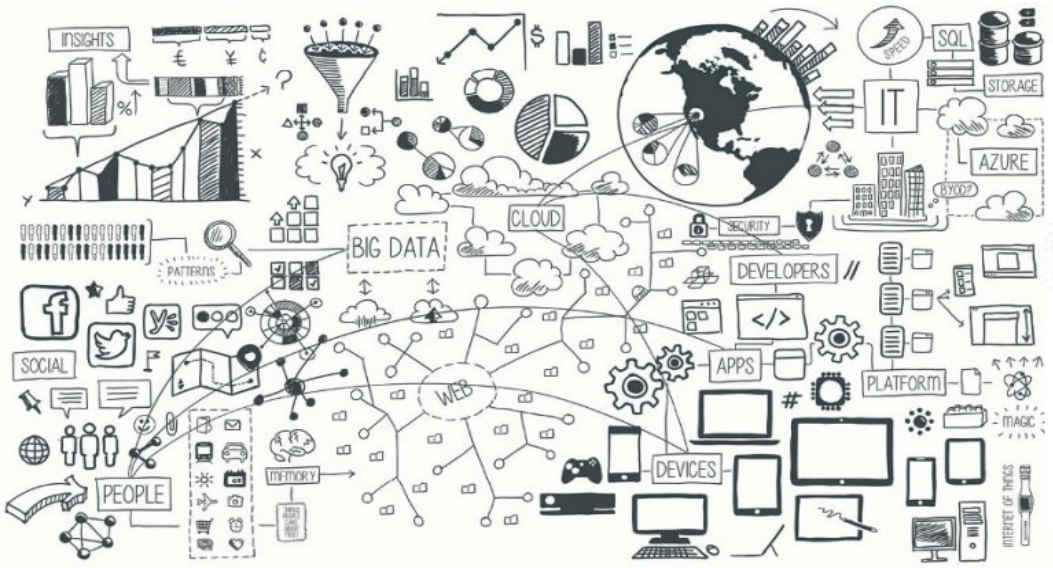
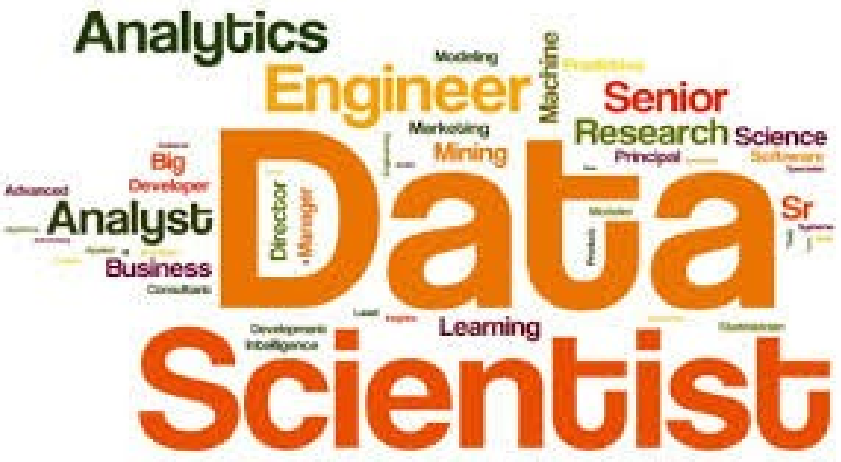
Data Mining (Minería de Datos)

CANONICAL PROBLEMS & LEARNING PARADIGMS



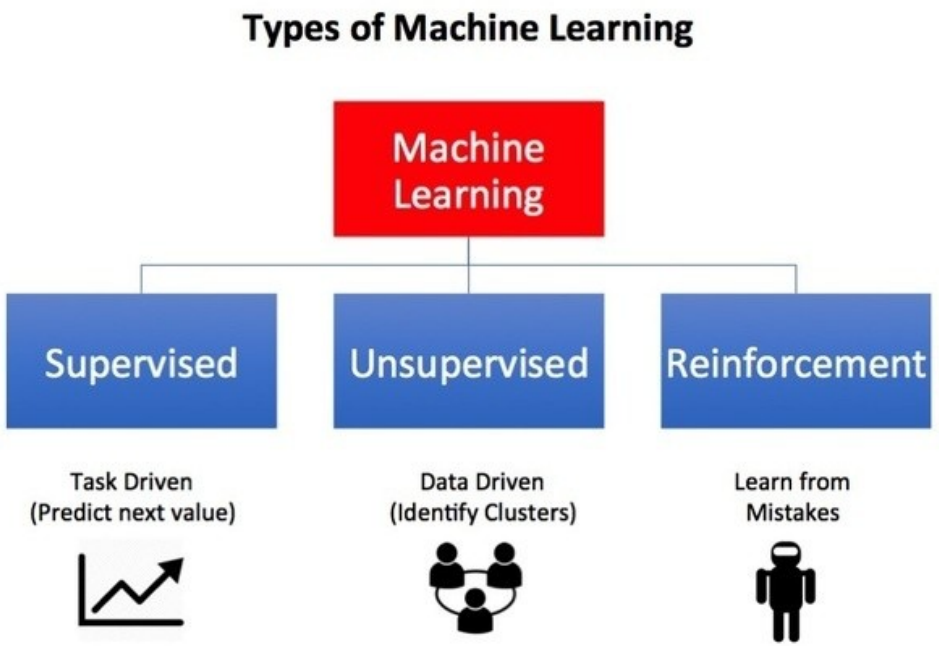


José Manuel Gutiérrez

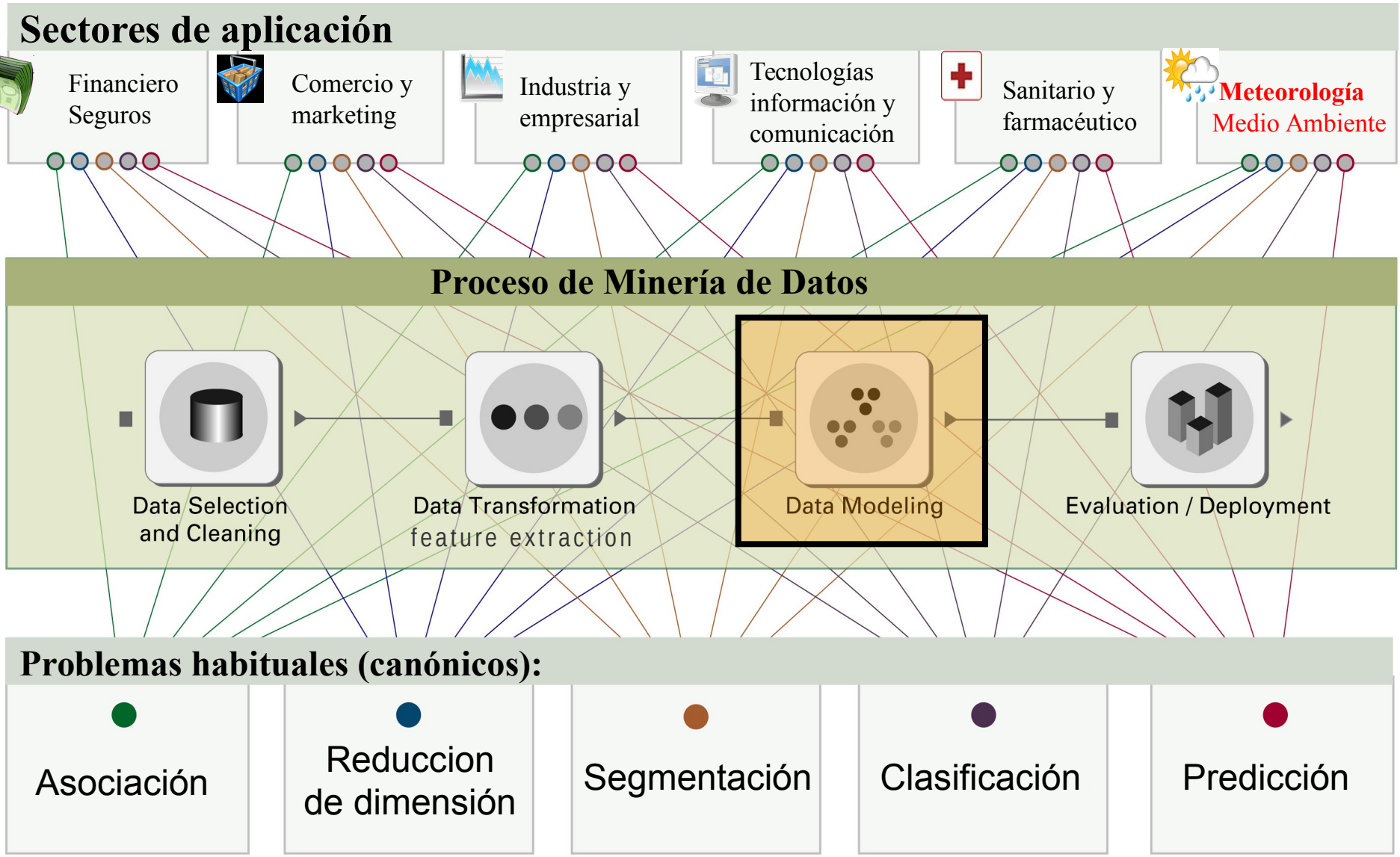
Grupo de Meteorología

Univ. de Cantabria – CSIC
MACC / IFCA

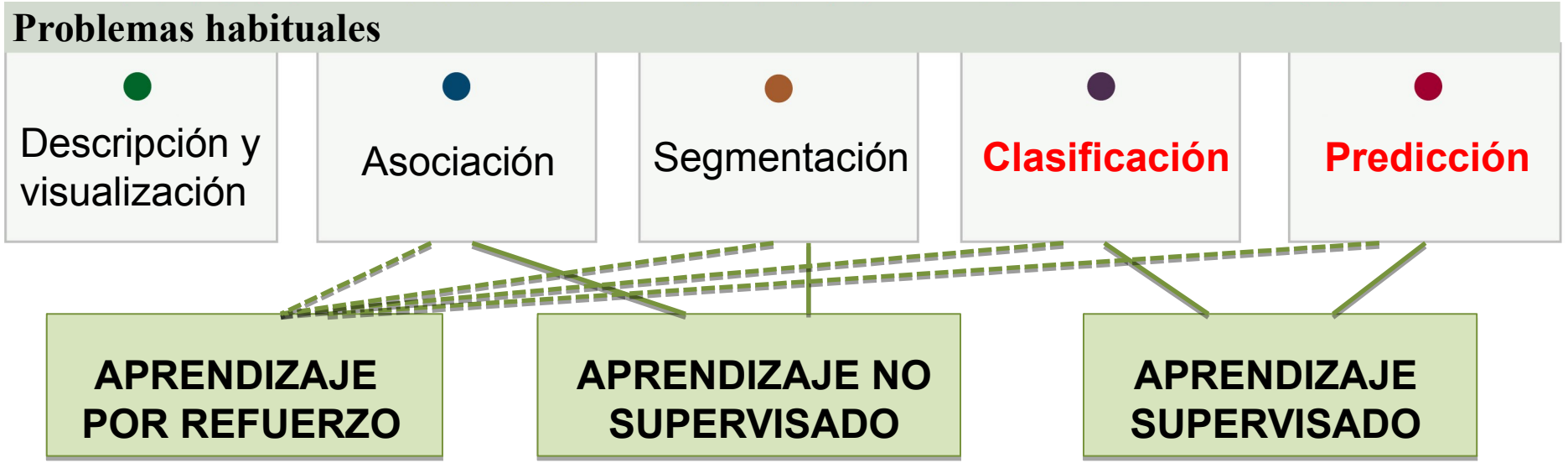
CANONICAL PROBLEMS & LEARNING PARADIGMS



Oct	30	Aplazada (sesión de refuezo)
Nov	6	Presentación, introducción y perspectiva histórica
	8	**Paradigmas, problemas canonicos y data challenges**
	13	Reglas de asociación
	15	Practica: Reglas de asociación
	20	Evaluación, sobrejuste y crossvalidacion
	22	Practica: Crossvalidacion
	27	Arboles de clasificacion y decision
	29	Practica: Arboles de clasificación
		T01. Datos discretos
Dic	4	Técnicas de vecinos cercano (k-NN)
	11	Práctica: Vecinos cercanos
	13	Reducción de dimensión lineal
	18	Practica: LDA y PCA
	20	Reducción no lineal
		T02. Clasificación
Ene	8	Arboles de clasificación y regresion (CART)
	10	Practica: CART
	15	Ensembles: Bagging and Boosting
	17	Practica Random Forests
		T03. Prediccion
	22	Practica Gradient boosting
	24a	Técnicas de agrupamiento
	24b	Practica: Técnicas de agrupamiento
	29a	Practica: El paquete CARET
	29b	Examen



Machine learning develop methods for data modelling and prognosis.



Datos de entrada *(X)*: $(X_1, X_2, \ldots, X_n)$

Aprendizaje supervisado: Se entrena con datos *(X)* que han sido etiquetados ("label") $(y_1, y_2, \ldots)$. Las etiquetas clasifican cada punto de datos en uno o más grupos, como "manzanas" o "naranjas". El sistema aprende cómo se estructuran estos datos, se entrena de manera que minimiza el error de predicción del sistema. El objetivo es predecir las categorías de datos nuevos o de "test".

Aprendizaje NO supervisado: Se trata de agrupar e interpretar los datos sólo con los datos de entrada *(X)*.

Aprendizaje por refuerzo: Se encuentra entre el aprendizaje supervisado y no supervisado. Se centra en ir aprendiendo de la experiencia. Recive recompensas o castigos $(r_1, r_2, \ldots)$ de las acciones $(a_1, a_2, \ldots)$ que realiza. El objetivo es maximizar las recompensas.

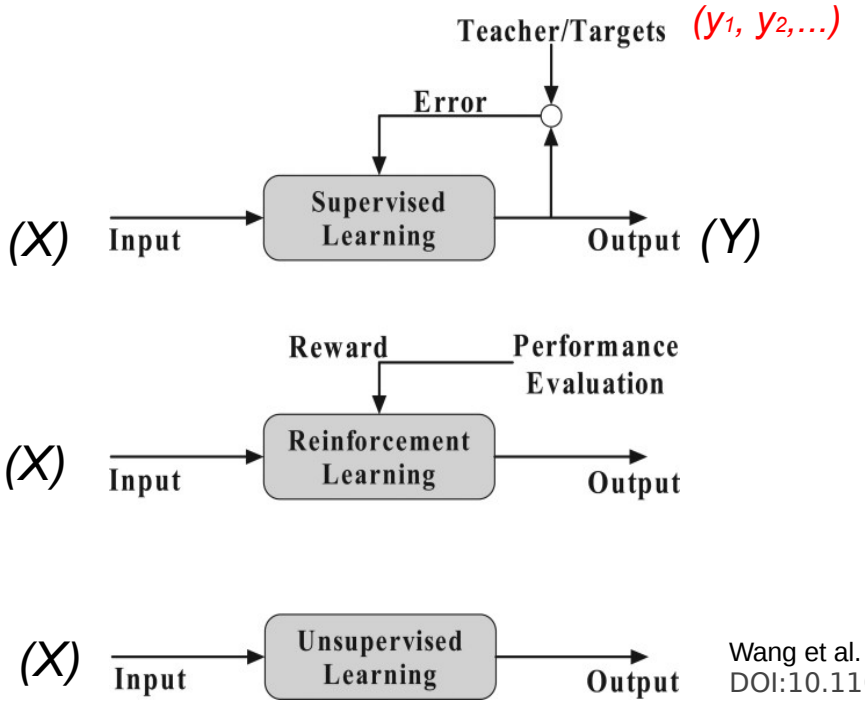
Problemas habituales

- Descripción y visualización
- Asociación
- Segmentación
- **Clasificación**
- **Predicción**

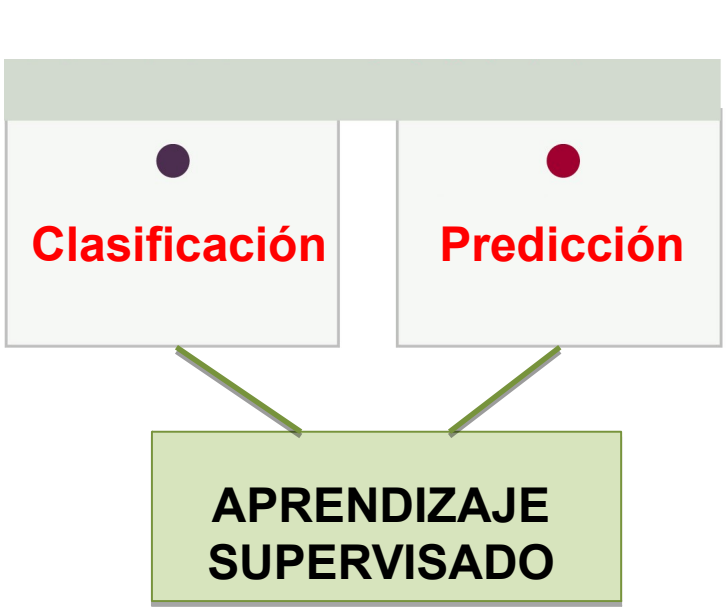
APRENDIZAJE POR REFUERZO

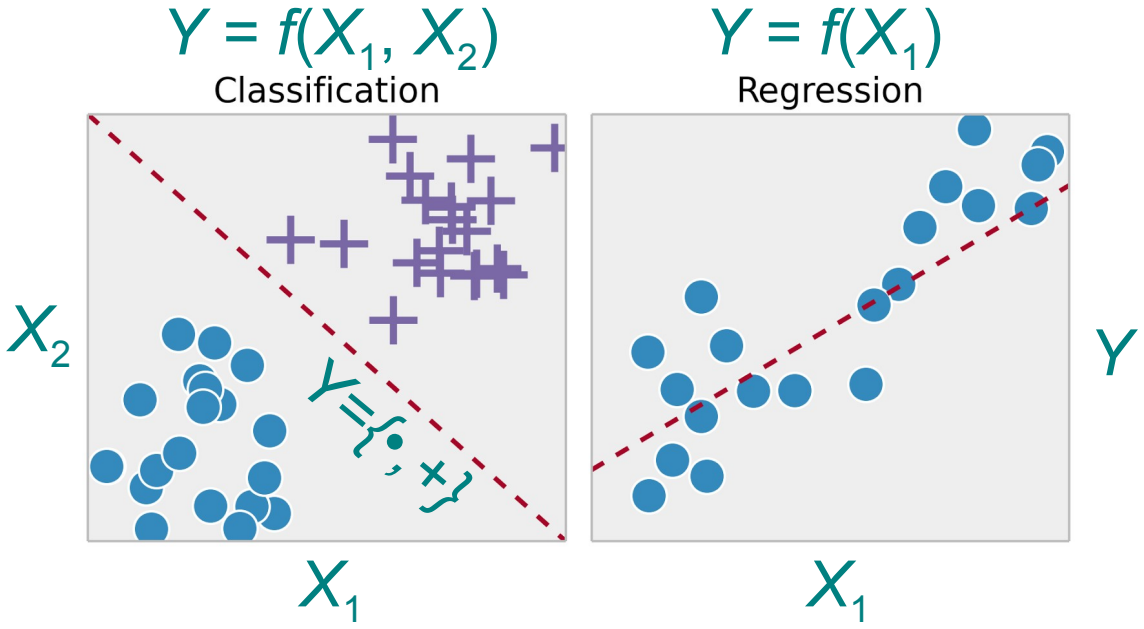
APRENDIZAJE NO SUPERVISADO

APRENDIZAJE SUPERVISADO

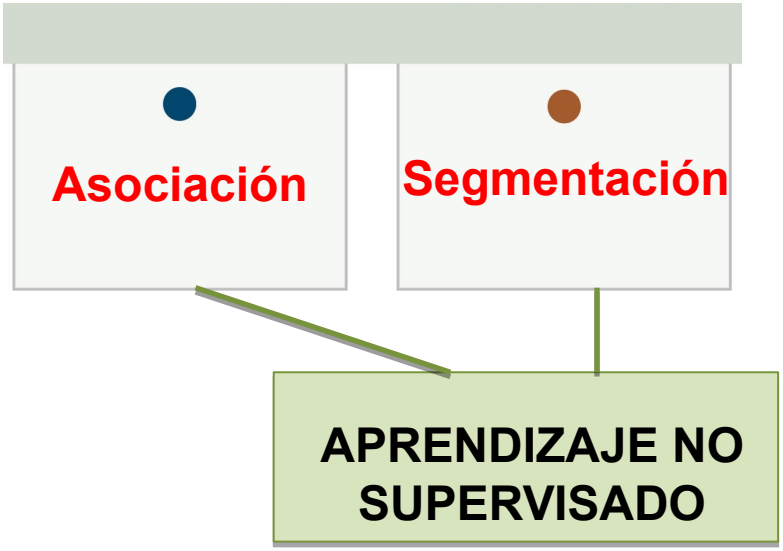


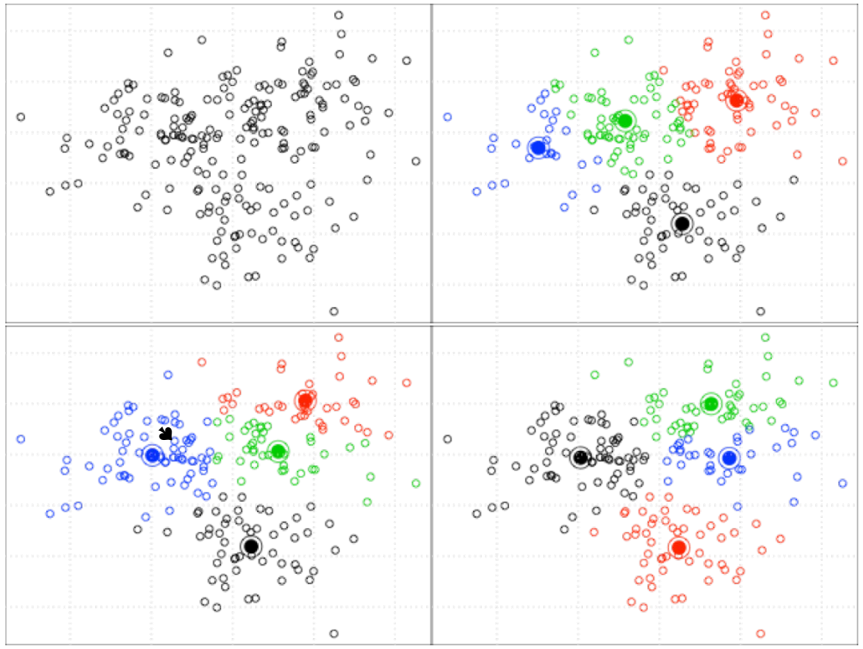
Wang et al. 2012
DOI:10.1109/TSMCC.2012.2186565





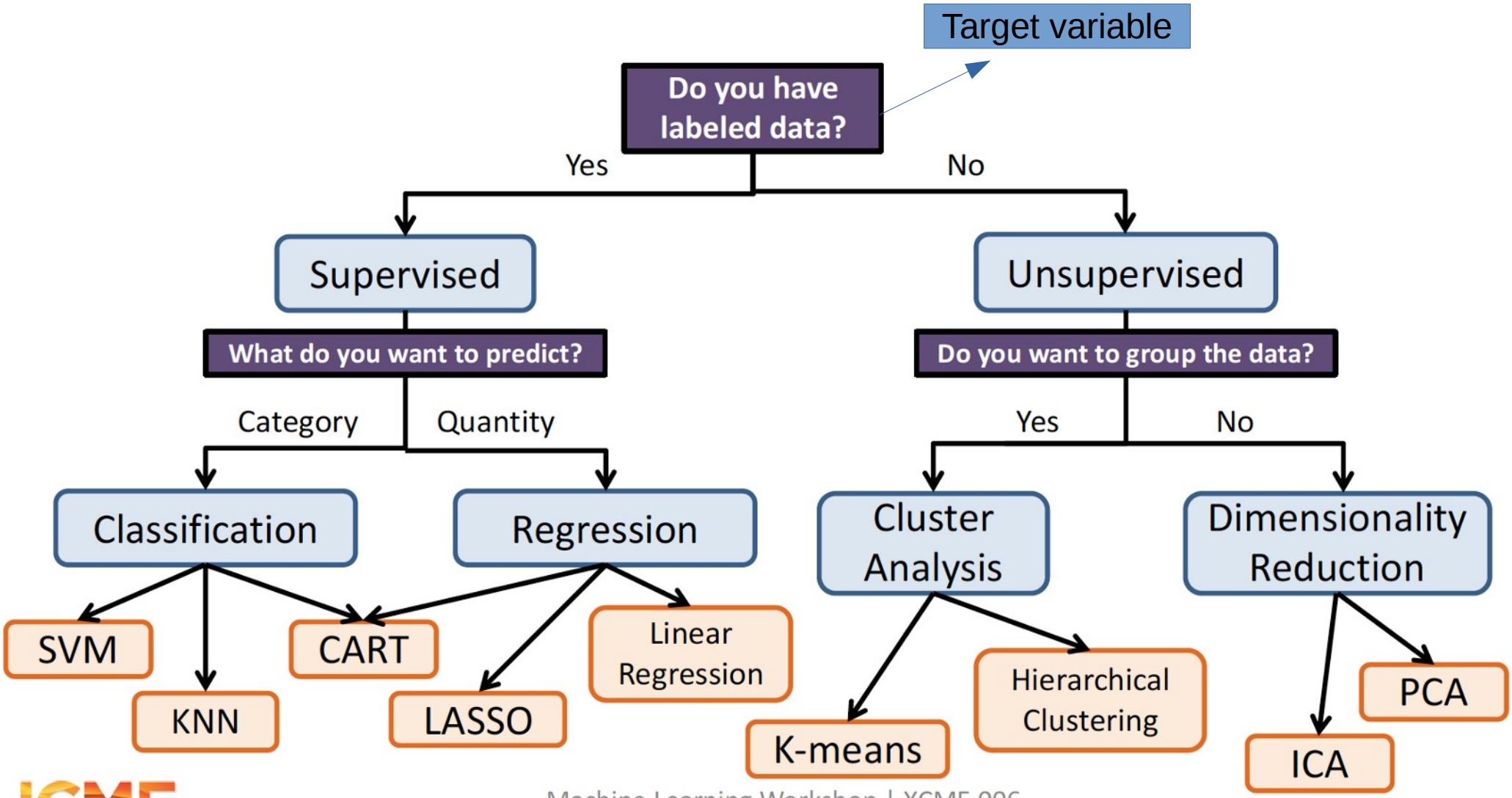
- Target Variable: Y : *categorical/factor* or *continuous*
  - What we are trying to predict.
- Predictive Variables: $\{X_1, X_2, \ldots, X_N\}$: *continuous or factor*
  - "Covariates" used to make predictions.
- Predictive Model: $Y = f(X_1, X_2, \ldots, X_N)$
  - "Learning engine" that estimates the f (or the parameters defining f).



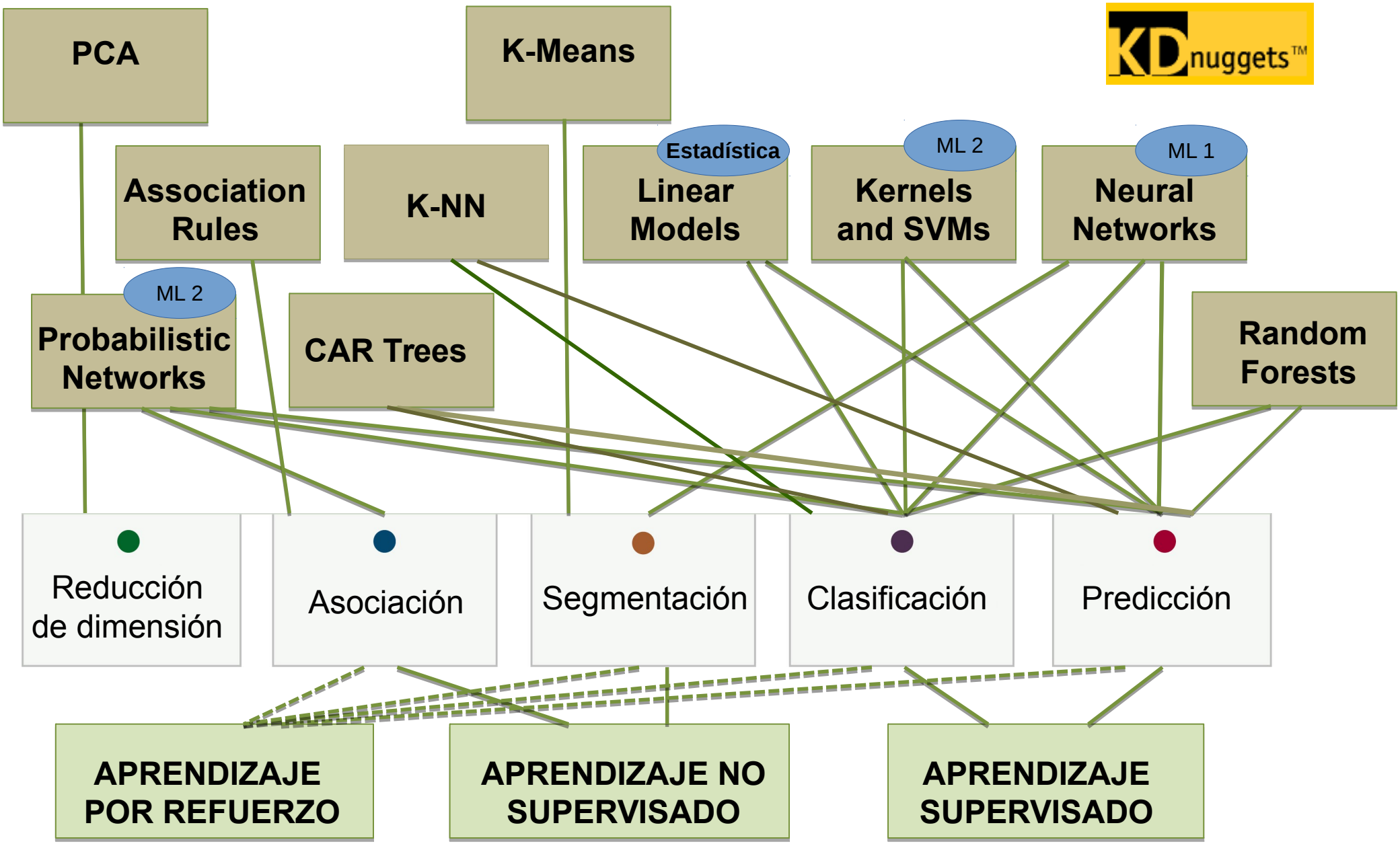


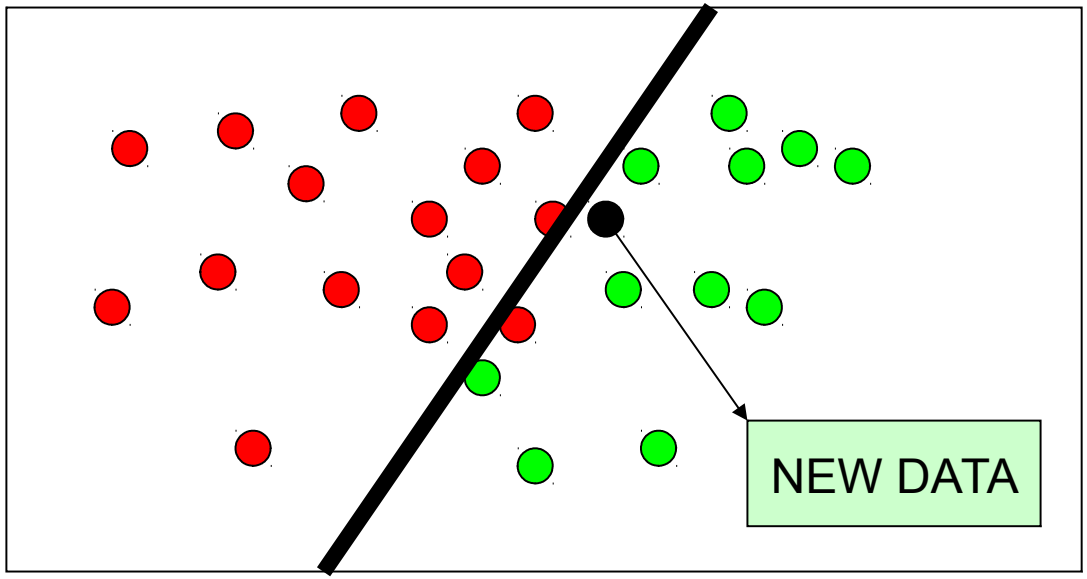
discrete: #clusters

- Target Variable: *There is no target variable*
- Variables: $\{X_1, X_2, \ldots, X_N\}$: *continuous* or *factor*
  - "Covariates" used to make predictions.
- Predictive Model: Algorithmic, based on $(X_1, X_2, \ldots, X_N)$.
  - Ad-hoc "learning" and "prediction" engine.

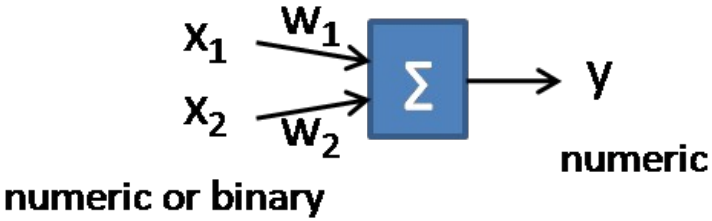

Target variable
Do you have labeled data?
Yes
No
Supervised
Unsupervised
What do you want to predict?
Do you want to group the data?
Category
Quantity
Yes
No
Classification
Regression
Cluster Analysis
Dimensionality Reduction
SVM
KNN
CART
LASSO
Linear Regression
K-means
Hierarchical Clustering
ICA
PCA

https://www.kdnuggets.com/2017/10/top-10-machine-learning-algorithms-beginners.html





GENERATIVE METHODS:
Linear models are the simplest family for machine learning and have good generalization properties.

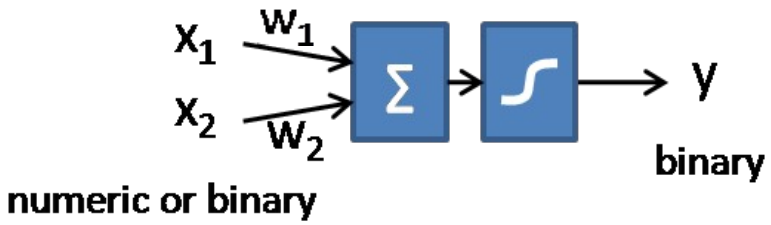


$$y = w_0 + w_1x_1 + w_2x_2$$

$$y = f(\mathbf{X},\mathbf{W}) = \mathbf{X}^T.\mathbf{W}$$

$$\mathbf{W} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$$

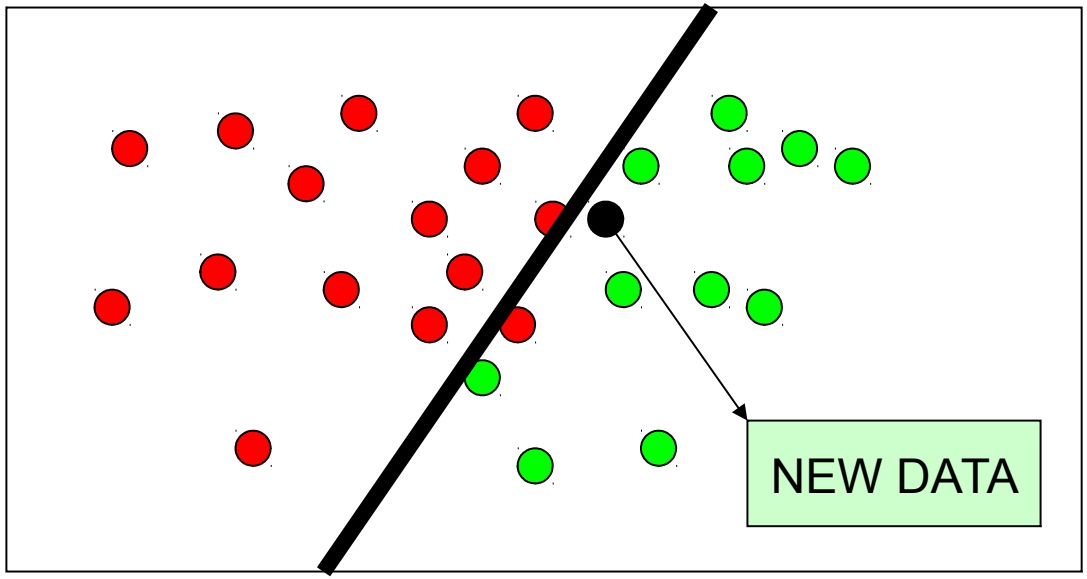
LINEAR REGRESSION



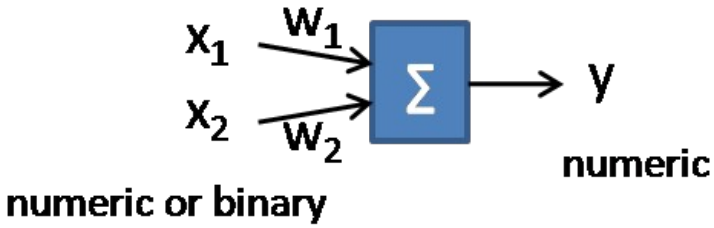
$$y = \text{sigmod}(w_0 + w_1x_1 + w_2x_2)$$
$$\ldots \text{ where sigmod}(k) = 1 / (1 + e^{-k})$$

$$y = f(\mathbf{X},\mathbf{W}) = \text{sigmod}(\mathbf{X}^T.\mathbf{W})$$

LOGISTIC REGRESSION

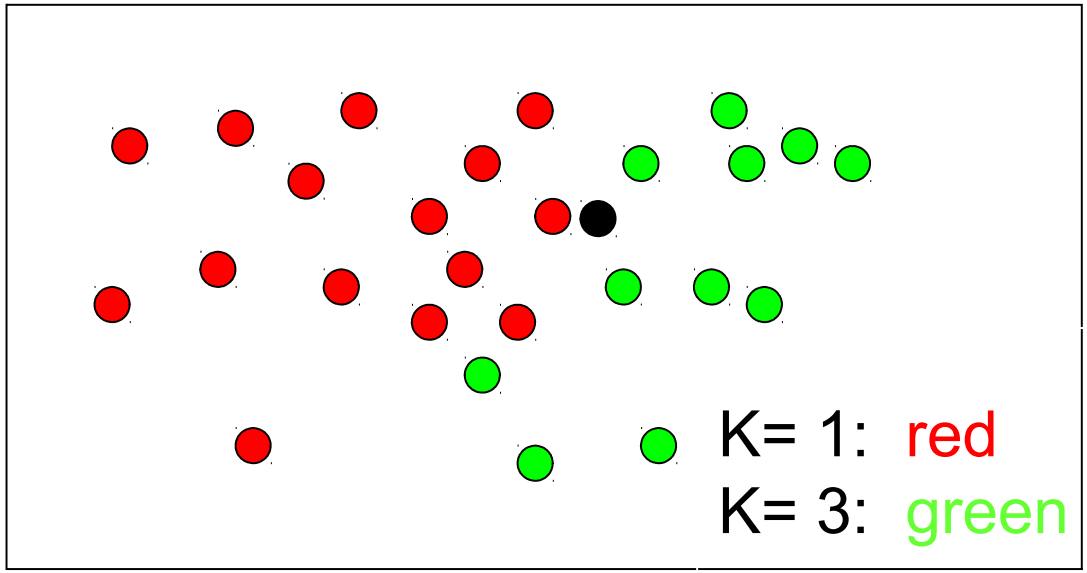


GENERATIVE METHODS:
Linear models are the simplest family for machine learning and have good generalization properties.



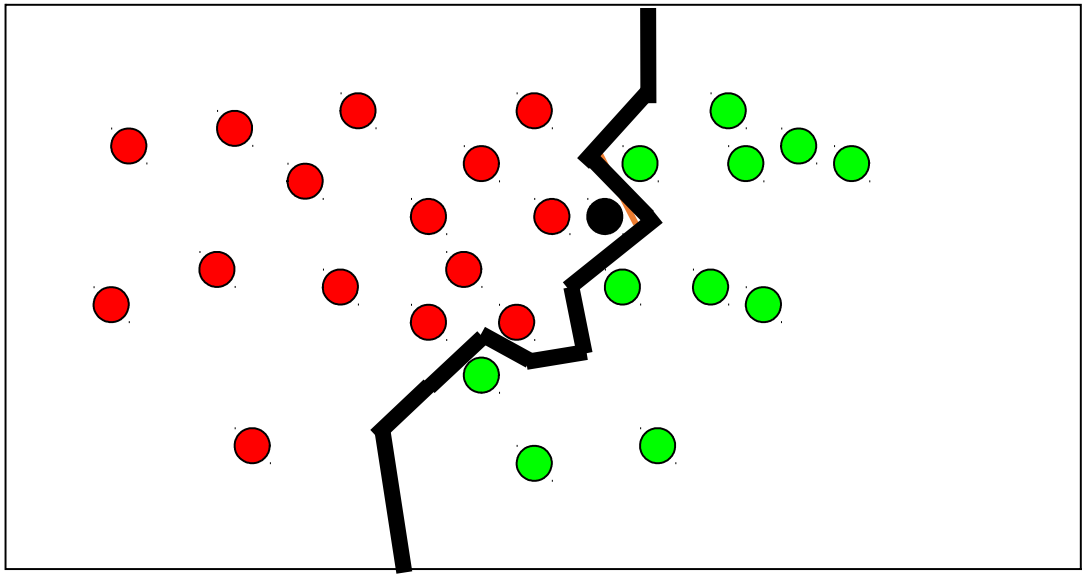
$$y = w_0 + w_1x_1 + w_2x_2$$

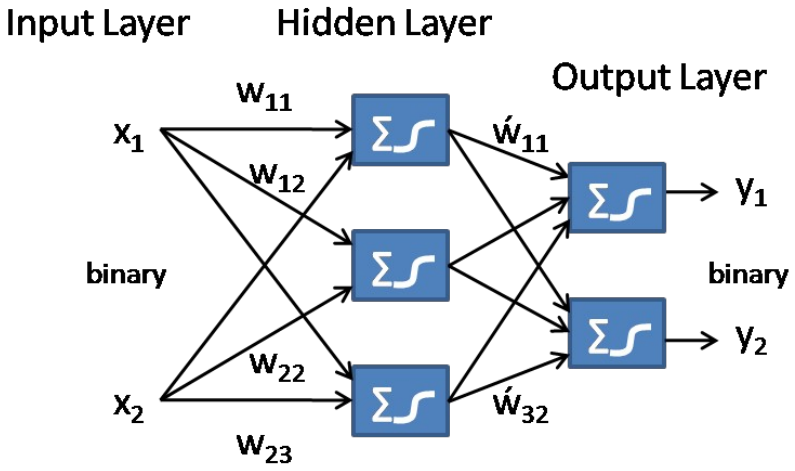
$$y = f(X,W) = X^T.W$$



NON-GENERATIVE (OR ALGORITHMIC)
K Nearest Neighbours is the simples non-generative method. It depends on a single parameter (K) to be tunned (generalization depends on K).

Increasing model complexity (e.g. number of parameters) can result in **overfitting** (lack of generalization).





Cuando se implementan sistemas de aprendizaje automático, existe una distinción clave entre los sistemas de aprendizaje en línea (**online**) y fuera de línea (**offline**):

- **Offline**

Los sistemas de aprendizaje **se entrenan y validan offline** y **se "congelan"** antes de empezar a ser utilizados por los usuarios.

Posteriores entrenamientos del sistema se realizarán de nuevo offline para congelar una actualización que da lugar a una **nueva versión**.

Este proceso es el más común ya que posibilita la **verificación humana del sistema antes de que el sistema interactúe con el usuario.**

- **Online**

Los sistemas de aprendizaje se entrenan y validan offline, pero **continúa entrenándose y validándose, es decir, actualizándose a medida que interactúa con los usuarios .**

El funcionamiento de los algoritmos de aprendizaje pueden mejorar en tiempo real.

No permite la verificación humana antes de que el sistema interactúe con el usuario.

Un ejemplo son los sistemas de detección de spam que se entrenan en respuesta a los patrones del correo entrante y al feedback que da el usuario sobre la precisión del sistema.