Introducción a

Ciencia de Datos

2025

Asignatura Optativa

LICENCIATURA EN INFORMÁTICA

FACET-UNT

CD2025

Docentes

- **Dra. María Graciela Molina**

  gmolina@herrera.unt.edu.ar

- **Lic. Jorge H. Namour**

  jnamour@herrera.unt.edu.ar

Horarios

- Miercoles de 9 a 13 hs (carga horaria 4 hs semanales)
- Lab. Redes - Dpto Cs de la Computación

facet virtual

- Curso: Ciencia de Datos
- Contraseña de matriculación: cd2025

Temas

- Conceptos de Ciencia de Datos
- Lenguajes de programación y Ciencia de Datos.
- Adquisición, exploración y preparación de datos
- Almacenamiento y procesamiento de grandes volúmenes de datos
- Modelado basado en datos: Ingeniería de características. Aprendizaje automático (clasificación/regresión). Redes neuronales artificiales.
- Evaluación del modelo. Interpretación y Explicación.

Herramientas

Teo/TP

- Para aprobar: Aprobar todos los TPs (entregas y rendidos de manera oral en las fechas asignadas por la cátedra)

- Al final (optativa) se pueden recuperar los TPs desaprobados

INTRO

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 13 COMMENT 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

1/3 FREE ARTICLES LEFT > REGISTER FOR MORE | SUBSCRIBE + SAVE!

Search Home My Network Jobs

#1. AI/Machine Learning Engineer

- **Role:** Designs and deploys AI and machine learning models to solve real-world problems.
- **Why It's the Hottest:** AI continues to be the cornerstone of innovation across industries.
- **Skills:** Python, TensorFlow, PyTorch, Deep Learning, NLP.
- **Industries:** Tech, Healthcare, Finance, Automotive.
- **Compensation:** $140,000–$200,000.

#2. Data Scientist

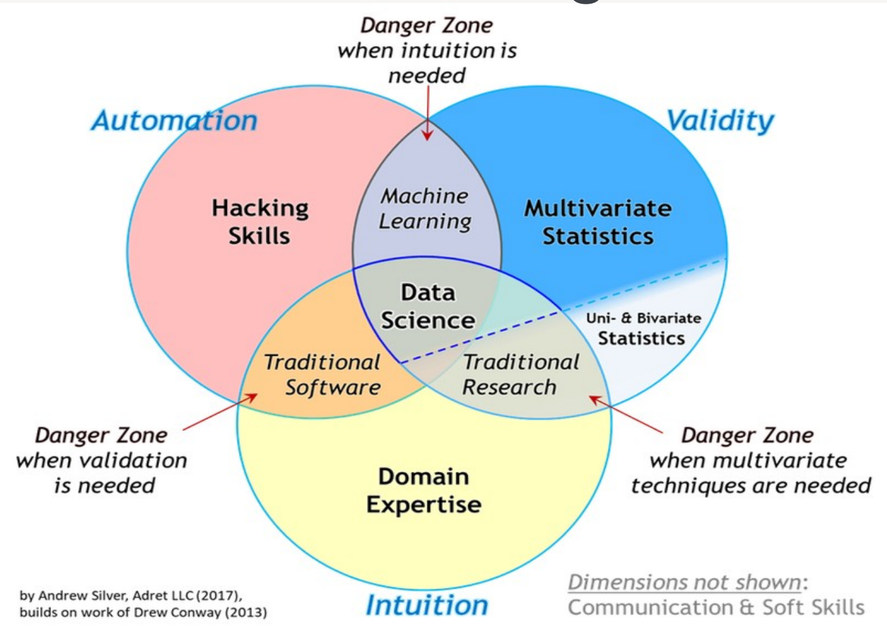
- **Role:** Uses advanced analytics and machine learning to derive actionable insights from data.
- **Why It's Hot:** Data drives decision-making in nearly every industry.
- **Skills:** Python, R, SQL, Data Visualization Tools.
- **Industries:** Finance, Healthcare, Retail, Tech.
- **Compensation:** $120,000–$180,000.

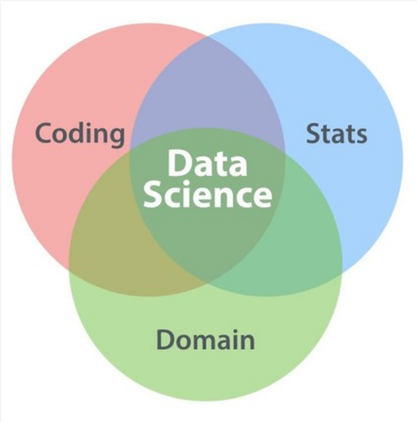
https://www.linkedin.com/pulse/top-25-hot-jobs-technology-2025-ranked-from-hottest-morgan-feldman-rksle/

Ciencia de Datos

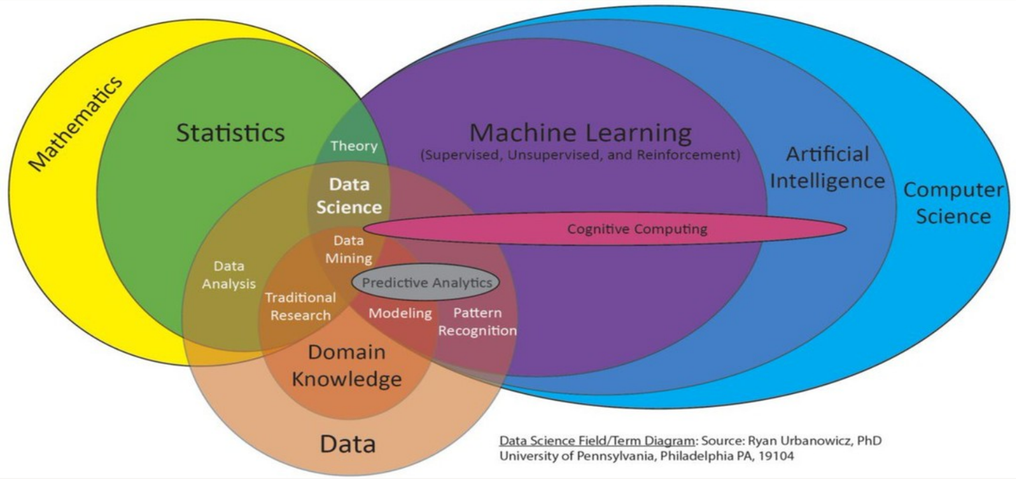
- Campo interdisciplinario
- Extraer conocimiento o mejor entendimiento de los datos
- Transformar los datos en información/conocimiento/ tomar decisiones

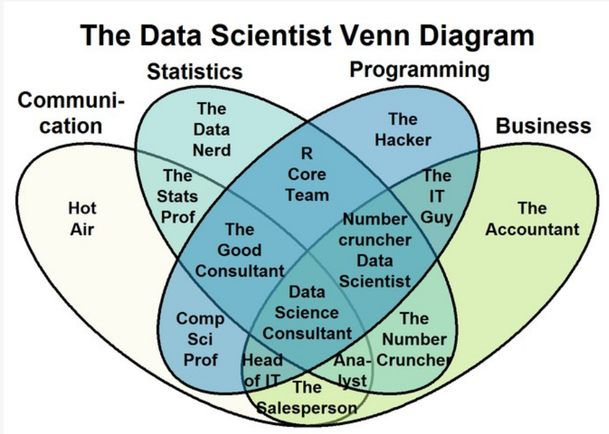
La batalla de los diagramas de Venn





Drew Conway's





Ciencia de datos e investigación científica

Ciencia, se denomina a un “cuerpo de doctrina, de validez universal y certeza objetiva, metódico y sistemático, que versa sobre un sector delimitado de la realidad y constituye un ramo particular del saber humano”.

Diccionario de la lengua española

Del lat. *scientia*.

1. f. Conjunto de conocimientos obtenidos mediante la observación y el razonamiento, sistemáticamente estructurados y de los que se deducen principios y leyes generales con capacidad predictiva y comprobables experimentalmente.

2. f. Saber o erudición. *Tener mucha, o poca, ciencia. Ser un pozo de ciencia. Hombre de ciencia y virtud.*

   SIN.: **conocimiento**, **saber**[2], **sabiduría**, **sapiencia**, **erudición**.

   ANT.: **ignorancia**, **incultura**, **nesciencia**.

3. f. Habilidad, maestría, conjunto de conocimientos en cualquier cosa. *La ciencia del caco, del palaciego, del hombre vividor.*

   SIN.: **habilidad**, **maestría**, **experiencia**.

4. f. pl. Conjunto de conocimientos relativos a las **ciencias** exactas, físicas, químicas y naturales.

Impone un camino o procedimiento de hallar la verdad y enseñarla -> un MÉTODO

↓

asegura los alcances de la ciencia, su proyección, que los conocimientos no nazcan y mueran con sus descubridores o quienes estuvieron próximos a ellos.

Los logros de La Ciencia son acumulativos

Ciencia de datos e investigación científica

TAREA: Leer el artículo y llenar cuestionario en la plataforma

METODO CIENTIFICO

Consiste en la observación sistemática, la medición, la experimentación, la formulación, el análisis y la modificación de las hipótesis. El método científico está sustentado por dos pilares fundamentales:

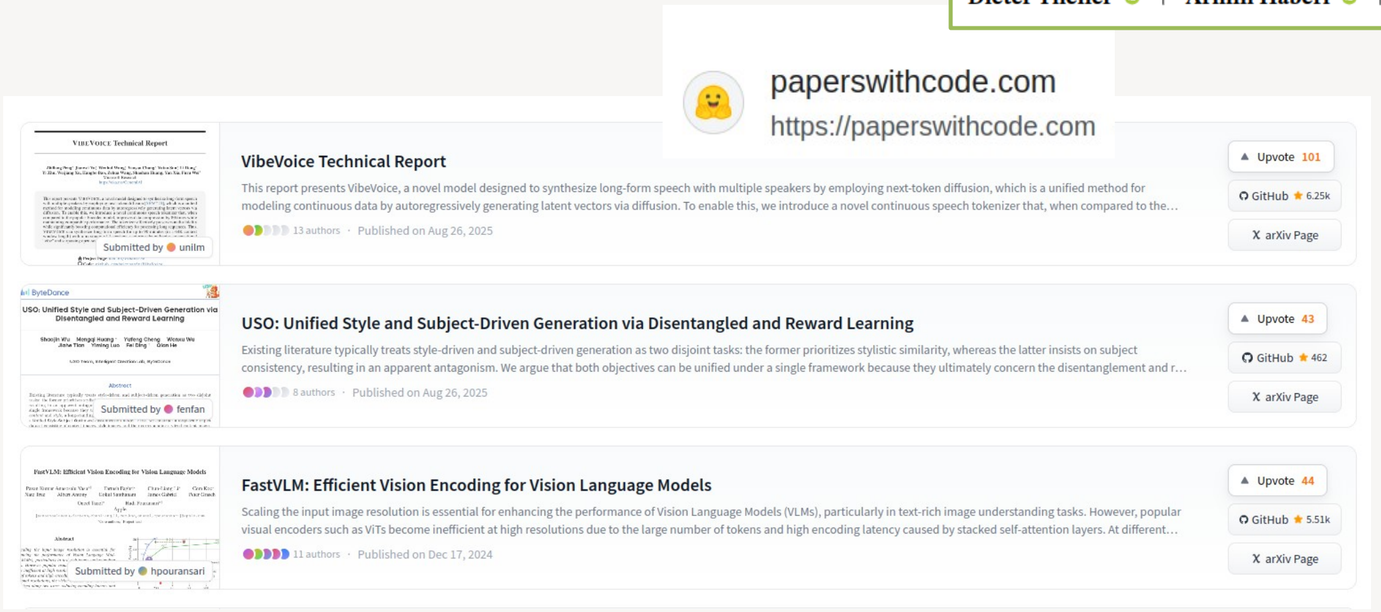
1 **Reproductibilidad**: capacidad de repetir un determinado experimento, en cualquier lugar y por cualquier persona.


Received: 2 July 2024 | Revised: 19 December 2024 | Accepted: 25 February 2025
DOI: 10.1002/aaai.70002

ARTICLE

Reproducibility in machine-learning-based research: Overview, barriers, and drivers

Harald Semmelrock[1] | Tony Ross-Hellauer[2] | Simone Kopeinik[2] | Dieter Theiler[2] | Armin Haberl[2] | Stefan Thalmann[3] | Dominik Kowald[1,2]



Ciencia de datos e investigación científica

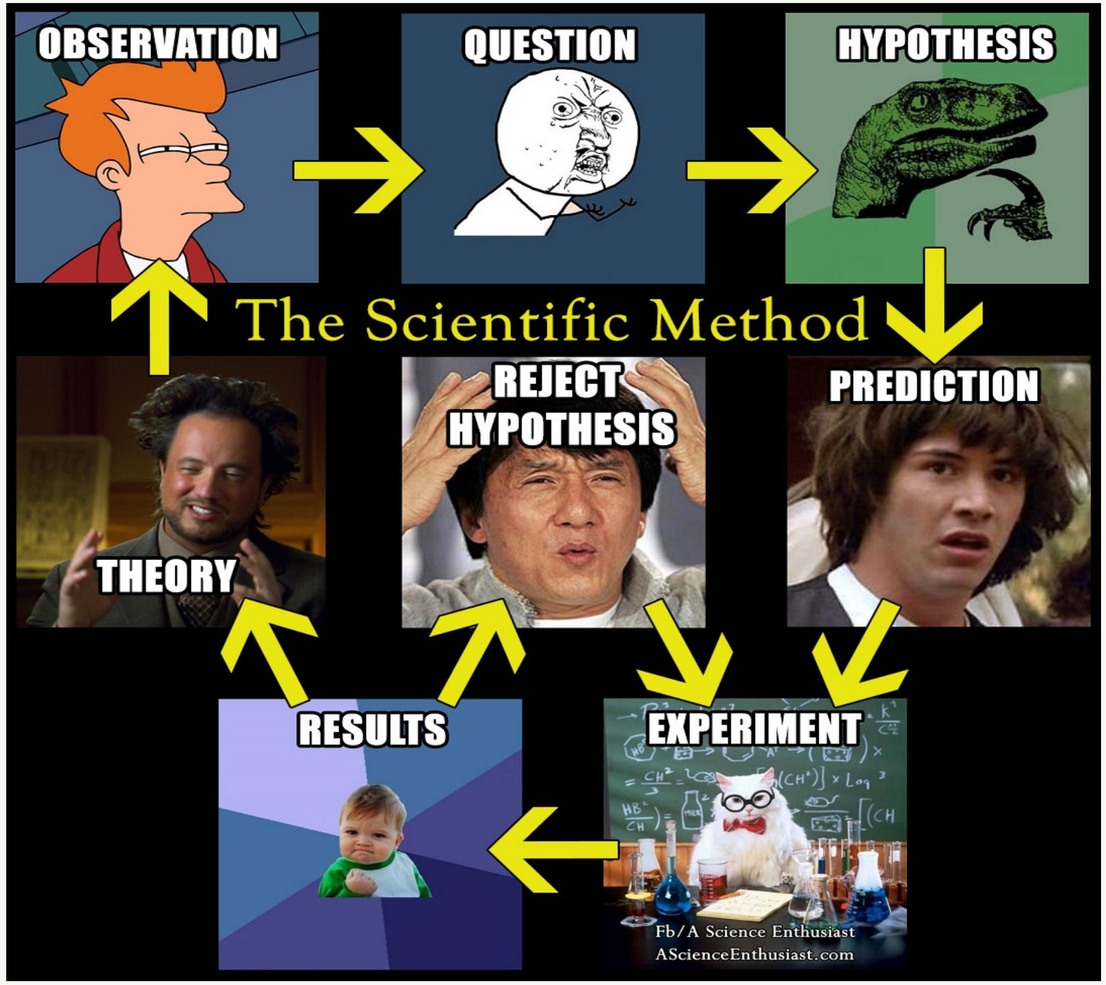
METODO CIENTIFICO

Consiste en la observación sistemática, la medición, la experimentación, la formulación, el análisis y la modificación de las hipótesis. El método científico está sustentado por dos pilares fundamentales:

2

Refutabilidad: toda proposición científica tiene que ser susceptible de ser falseada o refutada. Esto implica que se podrían diseñar experimentos, que en el caso de dar resultados distintos a los predichos, negarían la hipótesis puesta a prueba.

https://science.nasa.gov/dwarf-planets/pluto/



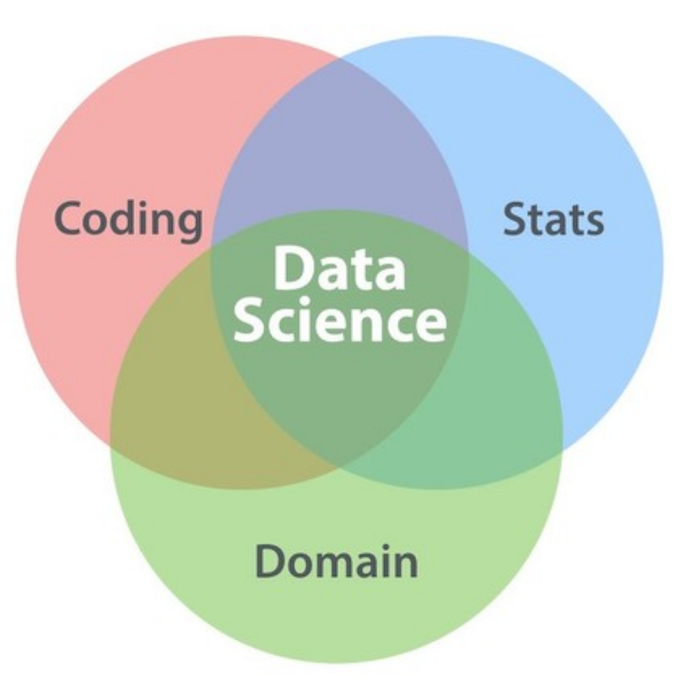
“un procedimiento para tratar un conjunto de problemas […]. Los problemas del conocimiento, a diferencia de los del lenguaje o los de la acción, requieren la invención o la aplicación de procedimientos especiales adecuados para los varios estadios del tratamiento de los problemas…”.

Bunge, M. (2018). *La ciencia: Su método y su filosofía* (Vol. 1). Laetoli.

El método científico

i. Enunciar preguntas bien formuladas y verosímilmente fecundas.
ii. Arbitrar conjeturas, fundadas y contrastables con la experiencia, para contestar las preguntas.
iii. Derivar consecuencias lógicas de las conjeturas.
iv. Arbitrar técnicas para someter las conjeturas a contraste.
v. Someter a contraste esas técnicas para comprobar su relevancia y la validez que merecen.
vi. Llevar a cabo la contrastación e interpretar sus resultados.
vii. Estimar la pretensión de verdad de las conjeturas y la fidelidad de las técnicas.
viii. Determinar los dominios en los cuales valen las conjeturas y las técnicas, y formular los nuevos problemas originados por la investigación.

Qué herramientas tenemos en Ciencia de Datos



- Programación ✓
- Matemática/ Estadística/ Métodos Numéricos ✓
- Conocimiento del problema/dominio ✓

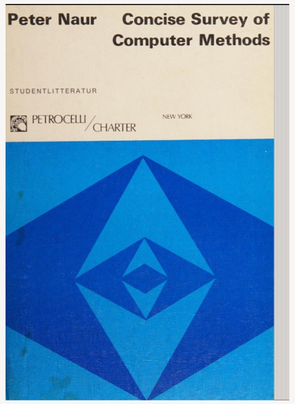
CD como metodología!

Cuando comenzó todo?

www.historyofdatascience.com/dartmouth-summer-research-project-the-birth-of-artificial-intelligence/

Nace el termino ‘Machine Learning’ (Arthur Samuel). h

1957



1974

Peter Naur usa el termino ‘Data Science’ en su paper ‘ The Concise Survey of Computer Methods’

https://archive.org/details/concisesurveyofc0000naur

Deep Blue (IBM) vence al campeón mundial de ajedrez Gary Kasparov

1997

https://www.chess.com/article/view/deep-blue-kasparov-chess

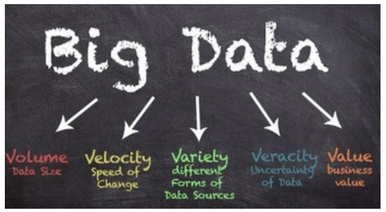
2012

Harvard declara el rol de científico de datos como ‘ the sexiest job of 21st century’

hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

https://onlinestemprograms.wpi.edu/blog/history-data-science-and-pioneers-you-should-know

Por Qué el boom en los últimos años?



- **Grandes datasets**
- **+ fácil acceder a data collection**
- **+ almacenamiento**

https://archive.ics.uci.edu/ml/index.php

https://www.kaggle.com/datasets

https://github.com/awesomedata/awesome-public-datasets

IMAGENET

Hardware:

- **Graphics Processing Units (GPUs)**
- **Paralelismo masivo**

HPC Cloud

TensorFlow

Software:

- **Técnicas mejoradas**
- **Nuevos modelos**
- **Toolboxes**
- **Algoritmos cada vez más maduros y cada vez más usado en diferentes campos de aplicación**

ChatGPT

Todos generan data

Medida	Simbología	Equivalencia
byte	b	8 bits
kilobyte	Kb	1024 bytes
megabyte	MB	1024 KB
gigabyte	GB	1024 MB
terabyte	TB	1024 GB
petabyte	PB	1024 TB
exabyte	EB	1024 PB
zetabyte	ZB	1024 EB
yottabyte	YB	1024 ZB
brontobyte	BB	1024 YB
geopbyte	GB	1024 BB

- Que conocimientos necesitamos?

Stack de conocimiento

- Experiencia
- Métodos
- Lenguages de programación
- Herramientas y librerías
- Acceso a adatos y transformación
- DBs

Stack de conocimiento

- **Experiencia**
- Métodos
- Lenguajes de programación
- Herramientas y librerías
- Acceso a datos y transformación
- DBs

✔

Conocer el problema:

- **Describir sus condiciones, alcances, parámetros, validar resultados, entender resultados, etc**

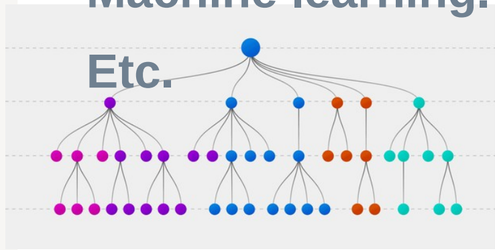
Ejemplos

- **Negocios: Finanzas, cadena de suministros, clientes, pronóstico de mercado,etc**
- **Ingeniería: Mantenimiento, producción, simulacionese, etc**
- **Ciencias Naturales: física experimental, biología, geología**
- **Ciencias Médicas, etc.**

Stack de conocimiento

- Experiencia
- **Métodos**
- Lenguajes de programación
- Herramientas y librerías
- Acceso a datos y transformación
- DBs

Visualización. Métodos estadísticos. Optimización. Machine learning. Deep learning. Etc.

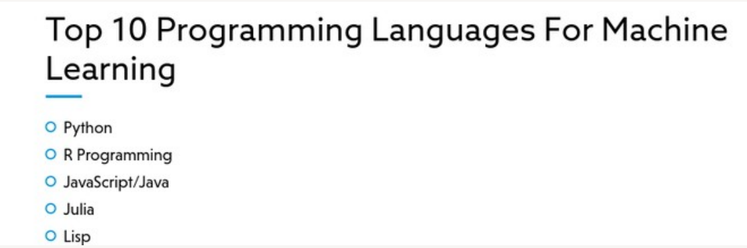


https://arxiv.org/pdf/1812.04948.pdf

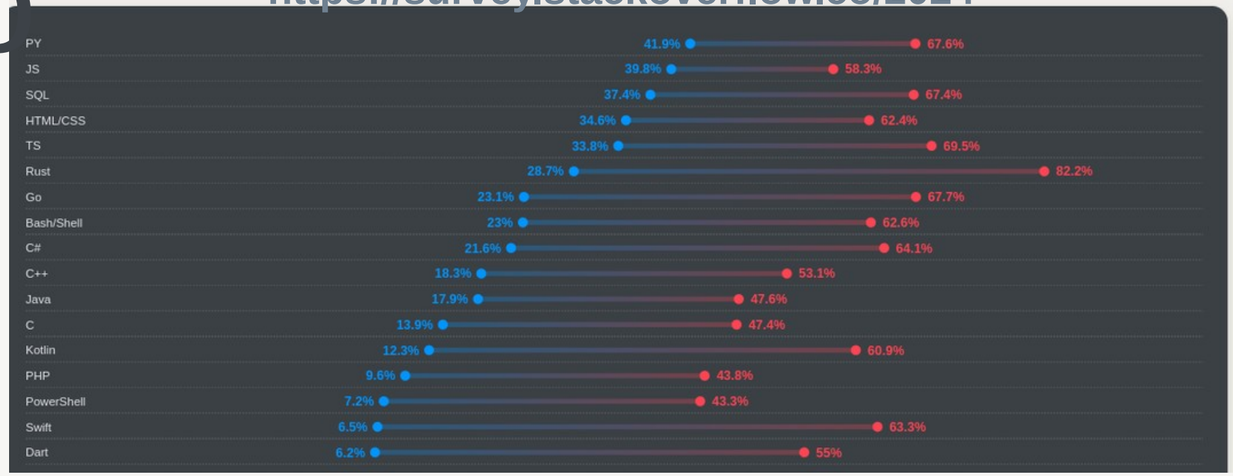
Stack de conocimiento

- Experiencia
- Métodos
- **Lenguajes de programación**
- Herramientas y librerías
- Acceso a datos y transformación
- DBs

https://www.spec-india.com/blog/programming-languages-for-machine-learning.



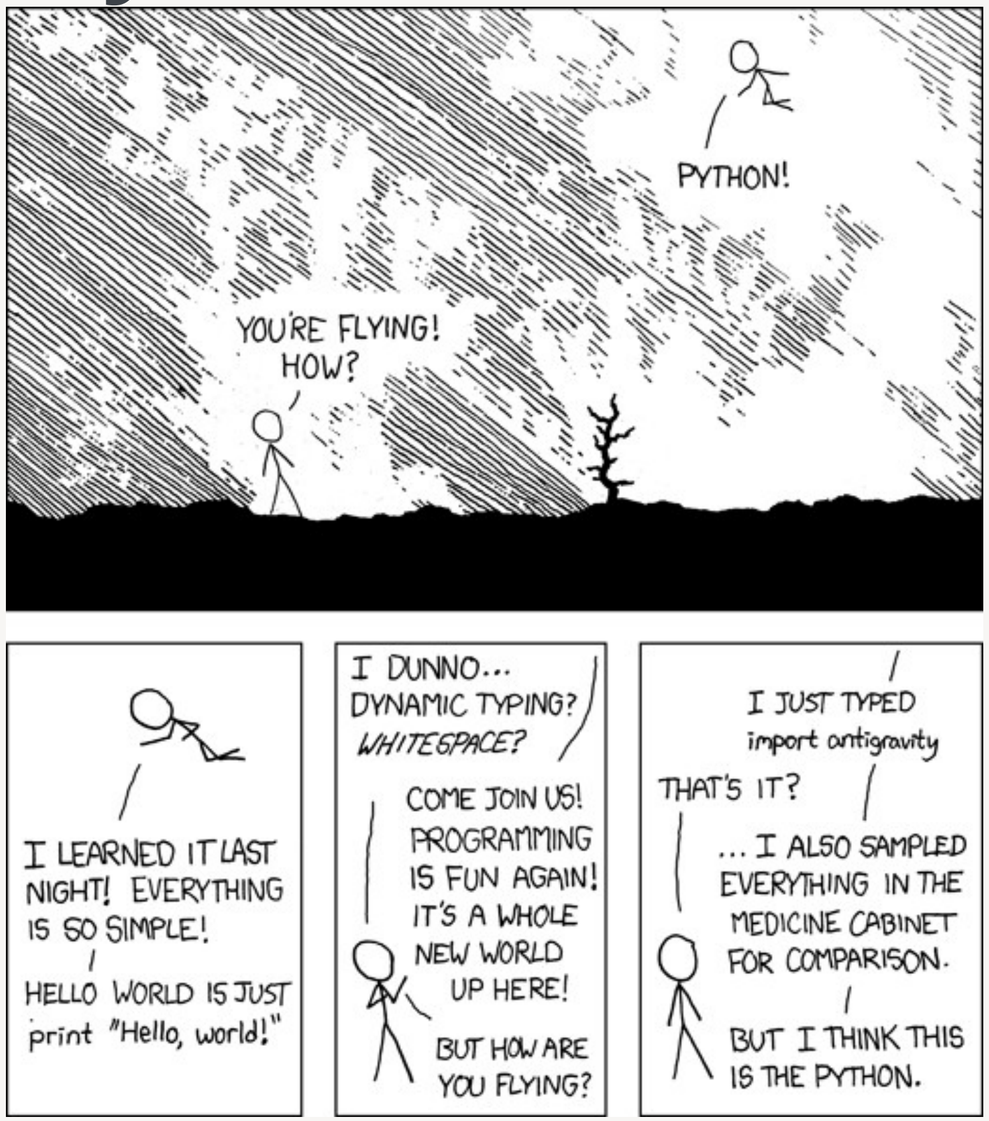
https://survey.stackoverflow.co/2024



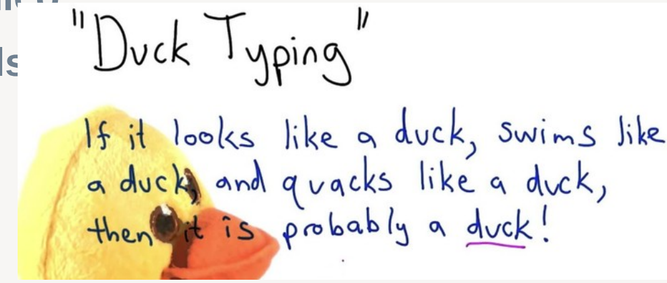
www.tiobe.com/tiobe-index

Aug 2024	Aug 2023	Change	Programming Language	Ratings	Change
1	1		Python	18.04%	+4.71%
2	3	▲	C++	10.04%	-0.59%
3	2	▼	C	9.17%	-2.24%

Python



- **Librerías ++**
- **Comunidad ++**
- **Curva de aprendizaje ++**
- **Free + Open Source**
- **Interpretado, multiparadigma,**
- **fuertemente tipado, tipado dinámico**
- **Frameworks, environments, shells**



- **Zen de Python**

Beautiful is better than ugly.

Explicit is better than implicit.

Simple is better than complex.

Complex is better than complicated.

...

Readability counts.

Special cases aren't special enough to break the rules.

...

- **Style Guide**

https://www.python.org/

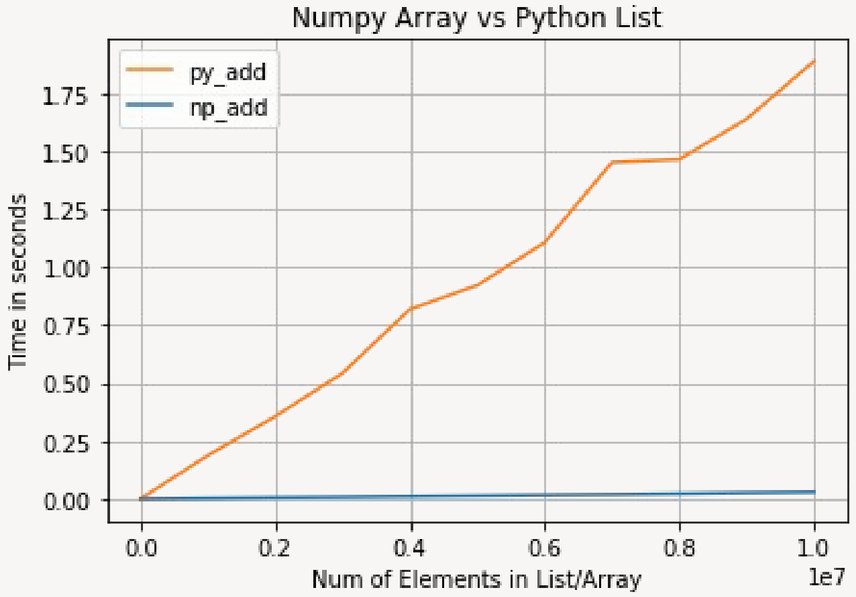
https://www.python.org/dev/peps/pep-0008/

TP0 - repaso Python

Stack de conocimiento

- Experiencia
- Métodos
- Lenguajes de programación
- **Herramientas y librerías**
- Acceso a datos y transformación
- DBs

Arreglos



Se compara la suma de 2 listas de hasta 10.000.000 elementos, con la suma de 2 arreglos con la misma cantidad de elementos

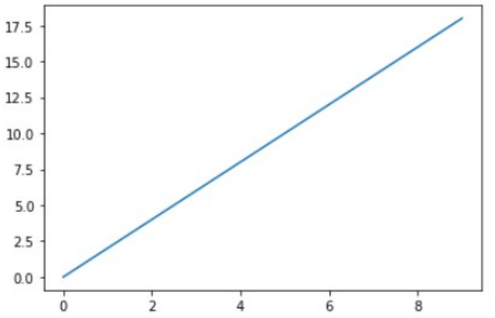
https://medium.com/@gough.cory/performance-

Listas	Array (Numpy)
• Pueden tener elementos de diferente tipo	• Solo pueden tener elementos de igual tipo
• No nec importar un modulo	• Nec importar un modulo
• No hace operaciones aritméticas directamente	• Hace operaciones aritméticas directamente
• Pensadas para poco elementos	• Pensadas para muchos elementos
• + flexibilidad	• < flexibilidad (op por elementos)
• Se puede mostrar sin nec de un loop	• Nec de un loop para el print
• Consume > memoria	• Consume < memoria que las listas

```
import numpy as np
import matplotlib.pyplot as plt

x = np.arange(10)
y = 2*x
print(y)
plt.plot(x,y)
```

```
[ 0  2  4  6  8 10 12 14 16 18]
[<matplotlib.lines.Line2D at 0x7fba635333d0>]
```



Stack de conocimiento

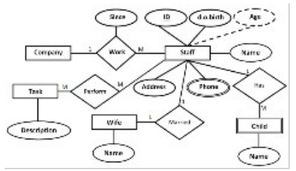
- Experiencia
- Métodos
- Lenguajes de programación
- Herramientas y librerías
- **Acceso a datos y transformación**
- DBs

- **Extrar/Transformar/Cargar/Data Streaming/Data Flow/Arquitectura de red/ Conectividad/Seguridad de datos/Criptografía**

Stack de conocimiento

- Experiencia
- Métodos
- Lenguajes de programación
- Herramientas y librerías
- Acceso a datos y transformación
- **DBs**

- **DB SQL/ERM/Normalización. DB NoSQL/InMemory File Formats (XML,JASON, etc)**



SQL

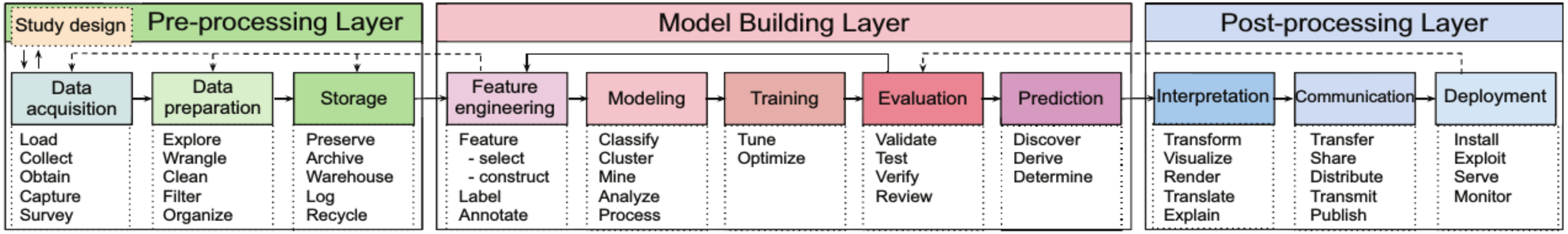
mongoDB.

cassandra

Data Science Pipeline

Ciclo de vida de un proyecto de CD

- Colección de etapas de tratamiento/procesamiento de los datos desde la adquisición, curado/limpieza, ingeniería de características, etc, hasta el modelado, evaluación y su puesta en operación
- Permite seguir el flujo de los datos y las tareas que se deben realizar sobre estos
- Mejora el diseño y ayuda a un desarrollo de software eficiente y de fácil mantenimiento
- No hay un único camino, depende en gran medida de la envergadura del proyecto del que se trate



Biswas et al., 2022

A Data Science Pipeline

Pregunta

DATA

wrangling/ munging

Limpieza de Datos

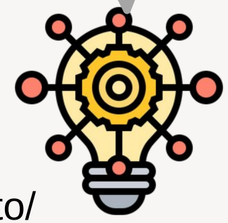
Exploración

Procesamiento/ Análisis/ Modelado

Validación

Pre-Proccesamiento

Monitoreo/ Deployment

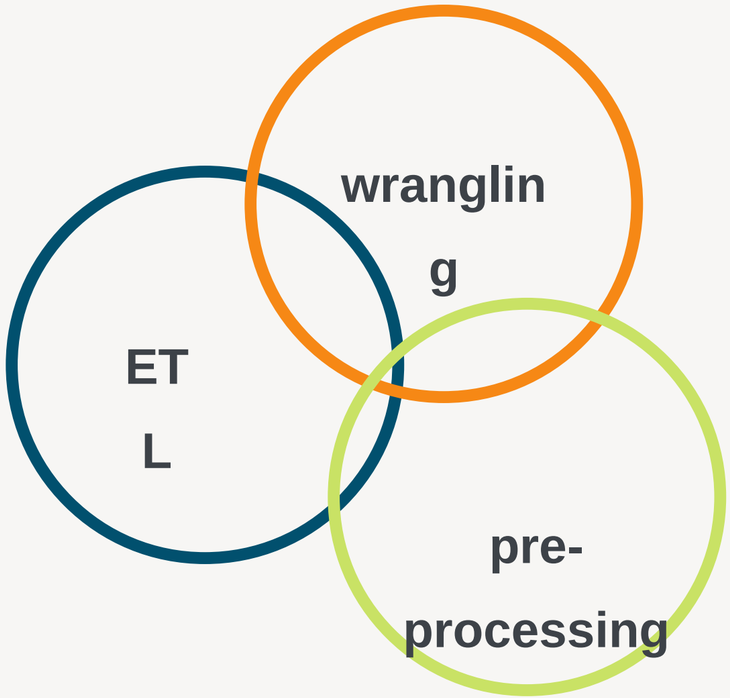


Conocimiento/ Decisión

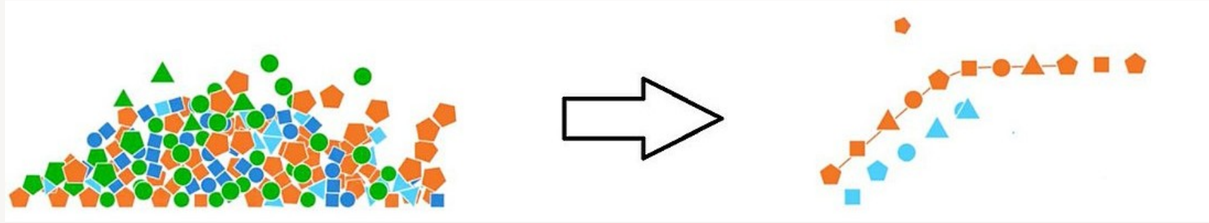
Adquisición de datos

- Fuente: instrumentos, dispositivos, bases de datos, experimentos, simulaciones, distribuidos en la nube, desde un archivo, sintéticos, etc
- De acuerdo a cómo se analizaran los datos: para procesamiento batch/micro-batch/ streaming (Big Data)
- Fuente: instrumentos, dispositivos, bases de datos, experimentos, simulaciones, distribuidos en la nube, desde un archivo, sintéticos, etc
- Modo: "manual", scrapping, via ftp, APIs, etc.
- Multiples fuentes + datos heterogeneos + estructutados/semi estructurados/ no estructurados + muchos datos.
- Depende del problema y de los objetivos finales del proyecto

Data wrangling/ munging



- Proceso de transformar y mapear datos "crudos" o sin procesar en otra forma con la intención de que sea más apropiado para realizar otras tareas
- Involucra a las tareas de: descubrimiento (entender mejor los datos), limpieza, re-estructuración de los datos, enriquecimiento (si es necesario agregarle información), validación (para asegurar consistencia, calidad y seguridad; por ejemplo haciendo un chequeo cruzado de los datos), y publicación (poner disponibles los datos).
- Existe un overlap con la integración de datos (ETL: Extract-Transform-Load) para combinar datos de diversas fuentes para proveer al usuario de una visualización unificada de los mismos.



Data Science Pipeline

Simplifiquemos

Adquisición

Pre-procesamiento

Almacenamiento

Procesamiento

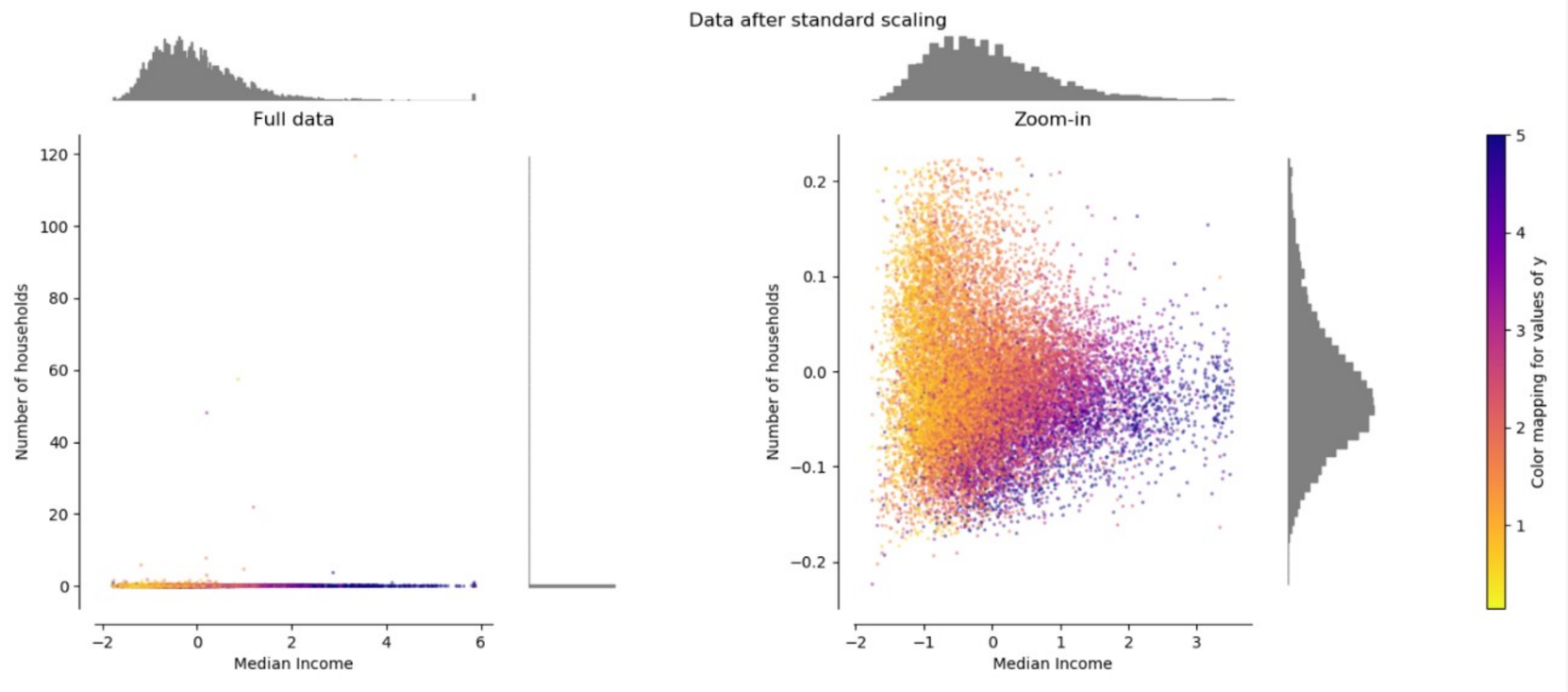
Deployment

Pre-procesamiento

- Normalización/ estandarización (transformación de datos)
- Exploración de los datos: Entender mejor los datos, serie de estadísticas, visualización.
- Filtrado de datos (parsing)
- Eliminación de duplicados (limpieza de datos)
- Tratamiento de nulos (limpieza de datos): NaN tienen un significado? se pueden corregir? o se deben eliminar?
- Tuplas perdidas (limpieza de datos): Cómo resolver cuando hay tuplas/filas perdidas en el dataset. Por ejemplo en mediciones horarias me falta las mediciones de una determinada hora.
- Datos dañados/corruptos (limpieza de datos): es mejor no tener datos que tener datos corruptos? conviene eliminarlos? es necesario generar algún reporte?
- Formato de datos (transformación): re-formatear los datos para ser transmitidos.
- Metadata (transformación): Organizarla/agregar si es necesario

Pre-procesamiento: Normalización/Estandarización

Objetivo: Escalar los datos.



https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html

- Escalar: "ajustar" los datos dentro de un rango específico
- Ayuda a mejorar la interpretación de los resultados del análisis de datos
- Obligatorio para la mayoría de los algoritmos de ML
- Normalizar es escalar los datos desde su valores originales a un rango entre 0 y 1
- Estandarizar se refiere a escalar la distribución de los datos de forma tal que la media de los valores observados sea igual a 0 y su desviación estándar igual a 1

https://seeing-theory.brown.edu/es.htm

Pre-procesamiento: Normalización/Estandarización

Reescalado de los datos de según su valor mín y máx

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

```
import numpy as np
from sklearn.preprocessing import MinMaxScaler
dataset = np.array([1.0, 12.4, 3.9, 10.4]).reshape(-1, 1)
scaler = MinMaxScaler(feature_range=(0, 1.5))
scaler.fit(dataset)
normalized_dataset = scaler.transform(dataset)
print(normalized_dataset)
```

```
[[0.        ]
 [1.5       ]
 [0.38157895]
 [1.23684211]]
```

```
print(np.mean(normalized_dataset))
print(np.std(normalized_dataset))
```

```
0.7796052631578947
0.611196249385709
```

Aún se conservan algunas diferencias en escala (diferentes unidades)

Reescalado de los datos de manera que su promedio=0 y su desviación estándar=1

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

```
import numpy as np
from sklearn.preprocessing import StandardScaler
dataset = np.array([1.0, 12.4, 3.9, 10.4]).reshape(-1, 1)
scaler = StandardScaler()
scaler.fit(dataset)
standardized_dataset = scaler.transform(dataset)
print(standardized_dataset)
print(np.mean(standardized_dataset))
print(np.std(standardized_dataset))
```

```
[[-1.27554   ]
 [ 1.17866354]
 [-0.65122506]
 [ 0.74810152]]
-1.1102230246251565e-16
0.9999999999999998
```

Las escalas son comparables (mediante la desviación estandar)

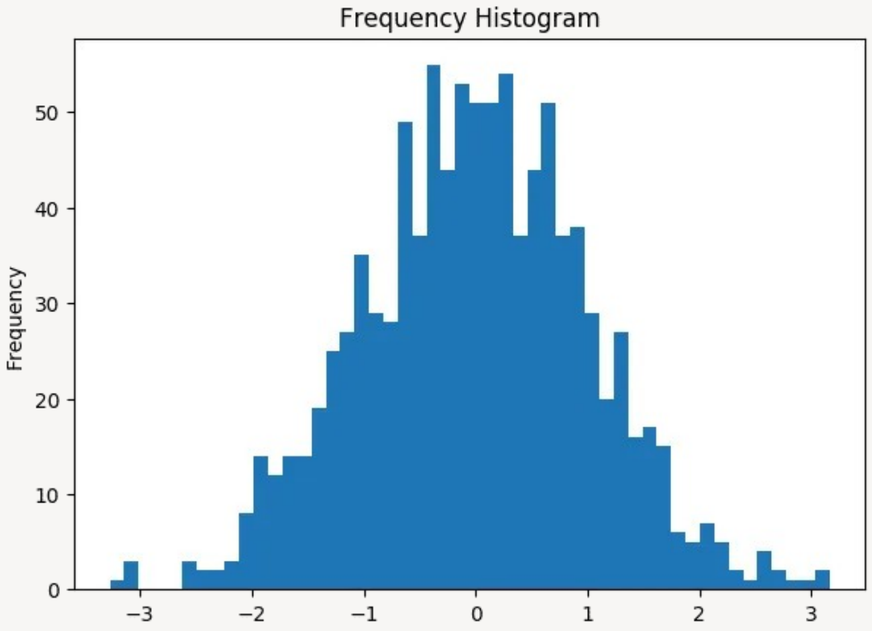
Pre-procesamiento: Exploración de los datos

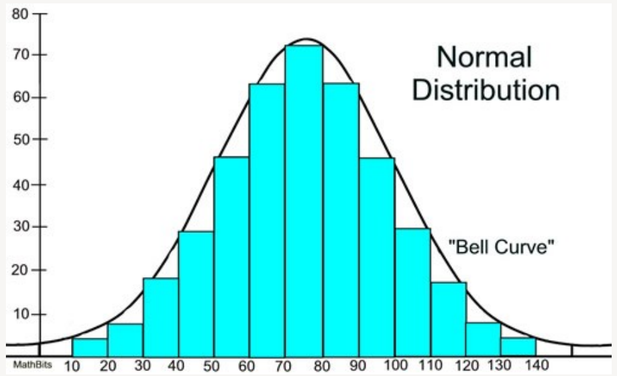
Objetivo: Entender la distribución de los datos,.

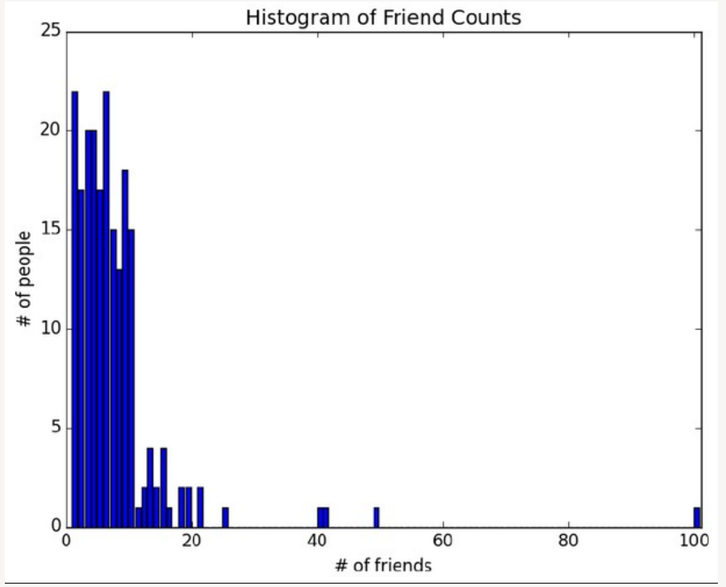
- Análisis de cantidad de datos nulos, rango, número de muestras, valores máximos y mínimos en el dataset, outliers.
- Si es posible, visualizar los datos como originales
- Histograma de frecuencia
- Tendencias centrales: queremos tener una noción de dónde están centrados nuestros datos (promedio, mediana, moda, cuartiles)
- Medidas de dispersión de los datos: rango intercuartil, varianza
- Hay muchas más estadísticas que iremos viendo a lo largo de los temas (por ejemplo correlación entre variables, estacionalidad en series de tiempo, etc)

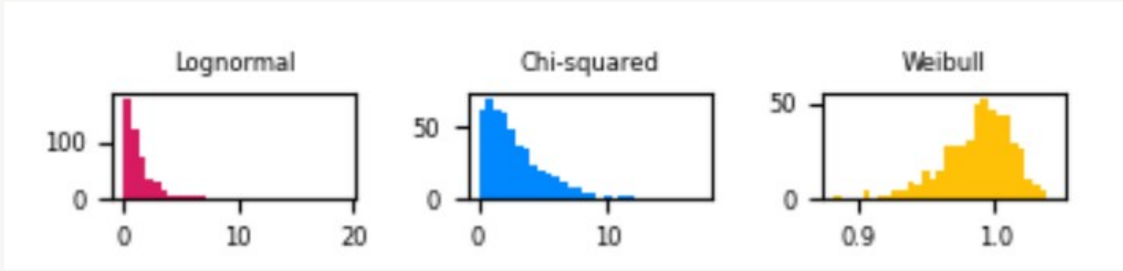
Pre-procesamiento: Exploración de los datos

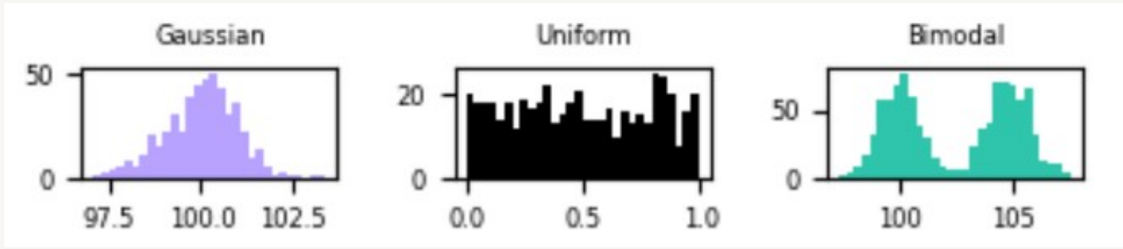
Histograma de frecuencias







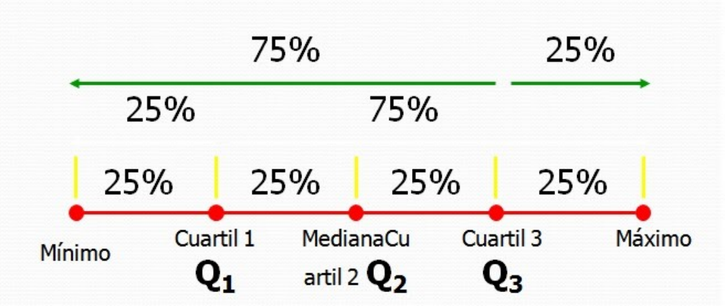




Pre-procesamiento: Exploración de los datos

- Tendencias centrales (promedio, moda, mediana)
- Cuartiles

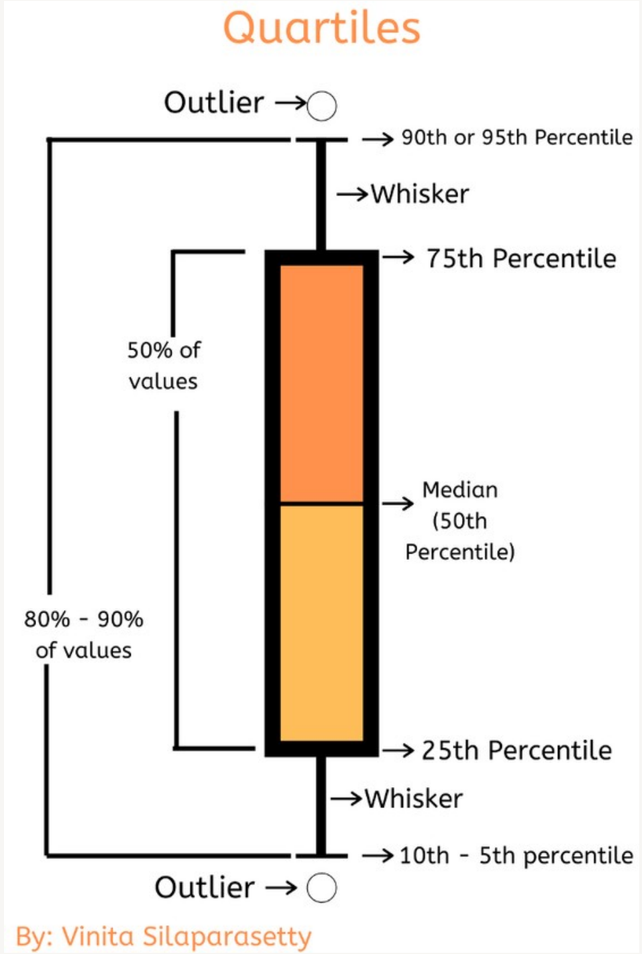
dividen al conjunto de datos ordenados en cuatro partes porcentualmente iguales



matplotlib.pyplot.boxplot

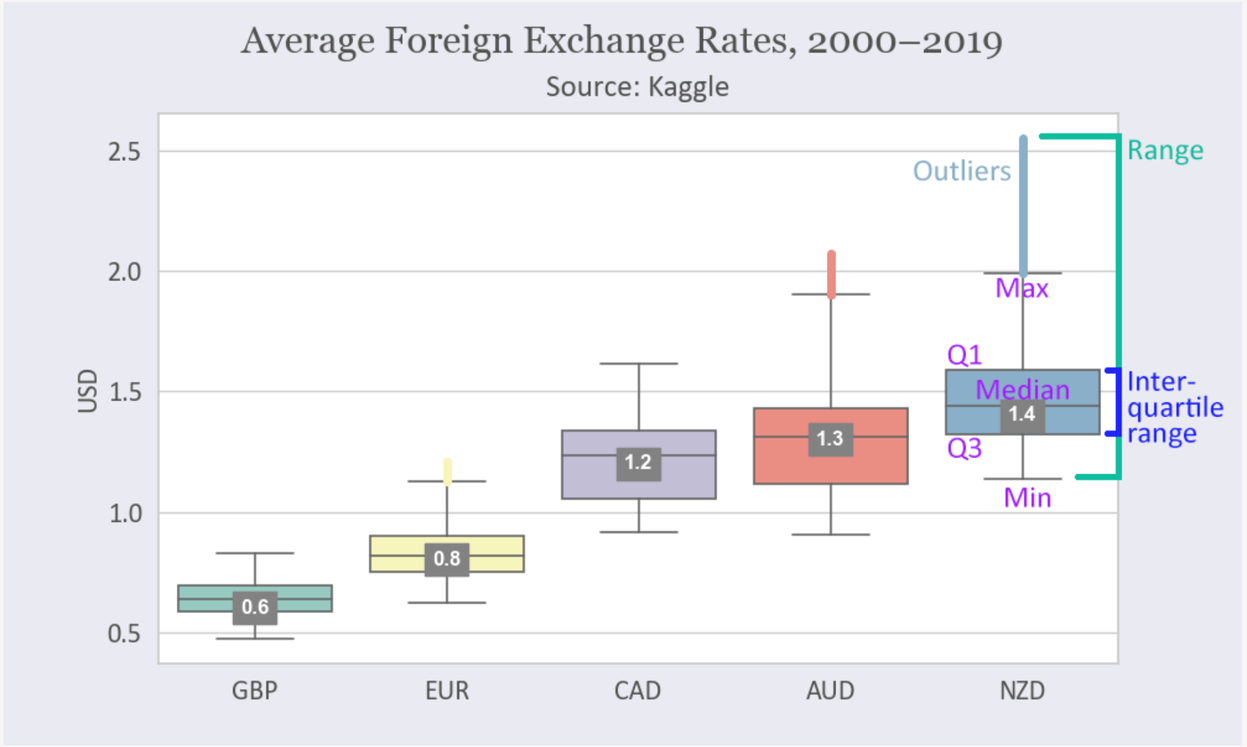
```
matplotlib.pyplot.boxplot(x, notch=None, sym=None, vert=None,
whis=None, positions=None, widths=None, patch_artist=None,
bootstrap=None, usermedians=None, conf_intervals=None, meanline=None,
showmeans=None, showcaps=None, showbox=None, showfliers=None,
boxprops=None, labels=None, flierprops=None, medianprops=None,
meanprops=None, capprops=None, whiskerprops=None, manage_ticks=True,
autorange=False, zorder=None, capwidths=None, *, data=None)
```
[source]

Draw a box and whisker plot.



Pre-procesamiento: Exploración de los datos

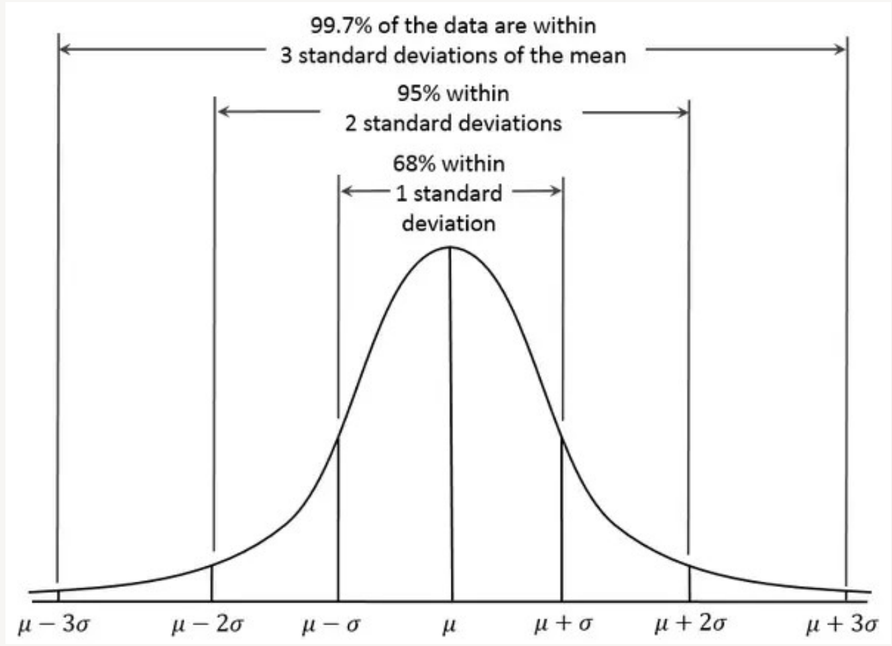
- Medidas de dispersión (varianza, rango intercuartil)



Rango intercuartil: más robusto que la varianza. Calcula la diferencia entre el 75avo percentil y el 25avo percentil. Es poco afectado por los valores alejados

La desviación típica o estándar (raíz cuadrada de la varianza) es una medida de la dispersión de los datos en relación al promedio, cuanto mayor sea la dispersión mayor es la desviación estándar. Así, si no hubiera ninguna variación en los datos, es decir (todos iguales), entonces la desviación estándar sería cero.

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$



Comenzamos con el TP1