TEGRA
cybersecurity center

Data Exploration

Marta Sestelo - msestelo@gradiant.org

Rafael Martínez - rmartinez@gradiant.org

TEGRA Skillshare Plan 2018

Installation Instructions

TEGRA

cybersecurity center

Docker installation

We are going to install Docker Community Edition (CE). It is ideal for developers and small teams looking to get started with Docker and experimenting with container-based apps. You can find instructions in the following links:

- Ubuntu: https://docs.docker.com/install/linux/docker-ce/ubuntu/
- Mac OS: https://store.docker.com/editions/community/docker-ce-desktop-mac
- Windows: https://store.docker.com/editions/community/docker-ce-desktop-windows

Once it has been installed, the following step will be download and run the *Jupyter Notebook Data Science Stack* image that contains the jupyter app.

Jupyter Notebook Data Science Image

- Includes libraries for data analysis from the Julia, Python, and R communities.

- Run the official Jupyter Notebook image in your Docker container:

  ```
  docker run --rm -it -p 8888:8888 -v "$(pwd):/notebooks" jupyter/datascience-notebook
  ```

- Run the following command to open up the application:

  ```
  jupyter notebook
  ```

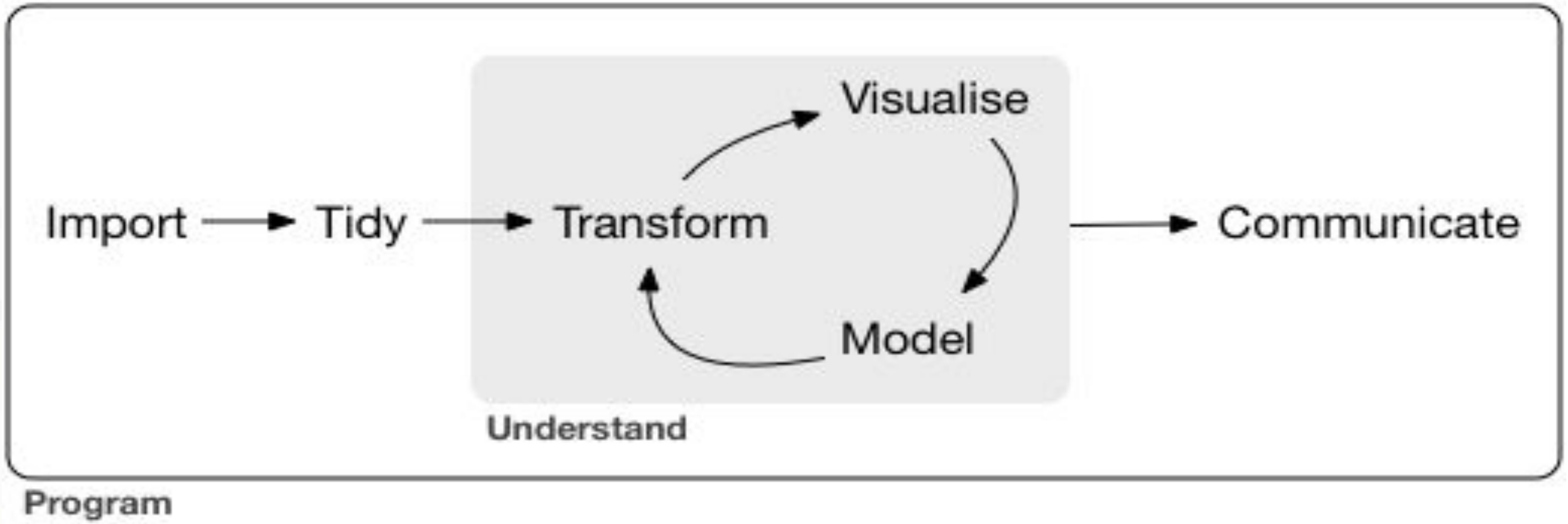
- Browse to http://localhost:8888

Note: First time, you need to copy the token from the terminal and paste it in the browser. Then generate a password and click Login.

Intro to Data Exploration

TEGRA
cybersecurity center

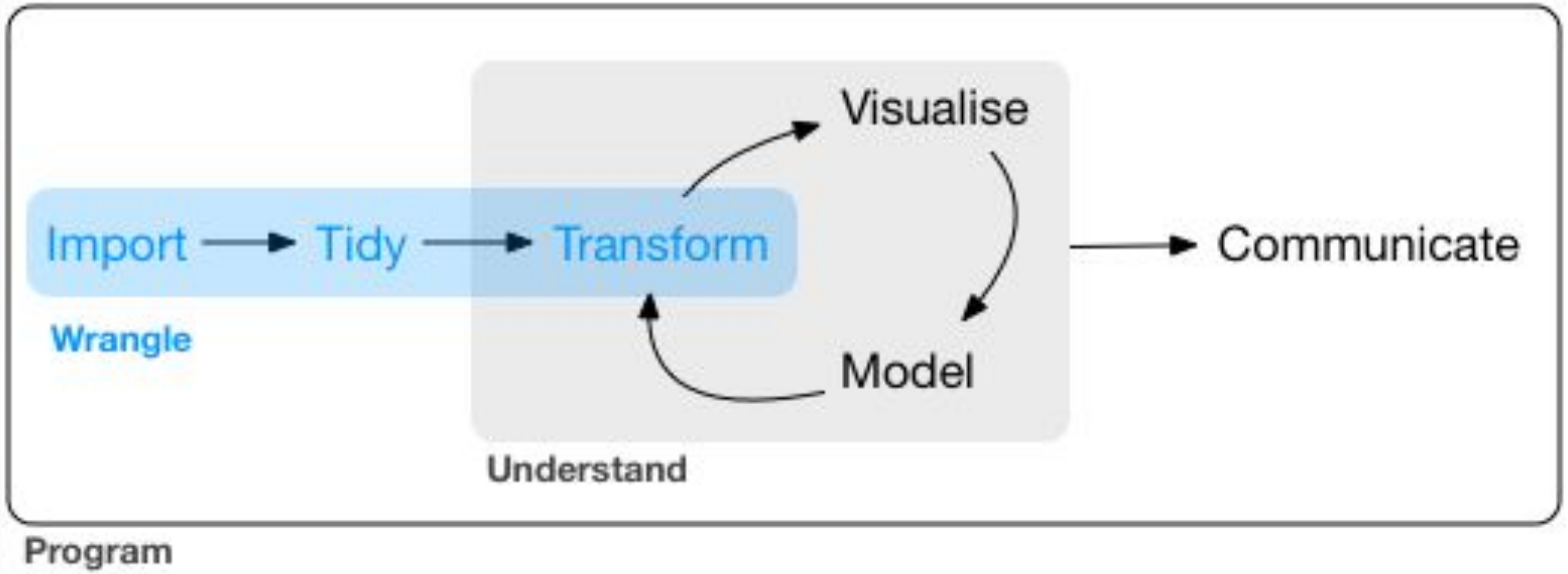
Intro to Data Analysis: Workflow



Intro to Data Exploration

Data wrangling is the art of getting your data into R or Python in a useful form for visualisation and modelling. This process can be divided in:

- data import - how to get your data from disk
- tidy data - a (consistent) way of storing your data that makes transformation, visualisation, and modelling easier



Intro to Data Exploration

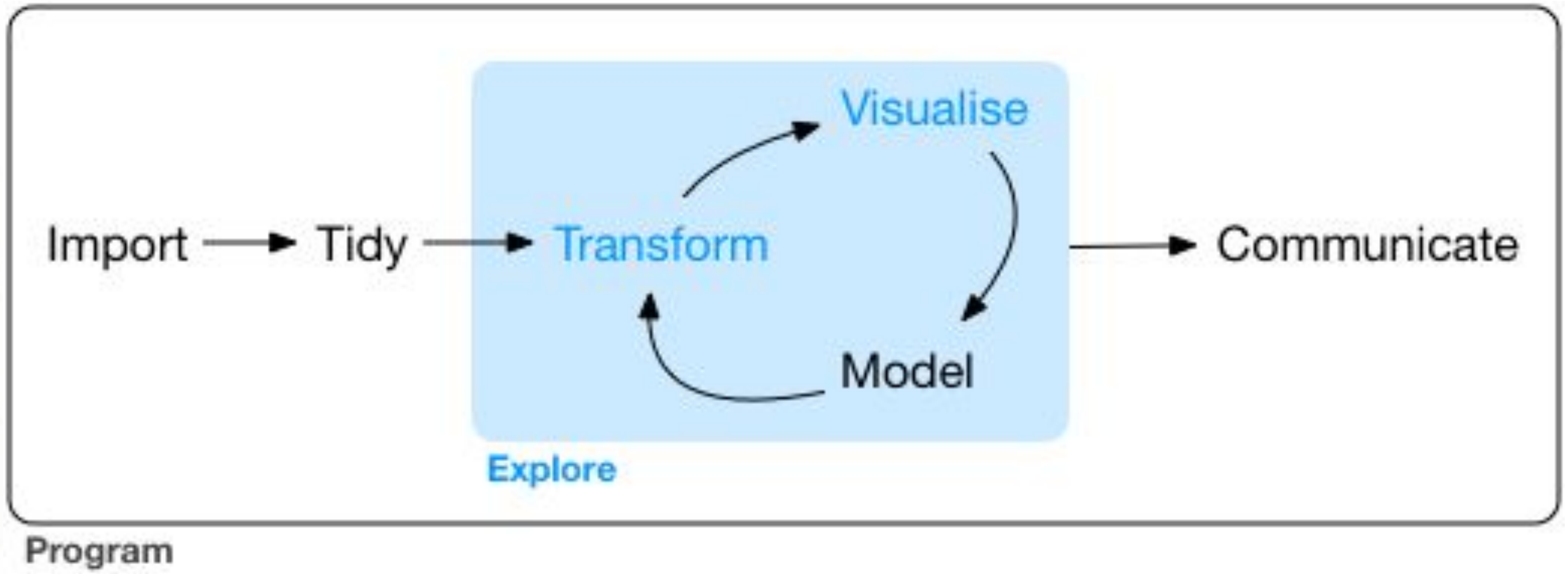
Data exploration is the art of looking at your data, rapidly generating hypotheses, quickly testing them, then repeating again and again and again. The goal of data exploration is to generate many promising ways that you can later explore in more depth.

And this art includes topics related to:

- Descriptive statistics
- Data transformation
- Data visualisation

Intro to Data Exploration

- Iterative cycle:
  - Generate questions about your data.
  - Search for answers by visualising, transforming, and modelling your data.
  - Use what you learn to refine your questions and/or generate new questions.

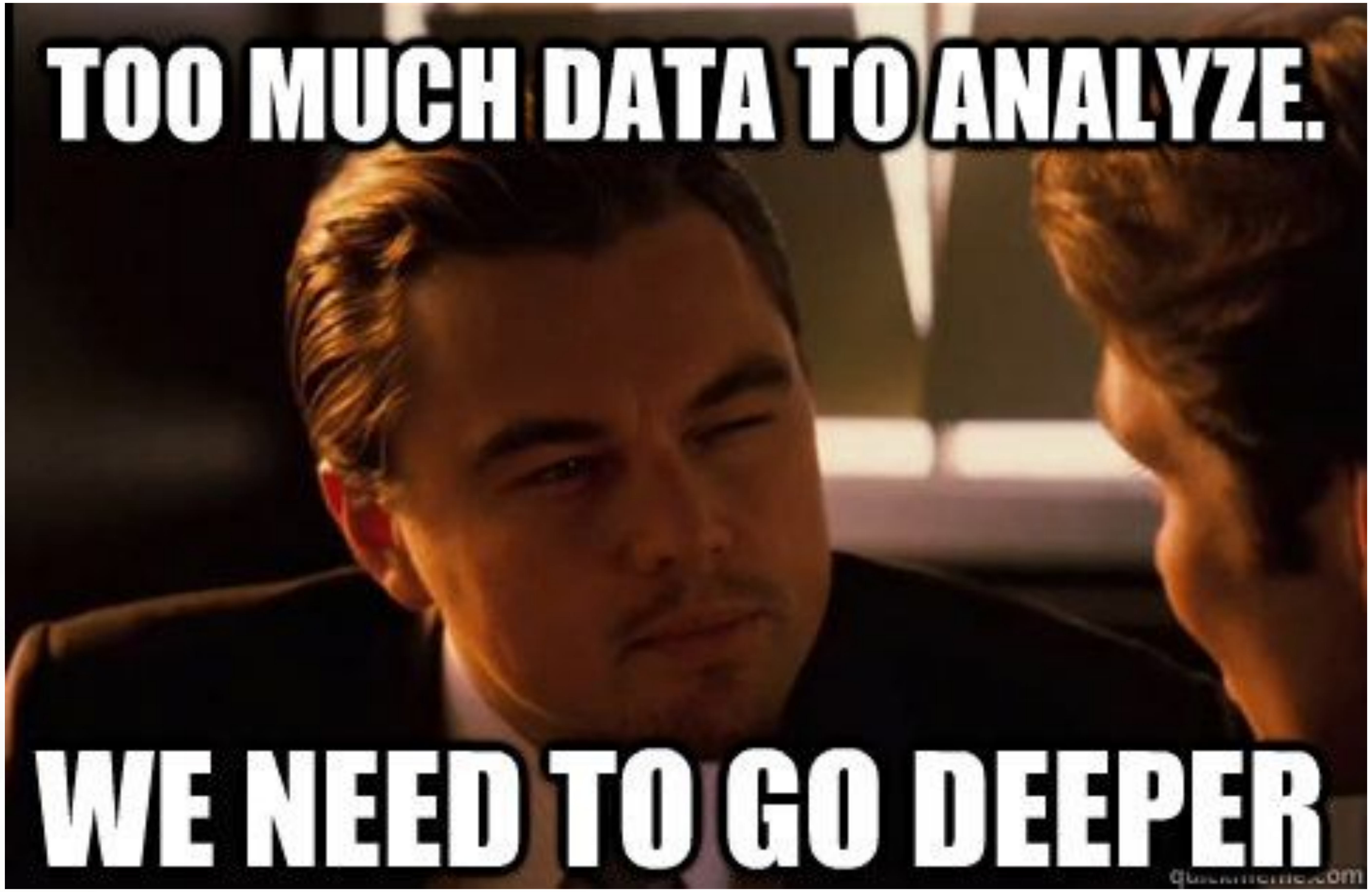


The **most important idea** is that *Exploratory Data Analysis (EDA)* is not a formal process with a strict set of rules. More than anything, EDA is a state of mind.

Jupyter Notebooks

TEGRA
cybersecurity center

Intro to Data Exploration



Notebooks for Data Science

A comprehensive comparison of Jupyter vs. Zeppelin

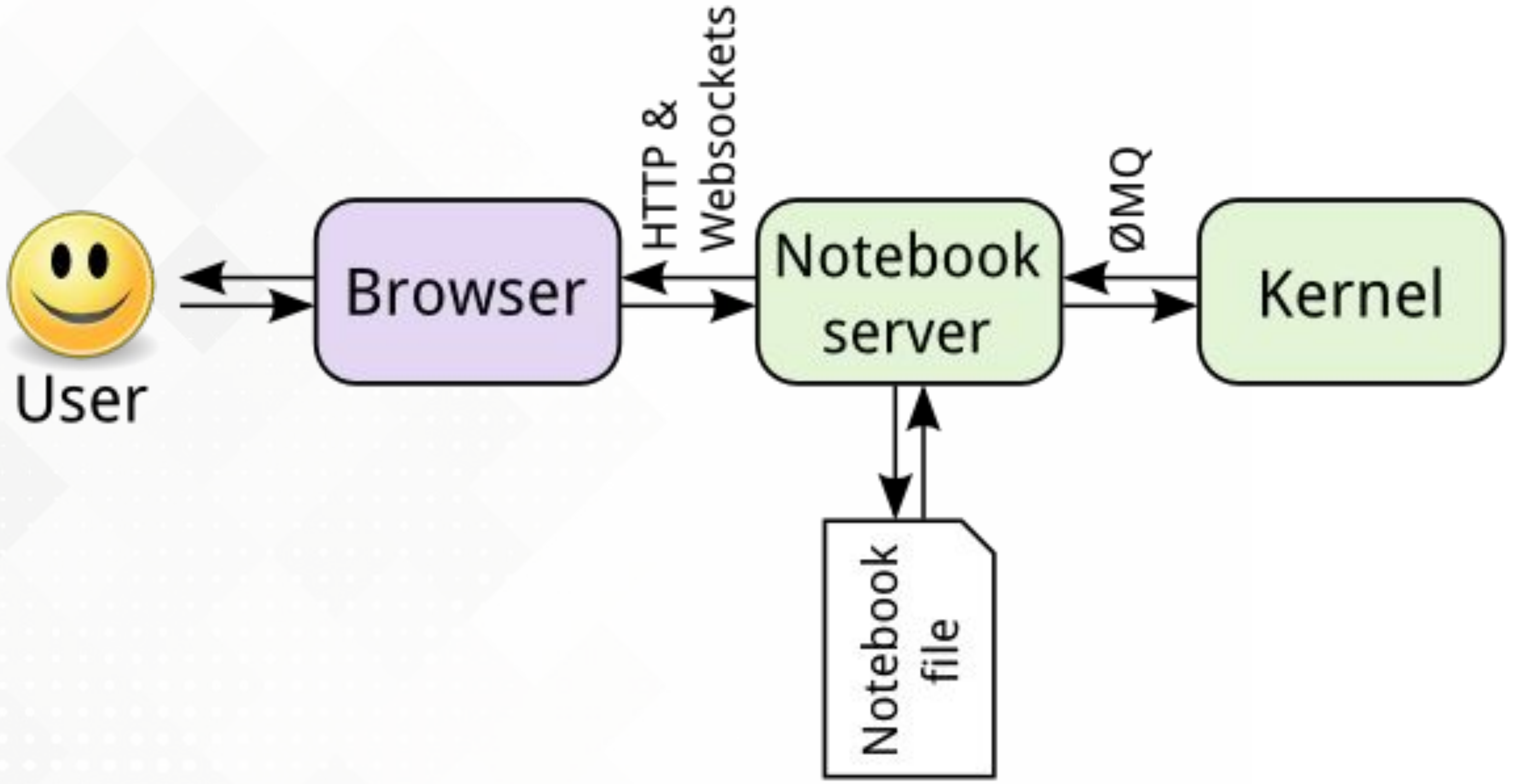
https://www.linkedin.com/pulse/comprehensive-comparison-jupyter-vs-zeppelin-hoc-q-phan-mba-/

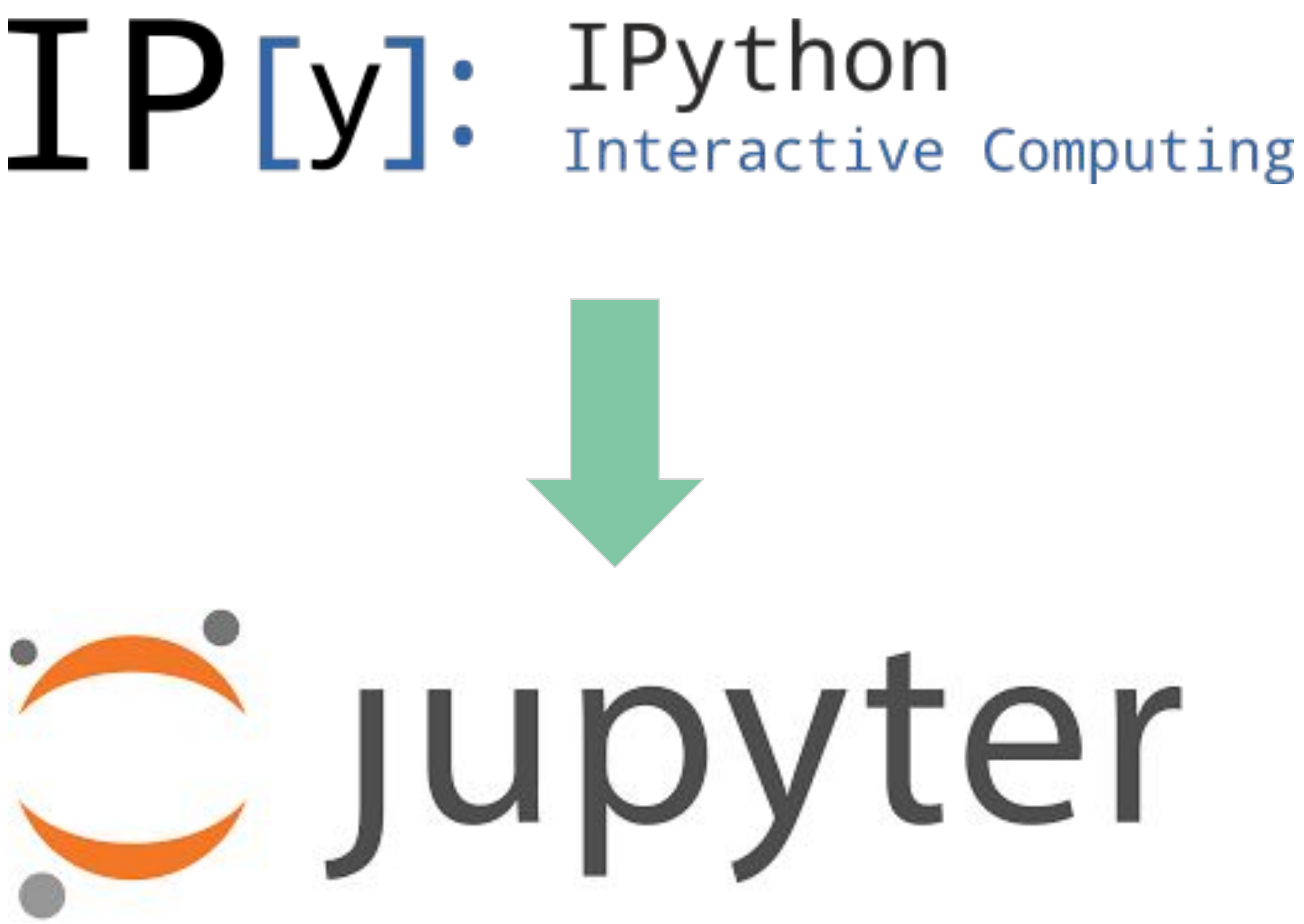
Jupyter, Zeppelin, Beaker: The Rise of the Notebooks | Open Data Science

https://medium.com/@alexisperrier/jupyter-zeppelin-beaker-the-rise-of-the-notebooks-open-data-science-1348f5eb14fe

Notebooks for Data Science

REPL = A Read–Eval–Print Loop, also known as an interactive toplevel or language shell, is a simple, interactive computer programming environment that takes single user inputs (i.e. single expressions), evaluates them, and returns the result to the user.





Notebooks for Data Science

jupyter

- Language of choice
- Share notebooks (nbviewer)
- Interactive output
- Big Data integration (Spark)

python™

R

julia

Scala

It’s time to...

… browse to https://mybinder.org/v2/gh/sestelo/tegra_skillshare/master

TEGRA

cybersecurity center

Marta Sestelo - msestelo@gradiant.org

Rafael Martínez - rmartinez@gradiant.org