

INTRODUCTION TO DATA SCIENCE WITH PYTHON

Introduction to Data Science

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-II



Outline

1 Data Science Basic Concepts

2 What is to be a Data Scientist

3 Foundations of Python



Outline

1 Data Science Basic Concepts

2 What is to be a Data Scientist

3 Foundations of Python



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.
 - ↳ observation
 - ↳ question
 - ↳ test
 - ↳ analysis
 - It is primarily used to make decisions and predictions making of predictive causal analytics, prescriptive analytics, and machine learning.
 - ↳ best
 - ↳ deploy
 - It is about uncovering findings from data, and it's about surfacing hidden insights that can help enable companies to make smarter business decisions.
- Agile*



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 **strategic decision making** using **statistical** and **machine learning** causal analytics, prescriptive analytics, and machine learning.
- It is about **uncovering findings from data**, and it's **not** enough to **Dev Ops**, **ML Ops**, **LLM Ops**, **business decisions**, **error**, **learning**



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.
- ↗ Pivot
↳ Only Foss



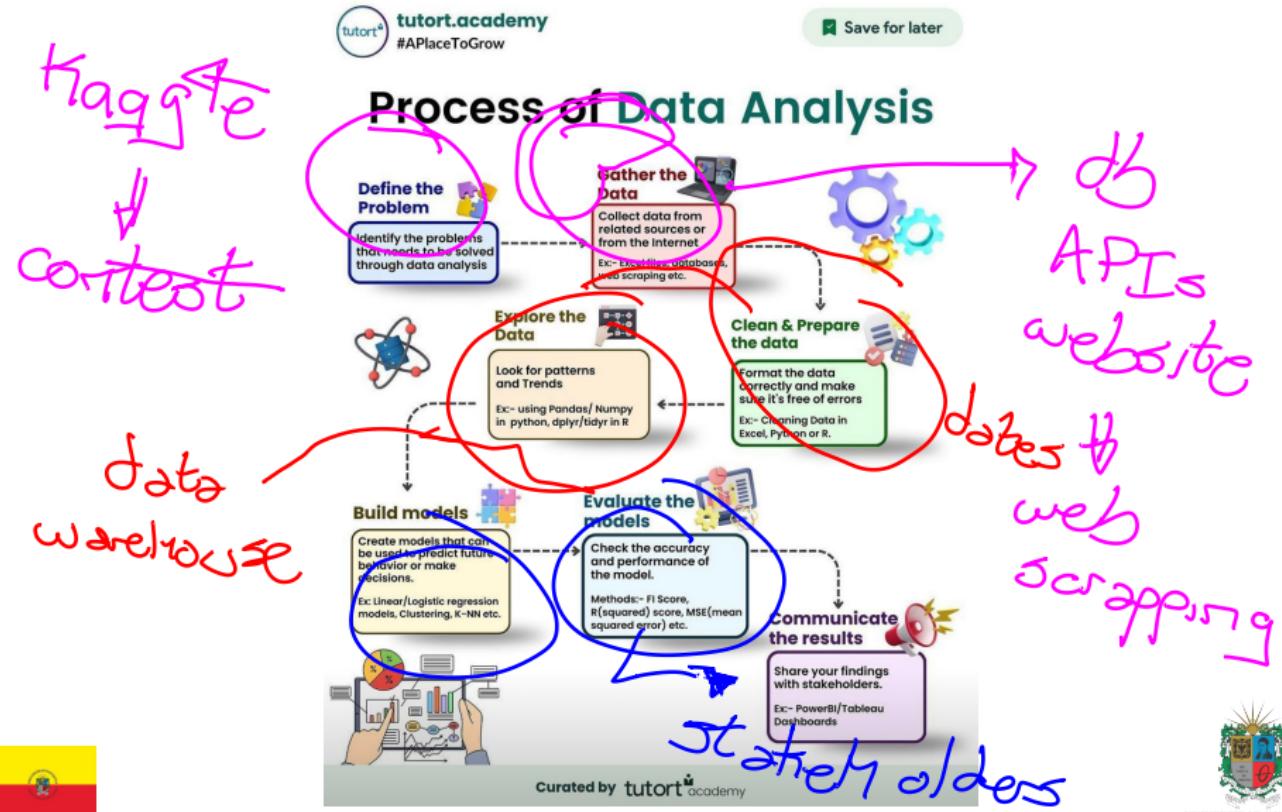
What is Data Science?

- **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**.

dark data



Process of Data Analysis



Types of Data Analysis

Forecasting

Different Types of Data Analysis

Averages
Frequencies
↓
facts



Descriptive Analysis

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

Answer "What happened?"



Predictive Analysis

Forecasts future events or values based on historical data patterns.

Answer "What might happen next?"



Diagnostic Analysis

Digs deeper, identifying factors influencing specific outcomes or trends.

Answer "Why did this happen?"



Prescriptive Analysis

Recommends optimal actions based on predicted outcomes and potential impacts.

Answer "What should we do now?"



Alex

wrong factors

understand

the business

business intelligence



variables
,
what if?



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 ~~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**.



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).
real-time



Data Systems & Big Data

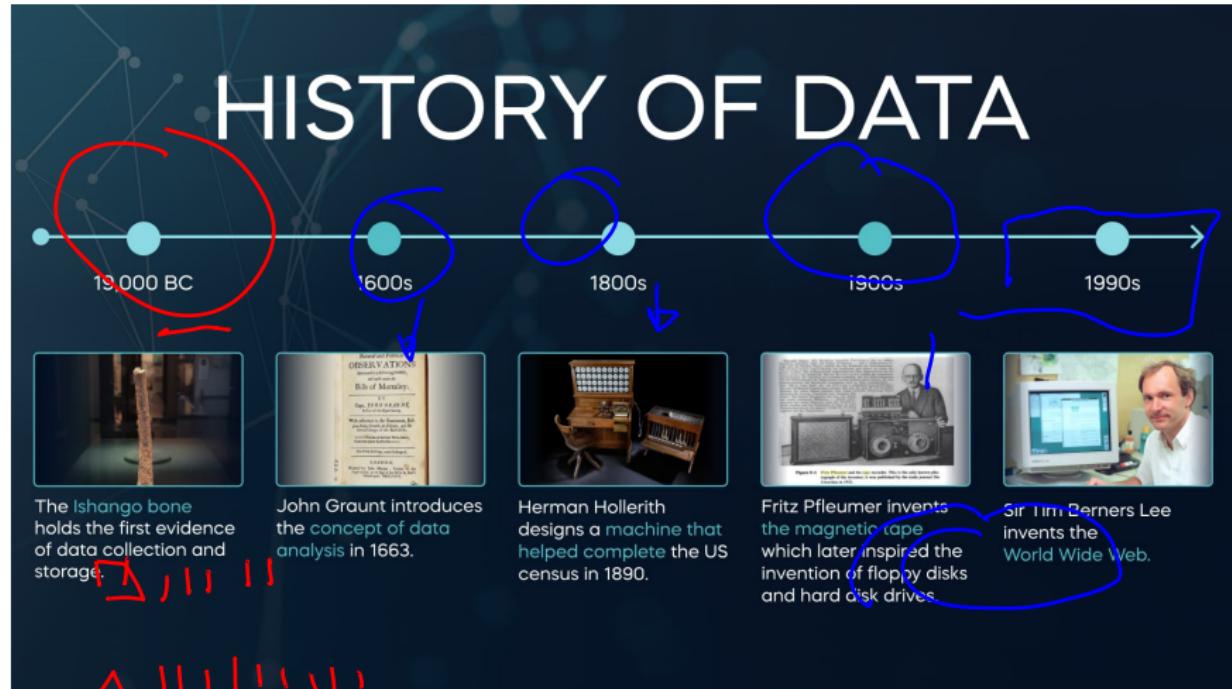
- **Big Data** refers to extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.
- **Data Systems** are the mechanisms to store, retrieve, and send data. They are crucial for handling **big data**.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

• HDFS

Storage (drive)



History of Data



Data Lake Vs. Data Warehouse

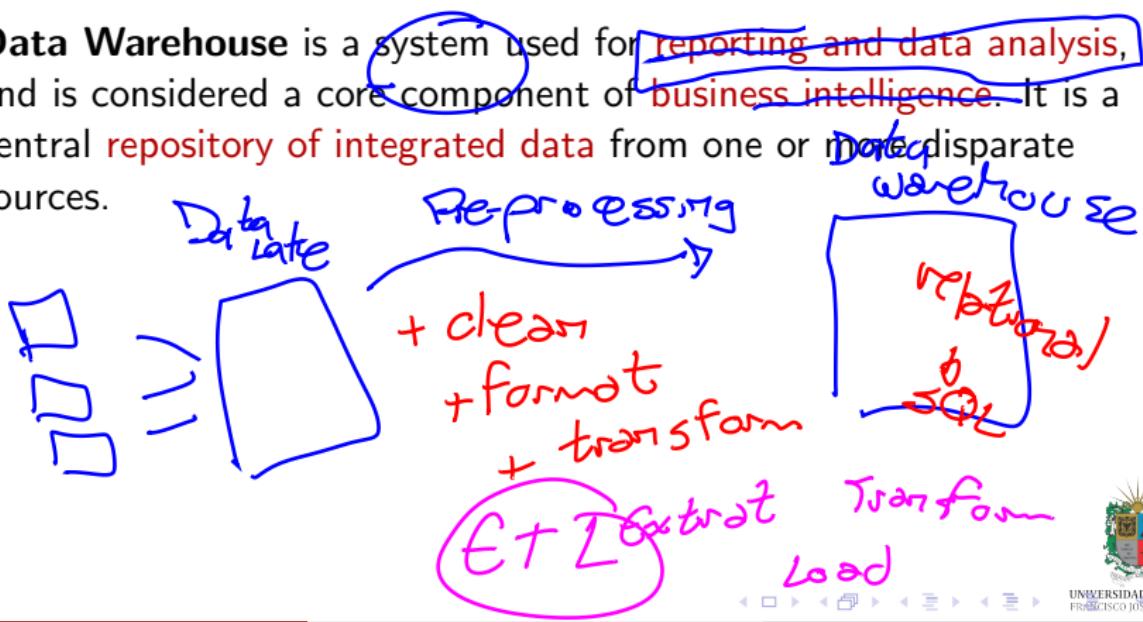
- **Data Lake** is a **storage repository** that holds a vast amount of **raw data** in its **native format** until it is needed. It is a place to **store every type of data in its native format** with no fixed limits on account size or file.

- **Data Warehouse** is a system used for reporting and data analysis, 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

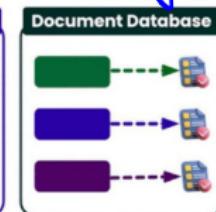
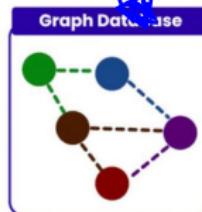
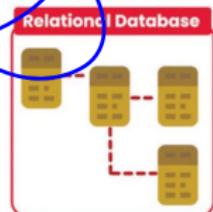
- **Data Lake** is a **storage repository** that holds a vast amount of **raw data** in its native format until it is needed. It is a place to **store every type of data in its native format** with no fixed limits on account size or file.
- **Data Warehouse** is a system used for **reporting and data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



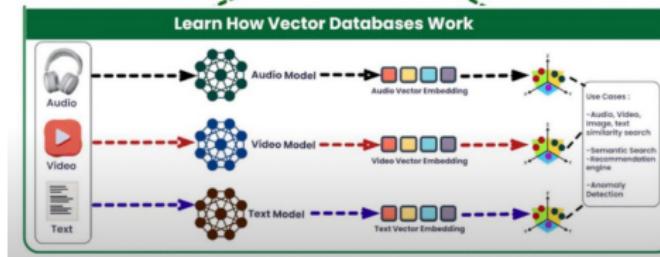
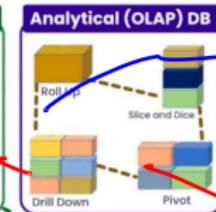
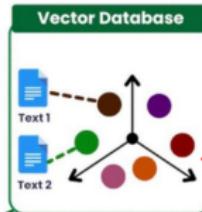
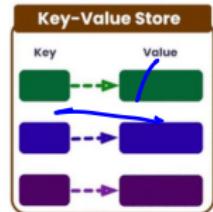
Types of Database

~~NoSQL~~
IMAP
cols

How Many Types of Database
Do You Know?



map
dict
hash



MongoDB

11

BSON

dog
cat

car

OLAP



Artificial Intelligence, Machine Learning, Deep Learning

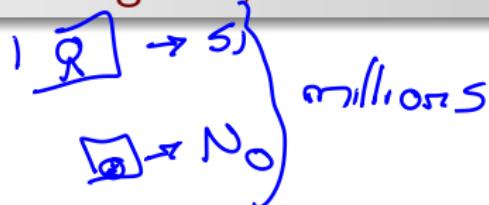
40^s => Alan Turing



- **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.
 - ↳ play
 - ↳ cryptography
- 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**.

Learn with examples

Train

Machine can solve a problem



Artificial Intelligence, Machine Learning, Deep Learning

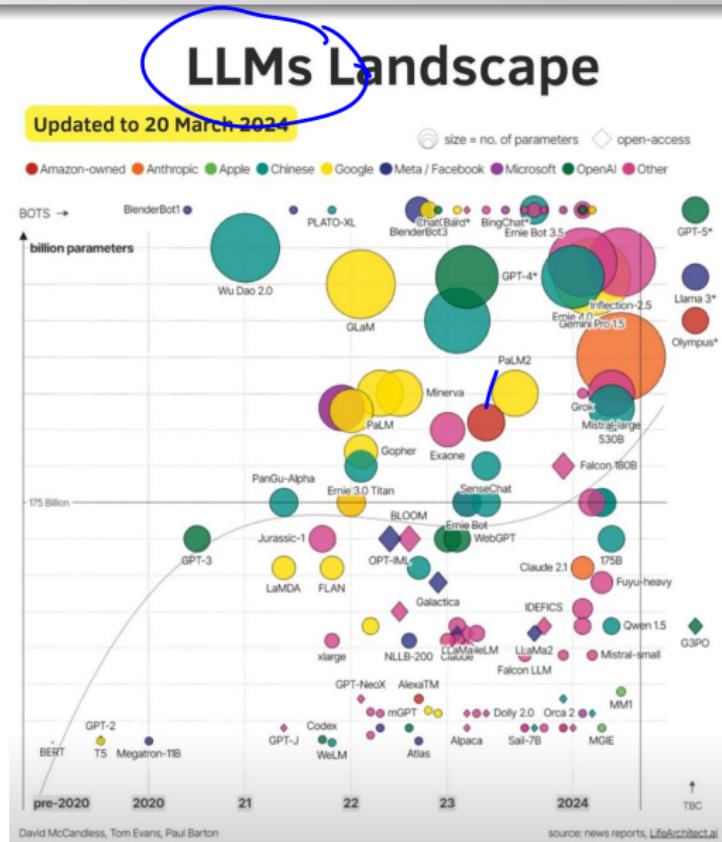


- **Artificial Intelligence** (AI) is the **simulation** of human intelligence processes by **machines**, especially computer systems.
- **Machine Learning** (ML) is a **subset of AI** that provides systems the ability to automatically learn and improve from experience **without being explicitly programmed**.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

Big Data



Large Language Models



3 Zettabytes
170 Billions

3 Zettabytes
20000 TA (GPo)



Data and MetaData

Data Lake

- **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 of a digital file.
- Metadata helps in **data discovery**, **organization**, and **interpretation**.
- Metadata is crucial in data management practices like **data governance**, **data cataloging**, and **dealing with A, B, C**.



↗ rows
↗ empty
disarr. button
groups ↗ A
↗ B
↗ C



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**.

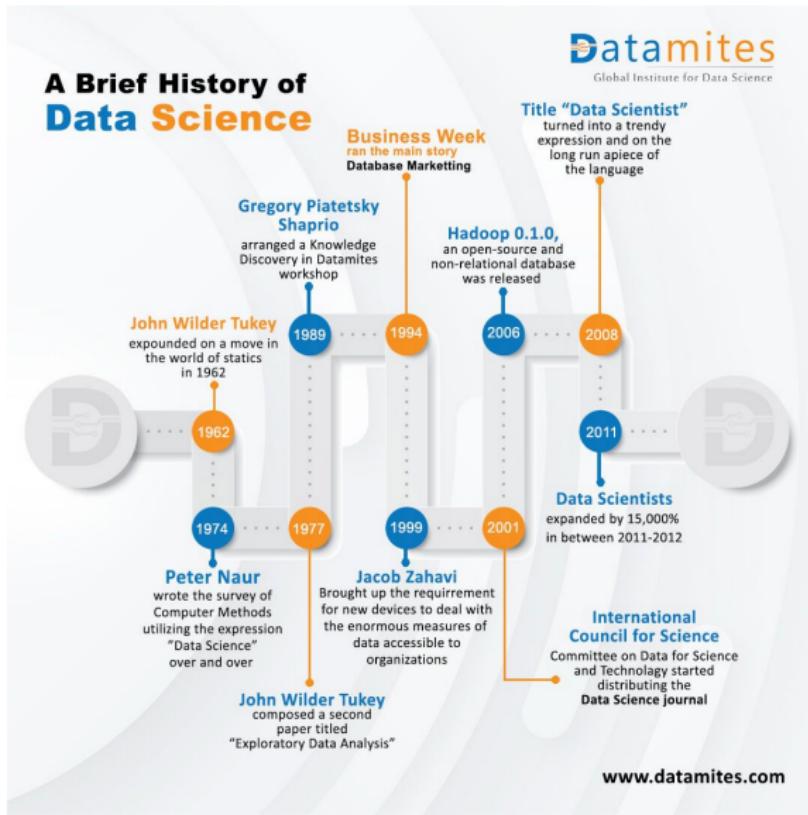


Data and MetaData

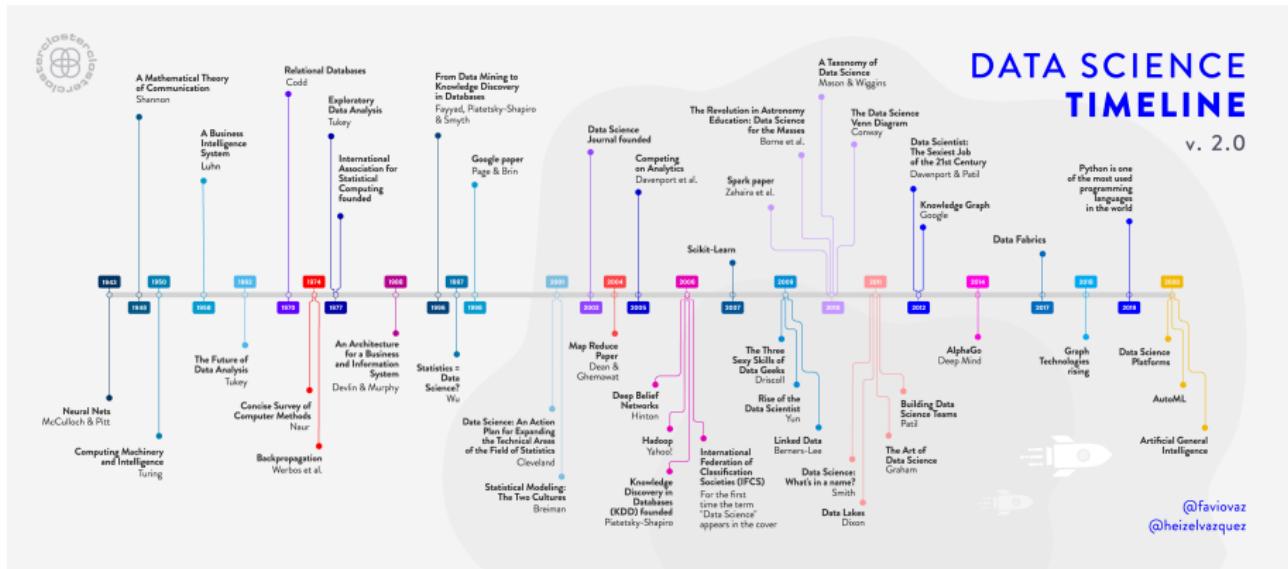
- **Data** refers to **raw**, **unprocessed**, and **unorganized** facts or details that alone might not make much sense or provide context.
- **Metadata** is **data about data**. It provides the who, what, where, when, why, and how of the data.
- Examples of metadata include file size, creation date, 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**.



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 personalise 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.
- Algo n'thnic
modelling } Bots*



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.

Social
Networks



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



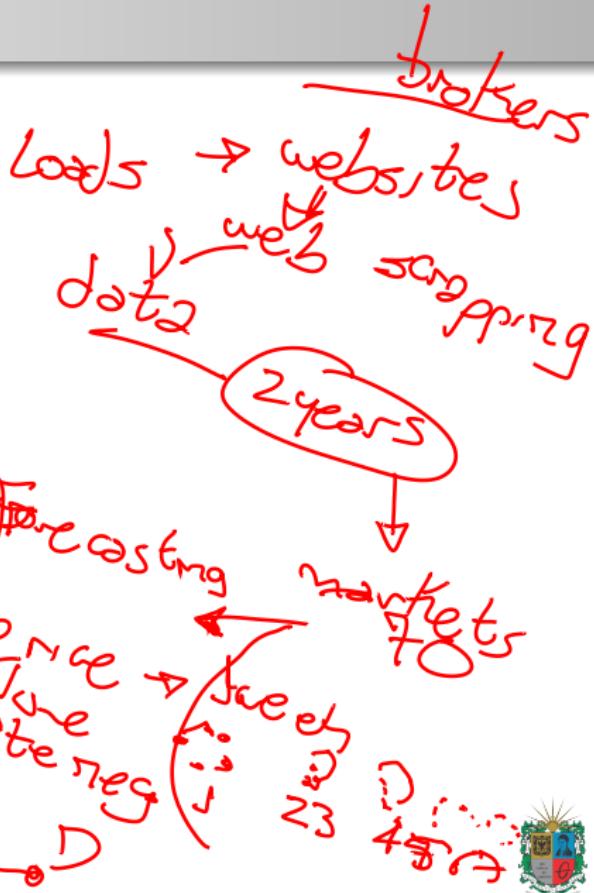
15
16 / -1

- 1 Data Science Basic Concepts

- 2 What is to be a Data Scientist

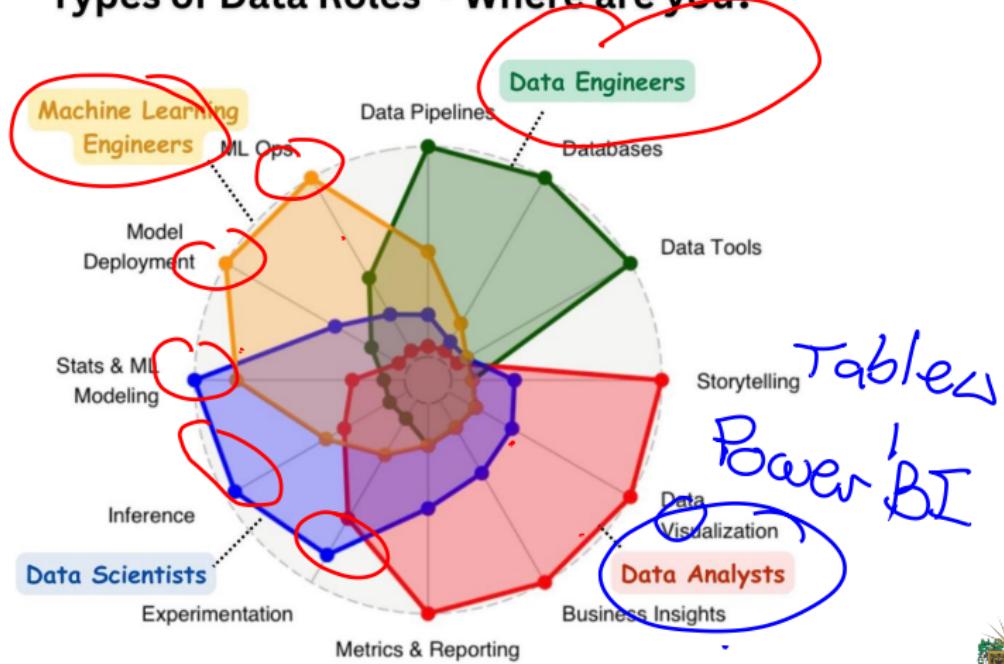
- 3 Foundations of Python

Reinforcement Learning
① Learning at B
C
D

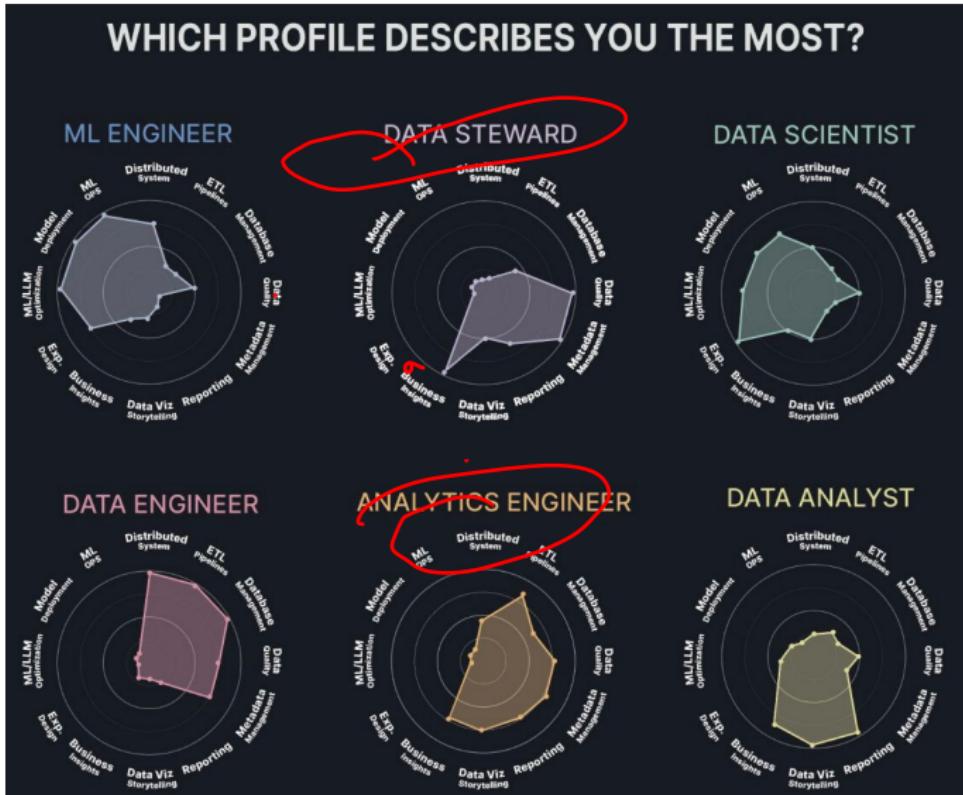


Tech Team — Roles

Types of Data Roles - Where are you?



Tech Team — Data Profiles



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

- PyTorch
- TensorFlow
- PyTorch Lightning
- learn
- Keras
- Hugging Face
- fast.ai
- OpenCV
- dmrc
- XGBoost
- AutoML
- DeepL
- Caffe2
- PostgresML
- DeepSpeed
- SuperAGI
- BabyAGI
- 飞桨
- mxnet
- PyTorch
- PyCARET
- JARVIS
- Chainer
- ONNX
- LWIG
- OPTUNA
- HOROVOD
- spark
- H2O.ai
- CoreNLP
- SINGA
- kedro
- opennn
- NIXTLA
- TimeGPT

AI MODELS

- Meta AI
- stability.ai
- Mistral 7B
- Gemma
- Falcon 180B
- BLOOM
- Hugging Face
- OpenAI
- OpenAI GPT2
- Mistral 7B Instruct
- Google AI
- EleutherAI
- together.ai
- EleutherAI Pythia
- together.ai
- EleutherAI GPT-7B
- EleutherAI
- together.ai
- dbnix
- mosaicml
- Hugging Face Dolly-7B
- Microsoft
- Cerebras-GPT
- Stanford Alpaca
- DeepMind
- Mamba-2.8B
- Google Research
- Google Research
- XLNet
- Meta AI
- Meta AI
- Meta AI
- RobERTa
- DistilBERT
- Google AI
- NVIDIA
- Google Research
- Dreamix
- codium
- Google DeepMind
- elastisearch
- mellisearch
- Sonic
- typesense
- OpenSearch
- swirl
- Toshi Search
- tantivy

LOGGING & MONITORING

- elastic
- CloudWatch Metrics
- Prometheus
- Sentry
- Cloud Telemetry
- logstash
- fluentd
- fluentbit
- VECTOR
- SigNoz
- openobservability
- graylog
- highlight.io
- syslog-ng

VISUALIZATION

- matplotlib
- plotly
- redash
- Superset
- TensorBoard
- PyGWall
- seaborn
- bokeh
- ggplot2
- Datawrapper
- RAWGraphs

COLLABORATION

- BeakerX
- jupyter
- Apache Zeppelin
- R Studio
- PolyNote

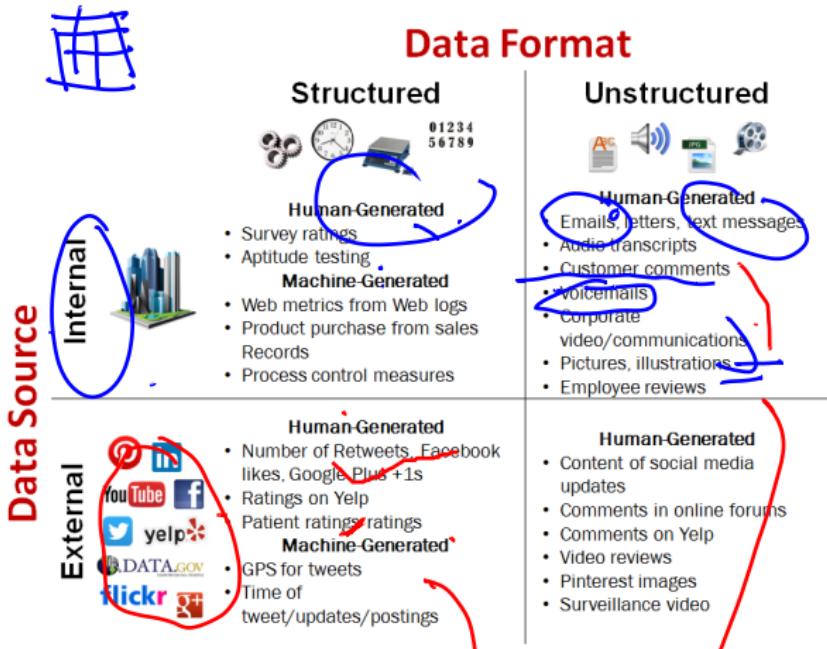
@alexwang

Data Science Python Tech Ecosystem



Data Sources and Formats

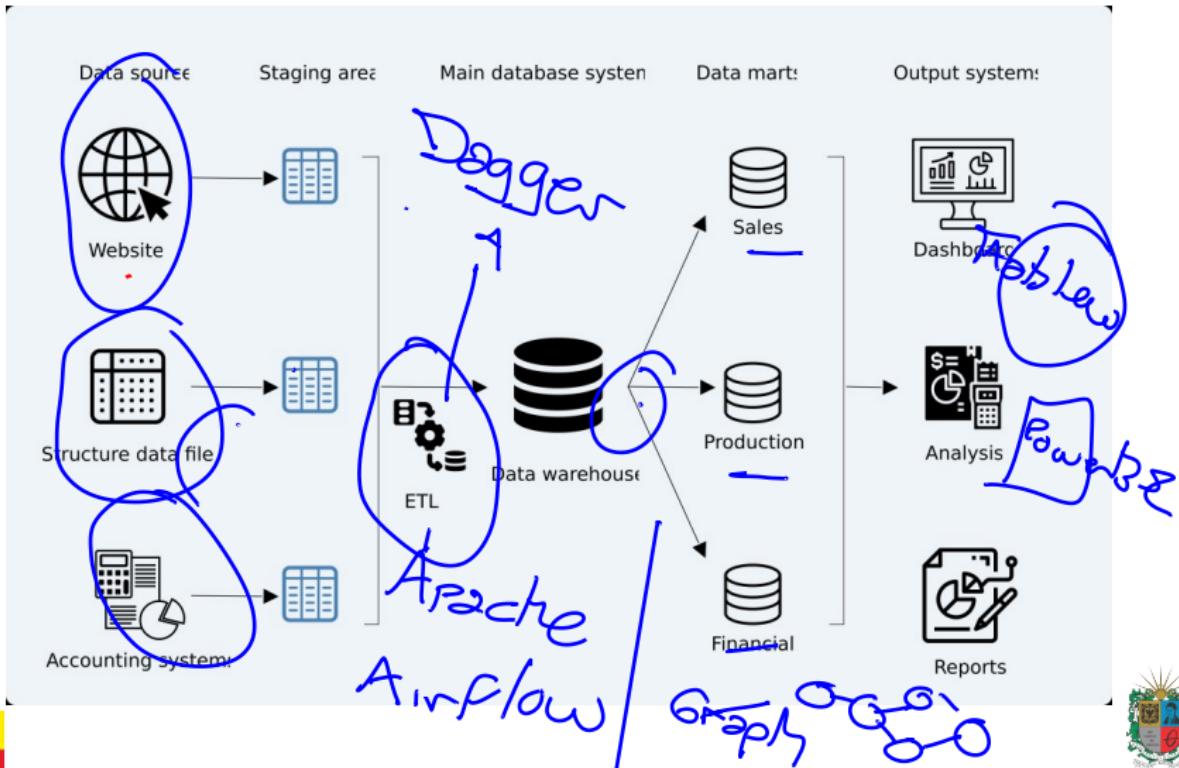
Data Definition Framework



text



Data Pipelines



Outline

- 1 Data Science Basic Concepts

→ Pipeline

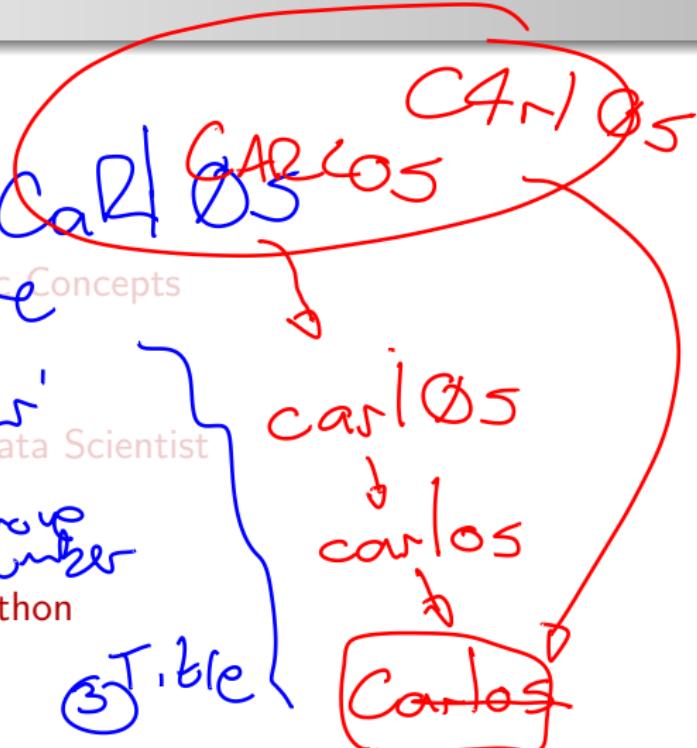
- 2 What is to be a Data Scientist

① lower'

② remove number

- 3 Foundations of Python

③ Table



Basic Definitions in Python

Guido van Rossum

- **Python** is a high-level, interpreted, and general-purpose programming language.
 - **High-level** means that Python is designed to be easy to read and write.
 - **Interpreted** means that Python code is executed line by line, rather than being compiled into machine code.
 - **Weakly typed** means that Python does not require you to declare the type of a variable.
 - **Multiparadigm** means that Python supports object-oriented, imperative, and functional programming styles.
 - **Snake-case** is the convention of writing variable names in lowercase, with underscores between words.
- Handwritten notes:
- **High-level**: web site, databases, AI / ML
 - **Interpreted**: IoT / MicroPython



Basic Definitions in Python

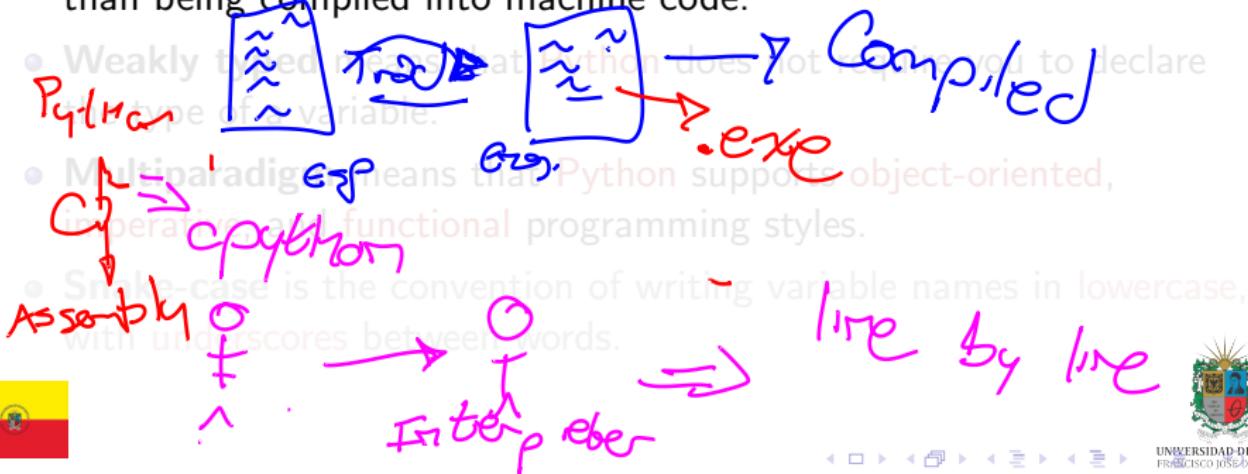
- **Python** is a **high-level**, **interpreted**, and **general-purpose** programming language.
 - **High-level** means that **Python** is designed to be easy to read and write.
 - **Interpreted** means that **Python** code is executed line by line, rather than being compiled into machine code.
 - **Weakly typed** means that **Python** does not require you to declare the type of a variable.
 - **Multiparadigm** means that **Python** supports object-oriented, imperative, and functional programming styles.
 - **Snake-case** is the convention of writing variable names in lowercase, with underscores between words.
- math
notation
natural
language*



Basic Definitions in Python

• `.py` \Rightarrow • `.pyc`

- **Python** is a **high-level**, **interpreted**, and **general-purpose** programming language.
- **High-level** means that **Python** is designed to be easy to read and write.
- **Interpreted** means that **Python** code is executed line by line, rather than being compiled into machine code.



Basic Definitions in Python

- **Python** is a **high-level**, **interpreted**, and **general-purpose** programming language.
- **High-level** means that **Python** is designed to be easy to read and write.
- **Interpreted** means that **Python** code is executed line by line, rather than being compiled into machine code.
- **Weakly typed** means that **Python** does not require you to declare the type of a variable.

- Multilingual means that **Python** supports object-oriented, imperative, and functional programming styles.
- Snake-case is the convention of writing variable names in lowercase, with underscores between words.



Basic Definitions in Python

- **Python** is a **high-level**, **interpreted**, and **general-purpose** programming language.
- **High-level** means that **Python** is designed to be easy to read and write.
- **Interpreted** means that **Python** code is executed line by line, rather than being compiled into machine code.
- **Weakly typed** means that **Python** does not require you to declare the type of a variable.
- **Multiparadigm** means that **Python** supports **object-oriented**, **imperative**, and **functional** programming styles.
- **Snake-case** is the convention of writing variable names in **lowercase**, with **underscores** between words.



Basic Definitions in Python

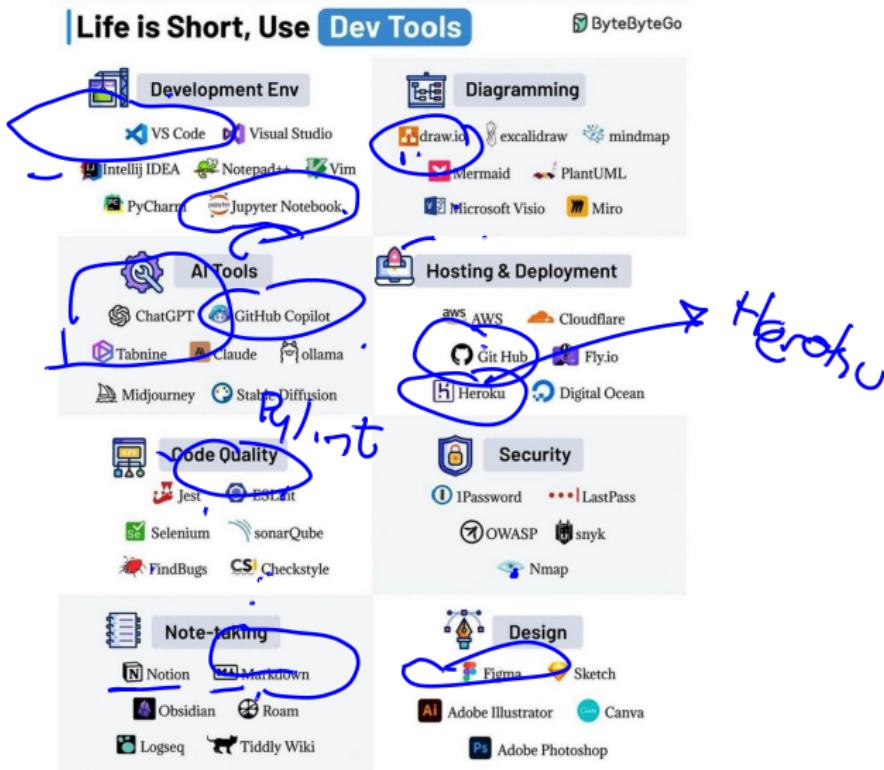
~~Case Case \Rightarrow my Variable~~

~~Snake Case
my variable~~

- **Python** is a **high-level**, **interpreted**, and **general-purpose** programming language.
- **High-level** means that **Python** is designed to be easy to read and write.
- **Interpreted** means that **Python** code is executed line by line, rather than being compiled into machine code.
- **Weakly typed** means that **Python** does not require you to declare the type of a variable.
- **Multiparadigm** means that **Python** supports **object-oriented**, **imperative**, and **functional** programming styles.
- **Snake-case** is the convention of writing variable names in **lowercase**, with **underscores** between words.



Popular Developer Tools



Virtual Environments

- **Virtual environments** are a way to create isolated spaces on your computer for **Python projects**.
 - Virtual environments allow you to install packages and dependencies for a specific project without affecting other projects.
 - Virtual environments are created using the `venv` module, which is included in the Python standard library.
 - Virtual environments are activated using the `source` command in the terminal.
 - Virtual environments are deactivated using the `deactivate` command in the terminal.
- P.2* *P3.J* *P.3z* *Numpy 0.3* *Numpy 1.1*
Numpy d. 2



Virtual Environments

- **Virtual environments** are a way to create isolated spaces on your computer for **Python projects**.
- **Virtual environments** allow you to **install packages** and **dependencies** for a specific project **without affecting other projects**.
- **Virtual environments** are created using the **venv** module, which is included in the **Python standard library**.
- **Virtual environments** are activated using the **source** command in the terminal.
- **Virtual environments** are deactivated using the **deactivate** command in the terminal.



Virtual Environments

- **Virtual environments** are a way to create isolated spaces on your computer for **Python projects**.
- **Virtual environments** allow you to **install packages** and **dependencies** for a specific project without affecting other projects.
- **Virtual environments** are created using **the venv module**, which is included in the Python standard library.
- **Virtual environments** are activated using the **source command** in the terminal.
- **Virtual environments** are deactivated using the **deactivate command** in the terminal.



Virtual Environments

- **Virtual environments** are a way to create isolated spaces on your computer for **Python projects**.
- **Virtual environments** allow you to **install packages** and **dependencies** for a specific project without affecting other projects.
- **Virtual environments** are created using the **venv** module, which is included in the Python standard library.
- **Virtual environments** are activated using the **source** command in the terminal.
.bin\Scripts\activate
- **Virtual environments** are deactivated using the **deactivate** command in the terminal.

source ~



Virtual Environments

- **Virtual environments** are a way to create isolated spaces on your computer for **Python projects**.
- **Virtual environments** allow you to **install packages** and **dependencies** for a specific project without affecting other projects.
- **Virtual environments** are created using the **venv** module, which is included in the Python standard library.
- **Virtual environments** are activated using the **source** command in the terminal.
- **Virtual environments** are deactivated using the **deactivate** command in the terminal.



Modules and Packages

- **Modules** are **Python files** that contain **functions, classes, and variables.**
- Modules are used to organize code and make it reusable.
- Packages are **directories** that contain **Python files** (modules).
- Packages are used to organize modules and make them reusable.



Modules and Packages

- **Modules** are Python files that contain functions, classes, and variables.
- **Modules** are used to organize code and make it reusable.
- Packages are directories that contain Python files (modules).
- Packages are used to organize modules and make them reusable.



Modules and Packages

Library → Package

- **Modules** are Python files that contain functions, classes, and variables.
- **Modules** are used to organize code and make it reusable.
- **Packages** are directories that contain Python files (modules).
- Packages are used to organize modules and make them reusable.



Modules and Packages

- **Modules** are Python files that contain functions, classes, and variables.
- **Modules** are used to organize code and make it reusable.
- **Packages** are directories that contain Python files (modules).
- **Packages** are used to organize modules and make them reusable.



Dependencies Management

- **Dependencies** are external libraries or packages that your Python project relies on.
- Dependencies are traditionally managed using a requirements.txt file, which lists the names and versions of the packages required by your project.
Con el fin de
- Dependencies can be installed using the pip package manager, which is included with Python.
- Currently, the poetry package manager is recommended for managing dependencies in Python projects. It follows the PEP 508 standard for specifying dependencies.



Dependencies Management

- **Dependencies** are external libraries or packages that your Python project relies on.
- **Dependencies** are traditionally managed using a ~~requirements.txt~~ file, which lists the names and versions of the packages required by your project.
- Dependencies can be installed using the pip package manager, which is included with Python.
- Currently, the poetry package manager is recommended for managing dependencies in Python projects. It follows the PEP 508 standard for specifying dependencies.



Dependencies Management

- **Dependencies** are external libraries or packages that your **Python project** relies on.
- **Dependencies** are traditionally managed using a **requirements.txt** file, which lists the **names** and **versions** of the packages required by your project.
- **Dependencies** can be installed using the **pip** package manager, which is included with Python.
- Currently, the **poetry** package manager is recommended for managing dependencies in Python projects. It follows the **PEP 508** standard for specifying dependencies.



Dependencies Management

- **Dependencies** are external libraries or packages that your **Python project** relies on.
- **Dependencies** are traditionally managed using a **requirements.txt** file, which lists the **names** and **versions** of the packages required by your project.
- **Dependencies** can be installed using the **pip** package manager, which is included with Python.
- Currently, the **poetry** package manager is recommended for managing dependencies in Python projects. It follows the **PEP 508** standard for specifying dependencies.



Jupyter Notebooks

- Jupyter Notebooks are a **web-based interactive computing environment** that allows you to create and share documents that contain **live code, equations, visualizations, and narrative text.**
- Jupyter Notebooks are used for data cleaning, data transformation, statistical modeling, data visualization, machine learning, and more.
- Jupyter Notebooks support multiple programming languages, including Python and Julia.

Handwritten notes:

Teradata R Python Python iPython

Julia Python

Python

Python



Jupyter Notebooks

- **Jupyter Notebooks** are a **web-based interactive computing environment** that allows you to create and share documents that contain **live code, equations, visualizations, and narrative text**.
- **Jupyter Notebooks** are used for **data cleaning, data transformation, statistical modeling, data visualization, machine learning, and more.**
- Jupyter Notebooks support multiple programming languages, including Python, R, and Julia.
- Jupyter Notebooks are used by data scientists, researchers, students, and educators to explore data, create reports, and teach programming.



Jupyter Notebooks

- **Jupyter Notebooks** are a **web-based interactive computing environment** that allows you to create and share documents that contain **live code, equations, visualizations, and narrative text**.
- **Jupyter Notebooks** are used for **data cleaning, data transformation, statistical modeling, data visualization, machine learning, and more**.
- **Jupyter Notebooks** support **multiple programming languages**, including **Python, R, and Julia**.
- **Jupyter Notebooks** are used by **data scientists, researchers, students, and educators** to **explore data, create reports, and teach programming**.



Jupyter Notebooks

- **Jupyter Notebooks** are a **web-based interactive computing environment** that allows you to create and share documents that contain **live code, equations, visualizations, and narrative text**.
- **Jupyter Notebooks** are used for **data cleaning, data transformation, statistical modeling, data visualization, machine learning, and more**.
- **Jupyter Notebooks** support **multiple programming languages**, including **Python, R, and Julia**.
- **Jupyter Notebooks** are used by **data scientists, researchers, students, and educators** to **explore data, create reports, and teach programming**.

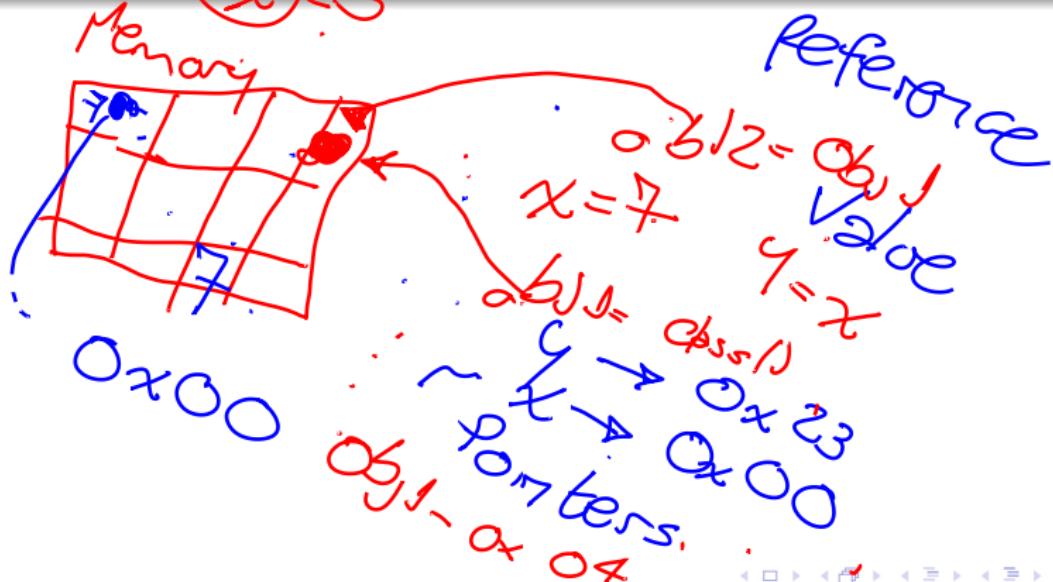
text \Rightarrow Markdown
code \Rightarrow Python



Variables and Memory Management

Variables Definition

Variables are used to store data values. Python has no command for declaring a variable. A variable is created the moment you first assign a value to it.



Conditionals

Definition

Conditionals are used to execute different code blocks based on *different conditions*.

Nested Conditionals

Nested conditionals are conditionals that are *inside* other conditionals.

Elif Conditionals

Elif conditionals are used to check *multiple conditions*.

block ->

`=`
`!=`

9MD

&

or

`<`
`<=`

`>`

`>=`

`else`

`not`



Conditionals

Definition

Conditionals are used to execute different code blocks based on *different conditions*.

Nested Conditionals

Nested conditionals are conditionals that are *inside* other conditionals.

If (new) {
Elif Conditional

Elif conditionals are used to check multiple conditions.

if ?:
... x=7

... y=5
if ?:
... z=3

else:
... x=z



Conditionals

Definition

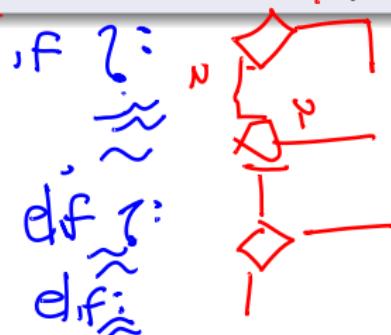
Conditionals are used to execute different code blocks based on *different conditions*.

Nested Conditionals

Nested conditionals are conditionals that are *inside* other conditionals.

Elif Conditionals

Elif conditionals are used to check *multiple conditions*.



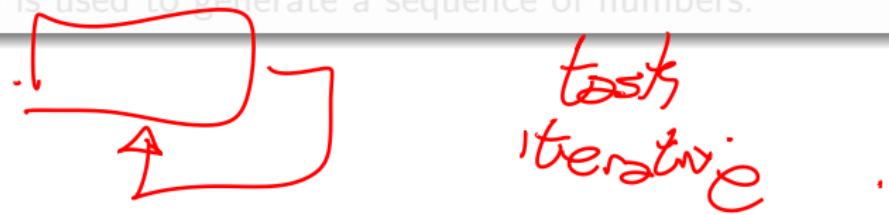
Loops and Range

Loops Definition

Loops are used to execute a block of code *multiple times*.

Range Definition

The range function is used to generate a sequence of numbers.



while => otherwise

for => set / sequence



Loops and Range

Loops Definition

Loops are used to execute a block of code *multiple times*.

Range Definition

The **range function** is used to generate a sequence of numbers.

2, 4, 6, 8
3, 5, 7, 9
5, 10, 15, 20

range (0
start, stop, *
inc / exc)

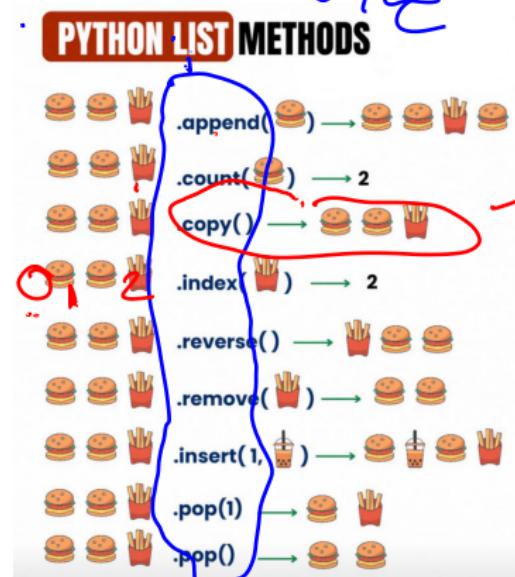


Lists

Definition

A **list** is a collection of items that are *ordered* and *changeable*. **Lists** are defined by enclosing the items in square brackets.

[]
list



Dictionaries

Definition

A **dictionary** is a collection of items that are *unordered*, *changeable*, and *indexed*. Dictionaries are defined by enclosing the items in curly braces.

```
{ "a": 1,  
  "b": 2,  
  "zed": 26  
}
```

pair key value
Java Hash map
(JS Object)
JSON



Sets and Tuples

Definition Sets

A **set** is a collection of items that are *unordered* and *unindexed*. **Sets** are defined by enclosing the items in curly braces.

$A = \{a, b, c\}$

$B = \{d, e, f\}$

A tuple is a collection of items that are *ordered* and *unchangeable*.

Tuples are defined by enclosing the items in parentheses.



Not repeated elements



Sets and Tuples

Definition Sets

A **set** is a collection of items that are *unordered* and *unindexed*. **Sets** are defined by enclosing the items in curly braces.

Definition Tuples

A **tuple** is a collection of items that are *ordered* and *unchangeable*.
Tuples are defined by enclosing the items in parentheses.

(*a*, *b*, *c*, *d*, *e*)



Lists Comprehensions

Definition

List comprehensions provide a concise way to create lists. Common applications are to make *new lists* where each element is the result of some operation applied to each member of another sequence or iterable, or to create a *subsequence* of those elements that satisfy a certain condition.

- $[\text{process } i) \text{ for } i \in \text{collection}]$
- $[\text{process } i) \text{ for } i \in \text{collection}$
 $\quad \text{if } i > 2]$



Functions

Definition

A **function** is a block of code that only runs when it is called. You can pass data, known as **parameters**, into a function. A **function** can return data as a result.

Type of Functions

- Built-in Functions
- User-defined Functions
 - Variadic Functions
 - Recursive Functions

def mane (**par1**, **par2**):
 ...
 ... **x = par1 + par2**
 ... **return**



Functions

Definition

A **function** is a block of code that only runs when it is called. You can pass data, known as parameters, into a function. A **function** can return data as a result.

Type of Functions

- **Built-in Functions**
- **User-defined Functions**
 - Variadic Functions
 - Recursive Functions

Python
list
, remove()
, sorted()

variable param
↓ less vs more



Python Iterators

Definition

An **iterator** is an object that contains a *countable number of values*. An **iterator** is an object that can be iterated upon, meaning that you can traverse through all the values.

Maps

The **map function** is used to apply a function to *all the items* in an input list.

Filters

The **filter function** is used to *select items* from an input list that meet a certain condition.



Python Iterators

Definition

An **iterator** is an object that contains a *countable number of values*. An **iterable** is an object that can be iterated upon, meaning that you can traverse through all the values.

Maps

The **map function** is used to apply a function to *all the items* in an input list.

Filters

The **filter function** is used to select items from an input list that meet a certain condition.



Python Iterators

Definition

An **iterator** is an object that contains a *countable number of values*. An **iterable** is an object that can be iterated upon, meaning that you can traverse through all the values.

Maps

The **map function** is used to apply a function to *all the items* in an input list.

Filters

The **filter function** is used to *select items* from an input list that meet a certain condition.



Lambda Functions

Definition

A **lambda function** is a small anonymous function. A **lambda function** can take any **number of arguments**, but can only **have one expression**.

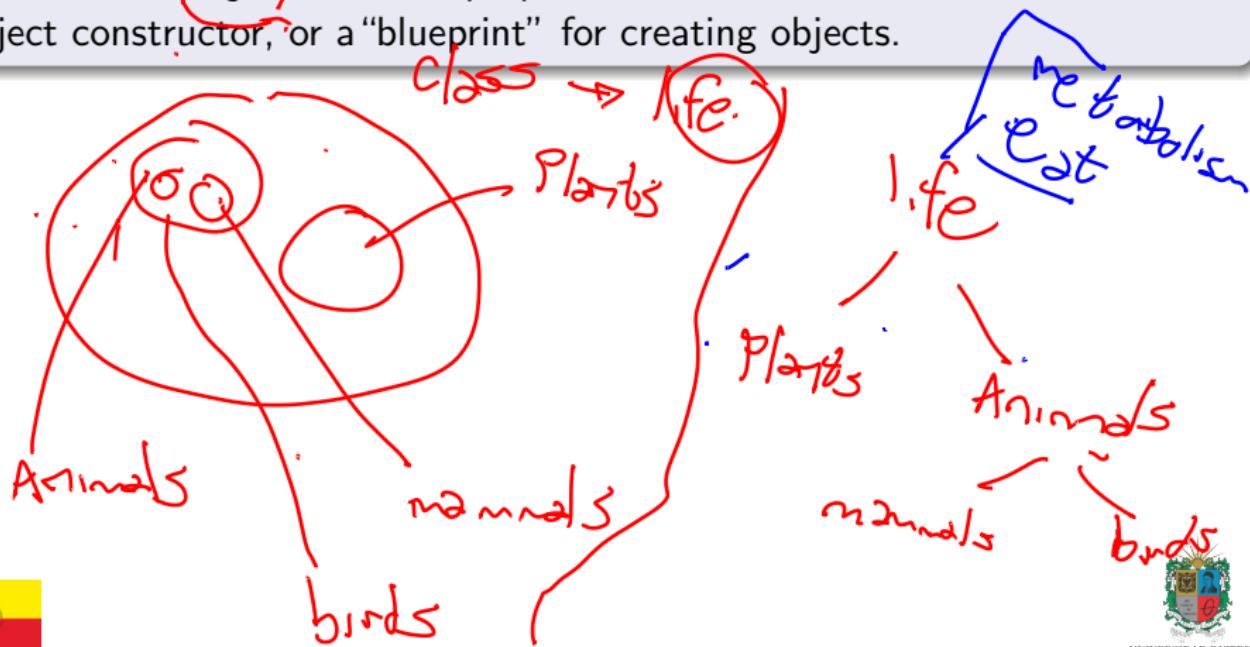
lambda par1, par2: expresion



Classes and Objects

Definition

Python is an *object-oriented programming* language. Almost everything in Python is an **object**, with its *properties* and *methods*. A **class** is like an object constructor, or a “blueprint” for creating objects.



Outline

- 1 Data Science Basic Concepts
- 2 What is to be a Data Scientist
- 3 Foundations of Python



Thanks!

Questions?



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

