

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

- 1 Data Science Basic Concepts
- 2 What is to be a Data Scientist



Outline

- 1 Data Science Basic Concepts
- 2 What is to be a Data Scientist



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**.



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**.



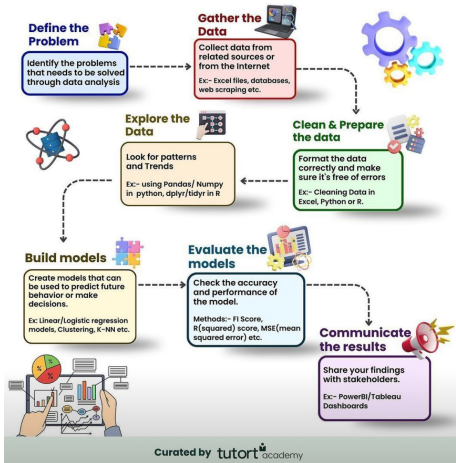
Process of Data Analysis



tutort.academy
#APlaceToGrow

Save for later

Process of Data Analysis



Types of Data Analysis

Different Types of Data Analysis



Descriptive Analysis

Summarizes basic data characteristics, like averages, frequencies, and distributions.

Answer 'What happened?'



Diagnostic Analysis

Digs deeper, identifying factors influencing specific outcomes or trends.

Answer 'Why did this happen?'



Predictive Analysis

Forecasts future events or values based on historical data patterns.

Answer 'What might happen next?'



Prescriptive Analysis

Recommends optimal actions based on predicted outcomes and potential impacts.

Answer 'What should we do now?'



Other Important Types That Are Widely Used

Exploratory Data Analysis (EDA)

Initial, open-ended investigation to understand data characteristics and uncover potential insights.

Inferential Analysis

Draws conclusions about a larger population based on a sample, using statistical methods to assess confidence levels.

Causal Analysis

Identifies factors causing specific outcomes, often using advanced techniques like causal inference methods.

Mechanistic Analysis

Examines underlying mechanisms and relationships within complex systems, often involving simulation or modeling.



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).



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).



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)**.



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

HISTORY OF DATA

19,000 BC



The **Ishango bone** holds the first evidence of data collection and storage.

1600s



John Graunt introduces the **concept of data analysis** in 1663.

1800s



Herman Hollerith designs a **machine that helped complete** the US census in 1890.

1900s



Fritz Pfleumer invents the **magnetic tape** which later inspired the invention of floppy disks and hard disk drives.

1990s



Sir Tim Berners-Lee invents the **World Wide Web**.



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.



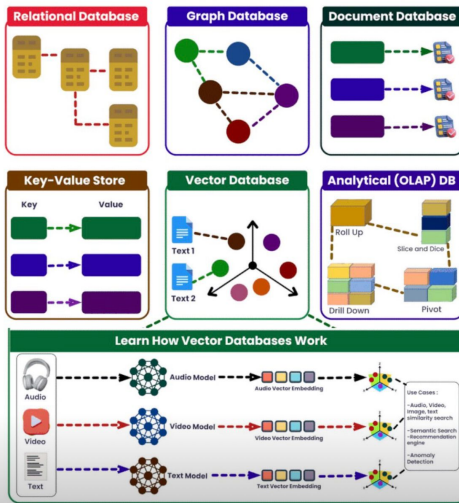
Data Lake Vs. Data Warehouse

- **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

How Many Types of Database Do You Know?



Artificial Intelligence, 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.



Artificial Intelligence, 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.



Artificial Intelligence, Machine Learning, Deep Learning

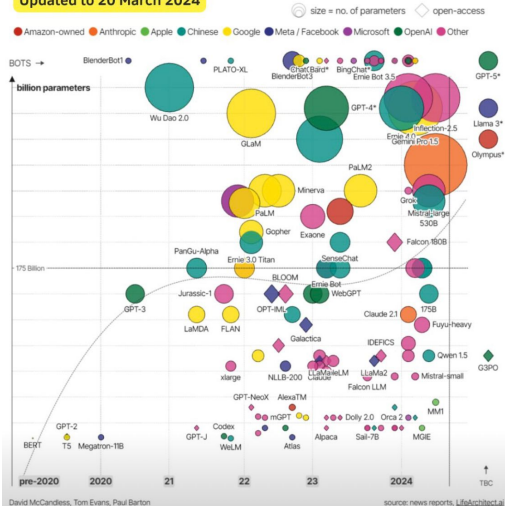
- **Artificial Intelligence (AI)** is the **simulation** of human intelligence processes by **machines**, especially computer systems.
- **Machine Learning (ML)** is a **subset of AI** that provides systems the ability to automatically learn and improve from experience **without being explicitly programmed**.
- **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

LLMs Landscape

Updated to 20 March 2024



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**.



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**.



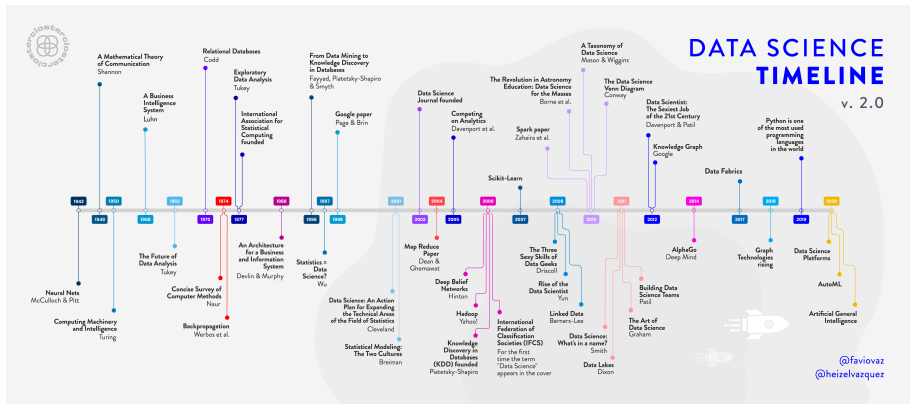
Data and MetaData

- **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 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.



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.



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.



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.



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.



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.



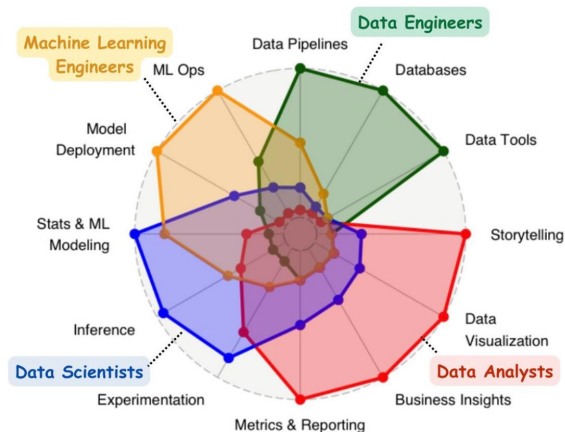
Outline

- 1 Data Science Basic Concepts
- 2 What is to be a Data Scientist



Tech Team — Roles

Types of Data Roles - Where are you?



Tech Team — Data Profiles

WHICH PROFILE DESCRIBES YOU THE MOST?

ML ENGINEER



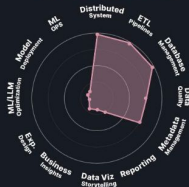
DATA STEWARD



DATA SCIENTIST



DATA ENGINEER



ANALYTICS ENGINEER



DATA ANALYST



Data Scientist Responsibilities

- 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**.



Artificial Intelligence Tech Ecosystem

AI Infrastructure Tools open source

AI FRAMEWORKS, TOOLS & LIBRARIES



AI MODELS



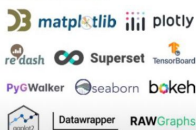
SEARCH



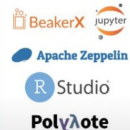
LOGGING & MONITORING



VISUALIZATION



COLLABORATION



@alexwang



Data Science Python Tech Ecosystem

Life is Short, I Use Python

Data Manipulation



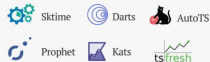
Statistical Analysis



Natural Language Processing



Time Series Analysis



Data Visualization



Machine Learning



Database Operations



Web Scraping



Data Sources and Formats

Data Definition Framework

Data Format

Data Source

Internal



Structured



Human-Generated

- Survey ratings
- Aptitude testing

Machine-Generated

- Web metrics from Web logs
- Product purchase from sales Records
- Process control measures

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

Unstructured



Human-Generated

- Emails, letters, text messages
- Audio transcripts
- Customer comments
- Voicemails
- Corporate video/communications
- Pictures, illustrations
- Employee reviews

Human-Generated

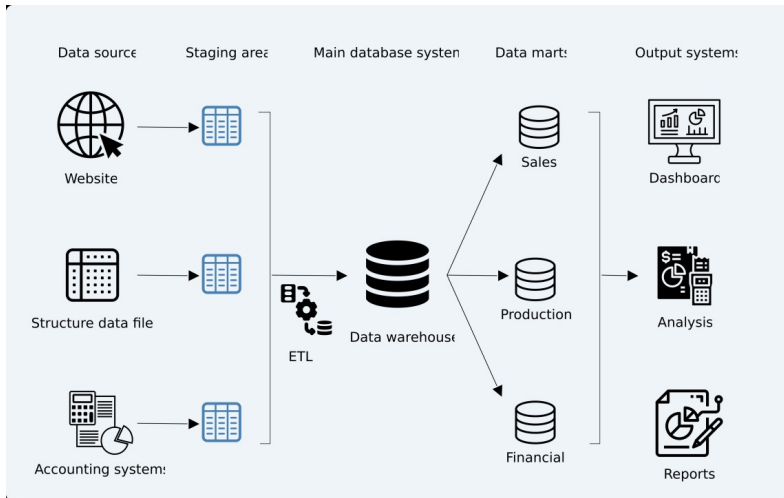
- Content of social media updates
- Comments in online forums
- Comments on Yelp
- Video reviews
- Pinterest images
- Surveillance video



Copyright 2014 Business Over Broadway

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Data Pipelines



Outline

- 1 Data Science Basic Concepts
- 2 What is to be a Data Scientist



Thanks!

Questions?



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

