Python Libraries

For Data Science

What is Data Science?

Data science is a field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
It employs techniques and theories drawn from many fields such as mathematics, statistics, information science, and computer science, including data mining, machine learning, clustering, statistical inference, and data visualization.

Data science is primarily used to make decisions and predictions making use of predictive analytics, and machine learning.

Structured Data vs Unstructured Data

Data is the lifeblood of business, and it comes in a huge variety of formats, everything from strictly formed relational databases to your last post on Facebook. All of that data, in all different formats, can be sorted into one of two categories: structured or unstructured data.

Structured vs. unstructured data can be understood by considering the who, what, when where, and the how of the data:

1. Who will be using the data?
2. What type of data are you collecting?
3. When does the data need to be prepared, before storage or when used?
4. Where will the data be stored?
5. How will the data be stored?

What is Structured Data?

Structured data is data that has been predefined and formatted to a set structure before being placed in data storage, which is often referred to as schema-on-write.

The best example of structured data is the **relational database**: the data has been formatted into precisely defined fields, such as credit card numbers or address, in order to be easily queried with **SQL**.

Pros of structured data:

- **Easy use by machine learning algorithms -** The specific and organized nature of structured data allows for easy manipulation and querying of that data making it the easiest data to use in machine learning.
- **Easy use by business users -** structured data can be used by an average business user with an understanding of the topic to which the data relates. There is no need to have an in-depth understanding of various different types of data or the relationships of that data.

What is Structured Data?

Pros of structured data:

- **Increased access to more tools -** Structured data also has the benefit of having been in use for far longer; historically, it was the only option. Data managers have more product choices when using structured data because there are more tools that have been tried and tested for using and analyzing structured data.

Cons of structured data:

- **A predefined purpose limits use** - While on-write-schema data definition is a large benefit to structured data, it is also true that data with a predefined structure can only be used for its intended purpose. This limits its flexibility and use cases.
- **Hard to change and less flexible -** Any change requirements means updating all of that structured data to meet the new needs.This results in massive expenditure of resources and time.

What is Unstructured Data?

Unstructured data is data stored in its native format and not processed until used, which is known as schema-on-read. It comes in a myriad of file formats, including email, social media posts, presentations, chats, IoT sensor data, and satellite imagery.

Pros of unstructured data:

- **Freedom of the native format -** Because unstructured data is stored in its native format, the data is not defined until it is needed. This leads to a larger pool of use cases, because the purpose of the data is adaptable. It allows for preparation and analysis of only the data needed. The native format also allows for a wider variety of file formats in the database, because the data that can be stored is not restricted to a specific format. That means the company has more data to draw from.
- **Faster accumulation rates -** There is no need to predefine the data, which means it can be collected quickly and easily.

What is Unstructured Data?

Pros of unstructured data:

- **Better pricing and scalability -** Unstructured data is often stored in cloud data lakes, which allow for massive storage. Cloud data lakes also allow for pay-as-you-use storage pricing, which helps cut costs and allows for easy scalability.

Cons of unstructured data:

- **Data science expertise -** The largest drawback to unstructured data is that data science expertise is required to prepare and analyze the data. A standard business user cannot use unstructured data as-is due to its undefined/non-formatted nature.
- **Specialized tools -** In addition to the required professional expertise, unstructured data requires specialized tools to manipulate. Standardized tools are intended for use with structured data, which leaves a data manager with limited choices in products for utilizing unstructured data.

Structured Data vs Unstructured Data

	Structured Data	**Unstructured Data**
Who	Self-service access	Requires data science expertise
What	Only select data types	Many varied types conglomerated
When	Schema-on-write	Schema-on-read
Where	Commonly stored in data warehouses	Commonly stored in data lakes
How	Predefined format	Native format

Data Science In Companies

As we learned, data is the most important asset that an organization has.

In most cases, the organization will want to use its own data in order to do one of 2 main things:

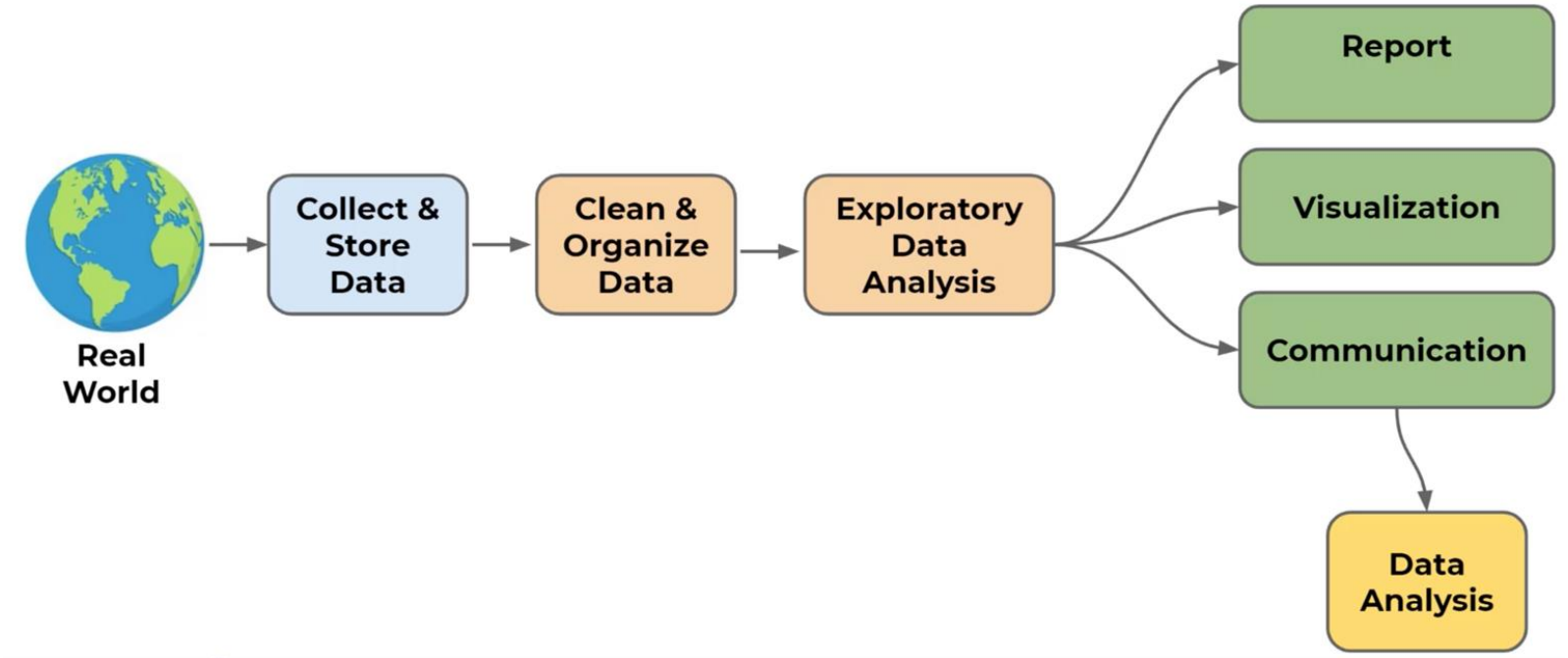
- **Answer a question in order to make a decision** → It can be a business related question, product question, management question and so on..
- **Solve a problem** → For example: provide password mechanism for customers or recommendation system.

  In most cases this type of use case will create eventually a **data product**.

In order to understand how data sciences manage to solve those problems we first need to understand the flow of data in the organization.

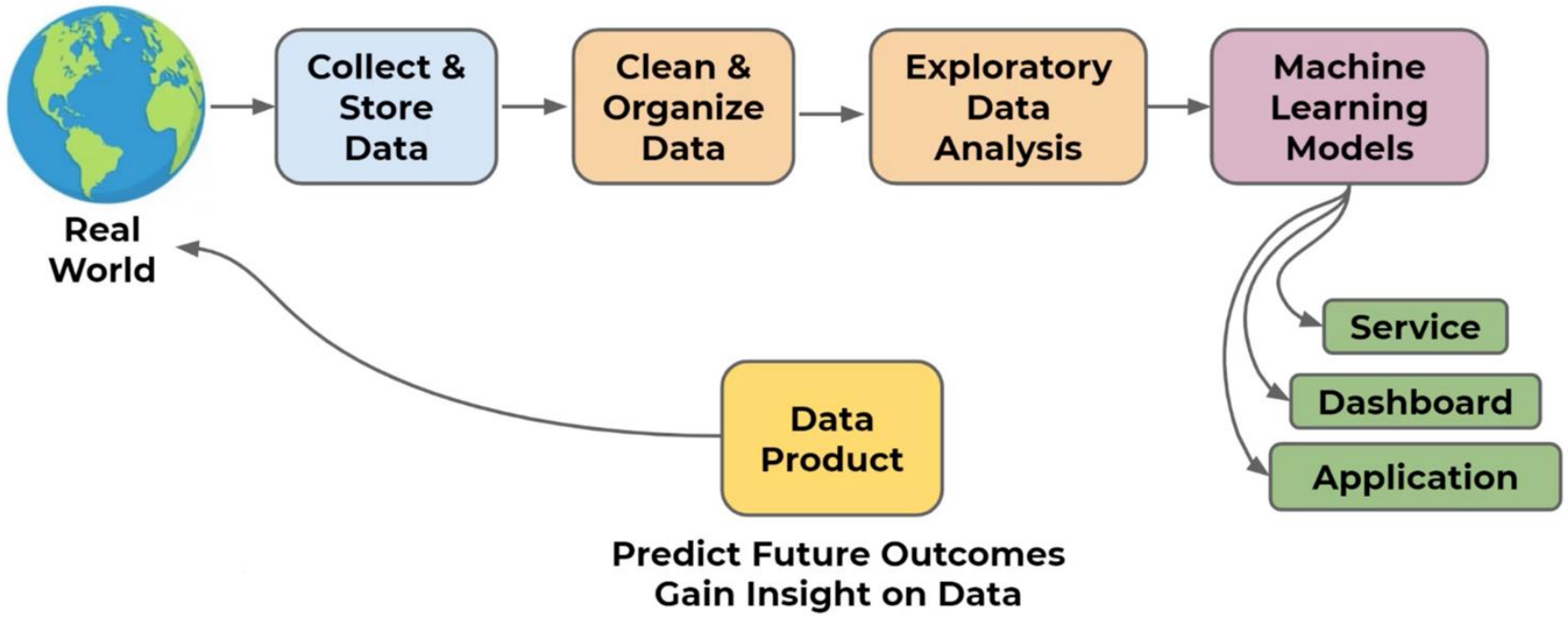
Data Science In Companies

Data flow in organizations in order to make decisions and answer key questions:



Data Science In Companies

Data flow in organizations in order to solve a problem and create data product:



Data Science In Companies

For **answering a question** by using the organization data we don't need data science or machine learning algorithms because it can be done manually just by explore the existing data and try to gather insights from it that will answer the question.

For **solving a problem** by using the organization data we will need to use machine learning algorithms and tools and for that we will need data scientists that will manage to execute it.

The reason for the different between the two flows is that for the first flow we don't need complex or automatic tool to achieve it - we just answering the question and moving to the next question.
Data products on the other hand must be automatic and should adapt to future changes in the data automatically so we don't have a choice but to use automatic tools and machine learning algorithms in order to achieve build and maintain it.

Other Related Roles In Organization

Beside The Data scientist and Data analyst roles that we already discussed their responsibility in the organization we also have other roles that we should be familiar with:

- **Product manager** → Responsible for guiding the development and execution of a product, from conception to launch.
- **Data engineer** → Responsible for the creating data flows and data pipeline in order to move and organized the data in different databases according to the organization needs. constructs, installs, tests, and maintains data management systems to help a business use data more effectively
- **Software developer** → Responsible for designing, coding, testing, debugging, and maintaining software applications according to a business's or client's needs.

Other Related Roles In Organization

- **AI developer** → Responsible for designing and implementing machine learning algorithms and artificial intelligence systems to automate and improve business operations.
  In the new world AI developer is expecting to also use Gen-AI tools and exploit their abilities in the product he develop.
- **Devops Engineer** → Works to unify software development (Dev) and operations (Ops) by automating and streamlining the integration and deployment process to allow faster and more reliable software delivery.

Nore: Not in every organization we will see every role and in some organizations specific role can be responsible for doing also the jobs of other roles.
For example: In small organizations the Data science can be also the Data engineer and the Data analyst.

What are Python Libraries?

Python libraries are pre-written codes designed to help resolve common programming tasks and challenges, reducing the amount of code developers need to write. They have a vast array functionalities, from web development and machine learning to artificial intelligence and data analysis.

By leveraging these libraries, developers can accelerate their development process, improve the readability of the code, and ensure its reusability.

Specifically, data science libraries in Python have revolutionized the way data is processed and analyzed.

For instance, **Pandas** enables high-level data manipulation and analysis, offering data structures and operations for manipulating time-series data and numerical tables.

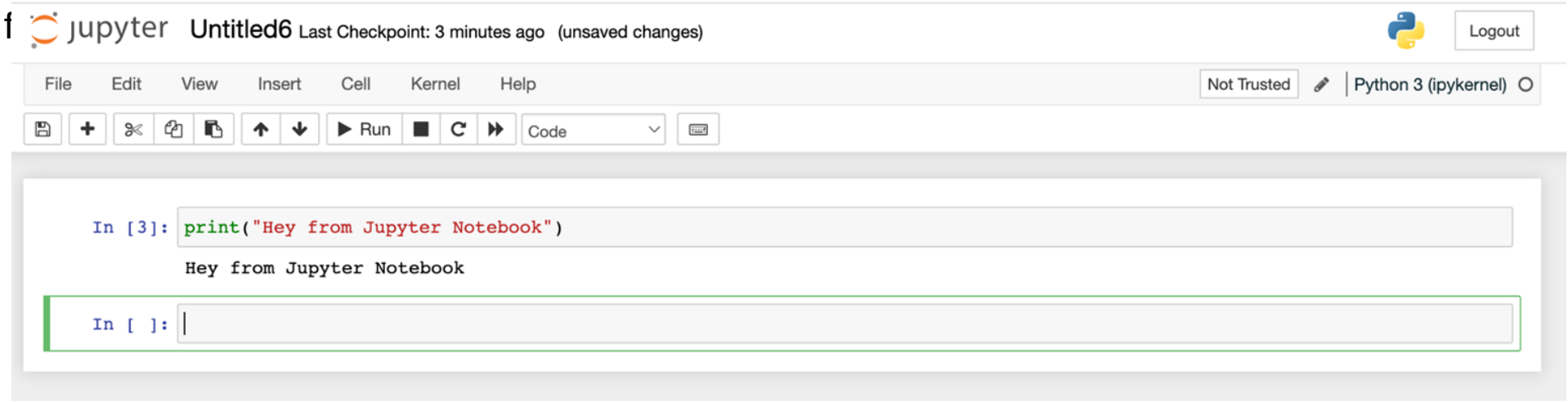
NumPy aids in complex mathematical computations and supports large multi-dimensional arrays and matrices.

Matplotlib is another critical library that enables efficient data visualization, creating static, animated,

Introduction to Jupyter Notebook

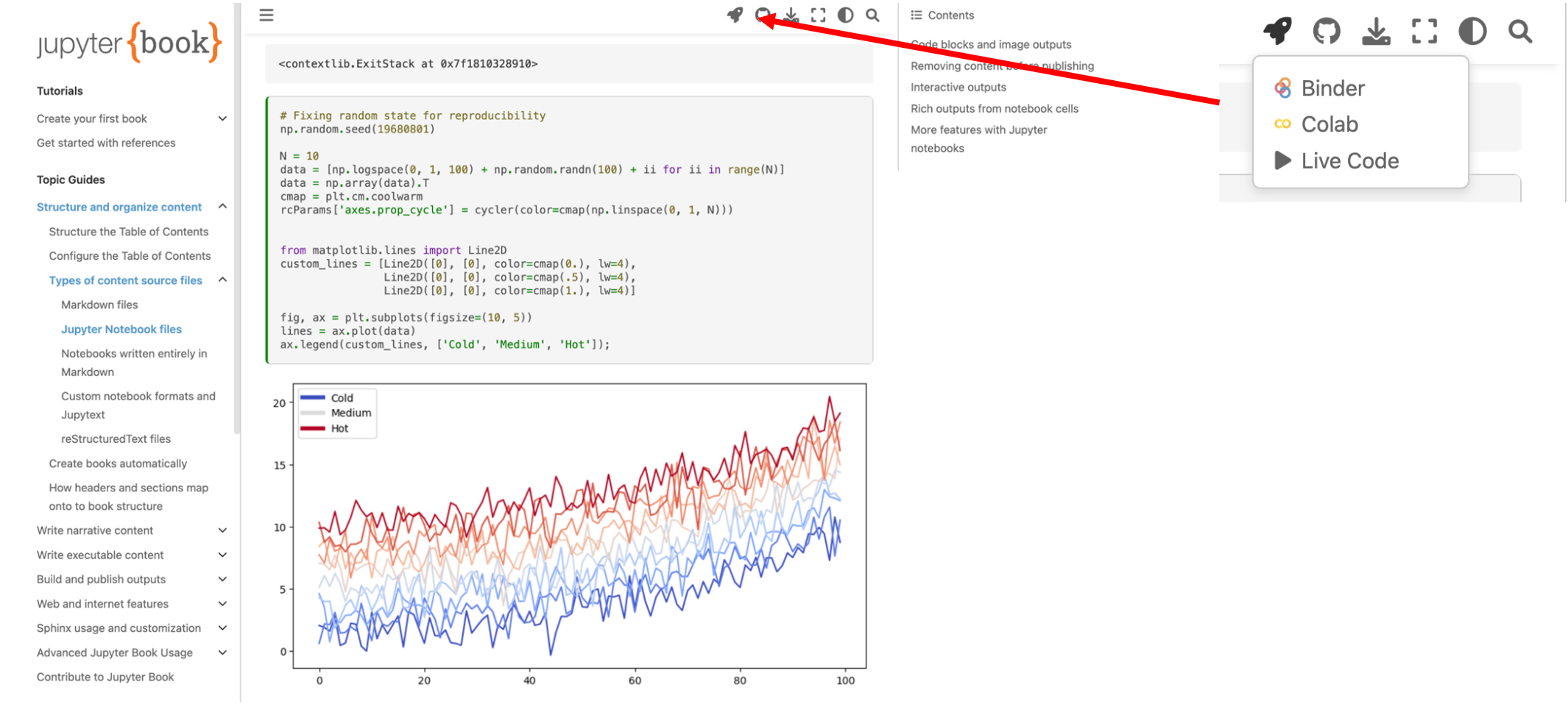
Jupyter Notebook is an open-source web application that allows the creation and sharing of document containing live code, equations, visualizations, and narrative text.

Jupyter Notebook enables us to create interactive, shareable notebooks with code snippets that can be run interactively, alongside explanations and visualizations. Due to these features, it has grown popular in



Introduction to Jupyter Notebook

Jupyter Notebook example - https://jupyterbook.org/en/stable/file-types/notebooks.html



Introduction to Anaconda

Anaconda is an open-source distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment.

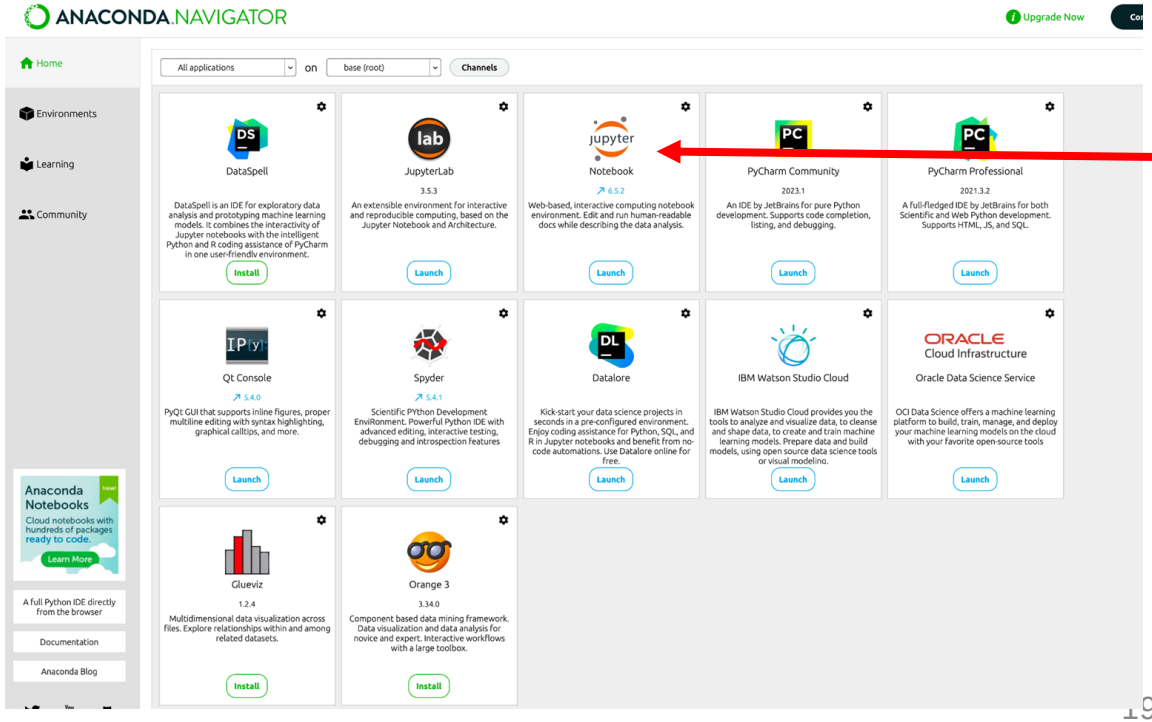
It is widely used for data science and machine learning tasks and it includes a range of tools for visualizing, cleaning, transforming and modelling data.

Jupyter Notebook is one of the many tools Anaconda provides and it’s easy to lunch and organize our Jupiter notebooks using the Anaconda distribution.

Anaconda & Jupyter Notebook Installation

In order to install Anaconda we need to follow those steps:

1. Go to Anaconda installation website - https://www.anaconda.com/download
2. Download and install the latest Anaconda version that match your OS (Windows / Mac)
3. Once finish, open Anaconda on your computer and you should see this opening screen:



Make sure you see Jupyter Notebook inside

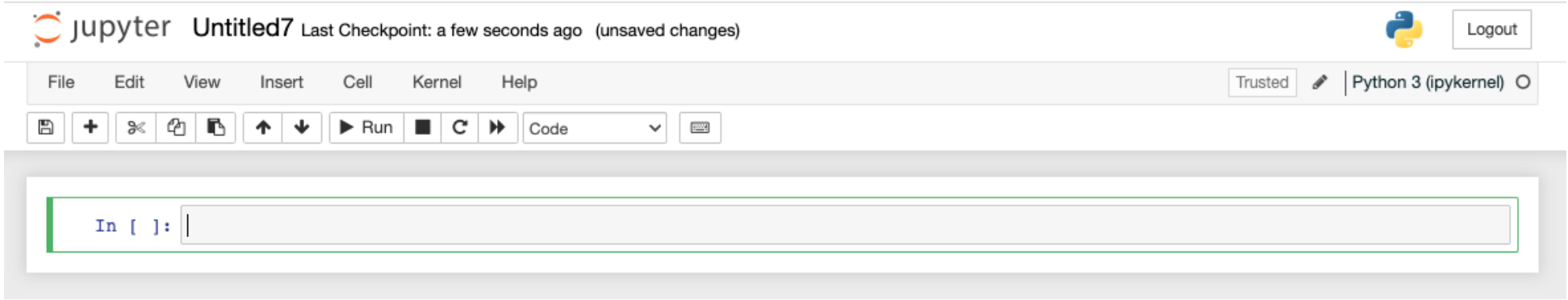
Jupyter Notebook Setup

Once we successfully install Anaconda we can now set up our Jupyter Notebook environment:

- On the Anaconda page press “Lunch” on Jupyter Notebook
- Once lunch you should see this page on your screen:

Jupyter Notebook Setup

- Go to “New” → “Python3”
- You should now see this empty Jupyter Notebook:



To make sure it’s working correctly type → print(“Hey from Jupyter Notebook”) and run the code by press on the “Run” button.