

Data Science, Big Data y Data Analytics

Introducción

Los datos están en todas partes. De hecho, la cantidad de datos digitales que existe está creciendo a un ritmo rápido, duplicándose cada dos años y cambiando la forma en que vivimos. Según IBM, se generaron 2.500 millones de gigabytes (GB) de datos cada día en 2012.



Introducción

Un artículo de Forbes afirma que Data está creciendo más rápido que nunca antes y para el año 2020, se crearán aproximadamente 1,7 megabytes de nueva información por segundo para cada ser humano en el planeta. Lo que hace que sea extremadamente importante al menos conocer los conceptos básicos del campo. Después de todo, aquí es donde reside nuestro futuro.



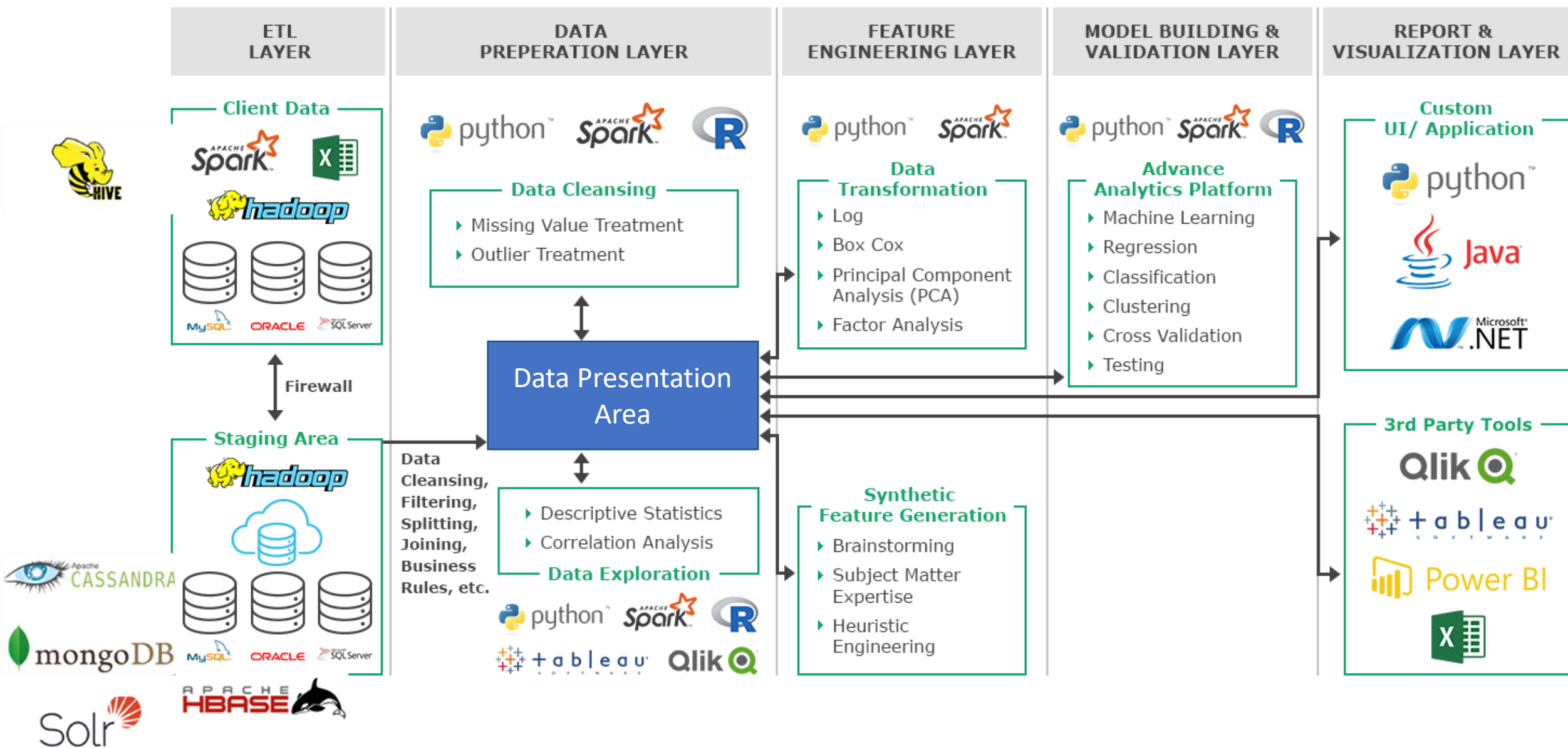
Introducción

Es importante que diferenciaremos entre Data Science, Big Data y Data Analytics, en función de qué es, dónde se utiliza, las habilidades que necesita para convertirse en un profesional en el campo y las perspectivas salariales en cada campo.

Data Science, Big Data, Data Analytics

- Data Science es la ciencia del estudio de datos.
- Big Data es un concepto teórico para definir los problemas que surgen del gran tamaño de los datos donde las herramientas tradicionales de manejo de datos no son lo suficientemente capaces.
- Data Analytics es un conjunto de herramientas y técnicas para realizar análisis de datos (grandes y pequeños).

Entonces, si tiene un problema de Big Data, usted, como Data Scientist, usaría Data Analytics para resolver esos problemas.



Data Science, Big Data, Data Analytics

Source Data

Store Data

Convert & ETL

Transform Data

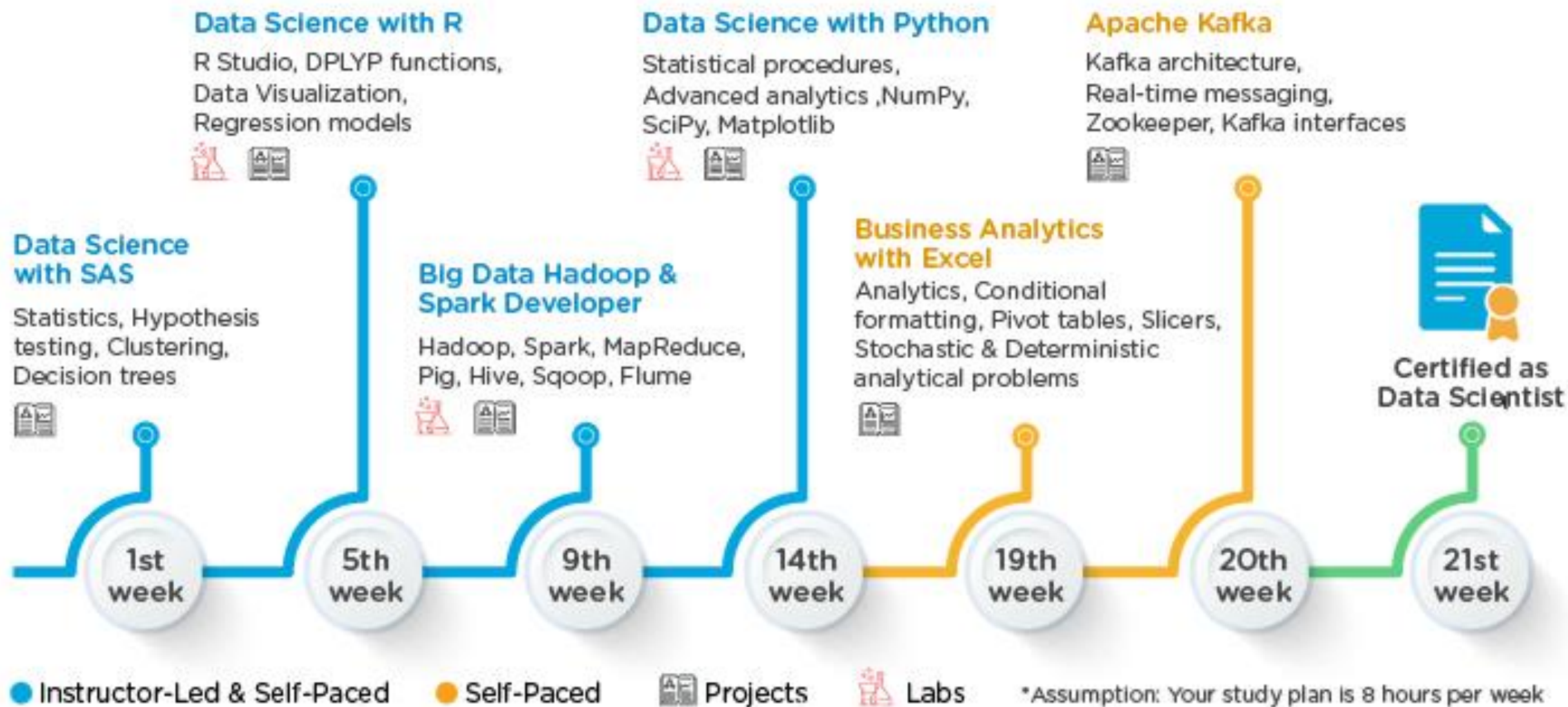
Exploratory Analysis

Model Build &
Generate Insights

Visualisation

Model Execution in
Production





Big Data Hadoop & Spark Developer

Hadoop, Spark, MapReduce, Pig, Hive, Sqoop, Flume



Apache Spark and Scala

Spark, Scala, RDD, SparkSQL, Spark Streaming, Spark ML, GraphX



Big Data & Hadoop Administrator

Scalability, Hadoop framework, Hadoop architecture, Cloudera manager, HUE, Hadoop Clusters



Apache Kafka

Kafka architecture, Real-time messaging, Zookeeper, Kafka interfaces



Apache Cassandra

Cassandra Architecture, Data Model Creation, Database Interfaces, Advanced Architecture

MongoDB Developer & Administrator

Replication, Sharding, Indexes, DB Notes, Master-Slave concepts, Replica Set

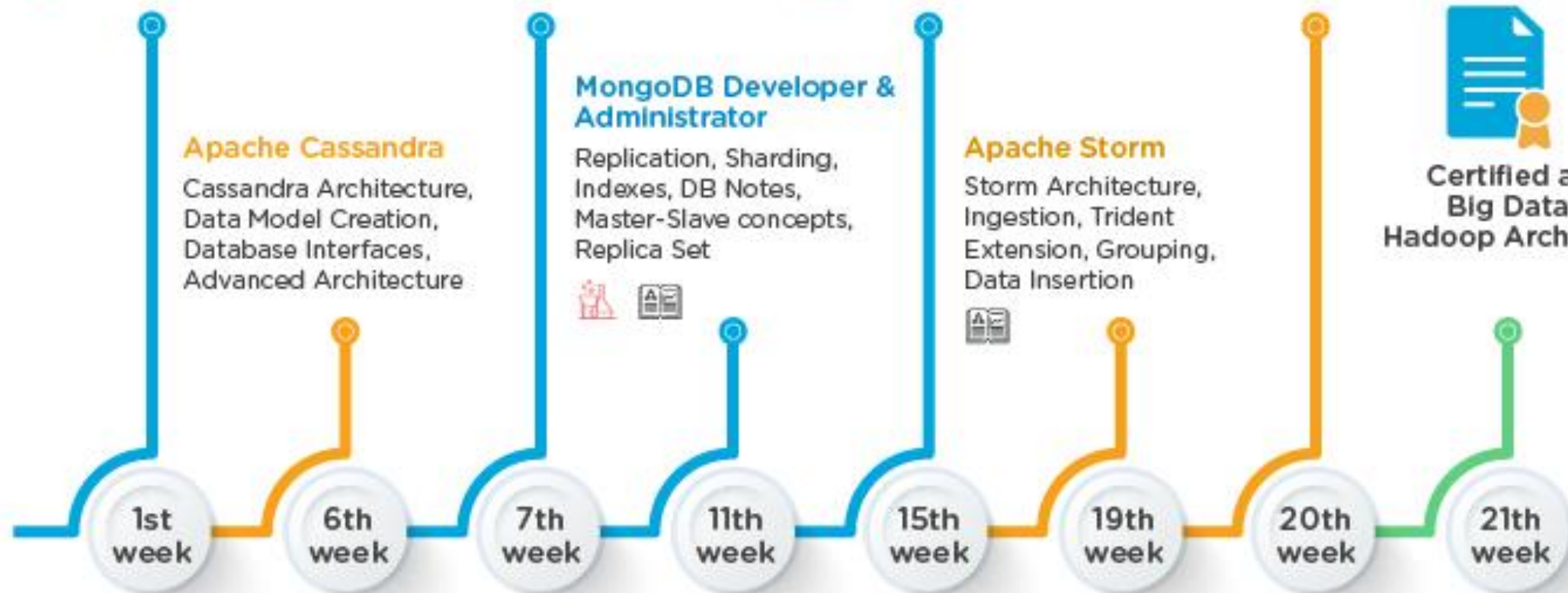


Apache Storm

Storm Architecture, Ingestion, Trident Extension, Grouping, Data Insertion



**Certified as
Big Data
Hadoop Architect**



● Instructor-Led & Self-Paced

● Self-Paced

📖 Projects

🧪 Labs

*Assumption: Your study plan is 8 hours per week

Business analytics with excel

Analytics, Conditional formatting, Pivot tables, Slicers, Stochastic & Deterministic analytical problems



Data Science with R

R Studio, DPLYR functions, Data Visualization, Regression models



Tableau Desktop 10 Qualified Associate Training

Data Blending, Data Extracts, Ad-hoc analytics, Heat map, Tree map, Waterfall, Pareto, etc



Data Science with Python

Statistical procedures, Advanced analytics, NumPy, SciPy, Matplotlib



Data Science with SAS training

Statistics, Hypothesis testing, Clustering, Decision trees



Certified as Business Analytics Expert

1st week

2nd week

6th week

10th week

15th week

19th week

● Instructor-Led & Self-Paced

● Instructor-Led

● Self-Paced

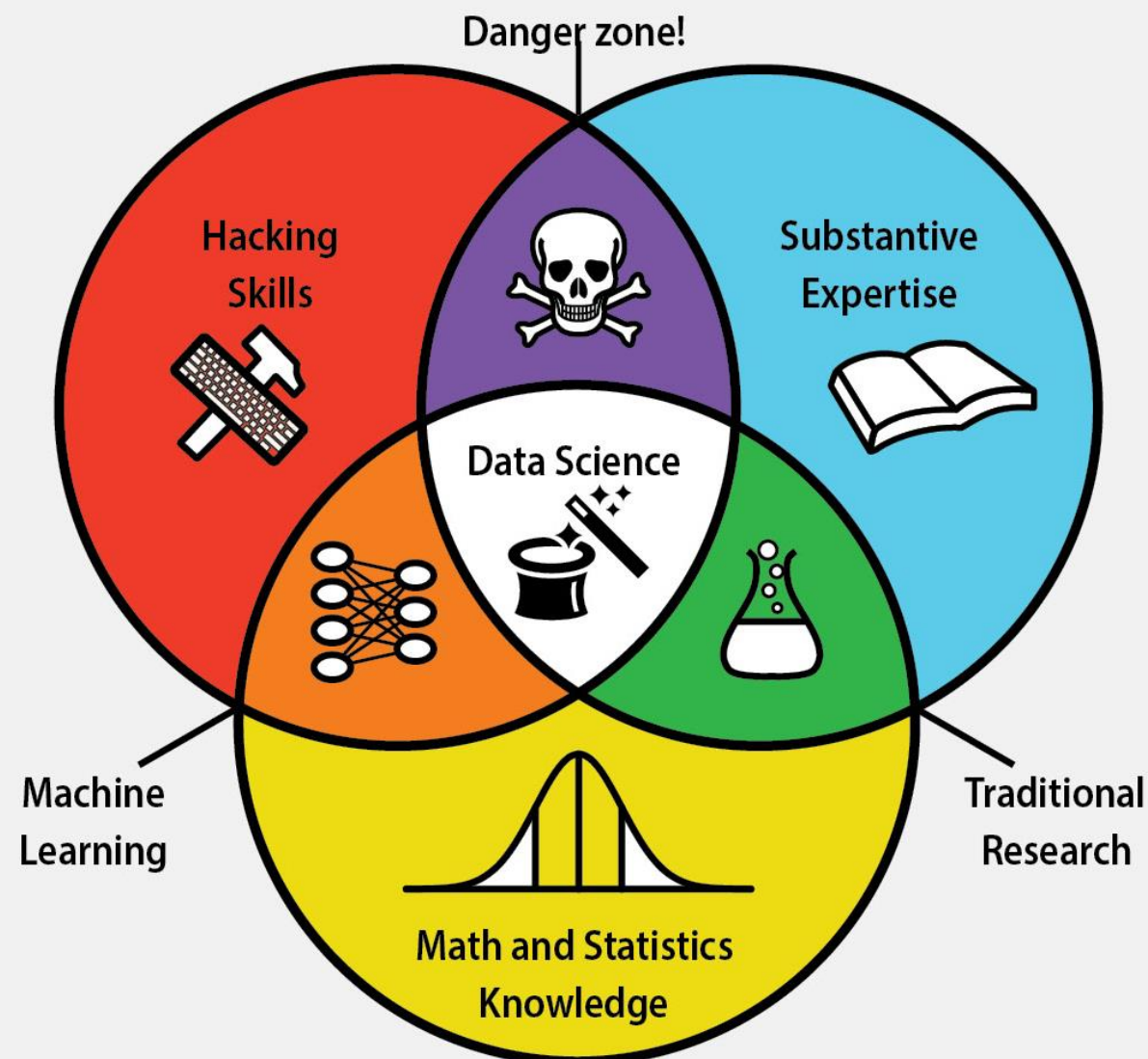
Projects

Labs

*Assumption: Your study plan is 8 hours per week

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.

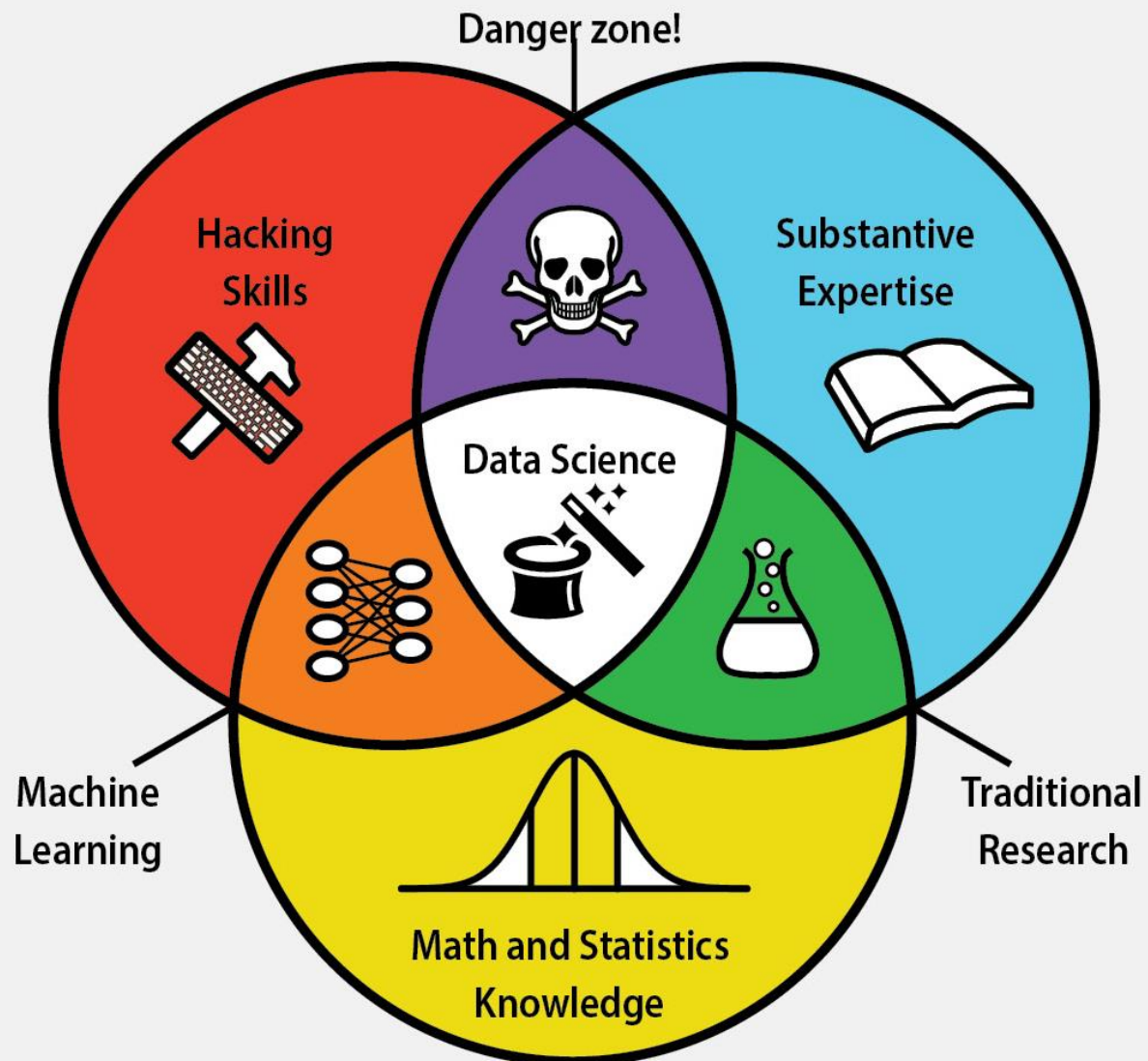


Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



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.

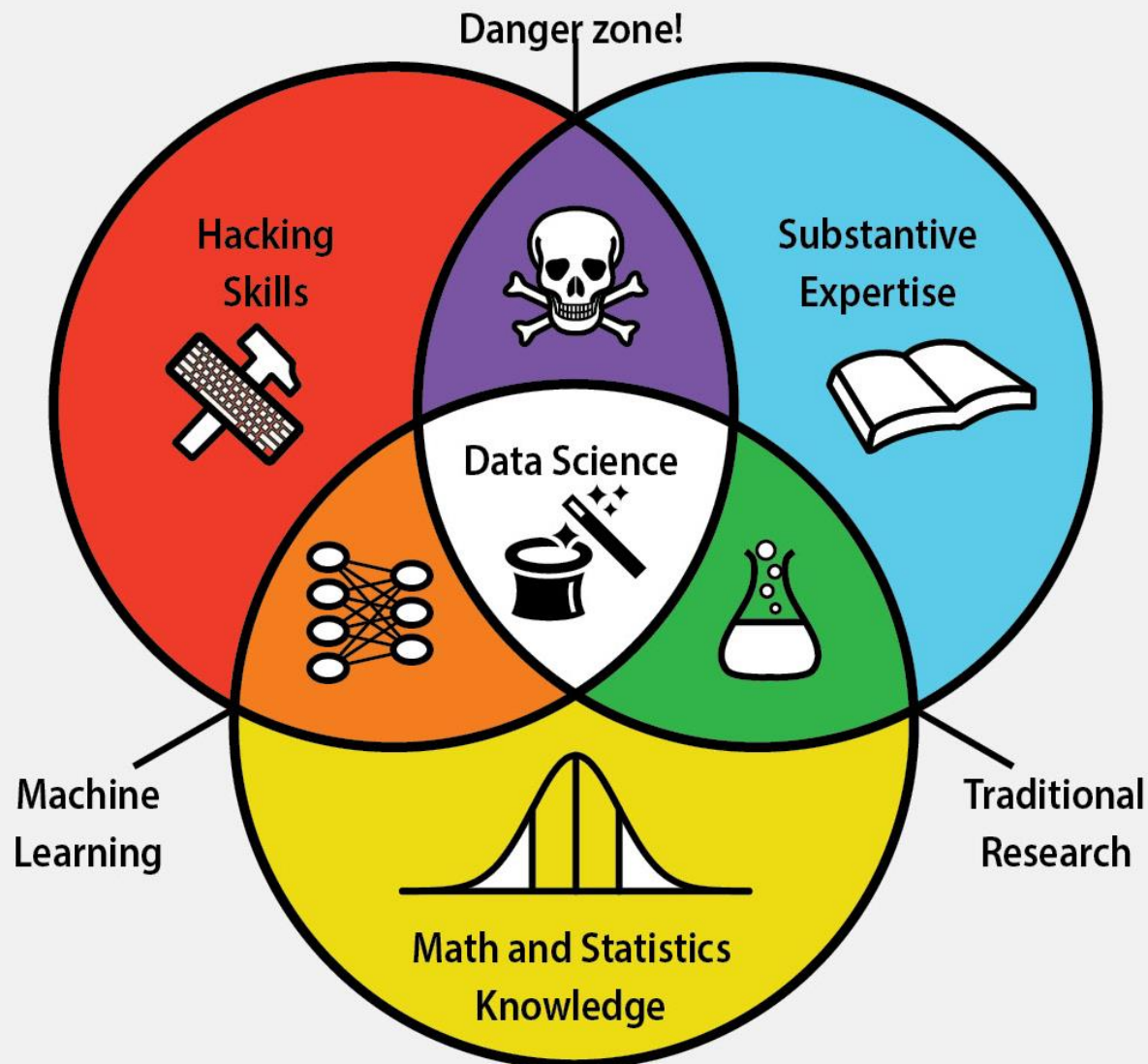


Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



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.

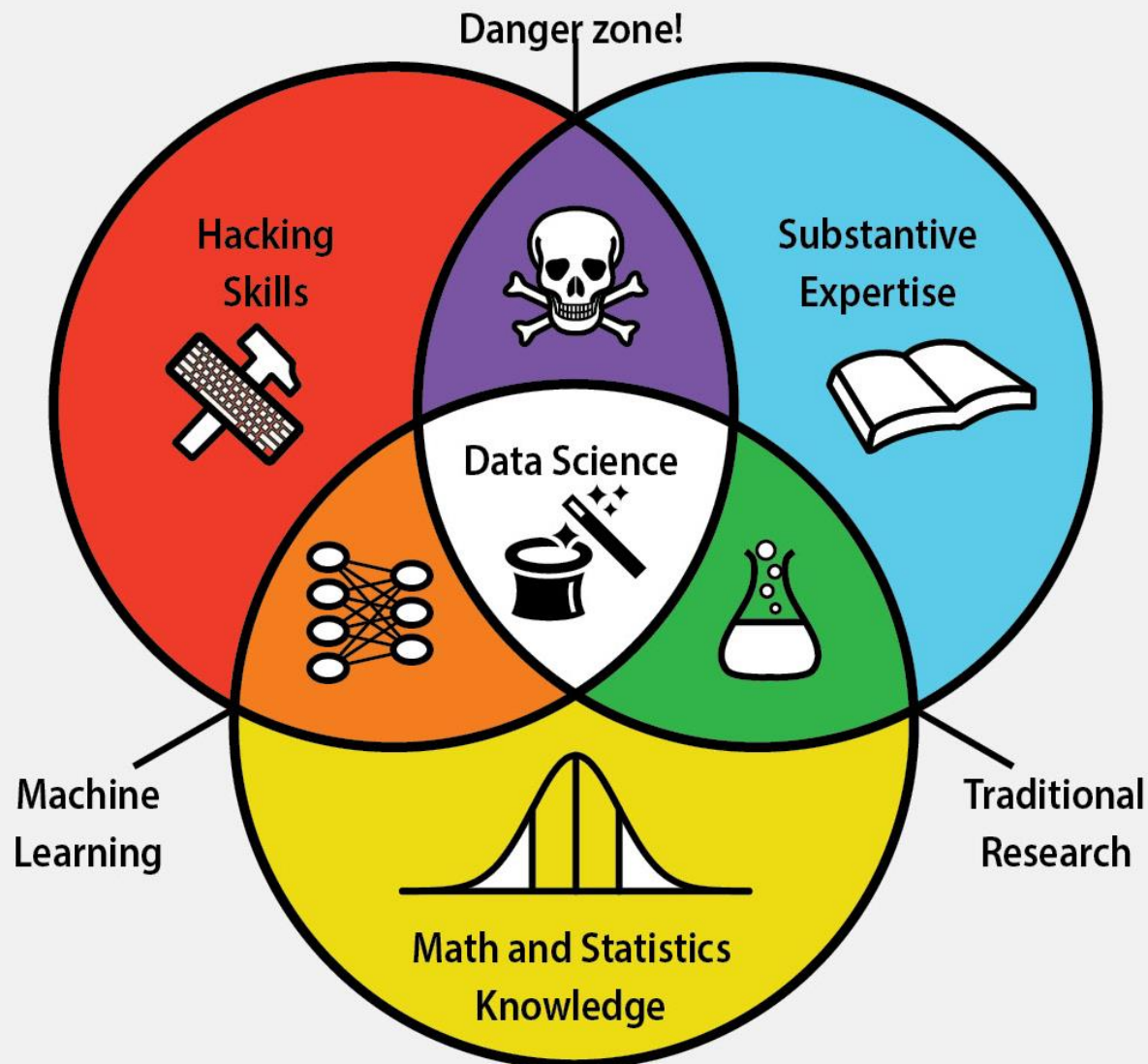


Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



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.

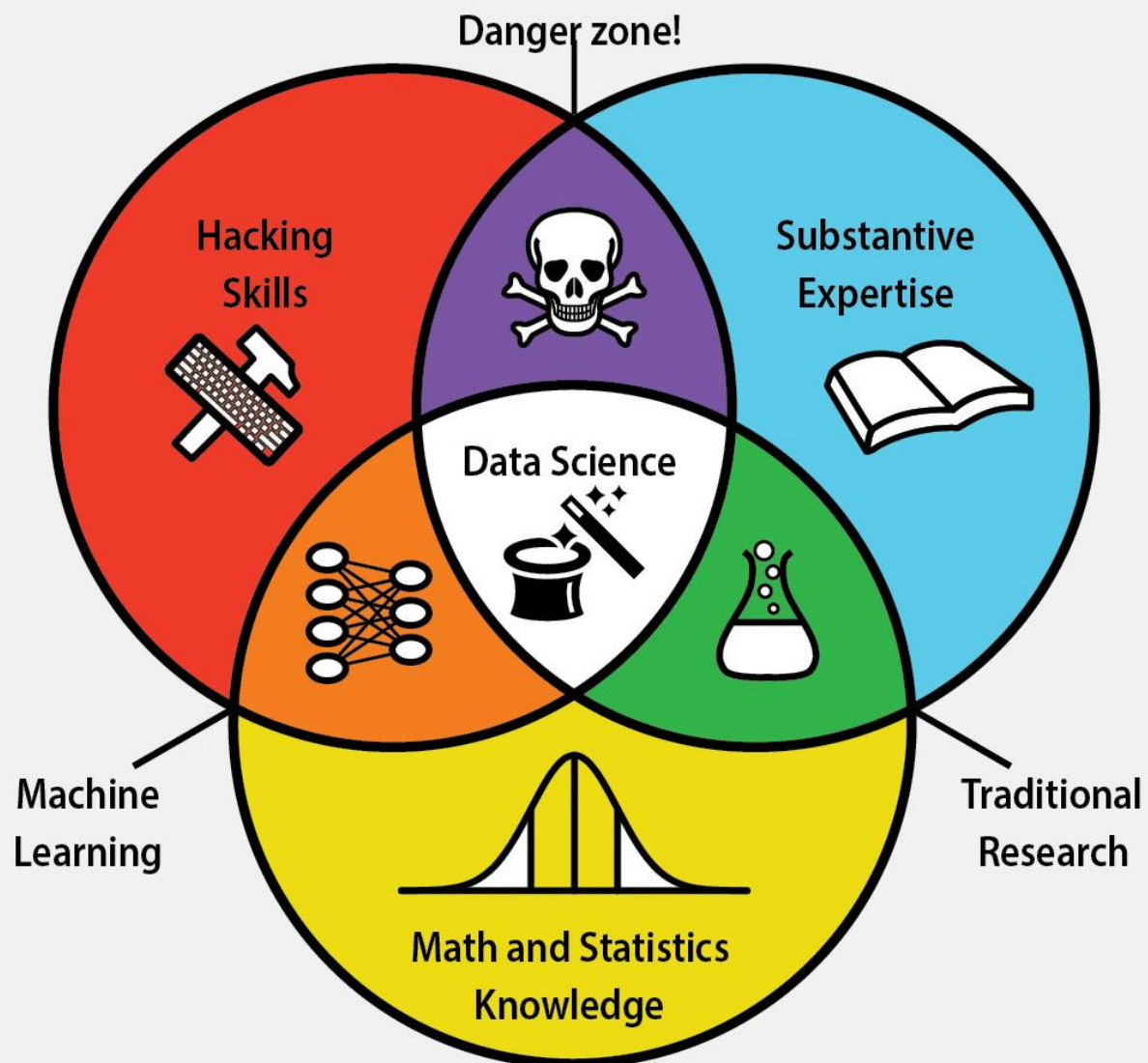


Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



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.

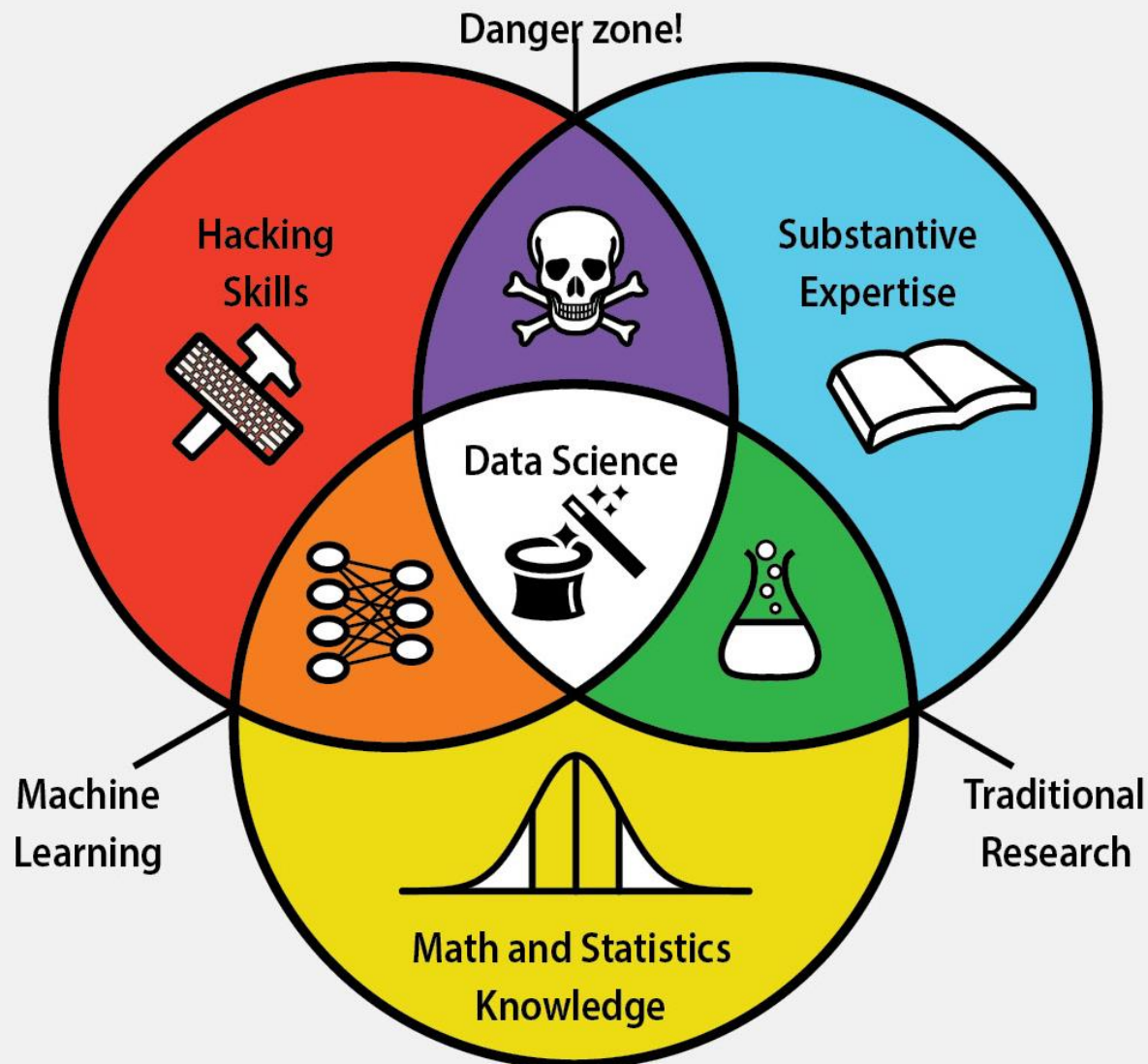


Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



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.

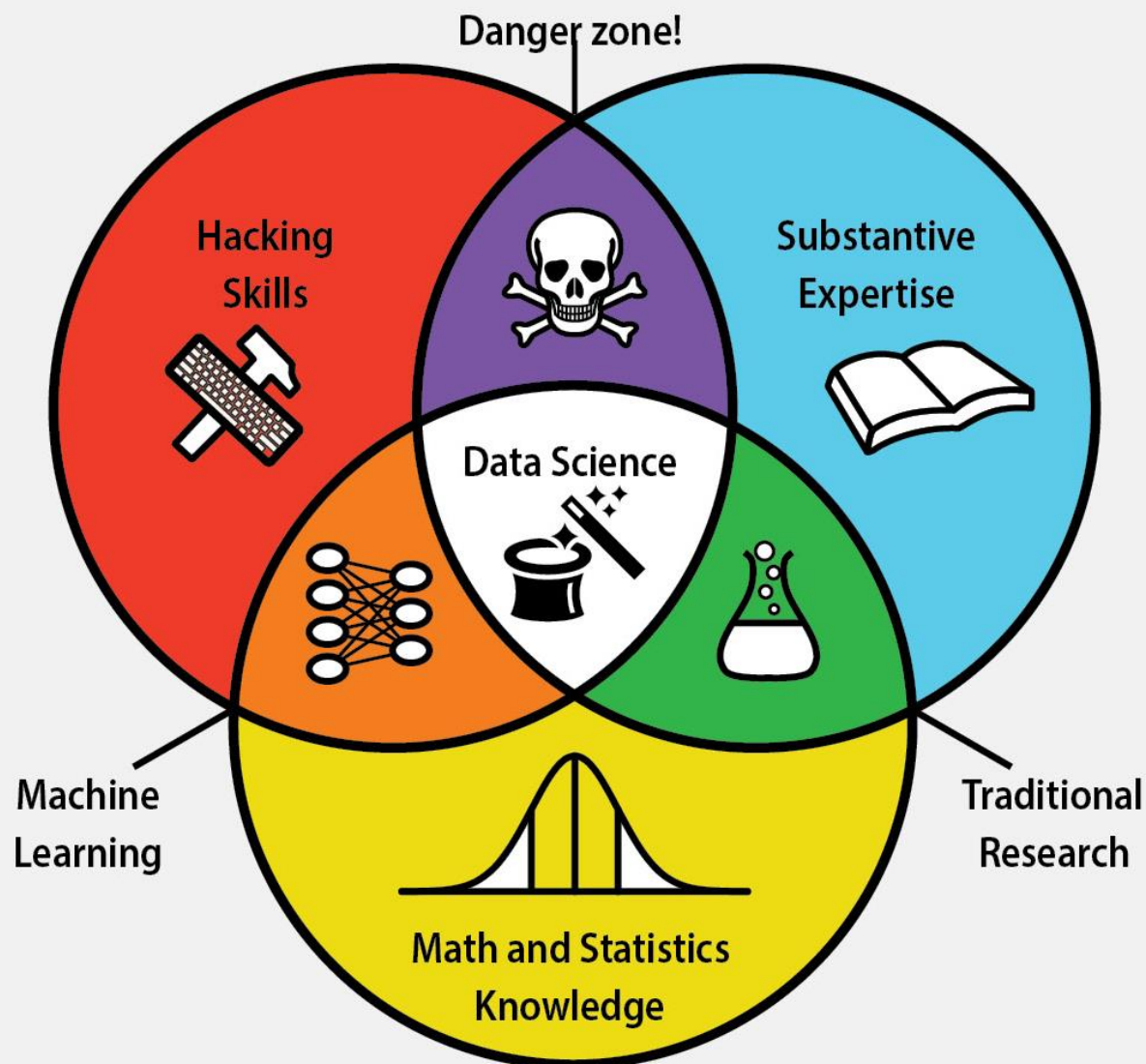


Machine learning stems from combining hacking skills with math and statistics knowledge, but does not require scientific motivation.



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.

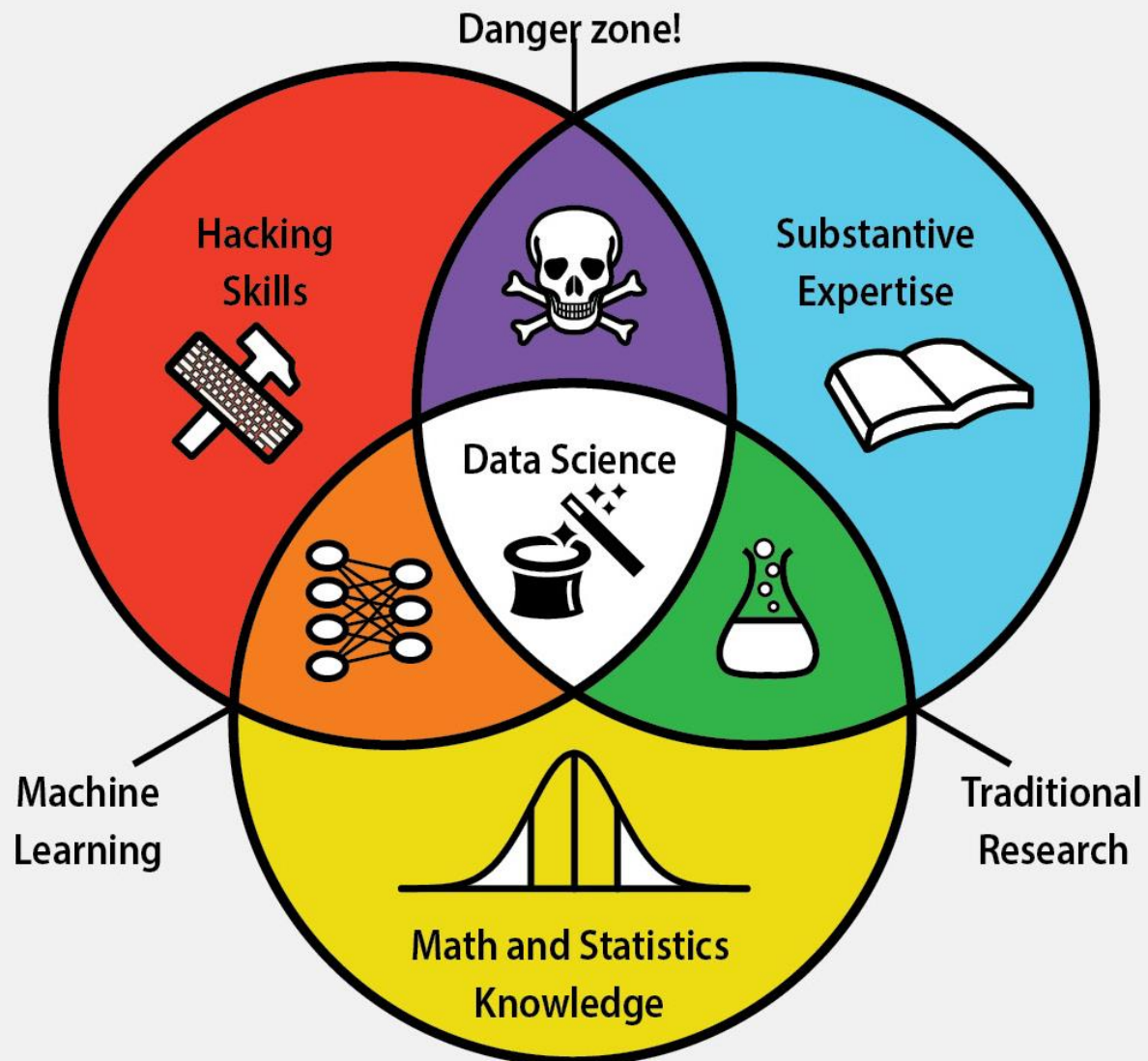


El **aprendizaje automático** se deriva de la combinación de las habilidades de hacking con las matemáticas y el conocimiento estadístico, pero no requiere motivación científica.



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.

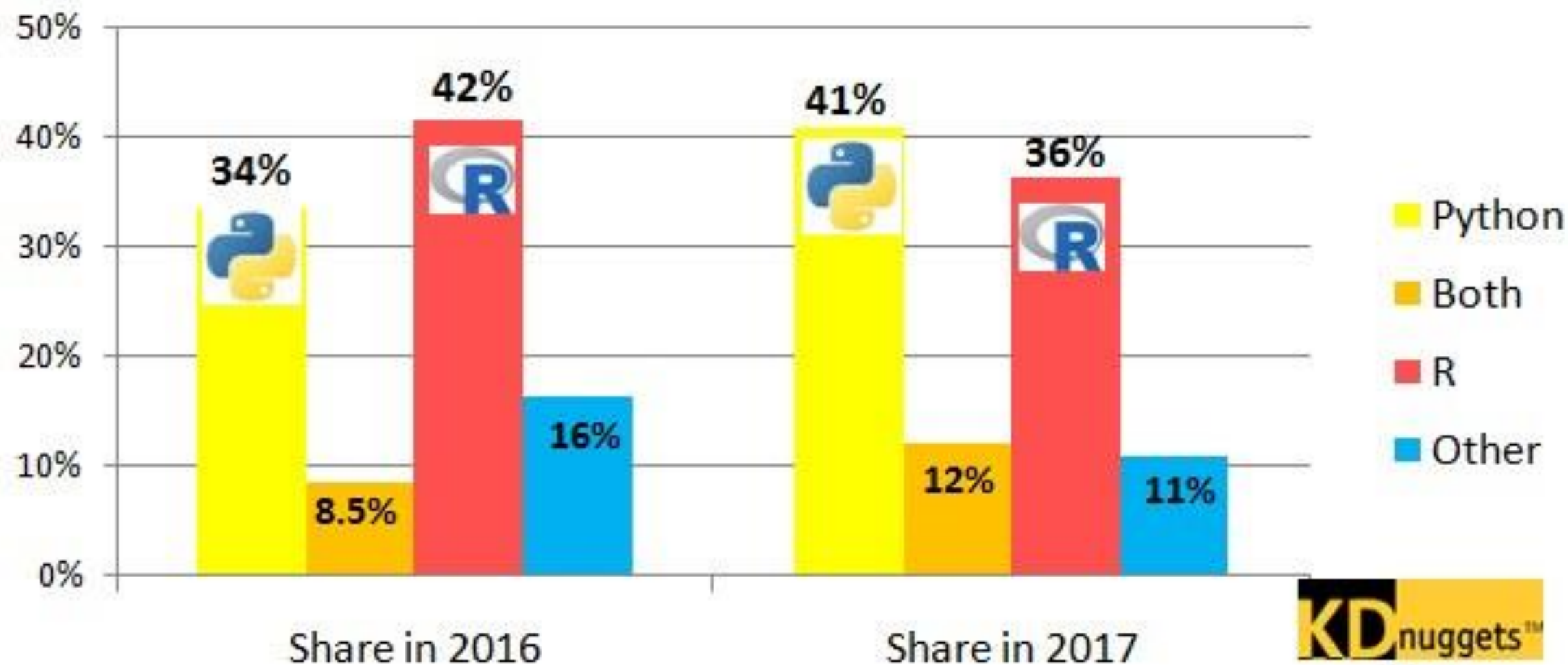


El **aprendizaje automático** se deriva de la combinación de las habilidades de hacking con las matemáticas y el conocimiento estadístico, pero no requiere motivación científica.



¡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



What does a data scientist do?



Raw Data

Processing

Dataset

Statistical
Models / Analysis

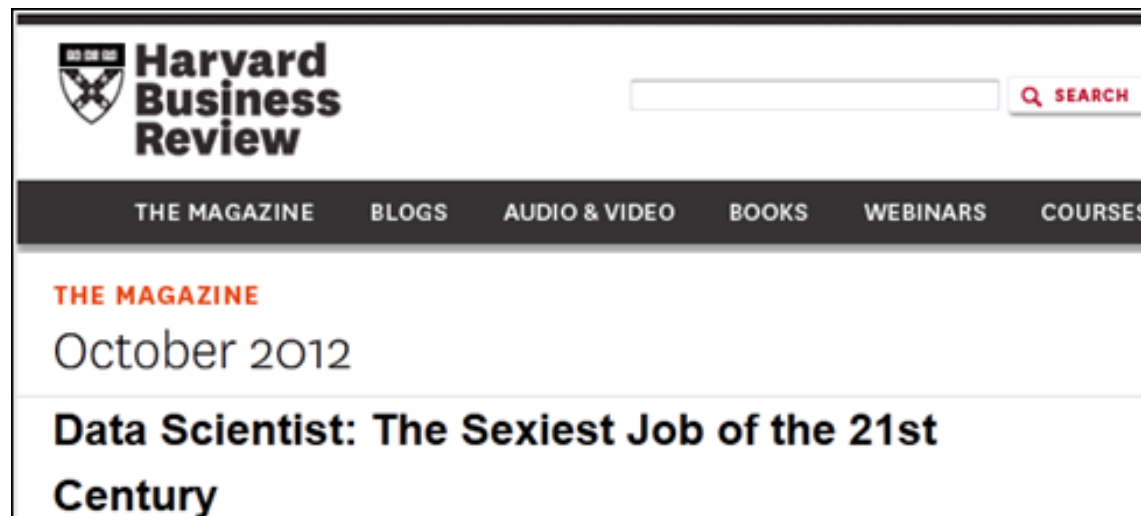
Machine Learning
Predictions

Data driven
Products

Reports
Visualization
Blogs


Data Scientist

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



Librerías mas populares para ciencias de datos en Python

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



NumPy

[Scipy.org](https://scipy.org)

NumPy

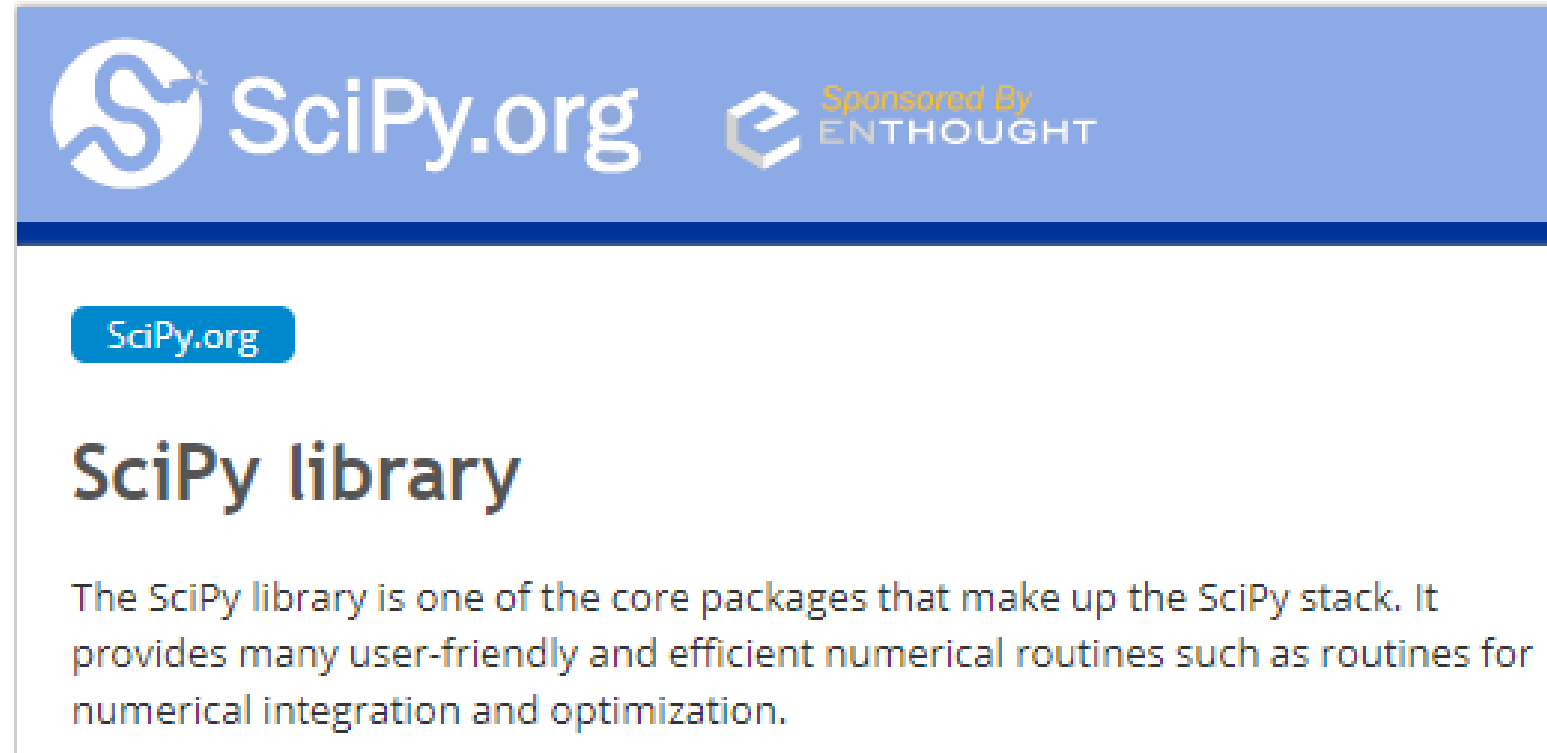
NumPy is the fundamental package for scientific computing with Python. It contains among other things:

- 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

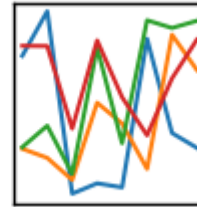
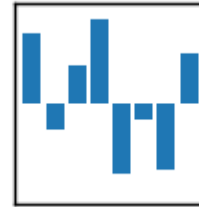


Librerías mas populares para ciencias de datos en Python

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

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



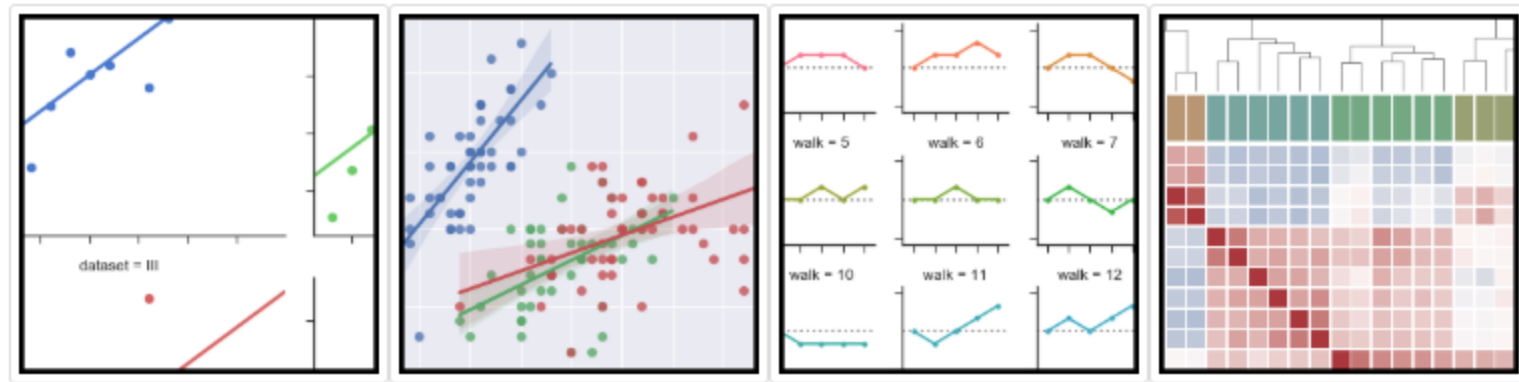
Python Data Analysis
Library

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

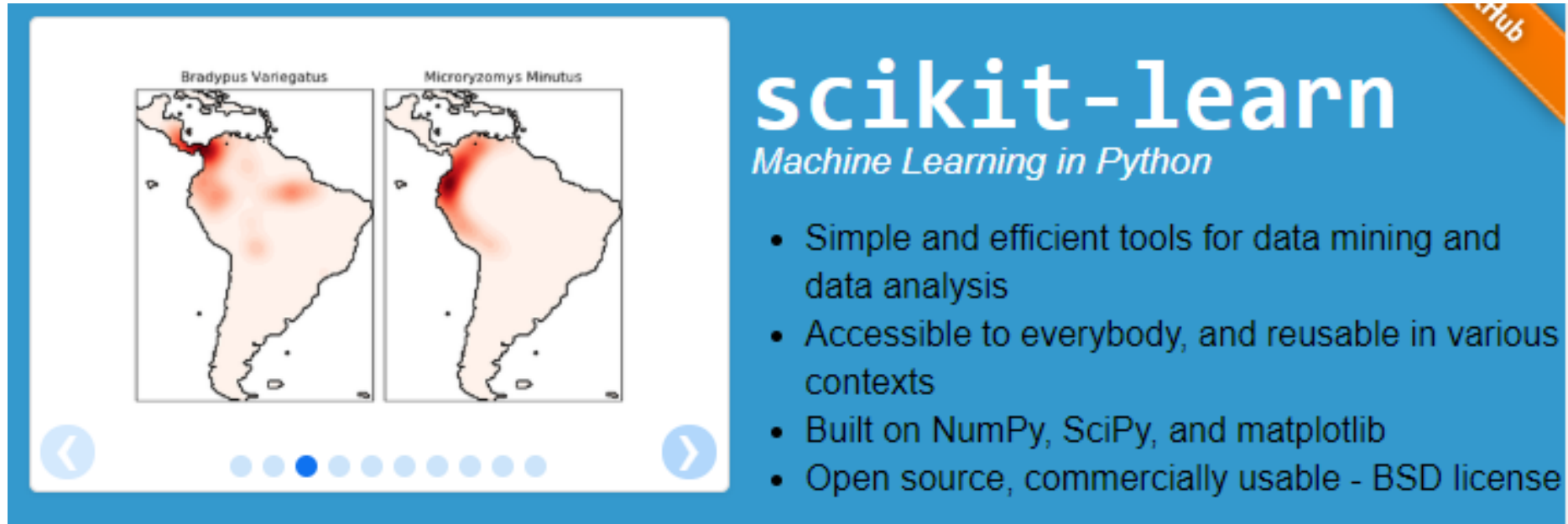
seaborn: statistical data visualization



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



The slide is divided into two main sections. The left section contains two maps of South America, each showing the distribution of a different species. The first map is titled "Bradypus Variegatus" and the second is titled "Microryzomys Minutus". Both maps show a high concentration of the species in the northern and central regions of South America, with a color gradient from light orange to dark red indicating the density. Below the maps are navigation arrows and a series of dots, with the third dot from the left being highlighted in blue. The right section features the "scikit-learn" logo in white text on a blue background, with the tagline "Machine Learning in Python" below it. A list of features is provided in white text on the blue background.

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



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

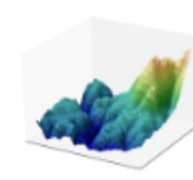
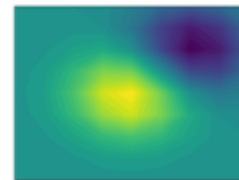
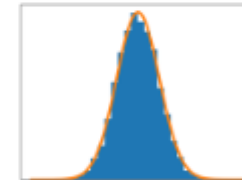
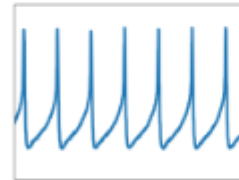
Librerías mas populares para ciencias de datos en Python

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



[home](#) | [examples](#) | [tutorials](#) | [pyplot](#) | [docs](#) »

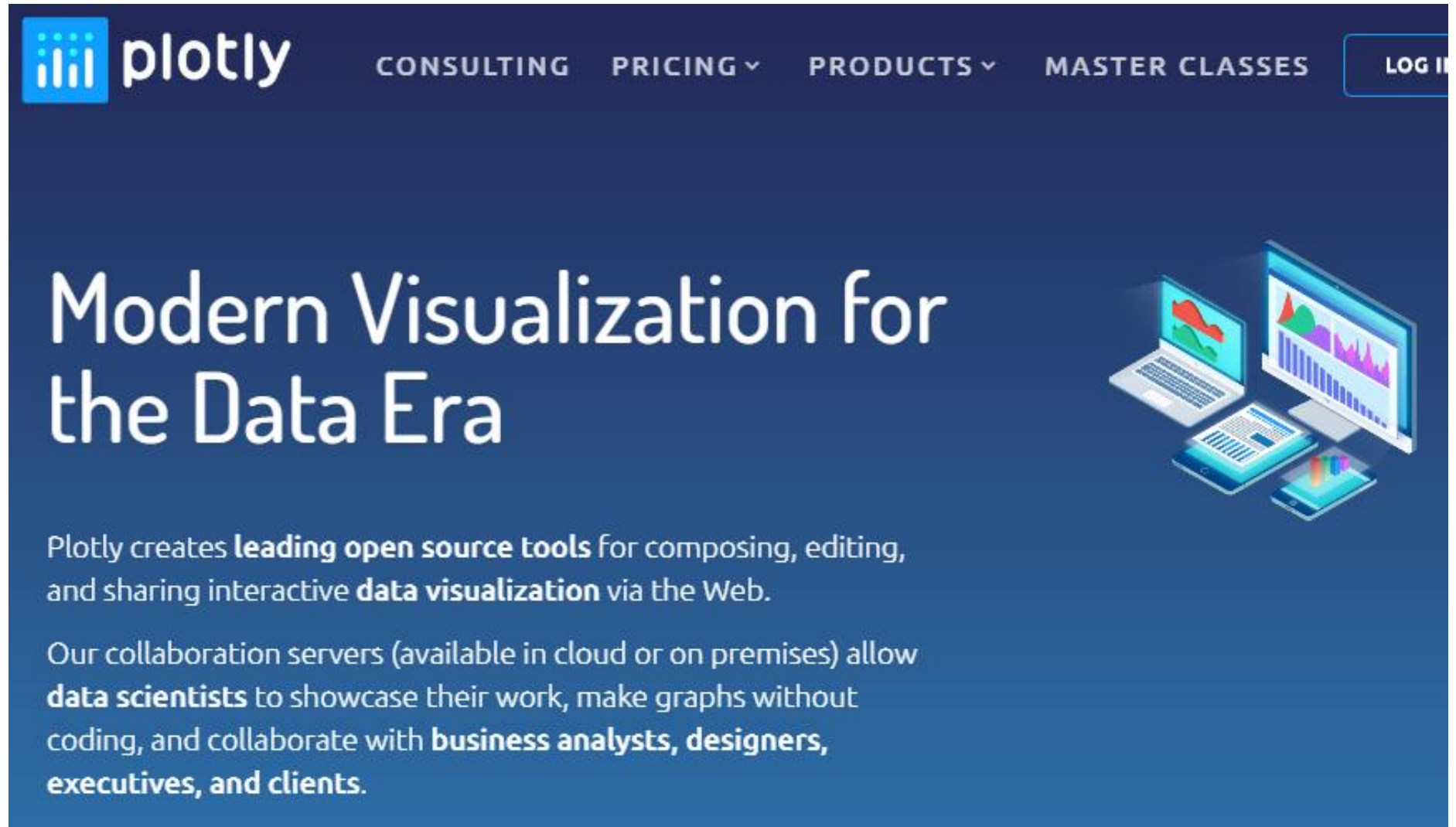
Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical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

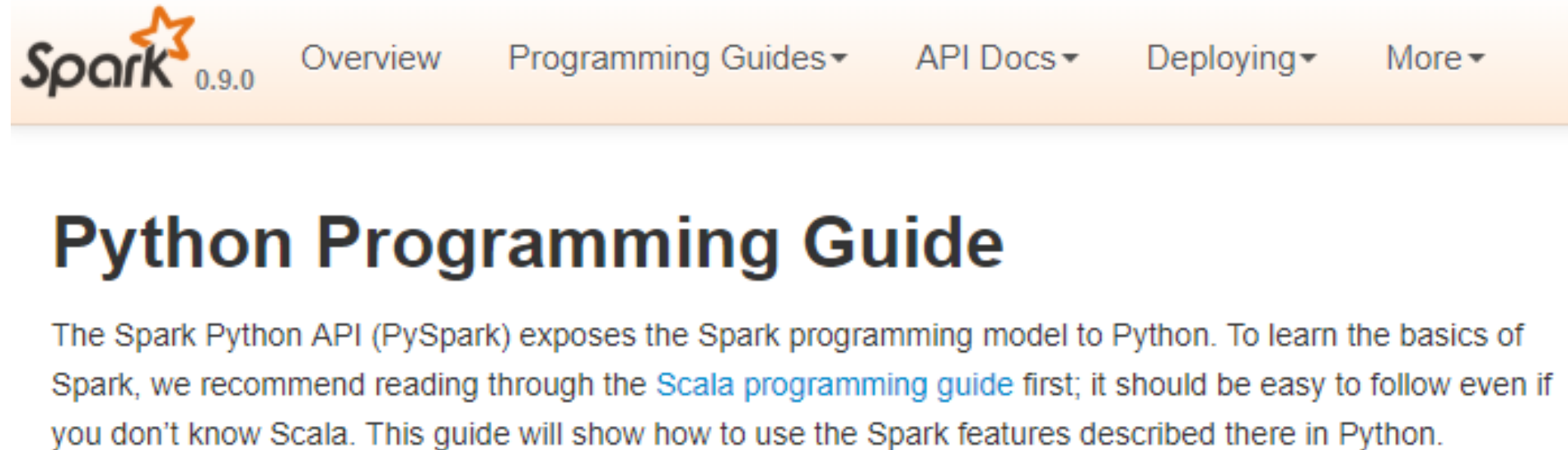
Librerías mas populares para ciencias de datos en Python

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

The image shows the Plotly website banner. At the top left is the Plotly logo, which consists of a blue square with three white vertical bars of increasing height, followed by the word "plotly" in white lowercase letters. To the right of the logo are navigation links: "CONSULTING", "PRICING", "PRODUCTS", and "MASTER CLASSES", each followed by a small downward arrow. Further right is a "LOG IN" button. The main heading "Modern Visualization for the Data Era" is in large white text. Below it is a paragraph: "Plotly creates **leading open source tools** for composing, editing, and sharing interactive **data visualization** via the Web." Below that is another paragraph: "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.**" On the right side of the banner is an illustration of a laptop, a desktop monitor, and two smartphones, all displaying various data visualizations like bar charts, line graphs, and area charts.

Librerías mas populares para ciencias de datos en Python

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



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

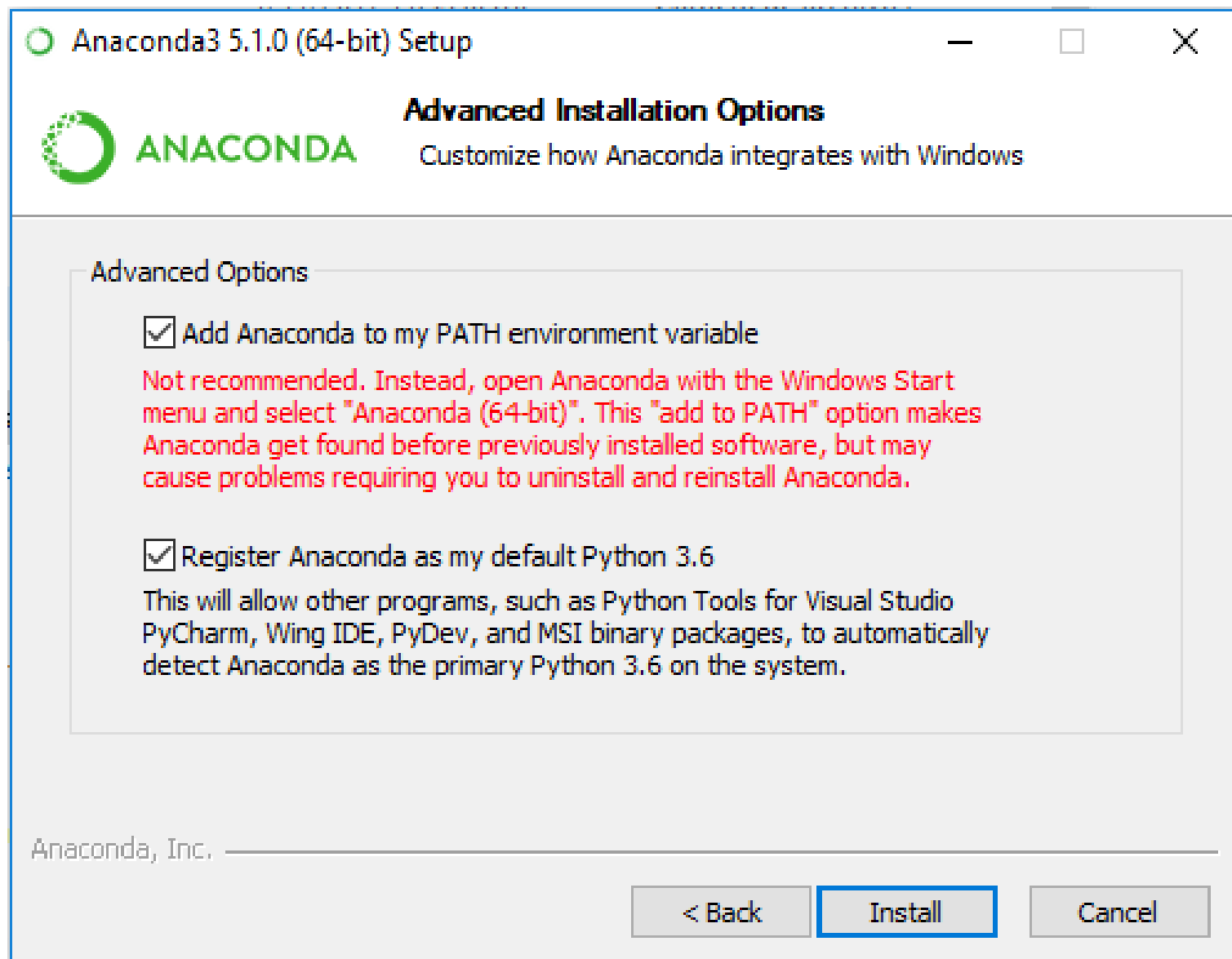
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

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





Applications on

base (root)

Channels

Refresh



jupyterlab

0.31.4

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

Launch



notebook

5.4.0

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

Launch



qtconsole

4.3.1

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

Launch



spyder

3.2.6

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

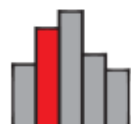
Launch



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

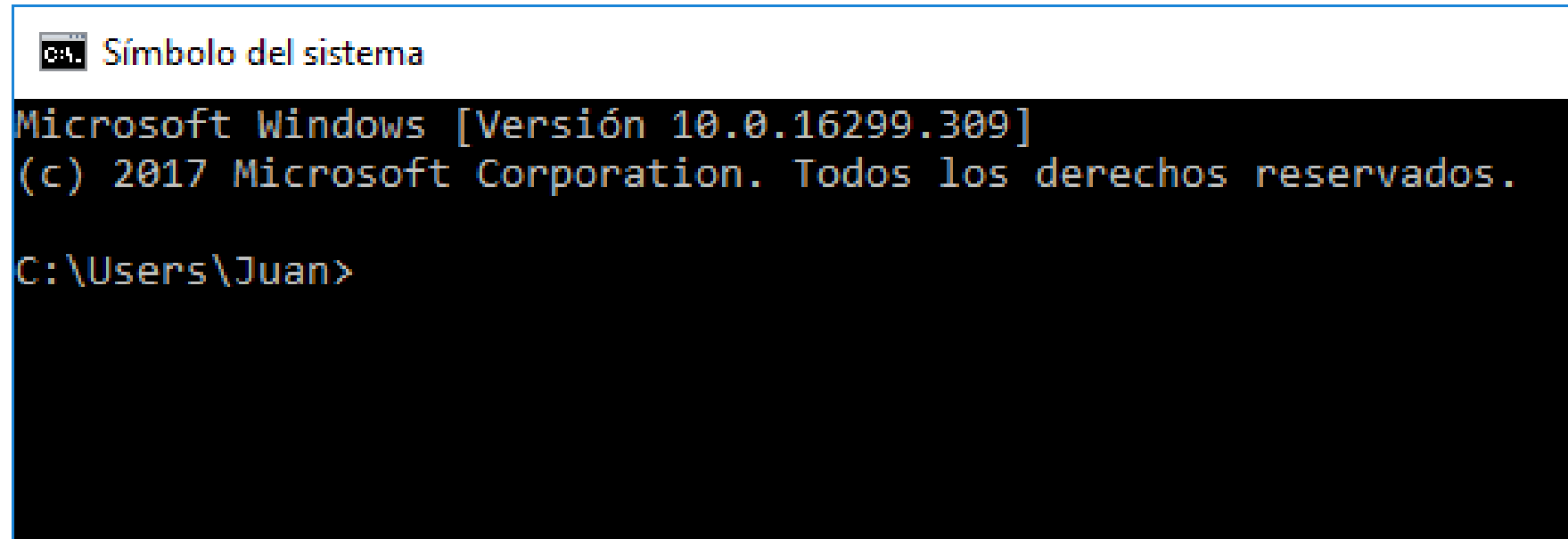


rstudio

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


A screenshot of a Windows Command Prompt window titled "Símbolo del sistema". The window has a black background with white text. The text displayed is: "Microsoft Windows [Versión 10.0.16299.309]" followed by "(c) 2017 Microsoft Corporation. Todos los derechos reservados." on the next line. The prompt "C:\Users\Juan>" is shown on the third line, indicating the current directory and user.

```
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:

Cambiar a la
carpeta
correspondiente

Este equipo ➤ OS (C:) ➤ CursoML

 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

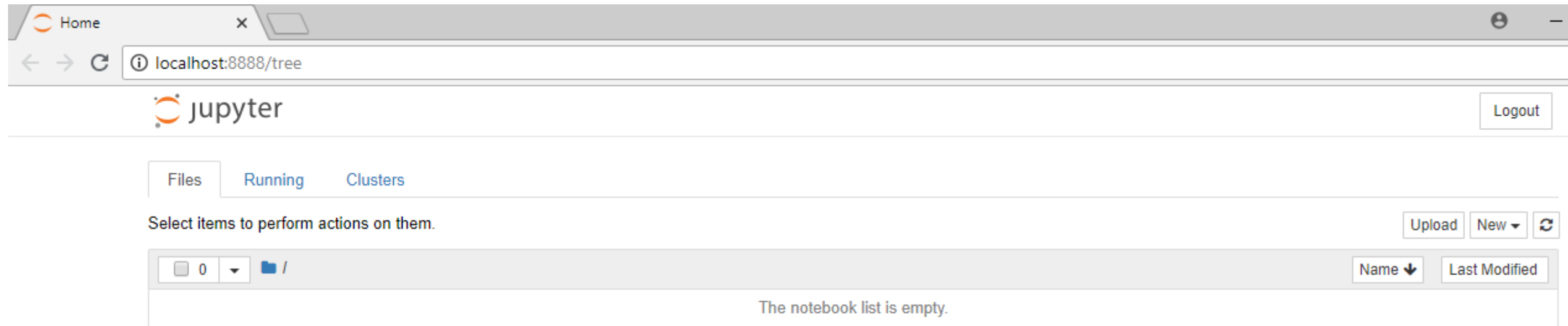
C:\CursoML>jupyter notebook
[I 23:13:18.960 NotebookApp] JupyterLab beta preview extension loading
JupyterLab
[I 23:13:18.961 NotebookApp] JupyterLab application directory is
[W 23:13:19.074 NotebookApp] Error loading server extension jupyterlab
Traceback (most recent call last):
```




pythonTM



Obtenemos:



The screenshot displays the JupyterLab web interface in a browser window. The browser's address bar shows the URL `localhost:8888/tree`. The JupyterLab header includes the logo, the text "jupyter", and a "Logout" button. Below the header, there are three tabs: "Files" (selected), "Running", and "Clusters". A message states "Select items to perform actions on them." To the right of this message are buttons for "Upload", "New" (with a dropdown arrow), and a refresh icon. Below the message, a file browser shows the root directory "/" with a count of "0" items. To the right of the file browser are two columns: "Name" with a downward arrow and "Last Modified". The main content area displays the message "The notebook list is empty."

Home x

localhost:8888/tree

jupyter Logout

Files Running Clusters

Select items to perform actions on them.

Upload New ↕ ↻

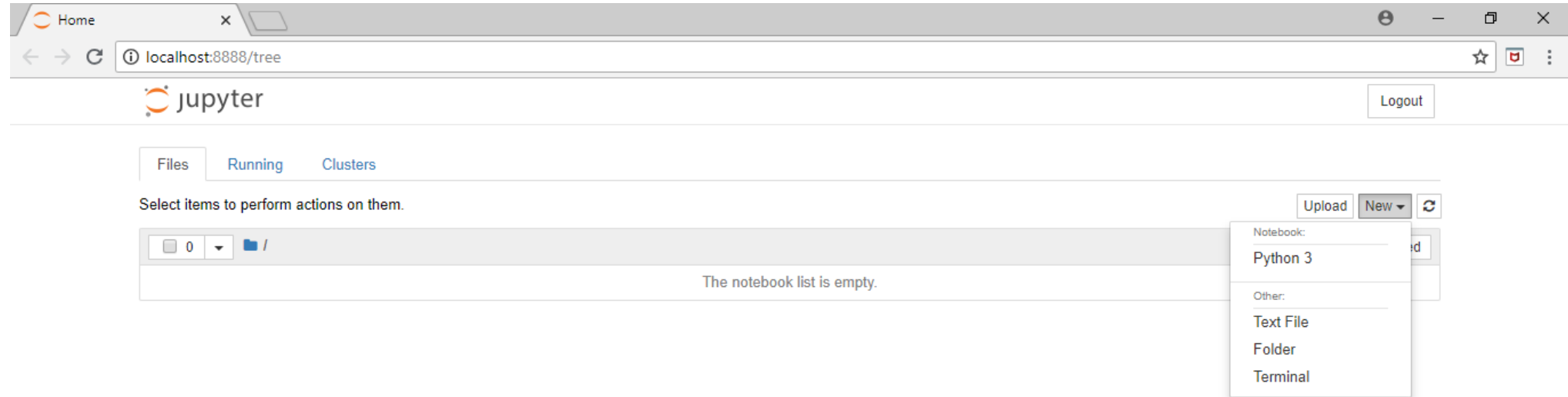
0 ▾ /

Name ↓ Last Modified

The notebook list is empty.

Para crear un block de notas

Se hace
clic en
New y se
elige
Python 3



El block de notas

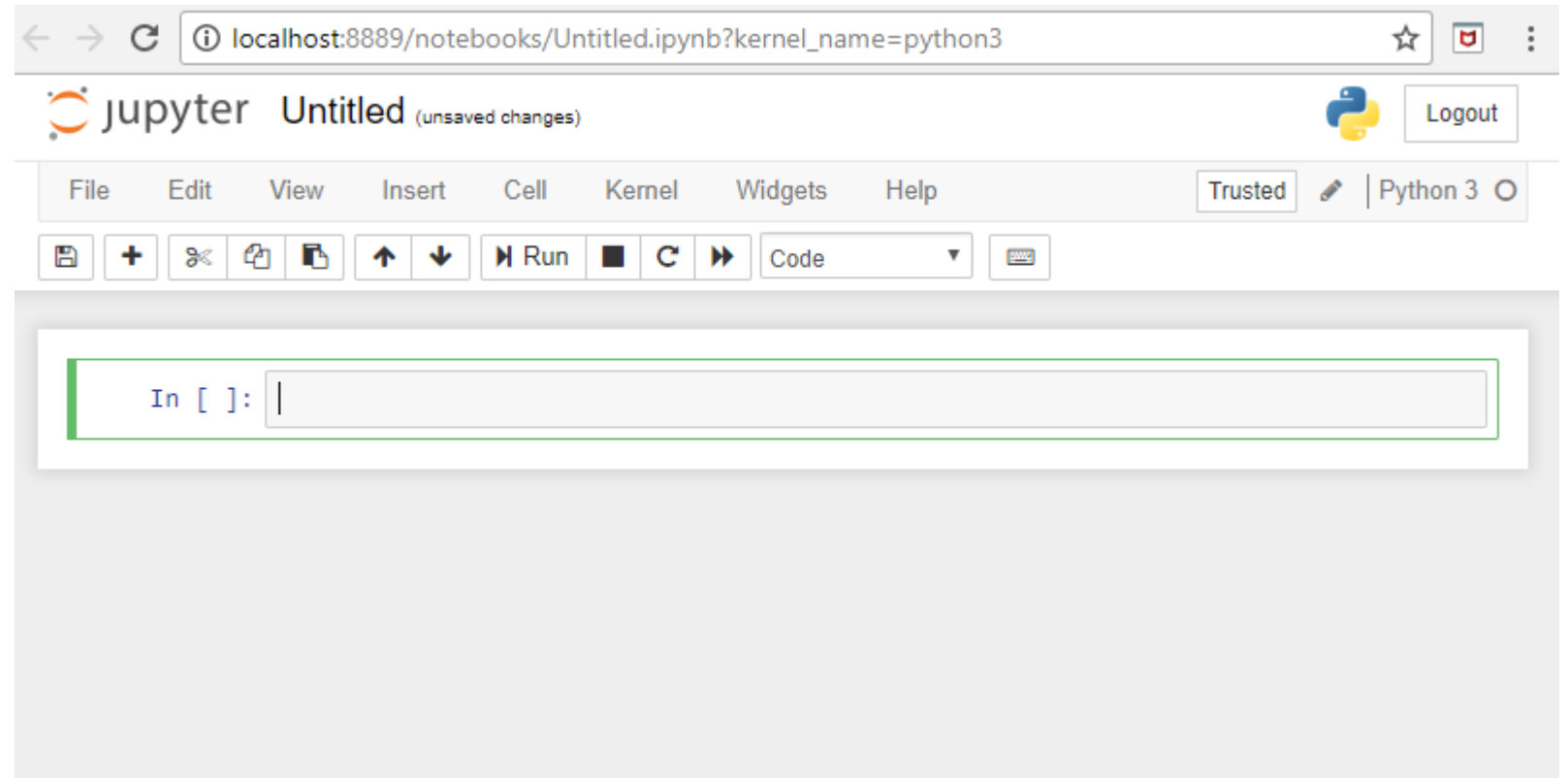
En el block de notas
tenemos distintos
tipos de celdas
como:

Code

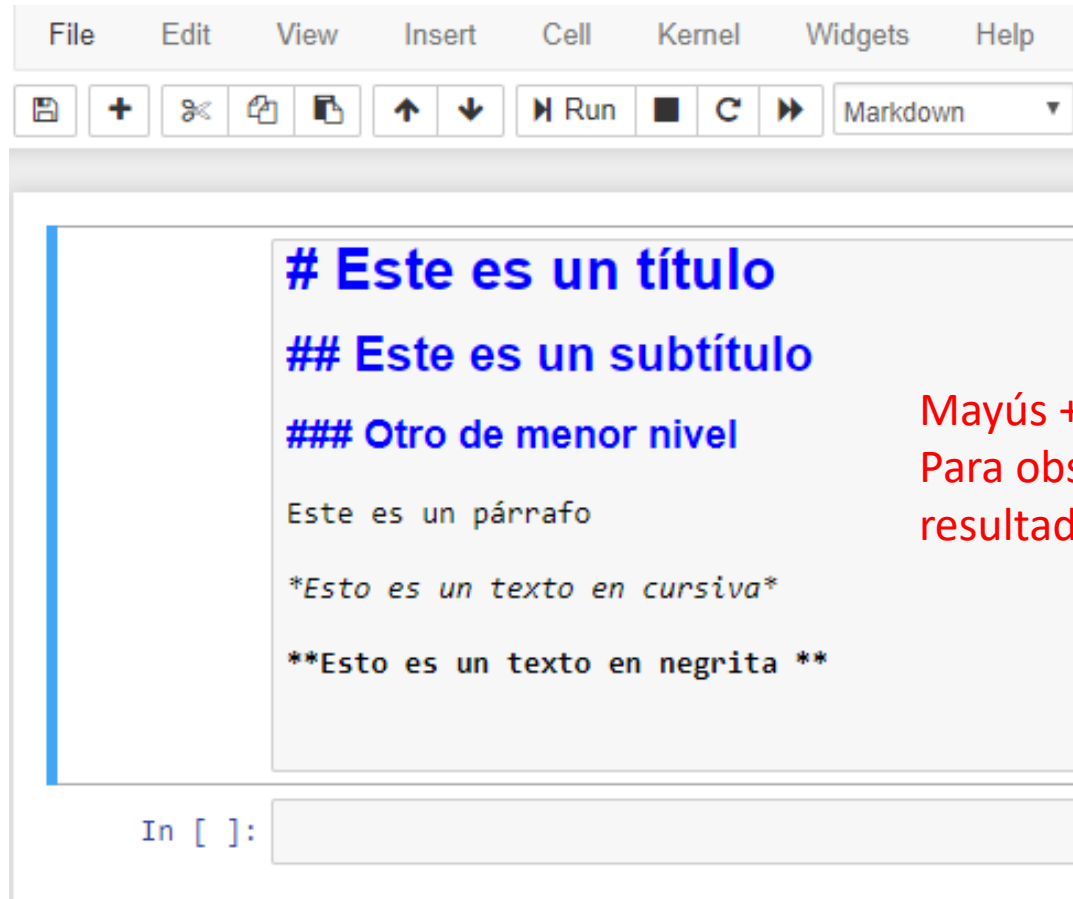
Markdown

Raw NBConvert

Heading



Celda Markdown



The image shows a Jupyter Notebook interface with a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for saving, adding, deleting, copying, pasting, undo, redo, and running. The selected cell is a Markdown cell, indicated by the 'Markdown' dropdown in the toolbar. The cell content is as follows:

```
# Este es un título
## Este es un subtítulo
### Otro de menor nivel

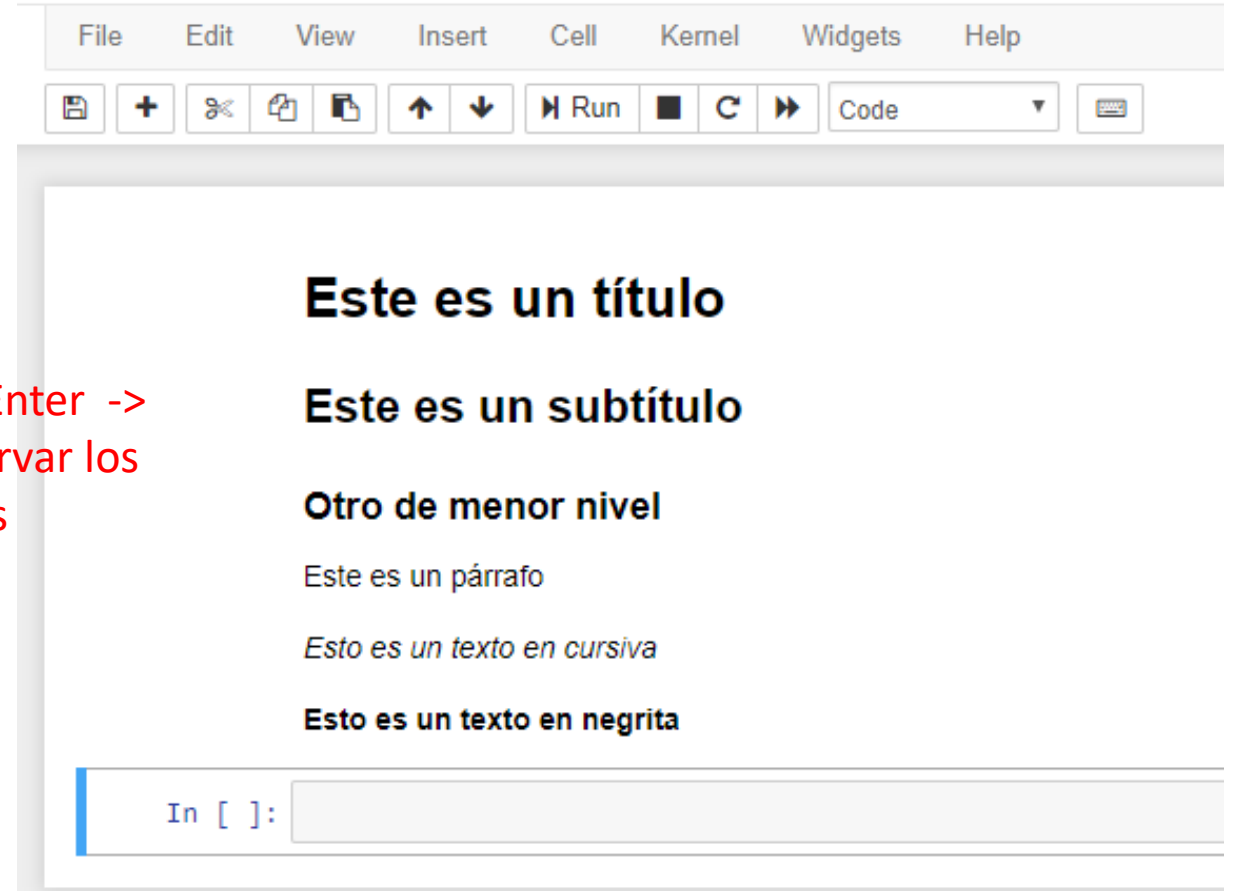
Este es un párrafo

*Esto es un texto en cursiva*

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

Below the cell, the prompt 'In []:' is visible next to an empty input field.

Mayús + Enter ->
Para observar los
resultados



The image shows the same Jupyter Notebook interface, but the cell is now in 'Code' mode, indicated by the 'Code' dropdown in the toolbar. The rendered output of the Markdown cell is displayed:

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

Below the output, the prompt 'In []:' is visible next to an empty input field.

Celda Code

En una celda
Code se puede
ejecutar y
probar código
Python

The screenshot displays a Jupyter Notebook interface. At the top is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. Below the menu is a toolbar containing icons for saving, adding a new cell, undo, redo, copy, paste, moving up/down, running the cell, clearing the cell, and a dropdown menu currently set to 'Code'. The main area of the notebook contains the following text:

- 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**

At the bottom, there is a code cell with the prompt 'In [1]:' followed by the code `print("FISI UNMSM")`. Below the code, the output 'FISI UNMSM' is displayed. A blue arrow points from the code to the output. To the right of the code cell, there are two yellow callout boxes with green text:

- Para ejecutar: Ctrl + Entrar
- Para ejecutar e insertar una nueva celda: Shift + Entrar

Python en pocos pasos

Temas a tratar

- Tipos de Datos
 - Números
 - Cadenas
 - Impresión Formateada
 - Listas
 - Diccionarios
 - Booleanos
 - Tuplas y Conjuntos
- Operadores de Comparación
- Sentencias If, elif y else
- Bucles For
- Bucles While
- range()
- Operadores de Comparación
- Sentencias If, elif y else
- Listas por comprensión
- Funciones
- Expresiones Lambda
- Map y Filter

Introducción a Machine Learning

Libro complementario

- Utilizaremos “Introduction to Statistical Learning” de Gareth James como libro complementario.
- Está disponible gratuitamente en línea, podemos conseguirlo en:

[Introduction to Statistical Learning - University of Southern California](http://www-bcf.usc.edu/~gareth/ISL/) ✓

www-bcf.usc.edu/~gareth/ISL/ ▼ Traducir esta página

Home, Download the book PDF (corrected 7th printing). Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani.


[Data Sets and Figures](#) · [R Code for Labs](#) · [Get the Book](#) · [About this Book](#)

Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

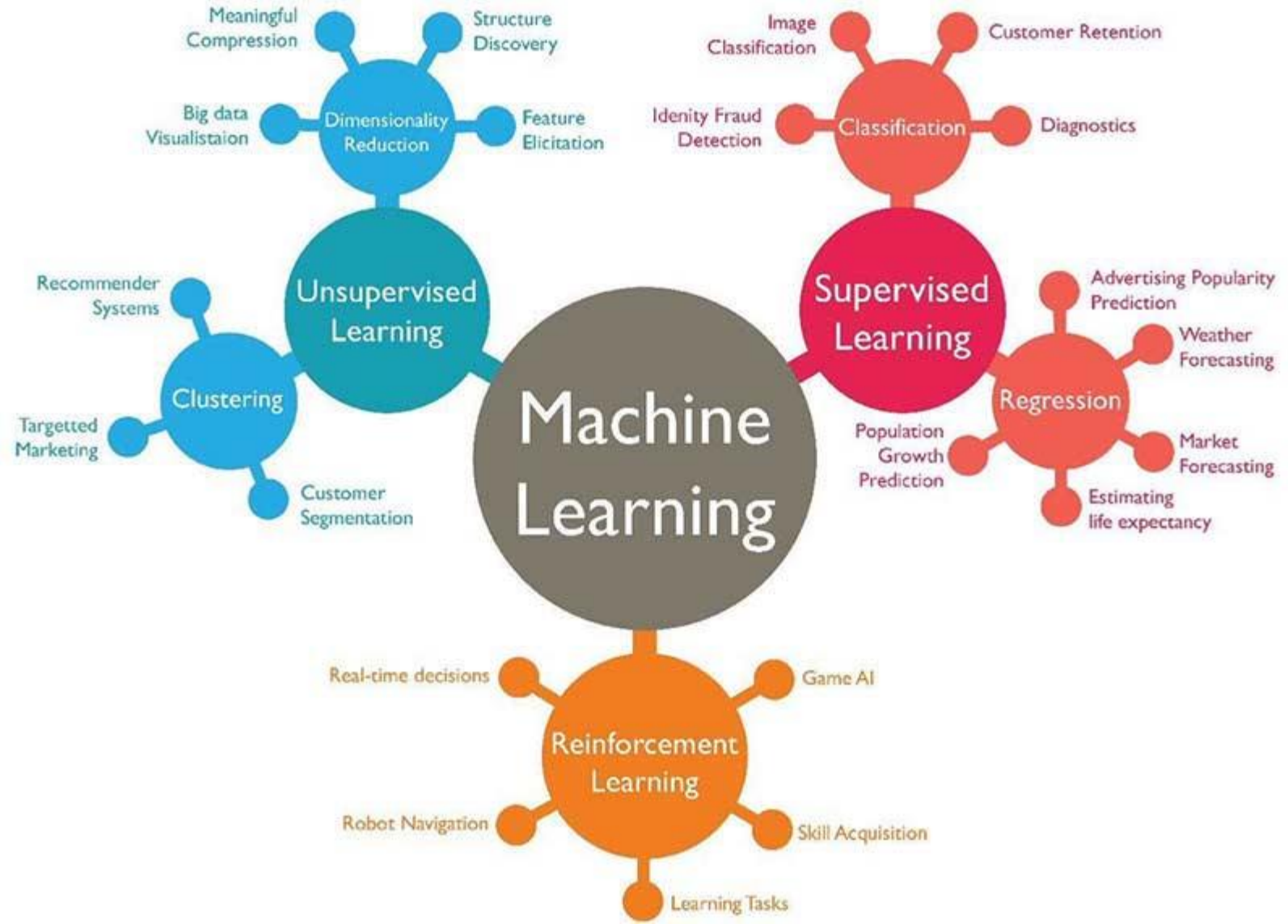
 Springer

Libro Complementario

- Los estudiantes que quieran la teoría matemática deben hacer las lecturas sugeridas.
- Los estudiantes que solo quieren aplicar los modelos y están más interesados en las aplicaciones de Python pueden simplemente enfocarse más en estos materiales.

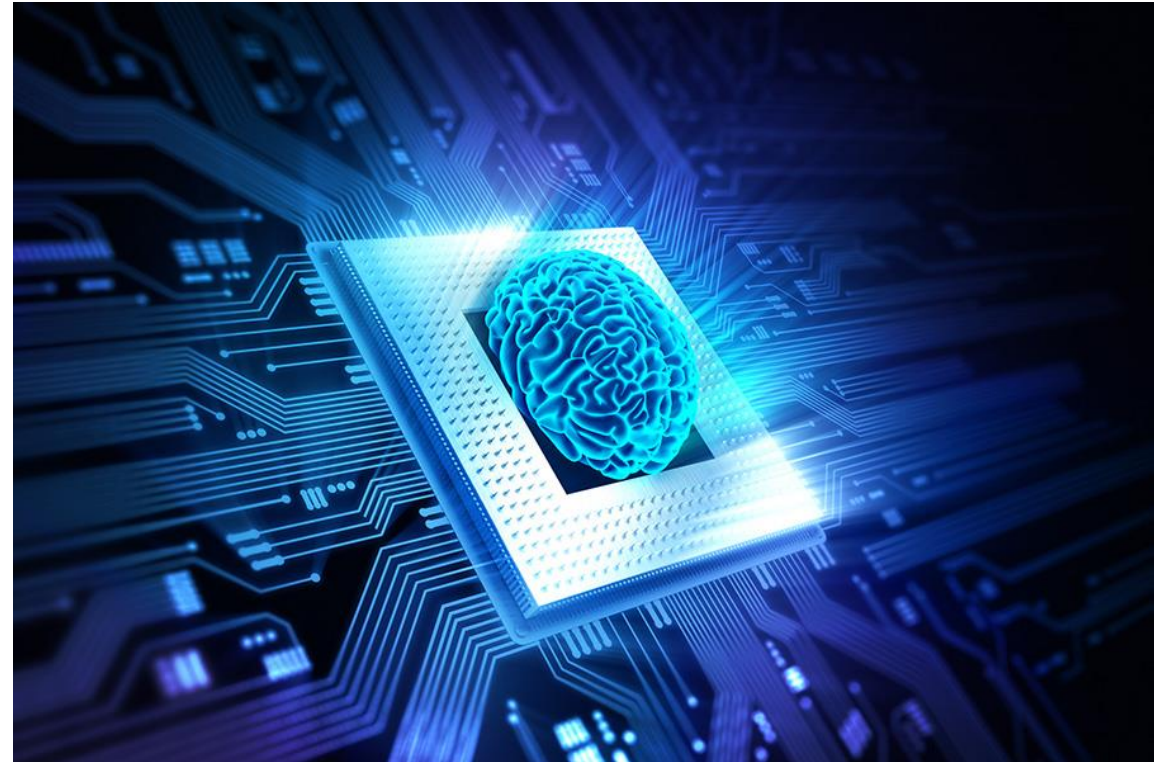
Libro Complementario

- Lea los Capítulos 1 y 2 si quiere obtener una mejor comprensión general antes de continuar con estos materiales.



¿Qué es Machine Learning o Aprendizaje Automático?

- El aprendizaje automático es un método de análisis de datos que automatiza la creación de modelos analíticos.
- Mediante el uso de algoritmos que aprenden iterativamente de los datos, el aprendizaje automático permite que las computadoras encuentren información oculta sin tener que programar explícitamente dónde buscar.

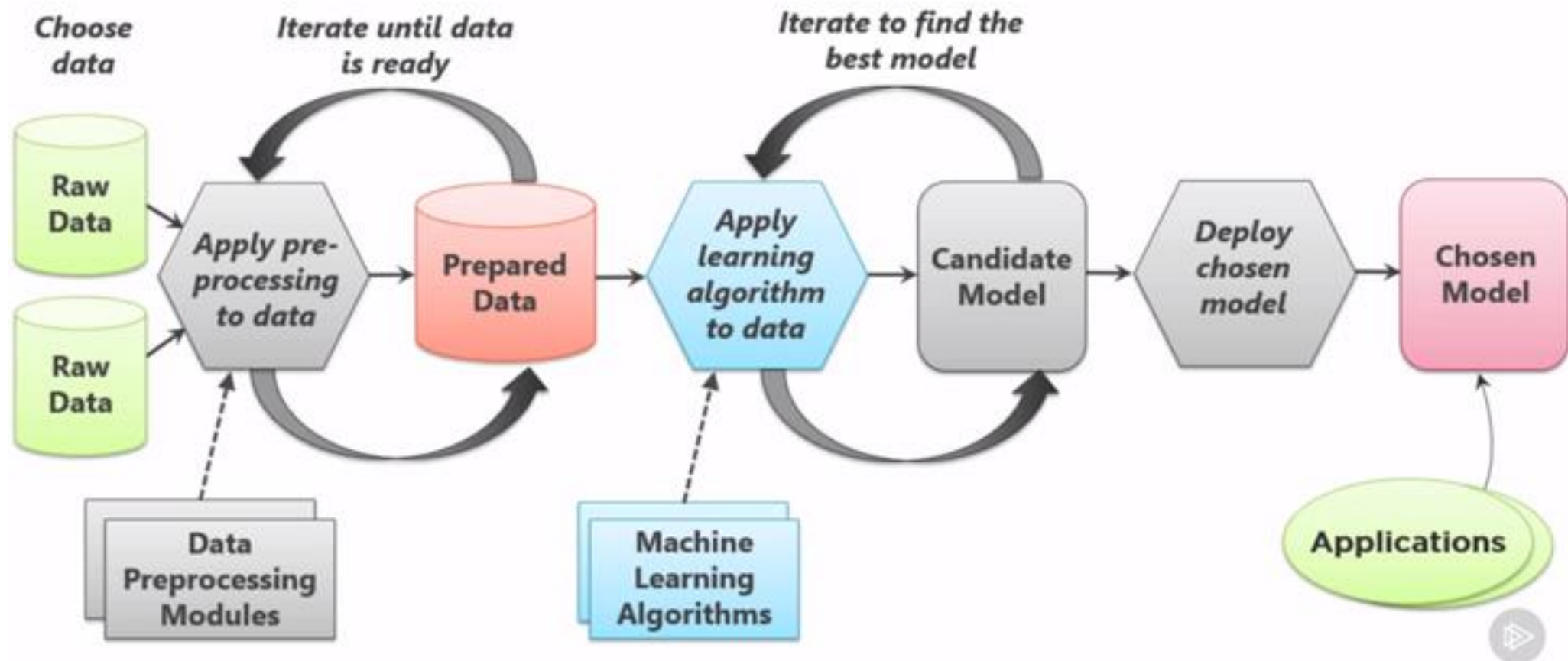


¿Para qué se usa?

- Detección de fraude.
- Resultados de búsqueda web.
- Anuncios en tiempo real en páginas web
- Calificación de crédito y las mejores ofertas siguientes.
- Predicción de fallas de equipos.
- Nuevos modelos de precios.
- Detección de intrusión de red.
- Motores de recomendación
- Segmentación del cliente
- Análisis de sentimiento de texto
- Predecir la rotación de clientes
- Reconocimiento de patrones e imágenes.
- Filtrado de spam de correo electrónico.
- Modelado financiero

Proceso del Aprendizaje Automático

The Machine Learning Process



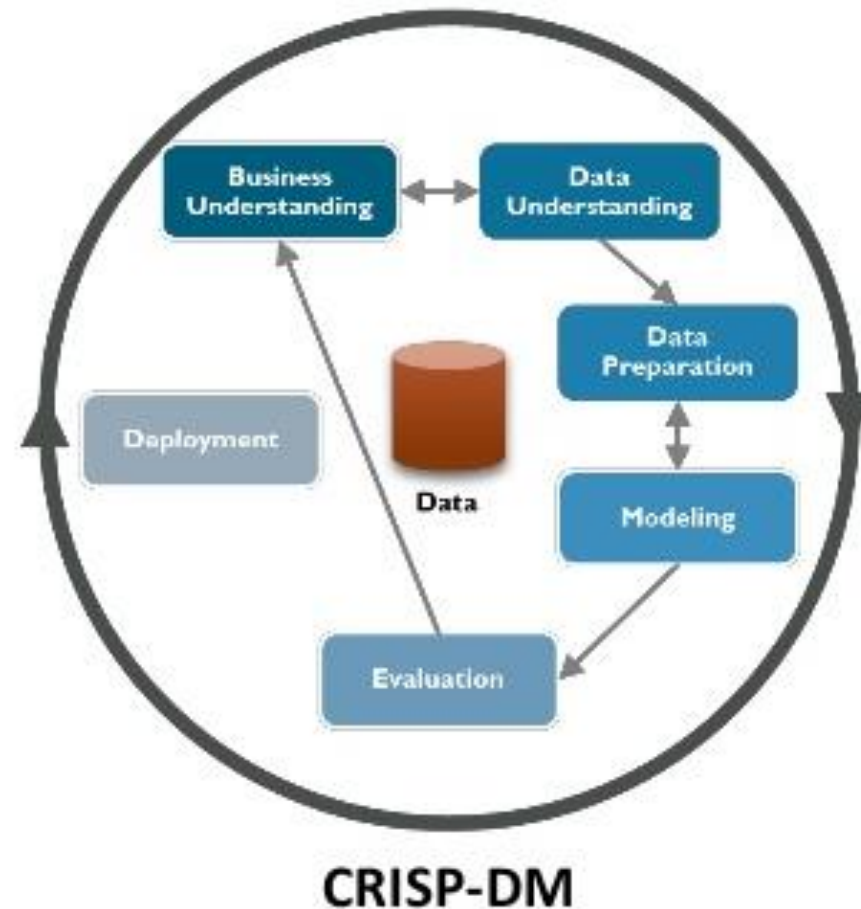
Proceso del Aprendizaje Automático

1. Data Engineering – 80%

- Data extraction
- Data cleaning
- Data transformation
- Data normalization
- Feature extraction

2. Machine Learning – 20%

- Model fitting
- Hyperparameters tuning
- Model evaluation



Aprendizaje Supervisado

- Los algoritmos de aprendizaje supervisados se entrenan usando ejemplos etiquetados, como una entrada donde se conoce el resultado deseado.
- Por ejemplo, unos objetos puede tener puntos de datos etiquetados como "M" (en mal estado) o "B" (buen estado).



buen estado



mal estado



¿Cuáles están en buen estado o mal estado?

Aprendizaje Supervisado

- El algoritmo de aprendizaje recibe un conjunto de entradas junto con las correspondientes salidas correctas, y el algoritmo aprende comparando su salida real con las salidas correctas para encontrar errores.
- Luego modifica el modelo en consecuencia.

Aprendizaje Supervisado

- A través de métodos como la clasificación, la regresión, la predicción y el aumento de gradiente, el aprendizaje supervisado usa patrones para predecir los valores de la etiqueta en datos adicionales no etiquetados.
- El aprendizaje supervisado se usa comúnmente en aplicaciones donde los datos históricos predicen eventos futuros probables.

Aprendizaje Supervisado

- A través de métodos como la clasificación, la regresión, la predicción y el aumento de gradiente, el aprendizaje supervisado usa patrones para predecir los valores de la etiqueta en datos adicionales no etiquetados.
- El aprendizaje supervisado se usa comúnmente en aplicaciones donde los datos históricos predicen eventos futuros probables.

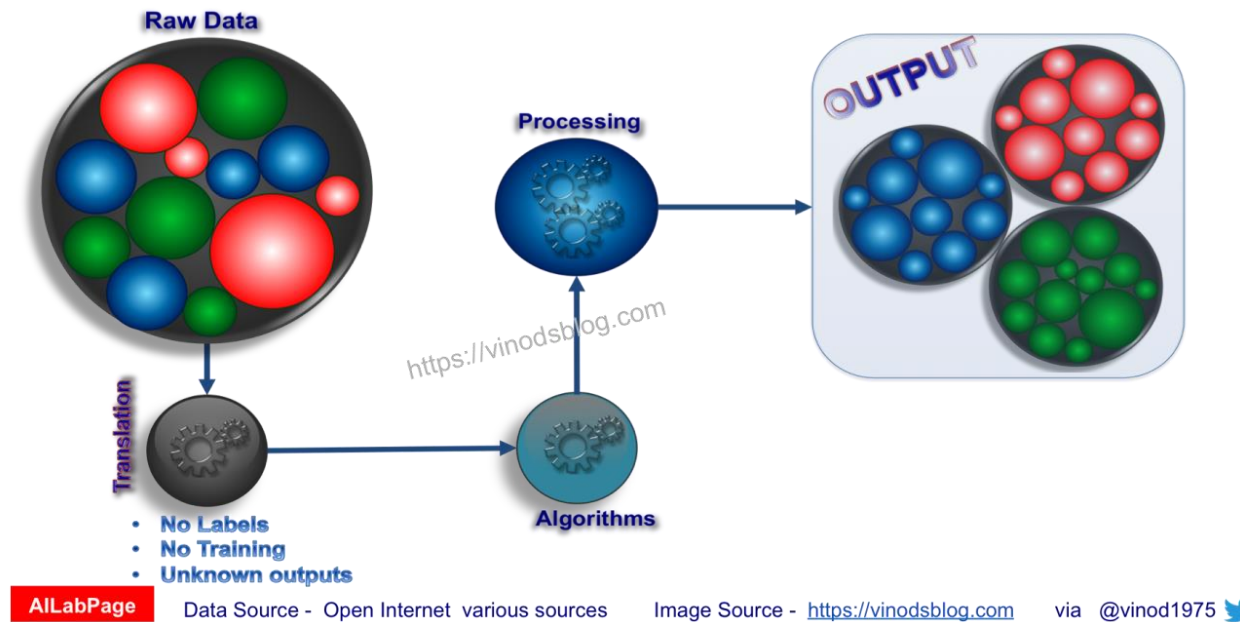
Aprendizaje Supervisado

- Por ejemplo, puede anticipar cuándo es probable que las transacciones con tarjeta de crédito sean fraudulentas o qué cliente de seguros es probable que presente un reclamo.
- O puede intentar predecir el precio de una casa en función de las diferentes características de las casas para las que tenemos datos de precios históricos.



Aprendizaje No Supervisado

- El aprendizaje no supervisado se usa con datos que no tienen etiquetas históricas.
- Al sistema no se le dice la "respuesta correcta". El algoritmo debe descubrir lo que se muestra.
- El objetivo es explorar los datos y encontrar alguna estructura dentro.



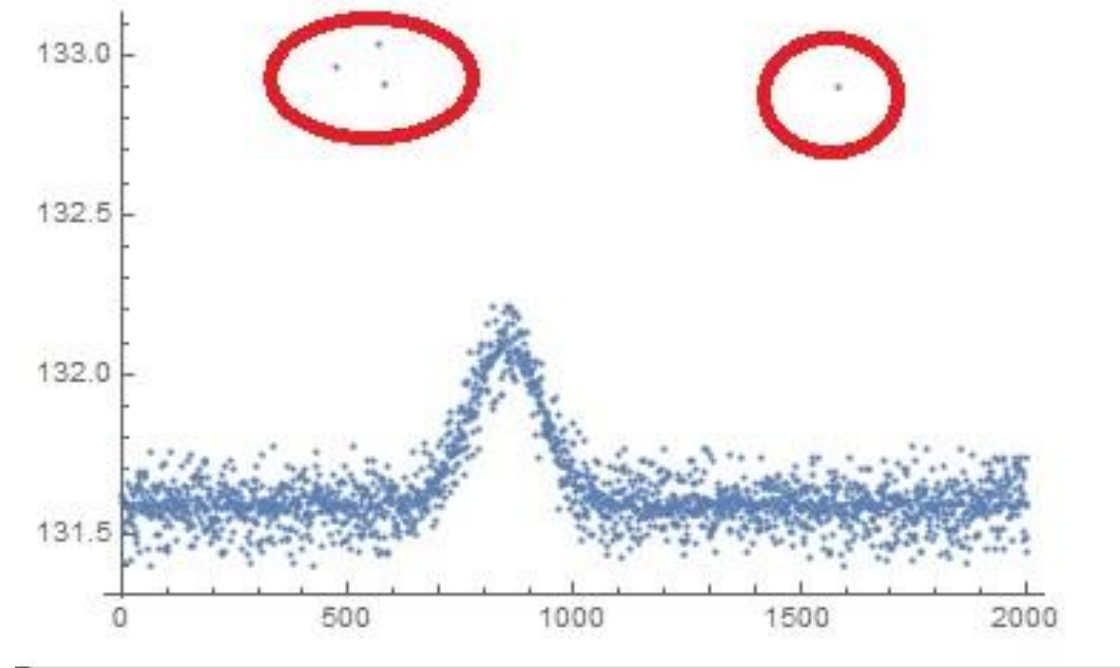
Aprendizaje No Supervisado

- O puede encontrar los principales atributos que separan segmentos de clientes entre sí.
- Las técnicas populares incluyen mapas autoorganizados, mapeo del vecino más cercano, clustering k-means y descomposición de valores singulares.



Aprendizaje No Supervisado

- Estos algoritmos también se utilizan para segmentar temas en un texto, recomendar elementos e identificar valores atípicos de datos.



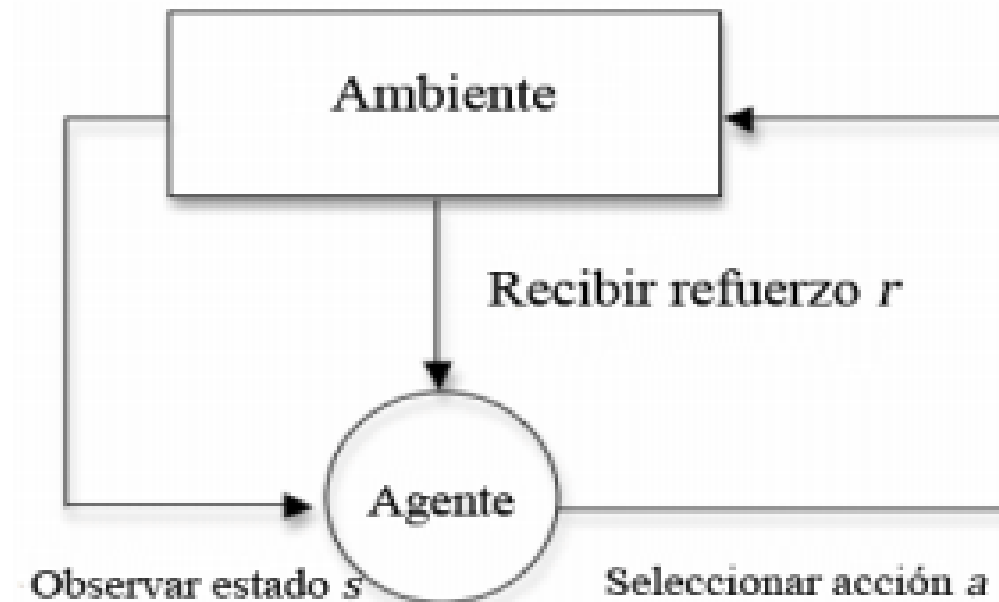
Aprendizaje Reforzado

- El aprendizaje reforzado a menudo se usa para robótica, juegos y navegación.
- Con el aprendizaje reforzado, el algoritmo descubre a través de prueba y error qué acciones rinden las mayores recompensas.



Aprendizaje Reforzado

- Este tipo de aprendizaje tiene tres componentes principales: el agente (el que aprende o el que toma las decisiones), el entorno (todo con lo que el agente interactúa) y las acciones (lo que el agente puede hacer).



Aprendizaje Reforzado

- El objetivo es que el agente elija acciones que maximicen la recompensa esperada durante un período de tiempo determinado.
- El agente alcanzará el objetivo mucho más rápido siguiendo una buena política.
- Entonces, el objetivo en el aprendizaje de refuerzo es aprender la mejor política.

