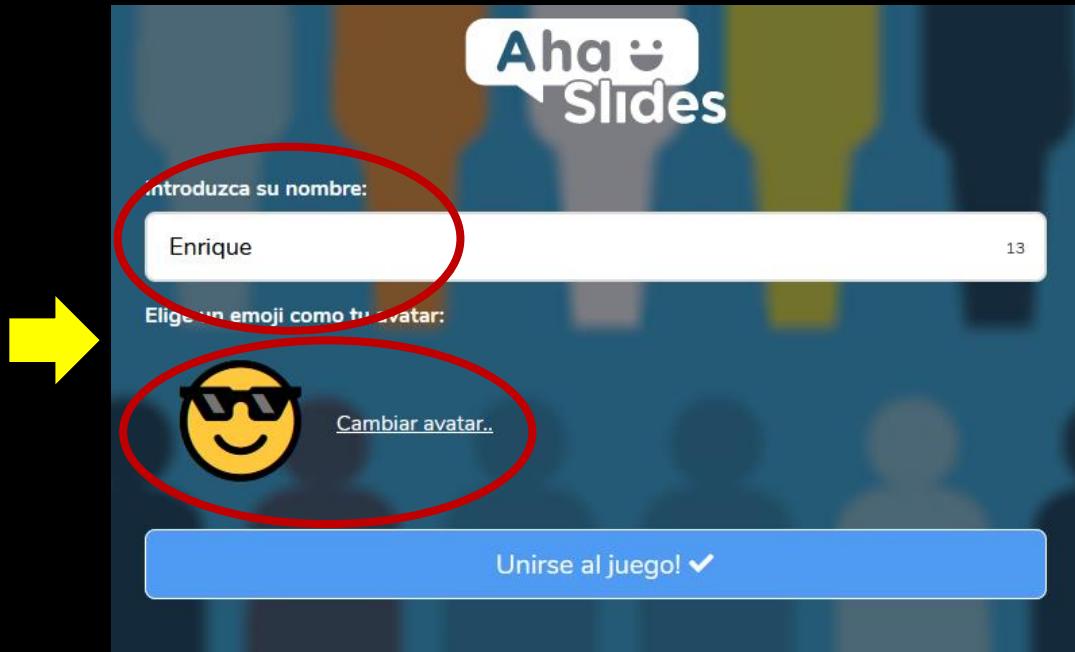


# Bienvenidos al curso de Ciencia de Datos

## Actividad de presentación



<https://ahaslides.com/CDDP01>

# CIENCIA DE DATOS

Profesor: Enrique Camacho  
[enrique.camacho@gmail.com](mailto:enrique.camacho@gmail.com)

CD - 01 - Presentación



**UADY**  
UNIVERSIDAD  
AUTÓNOMA  
DE YUCATÁN

FACULTAD DE INGENIERÍA

 **LAB**  
datos

# Contenido de la clase

- *Presentación individual*
- *Curso*
- *Reglas*
- *Temario*

# ¿Quiénes son los estudiantes?

- Carrera y semestre
- Libro, película, serie, anime
- Tiempo libre



# ¿Quién es el profesor?

**Enrique Camacho Pérez**

- Carrera: Ingeniería Física
- Posgrados: Maestría y Doctorado en Física
- Actualmente:
  - ✓ Evaluador de proyectos de innovación
  - ✓ Profesor e investigador en la FIUADY
  - ✓ LabDatos: proyecto personal
- Proyectos de IA, ciencia de datos y visión por computadora.
- Correr, leer y futbol americano



Científico de datos, la profesión  
más atractiva del siglo XXI

# Científico de datos, la profesión más atractiva del siglo XXI

**Data Scientist: The Sexiest Job of the 21st Century**

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

**W**hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the pressure to join. But users weren't executives had expected. manager put it, "It was like just stand in the corner sign

**Tesla Bot**

**Elon Musk**  
ON ARTIFICIAL INTELLIGENCE

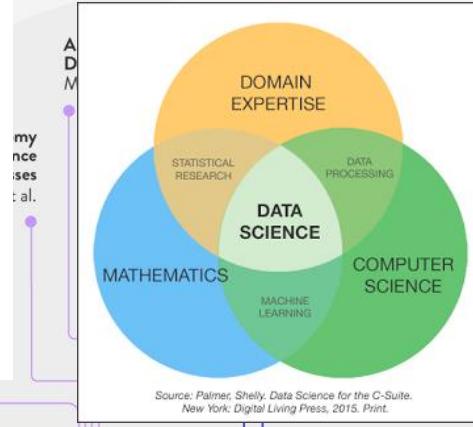
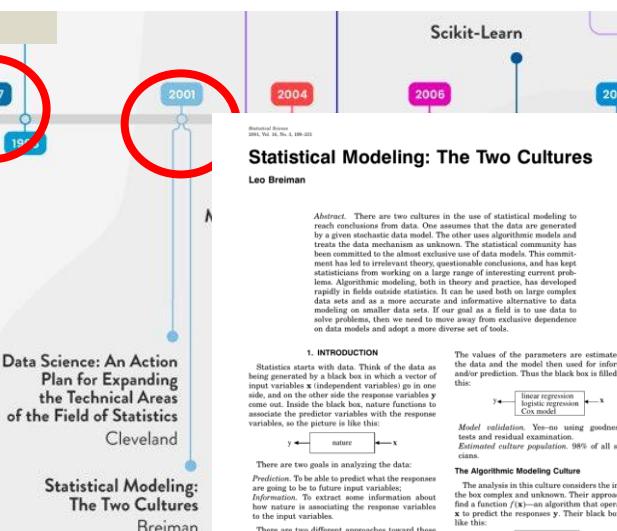
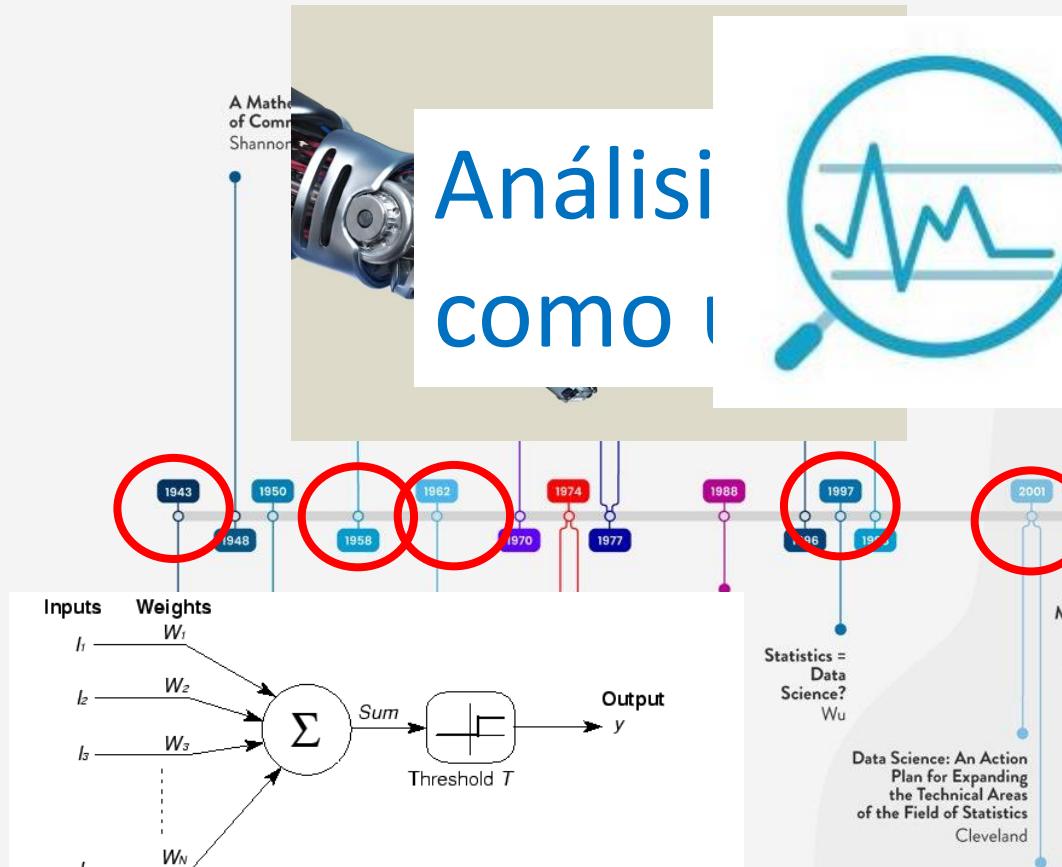
How Netflix Uses AI, Data Science, and Machine Learning — From A Product Perspective

Allen Yu [Follow](#)  
Feb 27, 2019 • 18 min read

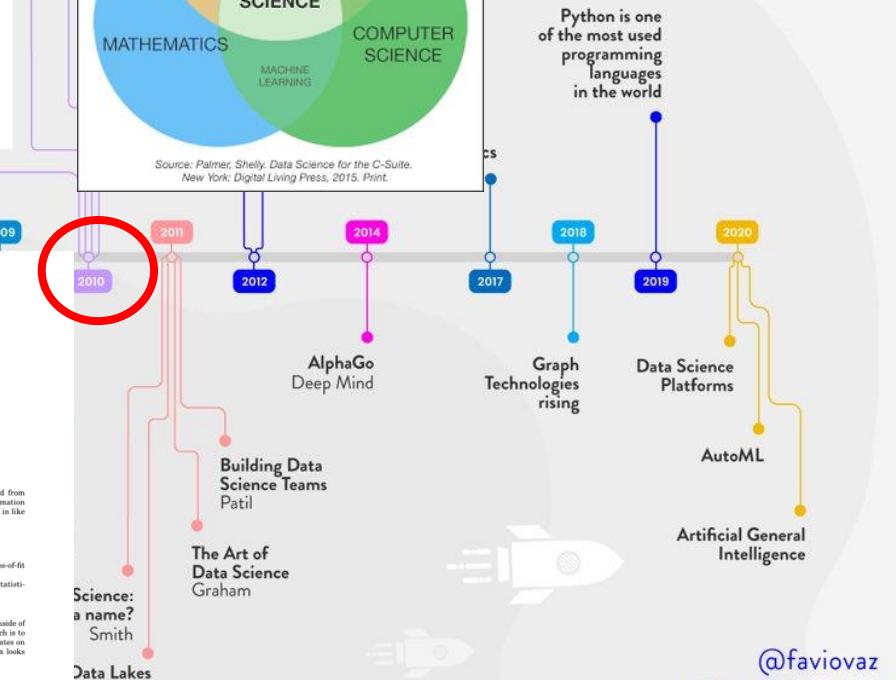


Netflix's machine learning algorithms are driven by business needs.

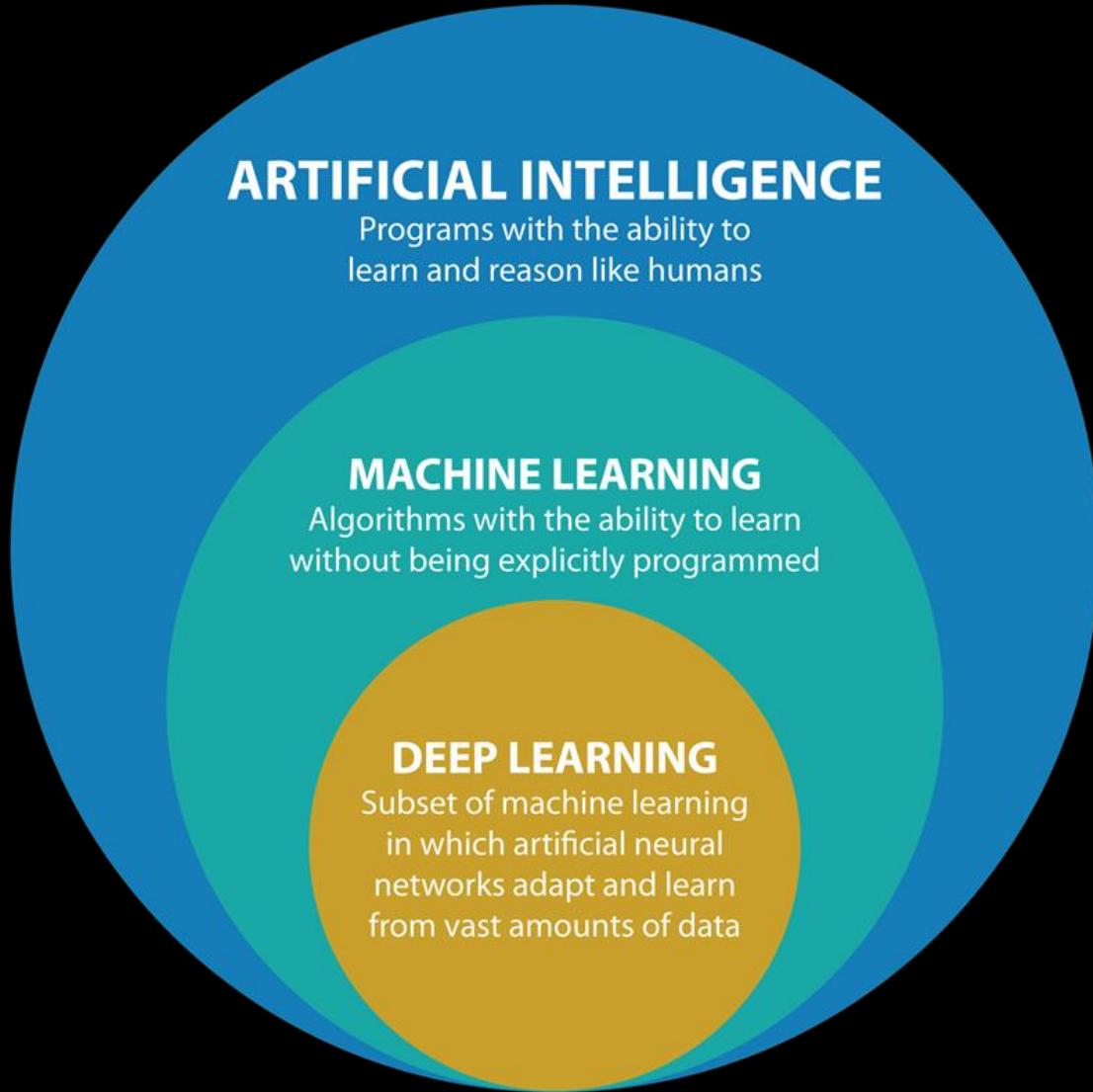
# ¿Desde cuando?



## A SCIENCE TIMELINE v. 2.0



# Estado actual

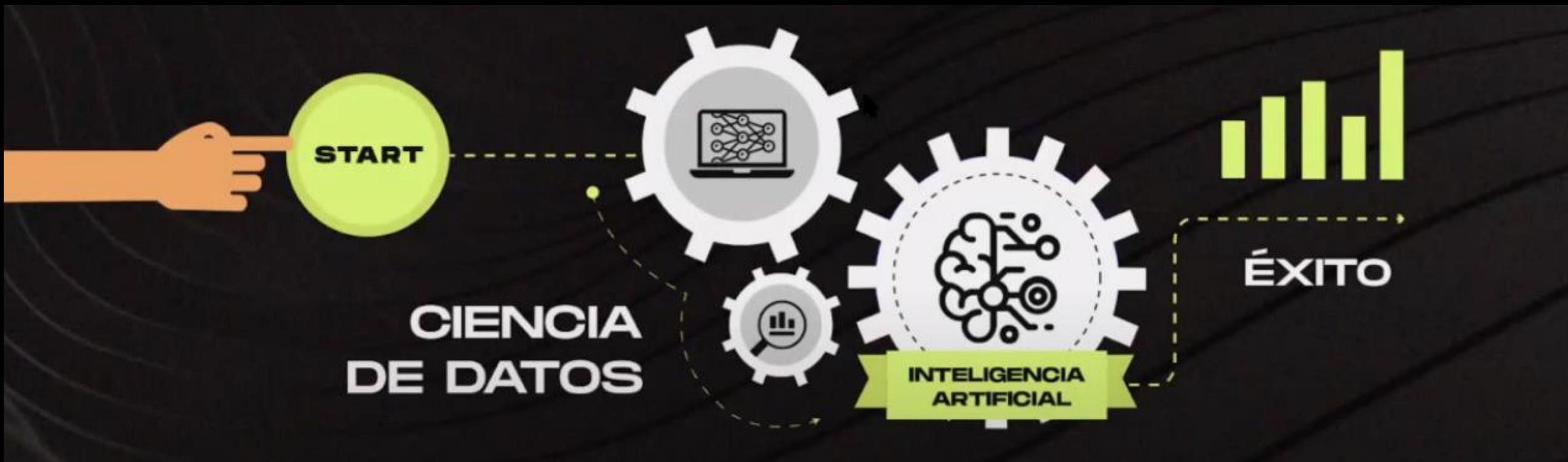


<https://medium.com/@experienciaIA/diferencias-entre-la-inteligencia-artificial-y-el-machine-learning-f0448c503cd4>

# ¿Cómo atacar un problema?



# Papel de la Ciencia de Datos



**ESTO SIGNIFICA QUE TODOS LOS  
PROYECTOS DE CIENCIA DE  
DATOS DEBEN SER:**

REPRODUCIBLES  
FALIBLES  
COLABORATIVOS  
CREATIVOS  
OBEDIENTES A REGULACIONES

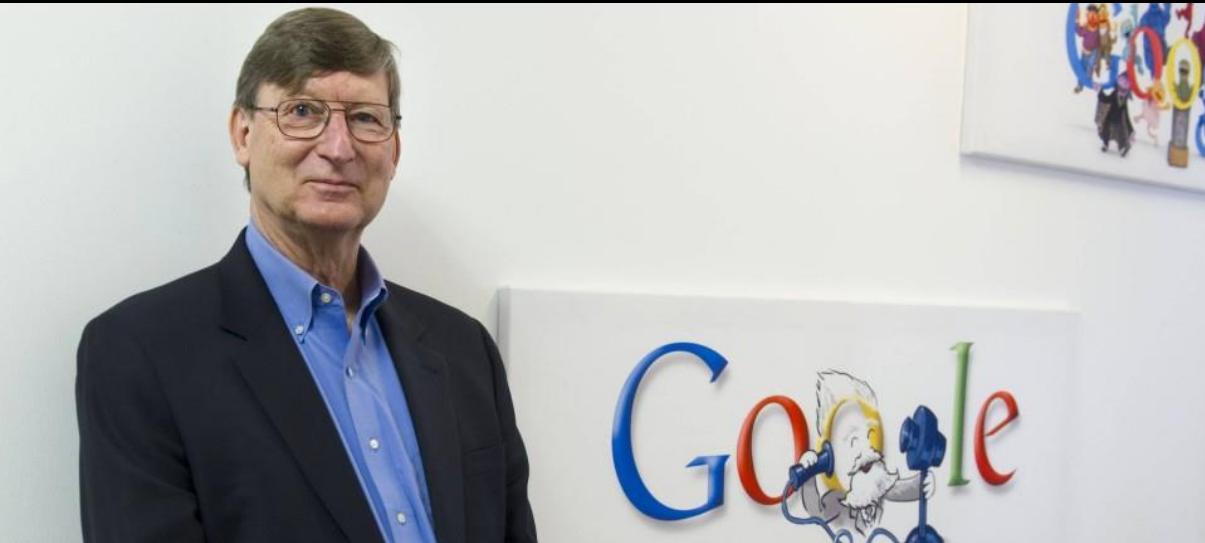


# ¿Qué es Ciencia de datos?

# ¿Qué es Ciencia de datos?

- Es el área que se encarga de la resolución a los problemas de organizaciones a través de las matemáticas, la programación y el método científico qué implica la creación de hipótesis, experimentos y pruebas a través del análisis de datos y generación de modelos predictivos.

# ¿Qué es Ciencia de datos?



“La capacidad de adquirir datos, procesarlos, entenderlos, visualizarlos, **extraer valor** y **saber comunicarlo**”

- Hal Varian economista de Google

**¿Cuáles consideras son las habilidades de los científicos de datos?**



### Habilidades

Programación  
Exploración de datos  
Soluciones creativas

### Conocimientos

Matemáticas  
Estadística

*Hacking Skills*

*Math & Statistics Knowledge*

**Substantive Expertise**

Machine Learning  
Data Science  
Traditional Research

Danger Zone!

### Experticia

Especialización  
Conocimiento de campo

Diagrama de Venn para Data Science  
Drew Conway (2010)



José Antonio Guerrero: uno de los mejores científicos de datos del mundo (Plataforma Kaggle)

“Es una persona con fundamentos en matemáticas, estadística y métodos de optimización, con conocimientos en lenguajes de programación y que además tiene una experiencia práctica en el análisis de datos reales y la elaboración de modelos predictivos. De las tres características quizás la más difícil es la tercera; no en vano la modelización de los datos se ha definido en ocasiones como un arte. Aquí no hay reglas de oro, y cada conjunto de datos es un lienzo en blanco.”

# Metodología de la Ciencia de datos



**¿Como construir un portafolio en Ciencia de datos?**

# How to Build a Data Science Portfolio?

KDnuggets Home » News » 2018 » Jul » Tutorials, Overviews » How to Build a Data Science Portfolio ( 18:n29 )

## How to Build a Data Science Portfolio



<= Previous post      Next post =>

- **How do you get a job in data science?** Knowing enough statistics, machine learning, programming, etc to be able to get a job is difficult. One thing I have found lately is quite a few people **may have the required skills to get a job, but no portfolio**. While a resume matters, having a portfolio of public evidence of your data science skills can do wonders for your job prospects.
- <https://www.kdnuggets.com/2018/07/build-data-science-portfolio.html>

**5. Skills:** Do list technical skills that the job description mentions. The order you list your skills in can suggest what you are best at.

**Proficient:** R, Python, Java, C, C++, SAS, SQL, Matlab, Caffe

**Familiar with:** Theano, MongoDB, Hadoop, JavaScript, HTML, CSS

### Programming:

**Data Science:** Python, Pandas, Numpy, Seaborn, scikit-learn,

**Others :** Shell, Perl, git, Expect/Tcl

**Basics:** HTML, CSS.

### Databases: MySQL

Technical **Machine Learning:** Classification, Regression, Feature engineering, Data scraping (see projects), manipulation and visualization (e.g. Matplotlib, Seaborn, Plotly)

**Statistics:** Regression, Confidence intervals, Bayesian and Monte Carlo methods (e.g. MCMC)

Coding **Python** (scikit-learn, NumPy, SciPy, Pandas, TensorFlow, Keras, PyMongo, Plotly), Git, SQL, MATLAB, R (intermediate), Linux, LaTeX

# Ejemplo de proyectos

Isaac Arroyo  
Draft · 1 min read

Edit

## Analysis of crimes in Mexico during 2017 with Machine Learning techniques (Cluster Analysis) : Comparative Elbow Method and Silhouette Method

Hello word

Import relevant libraries

EDA histogram

Complementation: Elbow Method and Silhouette Method

A map of Mexico showing state boundaries and names, with some states highlighted in different colors.

<https://towardsdatascience.com/analysis-of-crimes-in-mexico-during-2017-with-machine-learning-techniques-cluster-analysis-9c25147dfa86>

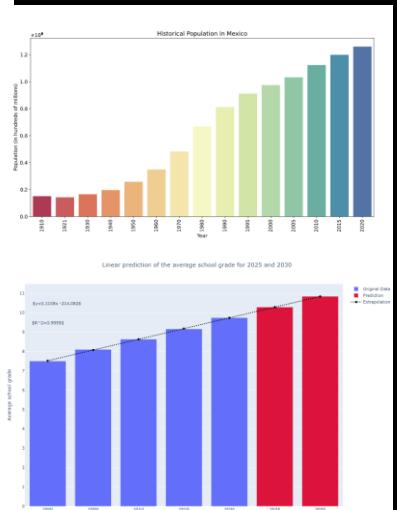


## Exploring the Mexican Census 2020

Aarón Hernández Arcique · Mar 26 · 7 min read

How to access the database and use it to analyze the results

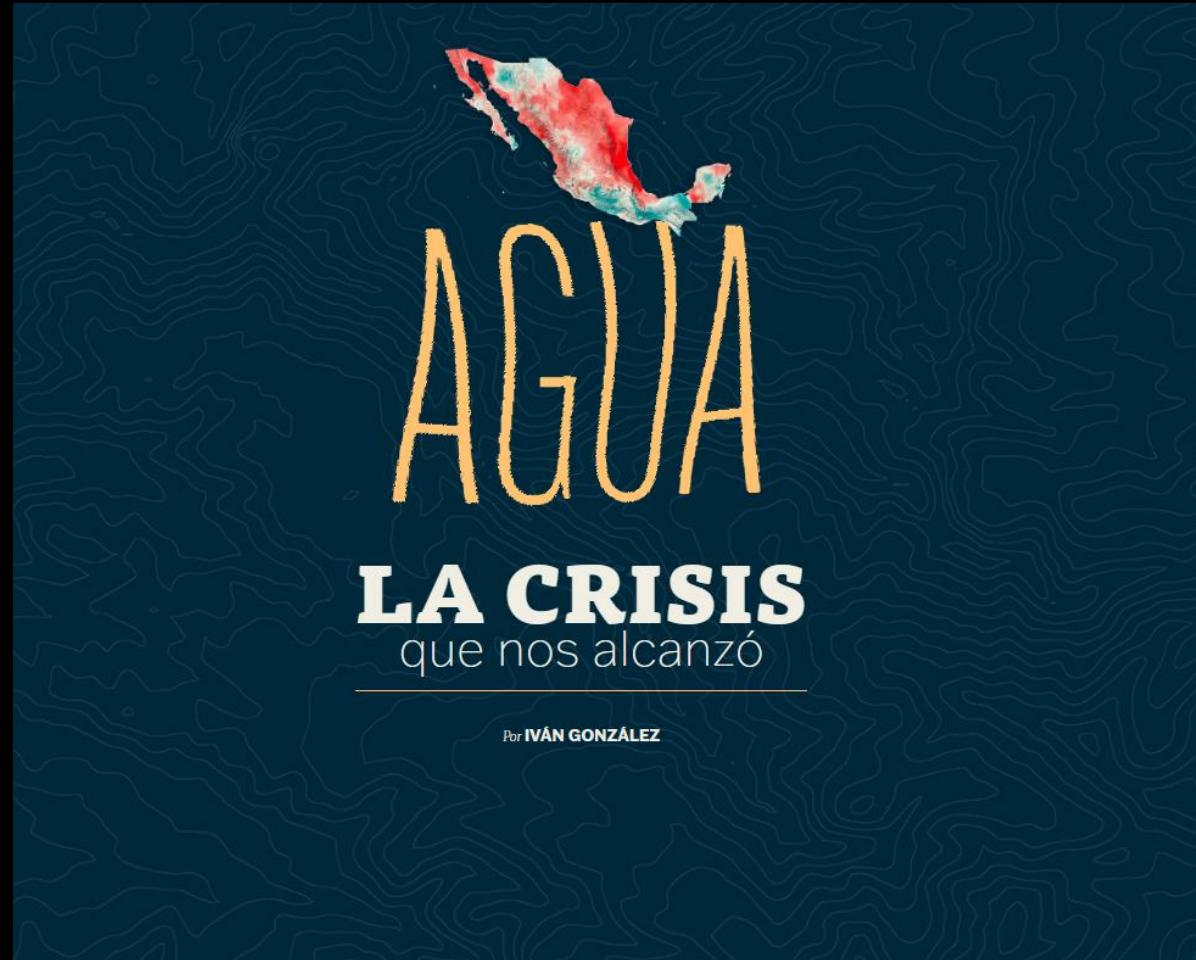
Hernández-Arcique A. and Camacho-Pérez E.

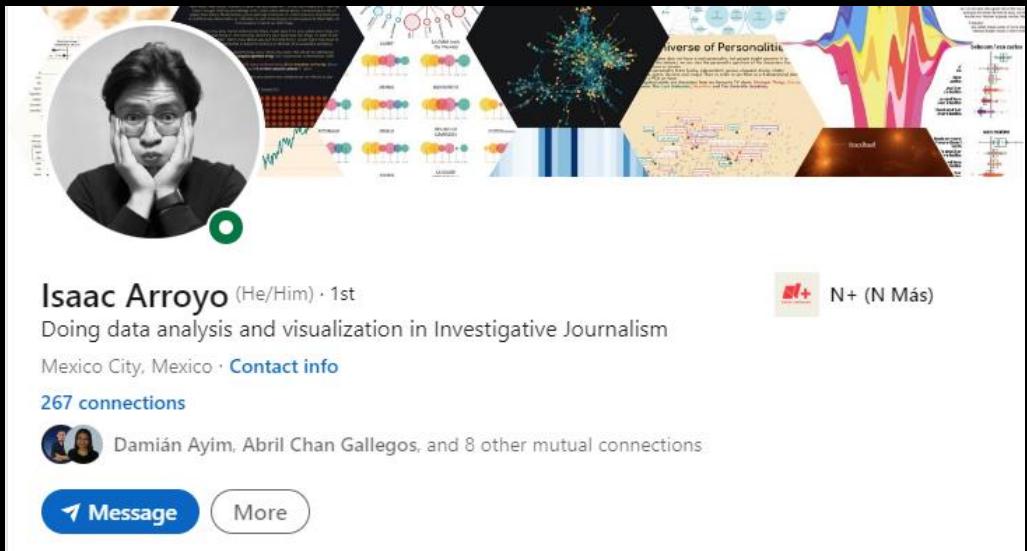


<https://ahernandezarcique.medium.com/exploring-the-mexican-census-2020-1643df67fa9b>

# Agua: la crisis que nos alcanzó | N+ Focus

(nmas.com.mx)





**Isaac Arroyo** (He/Him) · 1st  
Doing data analysis and visualization in Investigative Journalism  
Mexico City, Mexico · [Contact info](#)

267 connections

Damián Ayim, Abril Chan Gallegos, and 8 other mutual connections

[Message](#) [More](#)

## Experience

**Data Analyst**  
N+ (N Más) · Full-time  
Dec 2022 - Present · 9 mos  
Mexico City Metropolitan Area · On-site

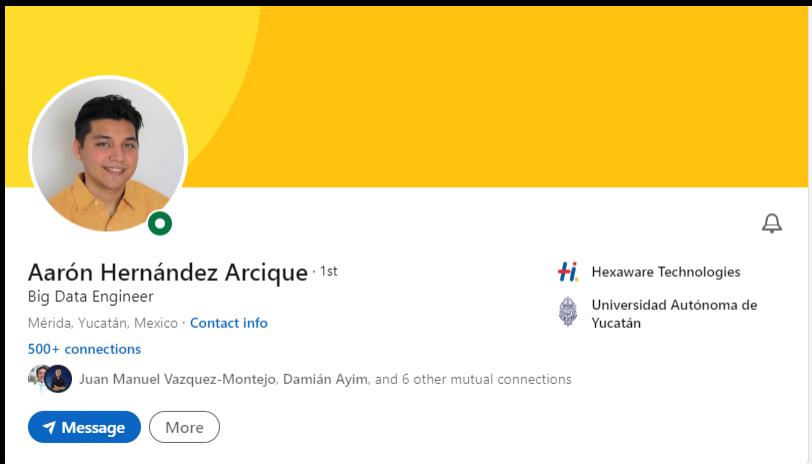
My responsibilities as a Data Analyst in an investigative journalism team are:

- Help reporters enhance their investigations through fact-checking, data gathering and statistics.
- Develop and program data processing scripts using APIs, downloads, and web scraping.
- Collaborate with the design team to create impactful data visualizations for videos and articles.
- Topics I've been involved in: Climate Change and Social Injustice.

I aim to get the audience's attention to communicate data, information and the investigation's results.

Skills: Python (Programming Language) · R (Programming Language) · Data Visualization · Statistics · Adobe Illustrator · Python





Aaron Hernandez Arcique · 1st  
Big Data Engineer  
Mérida, Yucatán, Mexico · [Contact info](#)  
500+ connections

 Hexaware Technologies  
 Universidad Autónoma de Yucatán

 [Message](#)  [More](#)

## Experience



### Big Data Engineer

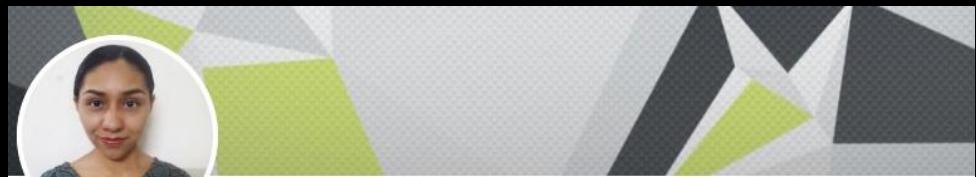
Hexaware Technologies · Full-time

Aug 2022 - Present · 1 yr 1 mo

Mérida, Yucatán, Mexico · Remote

- Design and development of data pipelines in cloud environments (AWS, Azure).
- Data extraction from API data sources.
- Structured and semi-structured databases maintenance.
- Development in Hadoop cluster (HDFS, Hive, Spark, Sqoop).
- Python programming for ETL processes.
- Development of SQL databases and writing applications to interface with SQL database.
- Development of POCs using AWS Machine Learning related services (Comprehend, Kendra, Transcribe).

**Skills:** Amazon Dynamodb · Amazon Kendra · Big Data · Data Pipelines · Python (Programming Language) · Hadoop · Amazon Web Services (AWS) · Gestión de API · NoSQL · PostgreSQL



**Sofía Briceño** · 2nd  
Data Science | Machine Learning  
Mérida, Yucatán, Mexico · [Contact info](#)

83 connections

 Damián Ayim, Isaac Arroyo, and 5 other mutual connections

[Connect](#) [Message](#) [More](#)

 **VinkOS**  
 Universidad Autónoma de  
Yucatán

## Experience

### Data Consultant

VinkOS · Full-time

Sep 2022 - Present · 1 yr

- ETL/ELT procedures and data science models.

**Skills:** Apache Spark · Neo4j · Pentaho

---

### Production Support Analyst

Tata Consultancy Services · Full-time

Apr 2022 - Sep 2022 · 6 mos

Guadalajara, Jalisco, Mexico

- Monitoring and follow-up of incidents, using Linux.
- Debugging in SQL developer.

**Skills:** Shell Scripting · Oracle SQL Developer

---

### Data scientist Jr.

Grupo Bimbo · Full-time

Nov 2021 - Feb 2022 · 4 mos

Mexico City, Mexico

- Extraction and transformation of data, development of time series models and preparation of reports.
- Manage complex projects based on Data Science and collaboration with various work teams for the. ...see more

**Skills:** Time Series Analysis



**Isabel Abigail Valadez Polanco** (She/Her) · 2nd  
Engineering physicist  
Mérida, Yucatán, Mexico · [Contact info](#)

163 connections

 Damián Ayim, Abril Chan Gallegos, and 3 other mutual connections

[Connect](#) [Message](#) [More](#)

## Experience

 **Commercial Analyst**  
Rappi · Full-time  
May 2023 - Present · 4 mos

 **Data Analyst**  
Materama · Full-time  
Sep 2022 - Apr 2023 · 8 mos  
On-site  
**Skills:** SQL · Analítica de datos · Transact-SQL (T-SQL) · MySQL · Microsoft Power BI · Microsoft Excel

 **Data Analyst Intern**  
BEDU · Internship  
Aug 2021 - Dec 2021 · 5 mos  
Mérida, Yucatán  
**Practicante como analista de datos en el área de nuevos productos.**  
**Skills:** Web scraping · Análisis de datos · Python

# Portafolio de cursos



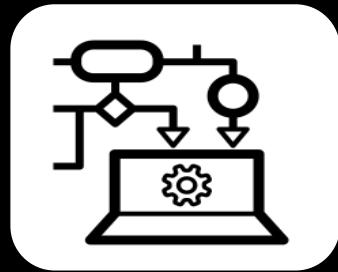
## 0. Adquisición de datos

- Diseño e implementación de Dataloggers para la adquisición de datos (Solo en verano)



## 1. Introducción a ciencia de datos

- Principales herramientas de Python



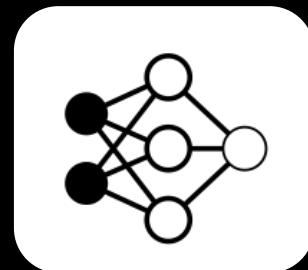
## 2. Ciencia de datos

- Herramientas avanzadas de manejo de datos
- Machine Learning



## 3. Conducción autónoma de vehículos

- Algoritmos básicos para la conducción autónoma



## 4. Deep Learning para visión por computadora

- Aplicación de arquitecturas de CNN

# Portafolio de cursos (Actualizado)



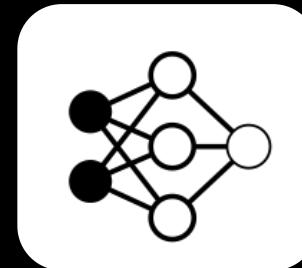
## 0. Aplicaciones con tarjetas de desarrollo

Diseño e implementación de proyectos en el Arduino UNO y en la Raspberry Pi  
*(Solo en verano)*



## 1. Ciencia de datos

- Principales herramientas de Python
- Herramientas avanzadas de manejo de datos
- Machine Learning



## 2. Deep Learning para visión por computadora

- Aplicación de arquitecturas de CNN
- Aplicación en la conducción autónoma de vehículos

## Taller de adquisición de datos

- Diseño e implementación de Dataloggers para la adquisición de datos (Solo en verano)

## Club de robots de competencia

- Asesorías
- Apoyo de material
- Herramientas de prototipado (corte laser e impresora 3D)



# ¿Qué esperan de este curso?



# Curso de Ciencia de Datos

# Compromisos de los estudiantes

- *Estar al pendiente del material y actividades que el profesor envié a la plataforma*
- *Preguntar cuando se tenga alguna duda a través del correo electrónico o la plataforma*
- *No copiar las tareas, en caso de detectarlo la calificación se repartirá.*
- *Subir en tiempo las tareas asignadas. (Las tareas tendrán una penalización de -5 puntos por cada día de retraso)*



# Compromisos del profesor

- Preparar material adecuado para el curso
- Colocar con tiempo suficiente el material en la plataforma
- Responder las dudas que existan
- Estar a tiempo para las clases



- *Número de WhatsApp: 9991730275*
- *Horario preferente para preguntas: Lunes a viernes de 9am-1pm*

# Compromisos del profesor

- Preparar material adecuado para el curso
- Colocar con tiempo suficiente el material en la plataforma
- Responder las dudas que existan
- Estar a tiempo para las clases

- *Número de WhatsApp: 9991730275*
- *Horario preferente para preguntas: Lunes a viernes de 9am-1pm*

# Evaluación del curso

- La calificación del curso es 100% con tareas
- La calificación final es el promedio de todas las tareas
- Todas las tareas tendrán cerca de 7 días para su realización (se contarán con al menos dos clases)

## *Consideraciones*

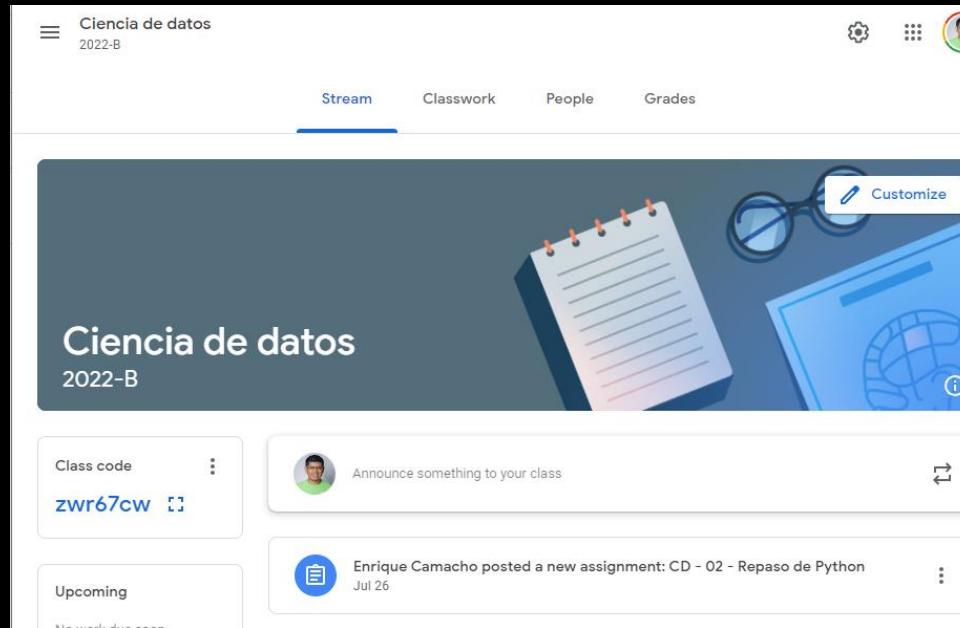
- *Por cada día que pase se penalizará con -5 puntos.*
- *El sistema detecta de manera automática cuando la tarea se entrega tarde.*
- *Todas las tareas tienen límite las 11:59pm*

# Acerca del material

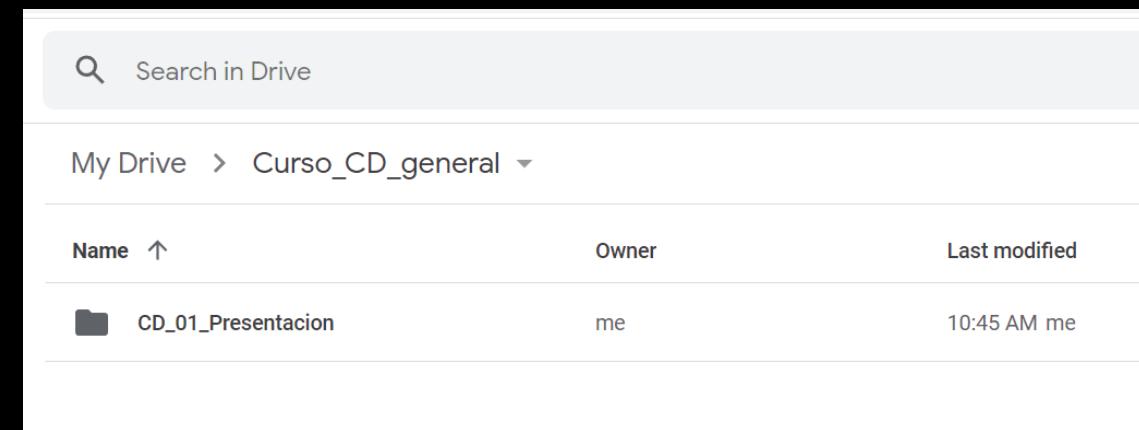
- Todo el material estará en la plataforma GoogleDrive y Google Class (código **cv6scrn**).
- **Utilizar nombre y apellido** para darse de alta en la plataforma



[https://drive.google.com/drive/folders/1iNRwzw3yf7xZzNAZvGkI0KEI0S9b\\_t-7?usp=sharing](https://drive.google.com/drive/folders/1iNRwzw3yf7xZzNAZvGkI0KEI0S9b_t-7?usp=sharing)



A screenshot of the Google Classroom Stream interface. At the top, it shows the class name "Ciencia de datos" and the year "2022-B". Below this, there are tabs for "Stream", "Classwork", "People", and "Grades". The "Stream" tab is selected and shows a large, stylized graphic of a notepad, glasses, and a globe. To the left of the graphic, the class name and year are displayed again. On the right side of the stream, there are two notifications: one from "zwr67cw" with the message "Announce something to your class" and another from "Enrique Camacho" with the message "posted a new assignment: CD - 02 - Repaso de Python".



A screenshot of the Google Drive interface showing a folder list. The search bar at the top contains "Search in Drive". Below it, the path "My Drive > Curso\_CD\_general" is shown. A table lists the contents of the folder:

Name	Owner	Last modified
CD_01_Presentacion	me	10:45 AM me

# Grupo de WhatsApp

- Para tener una comunicación más rápida se creará un grupo de WhatsApp.



<https://chat.whatsapp.com/J1LK3VKEExDo9vtyCVkNLuk>



# Ejemplo de actividades



2 Instrucciones

Descargar el archivo ZIP que contiene el archivo:

- CD\_05\_Pandas\_5\_DatosFaltantes

CIENCIA DE DATOS  
Dr. Enrique Camacho

Pandas -Parte 5  
Manejo de datos faltantes

Before

	set_of_numbers
0	1.0
1	2.0
2	3.0
3	4.0
4	5.0
5	6.0
6	7.0
7	8.0
8	9.0
9	10.0
10	NaN
11	NaN

After

	set_of_numbers
0	1.0
1	2.0
2	3.0
3	4.0
4	5.0
5	6.0
6	7.0
7	8.0
8	9.0
9	10.0
10	0.0
11	0.0

3 Instrucciones

- Ejecutar una a una cada celda y resolver los ejercicios propuestos en la notebook.
- Una vez terminada la revisión y los ejercicios, súbalos a la plataforma con el siguiente nombre:  
CD\_05\_Pandas-parte5\_Nombre\_Apellido.ipynb

CIENCIA DE DATOS  
Dr. Enrique Camacho

UADY  
UNIVERSIDAD AUTÓNOMA  
DE YUCATÁN  
FACULTAD DE INGENIERÍA

## Pandas -Parte 5 Manejo de datos faltantes

Before

	set_of_numbers
0	1.0
1	2.0
2	3.0
3	4.0
4	5.0
5	6.0
6	7.0
7	8.0
8	9.0
9	10.0
10	NaN
11	NaN

After

	set_of_numbers
0	1.0
1	2.0
2	3.0
3	4.0
4	5.0
5	6.0
6	7.0
7	8.0
8	9.0
9	10.0
10	0.0
11	0.0

• Vines McKinney, comenzó a desarrollar Pandas en el año 2008 mientras trabajaba en AQR Capital [https://www.aqr.com] por la necesidad que tenía de una herramienta flexible de alto rendimiento para realizar análisis cuantitativos en datos financieros.  
• Antes de dejar AQR convenció a la administración de la empresa de distribuir esta biblioteca bajo licencia de código abierto.  
Pandas es un acrónimo de Panel DATA analysis

### Librerías

```
In [1]: import numpy as np  
import matplotlib.pyplot as plt
```

### Introducción

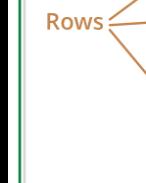
La diferencia entre los datos encontrados en muchos tutoriales y los datos del mundo real es que los datos del mundo real rara vez están limpios y homogéneos. En particular, muchos conjuntos de datos interesantes tendrán cierta cantidad de datos faltantes. Para complicar aún más las cosas, diferentes fuentes de datos pueden indicar datos faltantes de diferentes maneras.

En esta lección se discutirán algunas consideraciones generales para los datos faltantes, discutiremos cómo Pandas elige representarlos y se mostrarán algunas herramientas integradas de Pandas para manejar datos faltantes en Python. Se hará referencia a los datos faltantes de manera general como valores "nulo", "NaN" o "NA".

### Datos faltantes en pandas

En el desarrollo de Pandas se eligió usar sentinelas para los datos faltantes, y se eligió usar dos valores nulos de Python ya existentes: el valor especial de punto flotante `NaN` y el objeto Python `None`. Esta elección tiene algunos efectos secundarios, pero en la práctica termina siendo una buena elección en la mayoría de los casos de interés.

# Programa del curso

1. Pandas
  2. Visualización
  3. Machine Learning
  4. Tópicos actuales y proyecto

The diagram shows a green-bordered box containing a portion of a Pandas DataFrame. The columns are labeled '0' through '6'. The first column contains names: 'Av', 'John', 'Jori', 'Jori', 'Ter', 'Jan', and 'Eva'. The second column is labeled 'Rows' and has three orange arrows pointing to the first three rows of the first column.

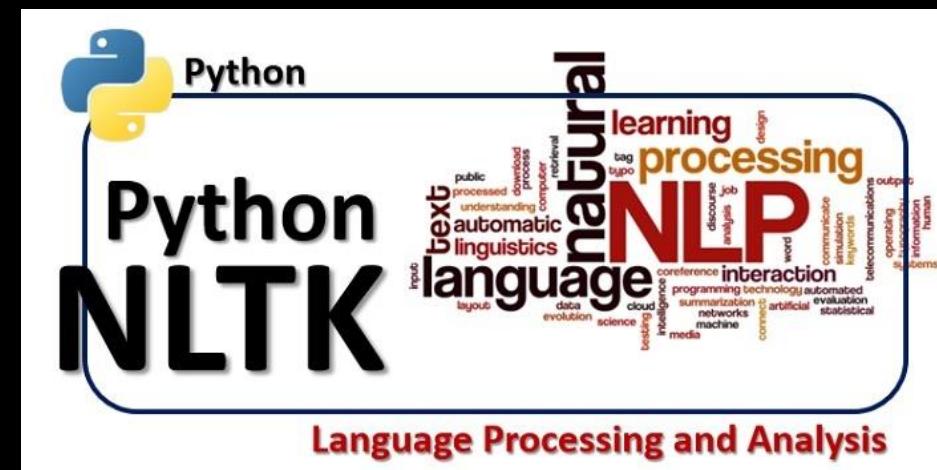
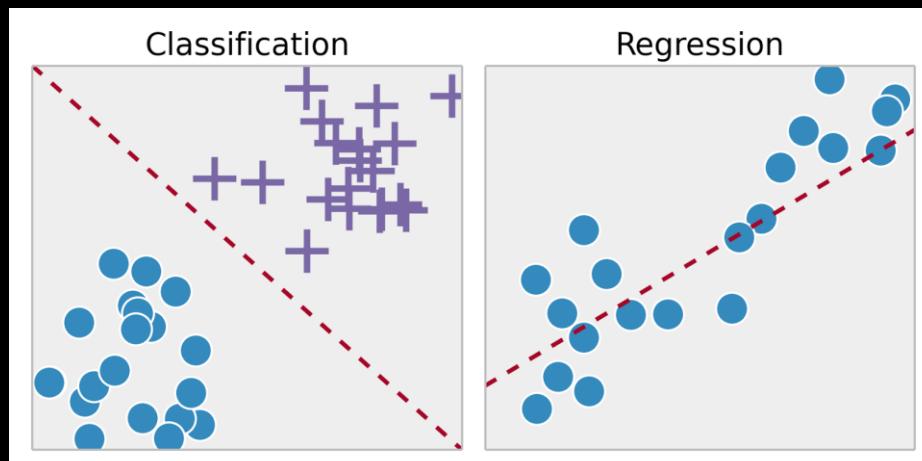
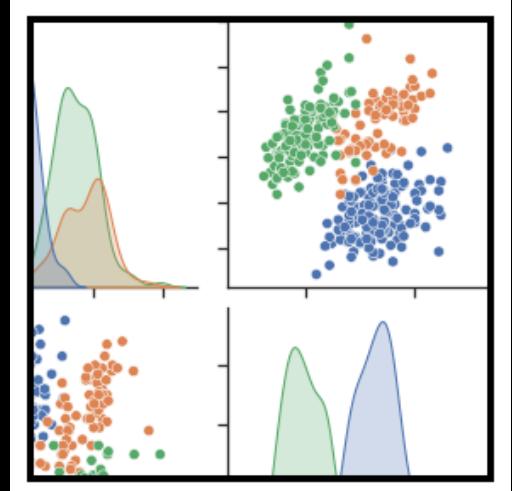
0	1	2	3	4	5	6
Av						
John						
Jori	Rows					
Jori						
Ter						
Jan						
Eva						

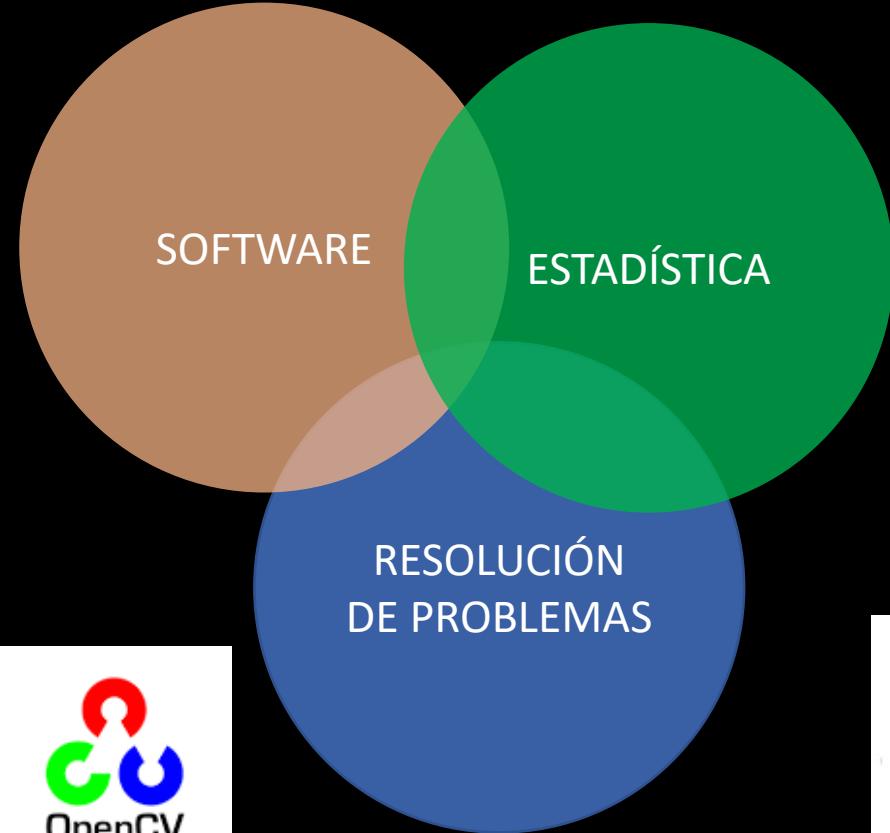
**Columns**

	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

**Rows**

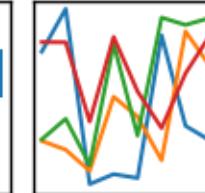
**Data**



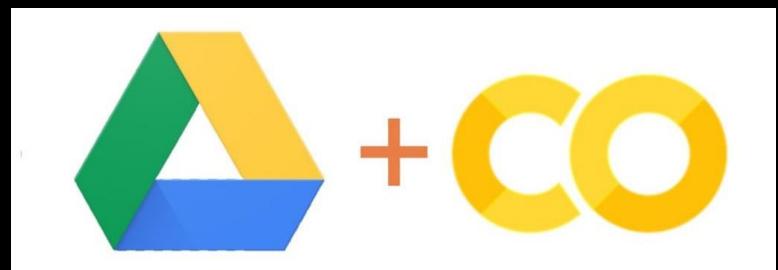
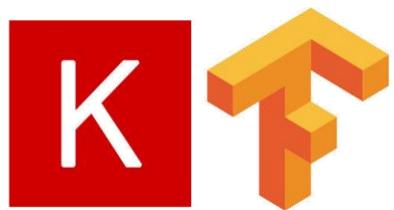


pandas

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



seaborn



# CALENDARIO DE ACTIVIDADES AGOSTO - DICIEMBRE DE 2023

							INICIO DEL PERÍODO ESCOLAR	FIN DEL PERÍODO ESCOLAR	DÍAS INHÁBILES
L	M	M	J	V	S	D			
					1	2			
3	4	5	6	7	8	9			
10	11	12	13	14	15	16			
17	18	19	20	21	22	23			
24	25	26	27	28	29	30			
31									
							AGOSTO 2023	SEPTIEMBRE 2023	
L	M	M	J	V	S	D	L	M	
		1	2	3	4	5		1	
7	8	9	10	11	12	13		2	
14	15	16	17	18	19	20		3	
21	22	23	24	26	27	28			
	29	30	31						
							OCTUBRE 2023	NOVIEMBRE 2023	DICIEMBRE 2023
L	M	M	J	V	S	D	L	M	
					1			1	
2	3	4	5	6	7	8		2	
9	10	11	12	13	14	15		3	
16	17	18	19	20	21	22		4	
23	24	25	26	27	28	29		5	
30	31								

- **30 sesiones**
- **9 Diciembre última clase/asesoría**
- **7 Diciembre presentación de proyectos o tarea final**
- **12 Diciembre entrega de calificación final**

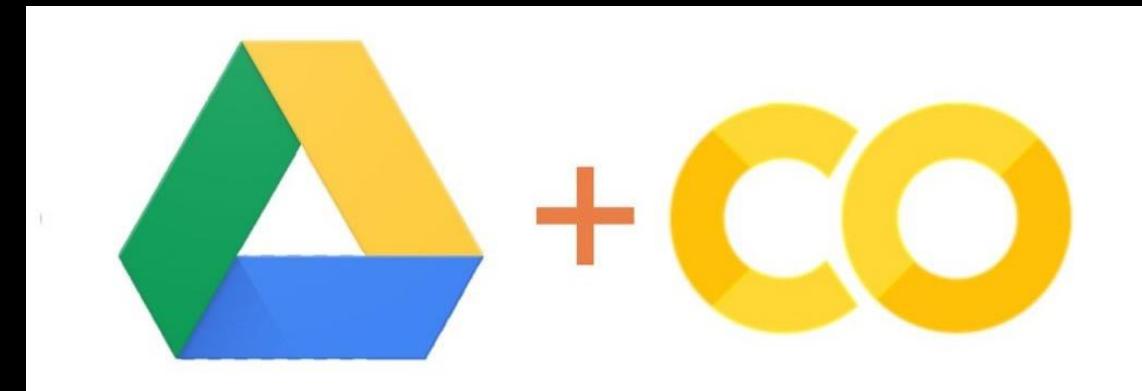
# Para iniciar el curso

Tener instalado en Windows, Linux o Mac:

- Python3
- Jupyter
- Numpy, SciPy, Matplotlib, Numpy y Pandas
- \*Notepad++

# Plataformas del curso

Familiarizarse con las plataformas de trabajo:



# Tarea 1

1. Realizar el tutorial de Markdown y capturar la pantalla final para evidencia. El archivo debe llamarse markdown\_Nombre\_Apellido. (jpg,png, pdf)

- <https://www.markdowntutorial.com/>
- <https://www.markdowntutorial.com/es/>

2. Realizar el tutorial del siguiente link (<https://youtu.be/O7KR0TEPiBs>) utilizando una grafica diferente(será necesario editar su propio archivo CSV). Subir a plataforma los siguientes archivos:

- El archivo de Python con el nombre idlePython\_Nombre\_Apellido.py
- El archivo de Jupyter con el nombre Jupyter\_Nombre\_Apellido.ipynb
- Descargue el archivo de Google Colab y suba a la plataforma el archivo con el nombre Colab\_Nombre\_Apellido.ipynb

Evidencia para subir:

Congratulations!

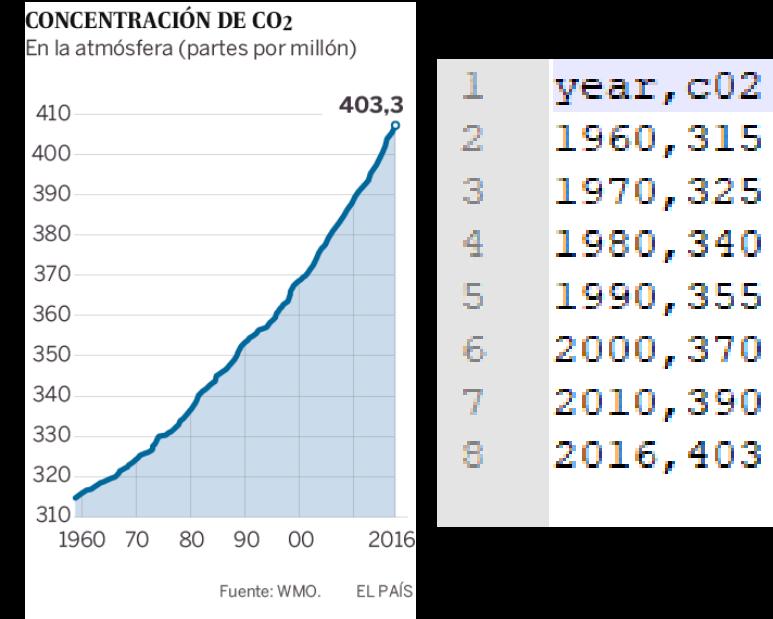
You've completed all the lessons!

Believe it or not, we've only just begun exploring what can be accomplished with Markdown. There are many "extended" implementations of Markdown that support formats like tables, definition lists, footnotes, and more. Because they're non-standard, they're not essential to learning the basics, as we've introduced here.

If you'd like to know more about these Markdown implementations, you're welcome to explore any number of other Markdown apps and tutorials. Here are just a few:

- <https://daringfireball.net/projects/markdown/>
- <https://spec.commonmark.org/dingus/>
- <https://johnmacfarlane.net/babelmark2/faq.html>
- <https://www.markdownguide.org>

Ejemplo de grafica y datos



¡NO SUBIR ARCHIVOS COMPRIMIDOS (ZIP, TAR, RAR, etc)!