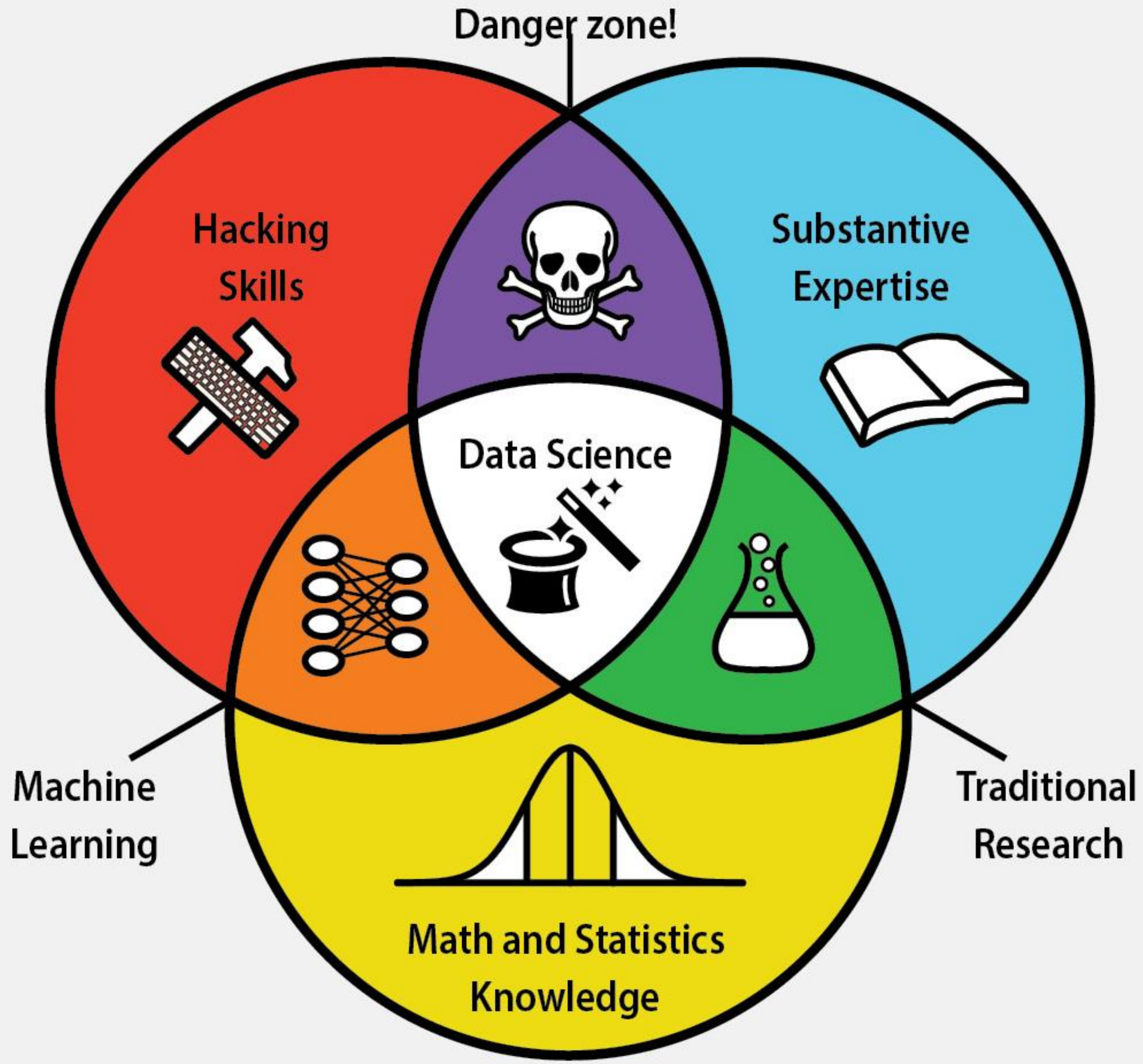
Python para Ciencia de Datos y Aprendizaje de Máquinas

DATA SCIENCE SKILLSET



Data science, due to its interdisciplinary nature, requires an intersection of abilities: **hacking skills**, **math and statistics knowledge**, and **substantive expertise** in a field of science.

Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

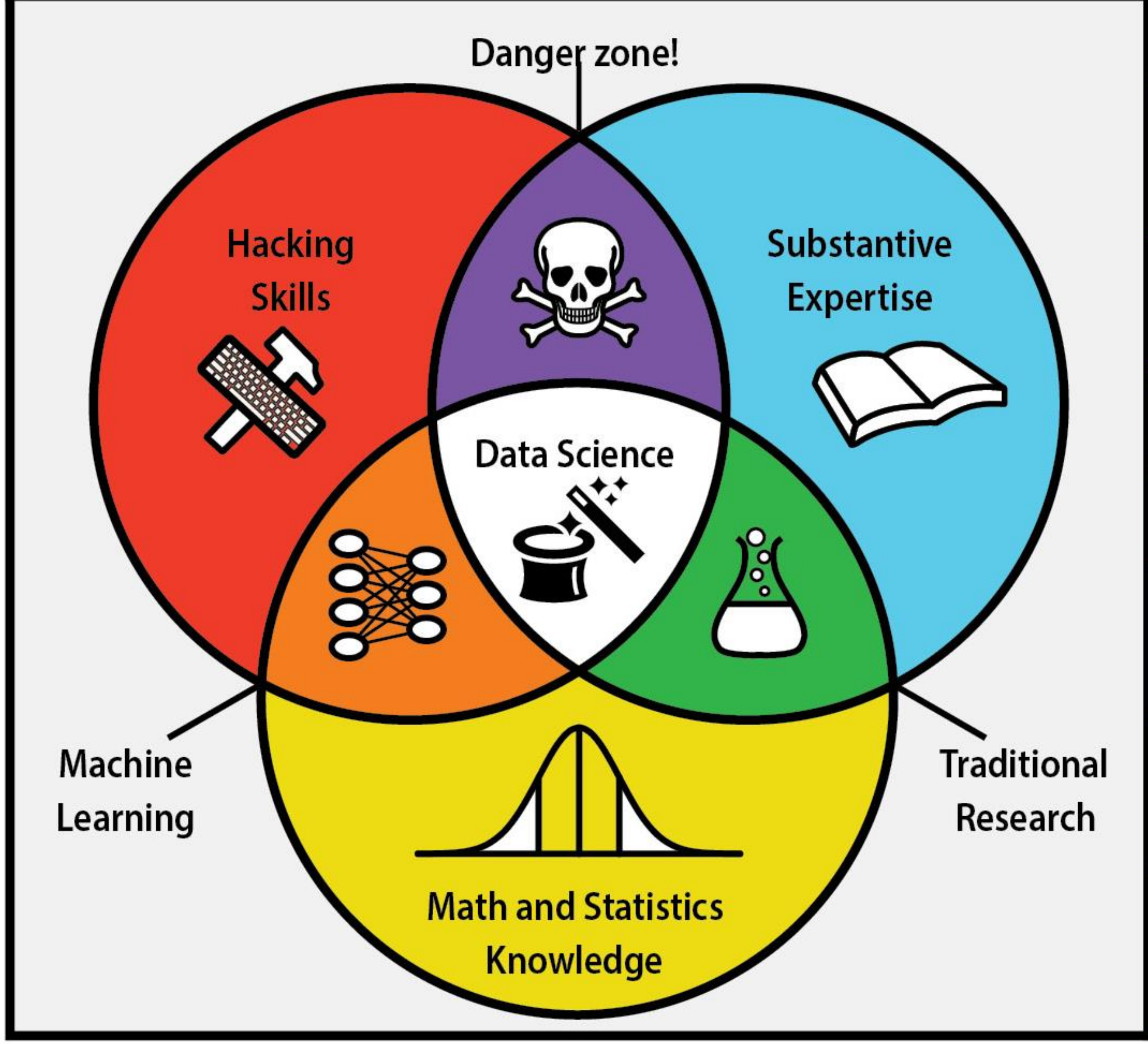
Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

CONJUNTO DE HABILIDADES EN DATA SCIENCE



La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: **habilidades de hacking**, **conocimiento de matemáticas y estadística**, y **experiencia sustantiva** en un campo de la ciencia.

Hacking skills are necessary for working with massive amounts of electronic data that must be acquired, cleaned, and manipulated.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

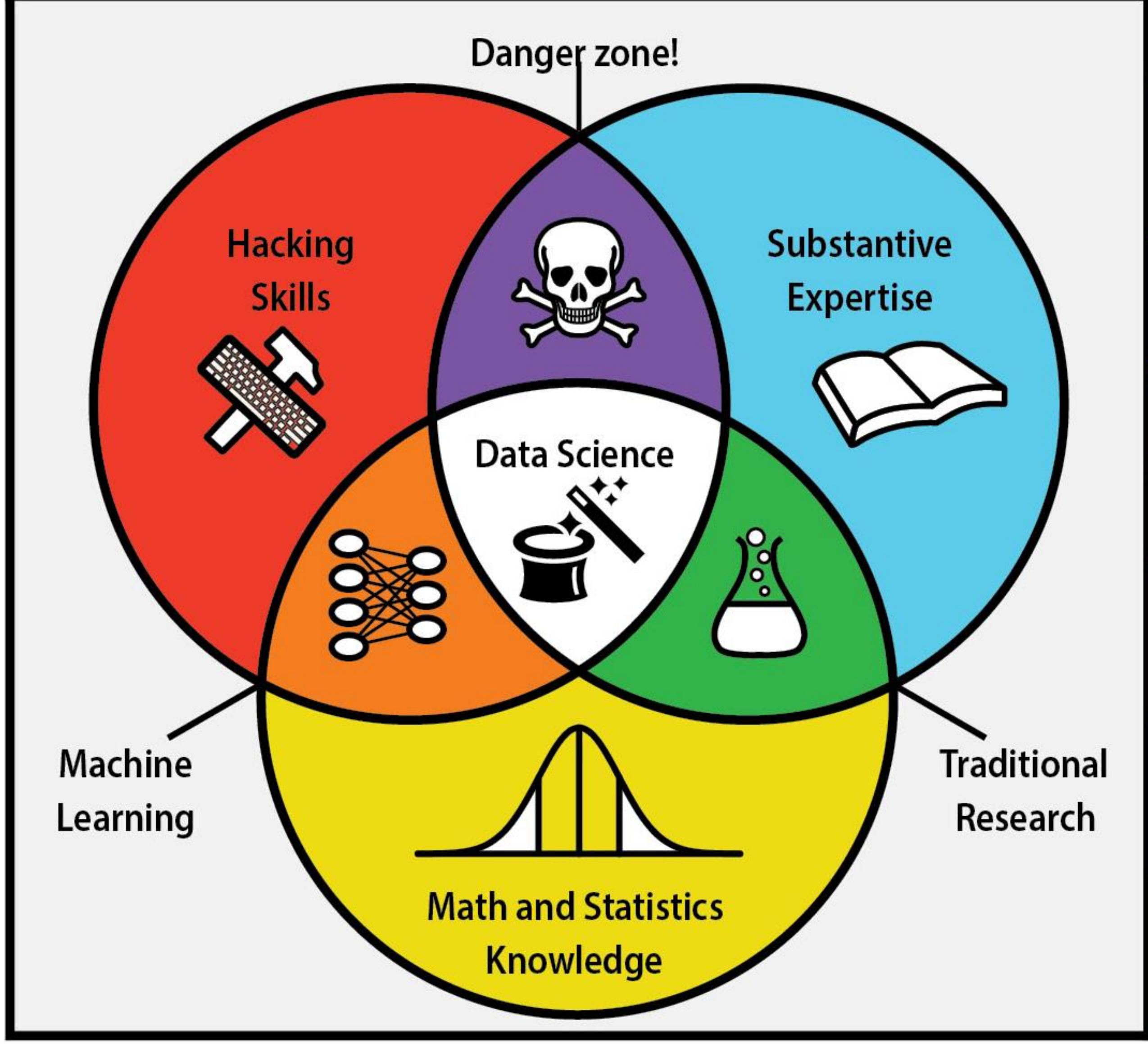
Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

CONJUNTO DE HABILIDADES EN DATA SCIENCE



La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: **habilidades de hacking**, **conocimiento de matemáticas y estadística**, y **experiencia sustantiva** en un campo de la ciencia.

Las **habilidades de hacking** son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.

Math and statistics knowledge allows a data scientist to choose appropriate methods and tools in order to extract insight from data.

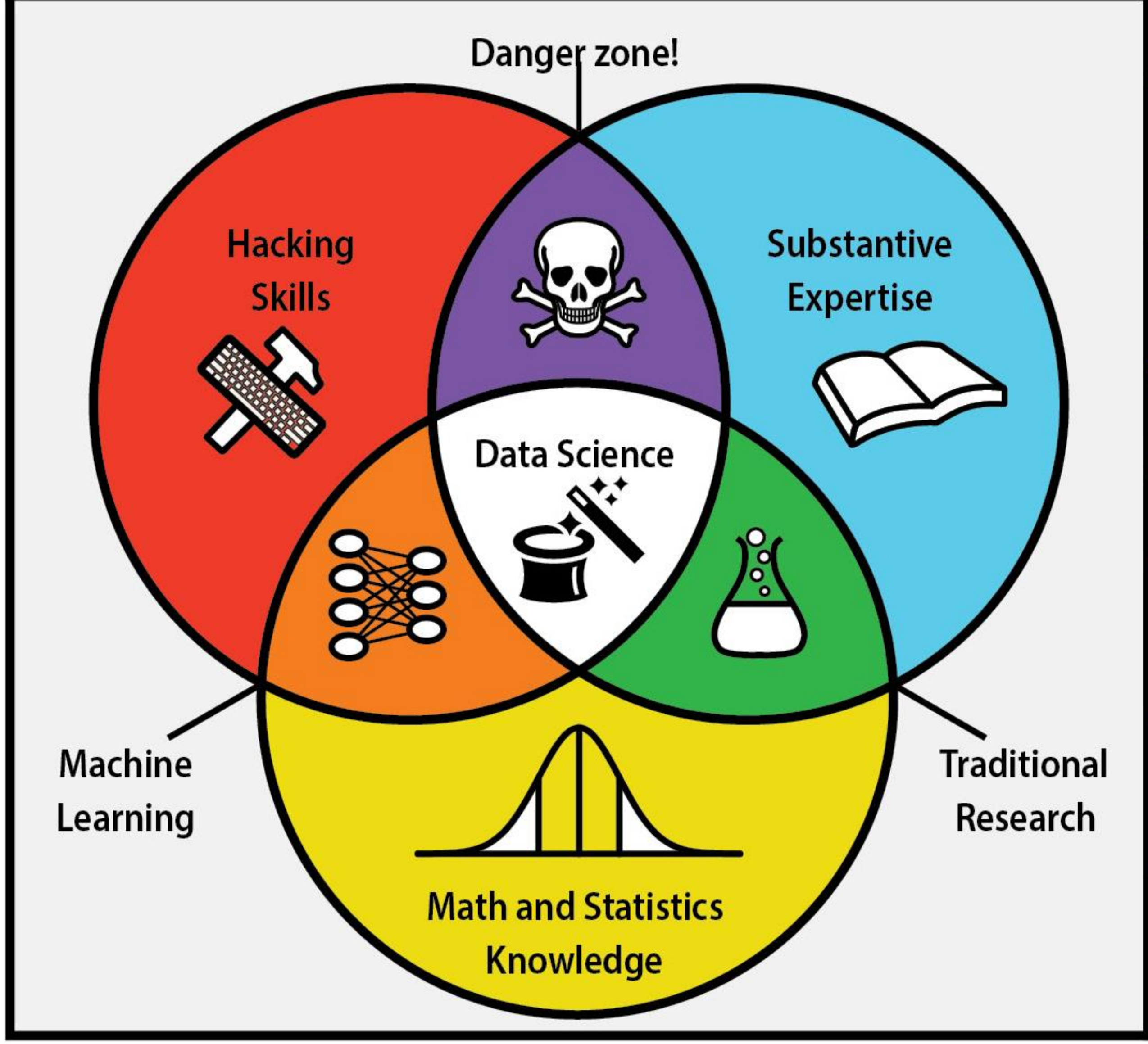
Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

CONJUNTO DE HABILIDADES EN DATA SCIENCE



La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: **habilidades de hacking**, **conocimiento de matemáticas y estadística**, y **experiencia sustantiva** en un campo de la ciencia.

Las **habilidades de hacking** son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.

El **conocimiento matemático y estadístico** permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.

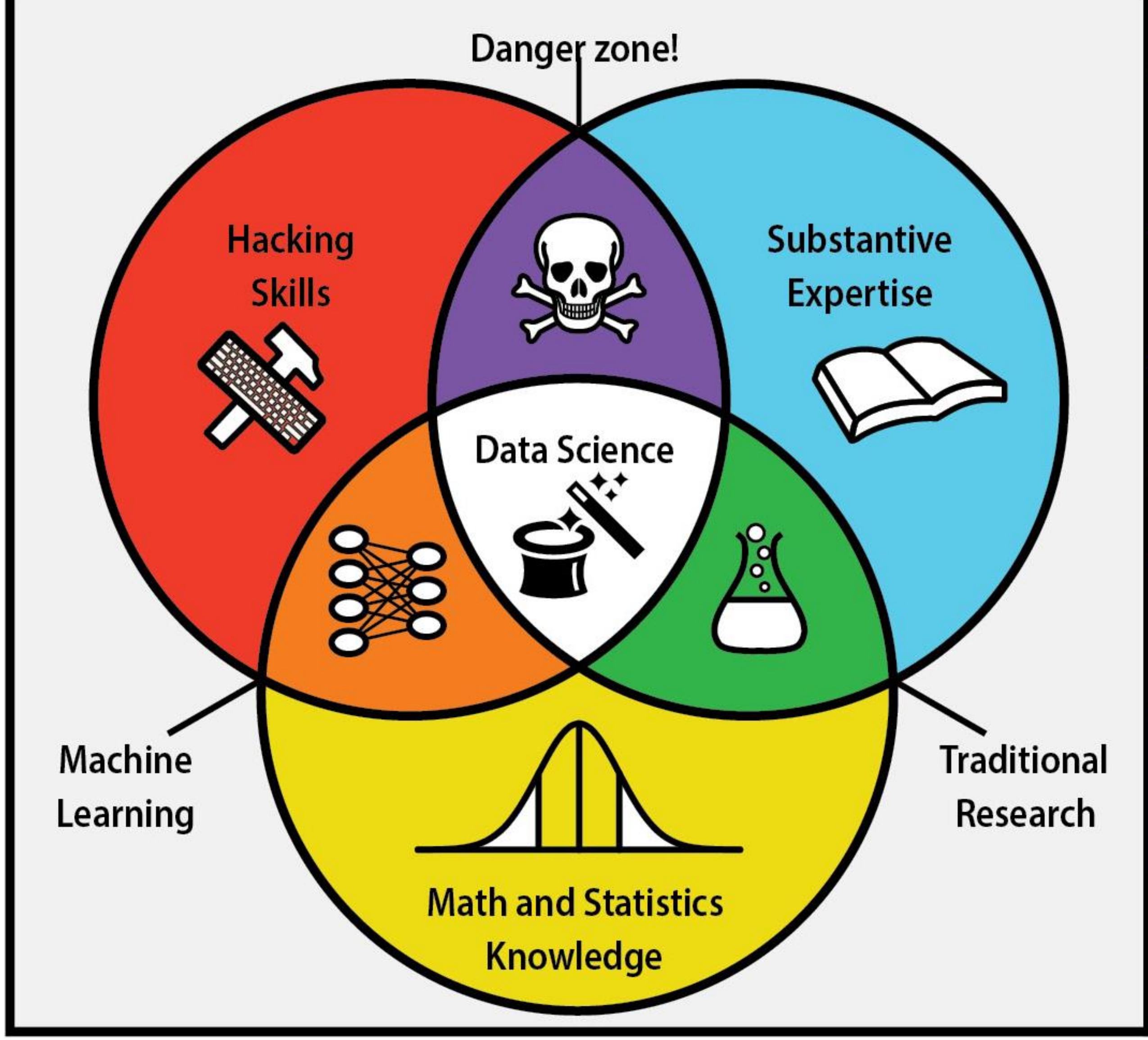
Substantive expertise in a scientific field is crucial for generating motivating questions and hypotheses and interpreting results.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

CONJUNTO DE HABILIDADES EN DATA SCIENCE



La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: **habilidades de hacking**, **conocimiento de matemáticas y estadística**, y **experiencia sustantiva** en un campo de la ciencia.

Las **habilidades de hacking** son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.

El **conocimiento matemático y estadístico** permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.

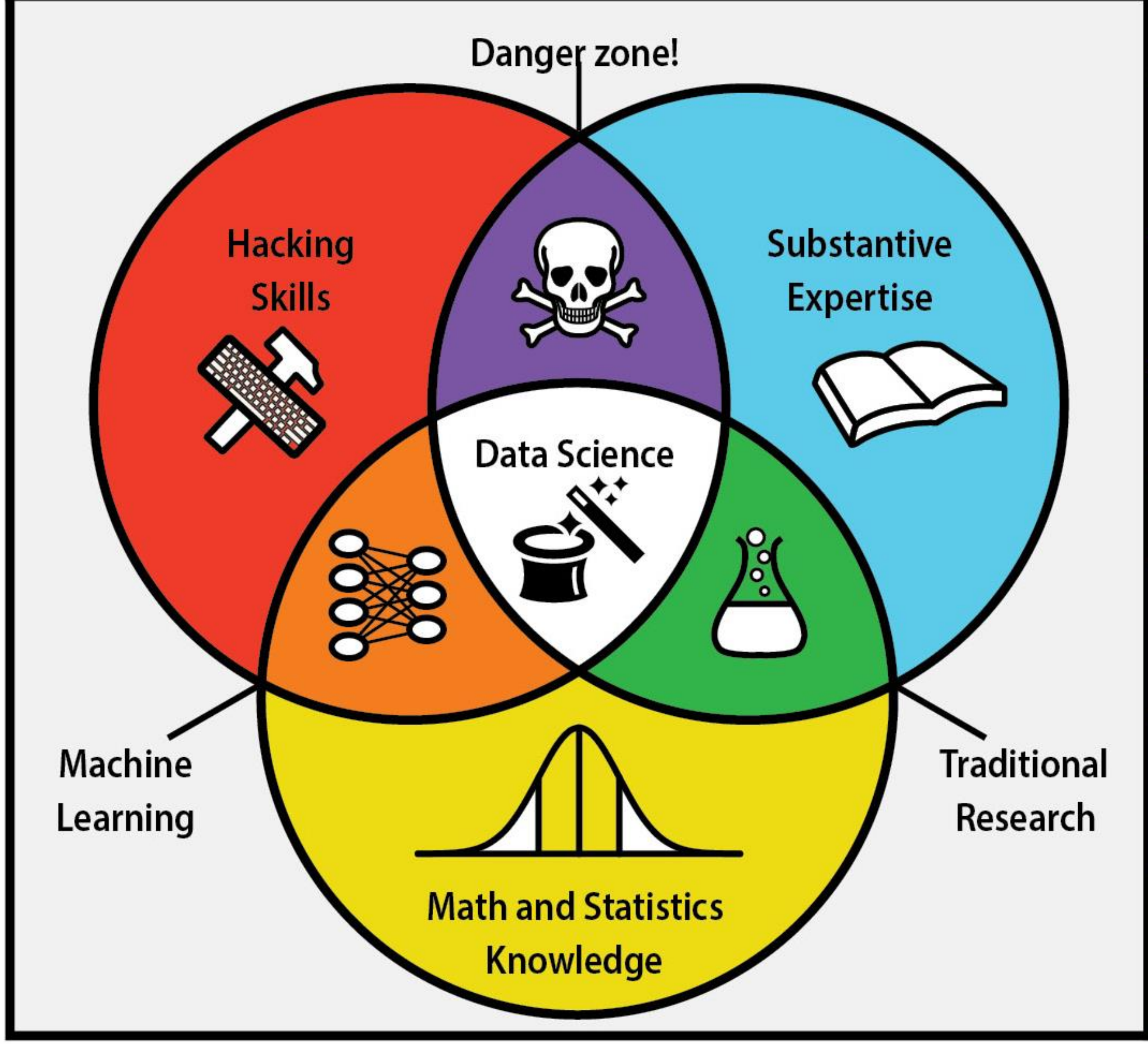
La **experiencia sustantiva** en un campo científico es crucial para generar preguntas motivadoras e hipótesis e interpretar los resultados.

Traditional research lies at the intersection of knowledge of math and statistics with substantive expertise in a scientific field.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

CONJUNTO DE HABILIDADES EN DATA SCIENCE



La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: **habilidades de hacking**, **conocimiento de matemáticas y estadística**, y **experiencia sustantiva** en un campo de la ciencia.

Las **habilidades de hacking** son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.

El **conocimiento matemático y estadístico** permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.

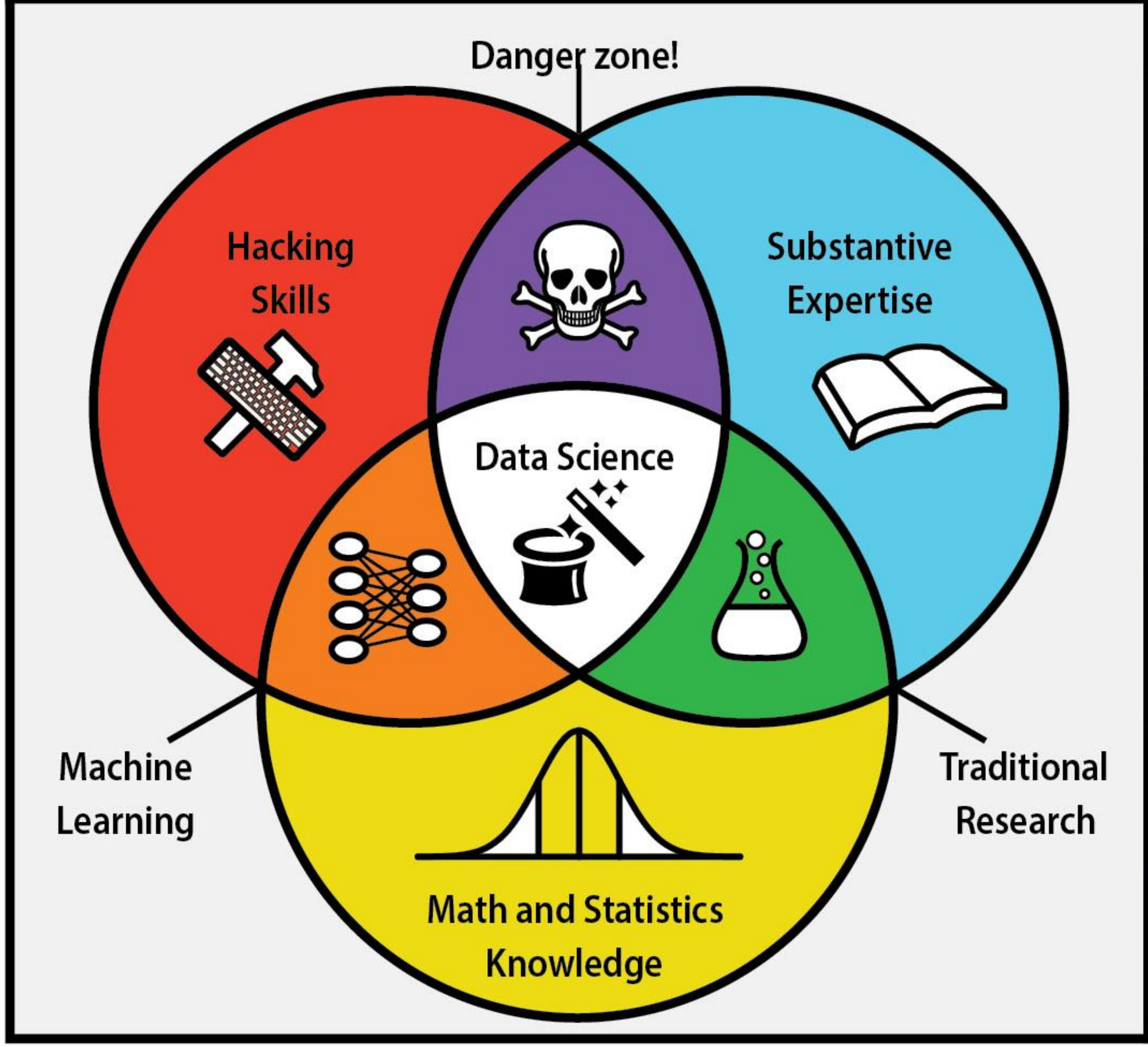
La **experiencia sustantiva** en un campo científico es crucial para generar preguntas motivadoras e hipótesis e interpretar los resultados.

La **investigación tradicional** se encuentra en la intersección del conocimiento de matemáticas y estadística con experiencia sustantiva en un campo científico.

Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

CONJUNTO DE HABILIDADES EN DATA SCIENCE



La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: **habilidades de hacking**, **conocimiento de matemáticas y estadística**, y **experiencia sustantiva** en un campo de la ciencia.

Las **habilidades de hacking** son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.

El **conocimiento matemático y estadístico** permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.

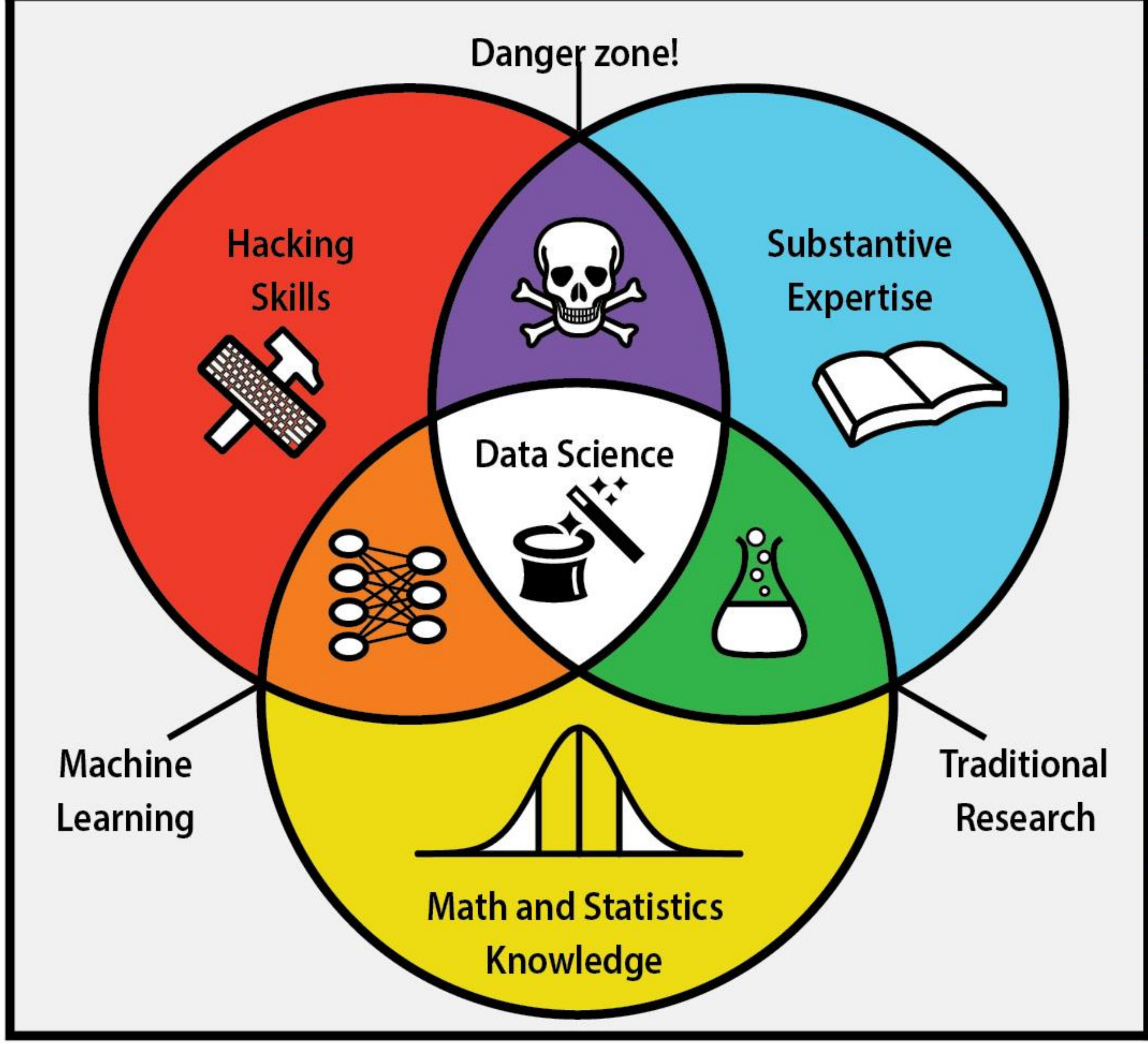
La **experiencia sustantiva** en un campo científico es crucial para generar preguntas motivadoras e hipótesis e interpretar los resultados.

La **investigación tradicional** se encuentra en la intersección del conocimiento de matemáticas y estadística con experiencia sustantiva en un campo científico.

El **aprendizaje automático** se deriva de la combinación de las habilidades de hacking con las matemáticas y el conocimiento estadístico, pero no requiere motivación científica.

Danger zone! Hacking skills combined with substantive scientific expertise without rigorous methods can beget incorrect analyses.

CONJUNTO DE HABILIDADES EN DATA SCIENCE



La ciencia de datos, debido a su naturaleza interdisciplinaria, requiere una intersección de habilidades: **habilidades de hacking**, **conocimiento de matemáticas y estadística**, y **experiencia sustantiva** en un campo de la ciencia.

Las **habilidades de hacking** son necesarias para trabajar con grandes cantidades de datos electrónicos que se deben adquirir, limpiar y manipular.

El **conocimiento matemático y estadístico** permite que un científico de datos elija los métodos y herramientas adecuados para extraer información de los datos.

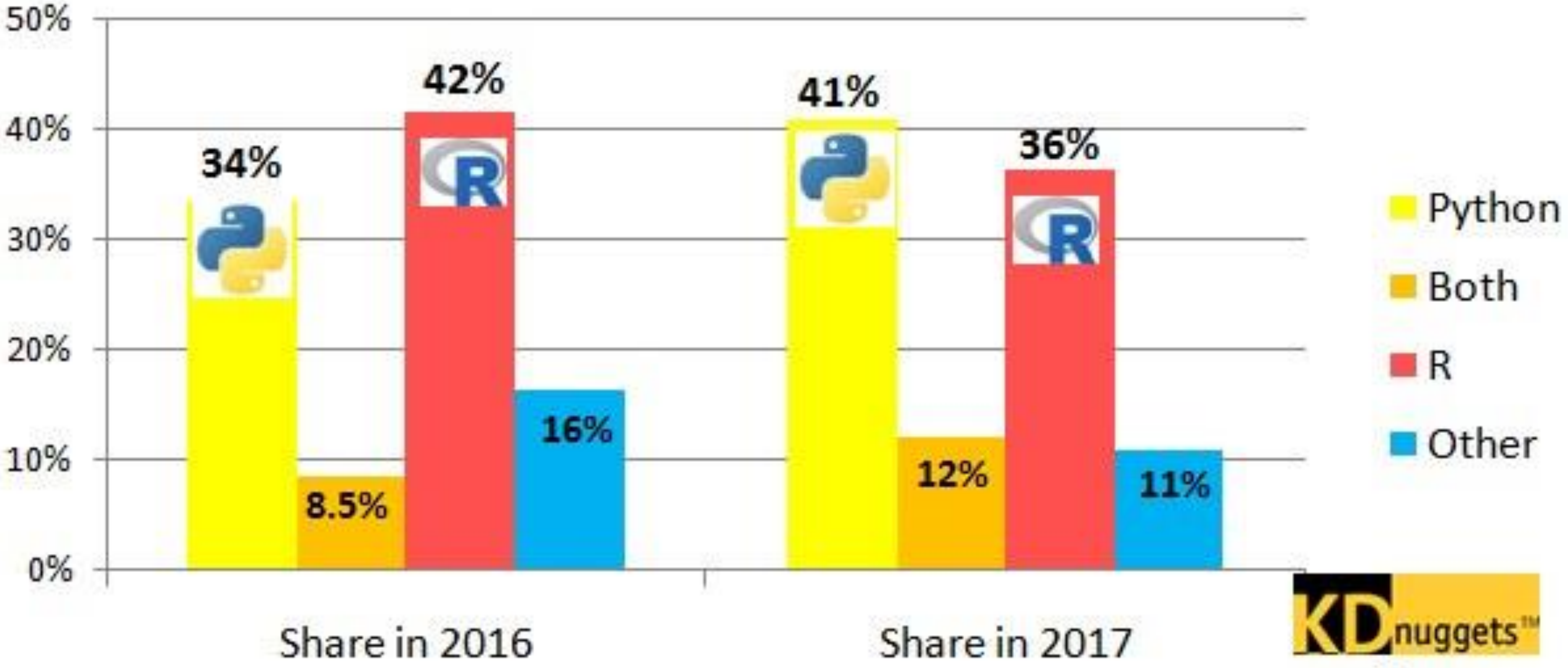
La **experiencia sustantiva** en un campo científico es crucial para generar preguntas motivadoras e hipótesis e interpretar los resultados.

La **investigación tradicional** se encuentra en la intersección del conocimiento de matemáticas y estadística con experiencia sustantiva en un campo científico.

El **aprendizaje automático** se deriva de la combinación de las habilidades de hacking con las matemáticas y el conocimiento estadístico, pero no requiere motivación científica.

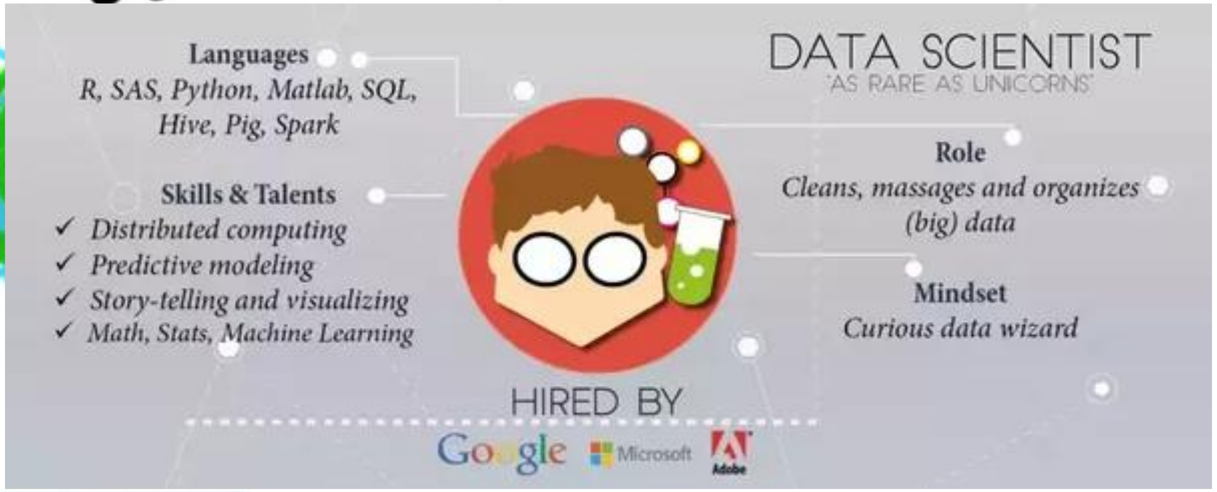
¡Zona peligrosa! Las habilidades de hackings combinadas con la experiencia científica sustantiva sin métodos rigurosos pueden obtener un análisis incorrecto.

Python, R, Both, or Other platforms for Analytics, Data Science, Machine Learning



Poll 2017

What does a data scientist do?



DATA SCIENTIST

"AS RARE AS UNICORNS"

Languages

R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents

- ✓ Distributed computing
- ✓ Predictive modeling
- ✓ Story-telling and visualizing
- ✓ Math, Stats, Machine Learning

Role

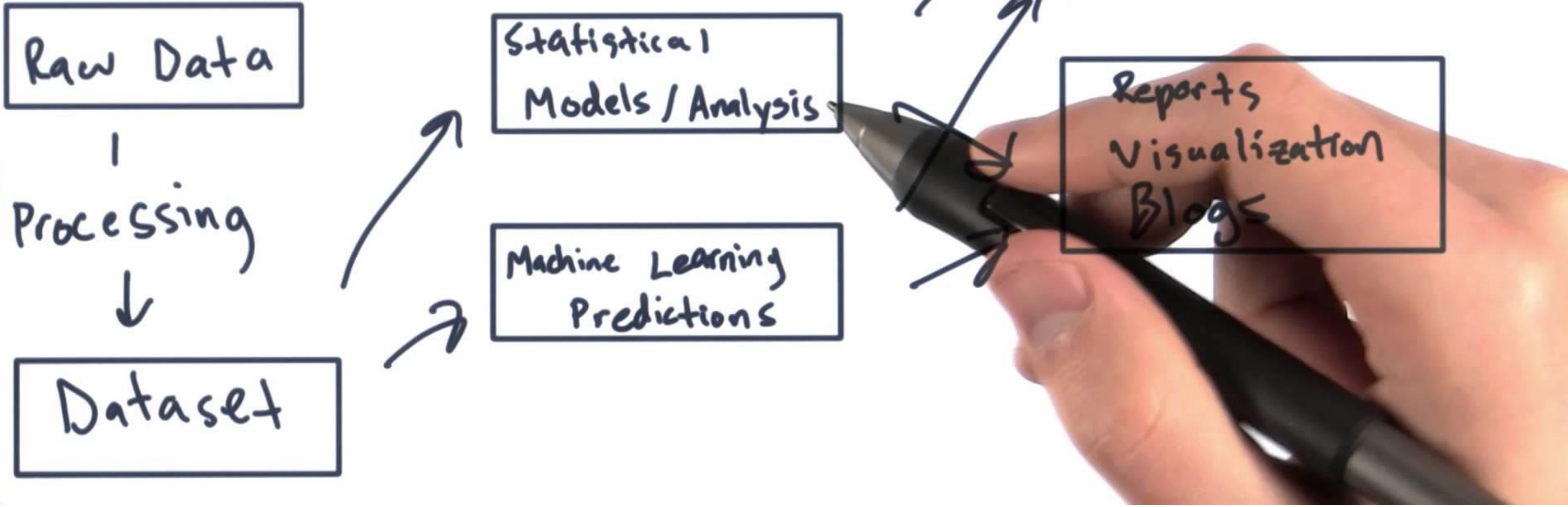
Cleans, massages and organizes (big) data

Mindset

Curious data wizard

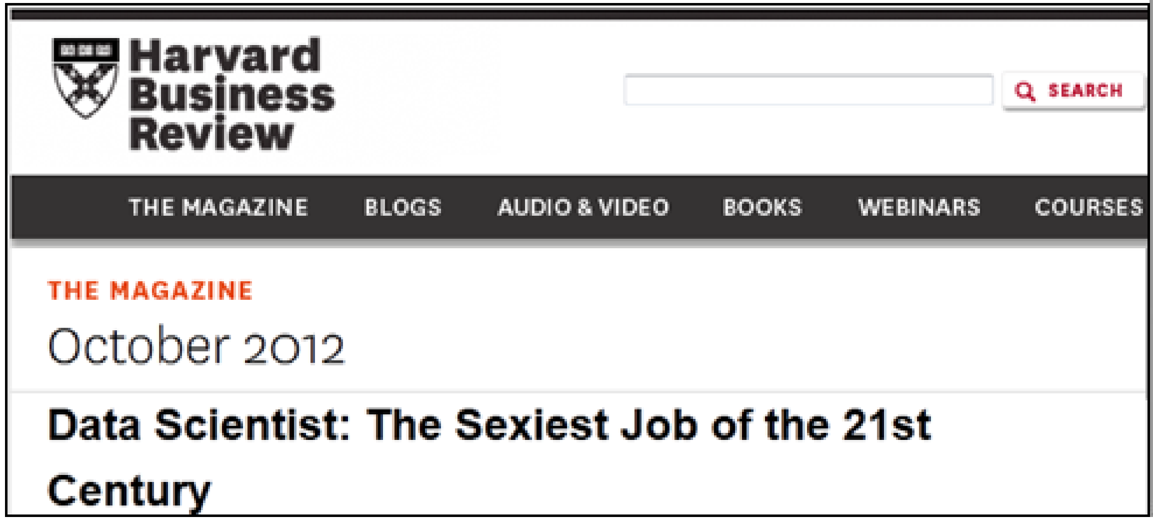
HIRED BY

Google
Microsoft
Adobe



Data Scientist

- Reconocido como uno de los mejores trabajos
- Grandes Salarios
- Solución de problemas interesantes



THE MAGAZINE

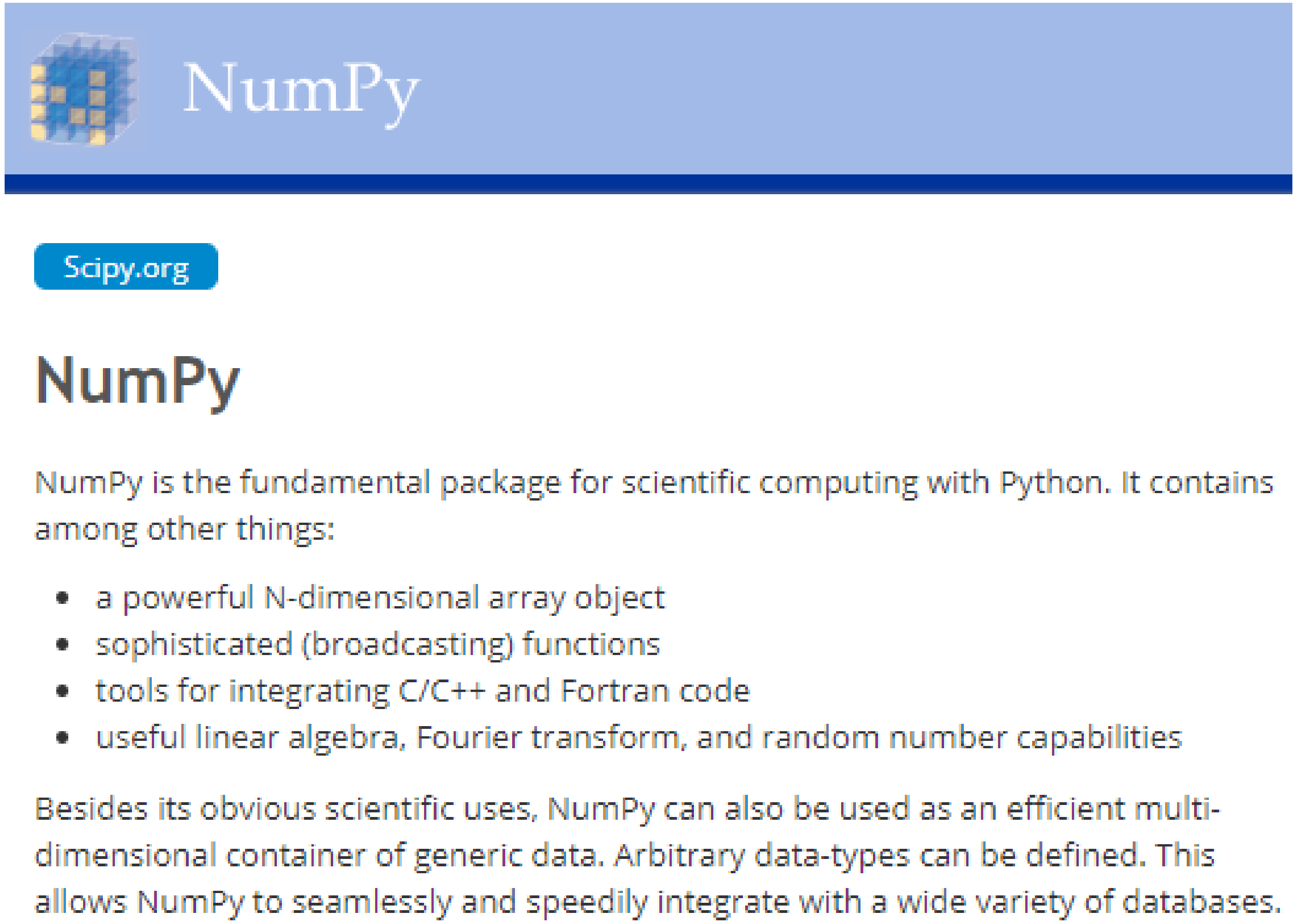
October 2012

Data Scientist: The Sexiest Job of the 21st Century

HBR.ORG
Harvard Business Review
OCTOBER 2012
46 The Big Idea
The True Measures Of Success
Michael J. Mauboussin
84 International Business
10 Rules for Managing Global Innovation
Keeley Wilson and Yves L. Doz
93 Leadership
What Ever Happened To Accountability?
Thomas E. Ricks
GETTING CONTROL OF
BIG DATA
How vast new streams of information are changing the art of management
PAGE 59

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- MatplotLib
- Plotly
- PySpark



NumPy

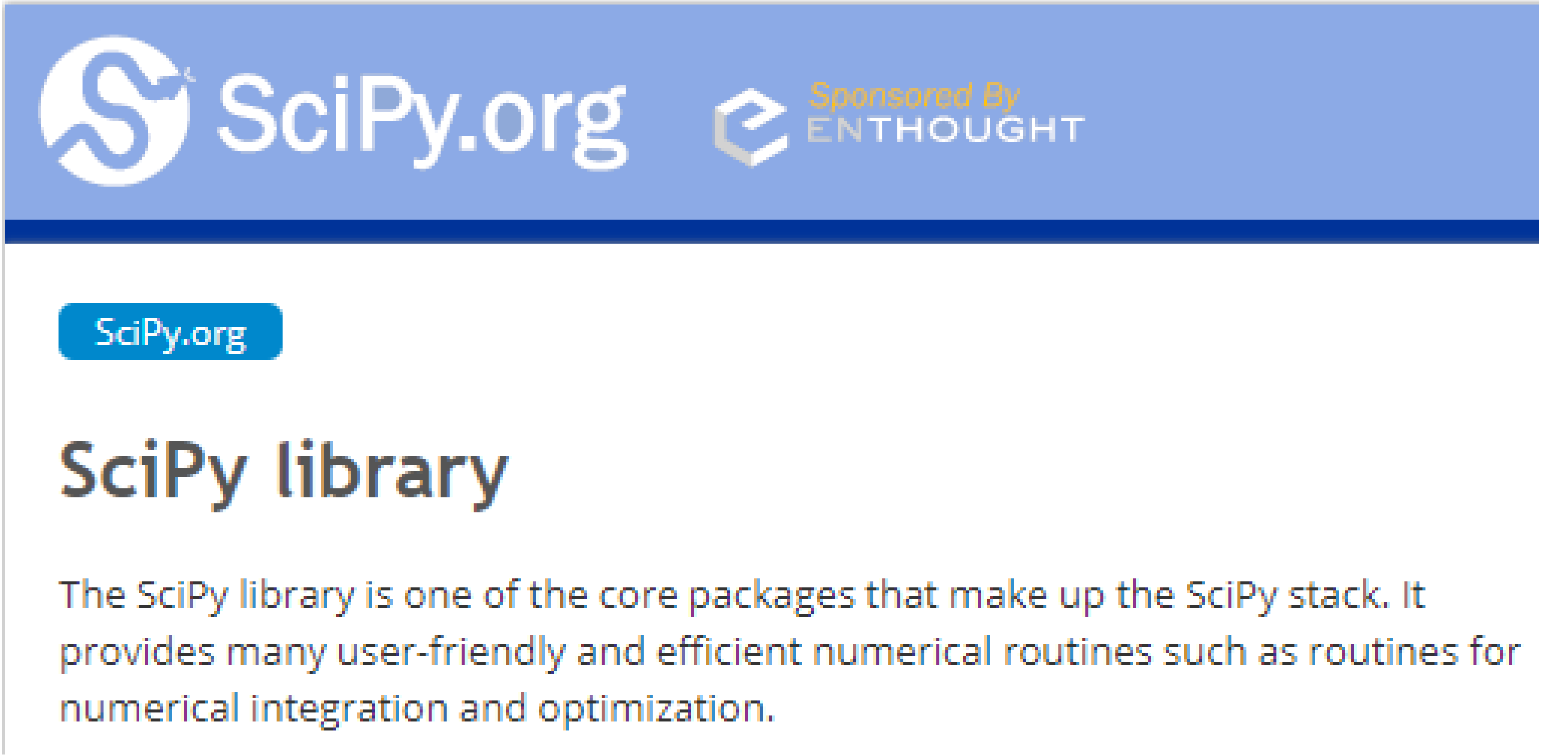
NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- a powerful N-dimensional array object
- sophisticated (broadcasting) functions
- tools for integrating C/C++ and Fortran code
- useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- MatplotLib
- Plotly
- PySpark

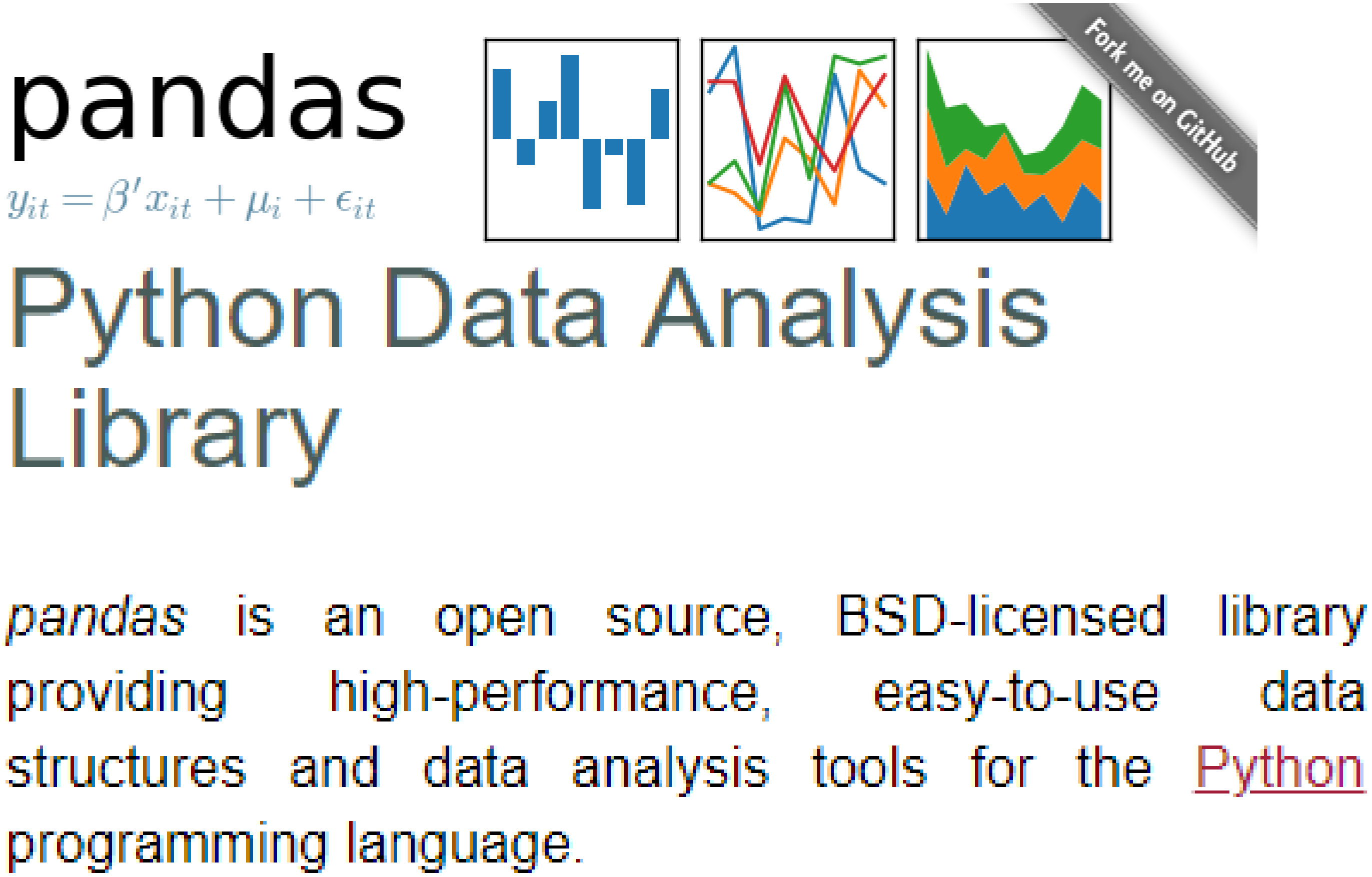


SciPy library

The SciPy library is one of the core packages that make up the SciPy stack. It provides many user-friendly and efficient numerical routines such as routines for numerical integration and optimization.

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- MatplotLib
- Plotly
- PySpark

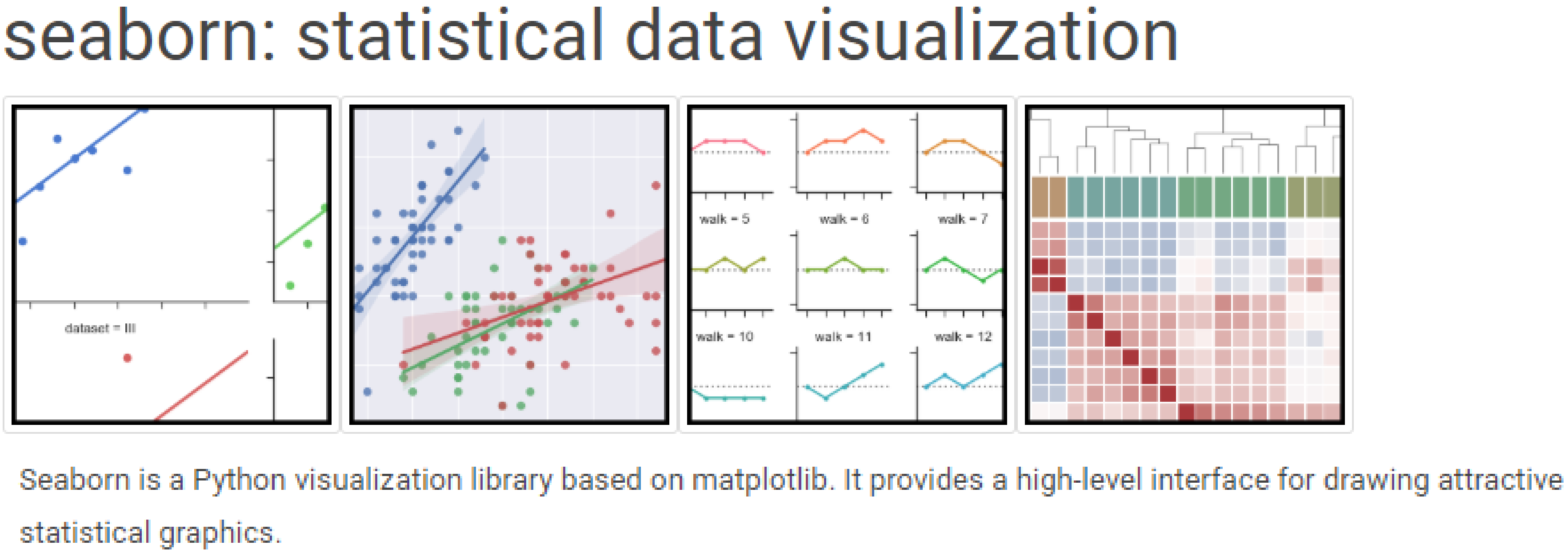


Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- MatplotLib
- Plotly
- PySpark

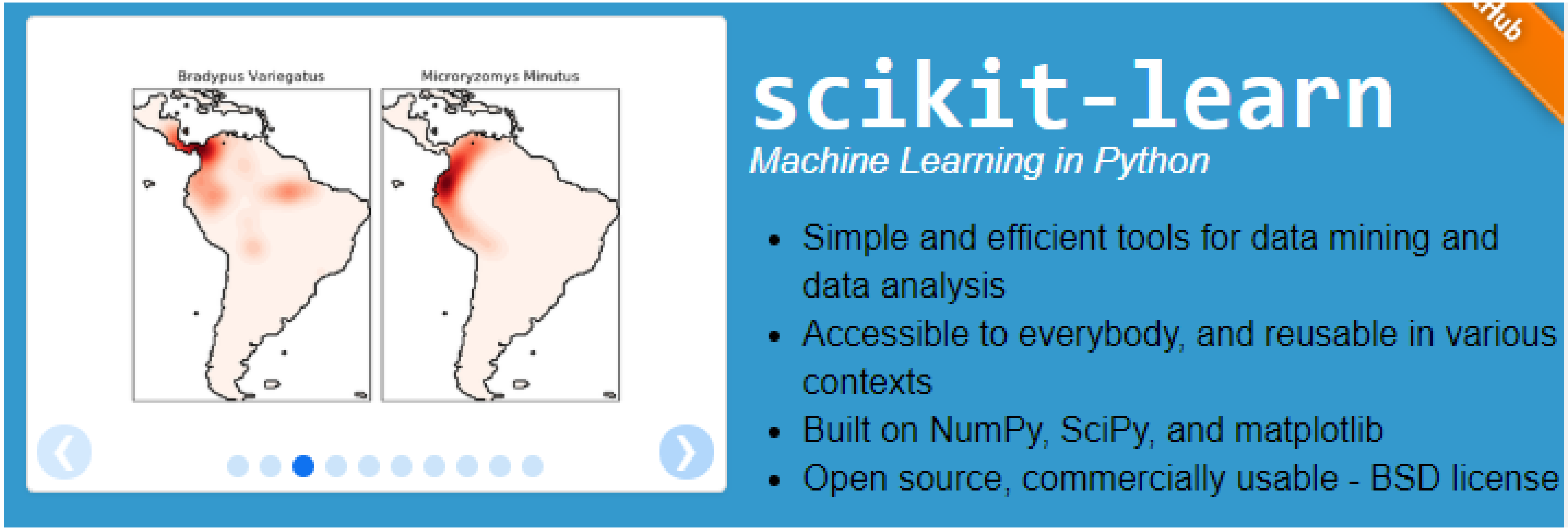


seaborn: statistical data visualization

Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- MatplotLib
- Plotly
- PySpark



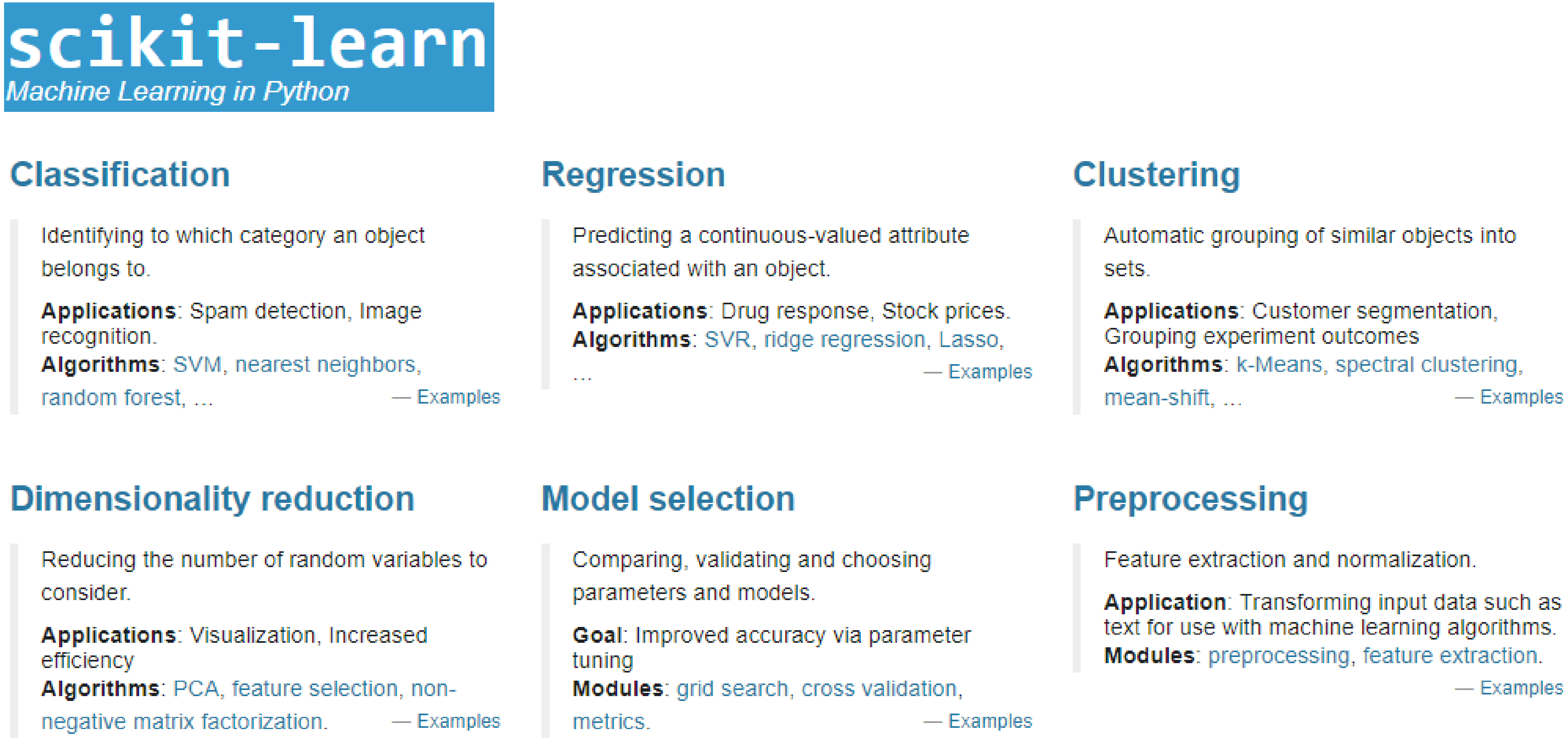
scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark



scikit-learn

Machine Learning in Python

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: grid search, cross validation, metrics. — Examples

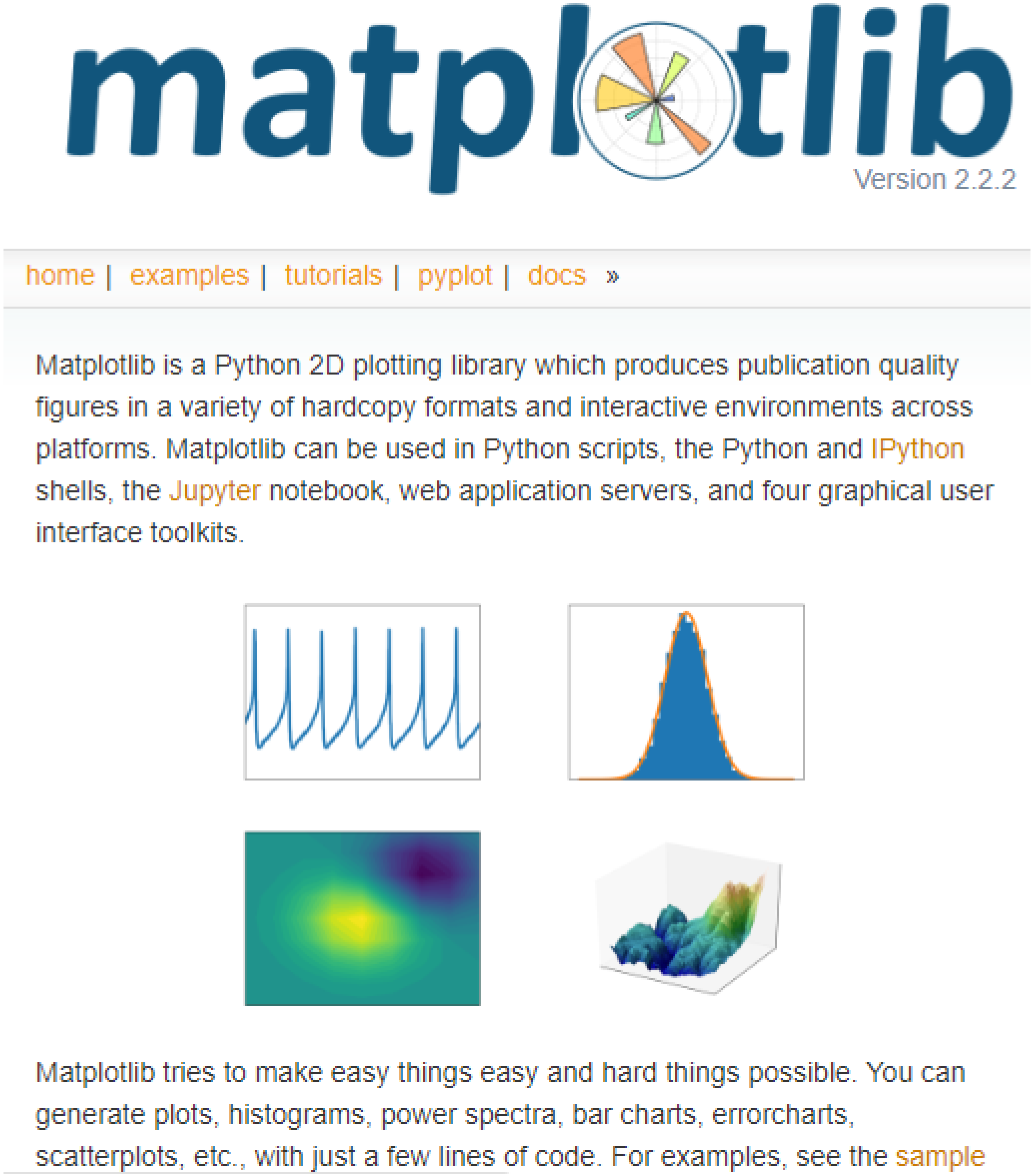
Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: preprocessing, feature extraction. — Examples

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark

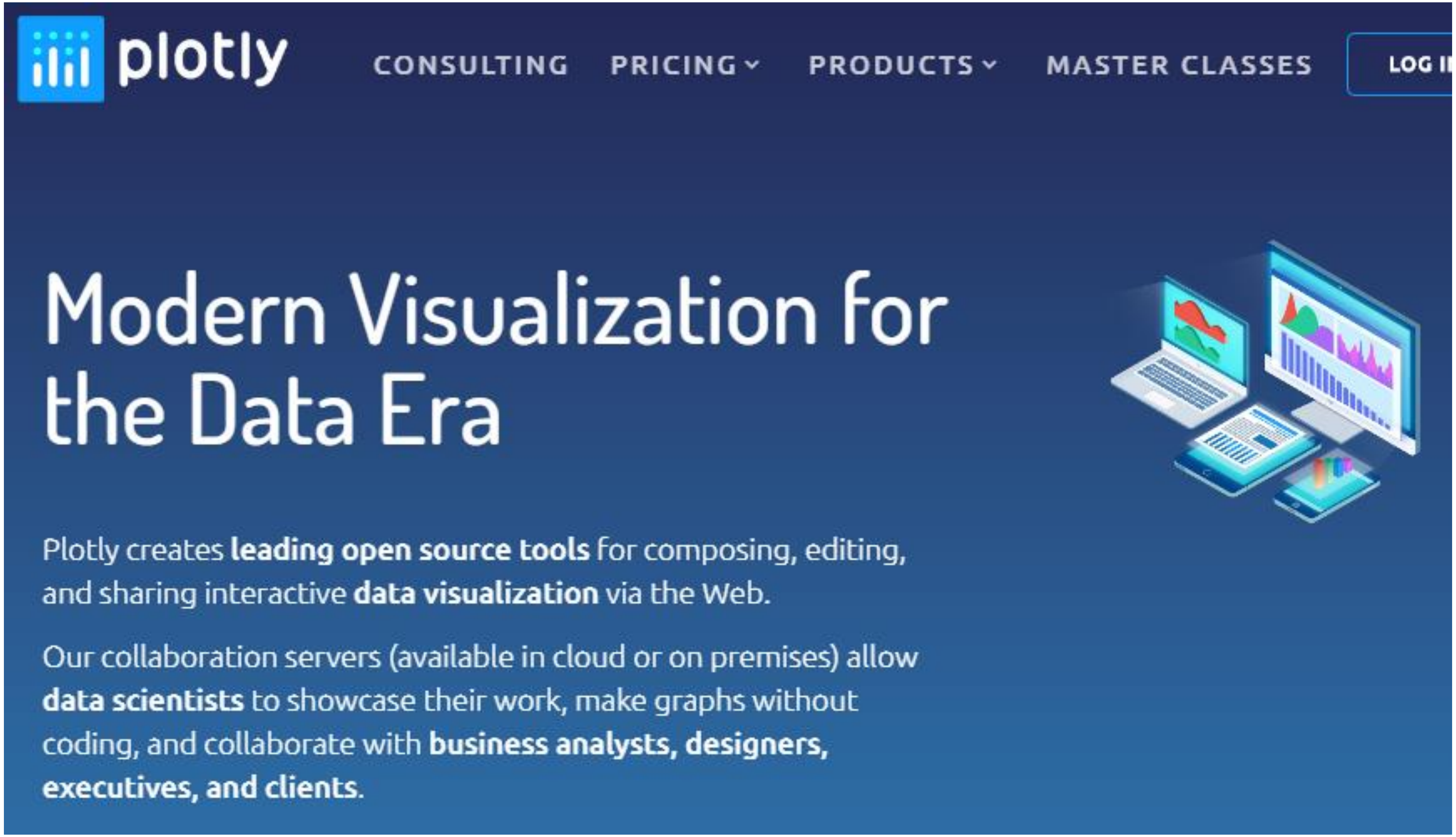


Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code. For examples, see the sample

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark



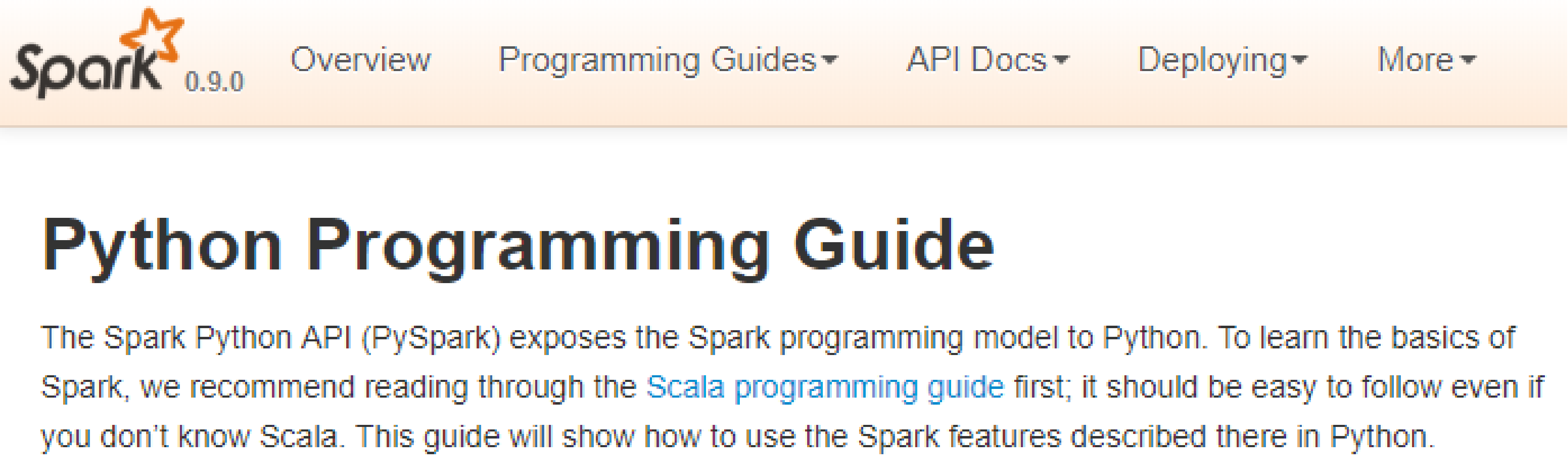
Modern Visualization for the Data Era

Plotly creates **leading open source tools** for composing, editing, and sharing interactive **data visualization** via the Web.

Our collaboration servers (available in cloud or on premises) allow **data scientists** to showcase their work, make graphs without coding, and collaborate with **business analysts, designers, executives, and clients**.

Librerías mas populares para ciencias de datos en Python

- NumPy
- SciPy
- Pandas
- Seaborn
- scikit-learn
- Matplotlib
- Plotly
- PySpark

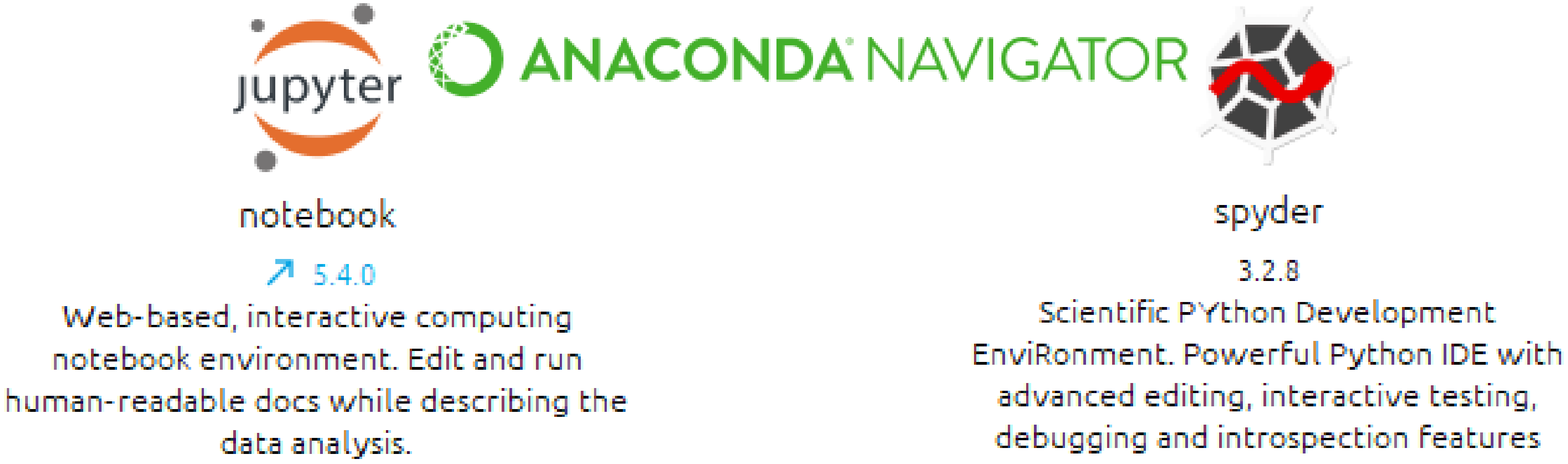


Python Programming Guide

The Spark Python API (PySpark) exposes the Spark programming model to Python. To learn the basics of Spark, we recommend reading through the Scala programming guide first; it should be easy to follow even if you don't know Scala. This guide will show how to use the Spark features described there in Python.

Configuración de Entorno

- En este taller usaremos Notebooks de Jupyter.
- Sin embargo usted es libre de usar el entorno de desarrollo que prefiera.
- Todas las notas pueden ser descargadas como archivos .py que son compatibles con cualquier IDE de Python o editor de texto.
- Usaremos la última versión de Python 3 a través de la distribución de Anaconda



notebook

↗ 5.4.0

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

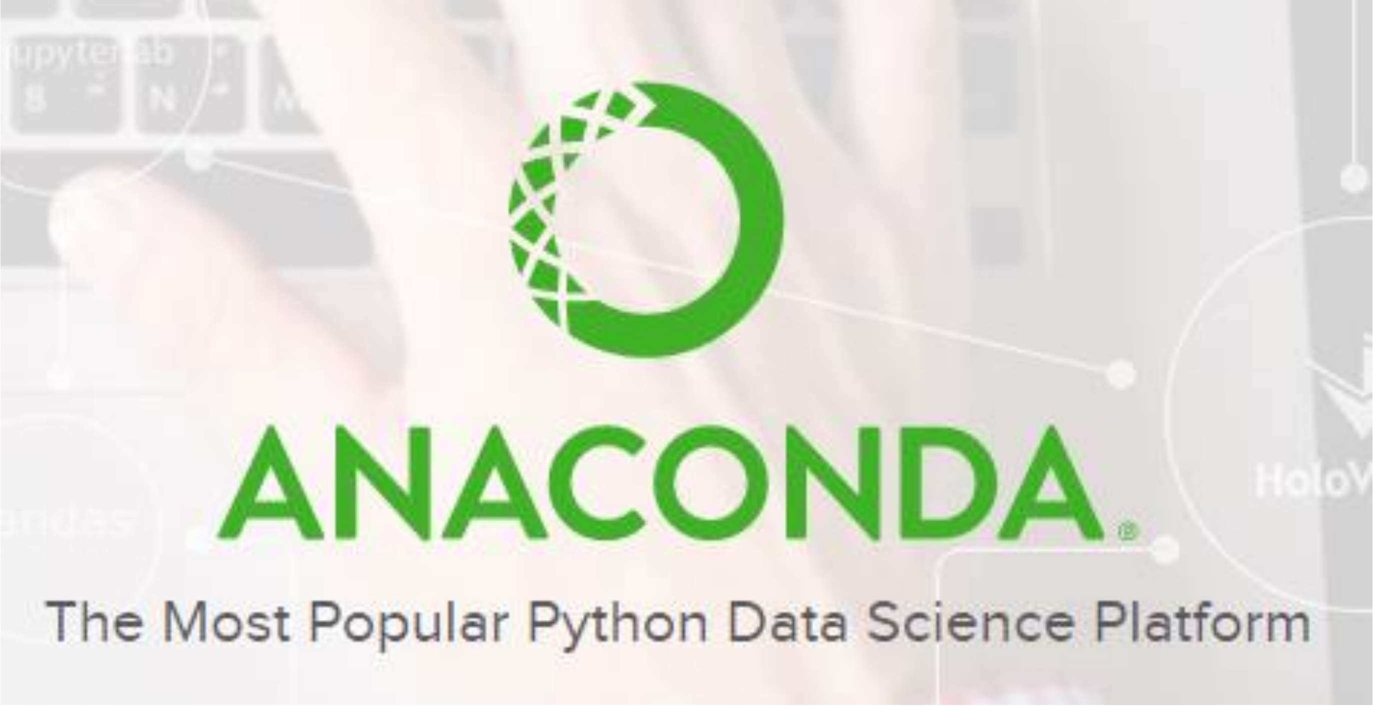
spyder

3.2.8

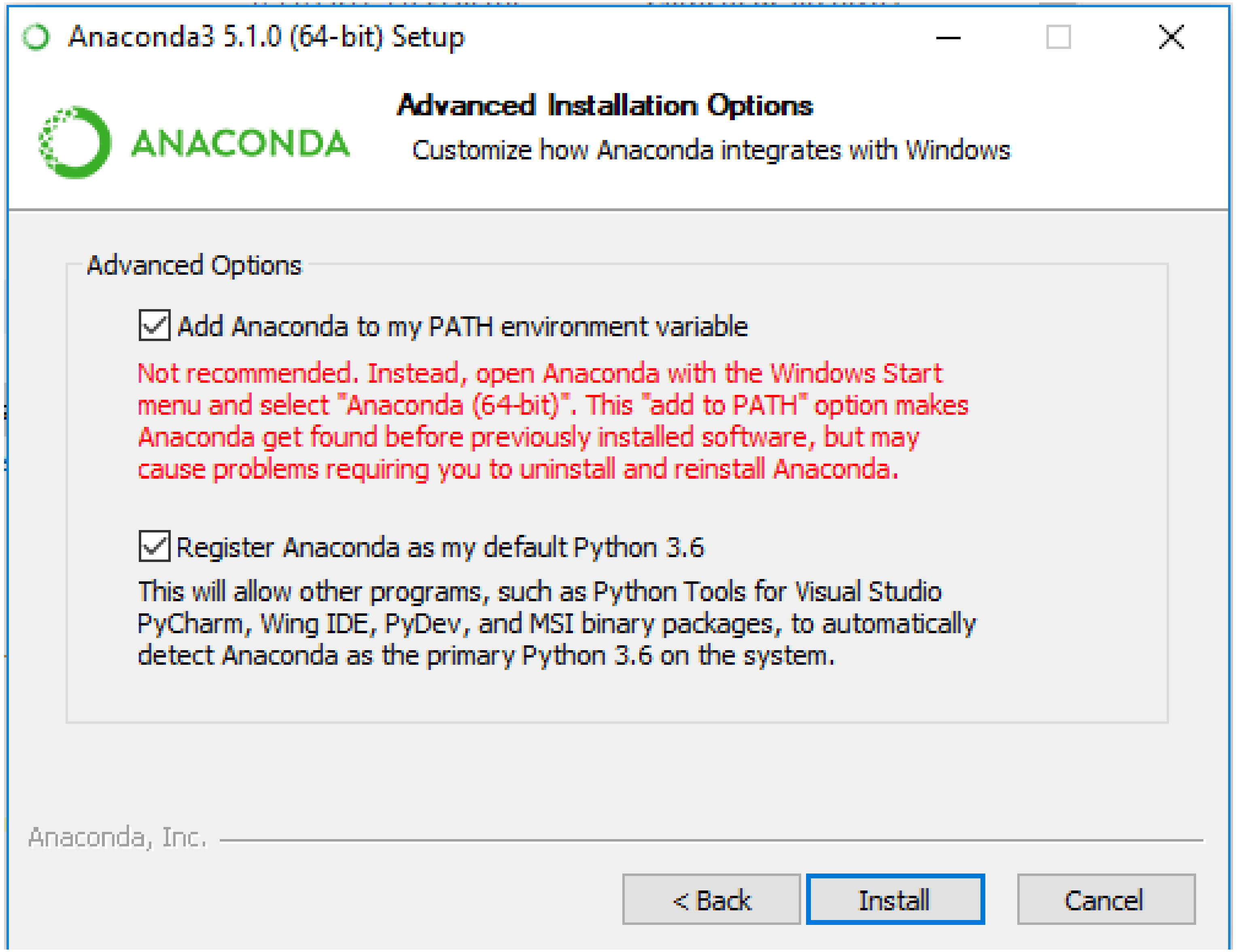
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Instalación de Anaconda Navigator

Desinstalar cualquier versión previa de Python, antes de instalar Anaconda.



Es muy importante considerar esta opción en la instalación para poder seguir los mismos pasos en los ejemplos



Anaconda3 5.1.0 (64-bit) Setup

Advanced Installation Options

Customize how Anaconda integrates with Windows

Advanced Options

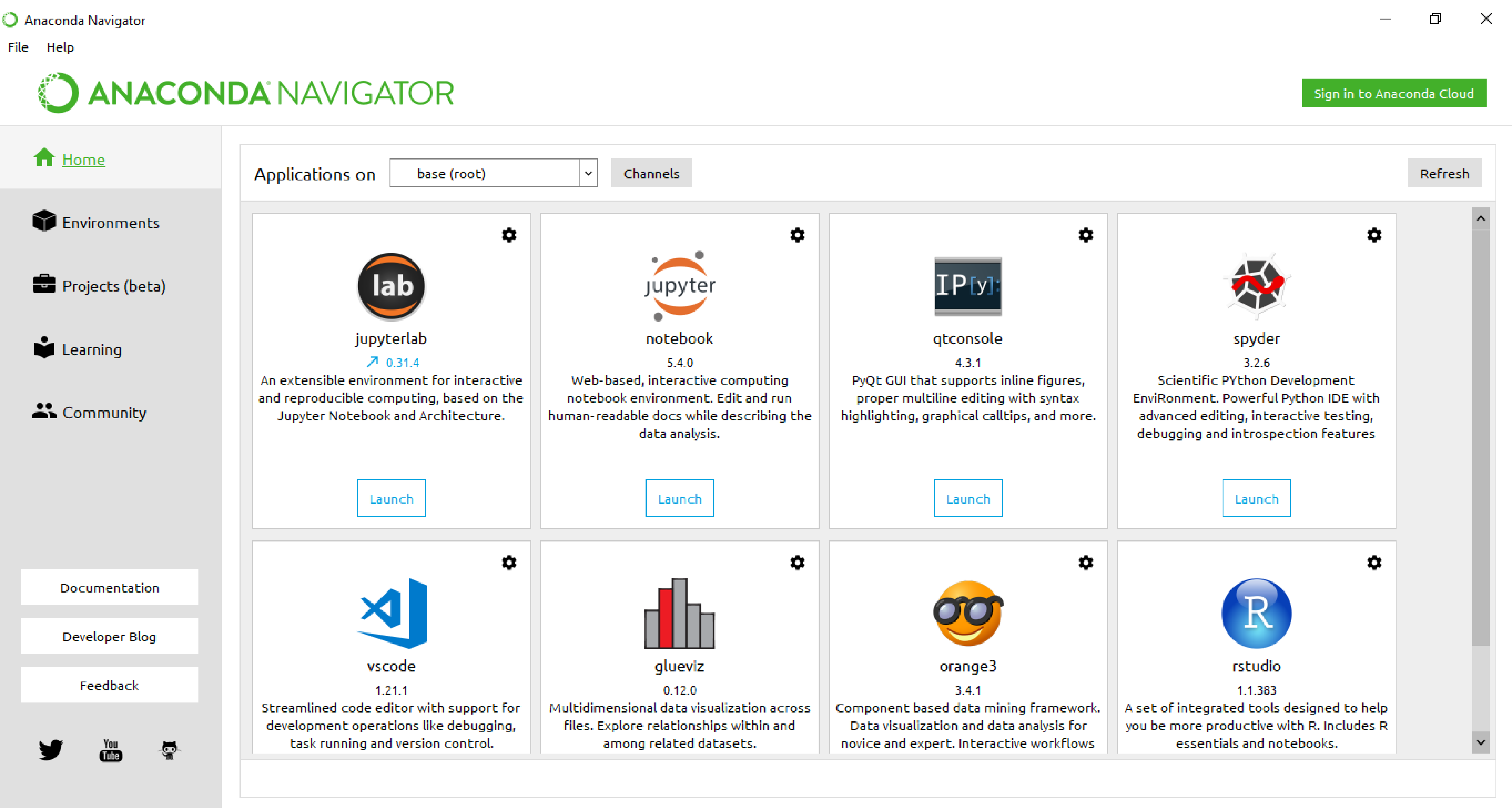
☑ Add Anaconda to my PATH environment variable

Not recommended. Instead, open Anaconda with the Windows Start menu and select "Anaconda (64-bit)". This "add to PATH" option makes Anaconda get found before previously installed software, but may cause problems requiring you to uninstall and reinstall Anaconda.

☑ Register Anaconda as my default Python 3.6

This will allow other programs, such as Python Tools for Visual Studio PyCharm, Wing IDE, PyDev, and MSI binary packages, to automatically detect Anaconda as the primary Python 3.6 on the system.

Anaconda, Inc.



jupyterlab
↗ 0.31.4
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

notebook
5.4.0
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

qtconsole
4.3.1
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

spyder
3.2.6
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

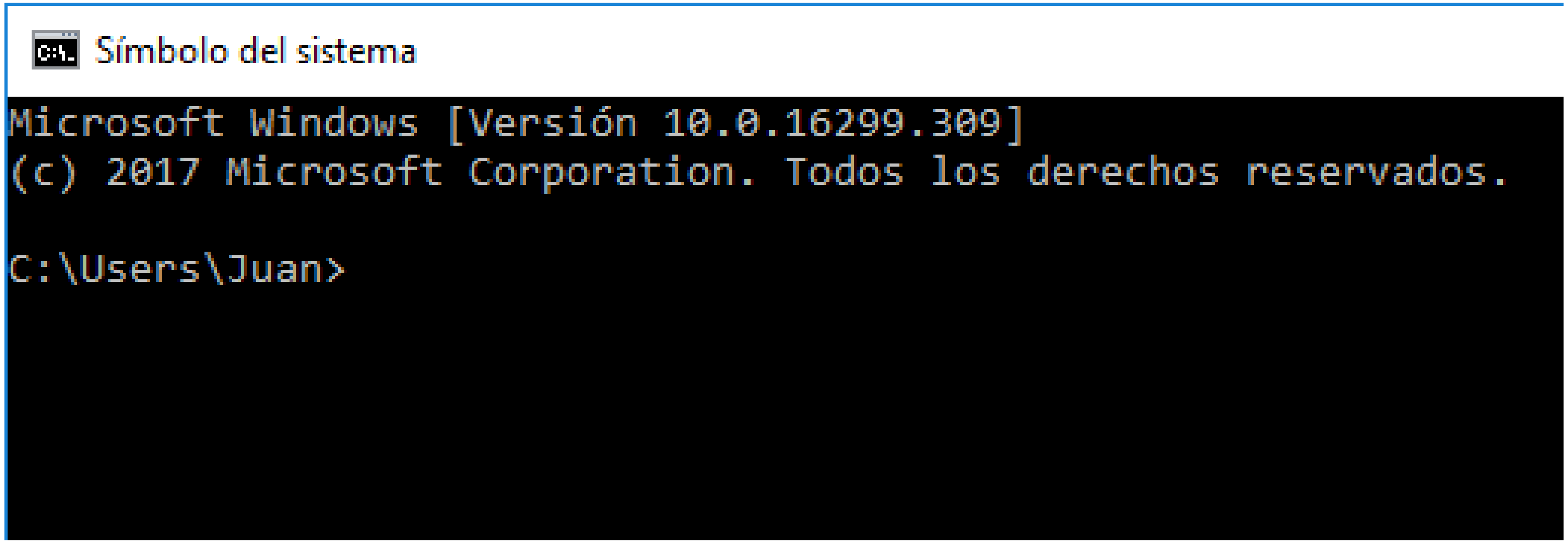
vscode
1.21.1
Streamlined code editor with support for development operations like debugging, task running and version control.

glueviz
0.12.0
Multidimensional data visualization across files. Explore relationships within and among related datasets.

orange3
3.4.1
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows

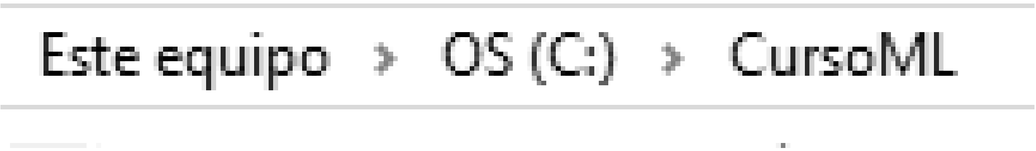
rstudio
1.1.383
A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Comprobar la instalación adecuada con la ventana de Símbolo del Sistema

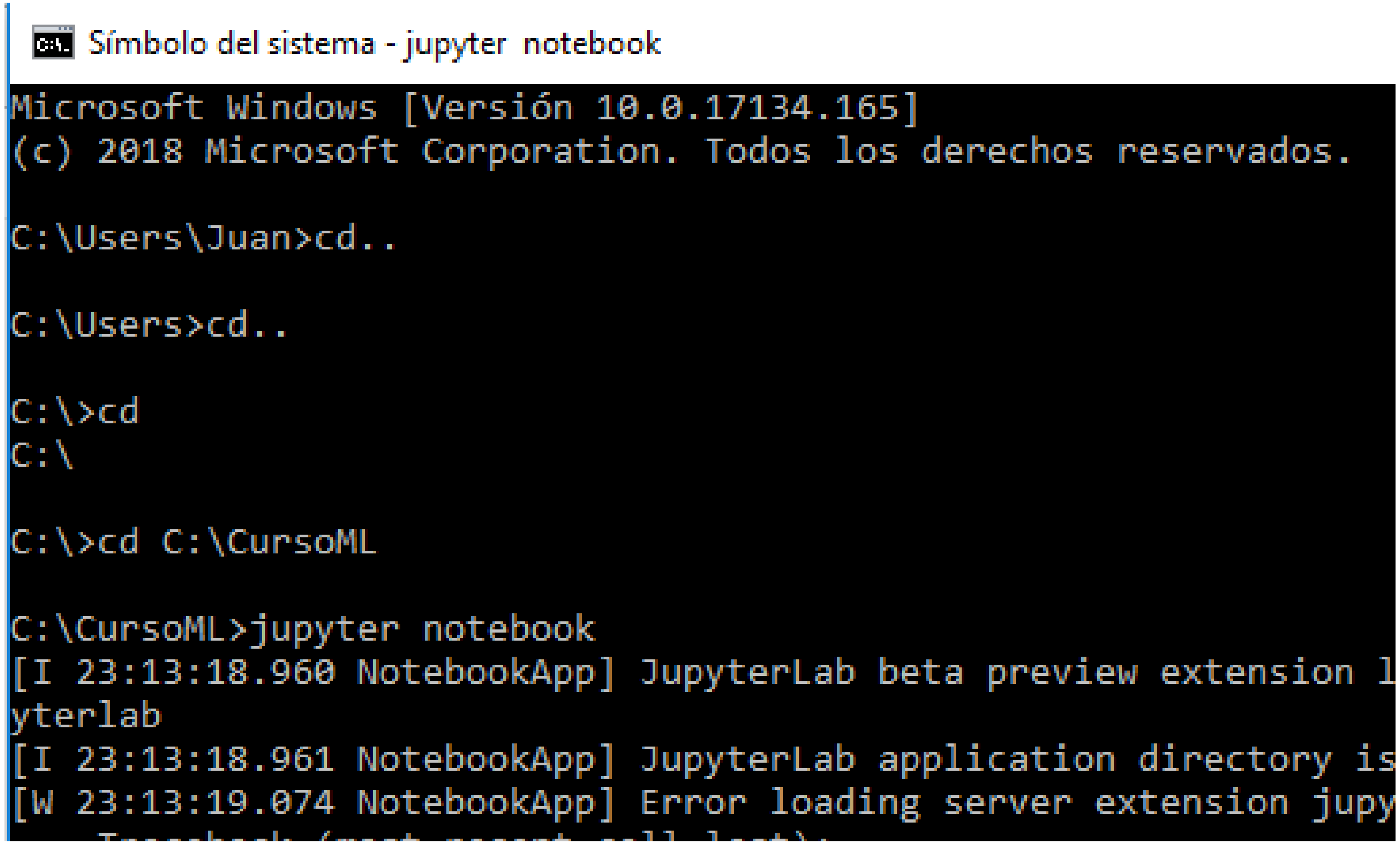


```
Microsoft Windows [Versión 10.0.16299.309]
(c) 2017 Microsoft Corporation. Todos los derechos reservados.

C:\Users\Juan>
```

Si tiene creado en la unidad C las siguientes carpetas:


Este equipo › OS (C:) › CursoML

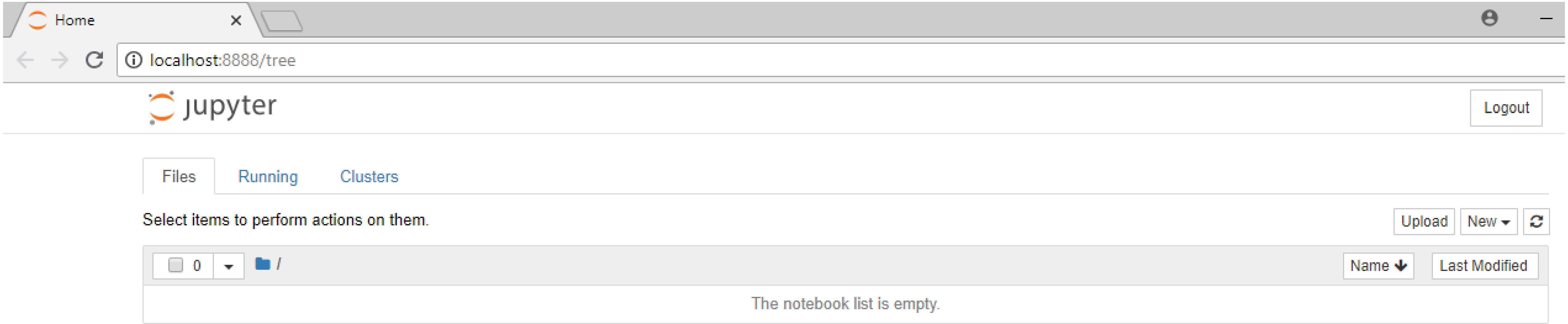
Cambiar a la carpeta correspondiente


Símbolo del sistema - jupyter notebook

```
Microsoft Windows [Versión 10.0.17134.165]
(c) 2018 Microsoft Corporation. Todos los derechos reservados.

C:\Users\Juan>cd..

C:\Users>cd..

C:\>cd
C:\

C:\>cd C:\CursoML

C:\CursoML>jupyter notebook
[I 23:13:18.960 NotebookApp] JupyterLab beta preview extension l
yterlab
[I 23:13:18.961 NotebookApp] JupyterLab application directory is
[W 23:13:19.074 NotebookApp] Error loading server extension jupy
    Traceback (most recent call last):
```

jupyter
python ™

Obtenemos:



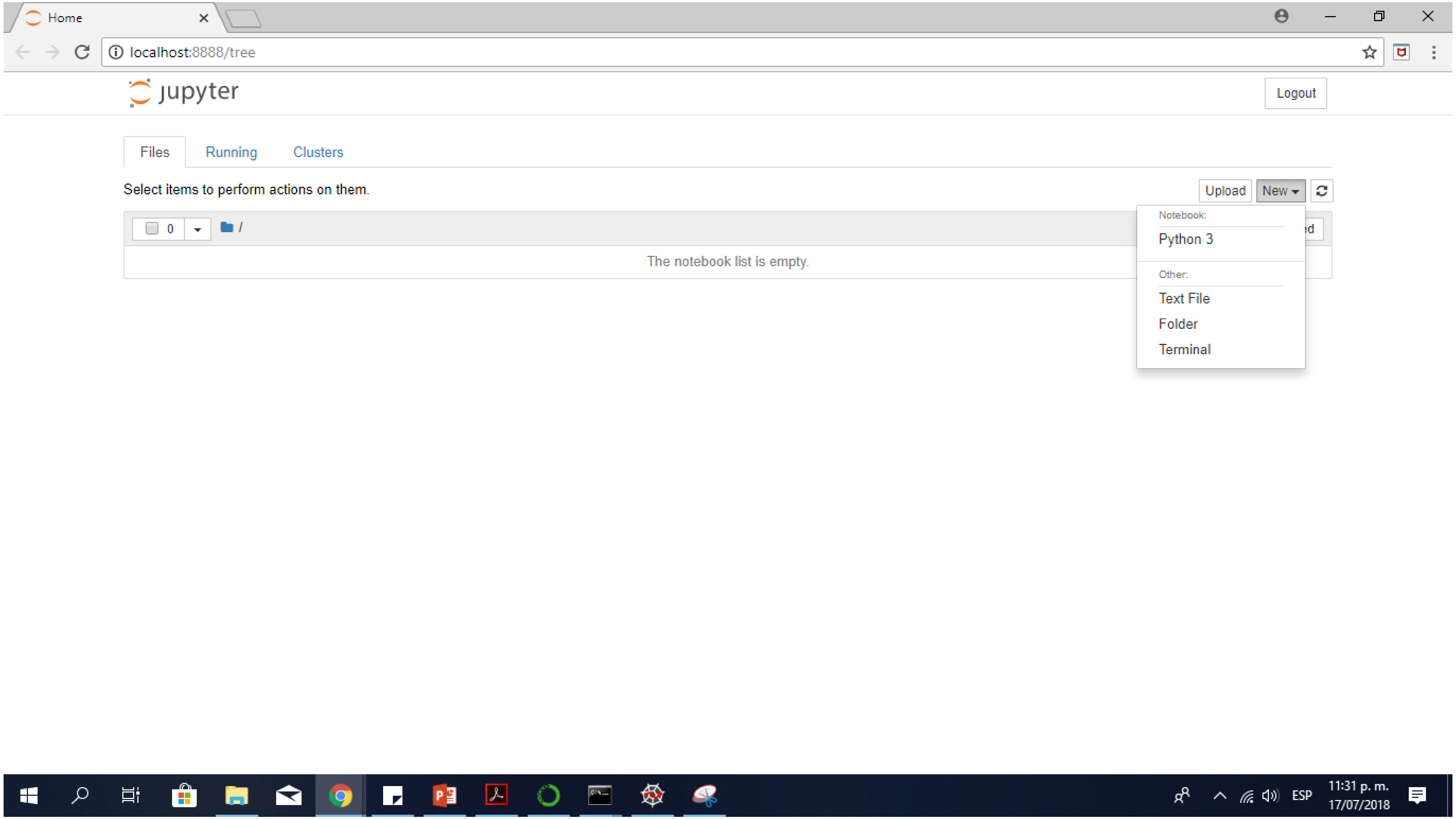
localhost:8888/tree

Select items to perform actions on them.

The notebook list is empty.

Para crear un block de notas

Se hace clic en New y se elige Python 3

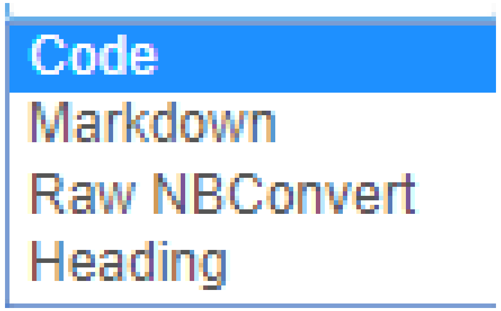


Select items to perform actions on them.

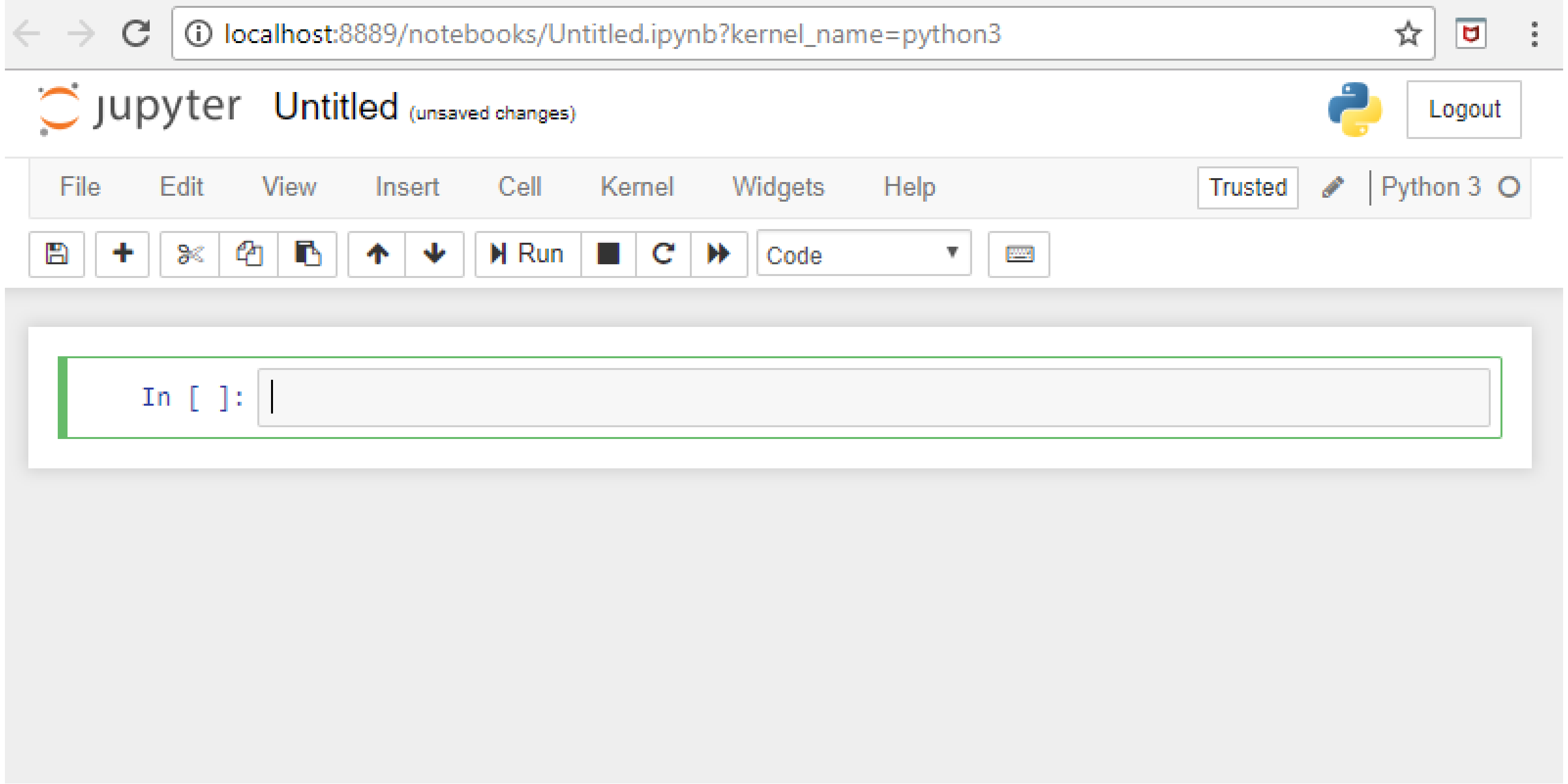
The notebook list is empty.

El block de notas

En el block de notas tenemos distintos tipos de celdas como:

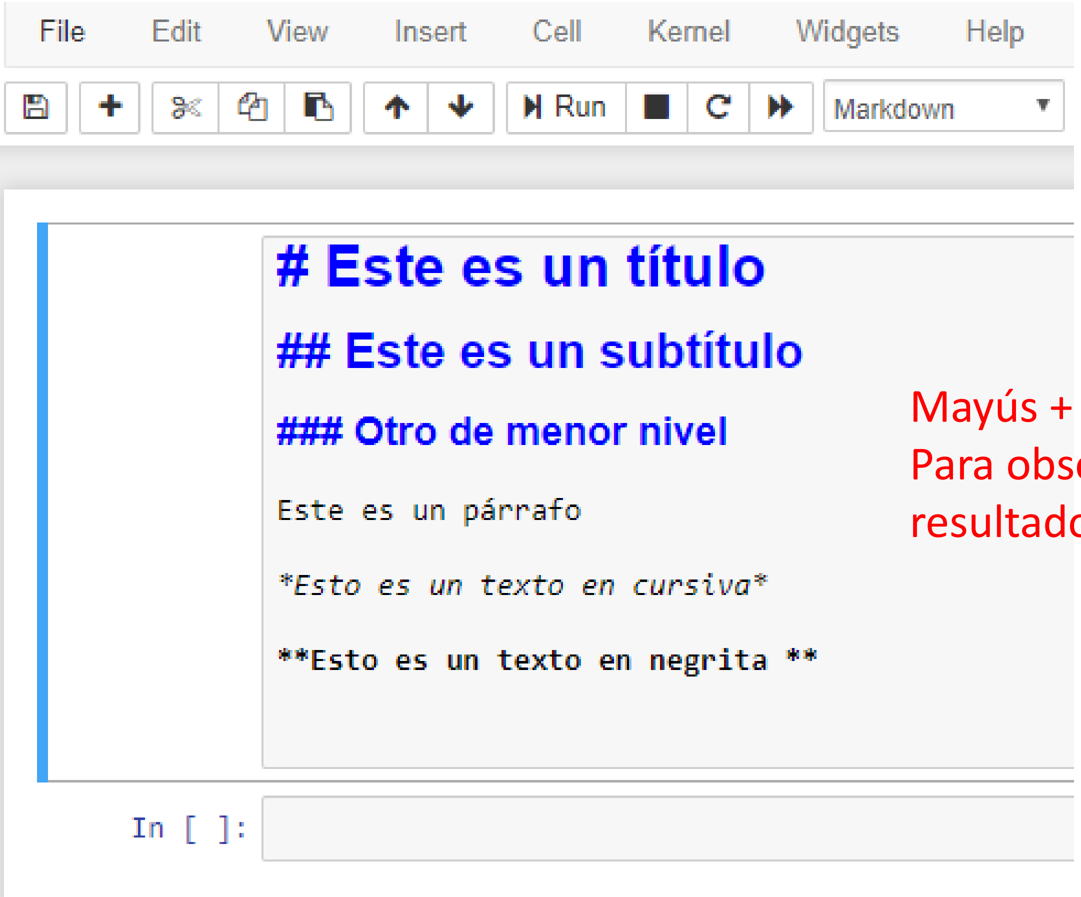


- Code
- Markdown
- Raw NBConvert
- Heading



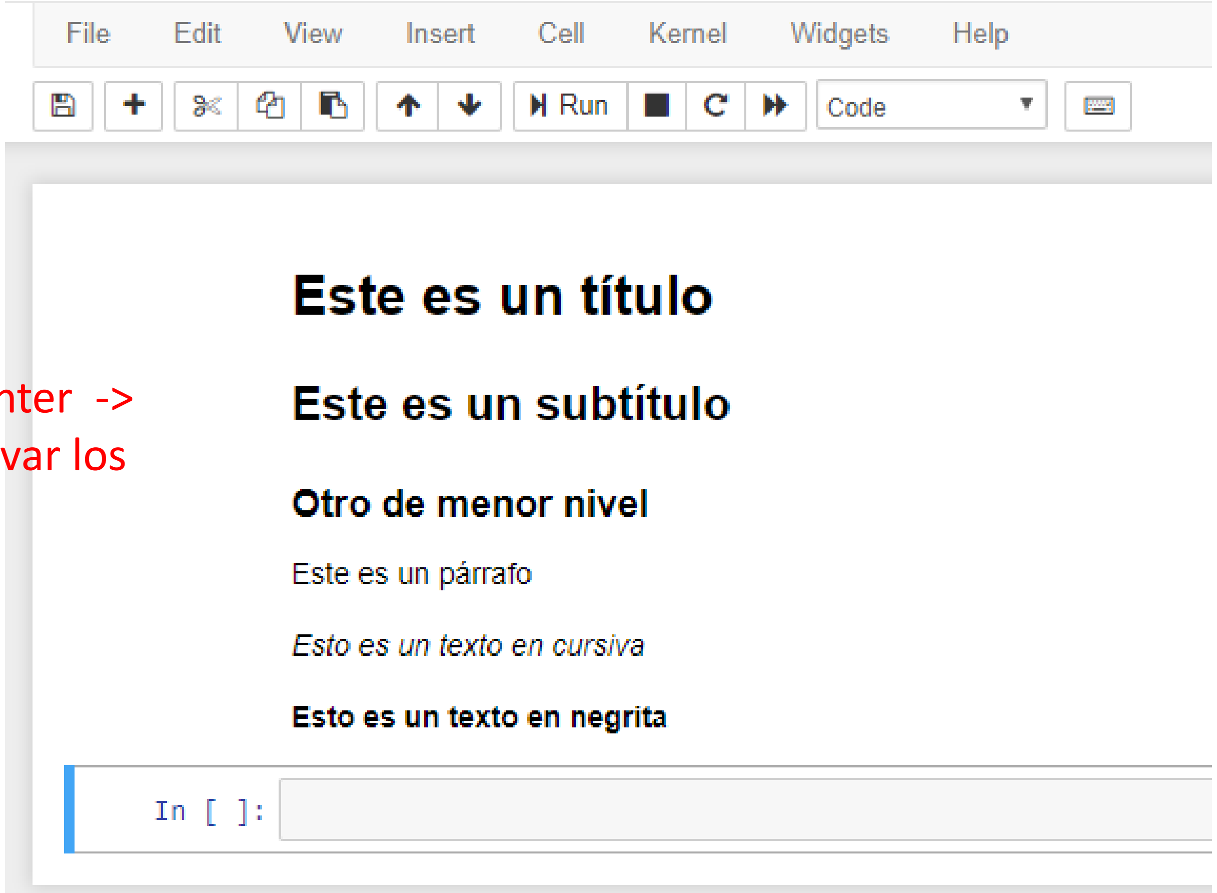
localhost:8889/notebooks/Untitled.ipynb?kernel_name=python3

Celda Markdown



```
# Este es un título

## Este es un subtítulo

### Otro de menor nivel

Este es un párrafo

*Esto es un texto en cursiva*

**Esto es un texto en negrita **
```

Mayús + Enter ->
Para observar los
resultados



Este es un título

Este es un subtítulo

Otro de menor nivel

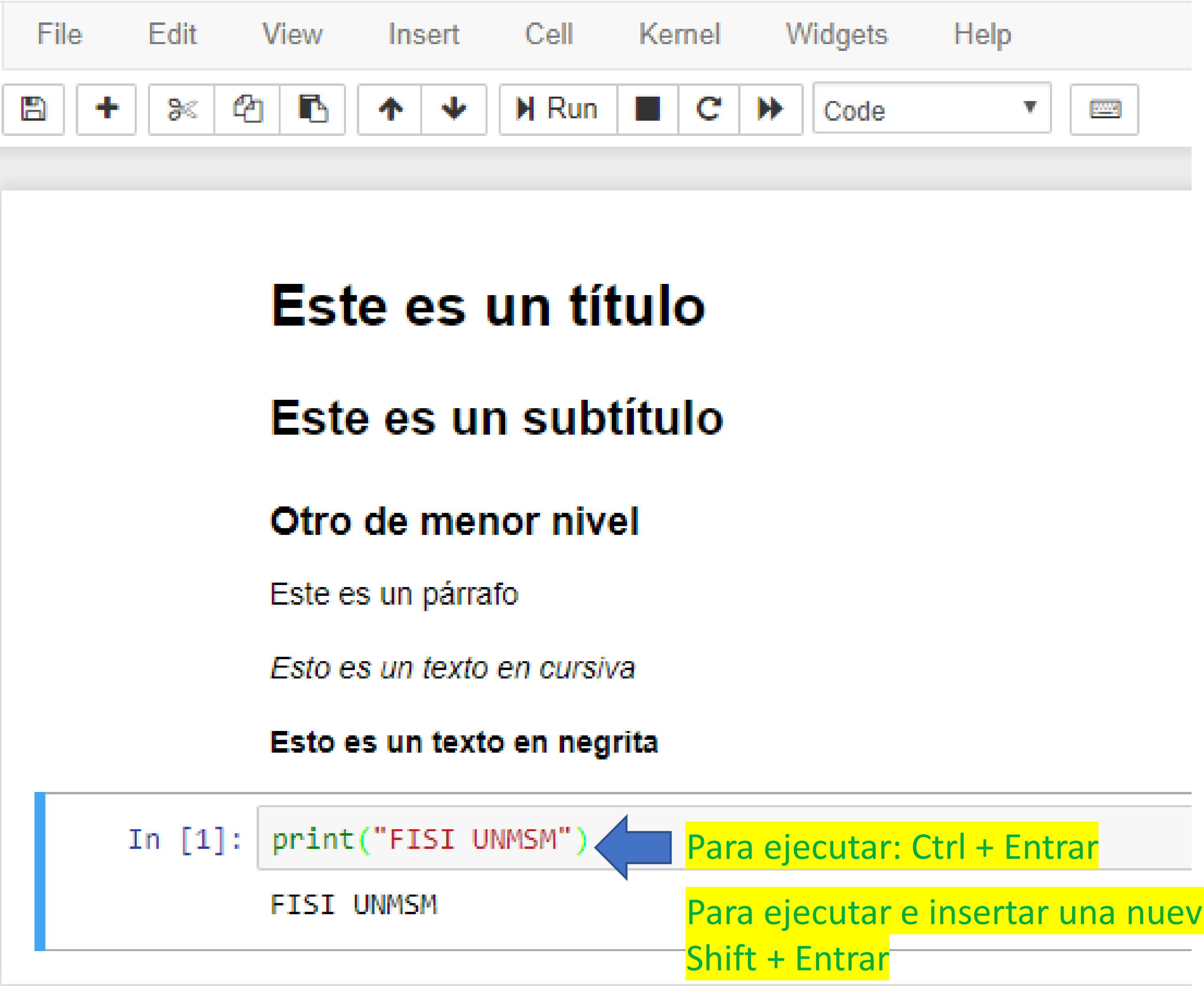
Este es un párrafo

Esto es un texto en cursiva

Esto es un texto en negrita

Celda Code

En una celda Code se puede ejecutar y probar código Python



Este es un título

Este es un subtítulo

Otro de menor nivel

Este es un párrafo

Esto es un texto en cursiva

Esto es un texto en negrita

```
In [1]: print("FISI UNMSM")
```

```
FISI UNMSM
```

Para ejecutar: Ctrl + Entrar

Para ejecutar e insertar una nueva celda: Shift + Entrar