Data Science y Machine Learning para ciberseguridad

Juan Anabalón R.

http://deoxyt2.livejournal.com

jar@monkeyslab.cl

Who I am?

Juan Anabalón R. *http://deoxyt2.livejournal.com*

Geek Code: GCS d@ s: a C++$@ UB>$ P L- !E--- !W+++@ !N !o K--? w$ O>$ M++>$!V PS+ PE-@ Y++ PGP !t !5 !X !R tv- b++++>$ DI !D G+++>$ e+++ h---- r+++ y+++

Nerdity Test MIT: NERD!

Especialista de Ciberseguridad en MonkeysLab.cl

Presidente ISSA Chile

Ingeniero de Ejecución en Informática

Magíster en Seguridad, Peritaje y Auditoría en Procesos Informáticos

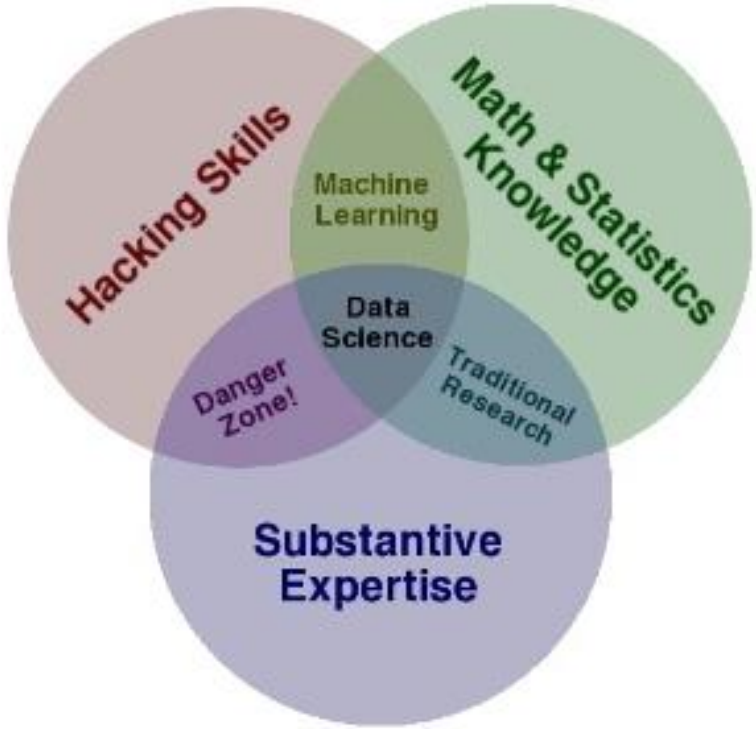
Varios años de experiencia

¿Qué es Data Science?

Definición

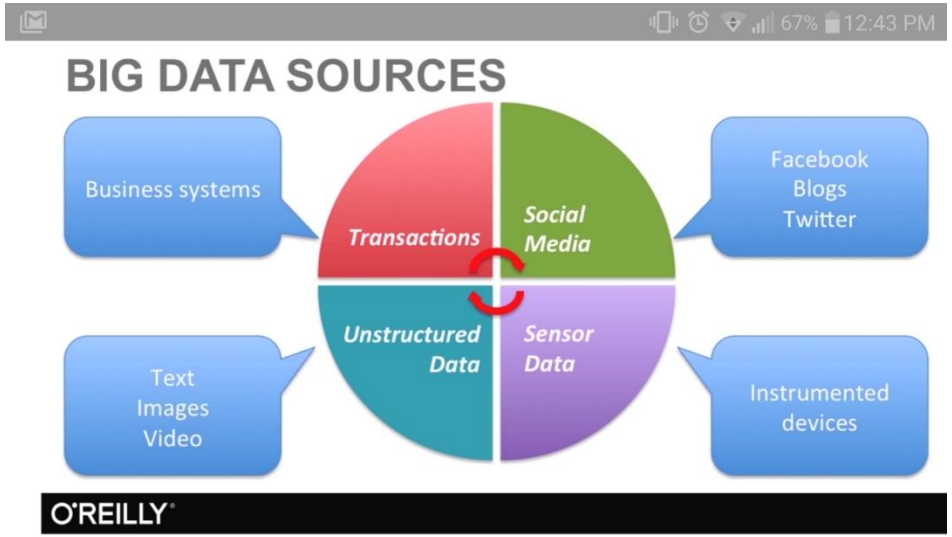
La ciencia de datos es el arte de convertir los datos en acciones. Esto se logra a través de la creación de productos de datos, que proporcionan información procesable sin exponer a los responsables de la toma de decisiones a los datos o análisis subyacentes.

Booz Allen Hamilton, Field Guide to Data Science, Pg. 17

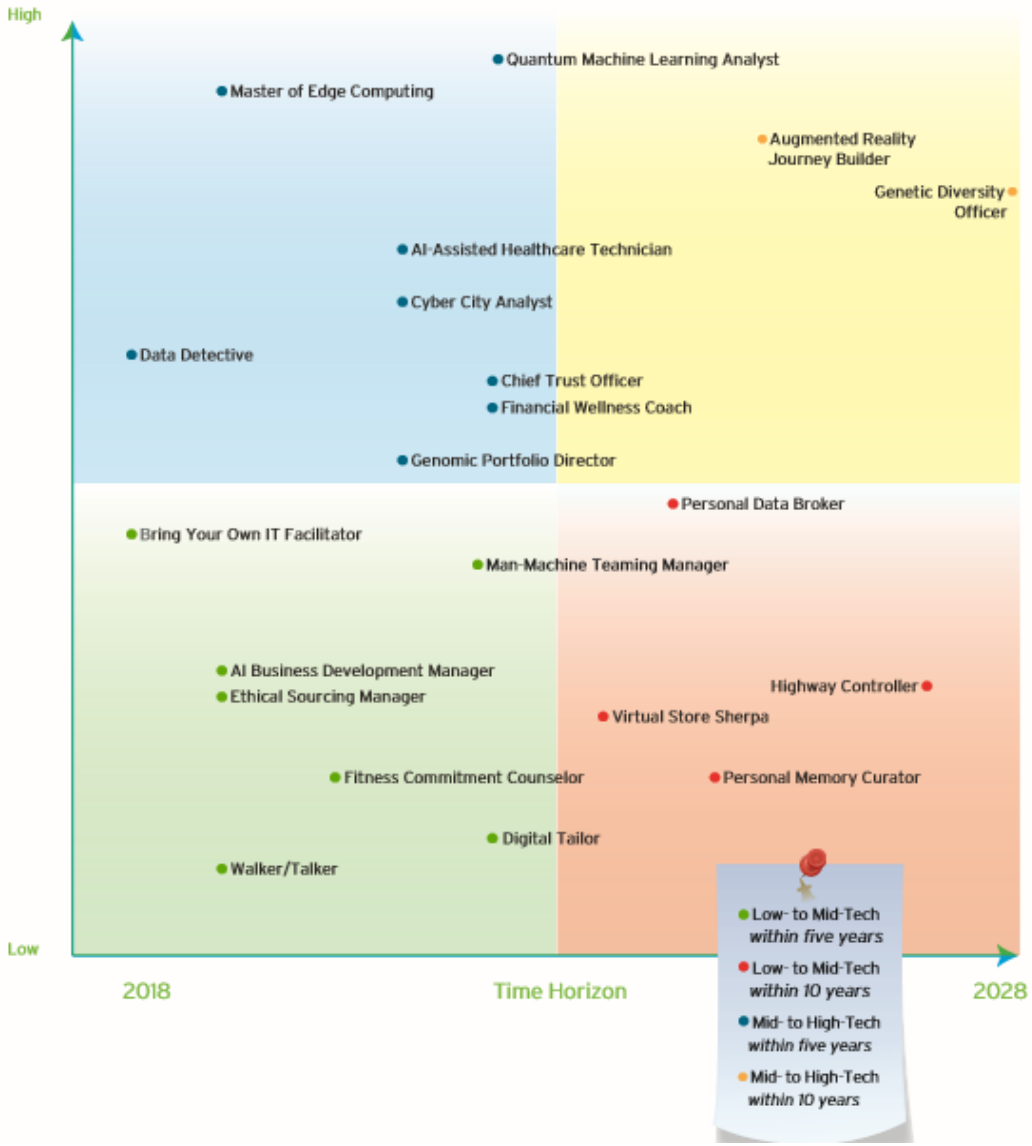


¿por qué?

- Rápido crecimiento de necesidades de la empresa.
- Explosión de datos.
- Big Data
- IoT
- Smart City



Cargos del futuro



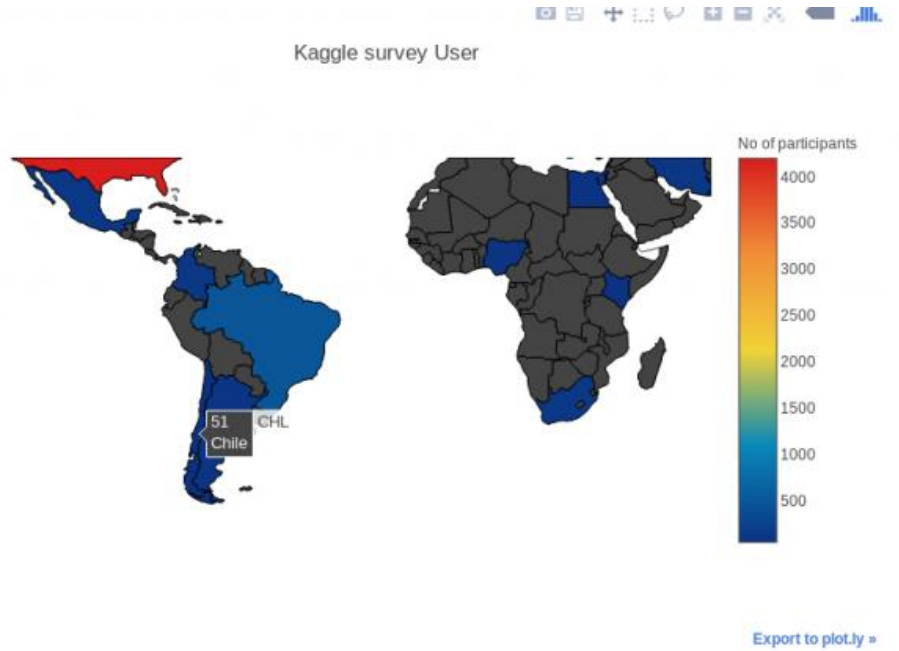
https://www.cognizant.com/whitepapers/21-jobs-of-the-future-a-guide-to-getting-and-staying-employed-over-the-next-10-years-codex3049.pdf

Actualmente…

Kaggle Survey

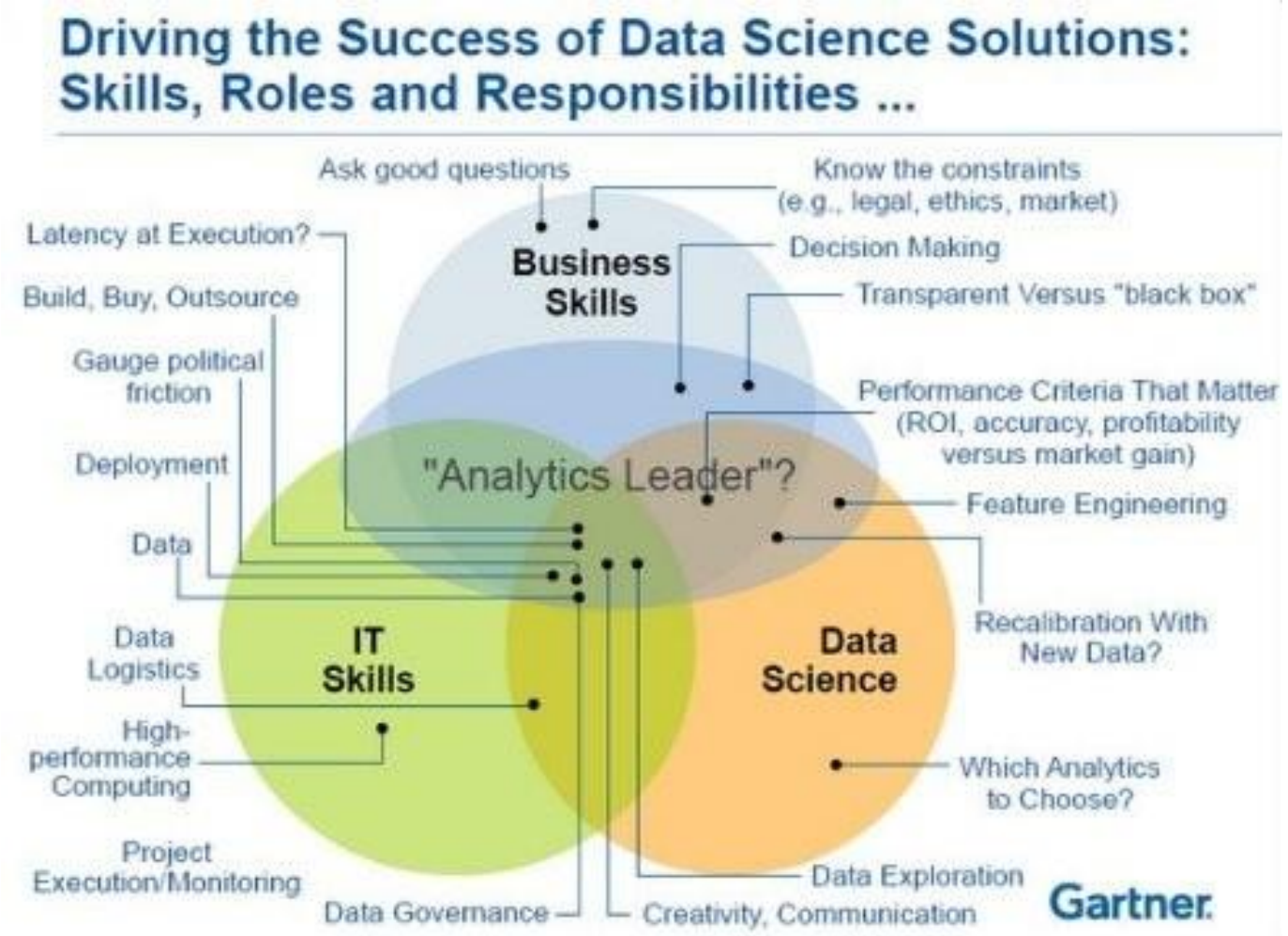
- Si bien **Python** es la herramienta más utilizada, hay más personas usando **R**.
- En promedio, los científicos de datos tienen alrededor de **30 años**, pero este valor varía según los países. Por ejemplo, el encuestado promedio de la India era aproximadamente nueve años más joven que el encuestado promedio de Australia.
- La Mayoría de los encuestados obtuvo un título de **maestría**, pero en los rangos salariales más altos ($ 150K +) ellos tienen un título de **doctorado**.

Chile



https://www.kaggle.com/surveys/2017

¿qué se necesita?



¿Qué es Machine Learning?

El aprendizaje automático o aprendizaje de máquinas (del inglés, "Machine Learning") es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender

Wikipedia

¿Qué es Machine Learning?

- **Arthur Samuel** lo describió como: "el campo de estudio que da a las computadoras la capacidad de aprender sin estar programadas explícitamente". Esta es una definición más antigua e informal.
  Tom Mitchell proporciona una definición más moderna: "Se dice que un programa de computadora aprende de la experiencia E con respecto a una clase de tareas T y la medida de performance P, si su desempeño en tareas en T, medido por P, mejora con la experiencia E"

Aplicaciones de ML

- Reconocimiento de imágenes
- Clasificación de correo basura (SPAM)
- Web searh engine
- Reconocimiento de voz
- Vehículo autónomo

Tipos de ML

- Supervised Learning
- Unsupervised Learning
- Reinforcement learning

Aprendizaje supervisado

- En el aprendizaje supervisado, se nos da un conjunto de datos y ya sabemos cómo debería ser nuestra salida correcta, teniendo la idea de que existe una relación entre la entrada y la salida.

- Los problemas de aprendizaje supervisado se clasifican en problemas de "regresión" y "clasificación". En un problema de **regresión**, estamos tratando de **predecir resultados** dentro de un resultado continuo, lo que significa que estamos tratando de asignar variables de entrada a alguna función continua. En un problema de **clasificación**, en cambio, estamos tratando de predecir los resultados en un **resultado discreto**. En otras palabras, estamos tratando de asignar variables de entrada en categorías discretas. Aquí hay una descripción sobre las matemáticas es divertido en datos continuos y discretos.

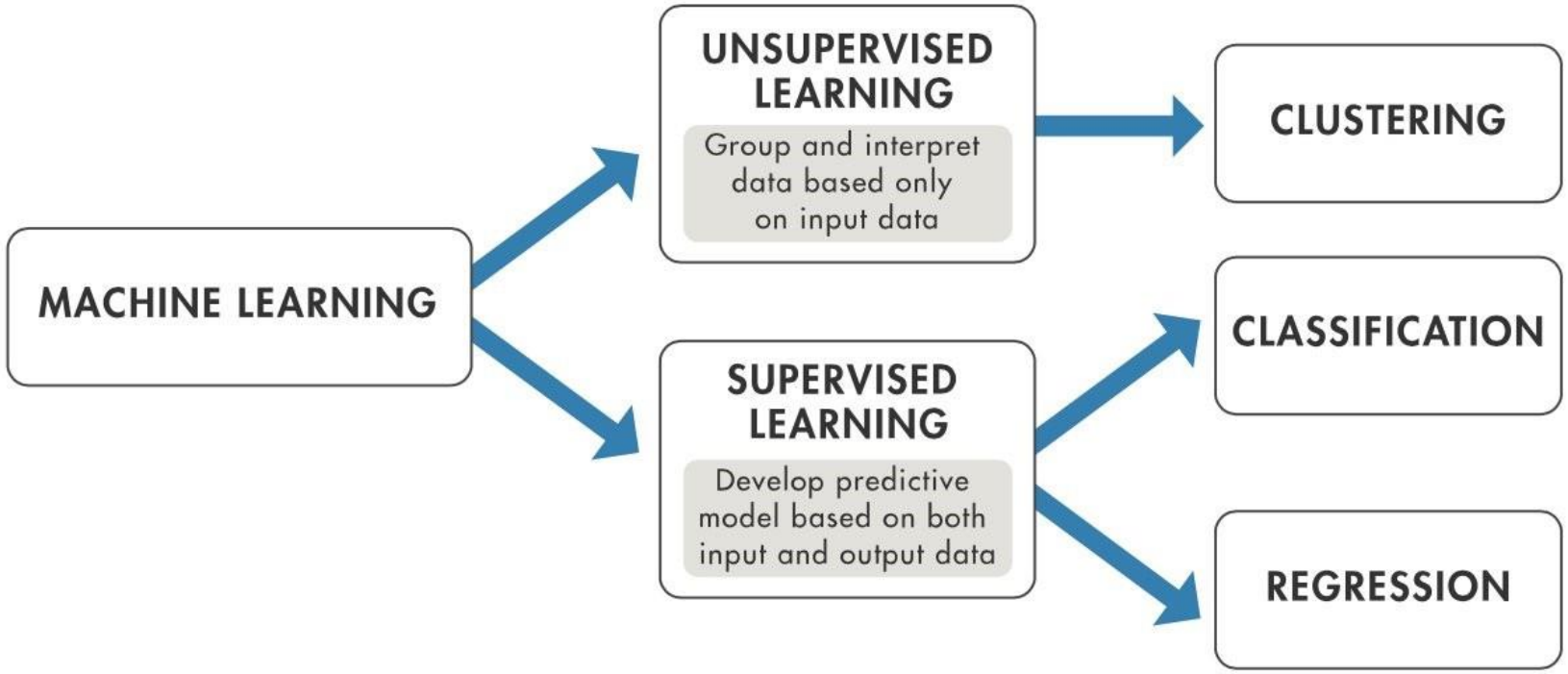
Aprendizaje no supervisado

- El aprendizaje no supervisado, por otro lado, nos permite abordar problemas **con poca o ninguna idea de cómo deberían ser nuestros resultados**.
  Podemos derivar la estructura de los datos donde no necesariamente conocemos el efecto de las variables.
  Con el aprendizaje no supervisado no hay comentarios basados en los resultados de la predicción, es decir, no hay un maestro que te corrija.

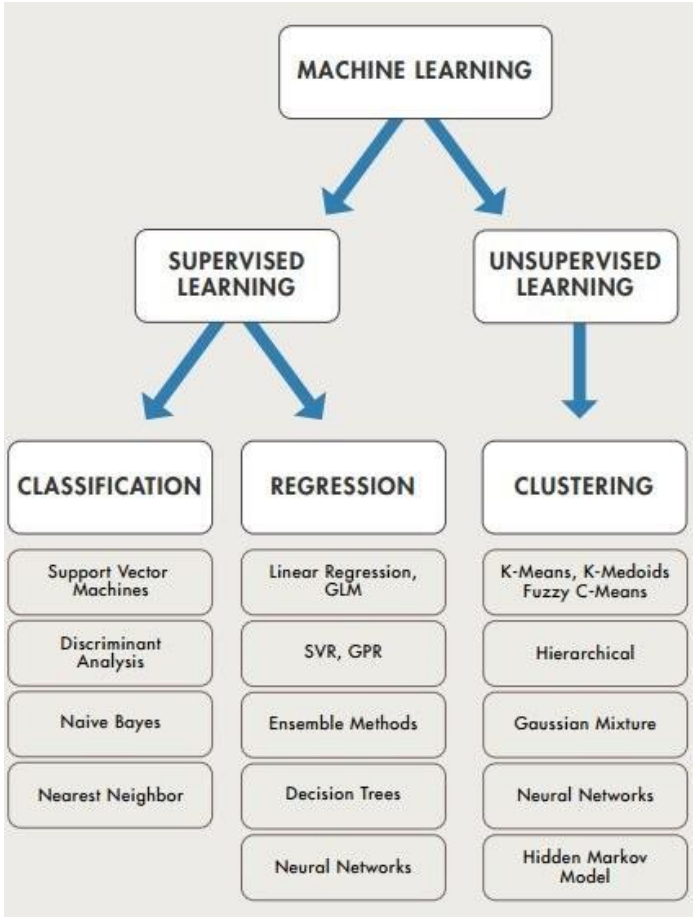
Aprendizaje por refuerzo

- En el aprendizaje de refuerzo, el objetivo es desarrollar un sistema (agente) que mejore su rendimiento en función de las interacciones con el entorno, podemos pensar en el aprendizaje de refuerzo como un campo relacionado con el aprendizaje supervisado. Sin embargo, en el aprendizaje por refuerzo, esta retroalimentación no es la etiqueta o el valor correcto, sino una medida de qué tan bien se midió la acción mediante una función de recompensa. Ejemplo: Ajedrés.

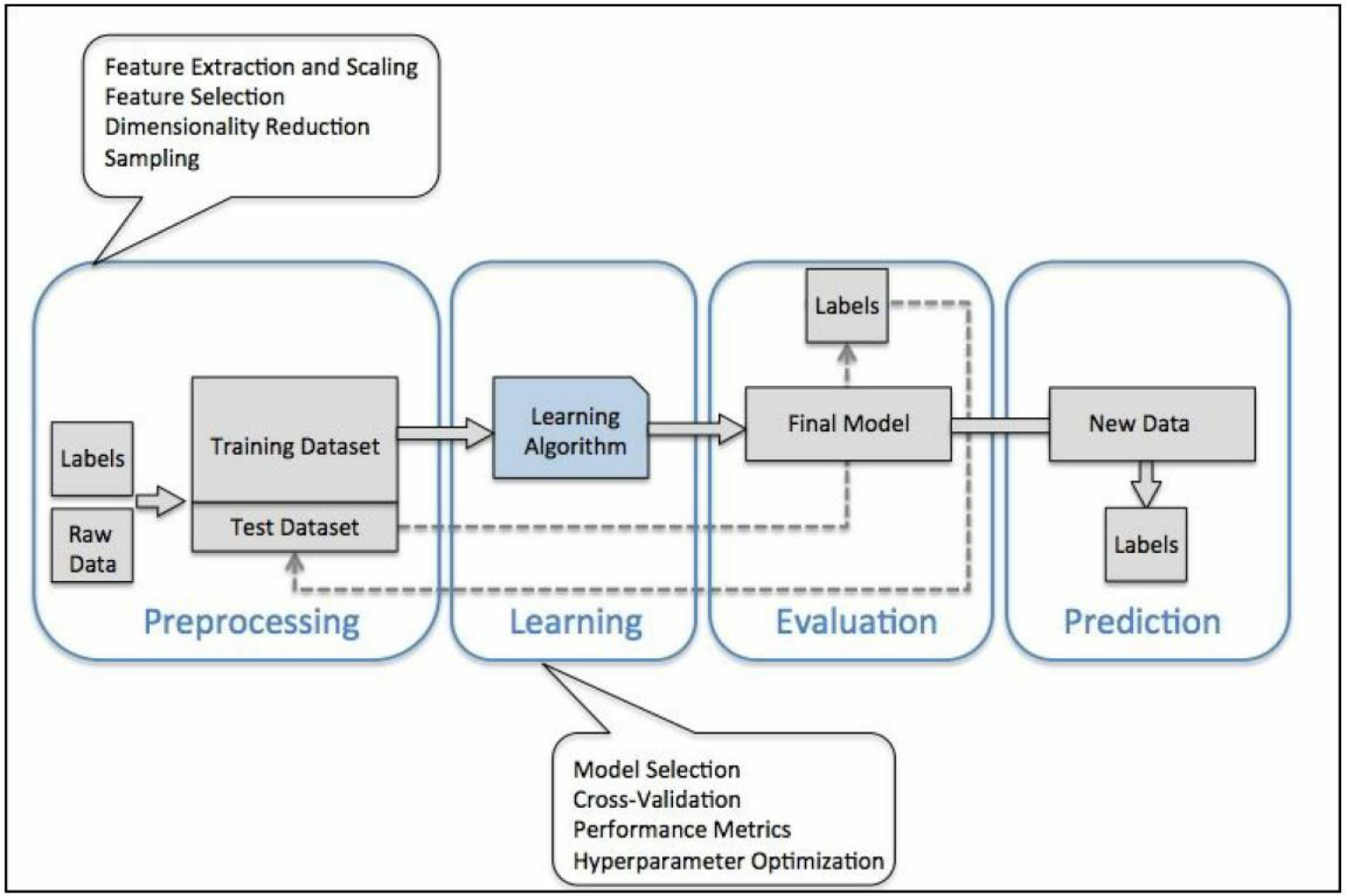
ML en una imagen



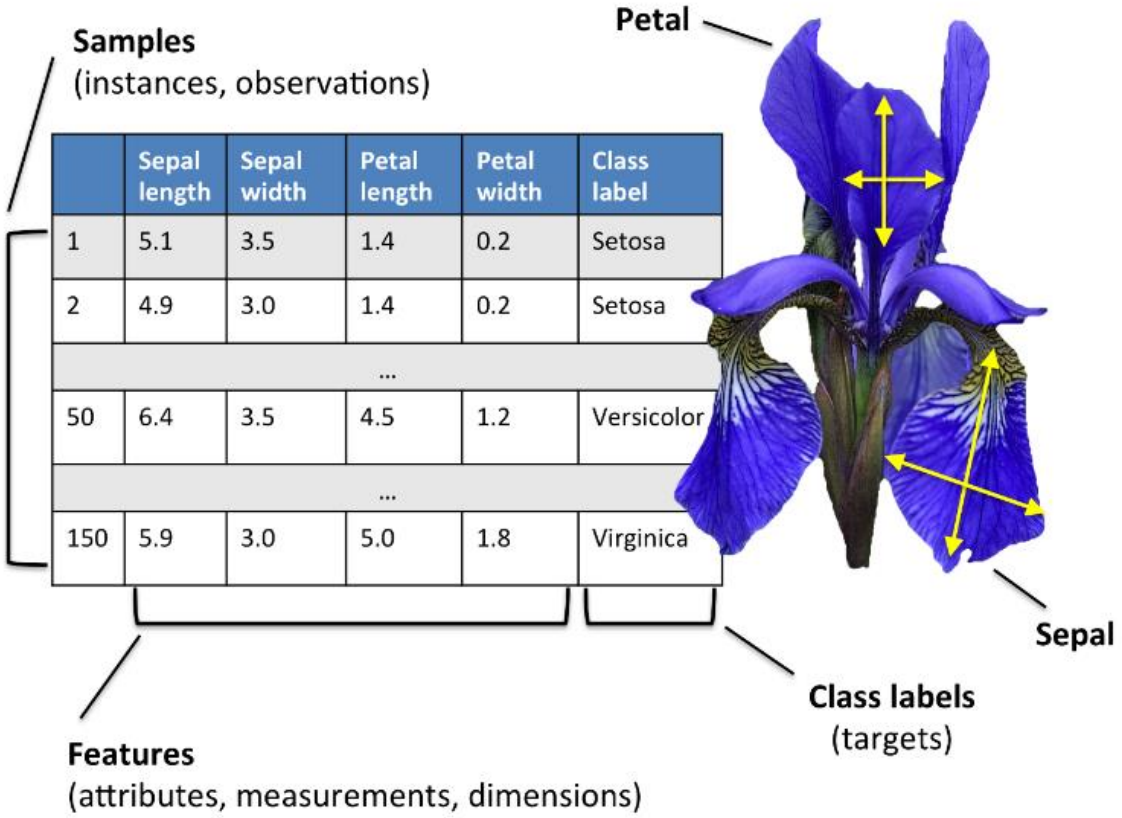
¿Cómo elegir un algoritmo?



Proceso


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Learning Algorithm
Labels
Final Model
New Data
Labels
Preprocessing
Learning
Evaluation
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

Iris Data Set



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

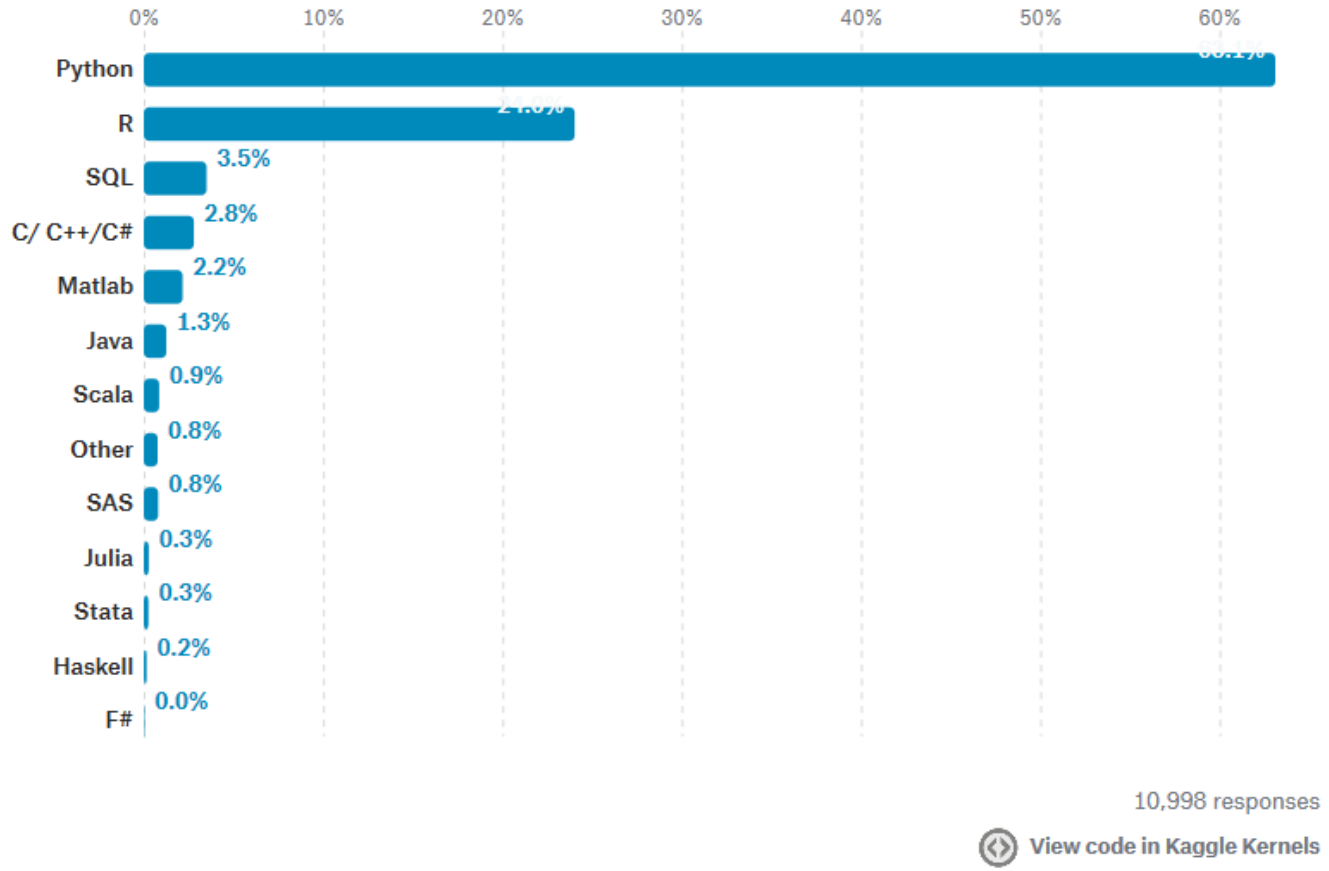
Ling-spam

Data Sets from Spam Studies

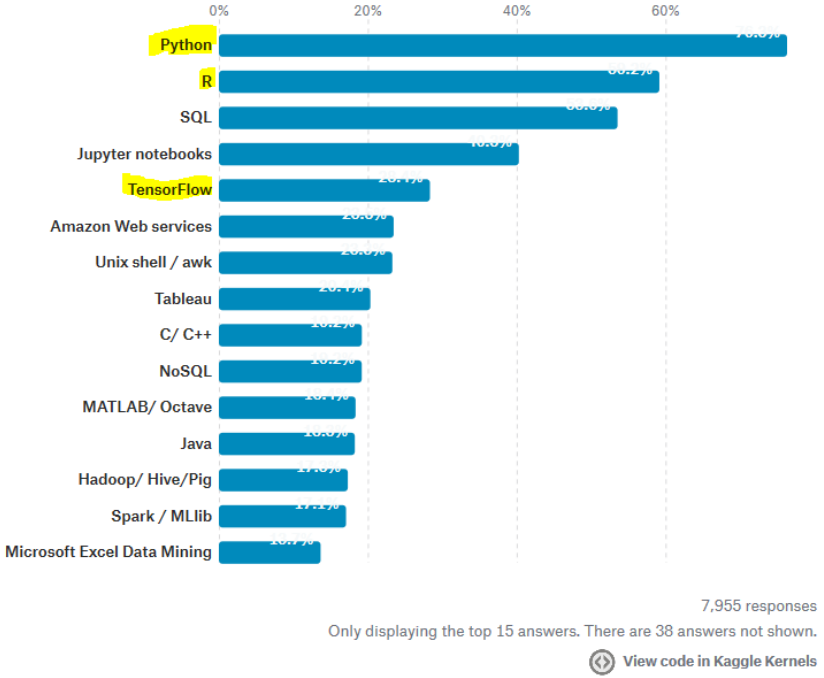
Name	Year Introduced	Times Used	Num Messages	Spam %
Ling-Spam	2000	5	3k	16%
PU1	2000	3	1k	45%
PU2 / PU3	2004	1	721 / 4k	20% / 44%
Non-Public	Various	4	1.7k – 11k	28% – 88%

- Other Data sets
  - Spam Assassin corpus
  - Spambase (from the UCI repository)

¿Qué lenguaje recomienda para empezar en DS?



¿Por qué Python?



Python tiene:

- Una sintaxis elegante y concisa
- Multiplataforma.
- Una amplia colección de bibliotecas eficaces y
- Herramientas de desarrollo perfeccionadas.

Python packages

- NumPy 1.9.1
- SciPy 0.14.0
- scikit-learn 0.15.2
- matplotlib 1.4.0
- pandas 0.15.2

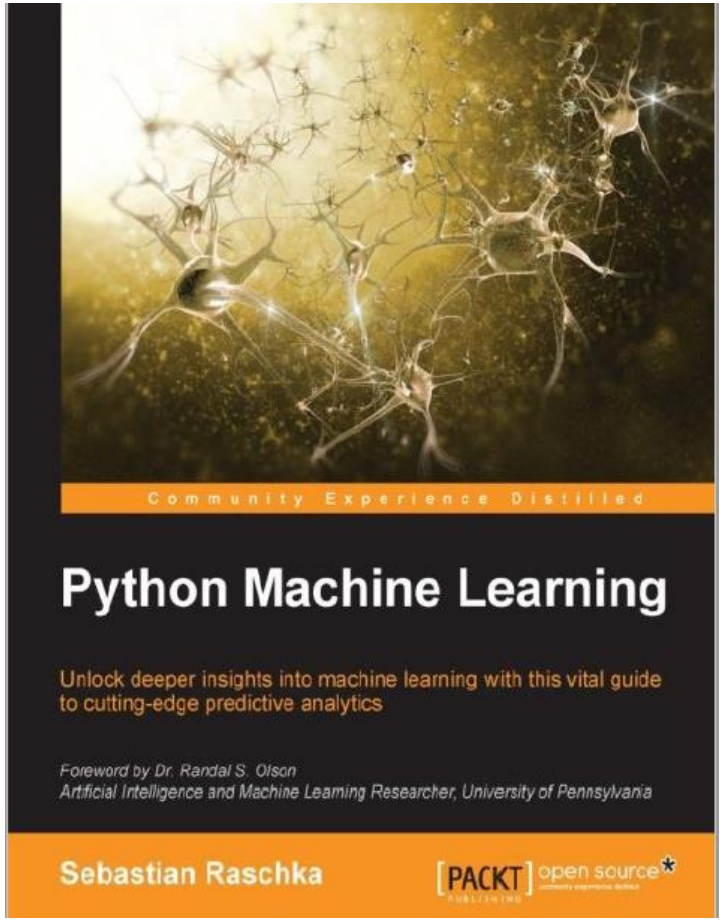
- Linux
  - **pip install SomePackage**
  - **pip install SomePackage –upgrade**

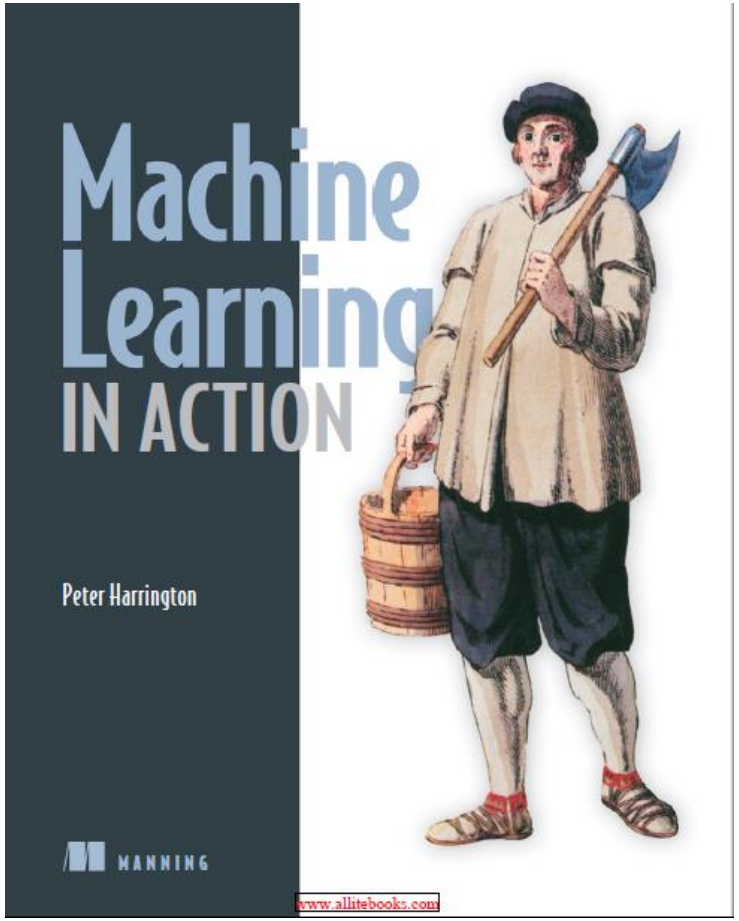
- Anaconda 
  - https://www.anaconda.com/download/

Demo

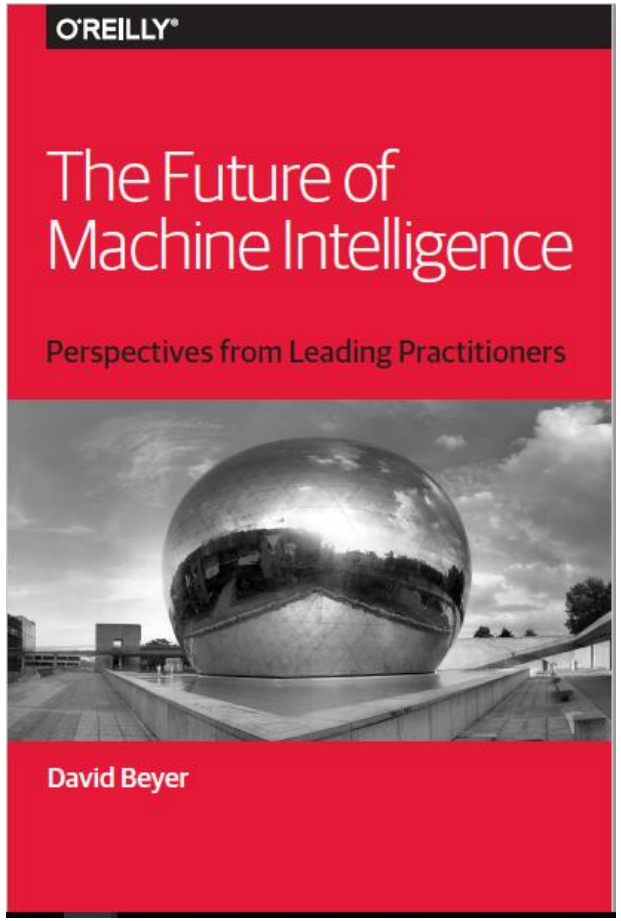

ISSA
Information Systems Security Association
The Global Voice of the Information Security Profession
Chapter-Chile

Recomendaciones





Más recomendaciones



coursera

UDACITY

edX