



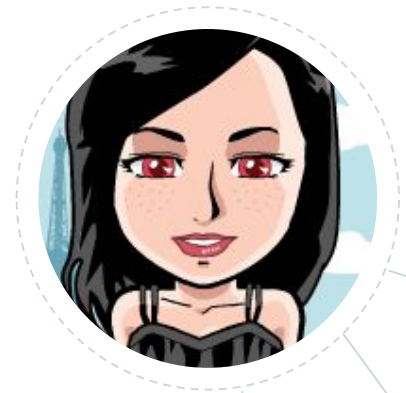
Intro Data Science con Python

Hello!

Soy Isabel Ruiz Buriticá

Me gustan las comunidades para
aprender y enseñar.

Me encuentras como @iris9112





¿Qué es data science?

Data science es una combinación multidisciplinaria de tratamiento de datos, desarrollo de algoritmos, estadística y tecnología para resolver problemas analíticamente complejos

*Data science
se trata de
descubrir que
hay escondido
dentro de los
datos*





**¿Quién es un
Data Scientist?**

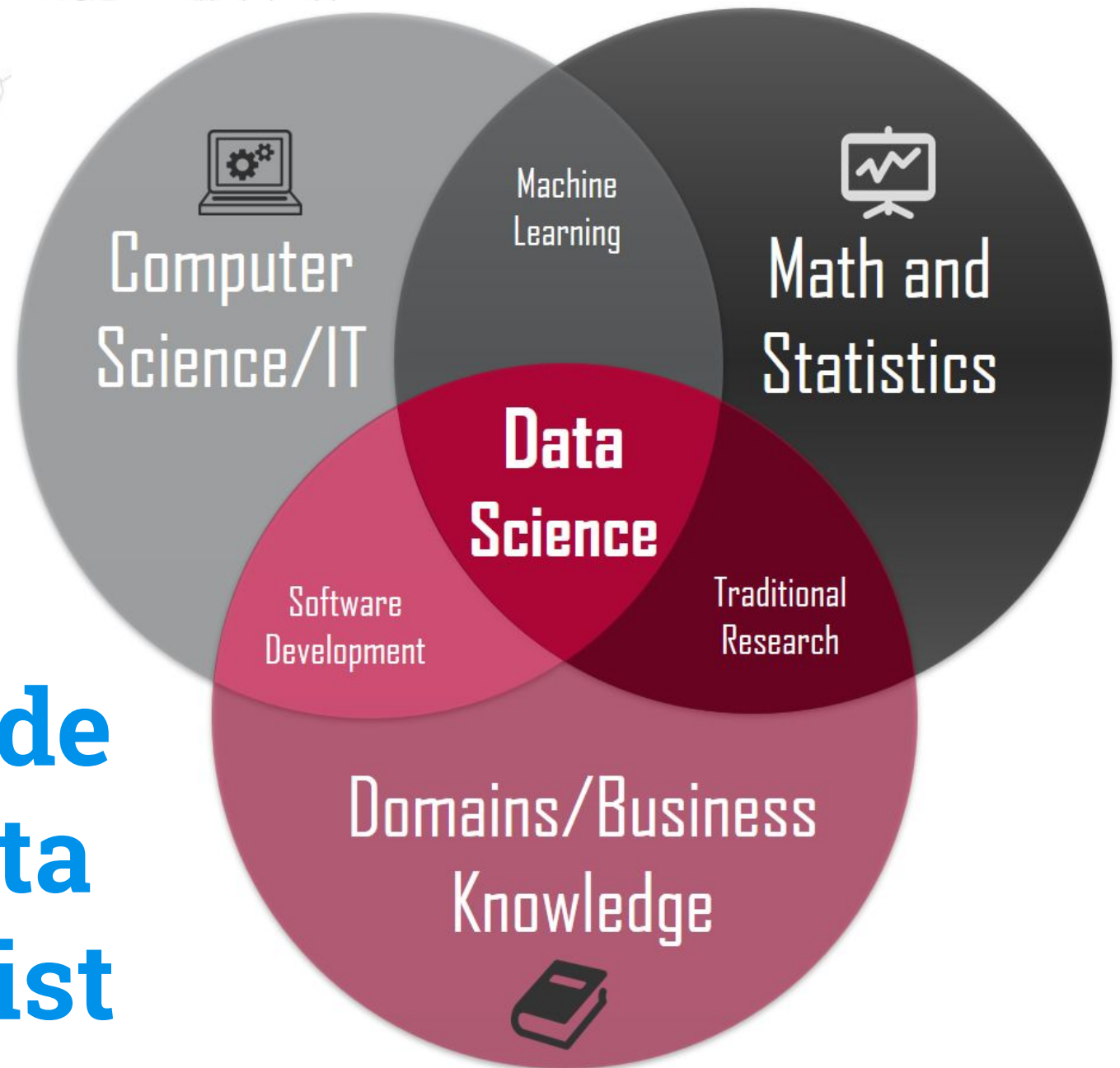


“

*A data scientist is someone who is
better at statistics than any
software engineer and better at
software engineer than any
statistician*

- Josh wills

Skills de un Data scientist



Perfiles

DATA Engineer

Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.



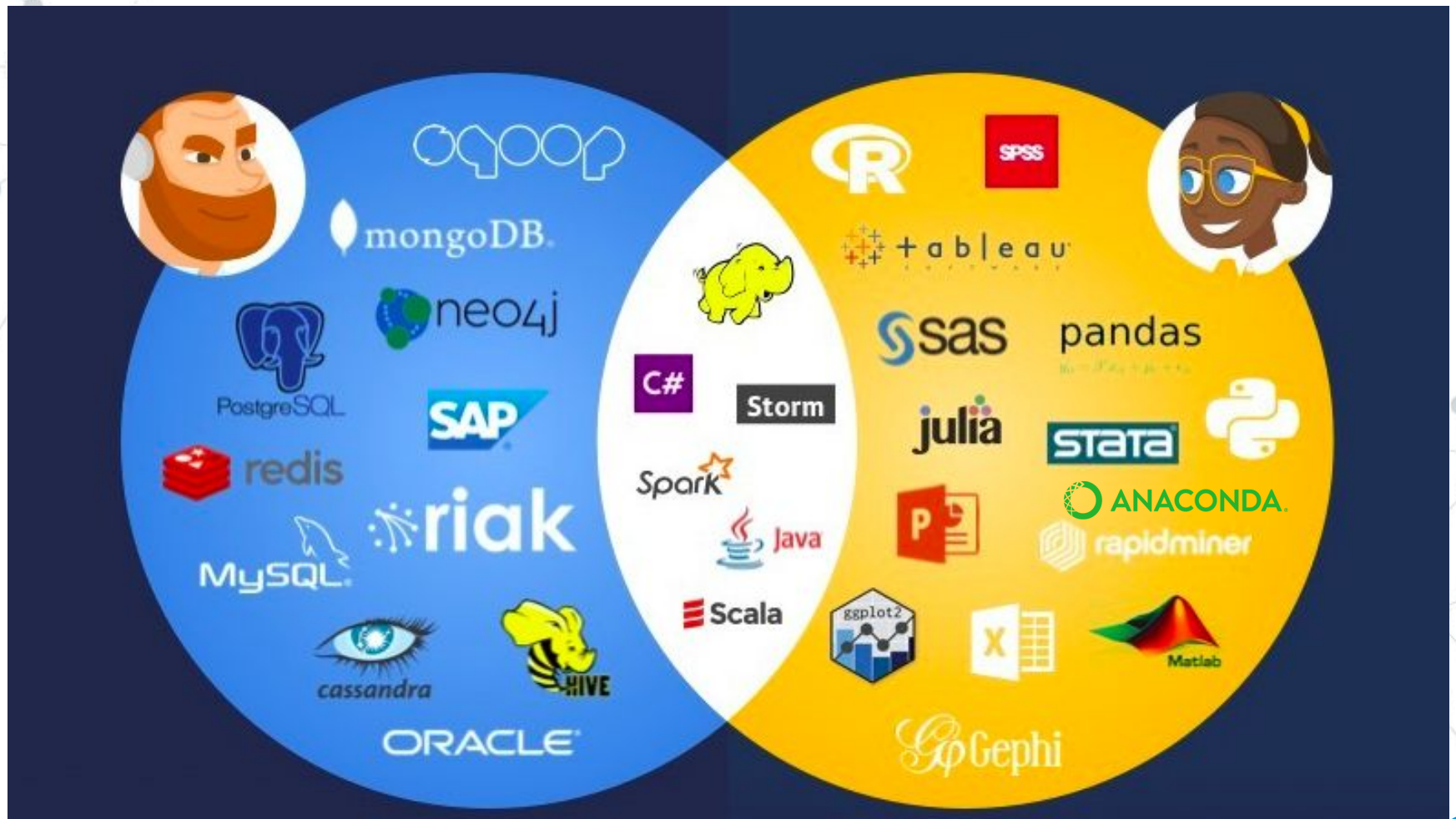
DataCamp
Learn Data Science By Doing

DATA Scientist

Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.



Lenguajes & Herramientas



¿Dónde aprender?



DataCamp



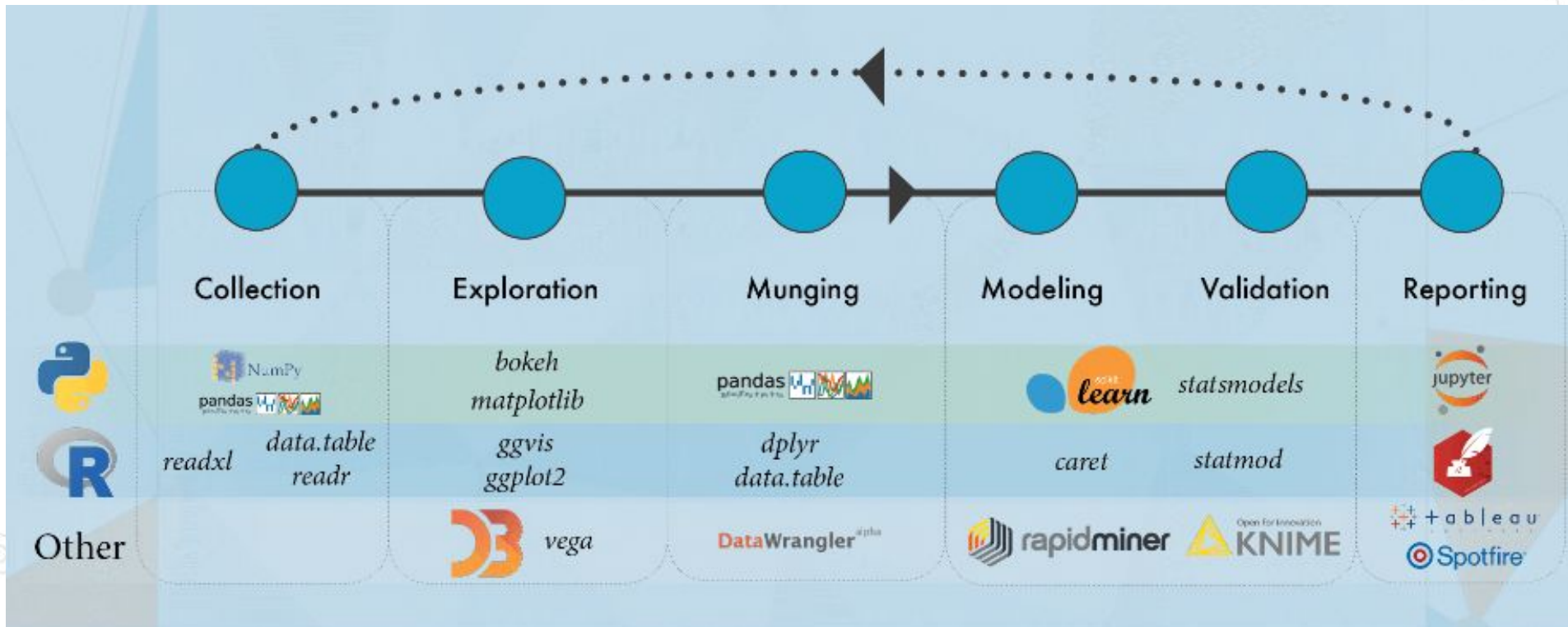
Platzi



“

Ha sido un tópico común que el 80% del valioso tiempo de un científico de datos se invierte simplemente buscando, limpiando y organizando datos, dejando solo un 20% para realizar el análisis.

Flujo de trabajo de un data Scientist



Principales librerías Python



Numpy

Es el módulo básico para la computación científica y la ciencia de datos en Python. Su objeto más usado son los arreglos multidimensionales.



matplotlib

Permite realizar gráficas de datos contenidos en listas o arrays. Proporciona una API, pylab diseñada para recordar a la de MATLAB.



Pandas

Es una extensión de NumPy para manipulación y análisis de datos. Principalmente para tablas numéricas y series temporales. Su principal tipo de dato es el DataFrame.



Scipy

Librería de cálculo numérico de gran capacidad basada en numpy, posee módulos para optimización de funciones, integración, funciones especiales, estadísticas, tratamiento de señales, entre otras.



Scikit-learn
















Este módulo está basado en NumPy y SciPy. Proporciona algoritmos para muchas tareas estándar de ML y minería de datos, como clustering, regresión, clasificación, reducción de dimensionalidad y selección de modelo.



Tensor Flow

Librería para Machine learning y deep learning. Desarrollado por Google. Permite entrenar redes neuronales para detectar y descifrar patrones y correlaciones, análogos al aprendizaje y razonamiento usados por los humanos

Github data Python 2018

Library Name	Type	Commits	Contributors	Releases	Watch	Star	Fork	Commits/ Contributors	Commits/ Releases	Star/ Contributors
 matplotlib	Visualization	25 747	725	70	498	7 292	398	36	368	10
 Bokeh	Visualization	16 983	294	58	363	7 615	2 000	58	293	26
 plotly	Visualization	2 906	48	8	198	3 444	850	61	363	72
Seaborn	Visualization	2 044	83	13	205	4 856	752	25	157	59
<i>pydot</i>	Visualization	169	12	12	17	193	80	14	14	16
 Orange3	Machine learning	22 753	1 084	86	2 114	28 098	14 005	21	265	26
XGBoost LightGBM CatBoost	Machine learning	3277	280	9	868	11 991	5 425	12	364	43
		1083	79	14	363	5 488	1 467	14	77	69
		1509	61	20	157	2 780	369	25	75	46
 eli5	Machine learning	922	6	22	39	672	89	154	42	112
 SciPy	Data wrangling	19 150	608	99	301	4 447	2 318	31	193	7
 NumPy	Data wrangling	17 911	641	136	390	7 215	2 766	28	132	11
 pandas	Data wrangling	17 144	1 165	93	858	14 294	5 788	15	184	12
 StatsModels <small>Statistics in Python</small>	Statistics	10 067	153	21	234	2 868	1 240	66	479	19
 TensorFlow	Deep learning	33 339	1 469	58	7 968	99 664	62 952	23	575	68
PYTORCH	Deep learning	11 306	635	16	816	15 512	3 483	18	707	24
 Keras	Deep learning	4 539	671	41	1 673	29 444	10 964	7	1111	44
dist-keras elephas spark-deep-learning	Distributed deep learning	1125	5	7	41	431	106	225	161	86
		170	13	5	97	913	189	13	34	70
		67	11	3	116	920	206	6	22	84
 Natural Language Toolkit	NLP	13 041	236	24	467	6 405	1 804	55	543	27
 spaCy	NLP	8 623	215	56	425	9 258	1 446	40	154	43
 gensim	NLP	3 603	273	52	415	6 995	2 689	13	69	26
 Scrapy	Data scraping	6 625	281	81	1 723	27 277	6 469	24	82	14 ⁹⁷

Datasets



**DATOS
ABIERTOS**
GOBIERNO DIGITAL COLOMBIA

<https://herramientas.datos.gov.co/es/blog/visualizaciones-de-los-mejores-conjuntos>

<https://www.kaggle.com/neuromusic/avocado-prices>

Kaggle is the place to do data science projects

[See how it works](#) ▶





\$105,000

Es en promedio es el salario anual de un data scientist

700,000 puestos

Es la cantidad de vacantes estimadas para el 2020

59%

De toda la demanda de trabajo está en finanzas y seguros, servicios profesionales e informática.

How does **NETFLIX** recommend movies? Matrix Factorization

Video:

<https://www.youtube.com/watch?v=ZspR5PZemcs>

Repo:

<https://github.com/yanneta/pytorch-tutorials/blob/master/collaborative-filtering-nn.ipynb>

Modelo	Aplicaciones (Ejemplo de uso)
Logistic Regression	Predicción de precios de inmuebles
Fully connected networks	Clasificación
Convolutional Neural Networks	Procesamiento de imágenes para poder encontrar gatitos en las fotos
Recurrent Neural Networks	Reconocimiento de Voz
Random Forest	Detención de fraude
Reinforcement Learning	Enseñarle a la máquina a jugar videojuegos y vencer!
Generative Models	Creación de imágenes
K-means	Crear Clusters a partir de datos sin etiquetar. Segmentar audiencias o Inventarios
k-Nearest Neighbors	motores de recomendación (por similitud/cercanía)
Bayesian Clasifiers	Clasificación de emails: Spam o no

A group of young women, likely students, are gathered in a hallway, smiling and looking towards a smartphone held up by one of them to take a selfie. They are wearing maroon blazers over white t-shirts. Some of the t-shirts have text and a green logo. The background shows a brightly lit hallway with white columns.

¡Vamos a practicar!
Numpy + Pandas + Matplotlib

Referencias

- ◎ Presentation template by [SlidesCarnival](#)
- ◎ DataCamp:
<https://www.datacamp.com/community/blog/data-scientist-vs-data-engineer>
- ◎ Datos Abiertos Col: <https://datos.gov.co/>
- ◎ Kaggle: <https://www.kaggle.com/datasets>
- ◎ <https://www.kaggle.com/learn/overview>
- ◎ IBM analisis:
<https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#f0d11517e3bd>
- ◎ Python Bootcamp Uniandes:
<https://github.com/PythonBootcampUniandes>