

# Ciencia de **datos**



**... aplicada**

**José Ramón Cajide**

DATA SCIENTIST en **El Arte de Medir**

MADRID - 17 Octubre '18

# CIENCIA DE DATOS APLICADA

## ¿Qué es?



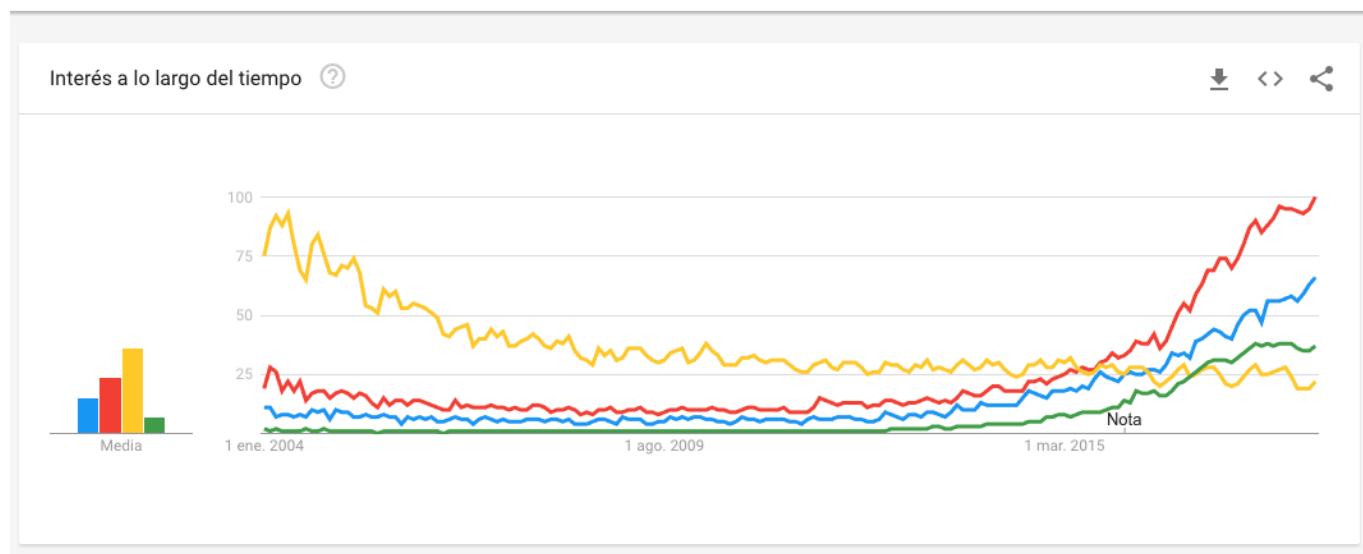
**La Minería de datos es un proceso para explorar grandes cantidades de datos con el fin de descubrir patrones relevantes**



**La Ciencia de datos es un campo de estudio que incluye análisis Big Data, minería de datos, modelos predictivos, visualización, matemáticas, y estadística**

# CIENCIA DE DATOS APLICADA

## Evolución

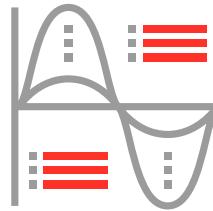


# CIENCIA DE DATOS APLICADA

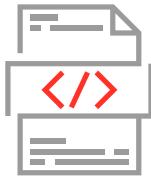
## Requisitos de un proyecto de ciencia de datos



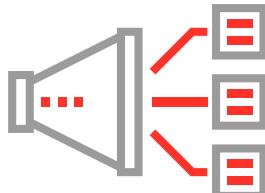
Conocimiento del negocio



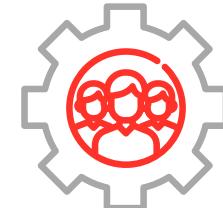
Análisis cuantitativo



Programación



Comunicación



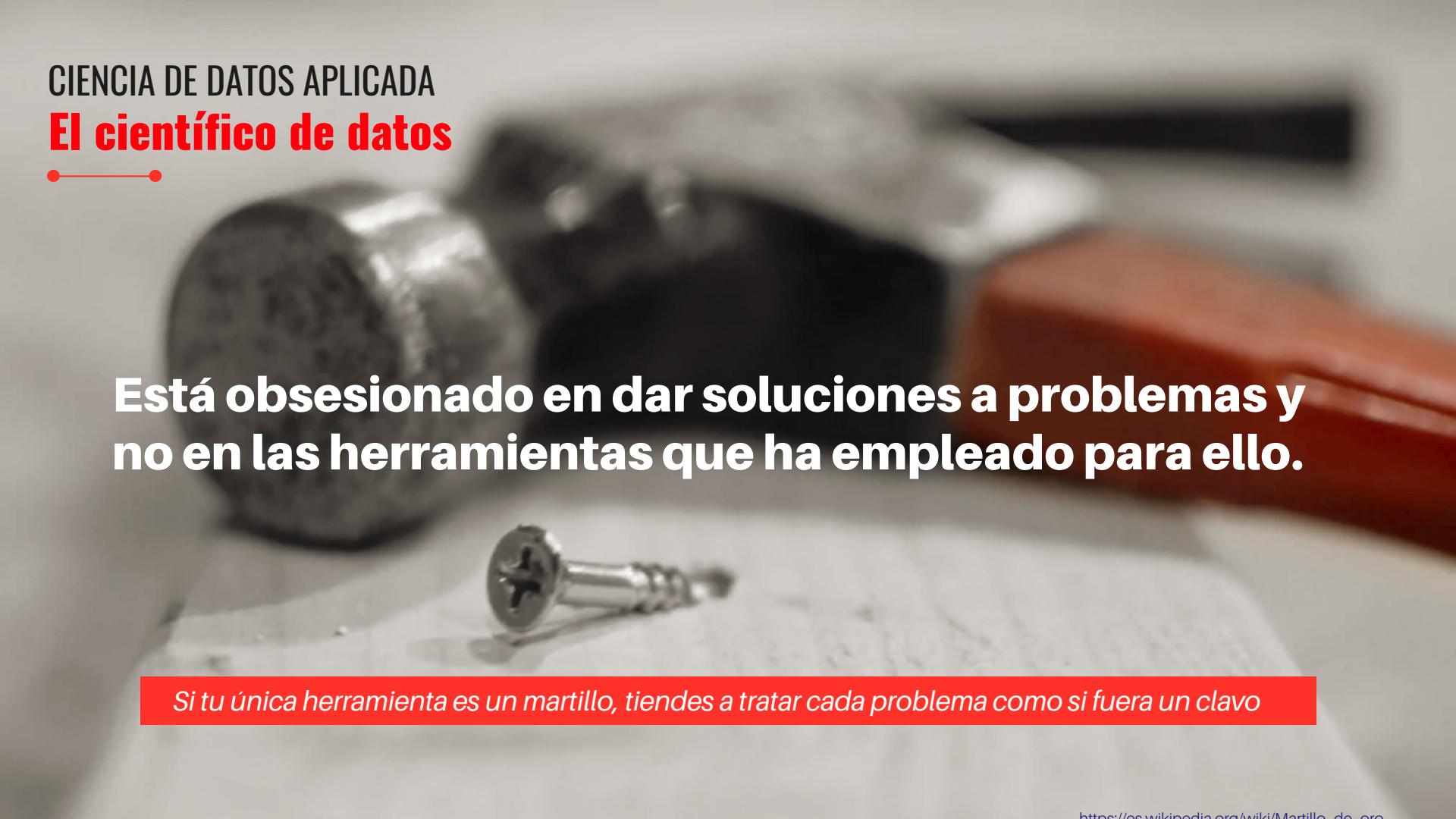
Trabajo en equipo



- Cálculo, álgebra
- Linux
- Python, R, SQL
- Manejo de datos
- Estadística
- Machine learning
- Visualización de datos
- Desarrollo de software
- ...

CIENCIA DE DATOS APLICADA

# El científico de datos

A black and white photograph of a hammer and a screwdriver lying on a light-colored wooden surface. The hammer is positioned horizontally in the upper left, and the screwdriver is lying diagonally across the lower center. The background is blurred.

Está obsesionado en dar soluciones a problemas y  
no en las herramientas que ha empleado para ello.

*Si tu única herramienta es un martillo, tiendes a tratar cada problema como si fuera un clavo*

CIENCIA DE DATOS APLICADA  
**El científico de datos**

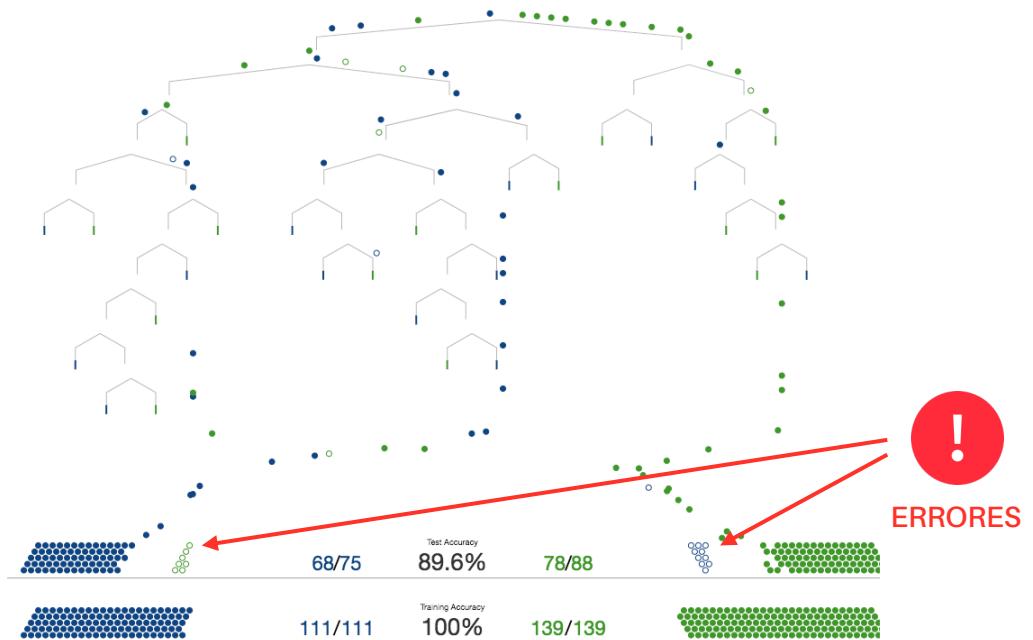


**No sufre parálisis por análisis**

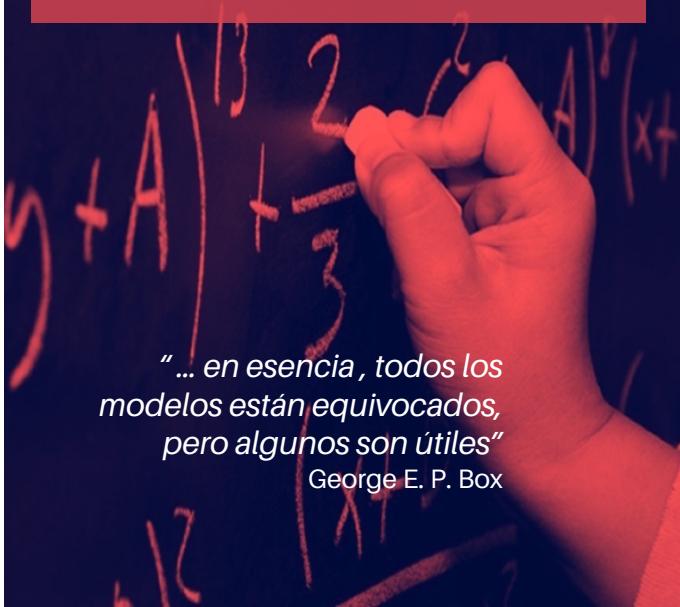


# CIENCIA DE DATOS APLICADA

## El científico de datos



Su prioridad es encontrar una solución ... aunque sabe que no es perfecta



“... en esencia, todos los modelos están equivocados, pero algunos son útiles”

George E. P. Box

CIENCIA DE DATOS APLICADA  
**El científico de datos**

A photograph showing several people's hands and arms as they work together on a project. One person in the center is pointing at a piece of paper with a pen. In the background, a laptop screen displays various data visualizations like bar charts and pie charts. The scene conveys a sense of teamwork and data analysis.

**Es un excelente comunicador**

*No sufre la "Maldición del Conocimiento"*

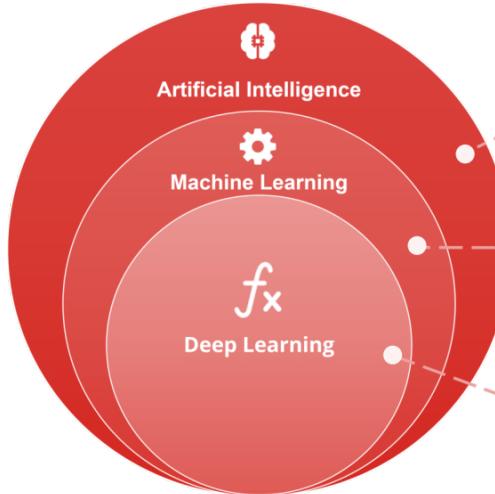
# CIENCIA DE DATOS APLICADA

## Metodología



# CIENCIA DE DATOS APLICADA

## AI Vs ML Vs DL



### ARTIFICIAL INTELLIGENCE

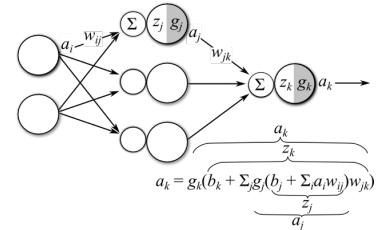
A technique which enables machines to mimic human behaviour

### MACHINE LEARNING

Subset of AI technique which use statistical methods to enable machines to improve with experience

### DEEP LEARNING

Subset of ML which make the computation of multi-layer neural network feasible



# MACHINE LEARNING

## Ejemplos

Retail	Marketing	Salud	Telco	Finanzas
<ul style="list-style-type: none"><li>• Predicción de la demanda</li><li>• Automatización de precios</li><li>• Segmentación de mercado</li><li>• Recomendaciones</li></ul>	<ul style="list-style-type: none"><li>• Análisis de comportamiento del usuario</li><li>• Social media</li><li>• Optimización de campañas</li></ul>	<ul style="list-style-type: none"><li>• Predicción de riesgo de enfermedad</li><li>• Diagnósticos y alertas</li><li>• Radiología</li></ul>	<ul style="list-style-type: none"><li>• Fuga de clientes</li><li>• Análisis de logs</li><li>• Detección de anomalías</li><li>• Mantenimiento preventivo</li></ul>	<ul style="list-style-type: none"><li>• Análisis de riesgo</li><li>• Detección de fraude</li><li>• Sistemas de scoring crediticio</li></ul>

# MACHINE LEARNING

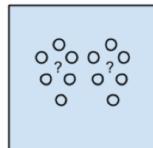
## Tipos de aprendizaje



ANÁLISIS

NO SUPERVISADO

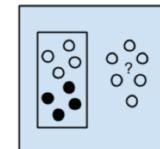
Descriptivo



*¿Podemos agrupar a nuestros usuarios en base a su comportamiento?*

SUPERVISADO

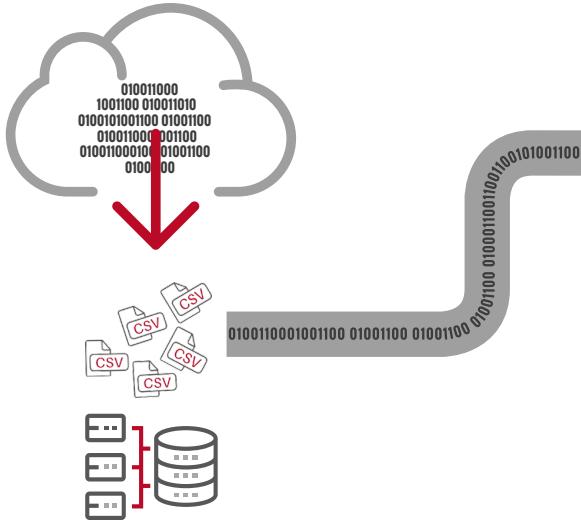
Predictivo



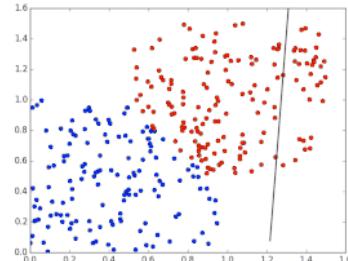
*¿Qué usuarios seguirán activos en 30 días?*

# MACHINE LEARNING

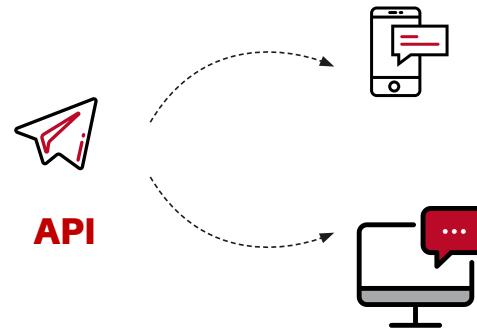
## Caso práctico: predicción de impago



### Modelado



### Acciones



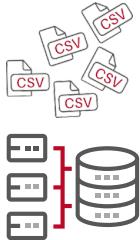
API

# MACHINE LEARNING

## Caso práctico: predicción de impago



¿Podemos predecir si una persona dejará de pagar su préstamo?



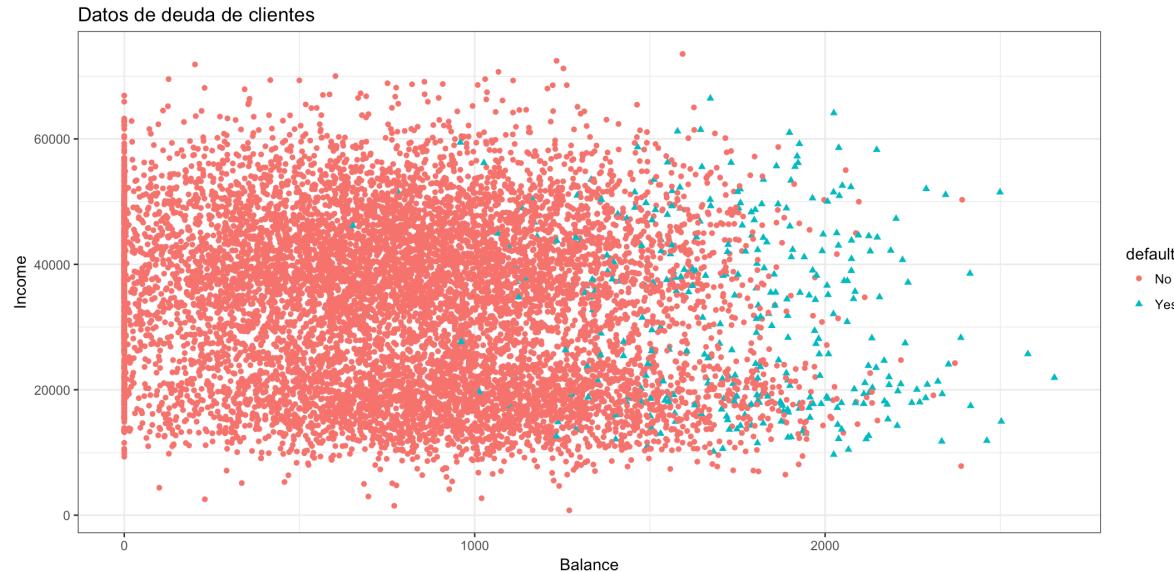
student	balance	income	default
Yes	1719	15752	Yes
No	1073	31767	No
No	825	24905	No
Yes	1856	33445	¿?

# MACHINE LEARNING

## Caso práctico: predicción de impago



¿Podemos predecir si una persona dejará de pagar su préstamo?



# MACHINE LEARNING

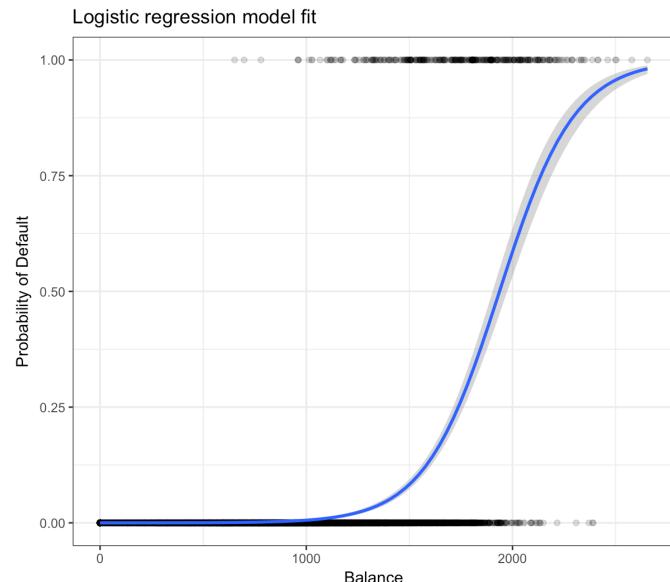
## Caso práctico: predicción de impago

### Regresión logística

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

Student (X1)	Balance (X2)	Income (X3)	Default (Y)
Yes	1719	15752	Yes
No	1073	31767	No
No	825	24905	No
Yes	1856	33445	??

- La variable de respuesta (Y) es categórica
- Se usan para predecir la probabilidad de una variable categórica en función de una o más variables predictivas (X)
- Nos permite decir en qué medida la presencia de una variable predictoria aumenta o disminuye la probabilidad de pertenecer a una clase (Y) en un determinado porcentaje.



Ejemplos: Y (aprobar/suspender, pagar/no pagar, abandonar/no abandonar)

# MACHINE LEARNING

## Caso práctico: predicción de impago



### Evaluación de los modelos de clasificación: la matriz de confusión

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN o Type II Error)
	Negativos	Falsos Positivos (FP o Type I Error)	Verdaderos negativos (VN)

- **VP** es la cantidad de *positivos* que fueron *clasificados correctamente* como positivos por el modelo.
- **VN** es la cantidad de *negativos* que fueron *clasificados correctamente* como negativos por el modelo.
- **FN** es la cantidad de *positivos* que fueron *clasificados incorrectamente* como negativos.
- **FP** es la cantidad de *negativos* que fueron *clasificados incorrectamente* como positivos.

# MACHINE LEARNING

## Caso práctico: predicción de impago

### Evaluación de los modelos de clasificación: Errores Tipo I y II

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN o Type II Error)
	Negativos	Falsos Positivos (FP o Type I Error)	Verdaderos negativos (VN)

Type I Error



Type II Error



# MACHINE LEARNING

## Caso práctico: predicción de impago

### Evaluación de los modelos de clasificación: Medidas de rendimiento

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN o Type II Error)
	Negativos	Falsos Positivos (FP o Type I Error)	Verdaderos negativos (VN)

$$\text{Exactitud} = \frac{VP + VN}{Total}$$

Porcentaje de los datos que son clasificados **correctamente**

$$\text{Tasa de error} = \frac{FP + FN}{Total}$$

Porcentaje de los datos que son clasificados **incorrectamente**

# MACHINE LEARNING

## Caso práctico: predicción de impago



		Predicción	
		Default	No default
Observación	Default	16 (VP)	43 (FN o Type II Error)
	No default	7 (FP o Type I Error)	1919 (VN)

### Sensibilidad (Recall/True Positive Rate):

Indica la capacidad de nuestro estimador para dar como casos positivos los casos realmente positivos. Mide la proporción de casos positivos (default) correctamente identificados.

$$\text{Sensibilidad} = \frac{VP}{\text{Total Positivos}} = \frac{VP}{VP + FN} = 0.27$$

# MACHINE LEARNING

## Caso práctico: predicción de impago



		Predicción	
		Default	No default
Observación	Default	16 (VP)	43 (FN o Type II Error)
	No default	7 (FP o Type I Error)	1919 (VN)

### Especificidad (True Negative Rate):

Indica la capacidad de nuestro estimador para dar como casos negativos los casos realmente negativos. **Mide la proporción de casos negativos (No default) correctamente identificados.**

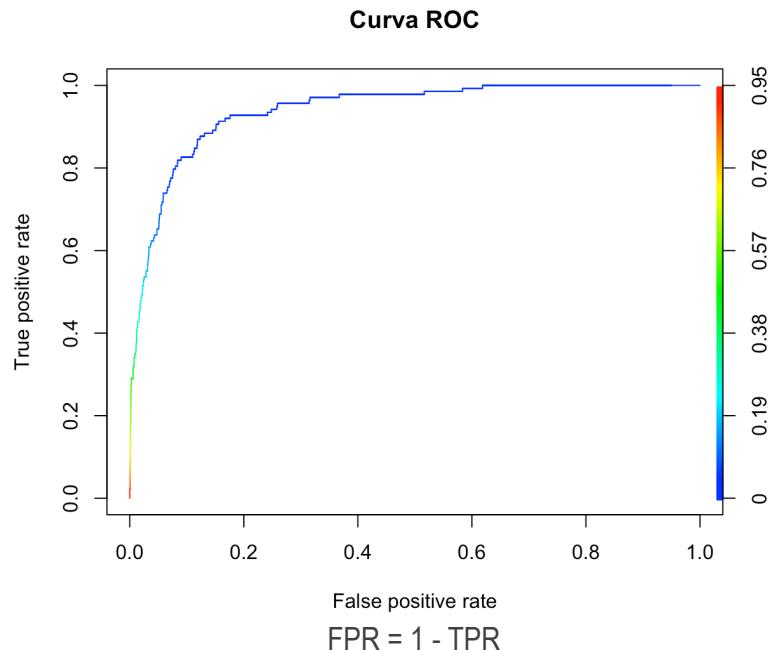
$$\text{Especificidad} = \frac{VN}{\text{Total Negativos}} = \frac{VN}{VN + FP} = 0.99$$

# MACHINE LEARNING

## Caso práctico: predicción de impago

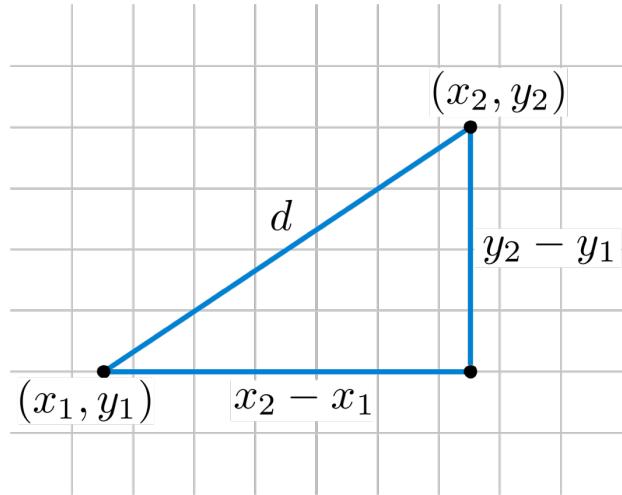


Evaluación de los modelos de clasificación: Curva Receiver Operating Characteristic



# MACHINE LEARNING

## Caso práctico: segmentación de clientes



$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

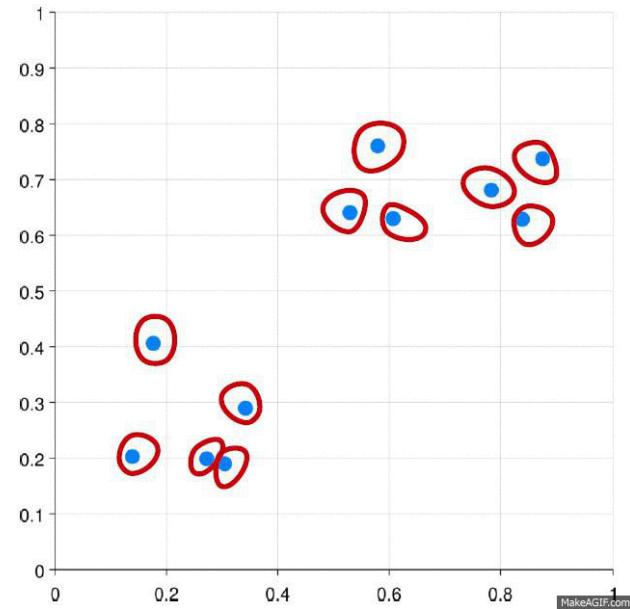
¿Cómo medimos la distancia entre dos clientes?

# MACHINE LEARNING

## Caso práctico: segmentación de clientes



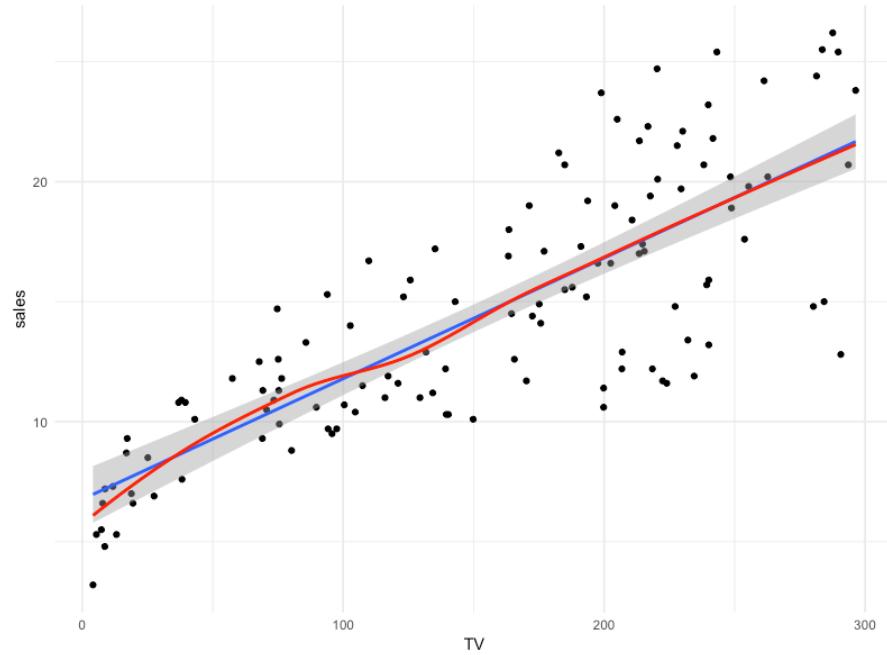
K-Means



Clústering jerárquico

# MACHINE LEARNING

## Caso práctico: marketing mix



Regresión lineal

# GRACIAS



**Jose Ramón  
Cajide**

DATA SCIENTIST

---

*Master en Analitica Web por la Universidad British Columbia. Master en Analitica Web. Web Analytics Certification (Market Motive). Master en Data Science. Google Regional Trainer part of the Google Partner Academy Programme.*



<https://www.linkedin.com/in/jrcajide/>



@jrcajide

Desde 2011  
potenciando las  
**ventajas  
competitivas**  
de nuestros  
clientes

PASIÓN Y PRECISIÓN

