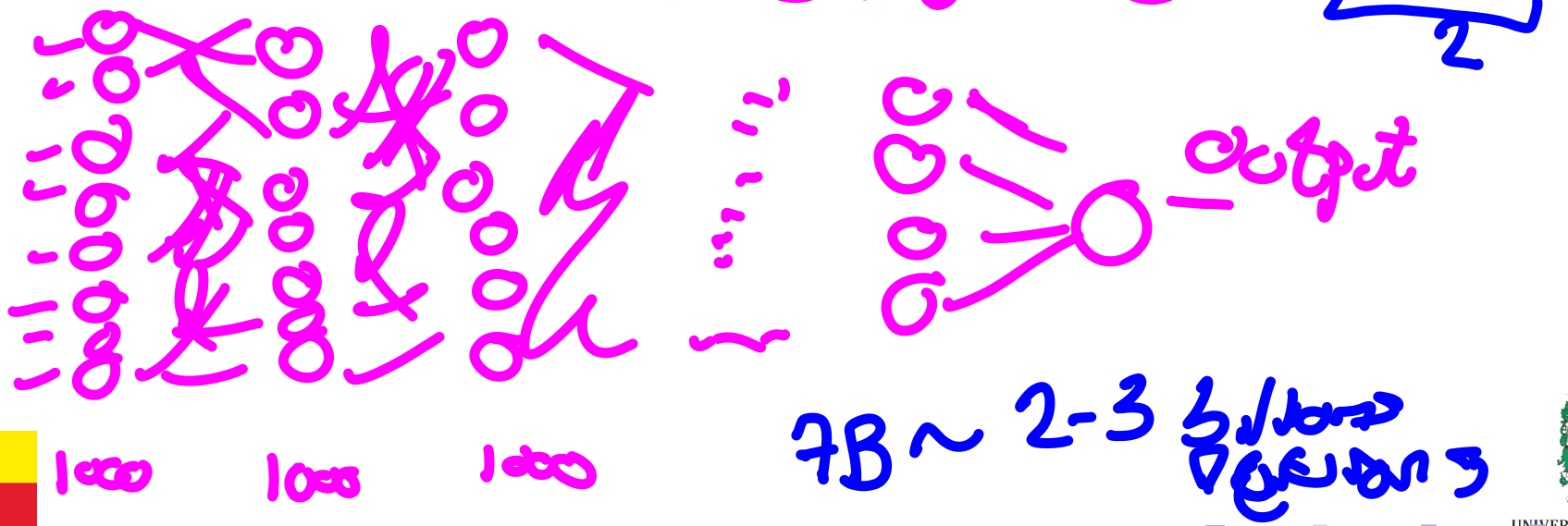
Artficial Inteligence, Machine Learning, Deep Learning

- **Artificial Intelligence** (AI) is the simulation of human intelligence processes by machines, especially computer systems.
- **Machine Learning** (ML) is a subset of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- **Deep Learning** (DL) is a subset of ML that uses neural networks with many layers. It is used to learn complex patterns in large amounts of data.

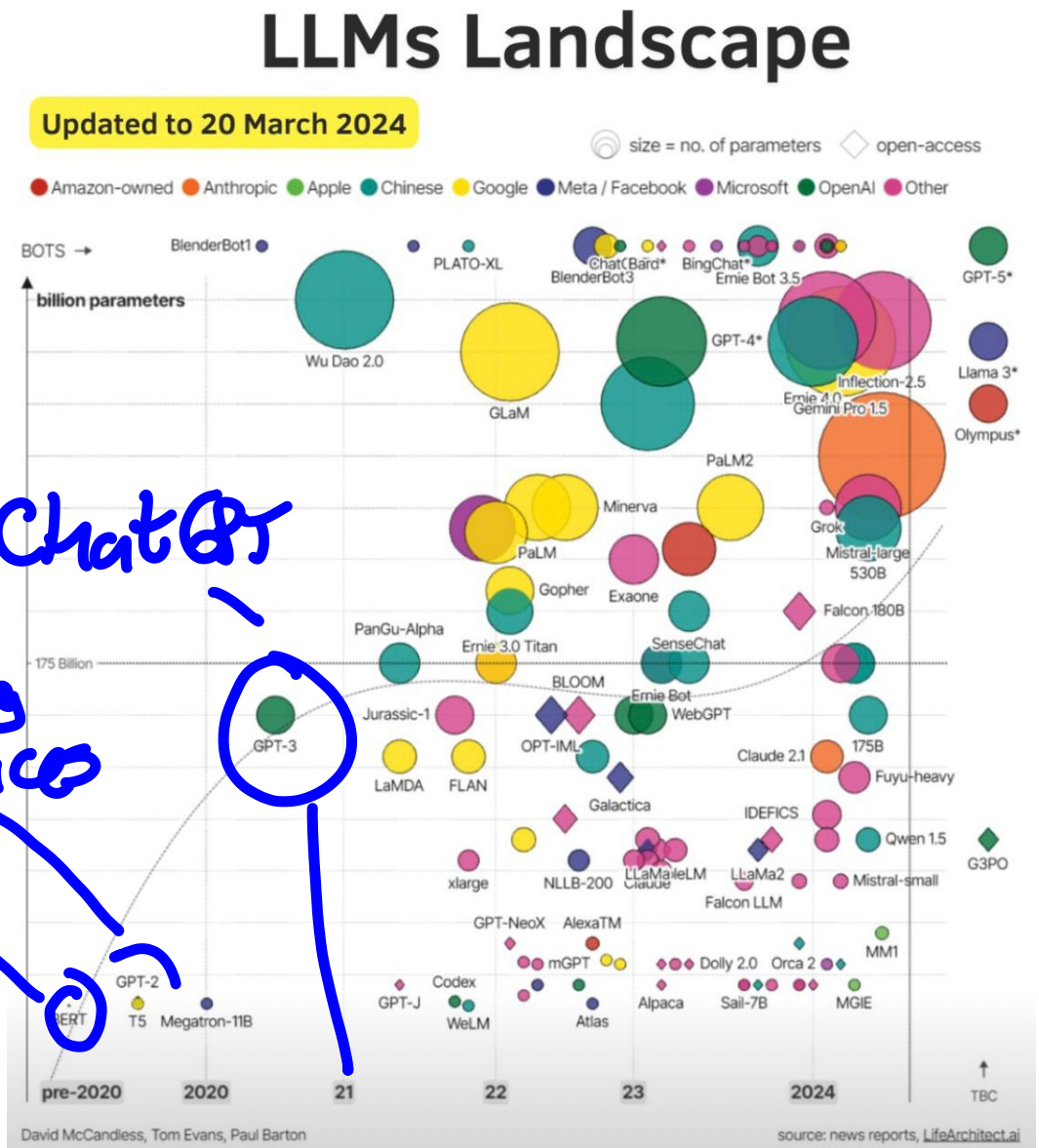


Artficial Inteligence, Machine Learning, Deep Learning

- **Artificial Intelligence** (AI) is the simulation of human intelligence processes by machines, especially computer systems.
- **Machine Learning** (ML) is a subset of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- **Deep Learning** (DL) is a subset of ML that uses neural networks with many layers. It is used to learn complex patterns in large amounts of data.



Large Language Models



Data and MetaData

- **Data** refers to raw, unprocessed, and unorganized facts or details that alone might not make much sense or provide context.
- **Metadata** is data about data. It provides the who, what, where, when, why, and how of the data.
- Examples of metadata include file size, creation date, modified date, and file type for a digital file.
- **Metadata** helps in data discovery, organization, and interpretation.
- **Metadata** is crucial in data management practices like data governance, data cataloging, and data lineage.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Data and MetaData

- **Data** refers to raw, unprocessed, and unorganized facts or details that alone might not make much sense or provide context.
- **Metadata** is data about data. It provides the who, what, where, when, why, and how of the data.
- Examples of metadata include file size, creation date, modified date, and file type for a digital file.
- **Metadata** helps in data discovery, organization, and interpretation.
- **Metadata** is crucial in data management practices like data governance, data cataloging, and data lineage.

Data and MetaData

- **Data** refers to raw, unprocessed, and unorganized facts or details that alone might not make much sense or provide context.
- **Metadata** is data about data. It provides the who, what, where, when, why, and how of the data.
- Examples of metadata include file size, creation date, modified date, and file type for a digital file.
- **Metadata** helps in data discovery, organization, and interpretation.
- **Metadata** is crucial in data management practices like data governance, data cataloging, and data lineage.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

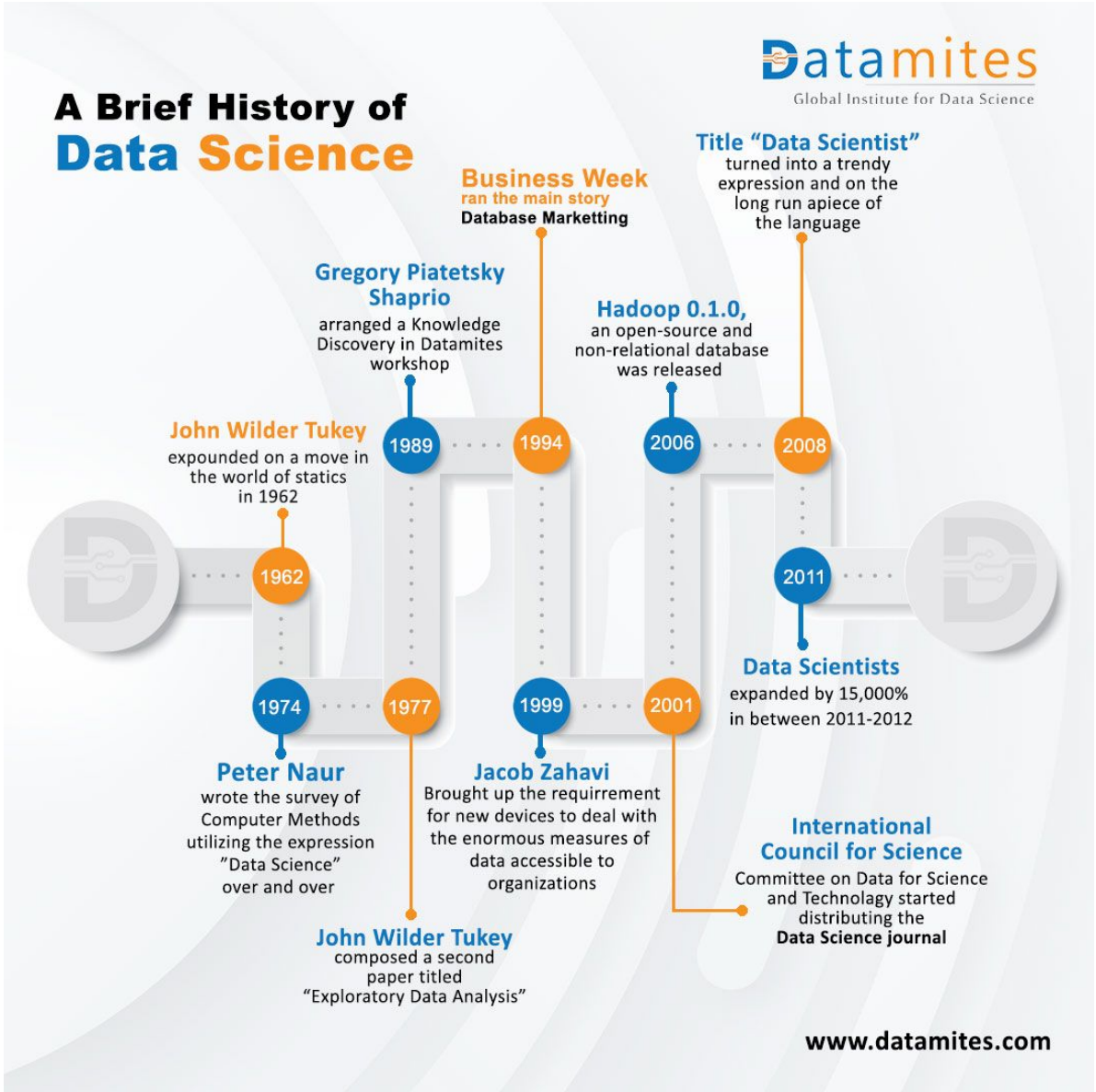
Data and MetaData

- **Data** refers to raw, unprocessed, and unorganized facts or details that alone might not make much sense or provide context.
- **Metadata** is data about data. It provides the who, what, where, when, why, and how of the data.
- Examples of metadata include file size, creation date, modified date, and file type for a digital file.
- **Metadata** helps in data discovery, organization, and interpretation.
- **Metadata** is crucial in data management practices like data governance, data cataloging, and data lineage.

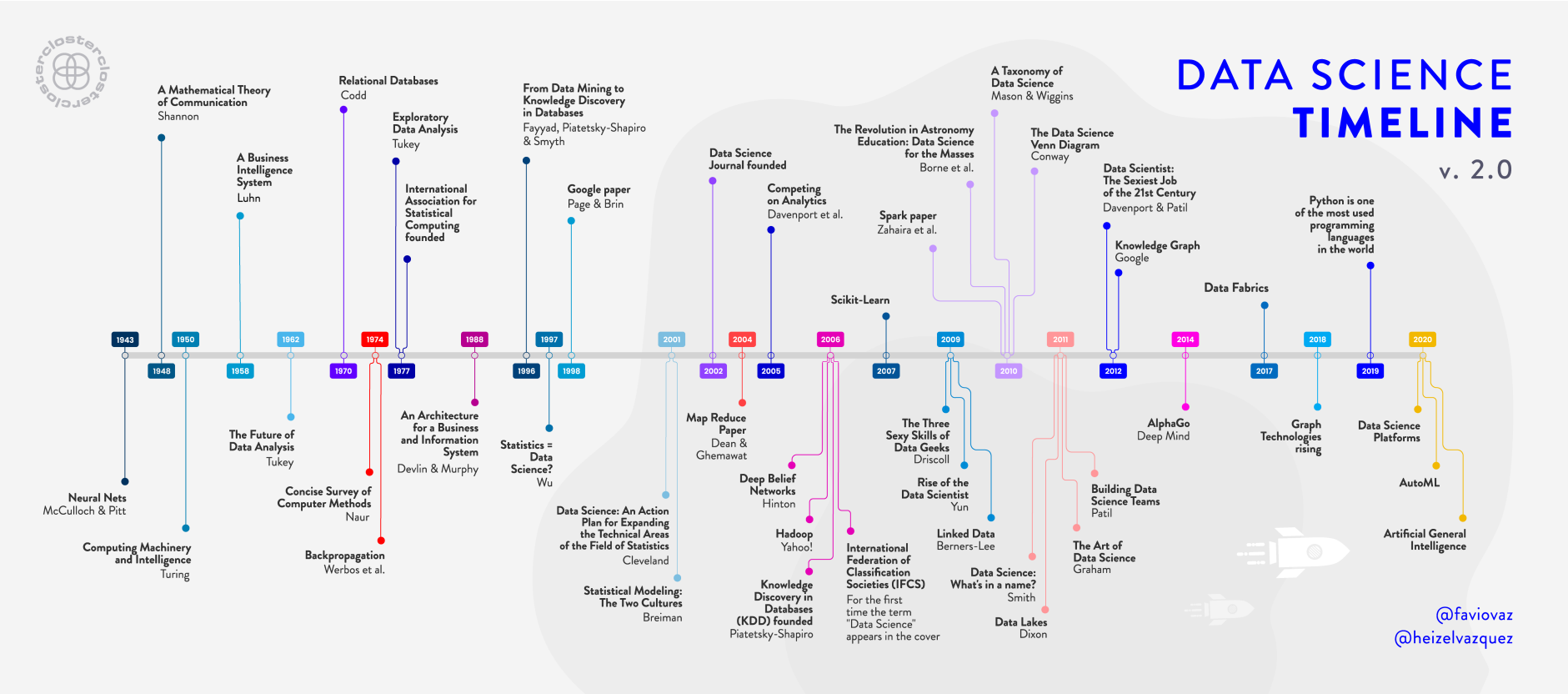
Data and MetaData

- **Data** refers to raw, unprocessed, and unorganized facts or details that alone might not make much sense or provide context.
- **Metadata** is data about data. It provides the who, what, where, when, why, and how of the data.
- Examples of metadata include file size, creation date, modified date, and file type for a digital file.
- **Metadata** helps in data discovery, organization, and interpretation.
- **Metadata** is crucial in data management practices like data governance, data cataloging, and data lineage.

Brief History of Data Science



Data Science Big Timeline



Data Science in Industry

- **Data Science** is used in many industries to make decisions, optimize processes, and increase efficiency.
- **Data Science** is used in healthcare to predict patient outcomes, optimize treatment plans, and personalize medicine.
- **Data Science** is used in finance to detect fraud, predict stock prices, and automate trading.
- **Data Science** is used in retail to optimize pricing, forecast demand, and personalize marketing.
- **Data Science** is used in manufacturing to predict equipment failures, optimize supply chains, and improve quality control.
- **Data Science** is used in transportation to optimize routes, predict maintenance, and improve safety.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Data Science in Industry

- **Data Science** is used in many industries to make decisions, optimize processes, and increase efficiency.
- **Data Science** is used in healthcare to predict patient outcomes, optimize treatment plans, and personalize medicine.
- **Data Science** is used in finance to detect fraud, predict stock prices, and automate trading.
- **Data Science** is used in retail to optimize pricing, forecast demand, and personalize marketing.
- **Data Science** is used in manufacturing to predict equipment failures, optimize supply chains, and improve quality control.
- **Data Science** is used in transportation to optimize routes, predict maintenance, and improve safety.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Data Science in Industry

- **Data Science** is used in many industries to make decisions, optimize processes, and increase efficiency.
- **Data Science** is used in healthcare to predict patient outcomes, optimize treatment plans, and personalize medicine.
- **Data Science** is used in finance to detect fraud, predict stock prices, and automate trading.
- **Data Science** is used in retail to optimize pricing, forecast demand, and personalize marketing.
- **Data Science** is used in manufacturing to predict equipment failures, optimize supply chains, and improve quality control.
- **Data Science** is used in transportation to optimize routes, predict maintenance, and improve safety.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Data Science in Industry

- **Data Science** is used in many industries to make decisions, optimize processes, and increase efficiency.
- **Data Science** is used in healthcare to predict patient outcomes, optimize treatment plans, and personalize medicine.
- **Data Science** is used in finance to detect fraud, predict stock prices, and automate trading.
- **Data Science** is used in retail to optimize pricing, forecast demand, and personalize marketing.
- **Data Science** is used in manufacturing to predict equipment failures, optimize supply chains, and improve quality control.
- **Data Science** is used in transportation to optimize routes, predict maintenance, and improve safety.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Data Science in Industry

- **Data Science** is used in many industries to make decisions, optimize processes, and increase efficiency.
- **Data Science** is used in healthcare to predict patient outcomes, optimize treatment plans, and personalize medicine.
- **Data Science** is used in finance to detect fraud, predict stock prices, and automate trading.
- **Data Science** is used in retail to optimize pricing, forecast demand, and personalize marketing.
- **Data Science** is used in manufacturing to predict equipment failures, optimize supply chains, and improve quality control.
- **Data Science** is used in transportation to optimize routes, predict maintenance, and improve safety.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

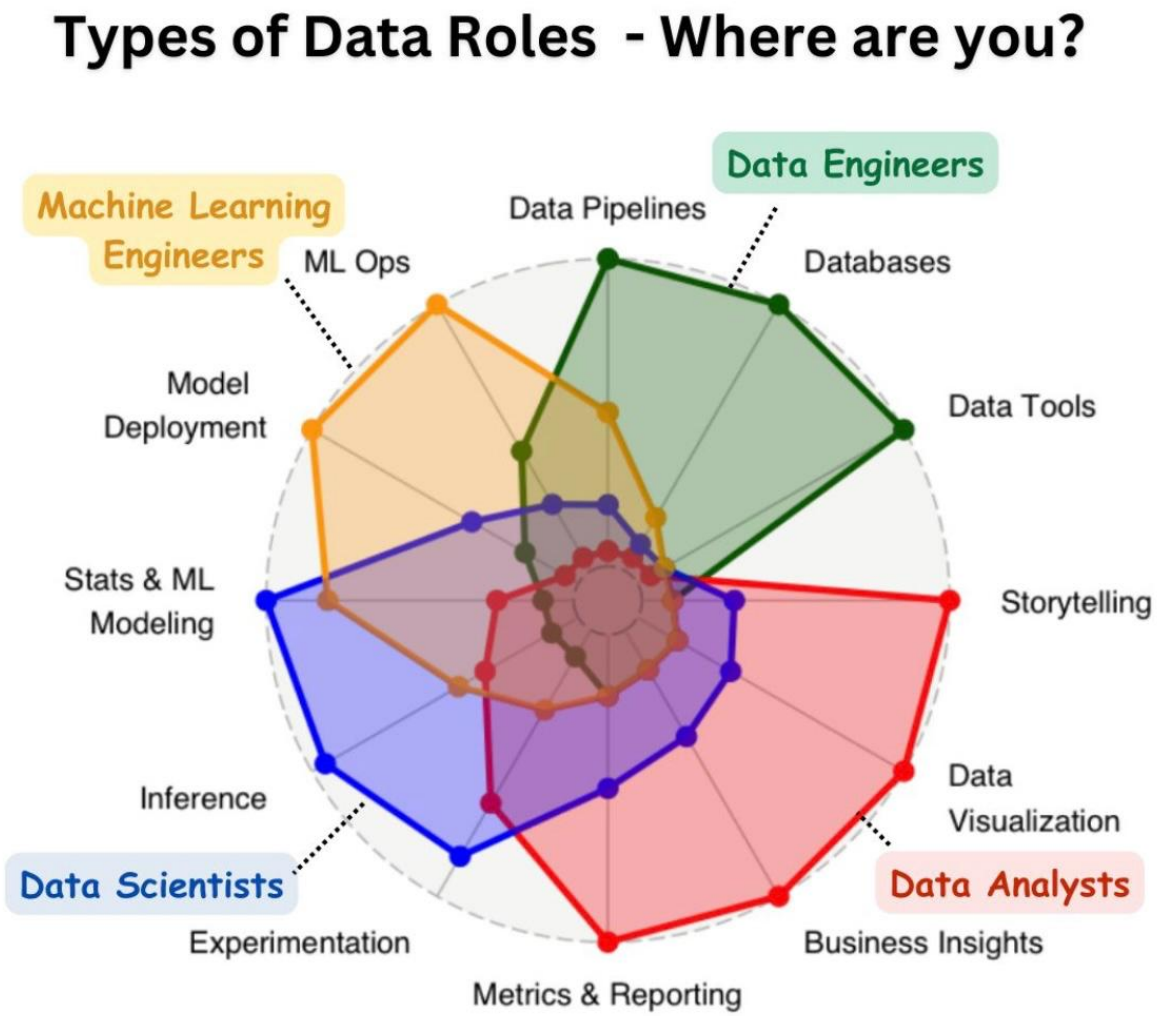
Data Science in Industry

- **Data Science** is used in many industries to make decisions, optimize processes, and increase efficiency.
- **Data Science** is used in healthcare to predict patient outcomes, optimize treatment plans, and personalize medicine.
- **Data Science** is used in finance to detect fraud, predict stock prices, and automate trading.
- **Data Science** is used in retail to optimize pricing, forecast demand, and personalize marketing.
- **Data Science** is used in manufacturing to predict equipment failures, optimize supply chains, and improve quality control.
- **Data Science** is used in transportation to optimize routes, predict maintenance, and improve safety.

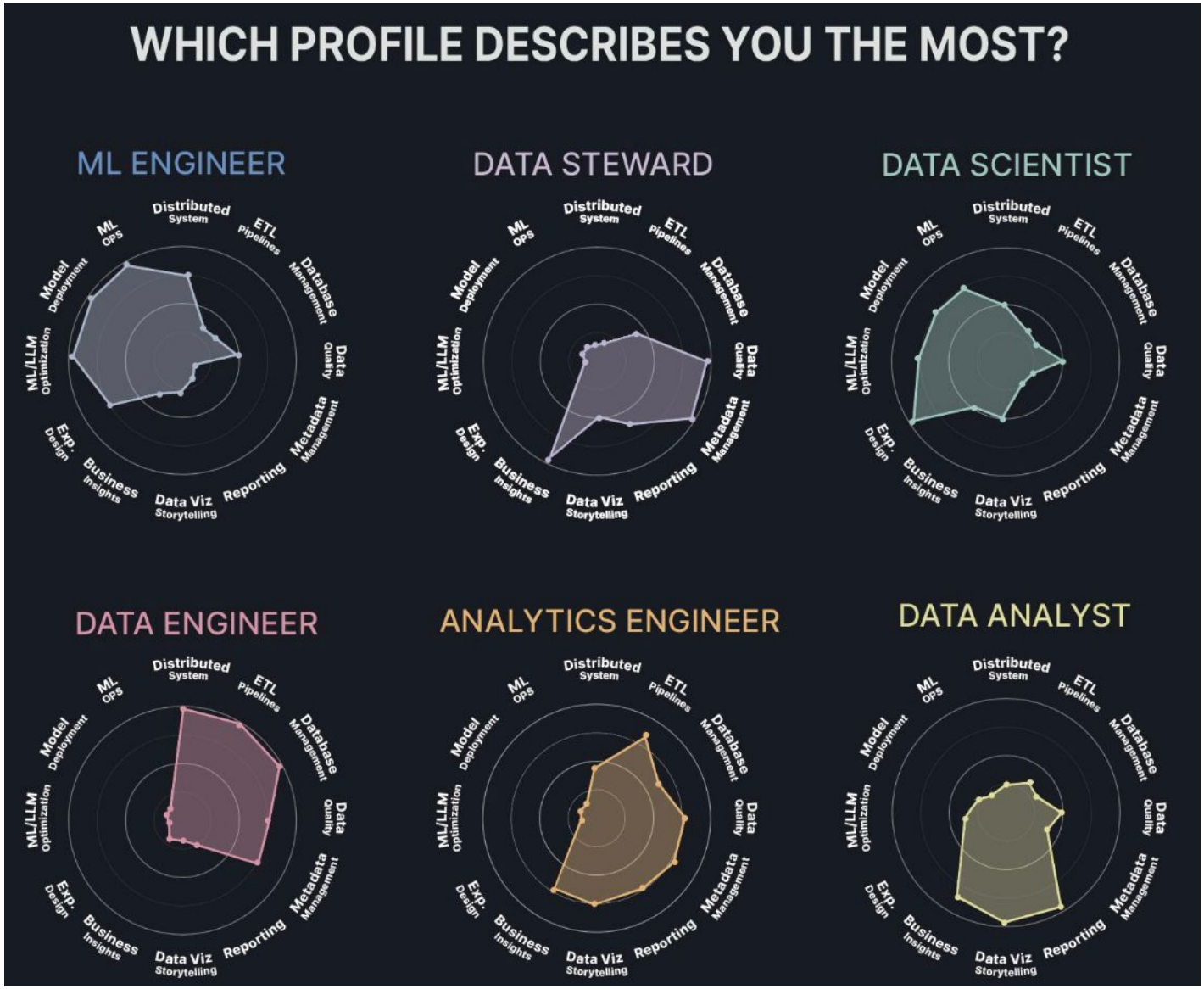
UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Outline

Tech Team — Roles



Tech Team — Data Profiles

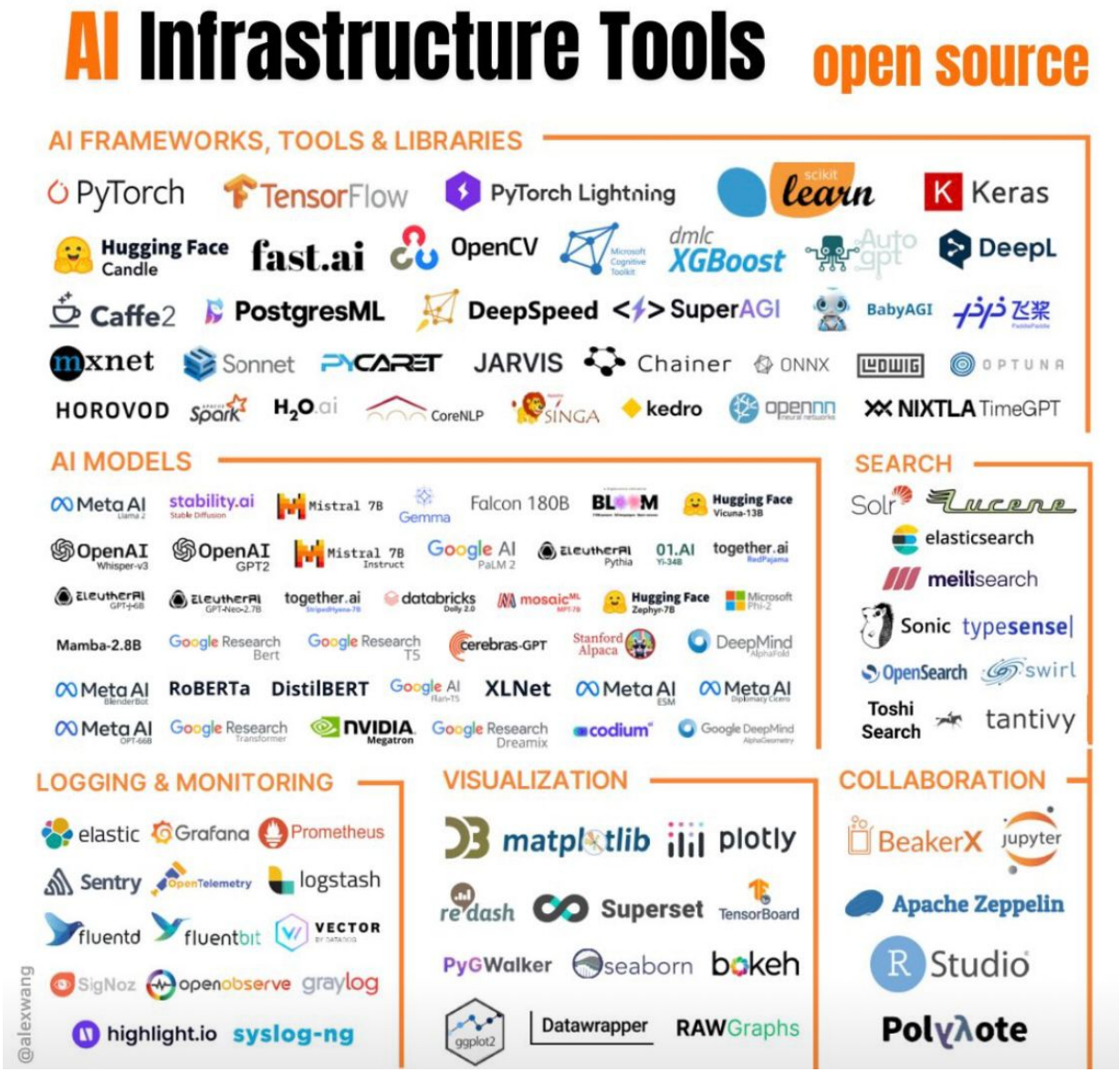


Data Scientist Responsabilities

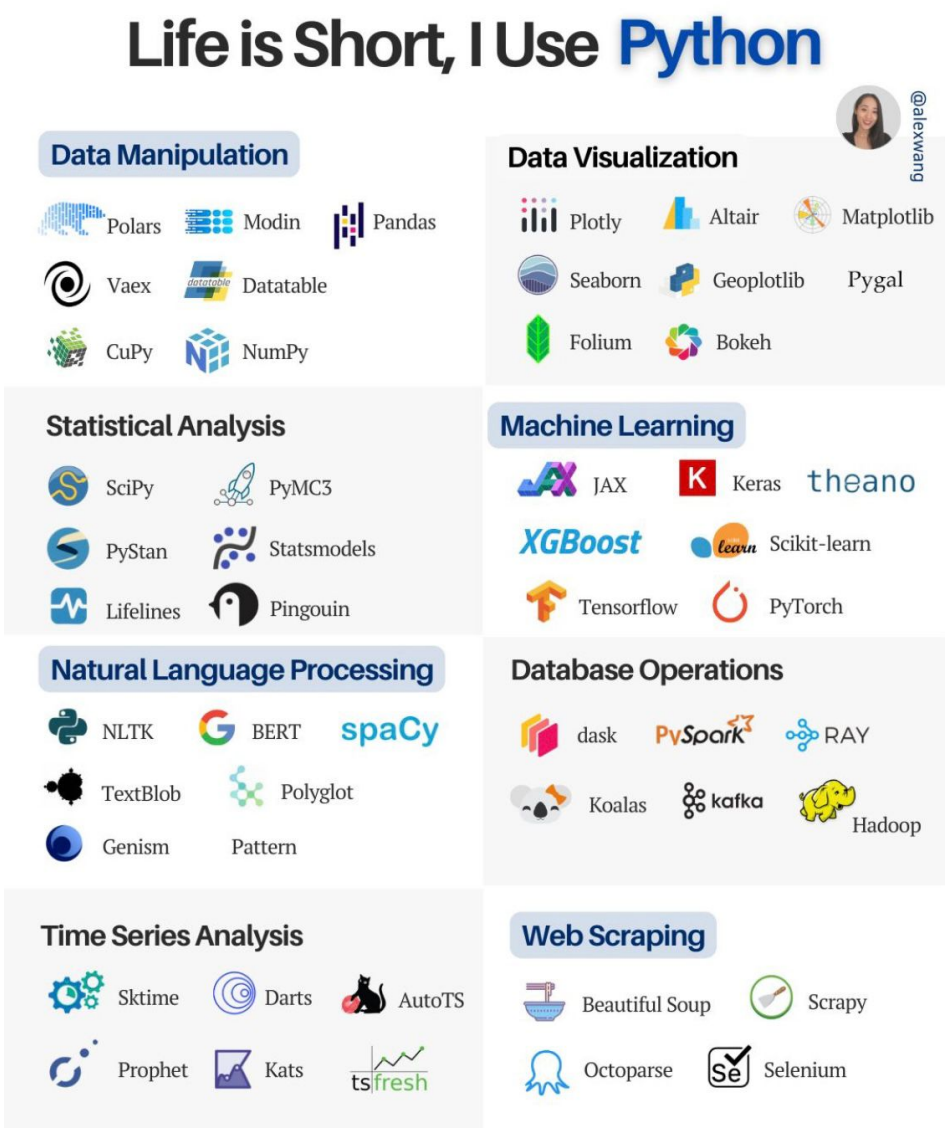
- Collecting large sets of structured and unstructured data from **disparate sources**.
- Cleaning and validating the data to ensure **accuracy**, **completeness**, and uniformity.
- Analyzing the data to identify **patterns** and trends.
- Interpreting the data to discover solutions and **opportunities**.
- Communicating findings to stakeholders using **visualization** and other means.
- Developing, prototyping, and implementing **machine learning models**.
- Staying current on techniques and tools in the field, and continually **improving skills**.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Artificial Intelligence Tech Ecosystem



Data Science Python Tech Ecosystem



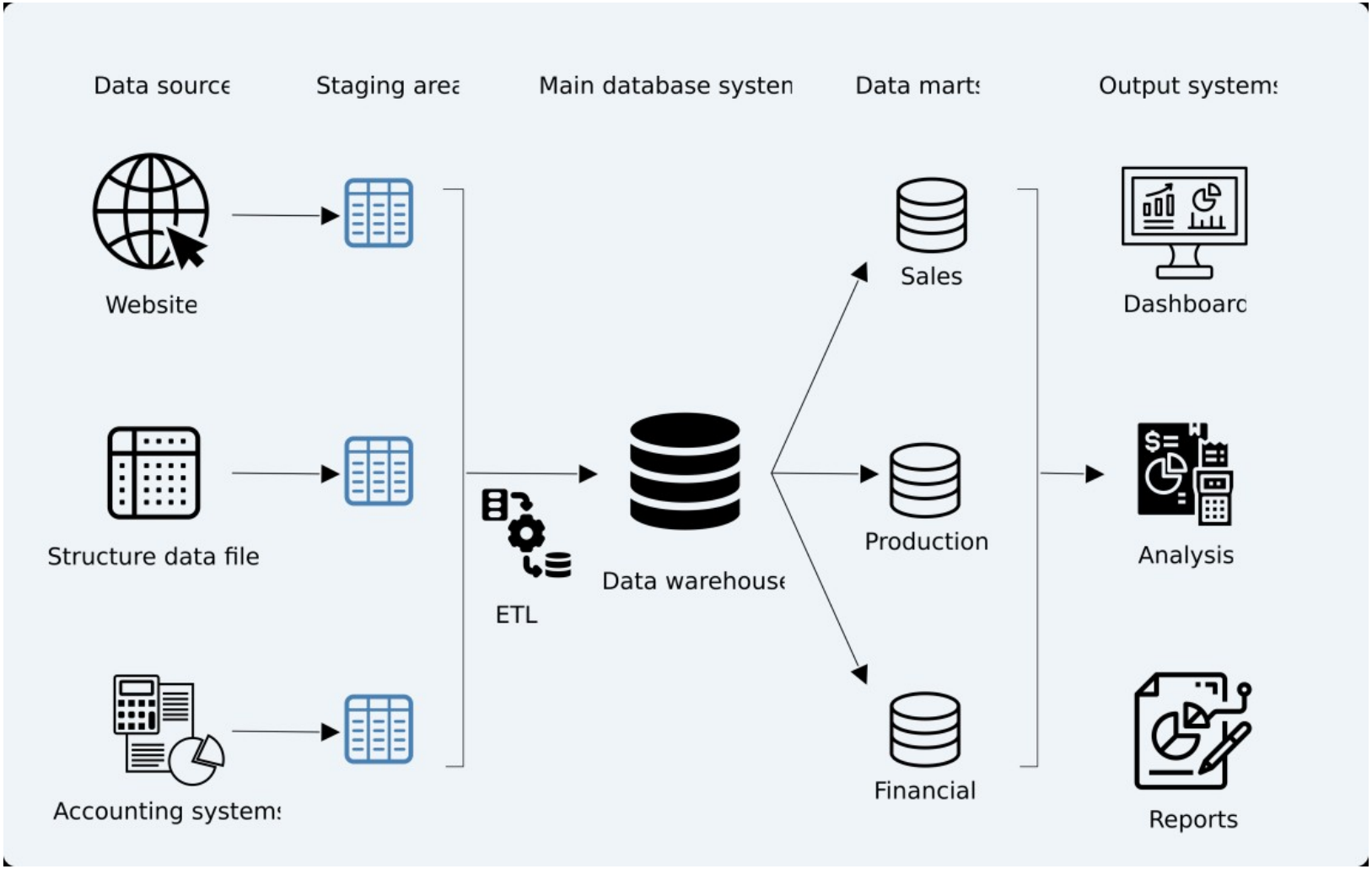
Data Sources and Formats

Data Definition Framework

Data Format

Data Source	Structured	Unstructured
Internal	**Human-Generated** • Survey ratings • Aptitude testing **Machine-Generated** • Web metrics from Web logs • Product purchase from sales Records • Process control measures	**Human-Generated** • Emails, letters, text messages • Audio transcripts • Customer comments • Voicemails • Corporate video/communications • Pictures, illustrations • Employee reviews
External	**Human-Generated** • Number of Retweets, Facebook likes, Google Plus +1s • Ratings on Yelp • Patient ratings ratings **Machine-Generated** • GPS for tweets • Time of tweet/updates/postings	**Human-Generated** • Content of social media updates • Comments in online forums • Comments on Yelp • Video reviews • Pinterest images • Surveillance video

Data Pipelines



Outline

Thanks!

Questions?

Repo: *https://github.com/EngAndres/ud-public/tree/main/courses/data-analysis-programming*

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS