



Data Science

WWW.DATASCIENCE.PE

SOBRE NOSOTROS

CONVERSATORIO

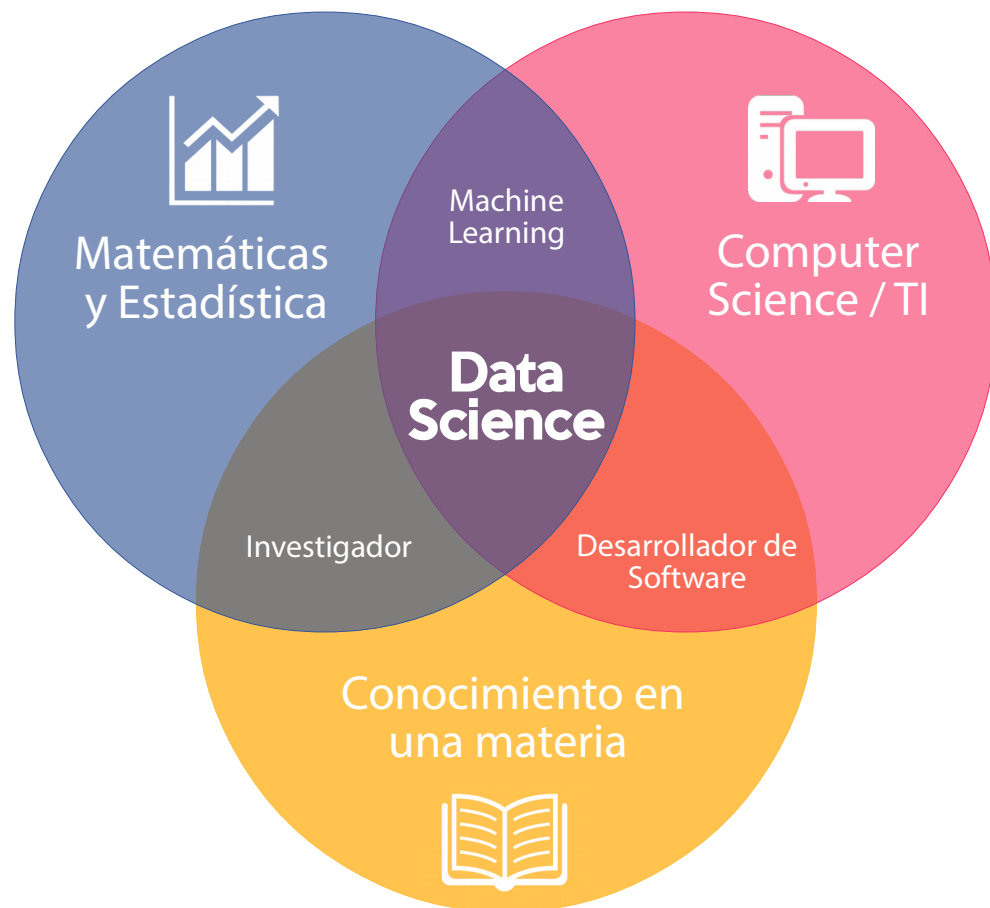
Python for Data Science



EXPOSITOR
WALTER CASAS

Data Science

¿Qué es Data Science?



Extraer conocimiento de los datos.

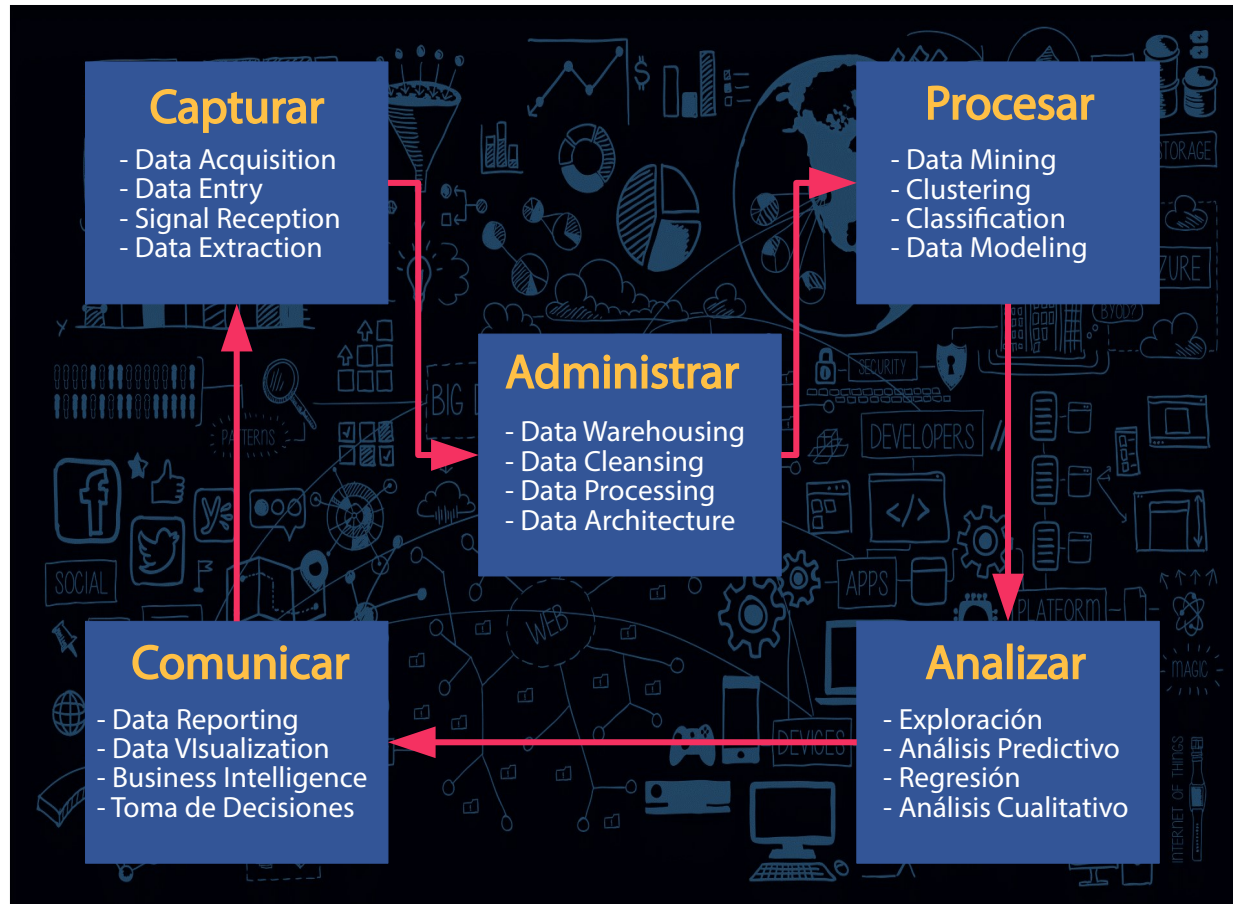
“Extraer datos, entenderlos, procesarlos,
obtener información valiosa, visualizarla y
comunicarla...”

- Hal Varian, jefe economista en Google y Profesor de Information Sciences,
Business and Economics en la UC Berkeley

WWW.DATASCIENCE.PE



¿Qué hace un Data Scientist?

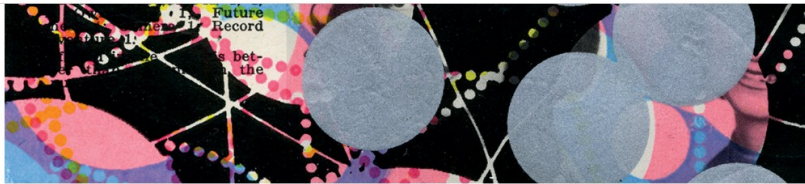


Crea algoritmos para organizar y sintetizar información que será usada para responder preguntas claves y dirigir futuras estrategias en la organización a la que hace parte. Además debe poseer una excelente comunicación para proporcionar resultados altamente técnicos, a diversos interesados en su organización o empresa, de una manera sencilla, que sea entendible por sus homólogos no técnicos.

¿Por qué ser un Data Scientist?

Según Harvard Business Review, la carrera de Data Scientist sería el trabajo más *Sexy* del siglo XXI

Harvard
Business
Review



DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE COMMENT 10 HH TEXT SIZE PRINT \$8.95 BUY COPIES

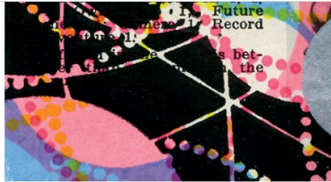
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early." Goldman, a PhD in physics from Stanford, was intrigued by the linking he did see going on and by the richness of the user profiles. It all made for messy data and unwieldy analysis, but as he began exploring people's connections, he started to see

WWW.DATASCIENCE.PE



¿Por qué ser un Data Scientist?

Harvard
Business
Review



DATA

Data Scientist: 1 the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE COMMENT HH TEXT SIZE

When Jonathan Goldman arrived for work, he still felt like a start-up. The company had quickly as existing members invited their connections with the people who were already on the missing in the social experience. As one LinkedIn member realizing you don't know anyone. So you just stand in Goldman, a PhD in physics from Stanford, was intriguing profile. It all made for messy data and unusual data analysis.

The most promising jobs of 2019

LinkedIn

Every January, millions of people start the new year searching for the next chapter in their professional journey, whether that means learning a new skill or finding a new job.

To help, we unveiled the [most in-demand skills of the year](#) last week, and this week we're back to share the year's Most Promising Jobs. Based on LinkedIn data, these positions come with high salaries, a significant number of job openings and year-over-year growth, and are more likely to lead to a promotion.

Whether you're interested in sales, design, or marketing, the perfect job for you might just be one of this year's Most Promising Jobs. Keep reading to learn more about the roles you'll want to keep an eye on, the most in-demand skills and the courses you need to learn them.

2019's Most Promising Jobs in the U.S.

1. Data Scientist

Median Base Salary: \$130,000

Job Openings (YoY Growth): 4,000+ (56%)

Career Advancement Score (out of 10): 9

Top Skills: Data Science, Data Mining, Data Analysis, Python, Machine Learning

Según Harvard Business Review, la carrera de Data Scientist sería el trabajo más *Sexy* del siglo XXI.

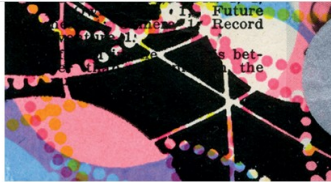
LinkedIn en 2018 y 2019 lo ha nombrado como uno de los trabajos más prometedores del año.

WWW.DATASCIENCE.PE



¿Por qué ser un Data Scientist?

Harvard
Business
Review



DATA

Data Scientist: 1 the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE COMMENT HH TEXT SIZE

When Jonathan Goldman arrived for work, he still felt like a start-up. The company had quickly as existing members invited their connections with the people who were already on the missing in the social experience. As one LinkedIn member realizing you don't know anyone. So you just stand in Goldman, a PhD in physics from Stanford, was intriguing. It all made for messy data and unusual data analysis.

The most promising jobs of 2019

LinkedIn

Every January, millions of people start the new year searching for the next chapter in their professional journey, whether that means learning a new skill or finding a new job.

To help, we unveiled the [most in-demand skills of the year](#) last week, and this week we're back to share the year's Most Promising Jobs. Based on LinkedIn data, these positions come with high salaries, a significant number of job openings and year-over-year growth, and are more likely to lead to a promotion.

Whether you're interested in sales, design, or marketing, the perfect job for you might just be one of this year's Most Promising Jobs. Keep reading to learn more about the roles you'll want to keep an eye on, the most in-demand skills and the courses you need to learn them.

2019's Most Promising Jobs in the U.S.

1. Data Scientist

Median Base Salary: \$130,000

Job Openings (YoY Growth): 4,000+ (56%)

Career Advancement Score (out of 10): 9

Top Skills: Data Science, Data Mining, Data Analysis, Python, Machine Learning

Según Harvard Business Review, la carrera de Data Scientist sería el trabajo más *Sexy* del siglo XXI.

LinkedIn en 2018 y 2019 lo ha nombrado como uno de los trabajos más prometedores del año.



WWW.DATASCIENCE.PE



¿Por qué ser un Data Scientist?

La demanda
aumentará en
28%
para el 2020

WWW.DATASCIENCE.PE



¿Por qué ser un Data Scientist?

La demanda
aumentará en
28%
para el 2020

4,524
Ofertas de
trabajo

W W W . D A T A S C I E N C E . P E



¿Por qué ser un Data Scientist?

La demanda
aumentará en
28%
para el 2020

4,524
Ofertas de
trabajo

\$ 120,931
Salario base
promedio

WWW.DATASCIENCE.PE



¿Por qué ser un Data Scientist?

La demanda
aumentará en
28%
para el 2020

4,524
Ofertas de
trabajo

\$ 120,931
Salario base
promedio

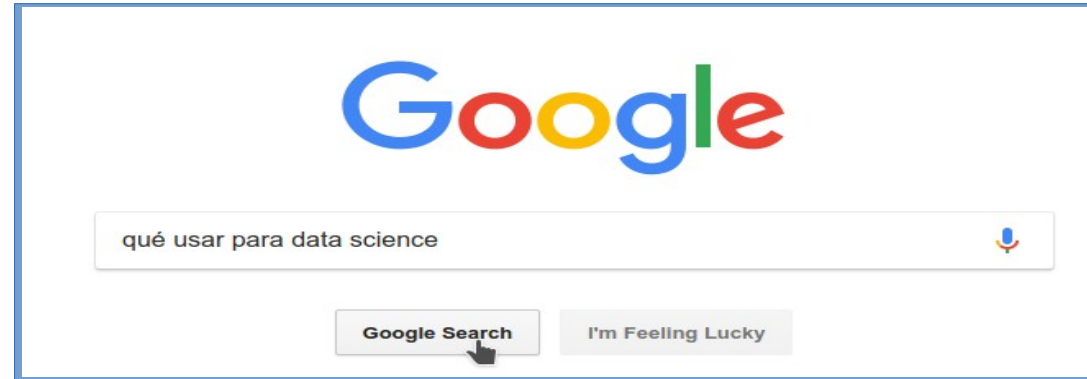
El trabajo
#1
En USA durante
el 2016,2017 y
2018

Fuente: Glassdor y Forbes

W W W . D A T A S C I E N C E . P E



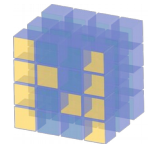
¿Por dónde empiezo?



W W W . D A T A S C I E N C E . P E



¿Por dónde empiezo?



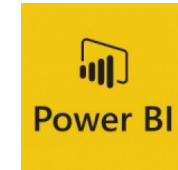
mongoDB

Google

qué usar para data science

Google Search

I'm Feeling Lucky



ORACLE®



+tableau

matplotlib

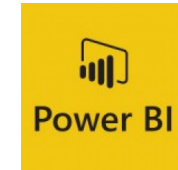
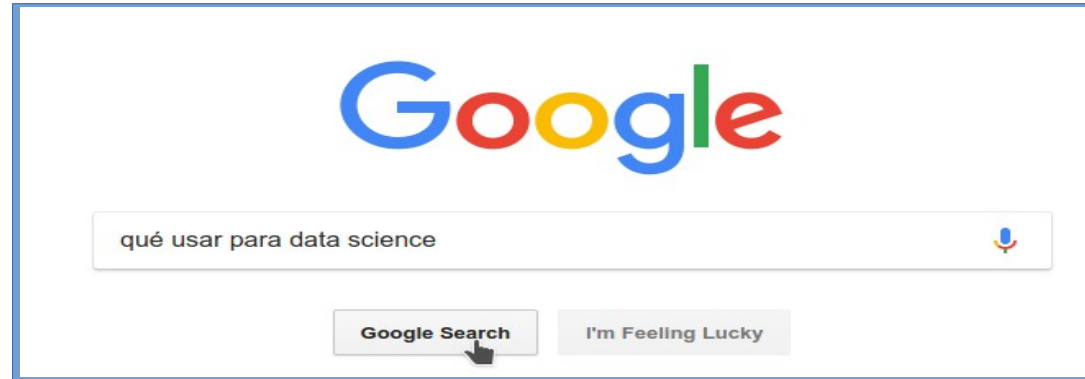
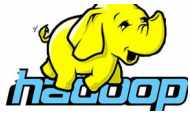


NumPy



WWW.DATASCIENCE.PE

¿Por dónde empiezo?



WWW.DATASCIENCE.PE

¿Por dónde empiezo?



WWW.DATASCIENCE.PE



¿Por dónde empiezo?



WWW.DATASCIENCE.PE



¿Por dónde empiezo?

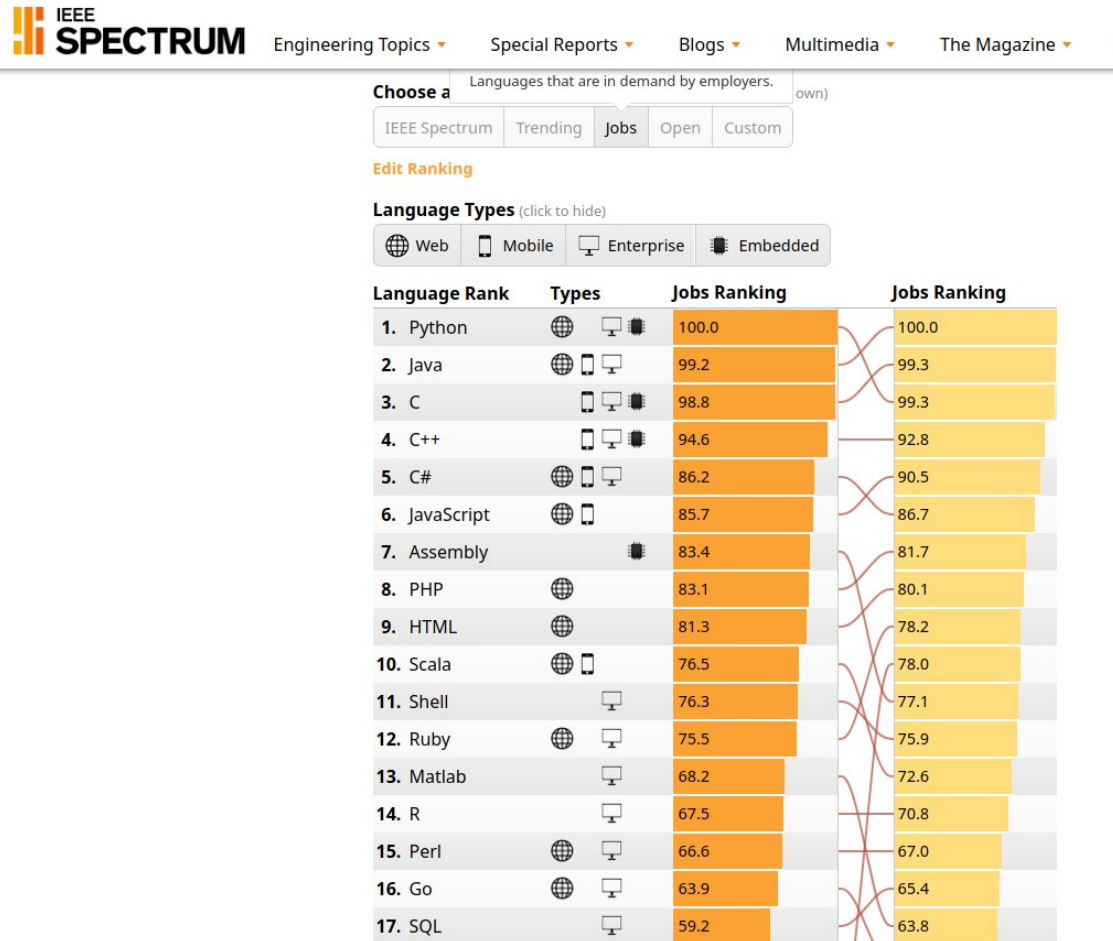


WWW.DATASCIENCE.PE



¿Por qué Python?

The Top Programming Languages 2018 – IEEE Spectrum



- Es de propósito general
- Curva de aprendizaje suave
- Gran comunidad de usuarios
- iPython
- Crecimiento en el ecosistema científico
- Lenguaje de alto nivel
- Muchas librerías

SCIENCE . PE



Librerías de Python para Data Science



Site oficial: www.numpy.org

- Permite crear objetos de arrays N-dimensionales y matrices.
- Mejorea las funciones, permitiendo el uso de matemática avanzada y estadística sobre esos objetos.
- Provee vectorización de operaciones en arrays y matrices.
- Otras librerías en Python están construidas sobre Numpy

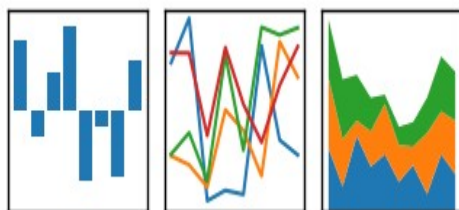
W W W . D A T A S C I E N C E . P E



Librerías de Python para Data Science

pandas

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



Site oficial: www.pandas.pydata.org

- Añade DataFrames y DataSeries como estructura de datos.
- Provee herramientas para manipulación de data: merging, sorting, slicing, etc.
- Permite el manejo de missing data.
- Permite hacer consultas como si fuese SQL.

W W W . D A T A S C I E N C E . P E



Librerías de Python para Data Science



Site oficial: www.scipy.org

- Colección de algoritmos para álgebra lineal, ecuaciones diferenciales, integración numérica, optimización, estadística, etc.
- Parte del Scipy stack junto Numpy, matplotlib, ipython, pandas.
- Construido sobre Numpy.

W W W . D A T A S C I E N C E . P E



Librerías de Python para Data Science



Site oficial: www.matplotlib.org

- Librería de ploteo en 2D que profucen figurars exportables en diversos formatos.
- Funcionalidades similares a la generación de gráficos en MATLAB
- Gráficos de barras, de dispersion, histogramas, gráficos circulares.
- Es muy fácil crear una visualización avanzada.

W W W . D A T A S C I E N C E . P E



Librerías de Python para Data Science

seaborn 0.9.0

Site oficial: www.seaborn.pydata.org

- Basado en matplotlib.
- Provee una interfaz de alto nivel para dibujar gráficos estadísticos.
- Un estilo similar a la librería ggplot2 en R.

WWW.DATASCIENCE.PE



Librerías de Python para Data Science



Site oficial: www.sickit-learn.org

- Herramientas simples y eficientes para data mining y data análisis.
- Provee algoritmos de machine learning: clasificación, regresión, clustering, model validation, etc.
- Construido en base a NumPy, SciPy y matplotlib

W W W . D A T A S C I E N C E . P E



Herramientas para Data Science

IP[y]:
IPython

Site oficial: www.ipython.org

- Un poderoso shell interactivo.
- Un kernel para Jupyter.
- Soporte para visualización interactiva de los datos y uso de herramientas de GUI.
- Fácil de usar, herramientas de alto rendimiento para computación paralela.

W W W . D A T A S C I E N C E . P E



Herramientas para Data Science

IP[y]:
IPython

Site oficial: www.ipython.org

- Un poderoso shell interactivo.
- Un kernel para Jupyter.
- Soporte para visualización interactiva de los datos y uso de herramientas de GUI.
- Fácil de usar, herramientas de alto rendimiento para computación paralela.

W W W . D A T A S C I E N C E . P E



Herramientas para Data Science



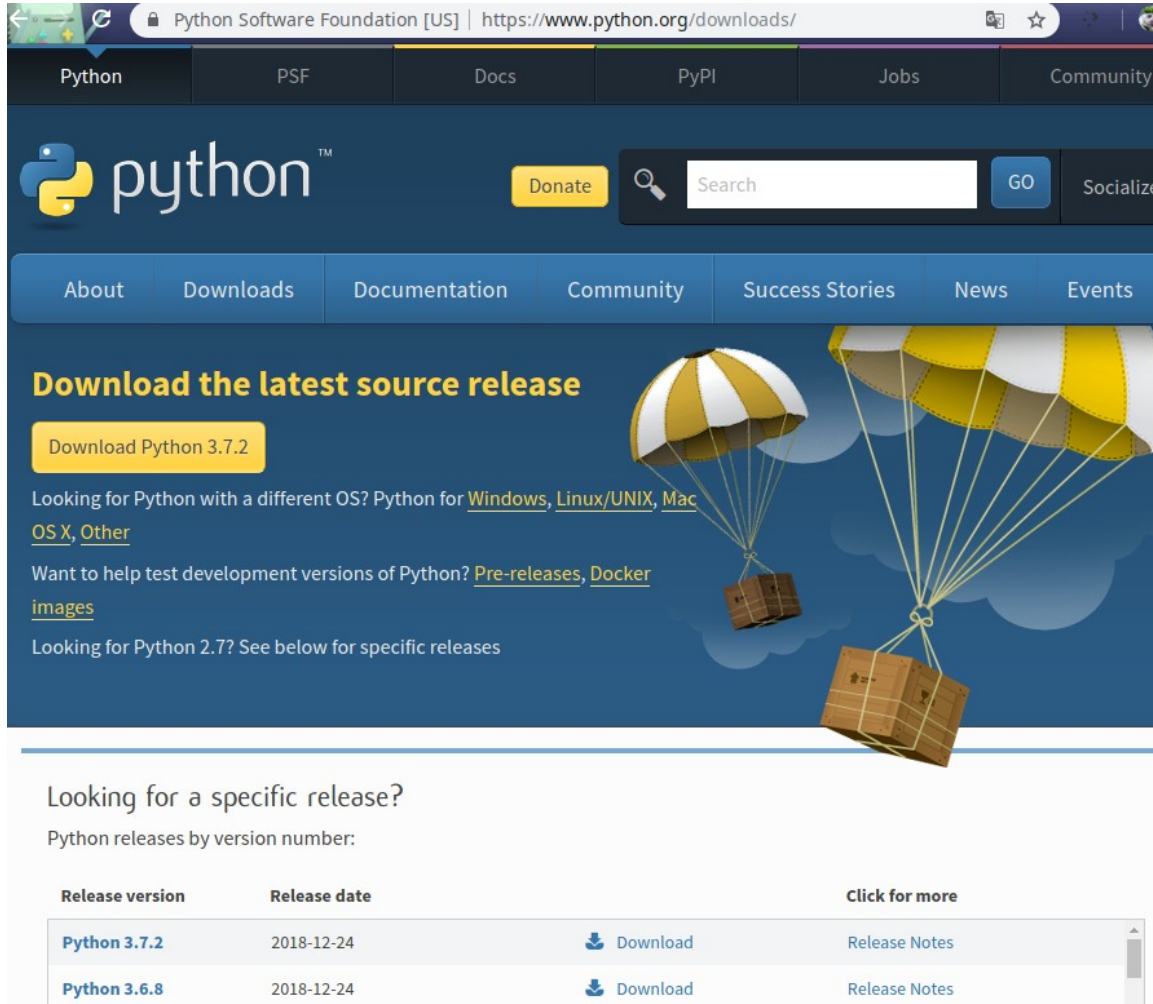
Site oficial: www.jupyter.org

- Poderosa shell interactiva vía web.
- La shell está organizada en pequeños bloques y cada bloque puede contener texto formateado en Markdown.
- Jupyter soporta integración con más de 40 lenguajes de programación.

W W W . D A T A S C I E N C E . P E



Preparando el entorno



The screenshot shows the Python Software Foundation website. The main navigation bar includes links for Python, PSF, Docs, PyPI, Jobs, and Community. Below this is a search bar and a 'Donate' button. The 'Downloads' section is highlighted, and the 'Download the latest source release' button is visible. The page features an illustration of two parachutes carrying boxes. Below the illustration, there are links for different operating systems and pre-releases. At the bottom, there is a table of Python releases by version number.

Download the latest source release

Download Python 3.7.2

Looking for Python with a different OS? Python for [Windows](#), [Linux/UNIX](#), [Mac OS X](#), [Other](#)

Want to help test development versions of Python? [Pre-releases](#), [Docker images](#)

Looking for Python 2.7? See below for specific releases

Looking for a specific release?

Python releases by version number:

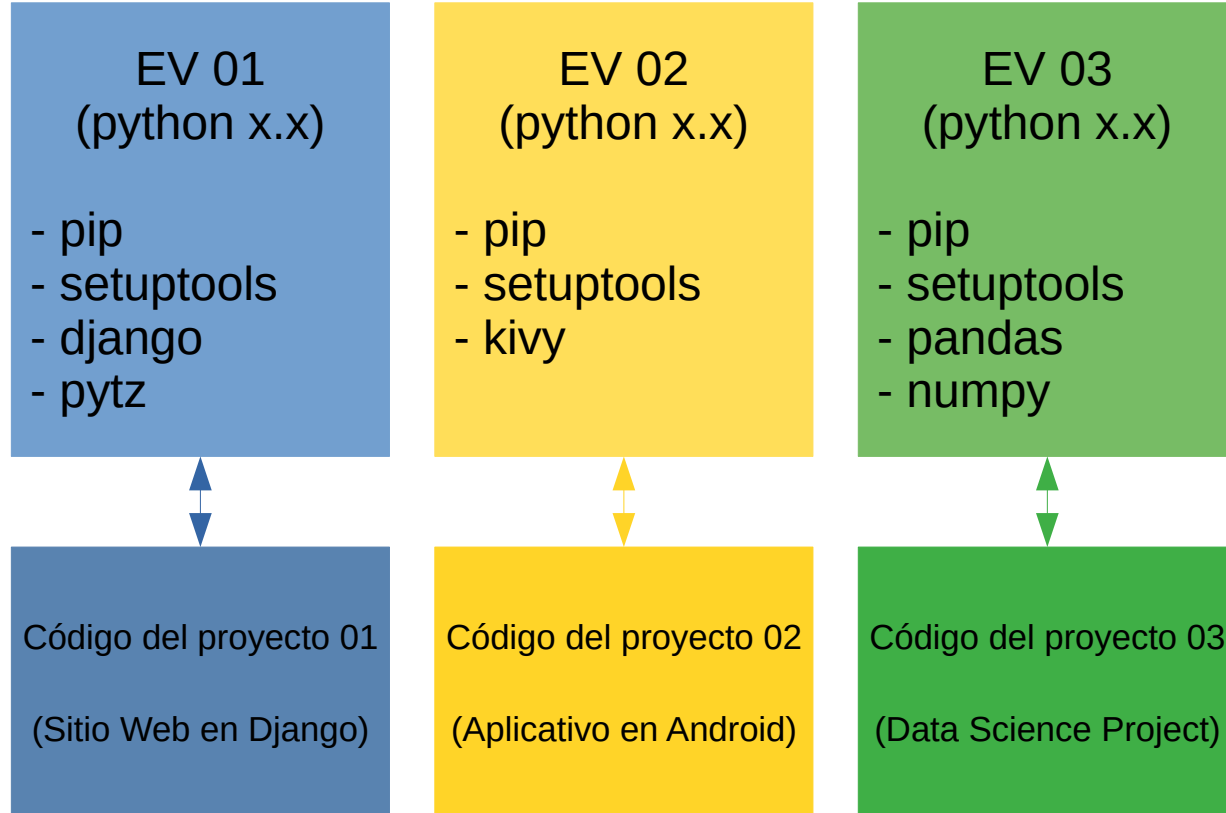
Release version	Release date	Click for more	
Python 3.7.2	2018-12-24	Download	Release Notes
Python 3.6.8	2018-12-24	Download	Release Notes

- Instalamos Python 3
- Instalamos pip
- Instalamos ipython
- Instalamos Jupyter

WWW.DATASCIENCE.PE



Preparando el entorno



Ejecución de proyectos en entornos virtuales:

- Instalamos virtualenv con pip

Luego creamos nuestro Kernel:

```
$ python -m venv proyecto
```

```
$ source proyecto/bin/activate
```

```
(venv) $ pip install ipykernel
```

```
(venv) $ ipython kernell install --user --name=proyecto
```

```
(venv) $ jupyter notebook
```

Seleccionar el Kernel creado

Dataset Iris



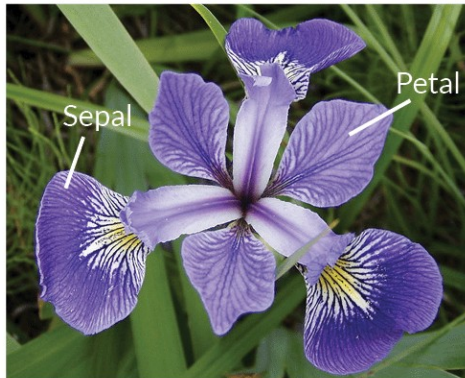
Iris Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Famous database; from Fisher, 1936



Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	2420814



Iris Versicolor



Iris Setosa



Iris Virginica

El dataset Iris es un ejemplo básico para demostrar los conceptos de Data Science.

El dataset contiene cuatro características (largo y ancho de los pétalos y sépalos) de 150 muestras de tres especies de Iris (Iris setosa, Iris, virginica e Iris versicolor).

Estas medidas se utilizaron para crear un modelo discriminante para clasificar la especie.

Go to



WWW.DATASCIENCE.PE



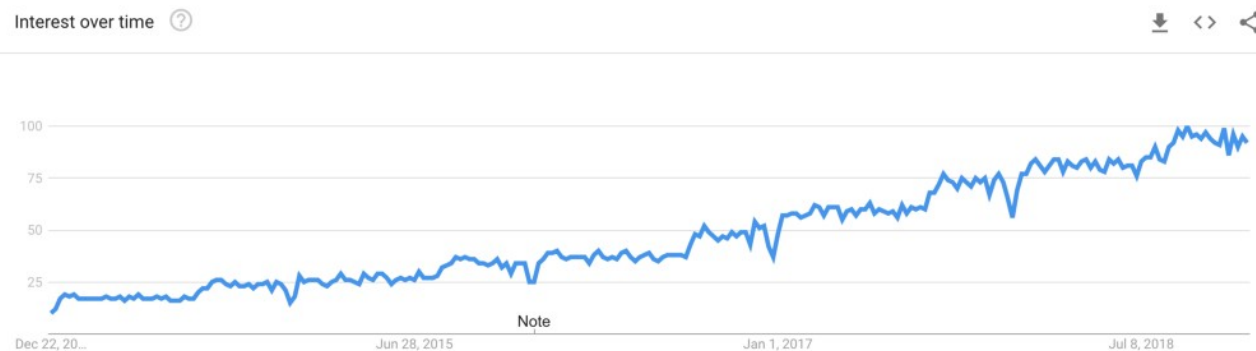
Hacia dónde va el Data Science



Borrador de
lineamiento de
ética para la IA
(Unión Europea)

La maduración de las tecnologías serán importantes en los años venideros, así como temas específicos que aún requieren mucha atención, tales como: Feature Engineering, Aprendizaje reforzado, Privacidad de datos, entre otros...

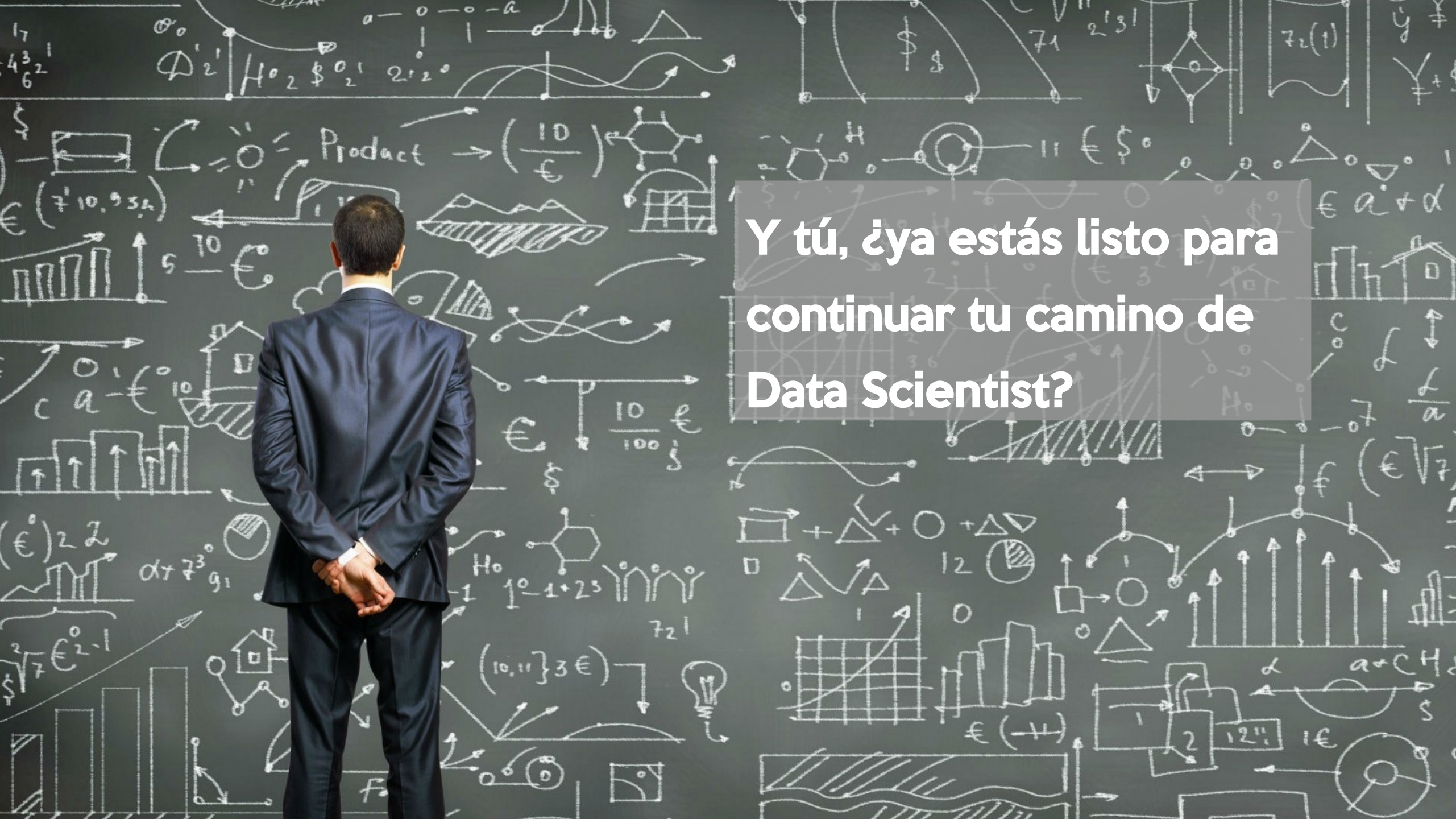
Quedan aún muchas interrogantes sobre lo que vendrá y de lo que aún falta hacer y se necesitan nuevos especialistas que puedan cubrir la demanda de un mundo donde cada día la labor del Data Scientist cobra mayor importancia.



Interés del término "Data Science" desde el 2013(Google Trends)

W W W . D A T A S C I E N C E . P E





**Y tú, ¿ya estás listo para
continuar tu camino de
Data Scientist?**

EVENTO ORGANIZADO POR DATA SCIENCE RESEARCH PERÚ



ENCUÉNTRANOS EN



Data Science Research



Data Science Research Perú



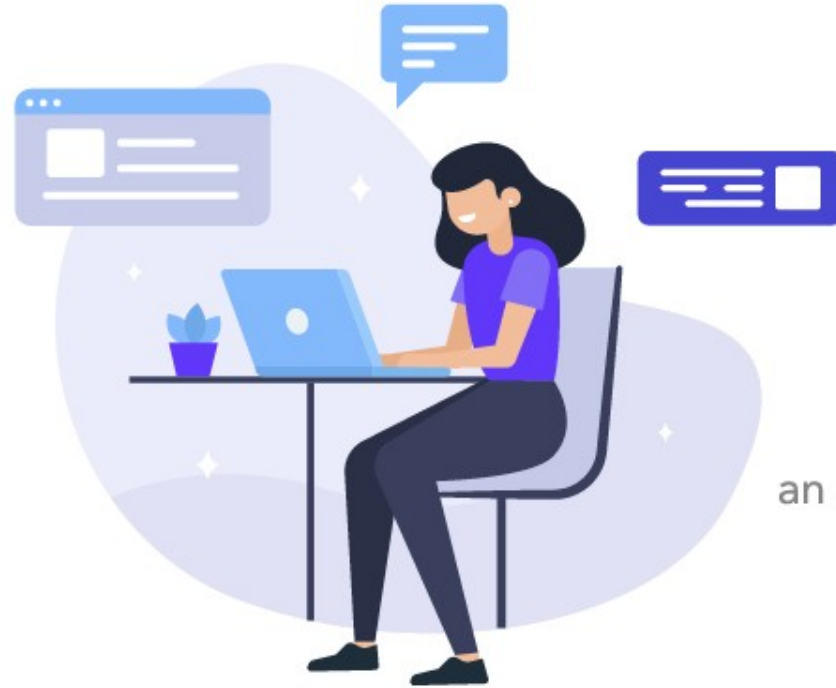
Data Science Research Perú

WWW.DATASCIENCE.PE

CON EL APOYO DE



UNIVERSIDAD
esan



an **NTT DATA** Company



MUCHAS GRACIAS POR VENIR

WWW.DATASCIENCE.PE