Ciencia y analítica de datos

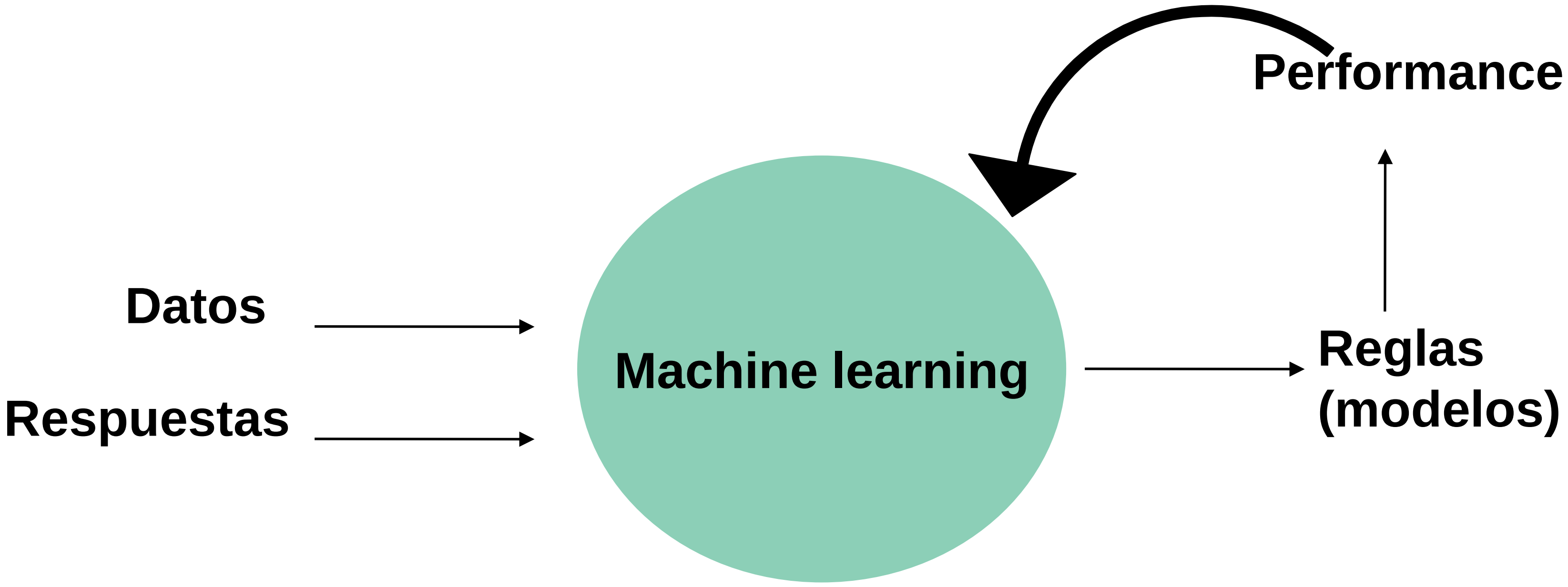
Dra. María de la Paz Rico Fernández.

Tecnológico de Monterrey

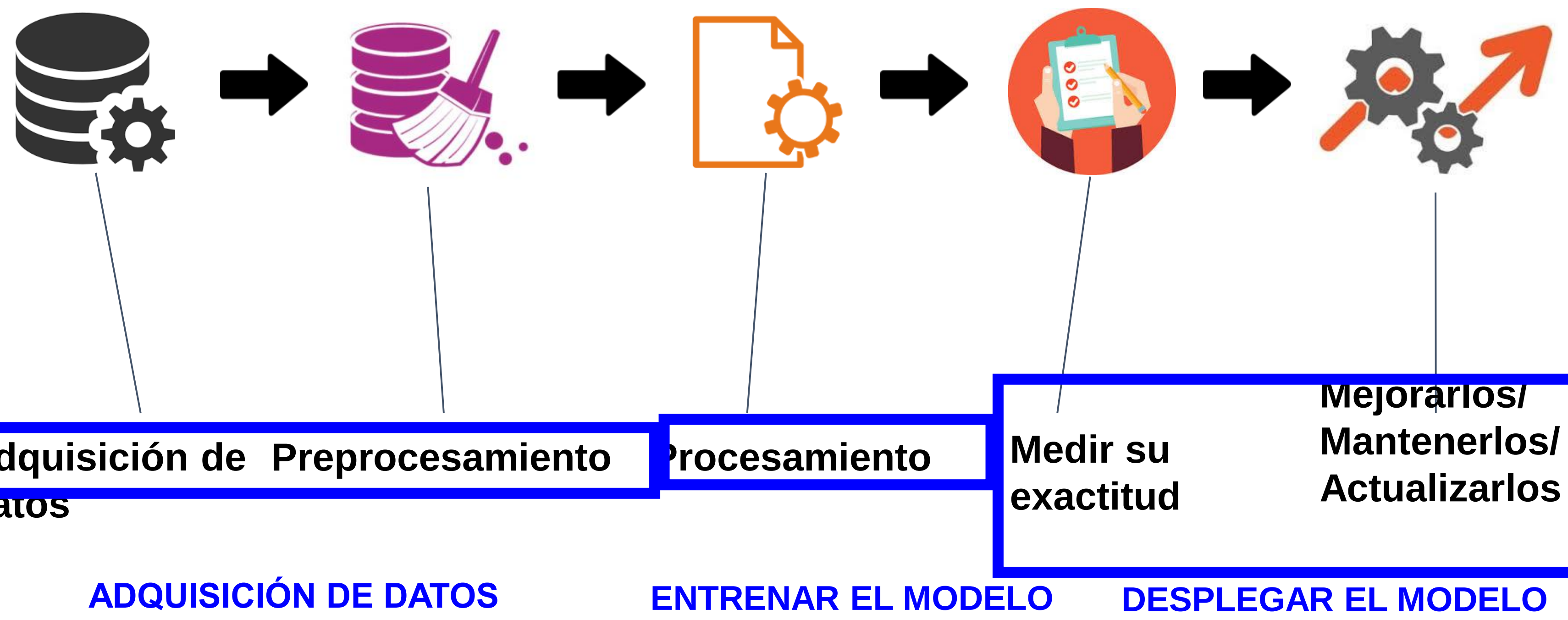
Presentación

- PhD. María de la Paz Rico Fdz.
- Doctorado en ciencias en Robótica y Manufactura avanzada con especialidad en visión por computadora e inteligencia artificial, Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional.
- AI Data Engineer por ANCUD IT, Berlin, Alemania.
- Instructora Certificada de NVIDIA.
- Trabajos:
  - Computer Vision Research Engineer en AIFI Inc, Sillicon Valley.
  - Chief Knoledge Officer en Centro de Innovación Industrial en Inteligencia Artificial.
  - City lead Monterrey Women in AI

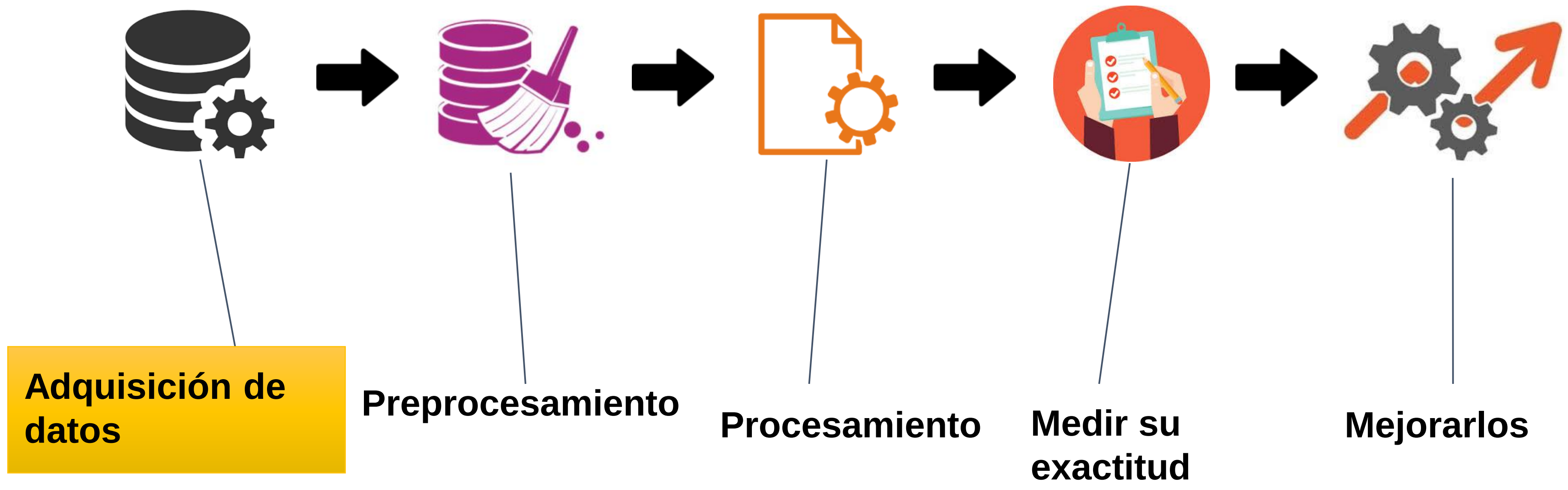
- ¿Qué necesitamos para ml?

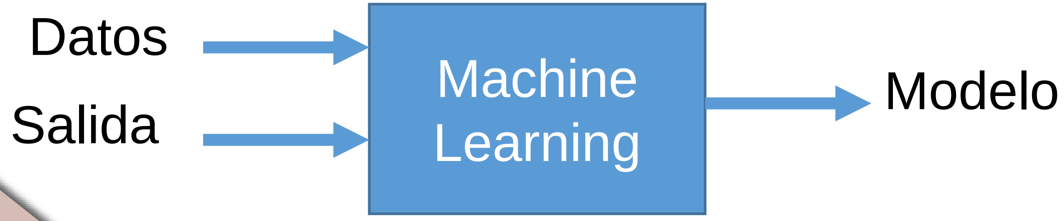


- ml pipeline


Adquisición de datos
Preprocesamiento
Procesamiento
Medir su exactitud
Mejorarlos/ Mantenerlos/ Actualizarlos
ADQUISICIÓN DE DATOS
ENTRENAR EL MODELO
DESPLEGAR EL MODELO

Entender el problema, identificar fuentes de datos (etiquetados) y resaltar posibles problemas con los datos.





P A S O S   C L A V E   P A R A   M L   P R O J E C T
(M L   P I P E L I N E)

- Adquisición de datos

Ejemplos de bases de datos

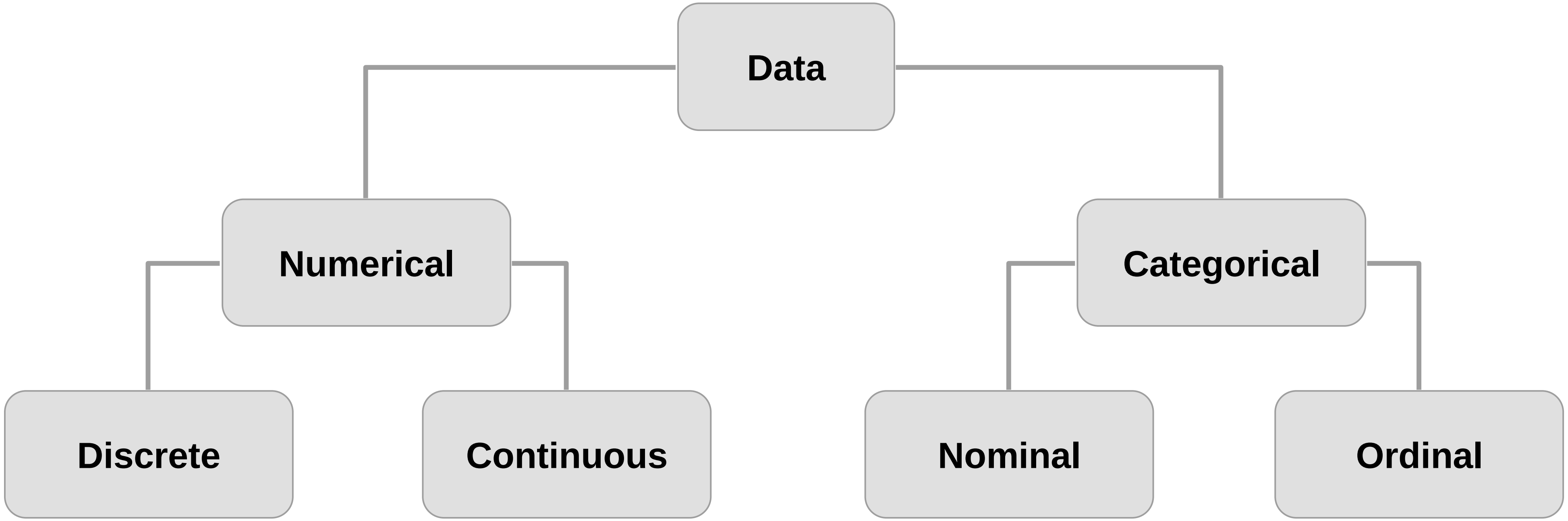
Home prices

size of house (square feet)	# of bedrooms	# of bathrooms	newly renovated	price (1000$)
523	1	2	N	115
645	1	3	N	150
708	2	1	N	210
1034	3	3	Y	280
2290	4	4	N	355
2545	4	5	Y	440

image	label
	cat
	not cat
	cat
	not cat

Machine Learning

TIPOS DE DATOS



Measurements are integers:
- age
- n° of students

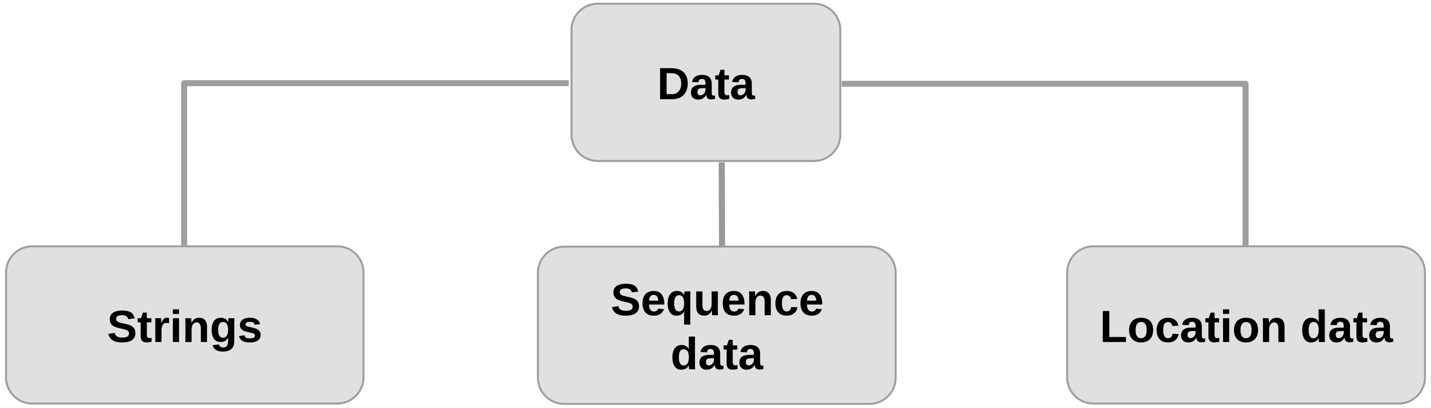
Measurements can take on any value, usually within a range:
- temperature
- weight

No natural order between categories:
- gender
- states / districts
- color names

An order between categories:
- T-shirt size (S, M, L)
- grades (A, B, C)
- time of the day (morning, afternoon, evening)

Machine Learning

Tipos de datos



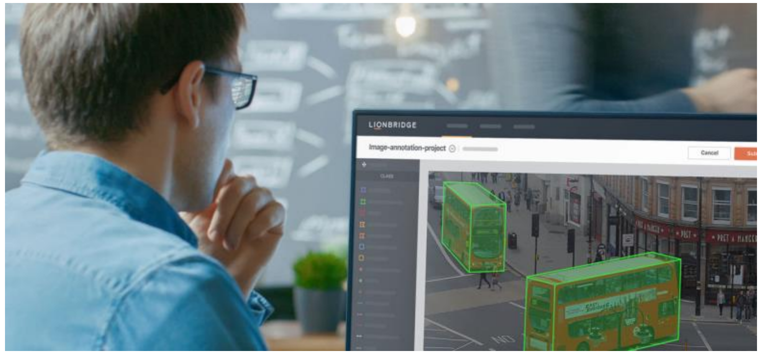
- time series (time order)
- sequences of strings (text data)

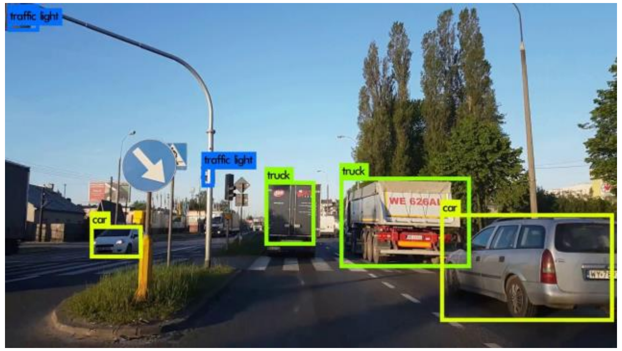
PASOS CLAVE PARA ML PROJECT (ML PIPELINE)

- Adquisición de datos

Podemos adquirir las bases de datos por:

1) Etiquetado manual:





2) De datos observados

user ID	time	price ($)	purchased
4783	Jan 21 08:15.20	7.95	yes
3893	March 3 11:30.15	10.00	yes
8384	June 11 14:15.05	9.50	no
0931	Aug 2 20:30.55	12.90	yes

machine	temperature (◦C)	pressure (psi)	machine fault
17987	60	7.65	N
34672	100	25.50	N
08542	140	75.50	Y
98536	165	125.00	Y

3) Descargandola de paginas web o partners.

PASOS CLAVE PARA ML PROJECT (ML PIPELINE)

- Adquisición de datos y sus problemas

1) Si entran datos ruidosos, estimaciones ruidosas saldr

2) Problemas en los datos
- Etiquetas incorrectas
- Datos faltantes

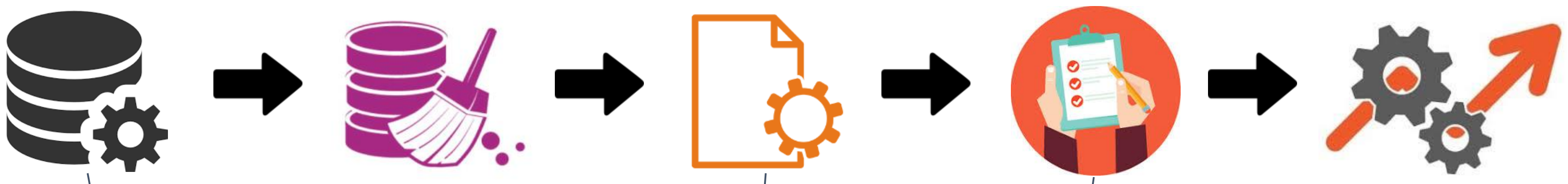

given: 5
corrected: 3

given: cat
corrected: frog

given: lobster
corrected: crab

3) Varios tipos
- Estructurados: Tablas de datos.

- No estructurados: imágenes, audio, video, texto



Adquisición de datos

Preprocesamiento

Procesamiento

Medir su exactitud

Mejorarlos

Limpiar los datos, transformarlos, reducirlos y separarlos

Entrenamiento

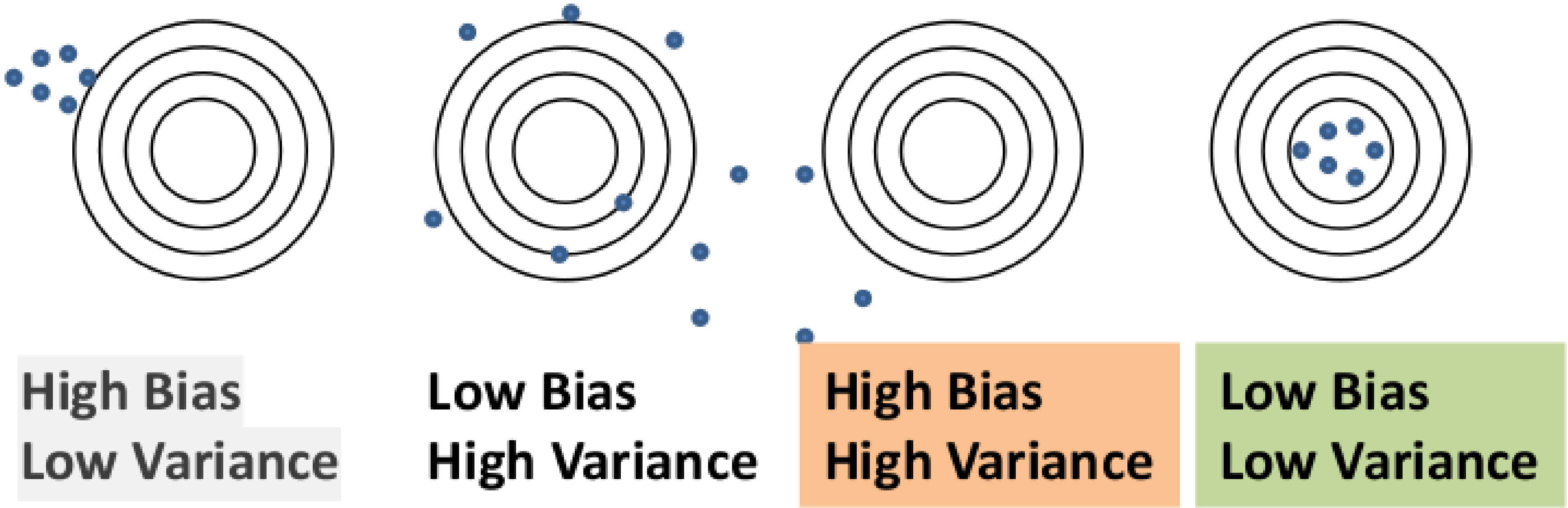
Evaluación

Machine Learning

Challenges of Machine Learning

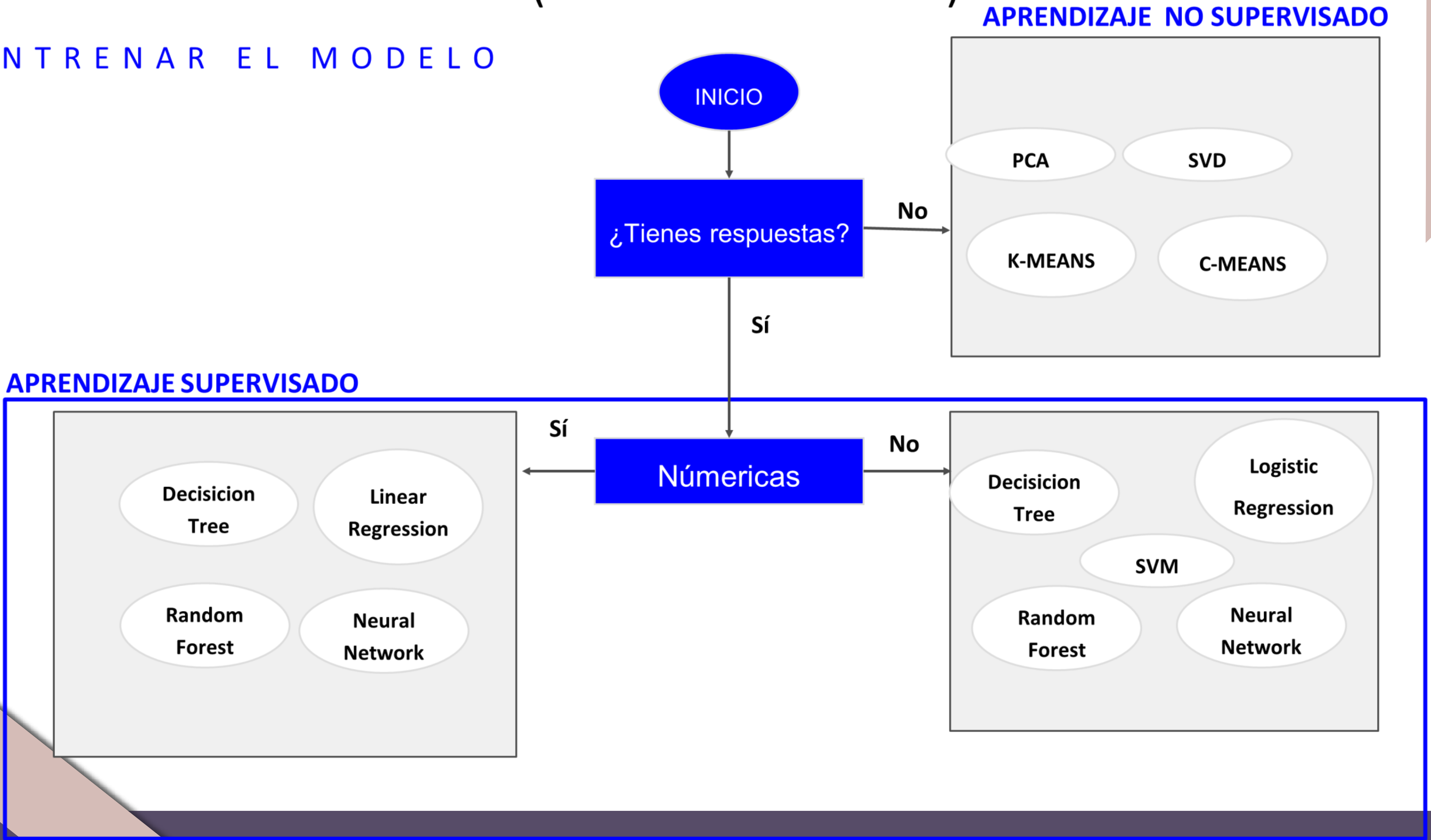
High-bias models: consistent but wrong predictions, prone to underfitting.

High-variance models: prone to overfitting.

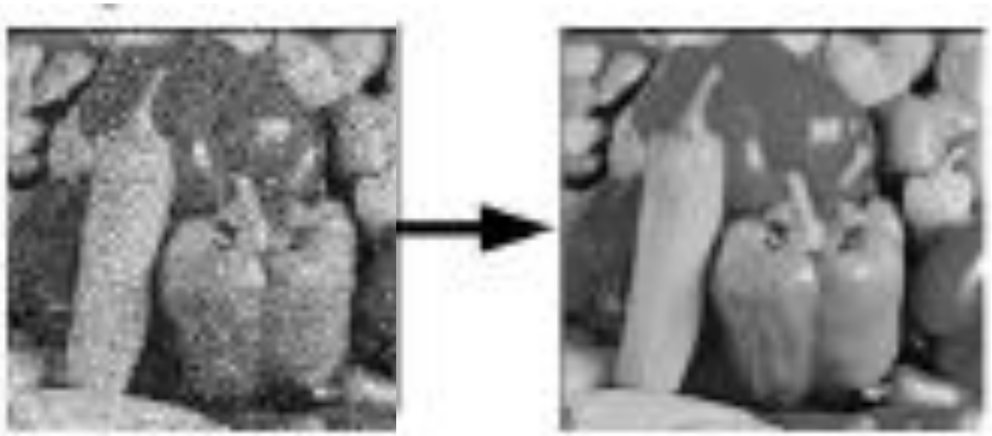


PASOS CLAVE PARA ML PROJECT (ML PIPELINE)

ENTRENAR EL MODELO

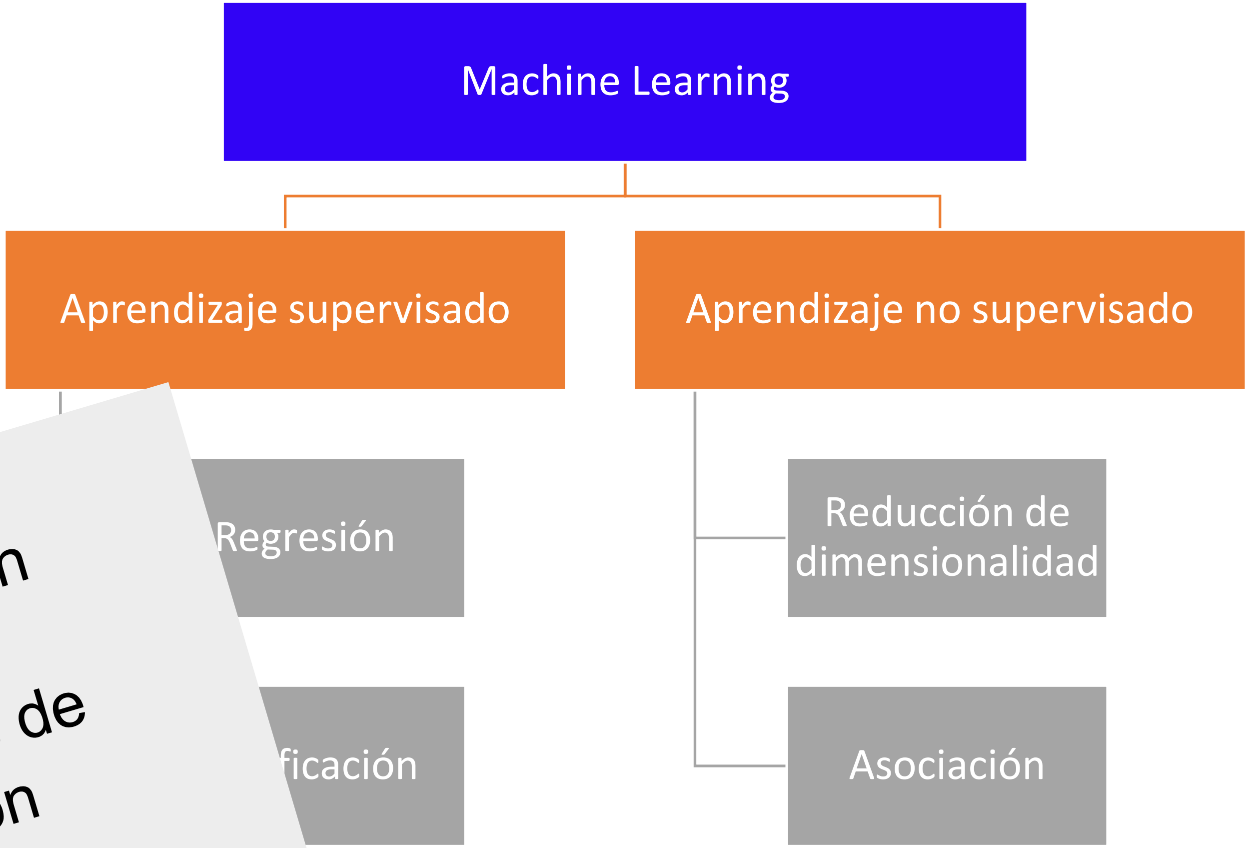


Machine Learning



Dog

From ‘Python Data Science Handbook’ by Jake Vander Plas.



Métodos:
- Regresión lineal
- Arboles de decisión
- SVM
- Neural network

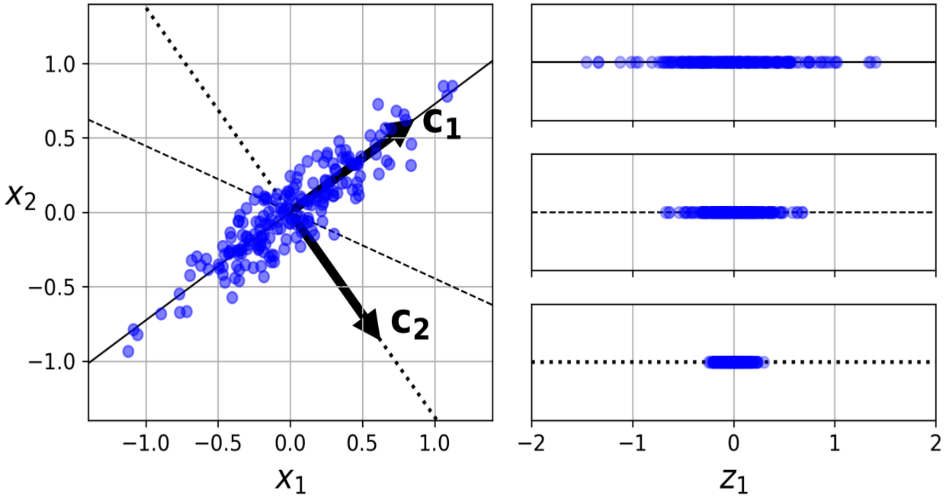
Métodos:
- K-means
- PCA

Machine Learning

Unsupervised Learning

Principal Component Analysis (PCA)

- **Dimensionality reduction**: when there are many features (e.g. thousands or millions) for each training instance it makes training slow and it could be hard to find a good solution.



- **PCA**:

  - identifies the hyperplane that lies closest to the data

  - projects the data onto the hyperplane

  - selects the projection that preserves the maximum amount of variance

Dimensionality Reduction

- When there are many features (e.g. thousands or millions) for each training instance it makes training slow and it could be hard to find a good solution.
- Reducing dimensionality of the training set before training a model speeds up training.
- Reducing dimensionality does reduce information.
- It is useful for data visualization.

The Curse of Dimensionality

- High-dimensional datasets are risk of being very **sparse**: most training instances are far away from each other making predictions less reliable since they will be based on larger extrapolations.

- **Sparsity** is a problem for statistical significance, the amount of data needed to support the result often grows exponentially with the dimensionality.

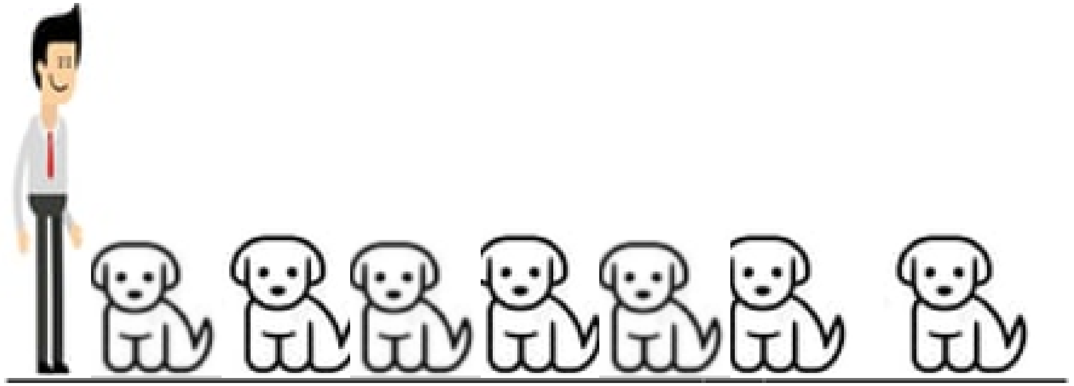


Figure 1 - One-dimension scenario

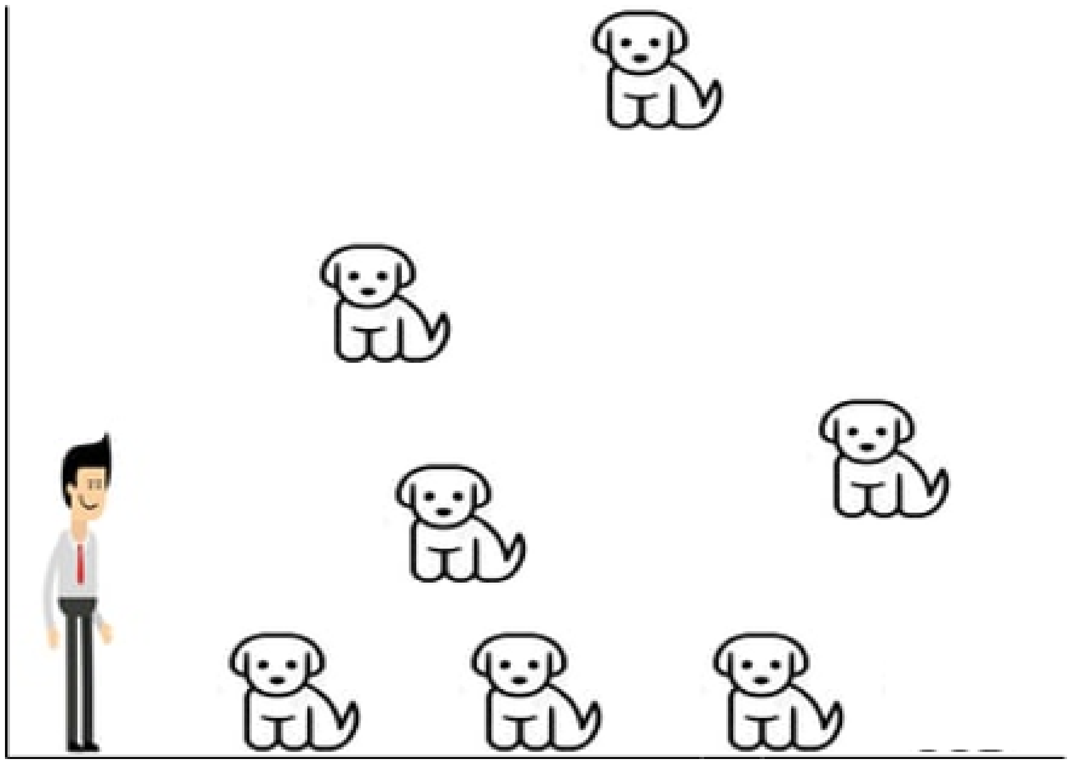


Figure 2- Two-dimension Scenario

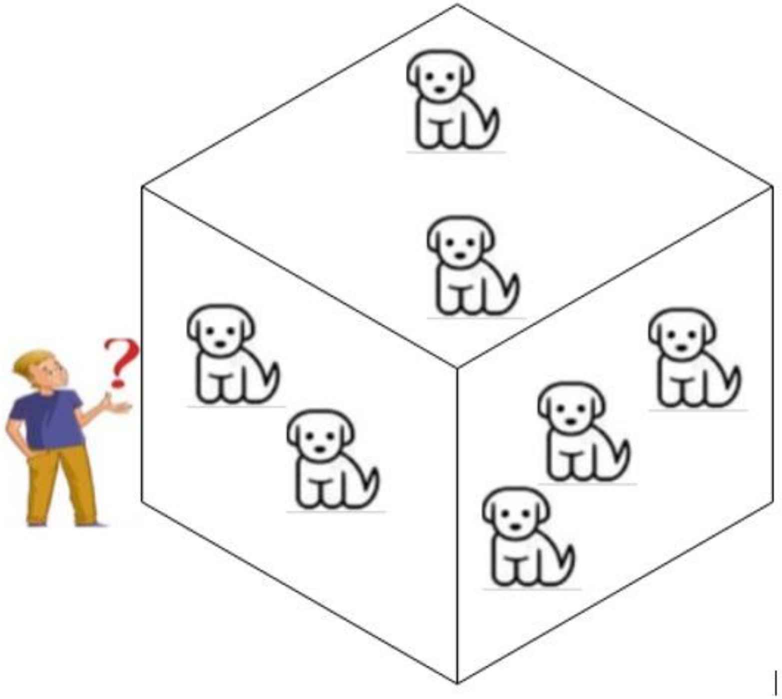
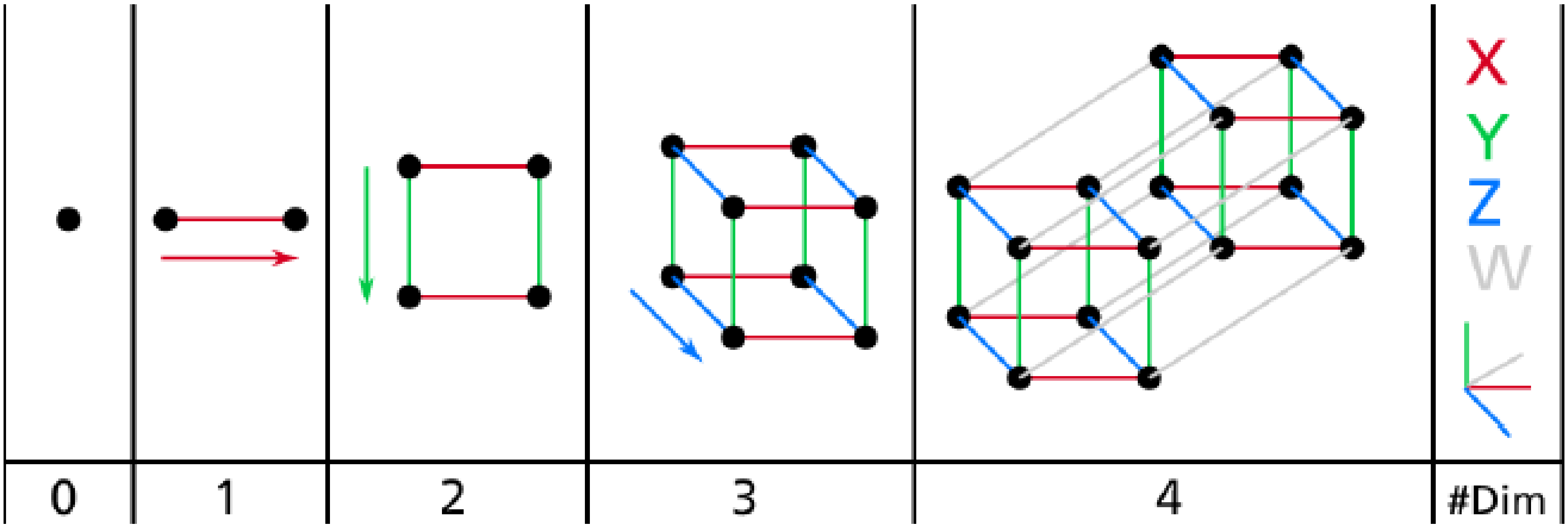


Figure 3 – Three-dimension scenario


X
Y
Z
W
0
1
2
3
4
#Dim

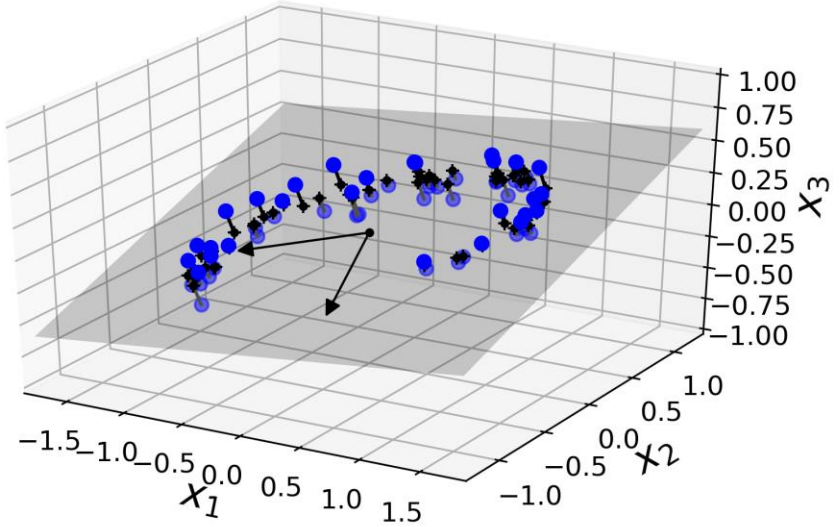
Approaches for Dimensionality Reduction

1- Projection

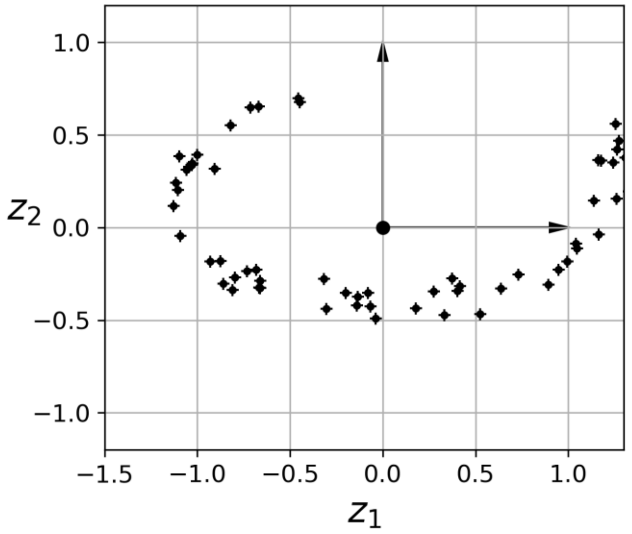
- In real world-problems, training instances are not spread out uniformly across all dimensions.

- Then, training instances lie within a lower-dimensional **subspace** of the high-dimensional space.

Before projection

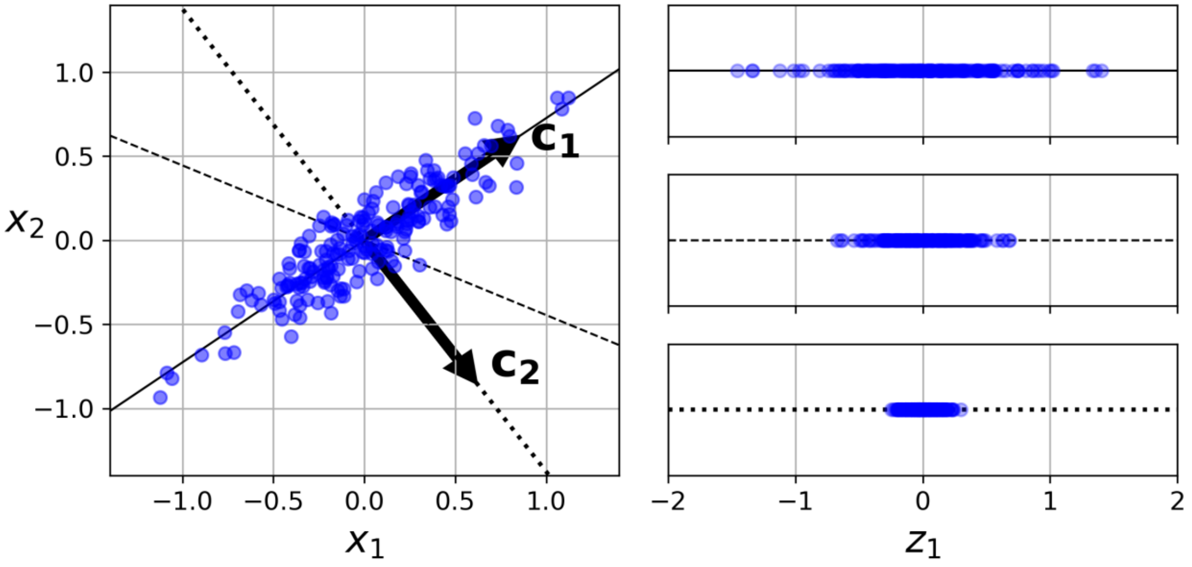


After projection



Principal Component Analysis(PCA)

- Unsupervised method for dimensionality reduction of the data
- Identifies the hyperplane that lies closest to the data.
- Then, it projects the data onto it.
- Selects the projection that preserves the maximum amount of variance (the axis that minimizes the mean squared distance between the original dataset and its projection onto that axis).



$$X_{dproj} = X.W_d$$

X: matrix training set

W_d: matrix containing the first *d* principal components

This projects the training set onto the space defined by the principal components.

From 'Hands-On Machine Learning with Scikit-Learn and TensorFlow' by Aurélien Géron.

Principal Component Analysis(PCA)

Principal Components

- PCA identifies the **axis** that accounts for the largest amount of **variance** in the training set.
- PCA finds a second axis, **orthogonal** to the first one, that accounts for the largest amount of remaining variance.
- PCA would also find a third axis, orthogonal to the both previous axis, etc.
- The unit vector that identifies the i^{th} axis is called **principal component**.
- PCA assumes that the dataset is centered around the origin. Scikit-Learn's PCA classes centers the data.

Principal Component Analysis(PCA)

Scikit-Learn

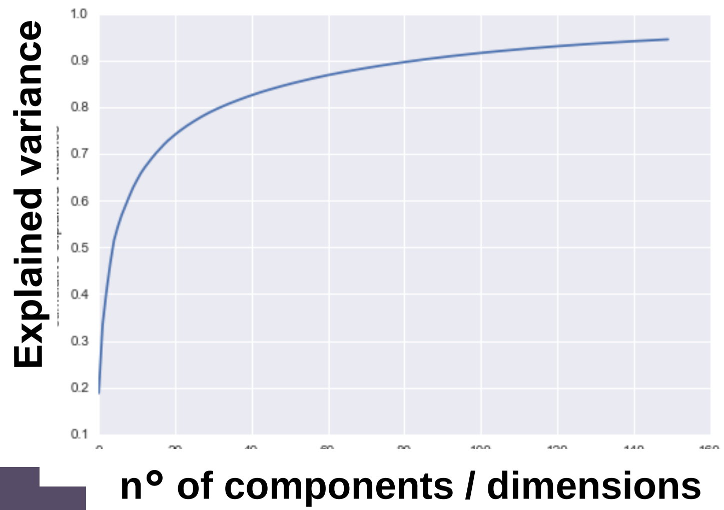
- After fitting PCA transformer to the dataset, the principal components can be accessed using ***components_*** variable (the first one: pca.components_T[:, 0]).

- Explained variance ratio of each principal component: the proportion of the dataset’s variance that lies along the axis of each principal component, ***explained_variance_ratio_*** variable.

PCA (Principal Component Analysis)

Choosing the Right Number of Dimensions

- It is generally useful to choose the number of dimensions that add up to a large proportion of the variance (i.e. 95%).
- In case of data visualization the dimensionality is usually reduced to 2 or 3 dimensions.
- **Scikit_Learn**: set n_components to a float between 0.0 and 1.0, indicating the ratio of variance to preserve, ***PCA*****(n-components=0.95)**.
- Plot the explained variance as a function of the number of dimensions.



From 'Python Data Science Handbook' by Jake Vander Plas.

Principal Component Analysis(PCA)

PCA for Compression

- After dimensionality reduction, the training set takes up less space.
- Speeds up an algorithm like SVM.
- It’s possible to decompress the reduced dataset by applying the inverse transformation of the PCA.

PCA (Principal Component Analysis)

Disadvantages

- PCA tends to be highly affected by **outliers** in the data.

- PCA assumes that the principle components are a **linear** combination of the original features.

- PCA assumes that the principle components are **orthogonal.**

- PCA uses **variance** as the measure of how important a particular dimension is.

- High variance axes are treated as principle components.

- Low variance axes are treated as noise.