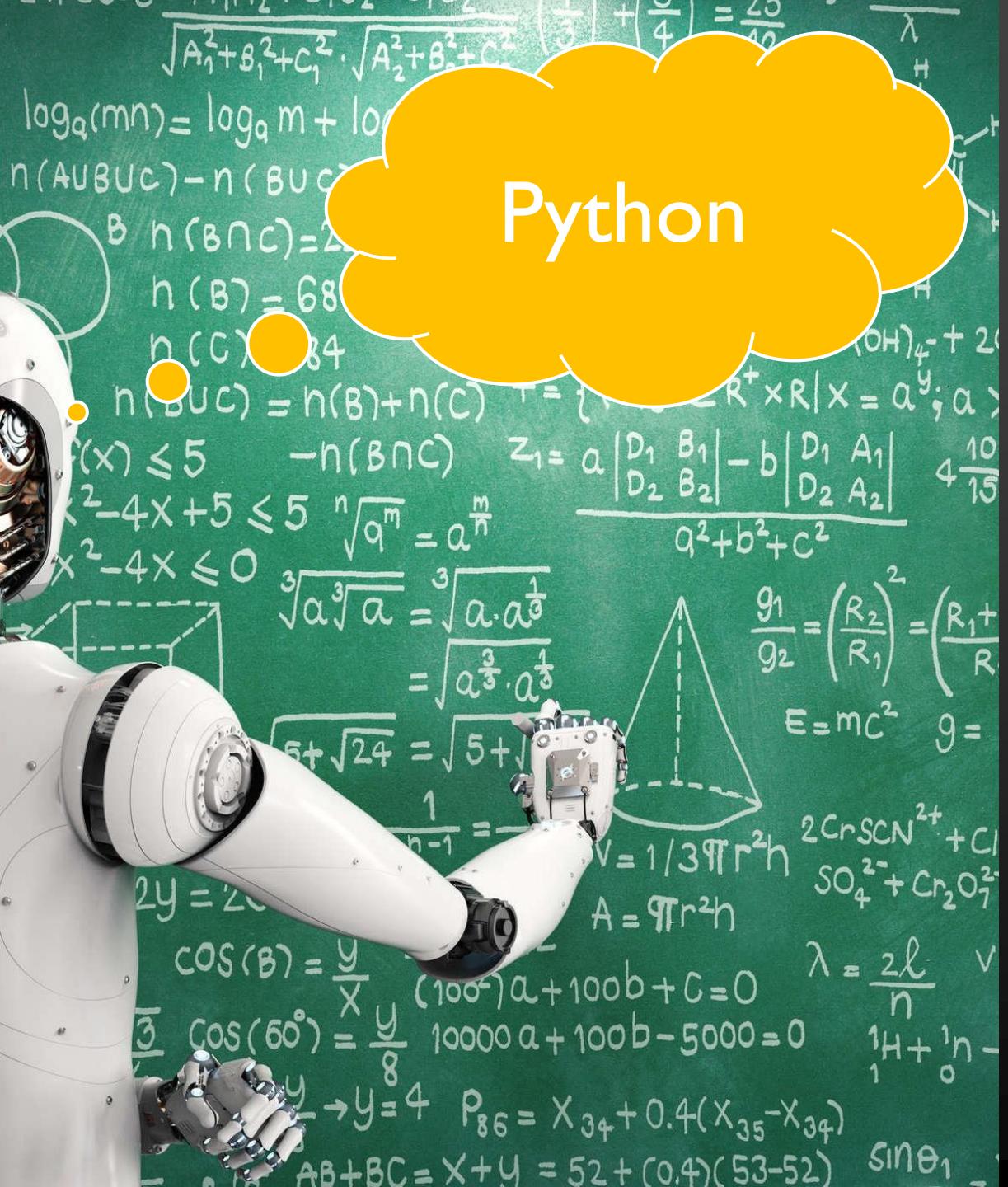
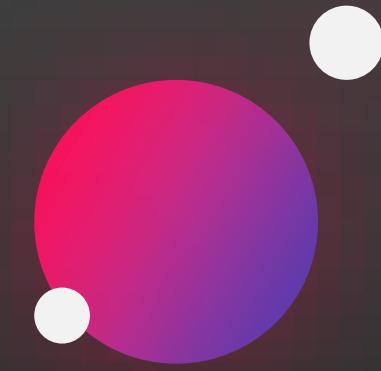


Introducción al aprendizaje automático en Python

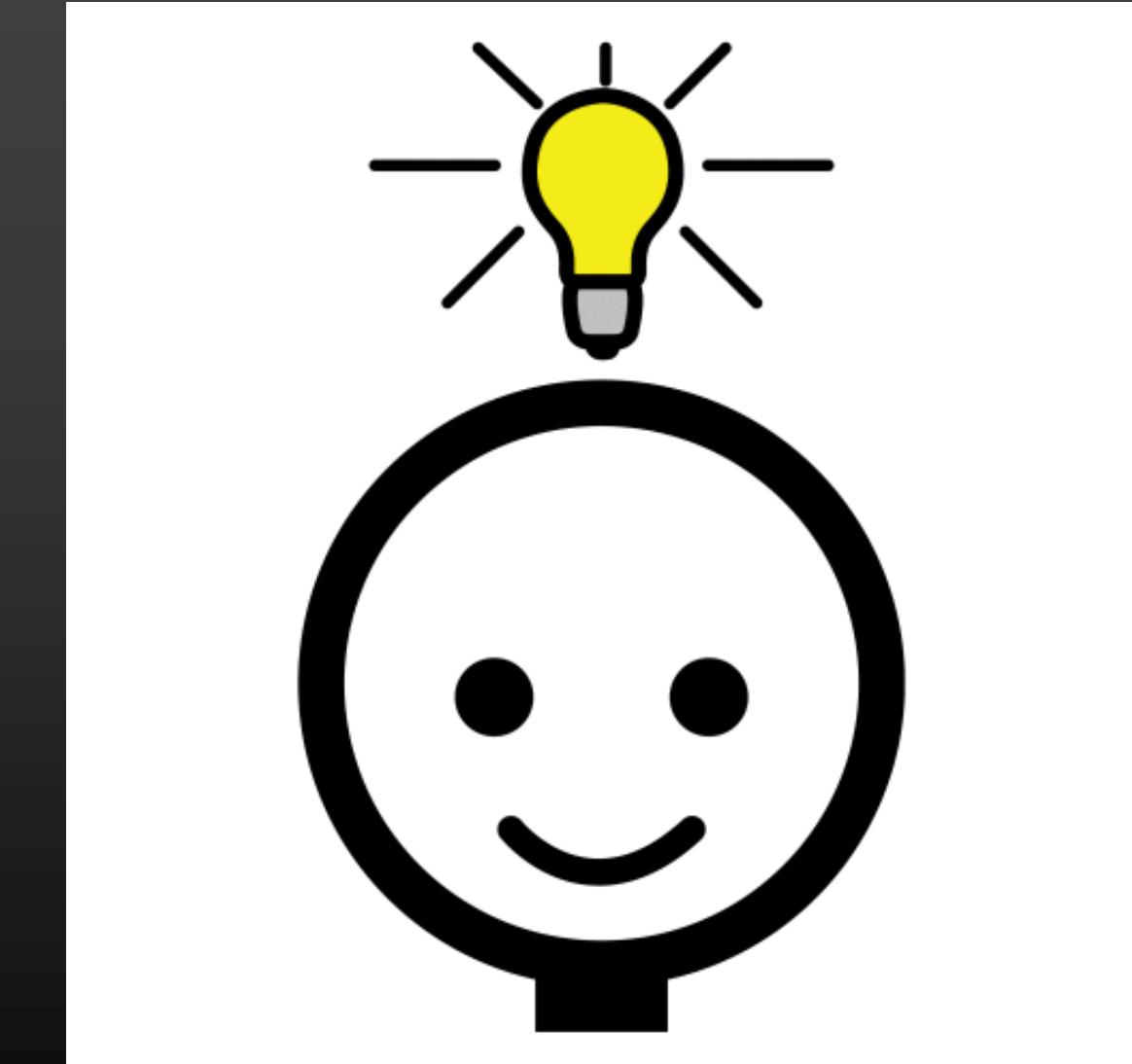
Ing. Jose Mariano Alvarez

<http://blog.josemarianoalvarez.com/>





Entendiendo de que hablamos



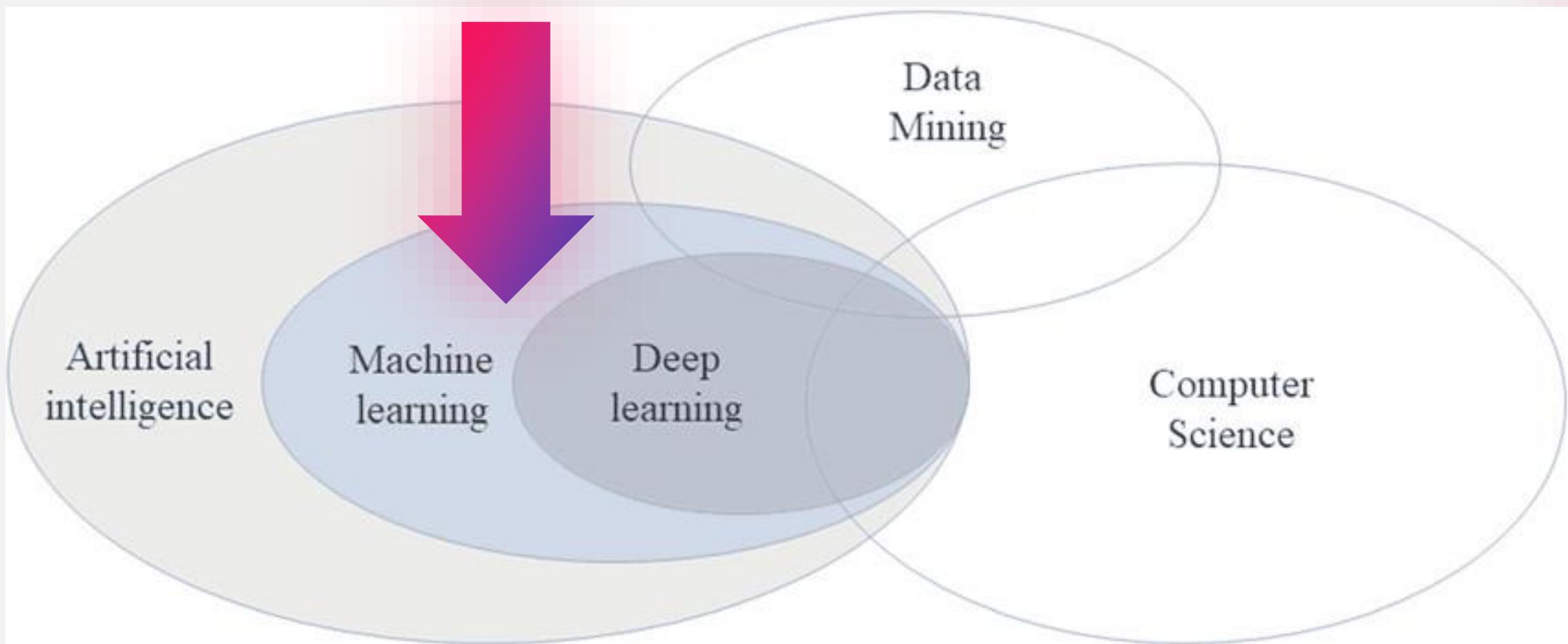


De todo esto !!!

Aplicaciones del Aprendizaje Automático o Machine Learning



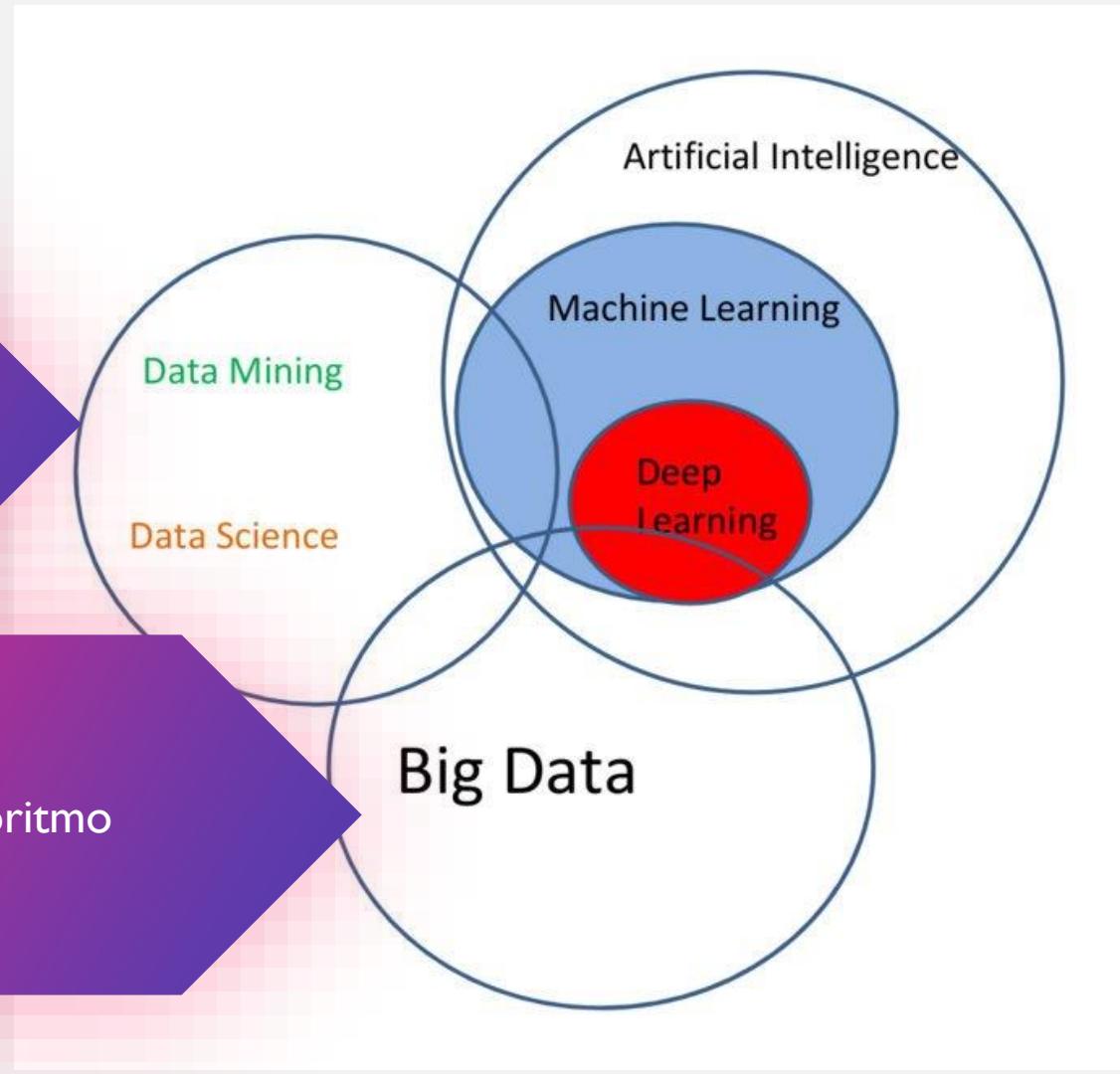
Donde está Machine Learning



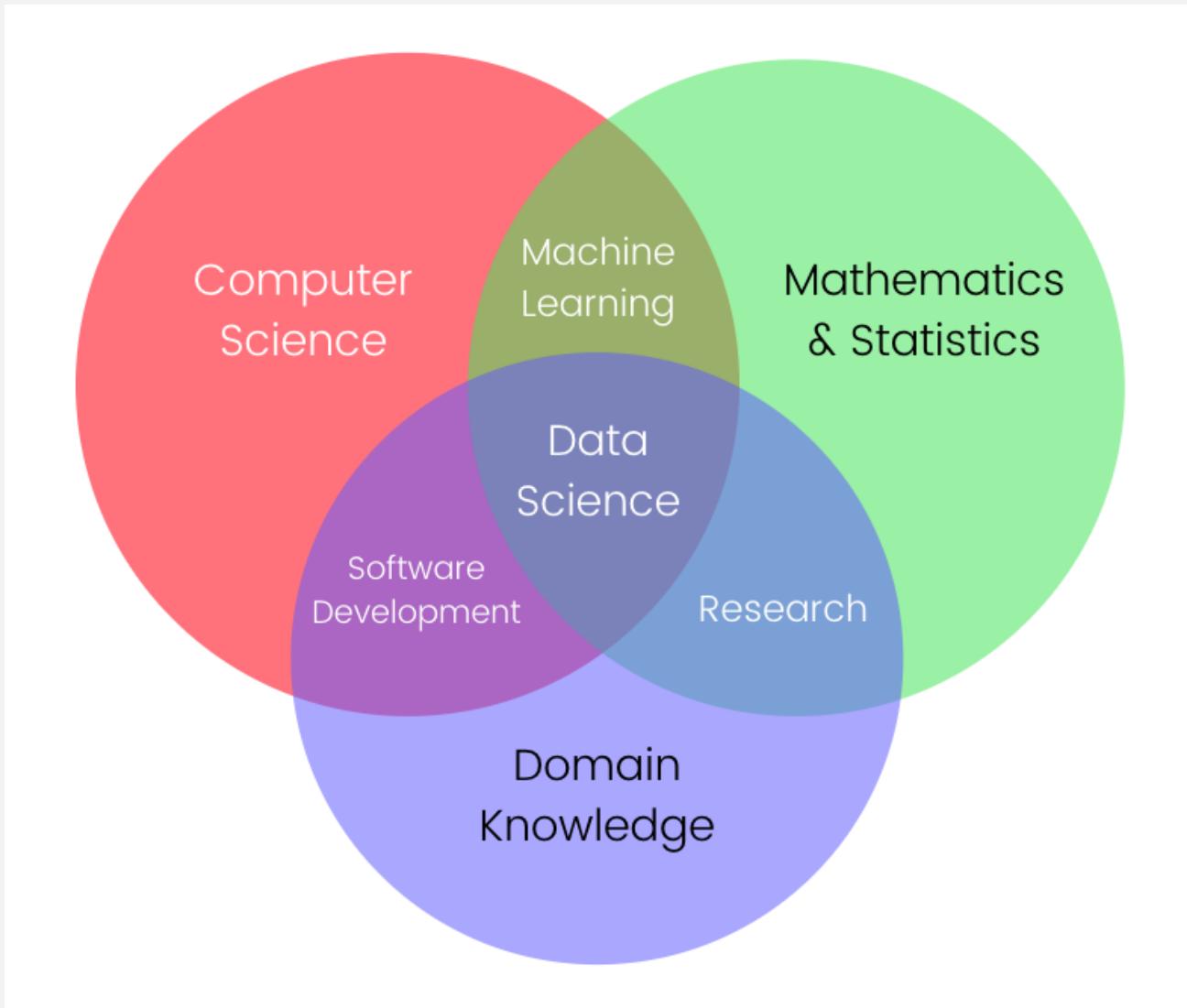
Algoritmos vs Inteligencia Artificial

No son solo algoritmos

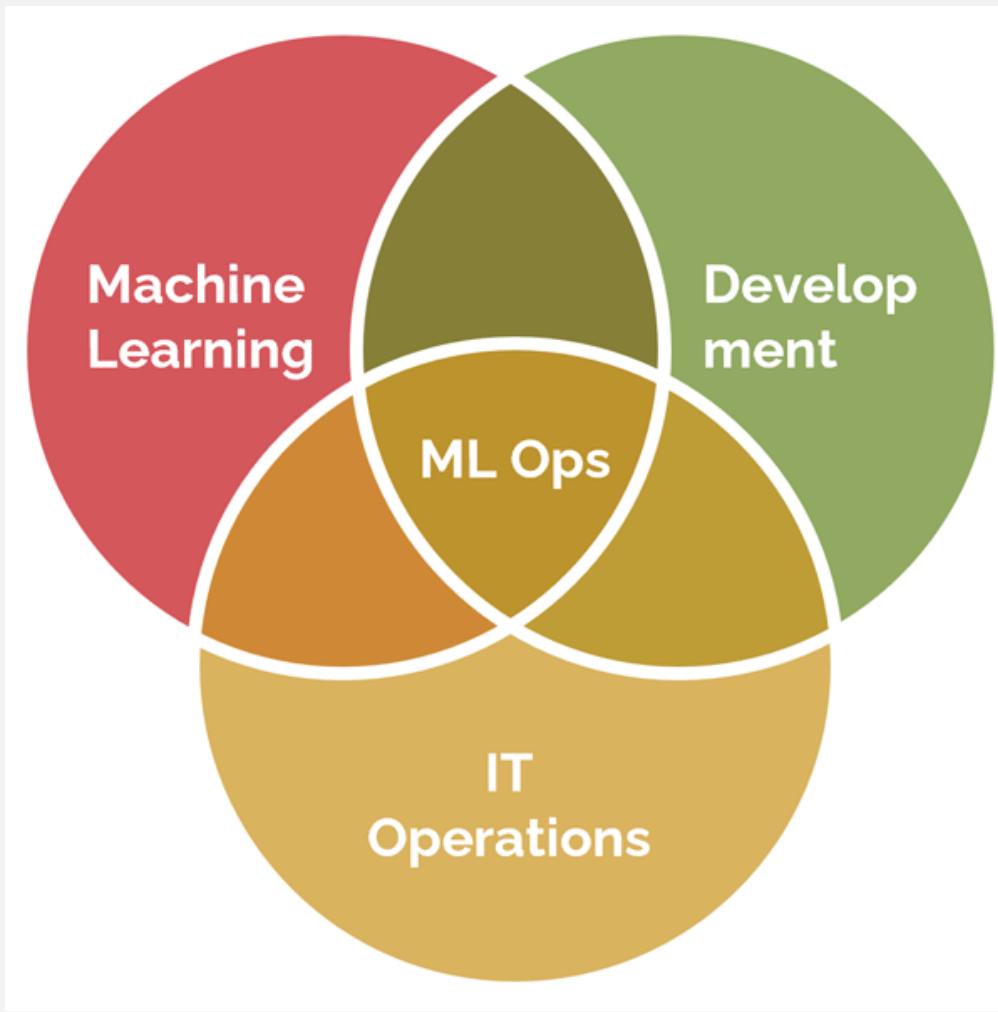
No se puede aplicar cualquier algoritmo



Donde está data science



MLOPS





```
conda env list  
conda env create -f stech.yml  
conda remove --name stech --all  
conda info --envs  
conda env list
```

```
$ cd project-directory  
$ conda activate stech  
(sabadostech) project-directory $  
conda deactivate
```

```
conda list  
conda list -n stech
```

```
Anaconda Powershell Prompt (anaconda3)  
(base) PS C:\Users\josem> conda activate stech  
(stech) PS C:\Users\josem> cd \code  
(stech) PS C:\code>
```

stech.yml

```
1 #conda env create -f stech.yml
2 name: stech
3 channels:
4   - anaconda
5   - conda-forge
6   - defaults
7 dependencies:
8   - python>=3.7.8
9   - pip
0   - pandas
1   - tensorflow=2.5
2   - py-xgboost
3   - lightgbm
4   - catboost
5   - numpy
6   - scipy
7   - scikit-learn
8   - statsmodels
9   - ipykernel
0   - jupyter
1   - jupyterlab
2   - matplotlib
3   - seaborn
4   - graphviz
5   - pyyaml
6   - jsonlines
7   - mlflow
8   - pip:
9     - mlflow
10
```





Anaconda

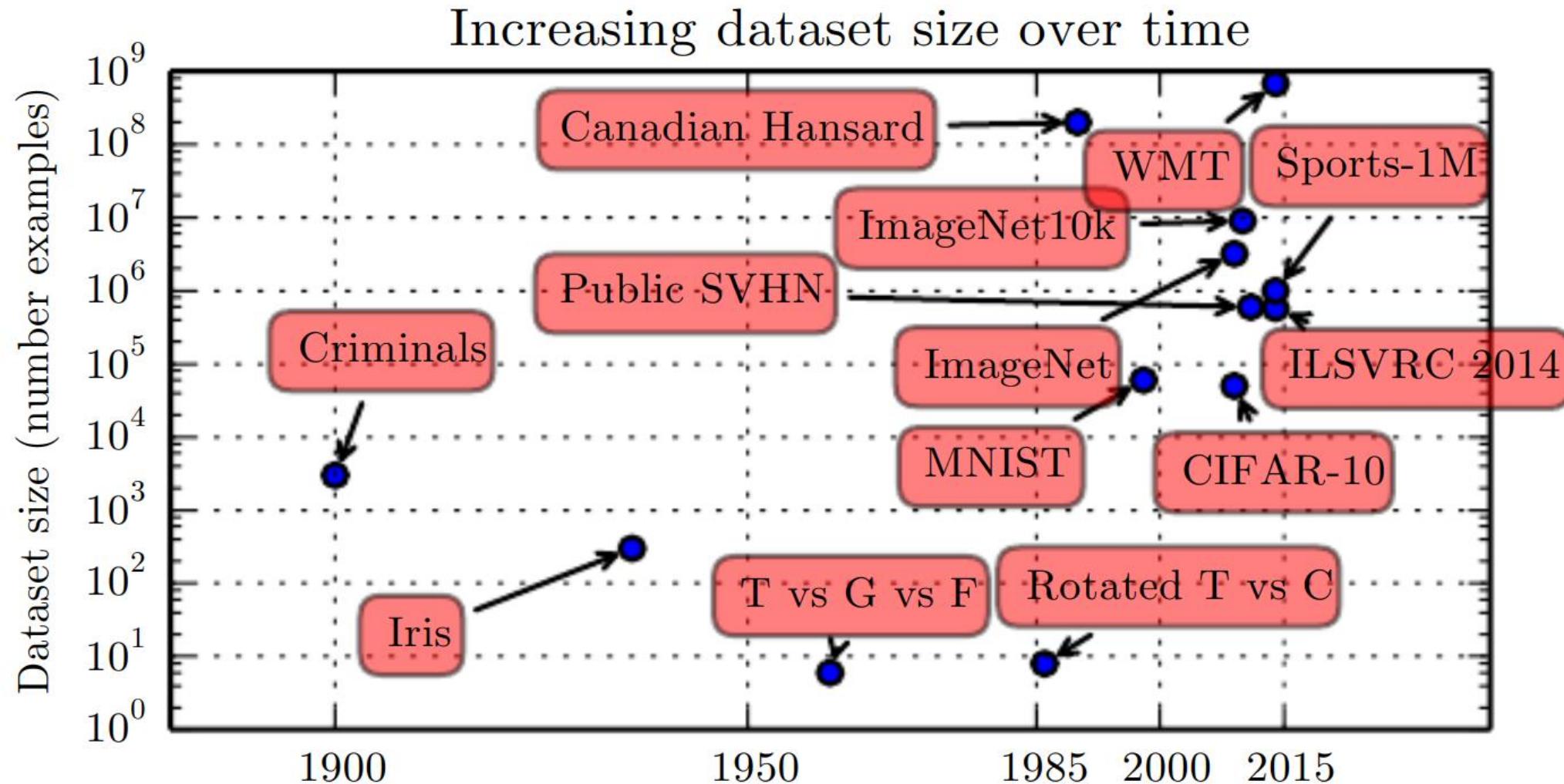


Problemas y algoritmos

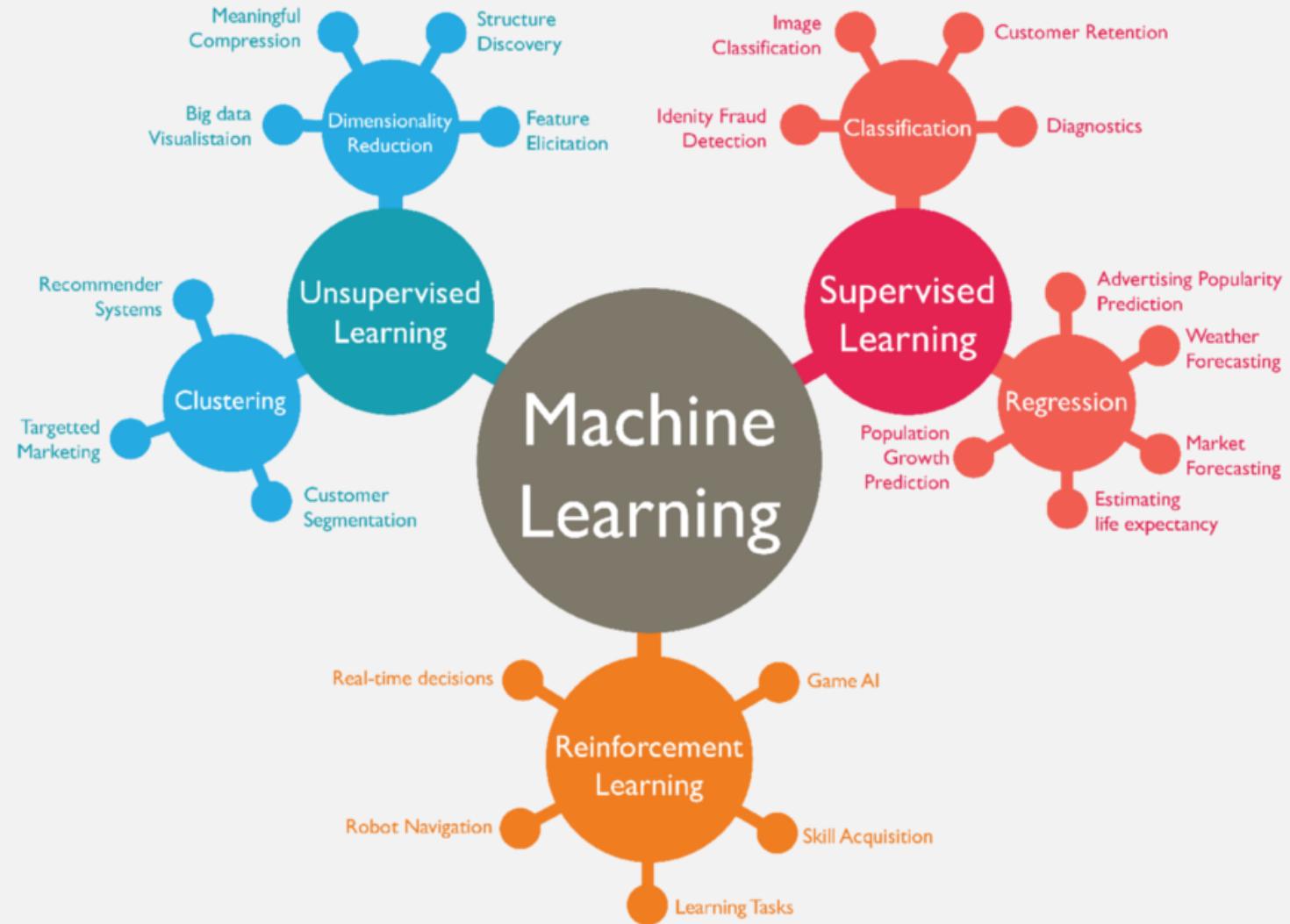


Escala logarítmica

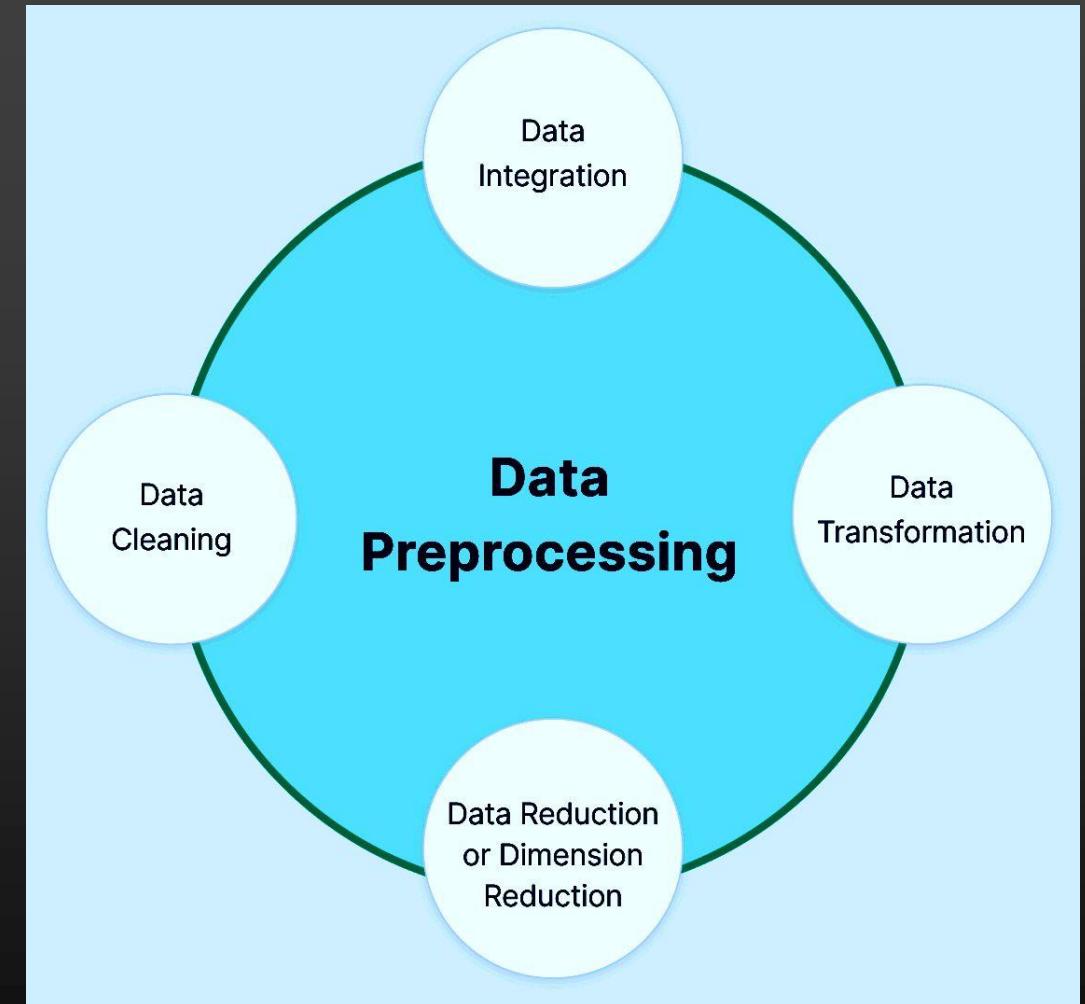
Tamaño de los Datasets



Tipo de problemas



Preprocesamiento



Técnicas de preprocessamiento

Faltantes

Ignorar

Constante

Calculado

Probabilidad

Ruido

Binning

Clustering

Regresión

ML

Removerlo

Inconsistentes

Datos externos

Bases de conocimiento

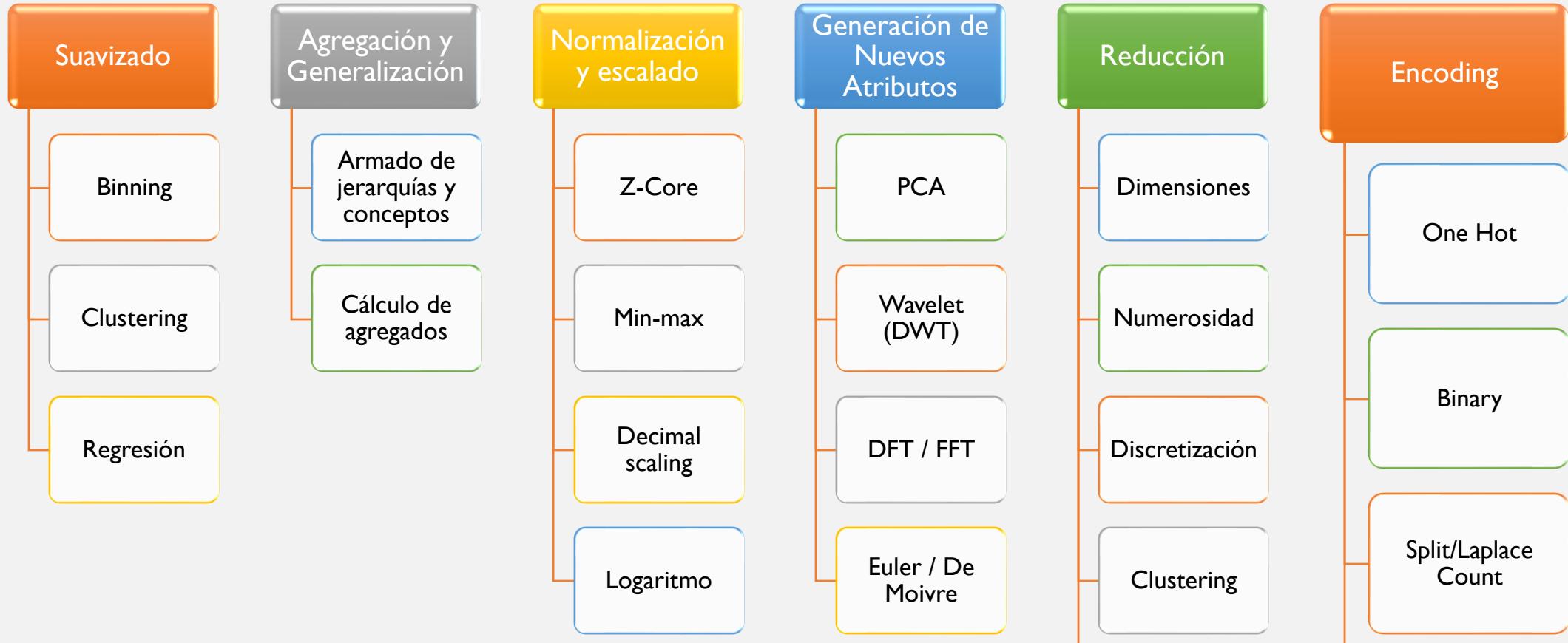
Escalado

Integración

En todos los casos se puede corregir manualmente si es posible



Transformación de datos



Estos son solo algunos ejemplos de las alternativas que existen para transformar los datos

Algunas advertencias



Estimadores segados

- Ley de De Moivre

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

 NEPC
NATIONAL EDUCATION
POLICY CENTER

Answer Sheet: Bill Gates Spent Hundreds of Millions of Dollars to Improve Teaching. New Report Says It Was a Bust

Valerie Strauss
June 29, 2018
◆ Foundation Involvement in School Reform, High-Stakes Testing and Evaluation, School Reform and Restructuring, Teacher Evaluation, Value-Added Assessment



El Principio de Bonferroni

Es posible que ciertos datos aleatorios se confundan con aquellos que estamos buscando realmente

Probemos con un ejemplo

Supongamos que queremos encontrar terroristas y que nuestra presunción es que estos se han reunido más de una vez en hoteles diferentes

RUIDO POSIBLE

- Probabilidad de que dos personas comunes visiten el mismo hotel en días diferentes

Análisis del ejemplo

- **Datos**
 - La población total es de 1000 millones de personas que viajan
 - Una persona visita un hotel una vez cada 100 días
 - Estudiamos un período de tiempo total de 100 días
 - Cada hotel puede alojar a 100 personas
 - En total tenemos 100.000 hoteles
- **Ruido: Hay una esperanza de que 250000 casos son simplemente producto de la casualidad**
- **Conclusión**

En un dataset muy grande es probable que cualquier proceso de filtrado arroje una cantidad importante de resultados falso-positivos que luego tenemos que de alguna manera detectar y eliminar

Notemos que 100.000 hoteles con capacidad para 100 personas cada uno implica que un total de 10^7 personas pueden alojarse al mismo tiempo y eso coincide con el 1% de 10^9 .

Calculemos primero la probabilidad de que dos personas visiten un hotel es $0.01^2 = 0.0001$ para que visiten el mismo hotel hay que dividir por la cantidad de hoteles es decir 100.000 por lo tanto la probabilidad de que dos personas cualesquiera visiten el mismo hotel el mismo día es 10^{-9} . Para que esto ocurra dos veces es decir en dos días diferentes elevamos esta probabilidad al cuadrado y nos da 10^{-18} .

Ahora tenemos que considerar la cantidad total de eventos posibles. La cantidad total de pares de personas es $\binom{10^9}{2} = 5 * 10^{17}$. El número total de pares de días es $\binom{1000}{2} = 5 * 10^5$. Estamos aproximando $\binom{n}{2} \approx n^2/2$ lo cual es válido cuando n es un número grande.

Por lo tanto el número total de sucesos es igual a la cantidad de pares de personas por la cantidad de pares de días por la probabilidad de que un par de personas visiten el mismo hotel en dos días diferentes es decir:

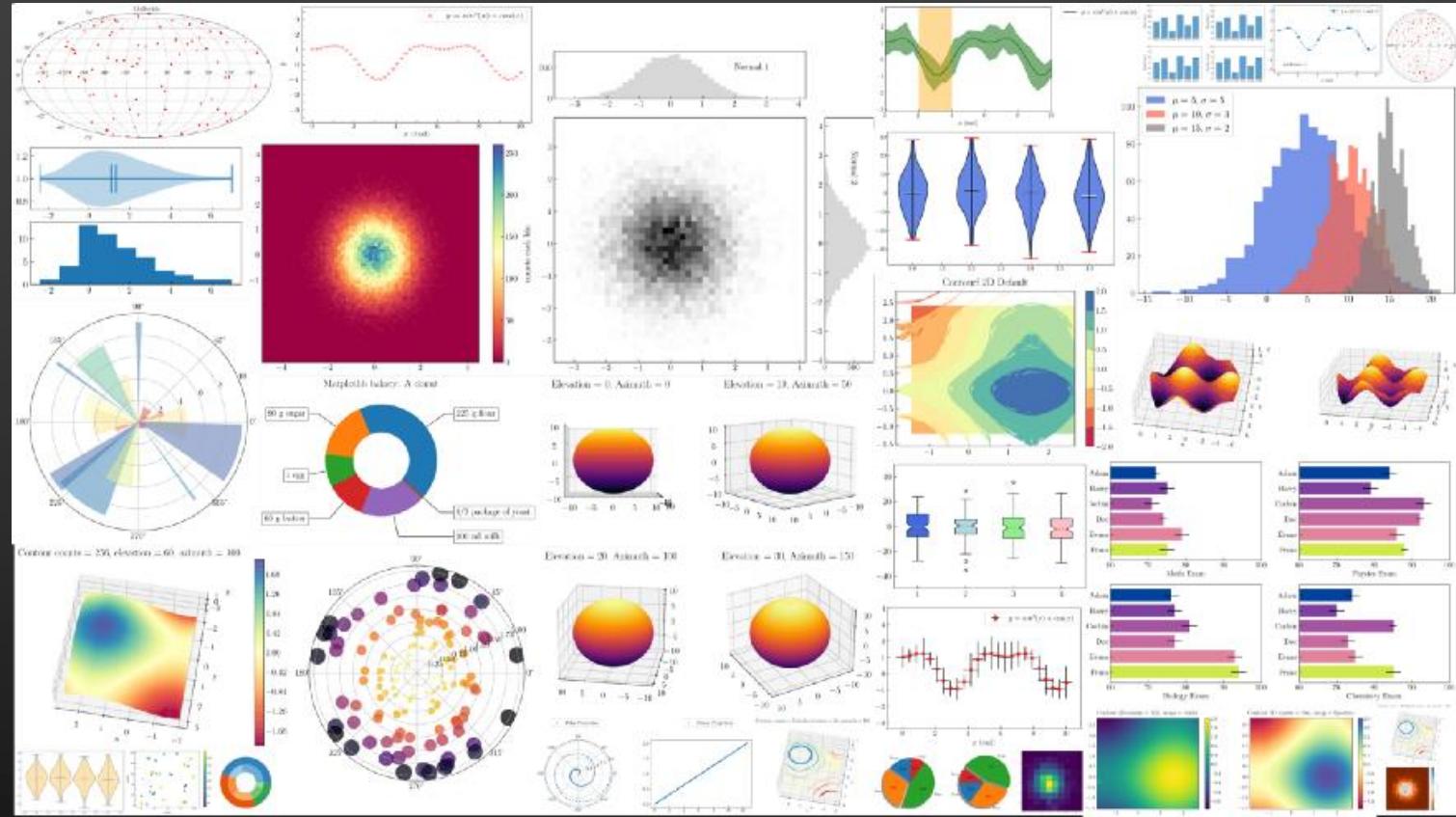
$$5 * 10^{17} * 5 * 10^5 * 10^{-18} = 250000$$

No Free Lunch Theorem

- Dos algoritmos de optimización cualesquiera son equivalentes si los promediamos sobre el set de todos los problemas posibles.
- Dado un problema de optimización, si un algoritmo funciona muy bien, entonces existe un problema en el cual el algoritmo funciona igual de mal.
- No existe un algoritmo que sea óptimo para cualquier problema de optimización.



Visualización de datos

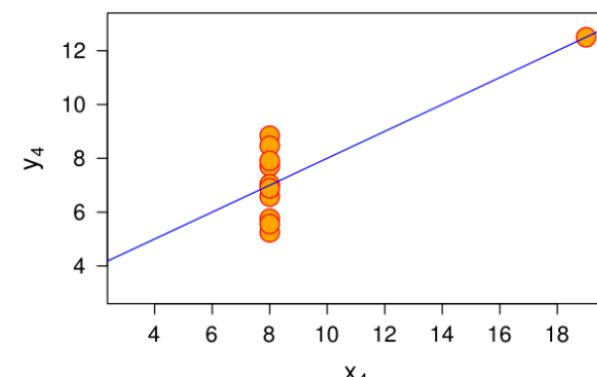
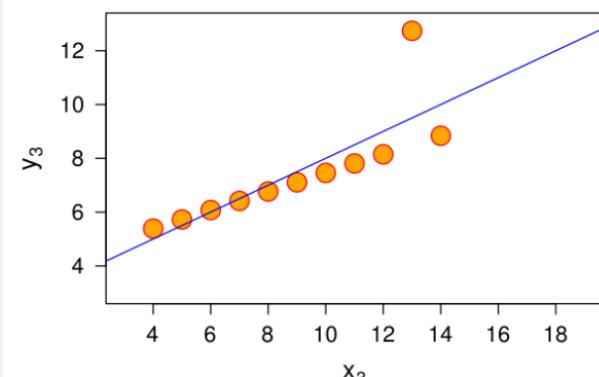
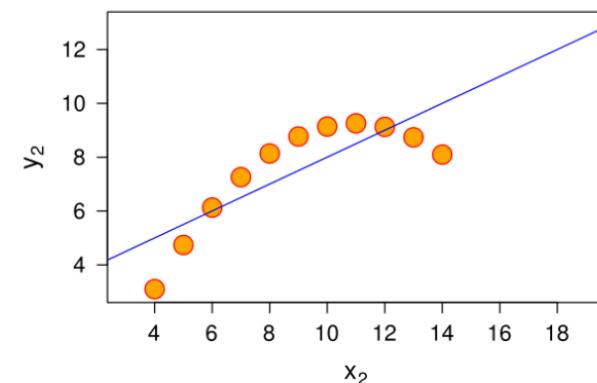
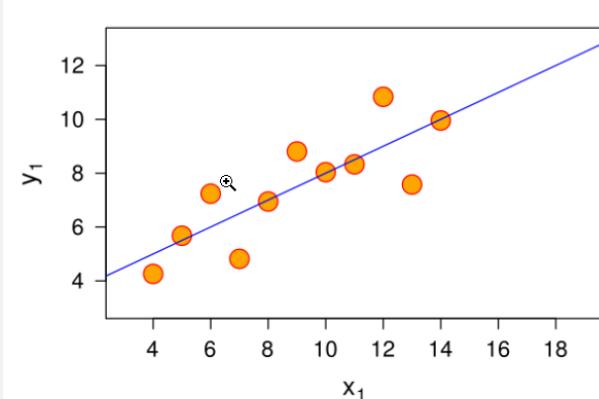


Cuarteto de Anscombe

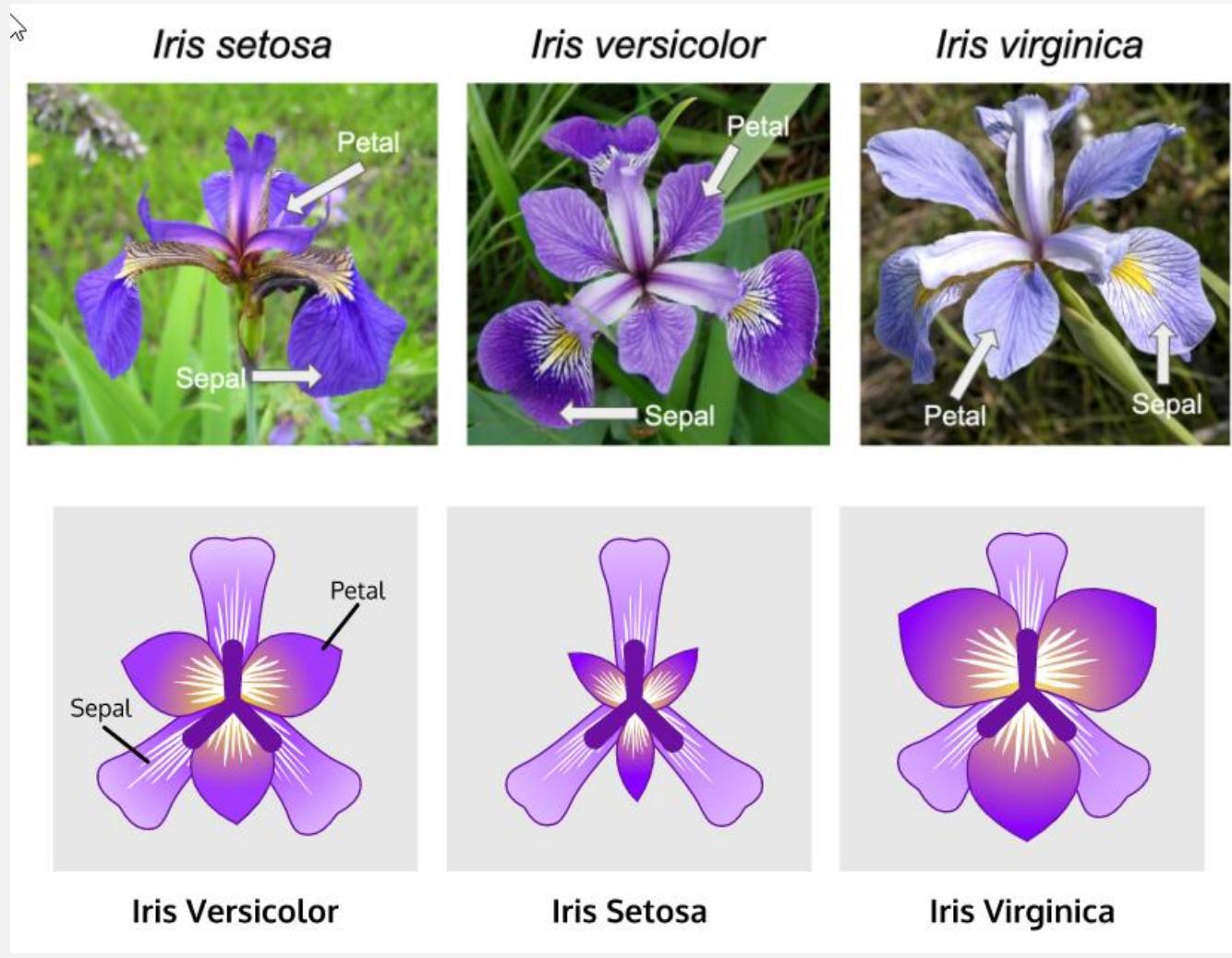
Cuarteto de Anscombe

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

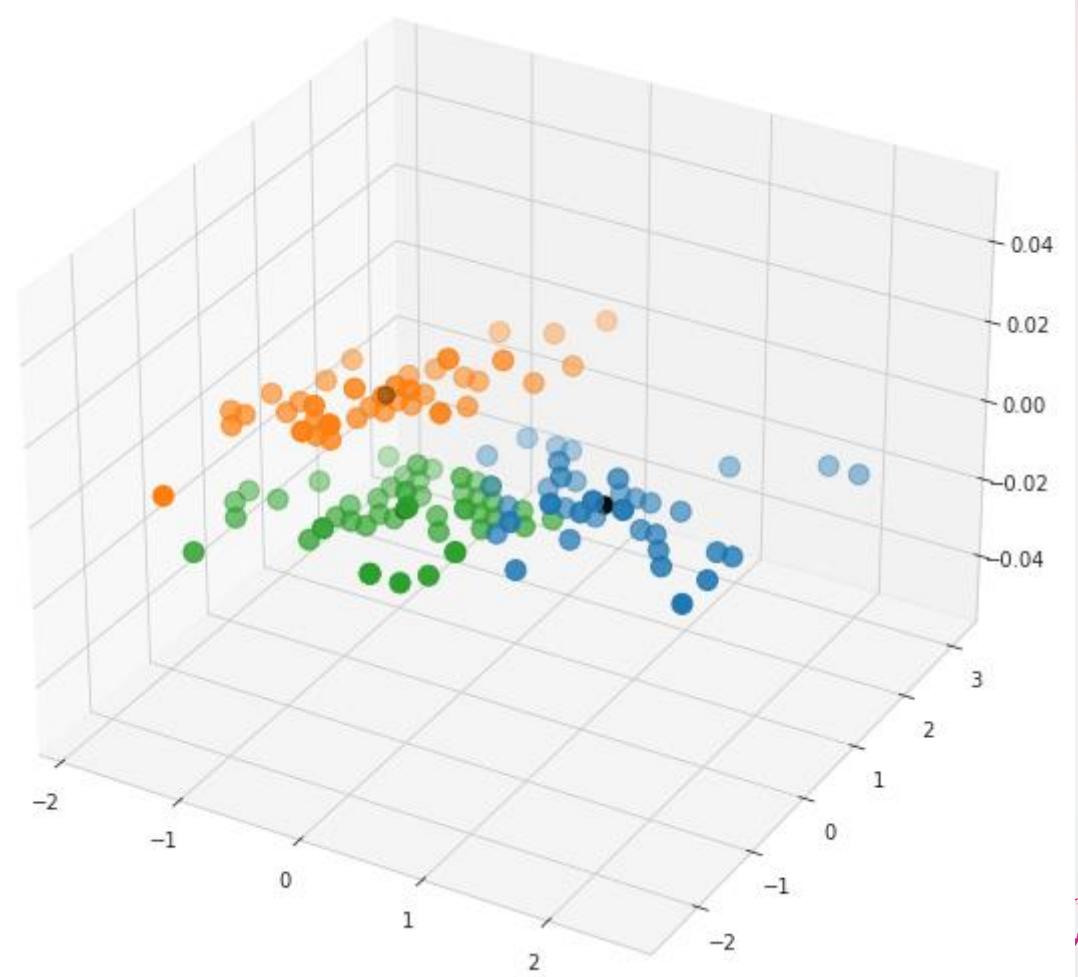
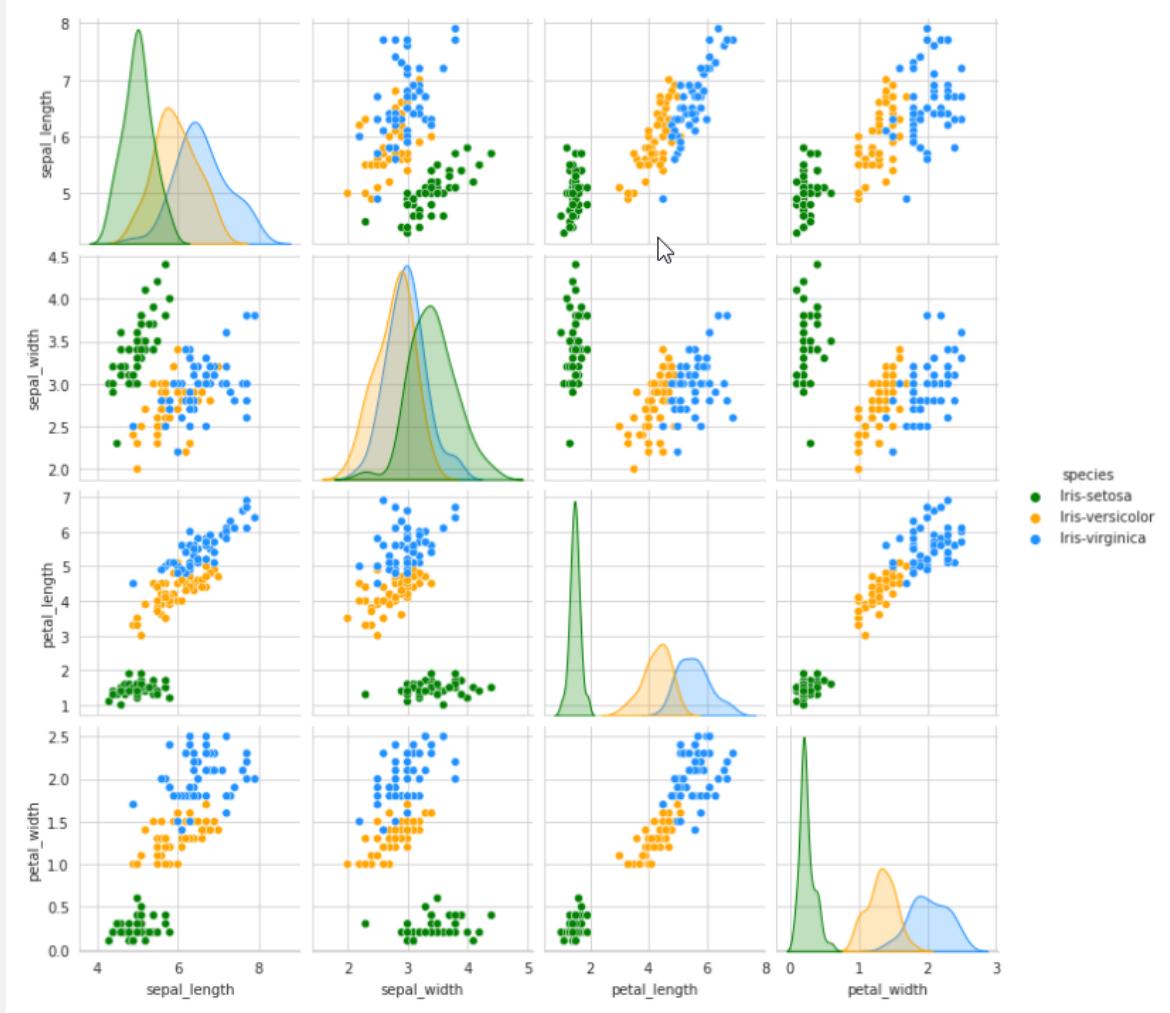
Propiedad	Valor
Media de cada una de las variables x	9.0
Varianza de cada una de las variables x	11.0
Media de cada una de las variables y	7.5
Varianza de cada una de las variables y	4.12
Correlación entre cada una de las variables x e y	0.816
Recta de regresión	$y = 3 + 0.5x$

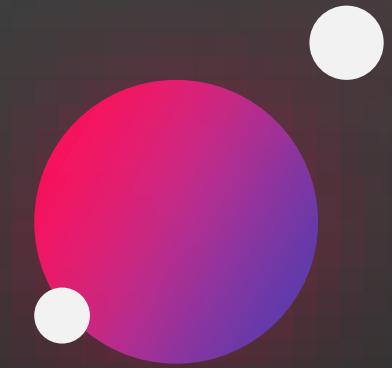


Iris Dataset



Visualizaciones - Iris Dataset





Algoritmos

Tarea T

- La idea es mejorar la realización de una tarea sin programar la forma fija en que se mejora esa tarea sino que el algoritmo “aprende”:
 - Clasificación
 - Regresión
 - Transcripción
 - Traducción
 - Detección de anomalías
 - Síntesis
 - Imputación de valores ausentes
 - Eliminación de ruido
 - Agrupamiento



Medida de performance P

- La idea es que la tarea T debe ser optimizada tomando como base la medida de performance o rendimiento P tratando de optimizar el valor de P

- Matriz de confusión
- Área bajo la curva ROC (AUC)
- Muchas otras

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^T (\hat{y}_t - y_t)^2}{T}}.$$



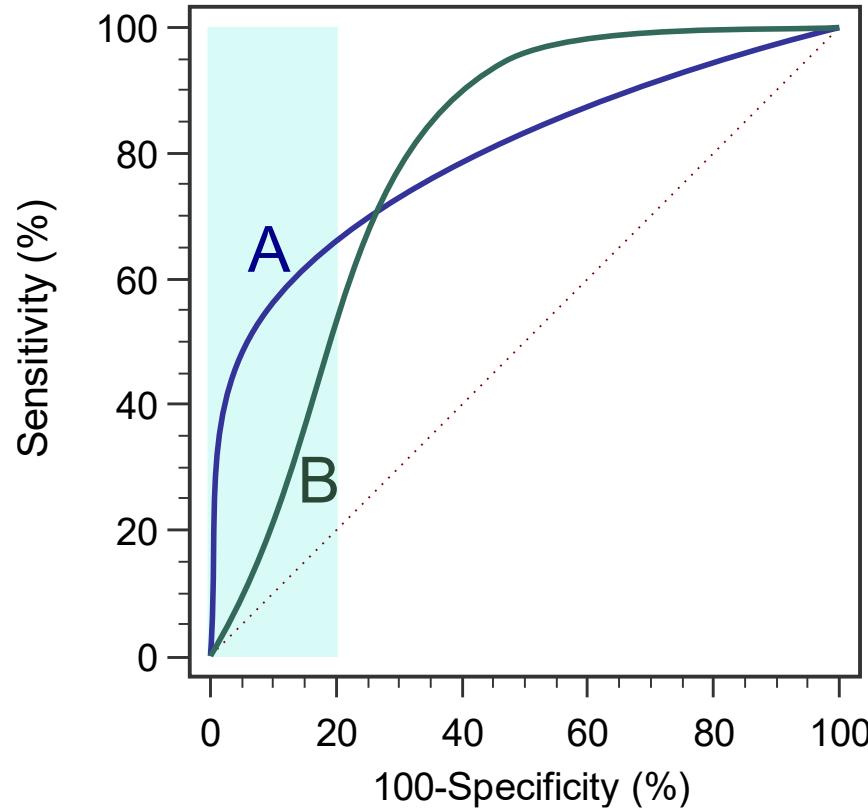
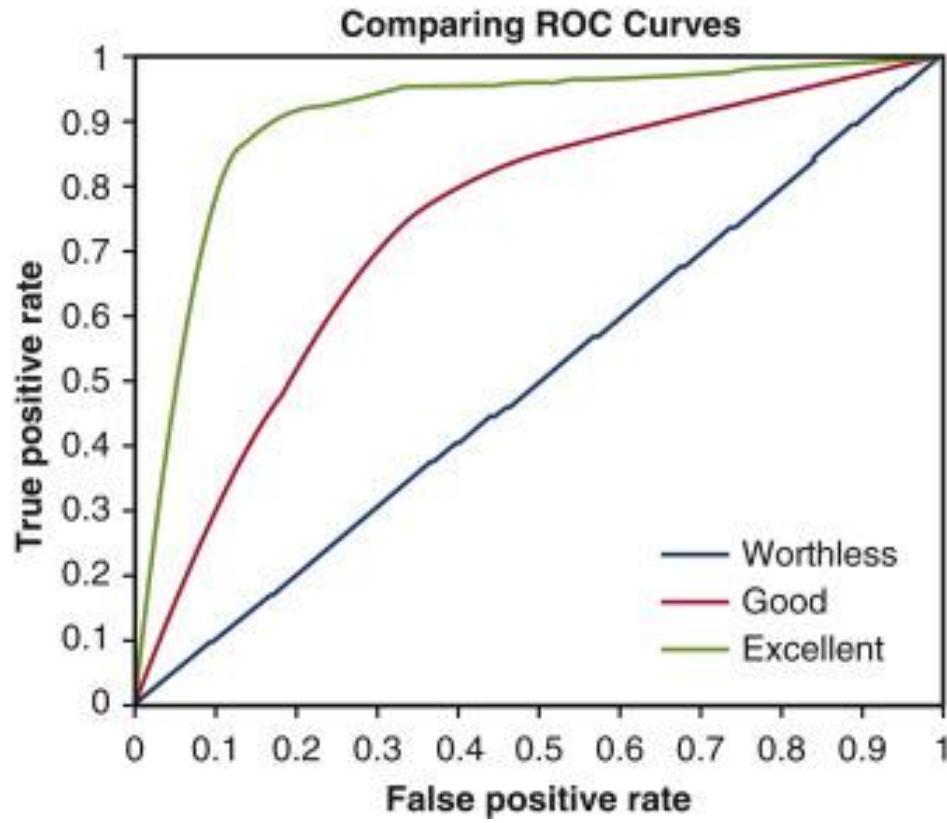
https://www.tensorflow.org/api_docs/python/tf/keras/metrics

<https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics>

Matriz de confusión

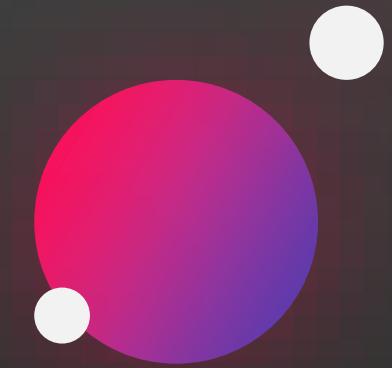
Predicted condition			Sources: [20][21][22][23][24][25][26][27] view · talk · edit	
Total population = P + N	Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	Positive (P)	False negative (FN), type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate = $\frac{FN}{P} = 1 - TPR$
Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection	False positive rate (FPR), probability of false alarm, fall-out = $\frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity = $\frac{TN}{N} = 1 - FPR$
Prevalence = $\frac{P}{P+N}$	Positive predictive value (PPV), precision = $\frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) = $\frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Negative likelihood ratio (LR-) = $\frac{FNR}{TNR}$
Accuracy (ACC) = $\frac{TP + TN}{P + N}$	False discovery rate (FDR) = $\frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) = $\frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) = PPV + NPV - 1	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$
Balanced accuracy (BA) = $\frac{TPR + TNR}{2}$	F_1 score = $\frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) = $\sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) = $\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times DFR}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP + FN + FP}$

ROC



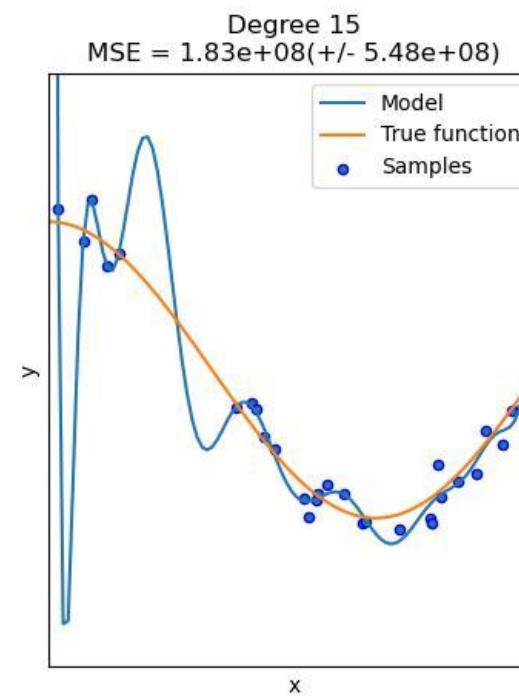
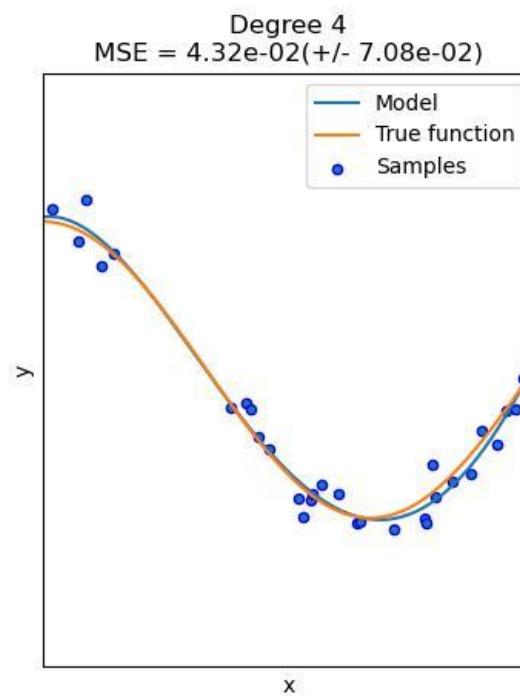
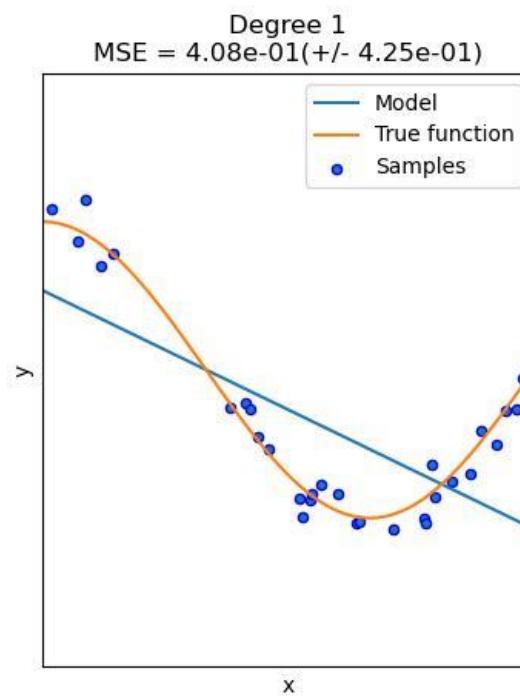
Experiencia E

- Basado en la experiencia requerida podemos clasificar los algoritmos en dos grandes áreas:
 - Algoritmos de aprendizaje NO SUPERVISADO
 - La experiencia E es nula y no existe en el dataset
 - Estos algoritmos requieren datasets con solo los datos que se usan para optimizar la medida de rendimiento P a medida que progresá el aprendizaje para realizar de la tarea T
 - Algoritmos de aprendizaje SUPERVISADO
 - Requiere de datos que contienen además los resultados o Experiencia E para a aprender
 - La experiencia E que es usada para mejorar la tarea T debe usar una medida de rendimiento P consistente con esta experiencia, el algoritmo y la tarea T.
 - La experiencia E puede contener ruido



Overfitting vs Underfitting

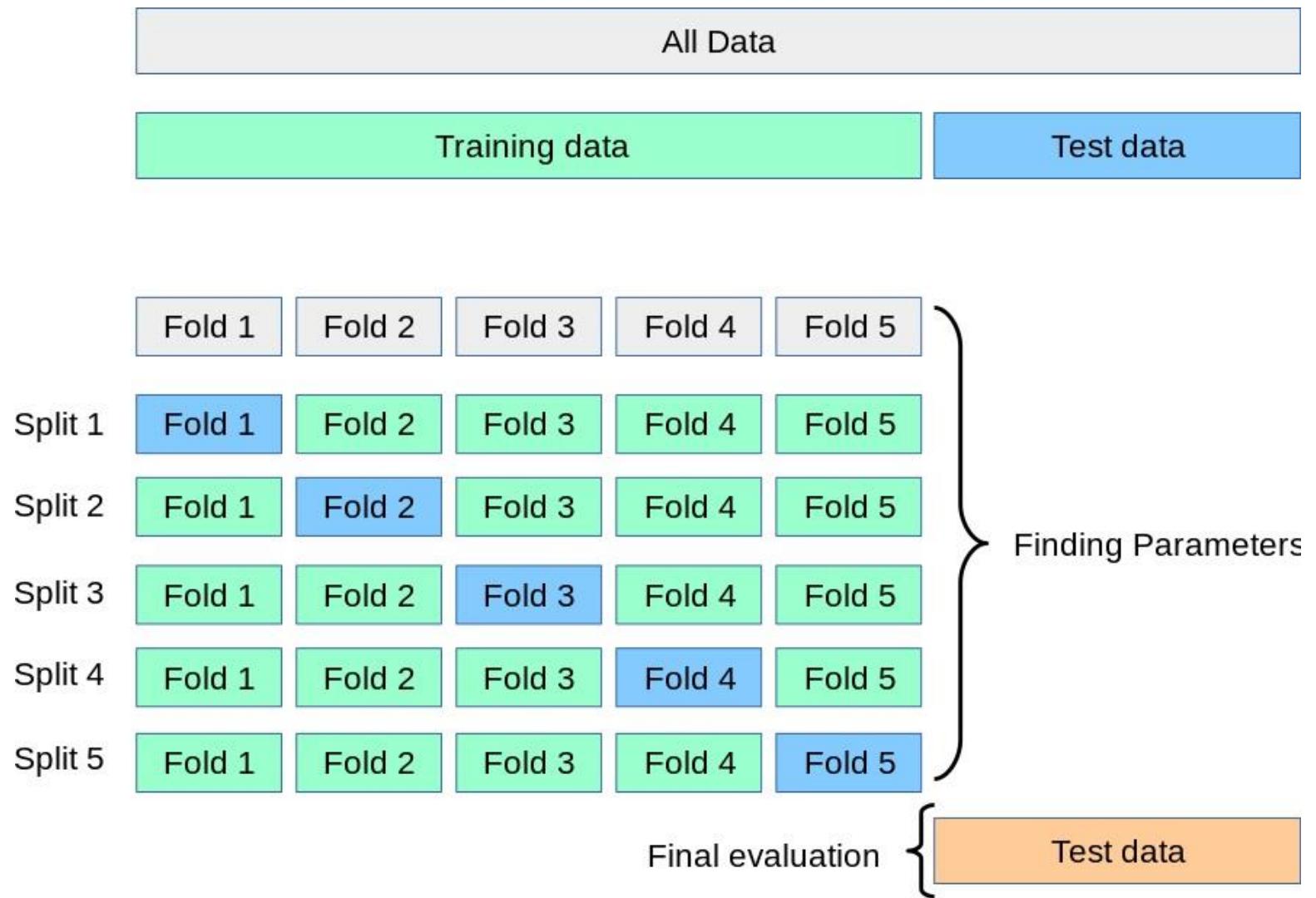
Overfitting vs Underfitting



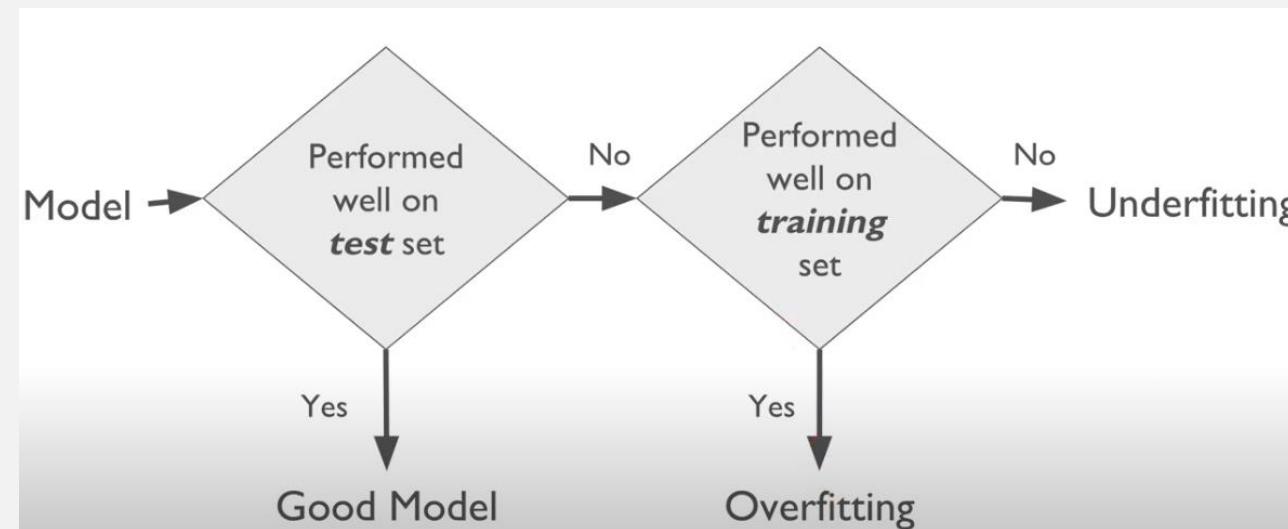
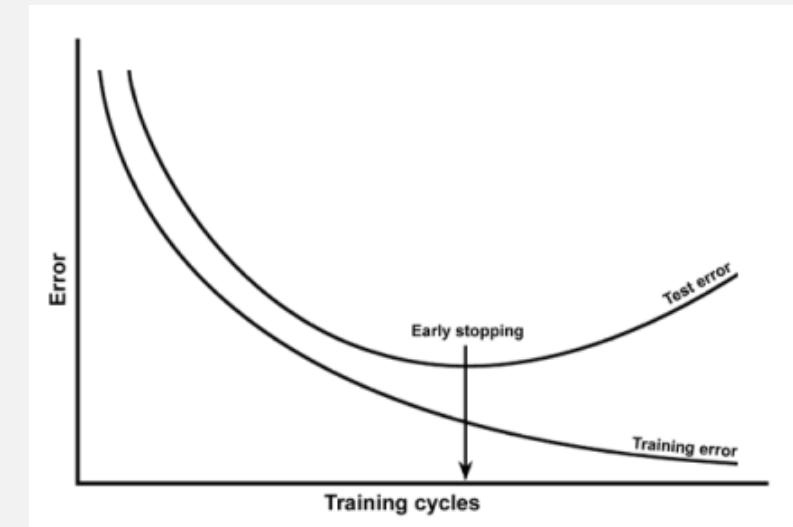
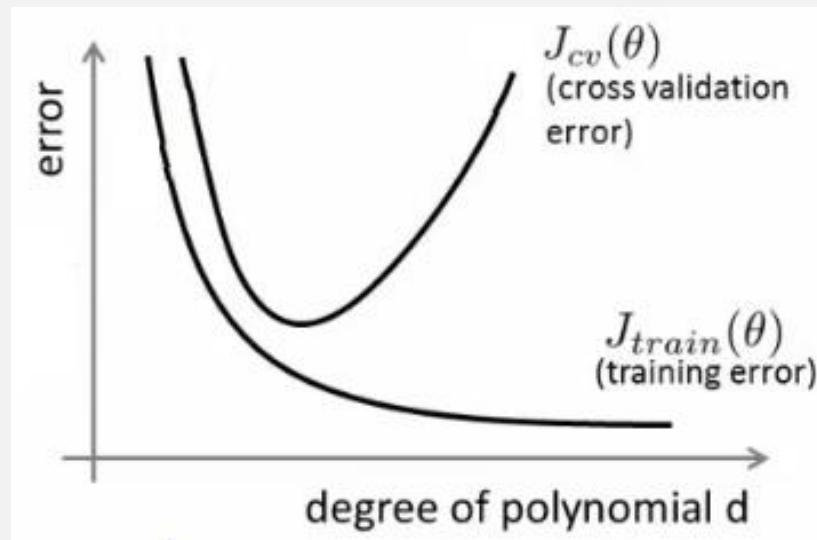
Train – Test – Validation



Cross Validation



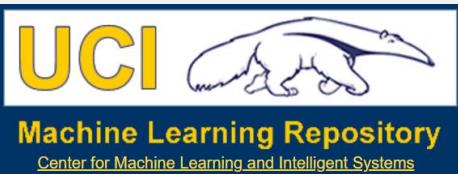
Overfitting vs Underfitting Ideas



Datasets públicos para aprender y probar los algoritmos

The Best Public Datasets
for Machine Learning

Datasets



<https://archive.ics.uci.edu/ml/datasets.php>



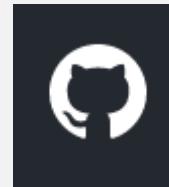
https://scikit-learn.org/stable/datasets/toy_dataset.html

Our World
in Data

<https://ourworldindata.org/>

<https://www.kaggle.com/datasets>

Datasets

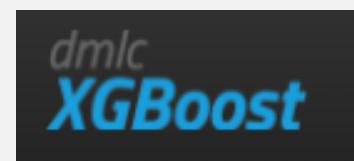


<https://github.com/awesomedata/awesome-public-datasets>

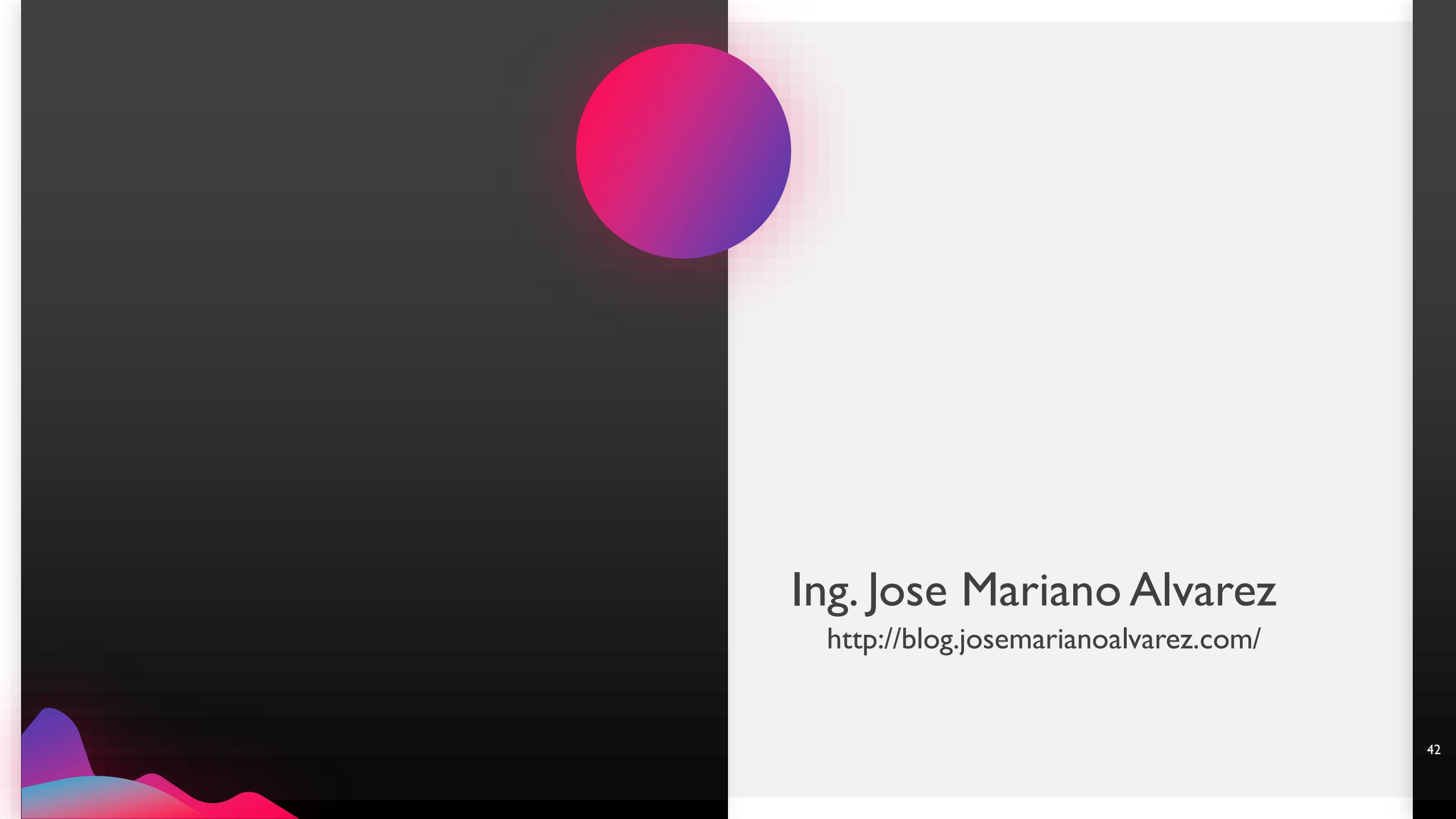
Algunas librerías



CatBoost



Demos



Ing. Jose Mariano Alvarez
<http://blog.josemarianoalvarez.com/>