Data Science en la Industria

Soy Cristhian Boujon

- Ingeniero en Sistemas de Información.
- Sr. Software Engineer en ecommerce.
- Profesor.

Desarrollo web hace 10 años atrás...

HTTP

HTML

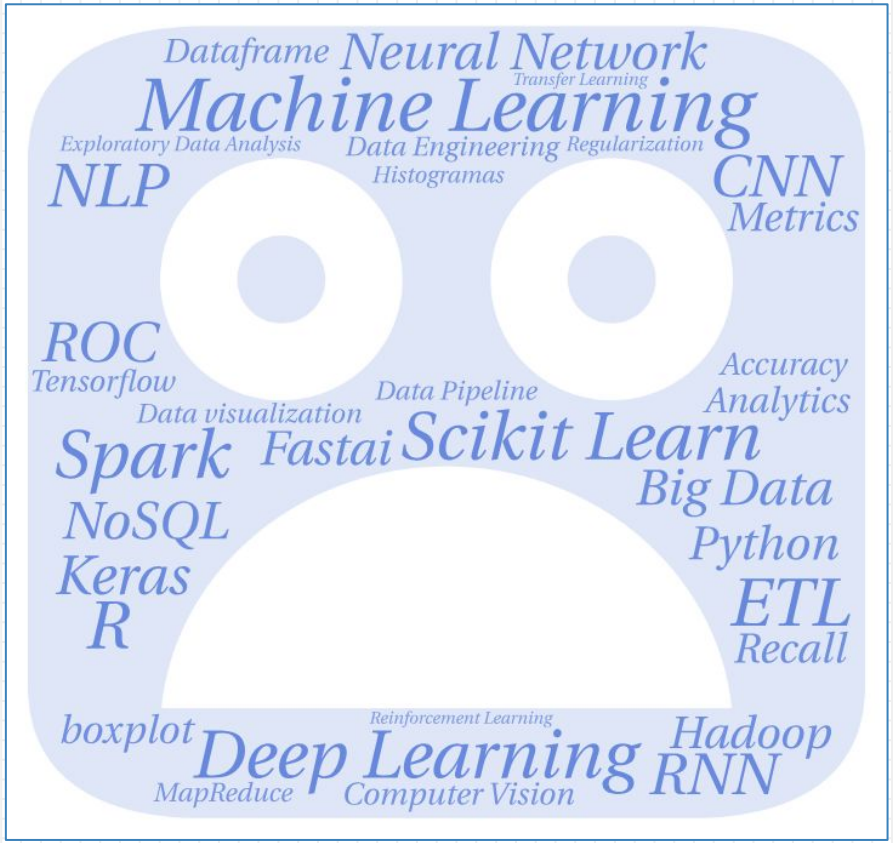
CSS

Javascript

PHP

MySQL

Data Science

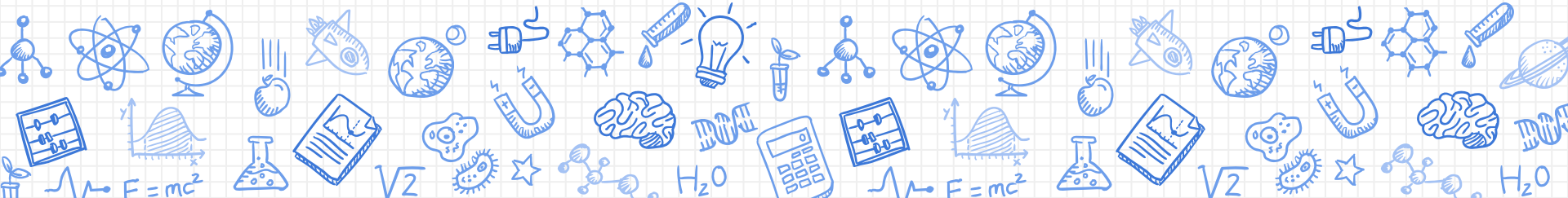


Data Science en la Industria

- Análisis de datos / Data Analysis
- Modelado / Modeling
- Ingeniería de datos / Data Engineering

Análisis de Datos

Se examinan los datos en bruto y así sacar conclusiones sobre la información.

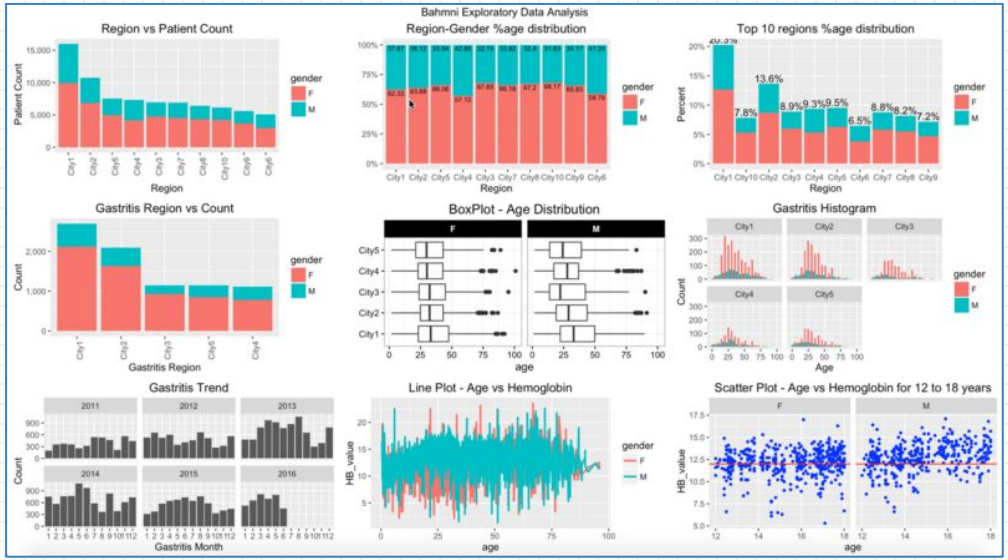


Análisis de Datos

- Conocimientos de Negocio.
- Conocimientos en estadística

Análisis de Datos – Exploración

Consiste en entender los datos con los que contamos.



Fuente: https://medium.com/bahmni-blog/introduction-to-exploratory-data-analysis-of-bahmni-using-r-6c186fd6f010

Análisis de Datos – Interpretación

Capturar el significado real de los datos en el negocio.

- Reportes
- Insights
- Visualización de datos
  - Rutina en EEUU
  - Marcas

Tools


jupyter


ggplot2


matplotlib


RStudio

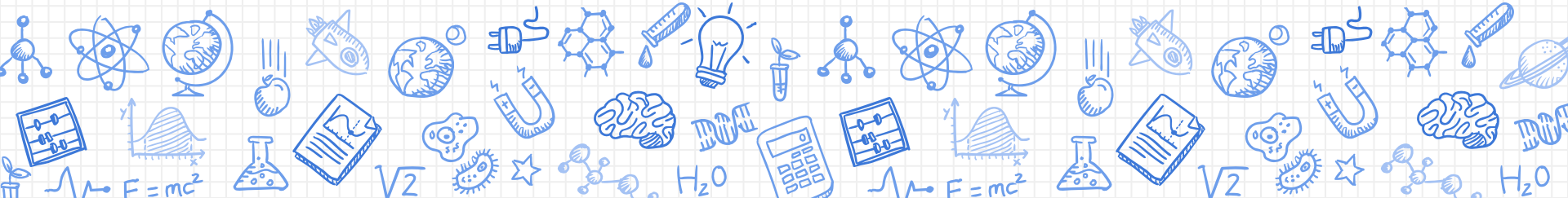

+ableau
SOFTWARE


plotly

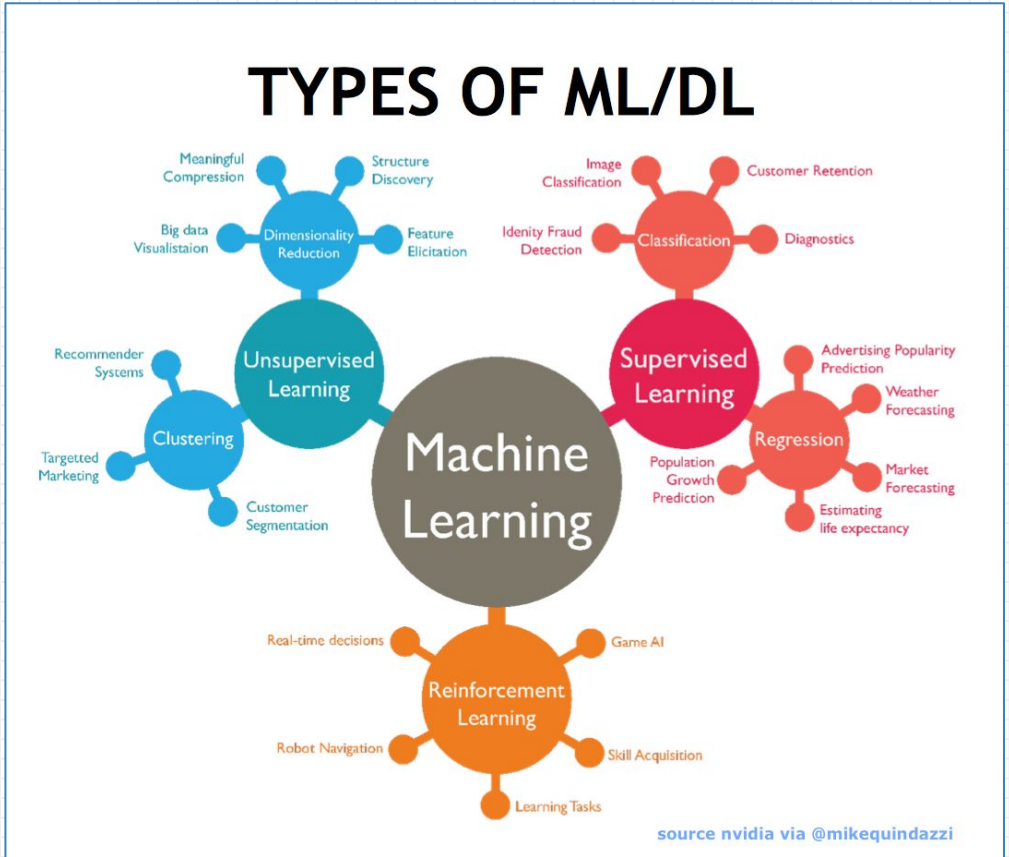

pandas
$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$

Modelado

Creación de modelos de Machine Learning



Modelado – Machine Learning Types



Modelado – Aprendizaje Supervisado

El modelo aprende en base a ejemplos que fueron previamente etiquetados por humanos.

Modelado – Clasificación

Si el target es discreto, hablamos de un modelo de clasificación.

Modelado – Ejemplo de Clasificación

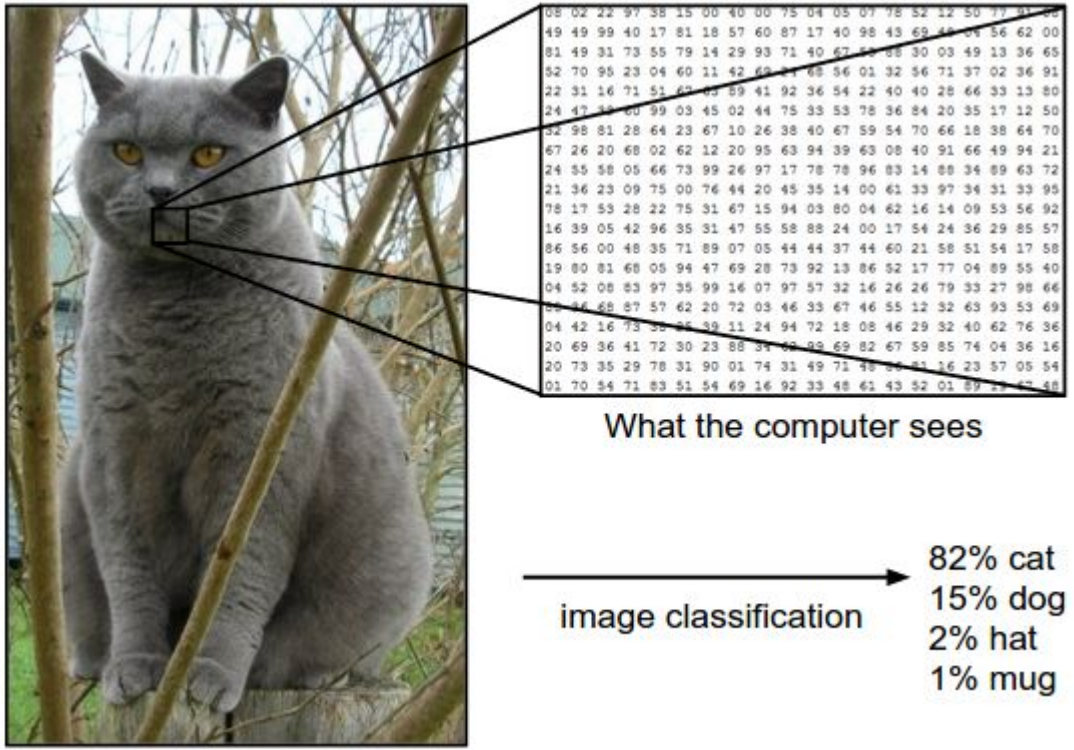
Features

Target

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunctio	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0

Fuente: https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/diabetes.csv

Modelado – Ejemplo de Clasificación



Fuente:
https://medium.com/the-data-experience/building-a-data-pipeline-from-scratch-32b712cfb1db

Modelado – Ejemplo de Clasificación

¿Features?

Target

v2	v1
Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat..	ham
Ok lar... Joking wif u oni...	ham
Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry q	spam
U dun say so early hor... U c already then say...	ham
Nah I don't think he goes to usf, he lives around here though	ham
FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still?	spam
Even my brother is not like to speak with me. They treat me like aids patent.	ham
As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune	ham
WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To cla	spam
Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for	spam
I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough to	ham
SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day,	spam
URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLA	spam
I've been searching for the right words to thank you for this breather. I promise i wont take your help for g	ham
I HAVE A DATE ON SUNDAY WITH WILL!!	ham

Fuente: https://www.kaggle.com/uciml/sms-spam-collection-dataset

Modelado – Regresión

Si el target son continuas, hablamos de un modelo de Regresión.

Modelado – Ejemplo de Regresión

Features

Target

Ejemplos

bedrooms	bathrooms	sqft_living	sqft_lot	floors	yr_built	lat	long	price
3	1	1180	5650	1	1955	47.5112	-122.257	221900
3	2	2570	7242	2	1951	47.721	-122.319	538000
2	1	770	10000	1	1933	47.7379	-122.233	180000
4	3	1960	5000	1	1965	47.5208	-122.393	604000
3	2	1680	8080	1	1987	47.6168	-122.045	510000
4	5	5420	101930	1	2001	47.6561	-122.005	1225000
3	2	1715	6819	2	1995	47.3097	-122.327	257500
3	2	1060	9711	1	1963	47.4095	-122.315	291850
3	1	1780	7470	1	1960	47.5123	-122.337	229500
3	3	1890	6560	2	2003	47.3684	-122.031	323000
3	3	3560	9796	1	1965	47.6007	-122.145	662500
2	1	1160	6000	1	1942	47.69	-122.292	468000
3	1	1430	19901	1.5	1927	47.7558	-122.229	310000

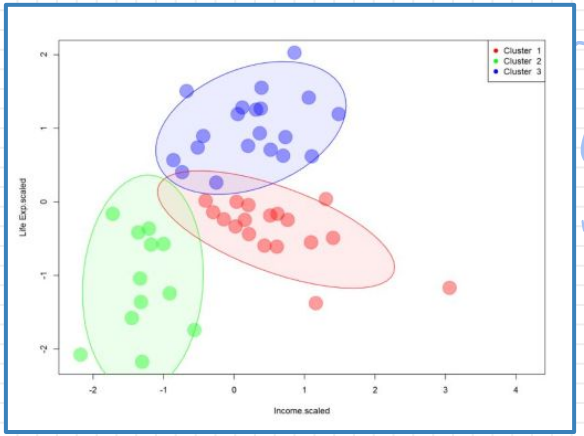
Fuente: https://www.kaggle.com/harlfoxem/housesalesprediction

Modelado – Aprendizaje No Supervisado

El modelo aprende en base a ejemplos que sin ningún tipo de etiqueta.

Modelado – Ejemplo de Clustering

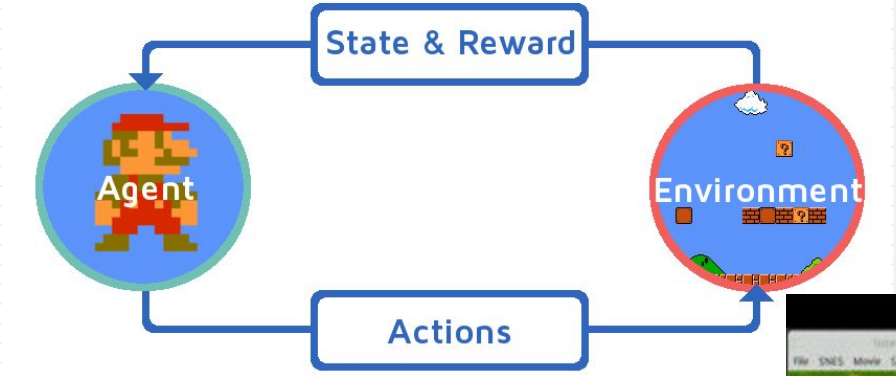
- Segmentación de Clientes basados en sus preferencias de consumo, datos demográficos, intereses, etc.
- Clasificación en la digitalización de documentos.

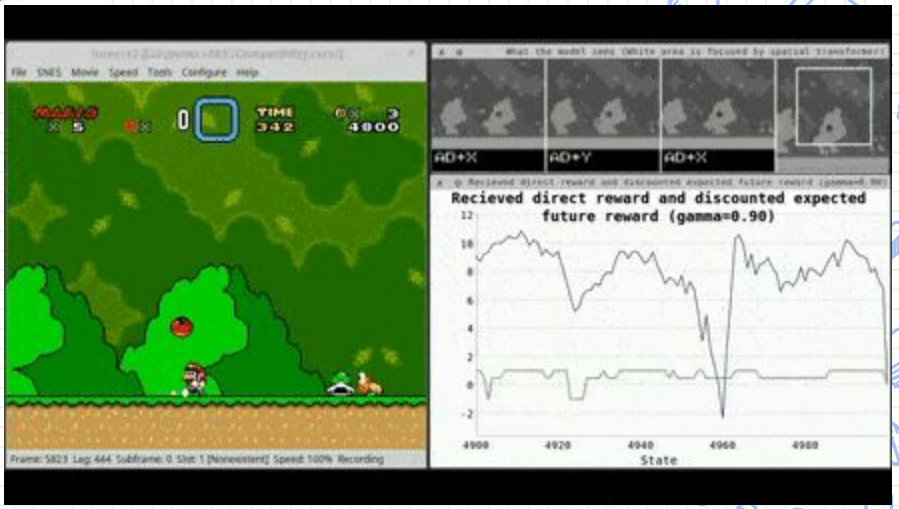


Aprendizaje por refuerzo

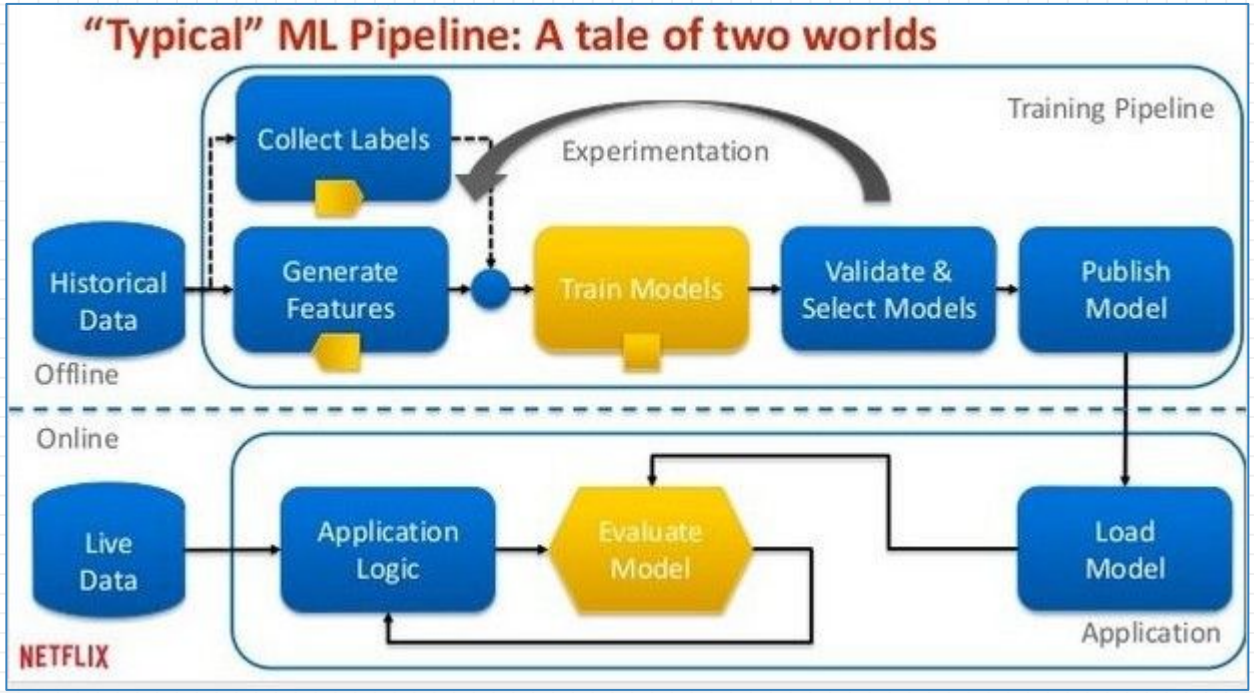
En el tercer tipo de Machine Learning, los modelos aprenden creando su propia experiencia realizando una tarea, optimizando la “recompensa” que pueda obtener ante determinada acción.

Aprendizaje por refuerzo – Ejemplos





Modelado – Ciclo de Vida de ML

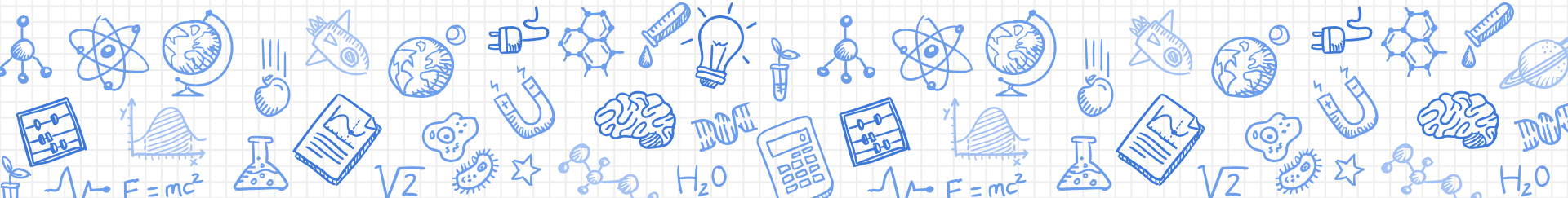


Tools

Natural Language Toolkit (NLTK)

Ingeniería de Datos

Se encarga de que todo funcione eficientemente, estable y escalable.



Gestión de los Datos

- Extracción de datos.
- Transformación de datos.
- Almacenamiento de datos.

Desarrollo de Software

- Data pipelines.
- Herramientas internas.
- APIs.
- Entrenamiento Automático de Modelos.
- Monitoreo.

Ingeniería de Software

- Adopción de Metodologías Ágiles.
- Versionado.
- Mantenibilidad.
- Escalabilidad.
- Modularidad.

Ingeniería de Datos – Desafíos

- Procesamiento Real-Time.
- Multiprocessing y Multithreading.
- Concurrencia y Asincronía.
- Tolerancia a Fallas.
- Almacenamiento Distribuido.

Tools


Spark

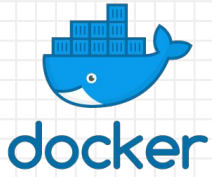

HIVE

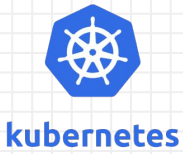

aws


hadoop

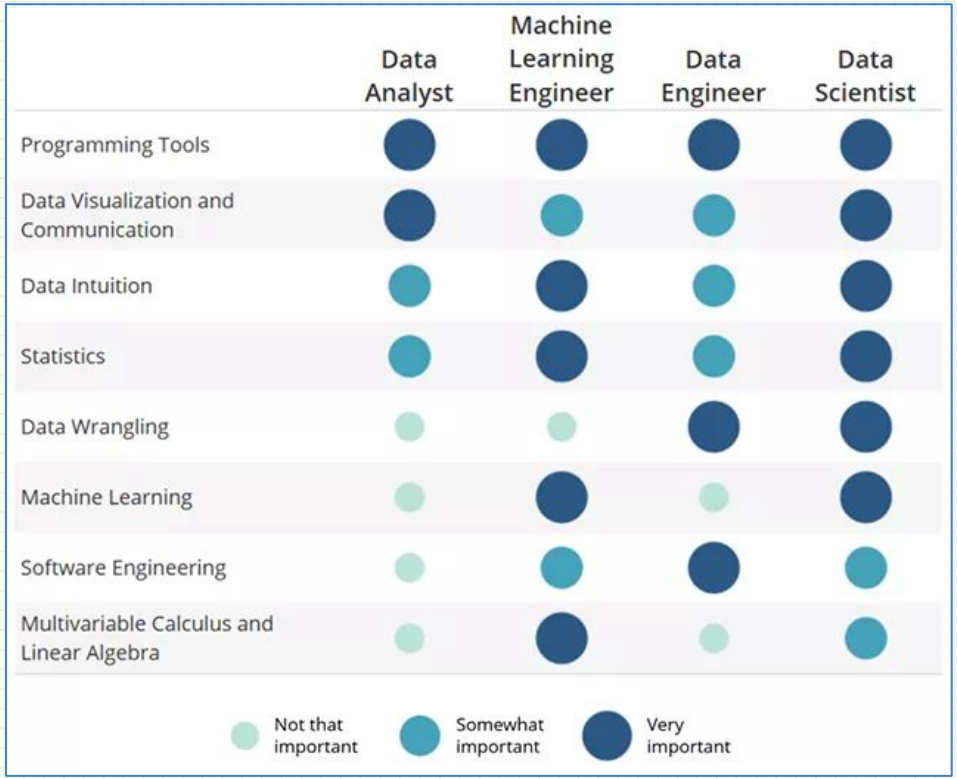

presto


Google Cloud


docker


kubernetes

Skills



Fuente: https://blog.udacity.com/2014/11/data-science-job-skills.html