eXplainable Machine Learning

Seminario (I+A)A

¿Qué vamos a ver?

¿Cuáles son mis objetivos?

¿Quién soy yo?

¿Qué es el eXplainable Machine Learning y por qué es necesario?

¿Qué es la explicación?

¿Cómo podemos conseguirla con ML?

¿Me lo demuestras?

Conclusiones

Mis Creencias, Deseos e Intenciones

¡Es mi **TFG**! Vuestra ayuda está genial

Crear **debate**

Enseñaros mi *Bitmoji*

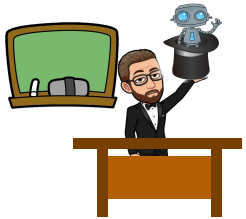
Haceros ver la importancia del **XML**

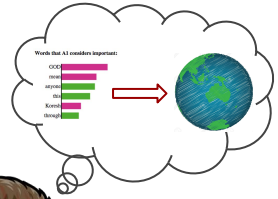
Haceros **pensar** cuál es el camino correcto

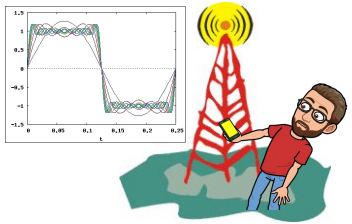
Relajar el **nivel**

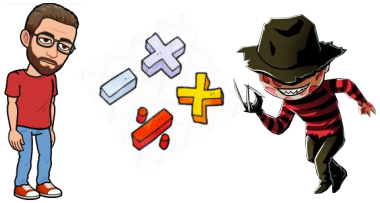
¿Quién soy yo?

7 de Junio de 1991 - 2,7 Kg



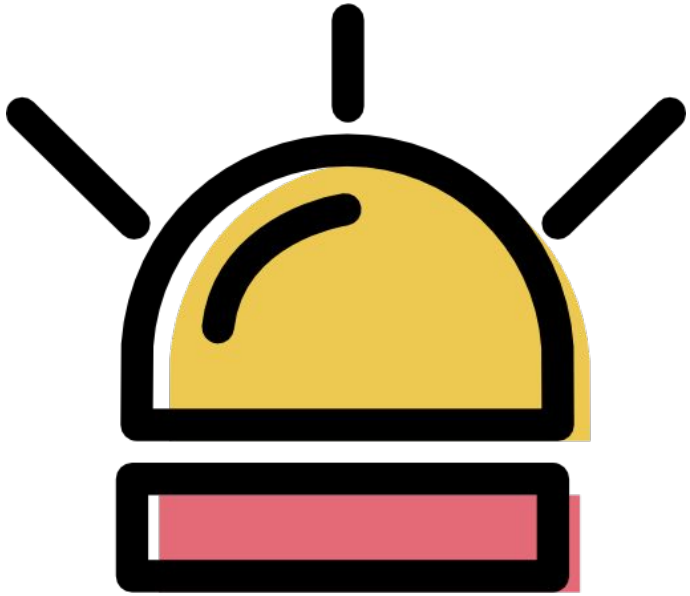






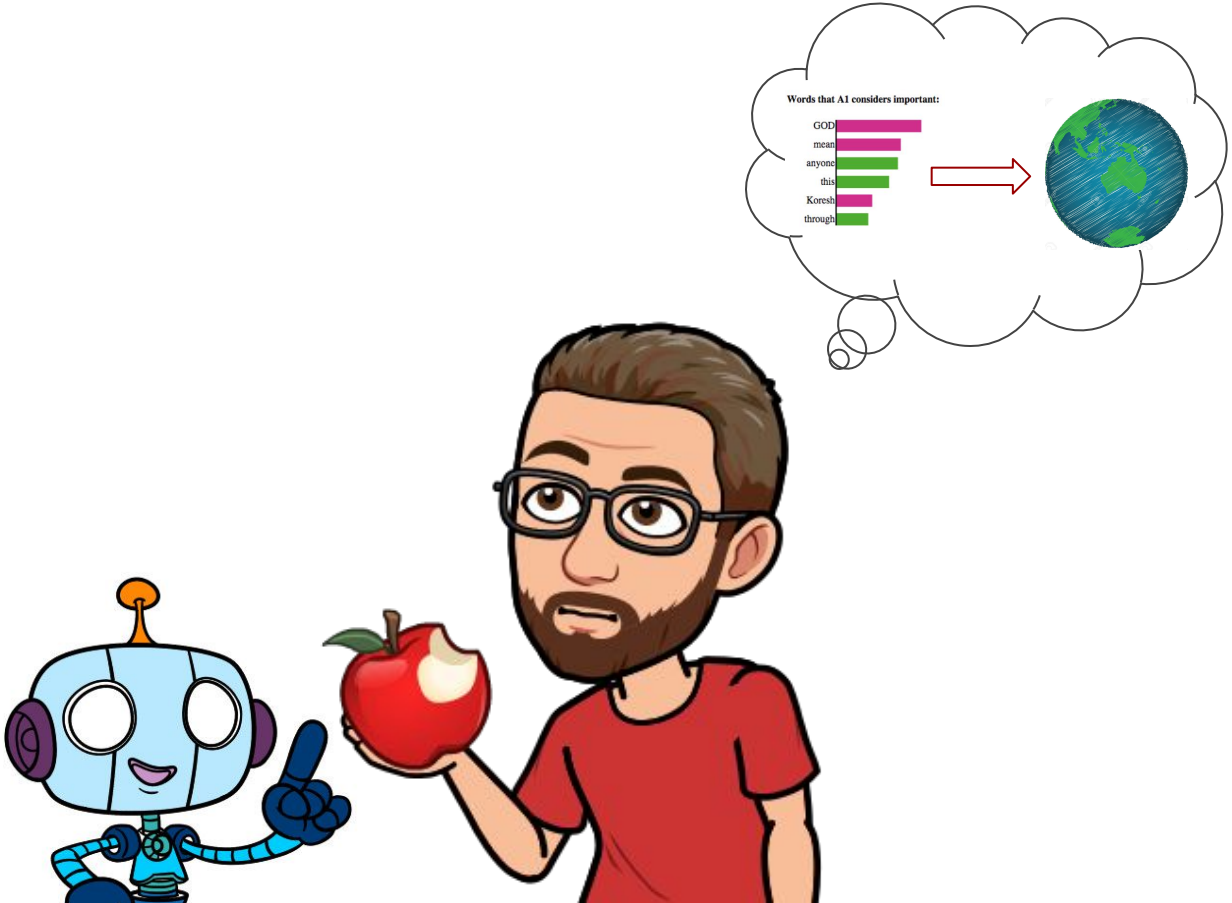
¿Qué es la pregunta correcta?

vs.



TU TURNO

ML como herramienta al conocimiento

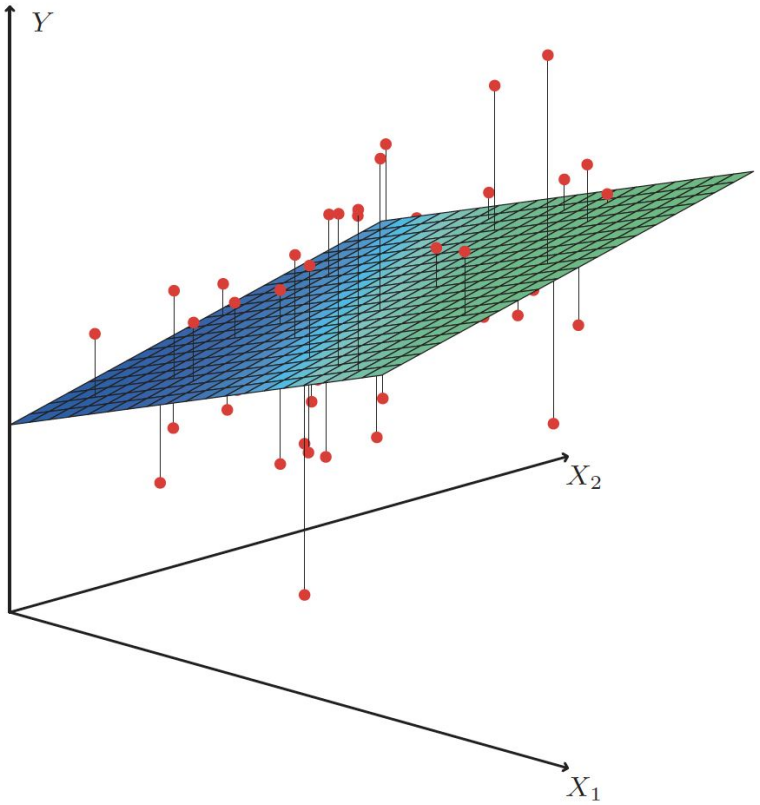


Ejemplos

vs.

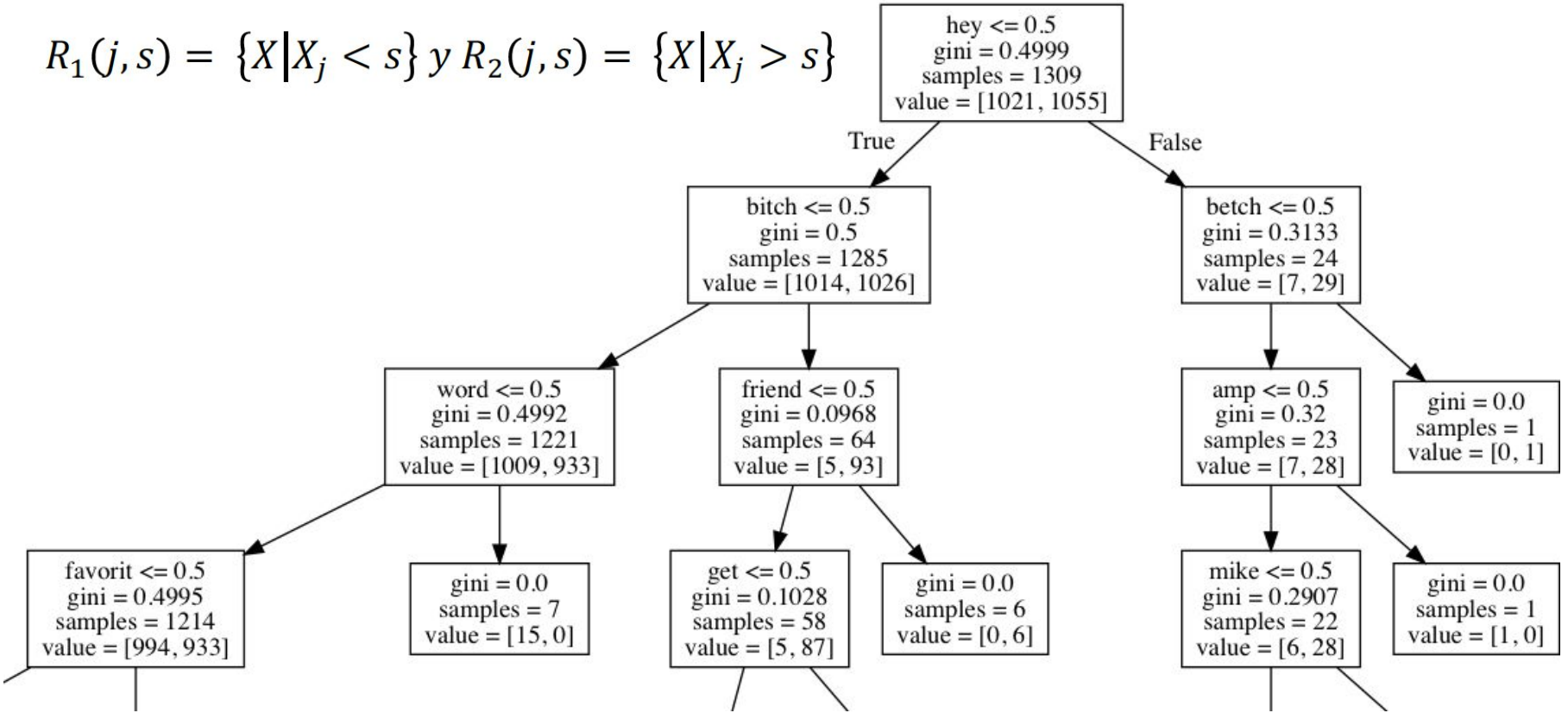
Regresión Lineal

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

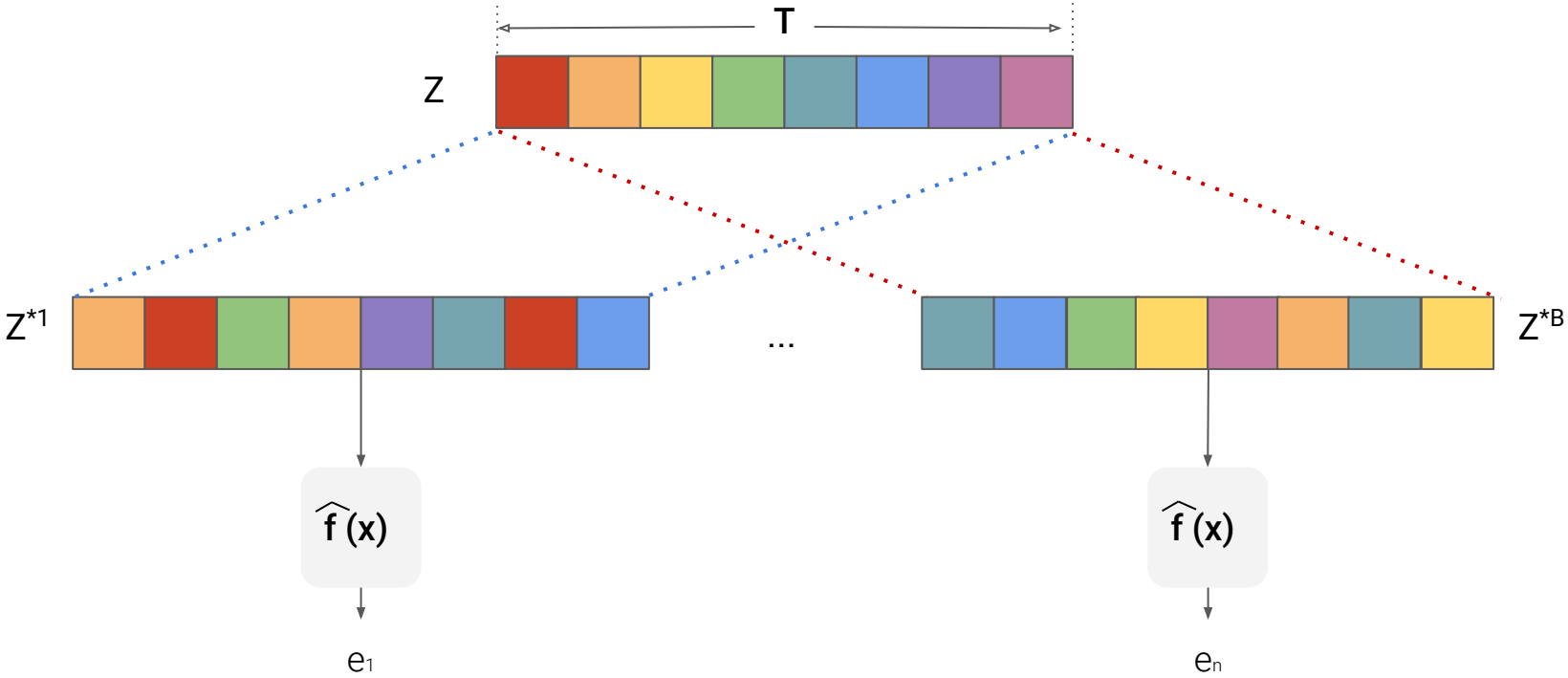


Árboles de decisión

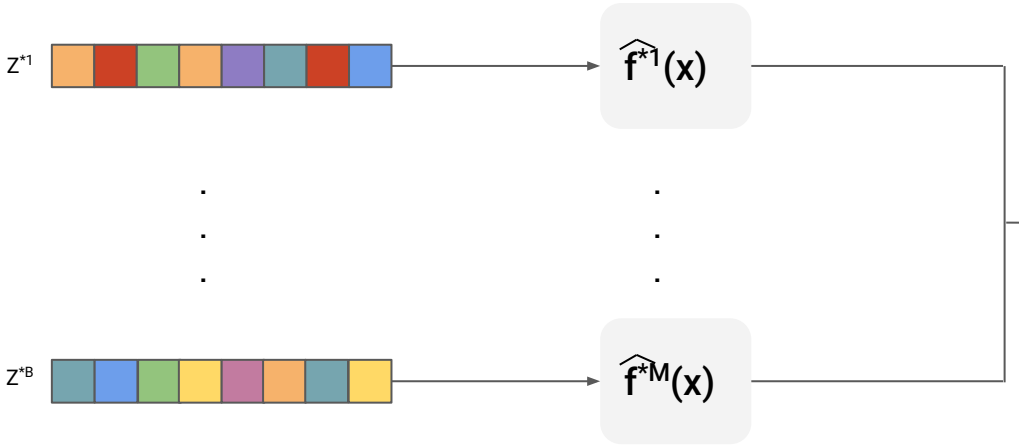
$$R_1(j,s) = \{X|X_j < s\}\ y\ R_2(j,s) = \{X|X_j > s\}$$



Bootstrap

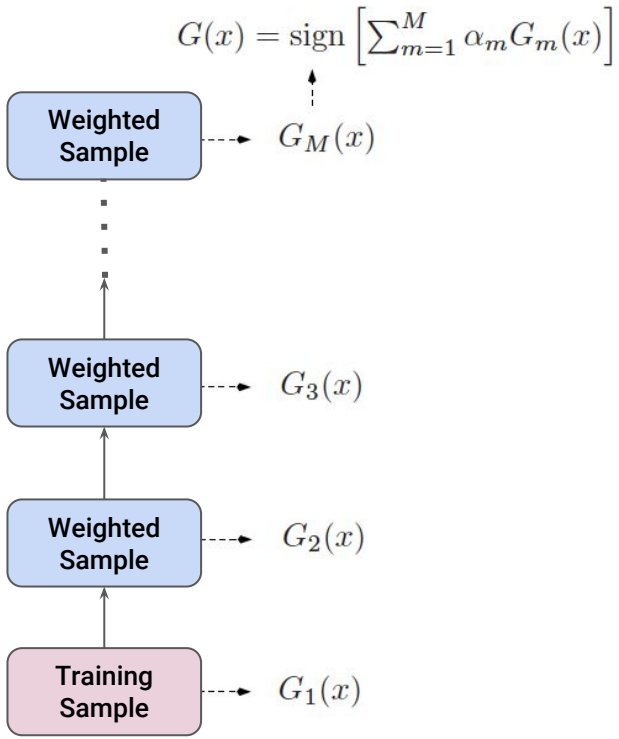


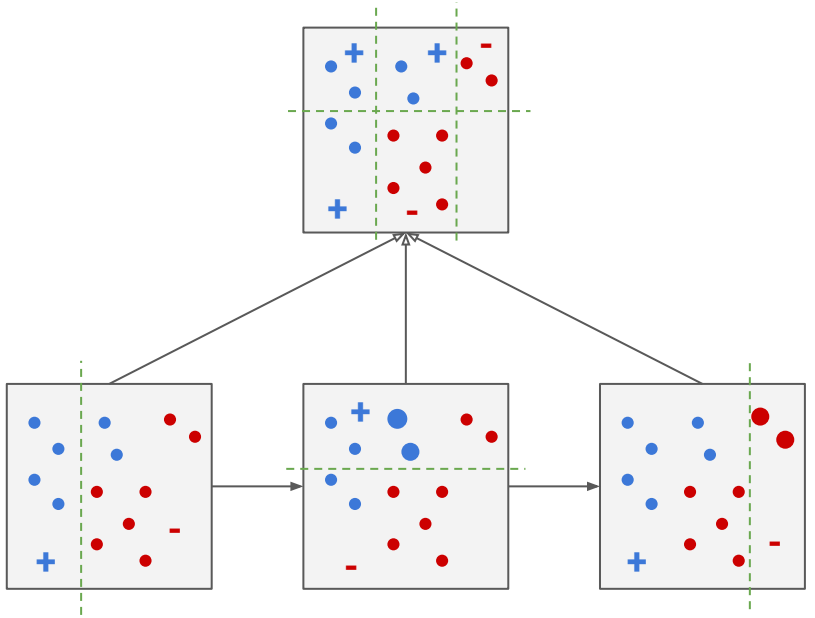
Bagging



$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x).$$

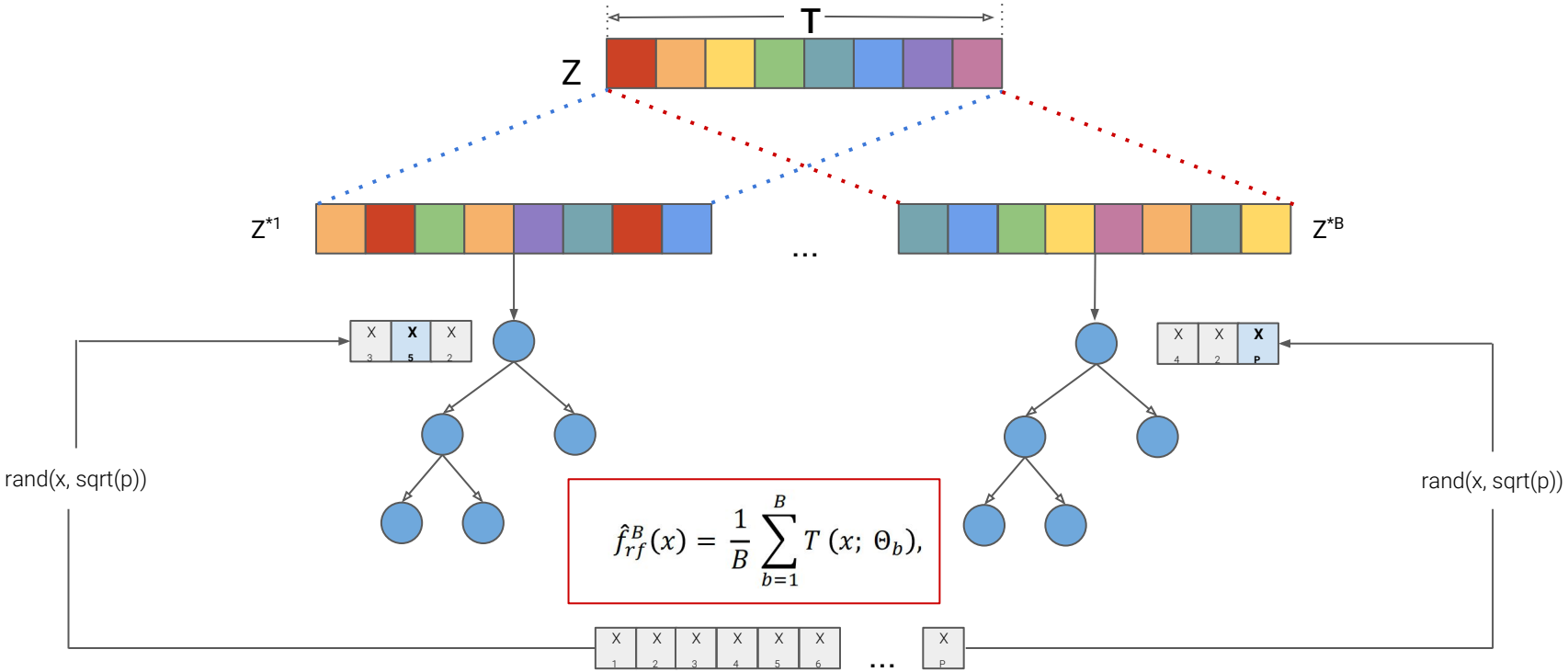
Boosting


$G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$
Weighted Sample
$G_M(x)$
Weighted Sample
$G_3(x)$
Weighted Sample
$G_2(x)$
Training Sample
$G_1(x)$



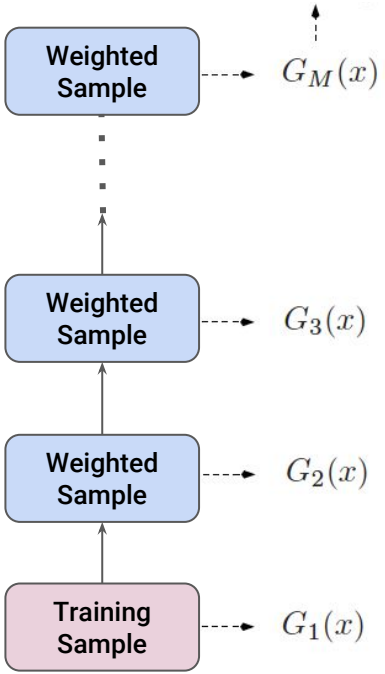
Random Forest

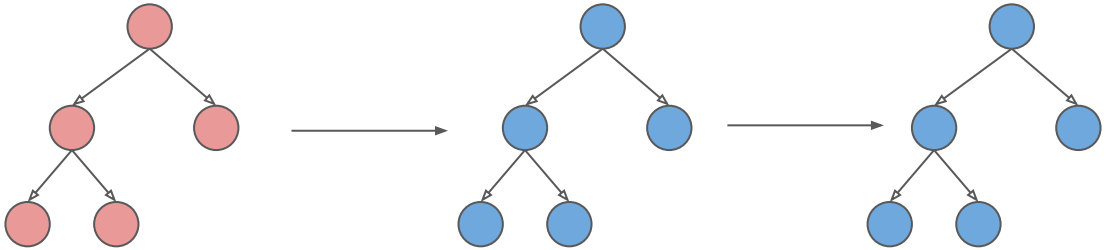
Árboles de decisión + Bootstrap + Bagging



Gradient Boosting

Árboles de decisión + Boosting + Descenso de gradiente

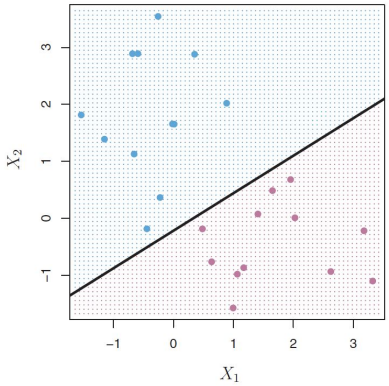


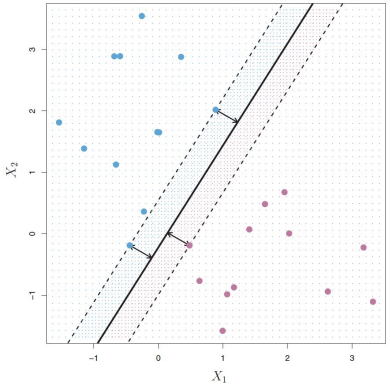


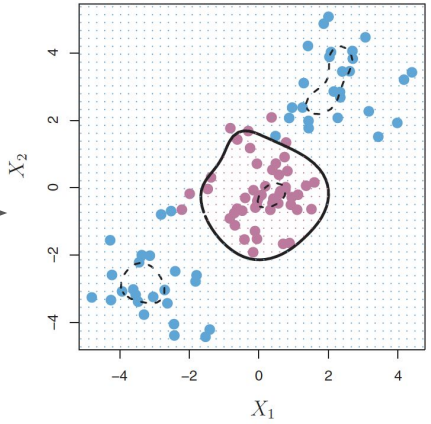
$\tilde{\Theta}_1$ $\tilde{\Theta}_2$ $\tilde{\Theta}_m$

$$= \arg\min \sum_{i=1}^{N} (-g_{im} - T(x_i;\ \Theta_m))^2.$$

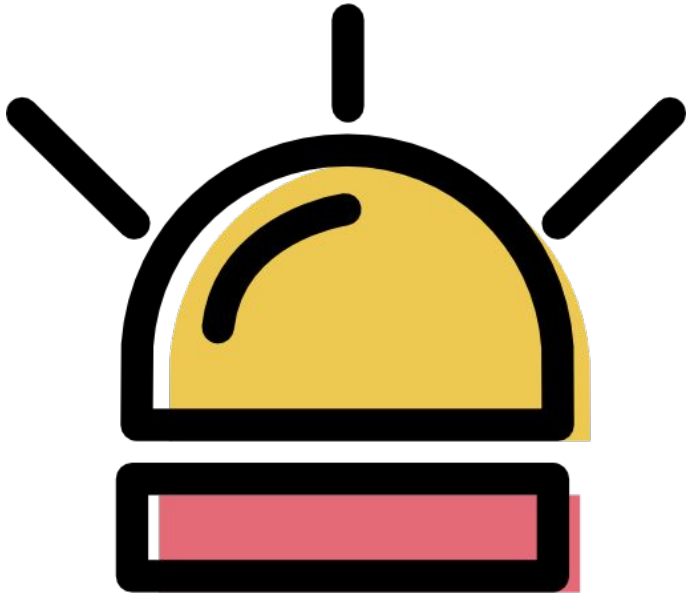
SVM





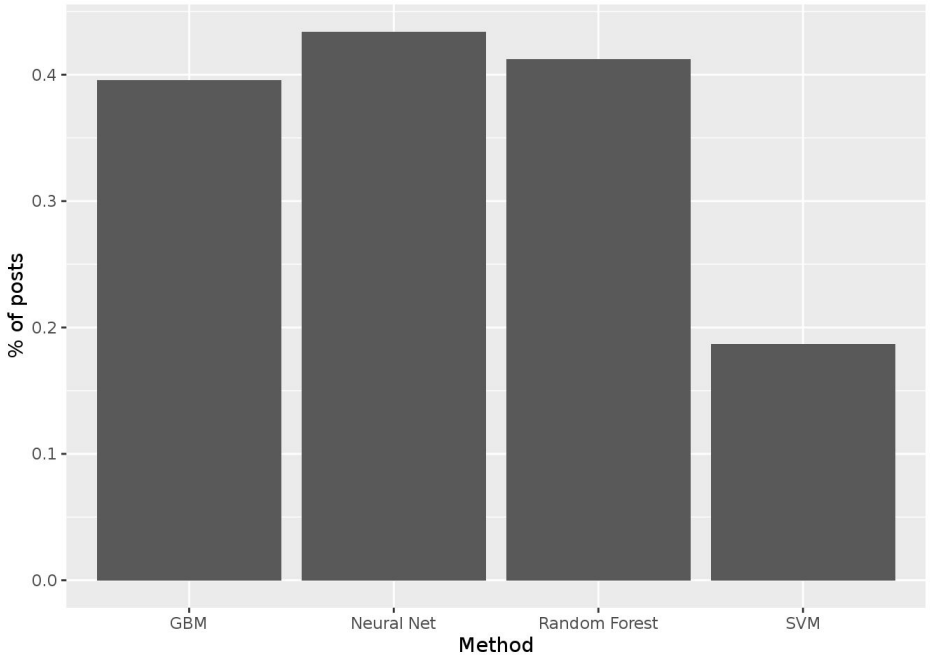


$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) .$$



TU TURNO

Y ¿Qué usamos?



https://www.kaggle.com/msjgriffiths/r-what-algorithms-are-most-successful-on-kaggle/notebook

Vale, vale… ya entiendo…

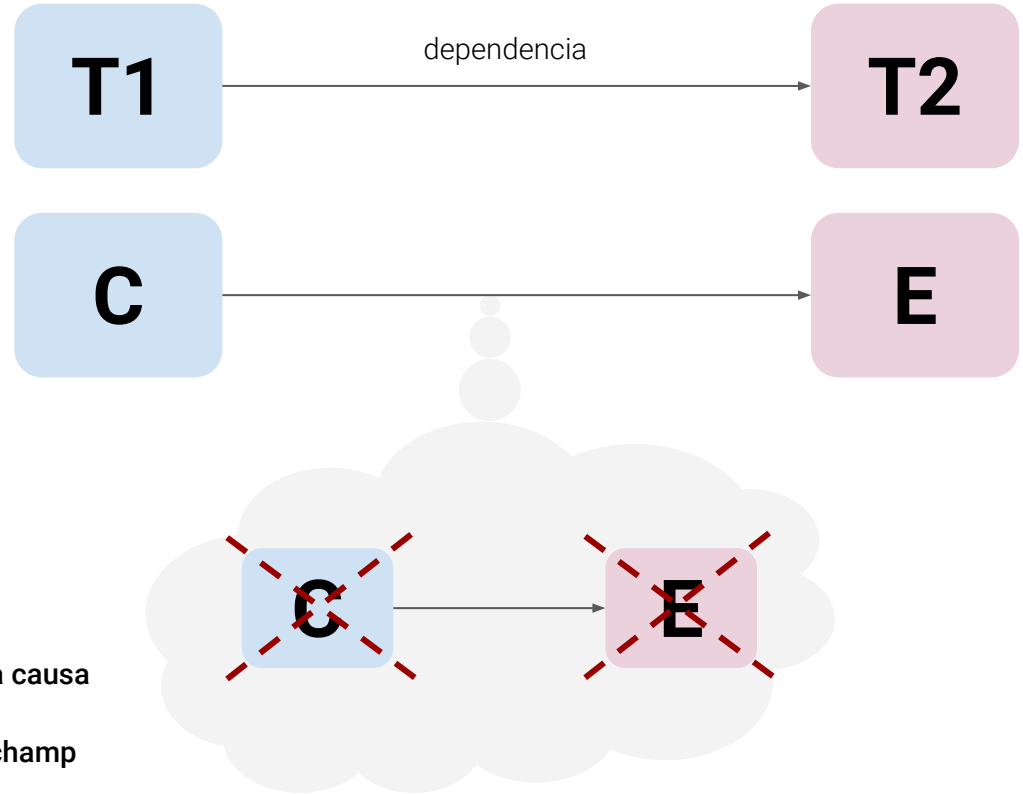
vs.

¿Qué es la explicación? Según la filosofía...

"Explicar un evento es proveer alguna información sobre su historia **causal**. En un acto de explicación, alguien que está en posesión de alguna información sobre la historia **causal** de algún evento – información explicativa, lo llamaré – intenta transmitírselo a otra persona".

- **Lewis**

¿Qué es la causalidad? Dependencia y Contrafácticos



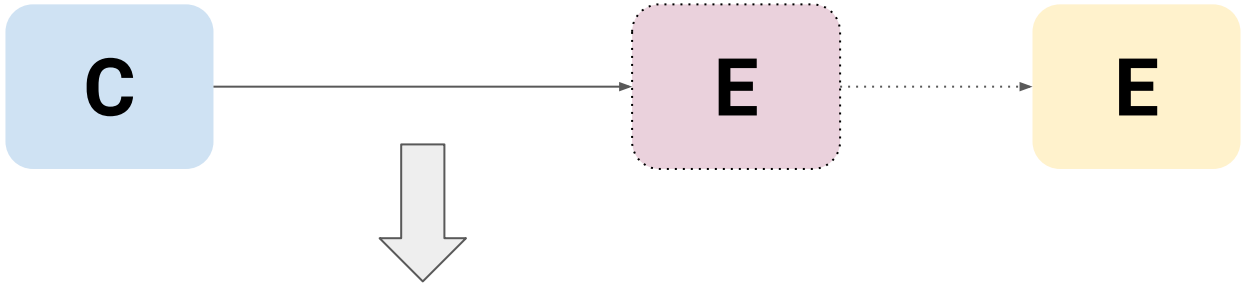
Teoría de la regularidad de la causa

Hume, Beauchamp

¿Qué es la causalidad? Dependencia y Contrafácticos

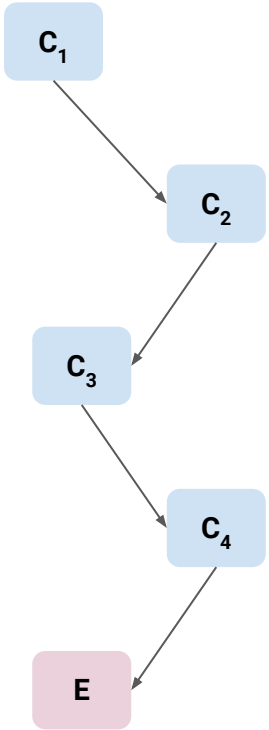
¿Qué es la causalidad? Dependencia y Estadística

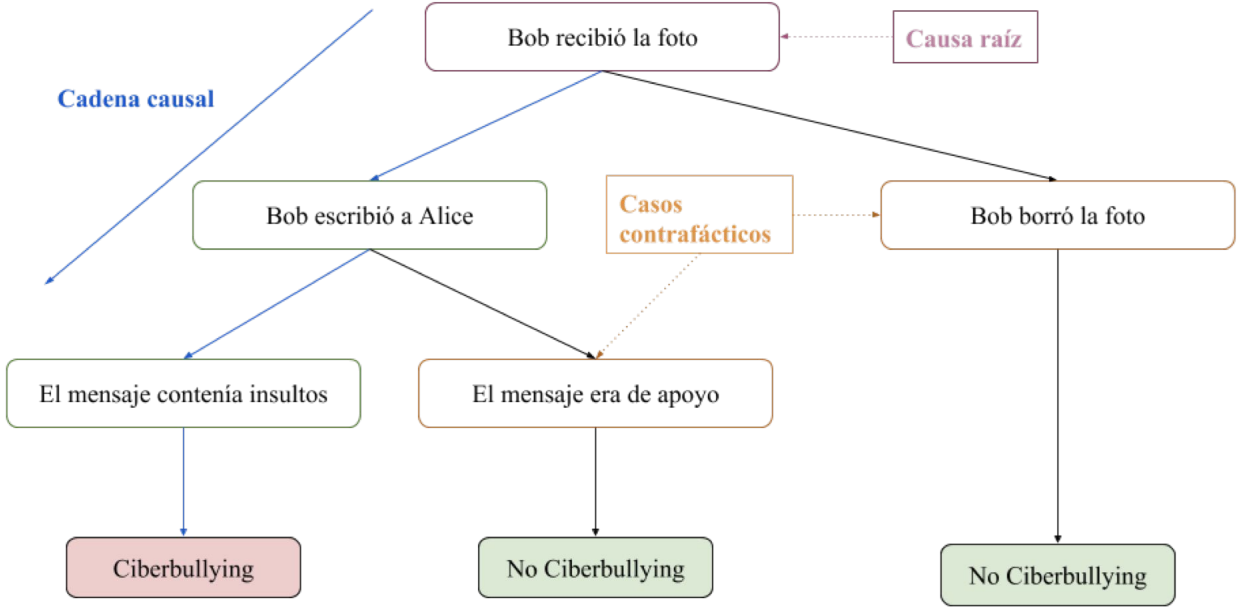
La teoría del intervencionismo de la causalidad.

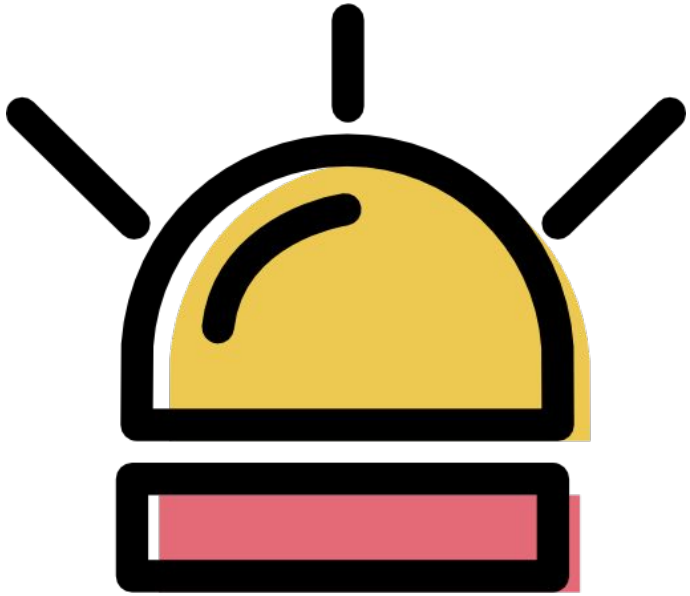


La teoría probabilística.

¿Qué es la causalidad? Cadenas causales

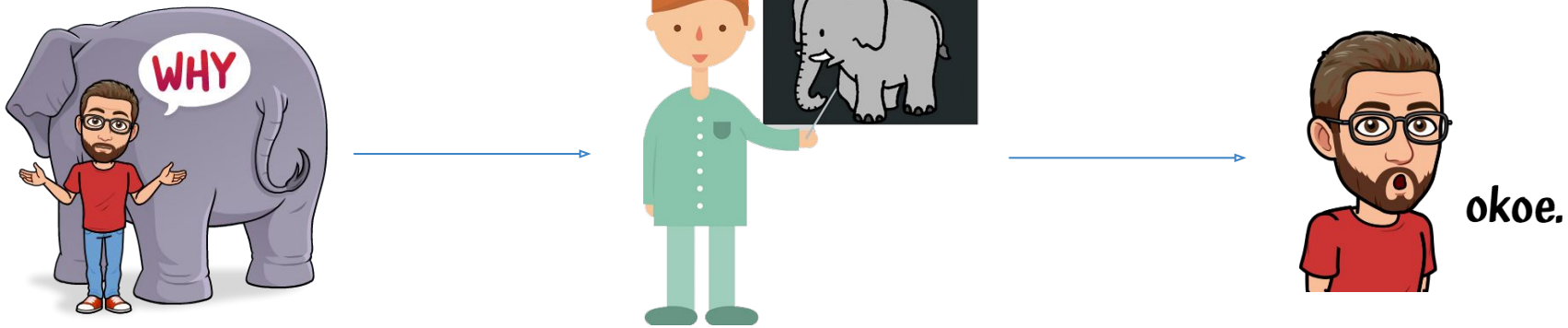






TU TURNO

¿Por qué necesitamos explicación?



PARA APRENDER

¿Por qué necesitamos explicación?

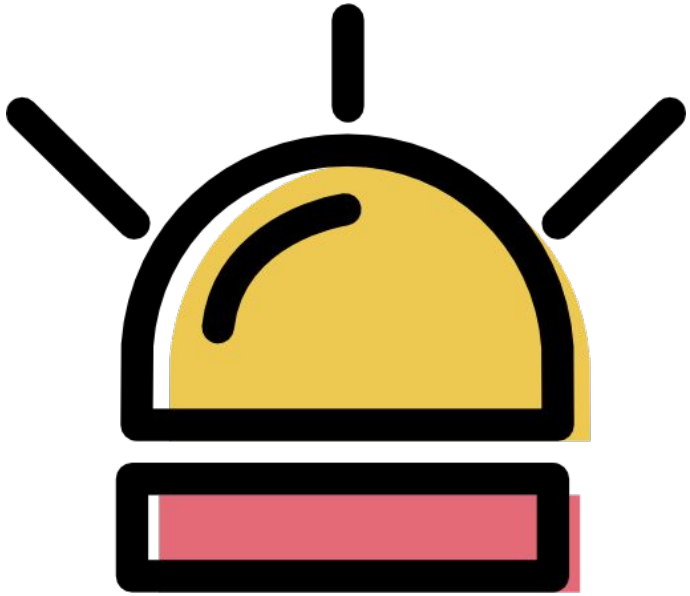
"Las explicaciones tienen un rol en el aprendizaje inferencial *porque* son explicaciones, no sólo revelan información causal"

- **Lombrozo**

¿Por qué necesitamos explicación?

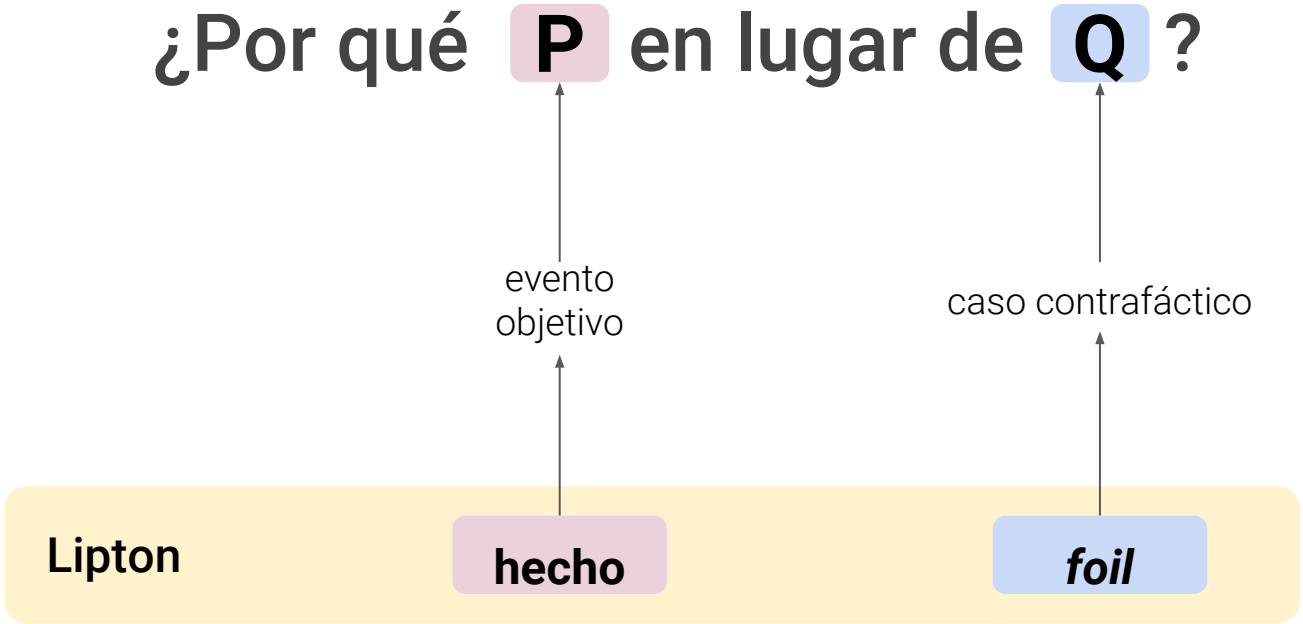
- **Encontrar sentido** para disminuir las contradicciones o inconsistencias entre elementos de nuestra estructura de conocimiento.
- **Gestionar interacciones sociales** para crear un significado compartido de algo

- **Malle**



TU TURNO

La explicación contrastiva



Procesos cognitivos. Atribución social

La atribución social es la **percepción de la persona.**

- **Helder**

Procesos cognitivos. Psicología popular

Aunque no son las verdaderas causas del comportamiento humano, son las que se usan para modelar y predecir el comportamiento de los demás.

1. Precondición de la acción

2. La acción en sí misma

3. Los efectos de la acción

Procesos cognitivos. BDI

- Creencias
- Deseos
- Intenciones

COMPRENDER

PREDECIR

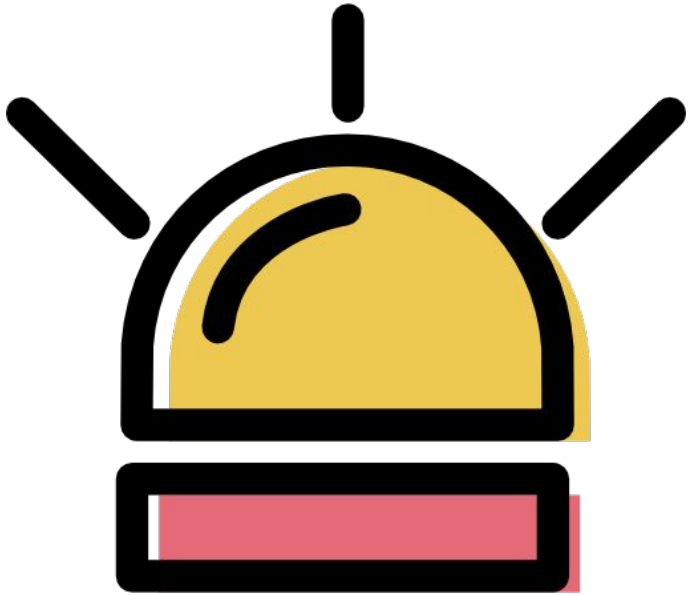
EXPLICAR

- Kashima

Procesos cognitivos. Normas y Moral

Efecto de Knohe

- Uttich y Lombrozo



TU TURNO

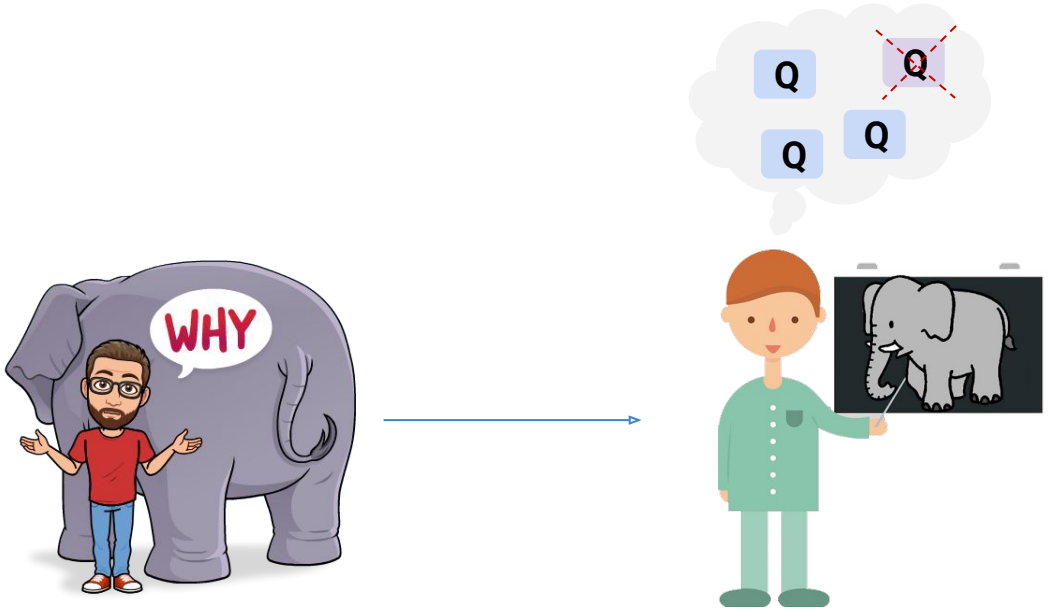
Explicación Social.

“Hay tantas causas de x como explicaciones de x. Considerar cómo la causa de la muerte pudo haber sido establecida por el médico como ‘hemorragia múltiple’, por el abogado como ‘negligencia por parte del conductor’, por el fabricante de coches como ‘defecto en la construcción de los frenos’, por un ingeniero civil como ‘la presencia de arbustos altos en ese momento’. Ninguno es más cierto que cualquiera de los otros, pero el contexto particular de la pregunta hace de algunas explicaciones más relevantes que otras.”

- **Hanson**

Explicación Social. Mutabilidad

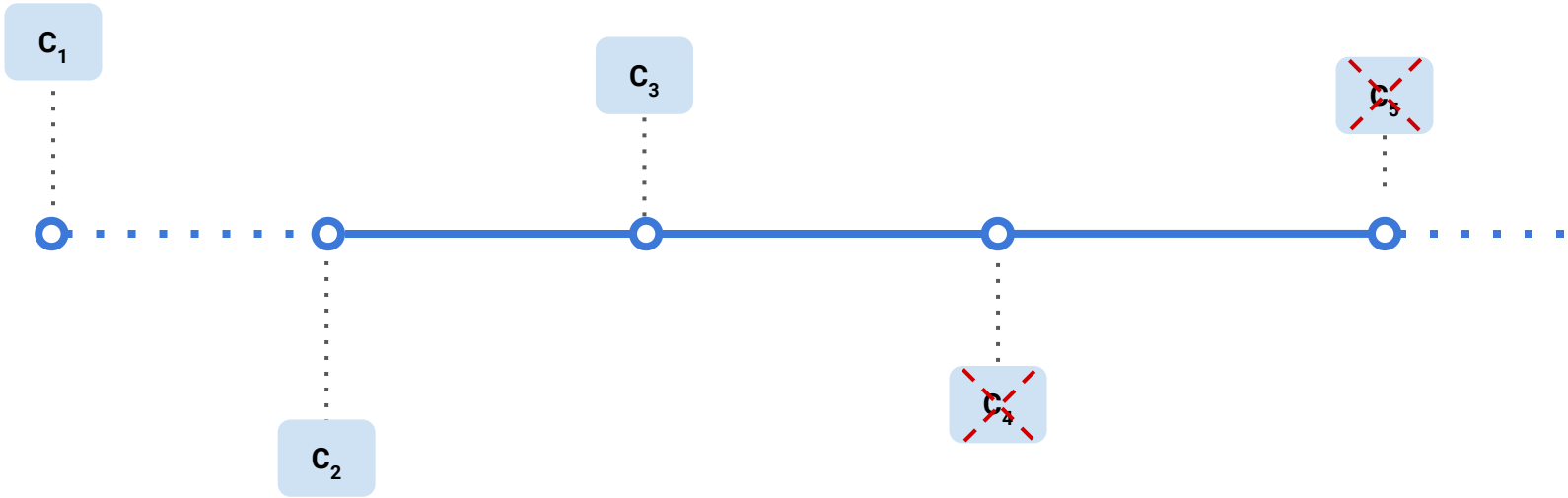
Anormalidad



- *Heurísticos de simulación*. Kahneran y Tversky

Explicación Social. Mutabilidad

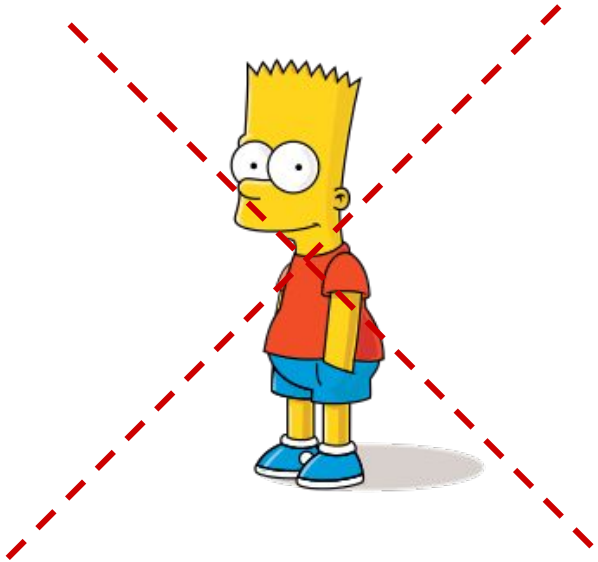
Temporalidad



- Miller y Gunasegaram

Explicación Social. Mutabilidad

Norma

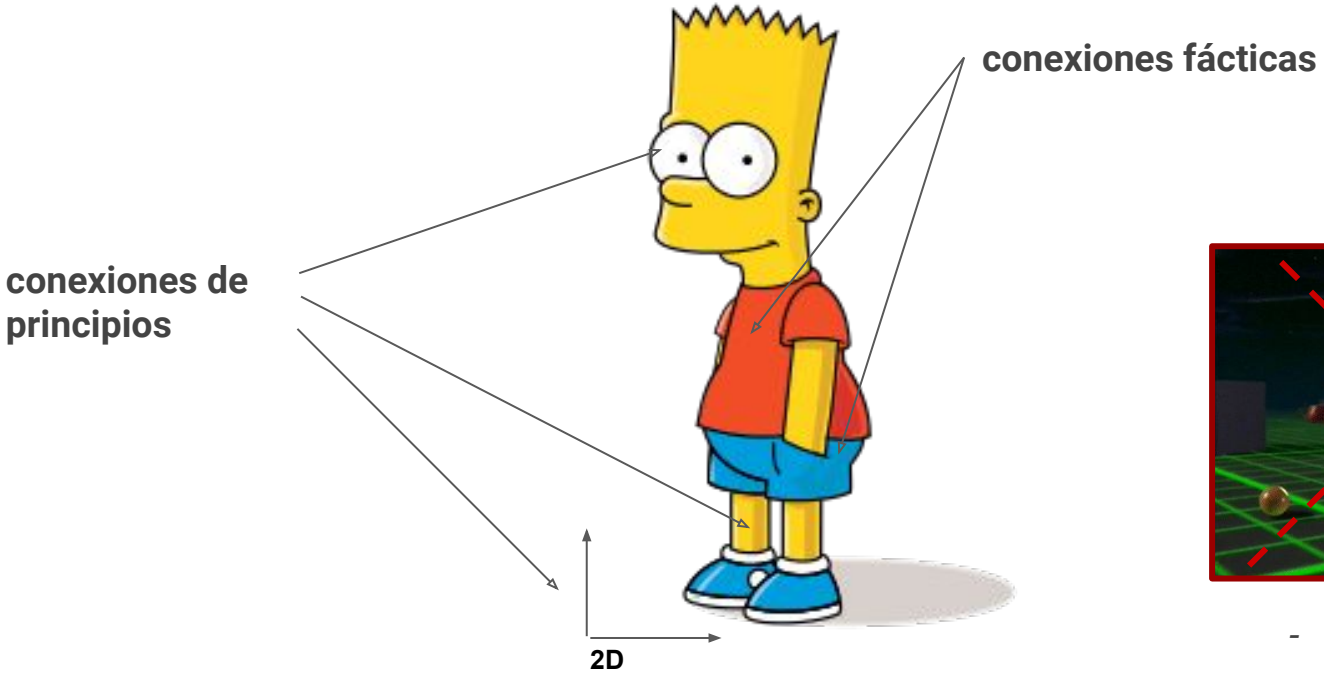


- McCloy y Byme

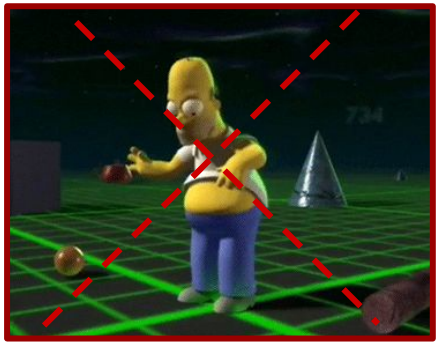
Explicación Social. Mutabilidad

k-propiedades y t-propiedad

Eres un personaje de los Simpsons si...



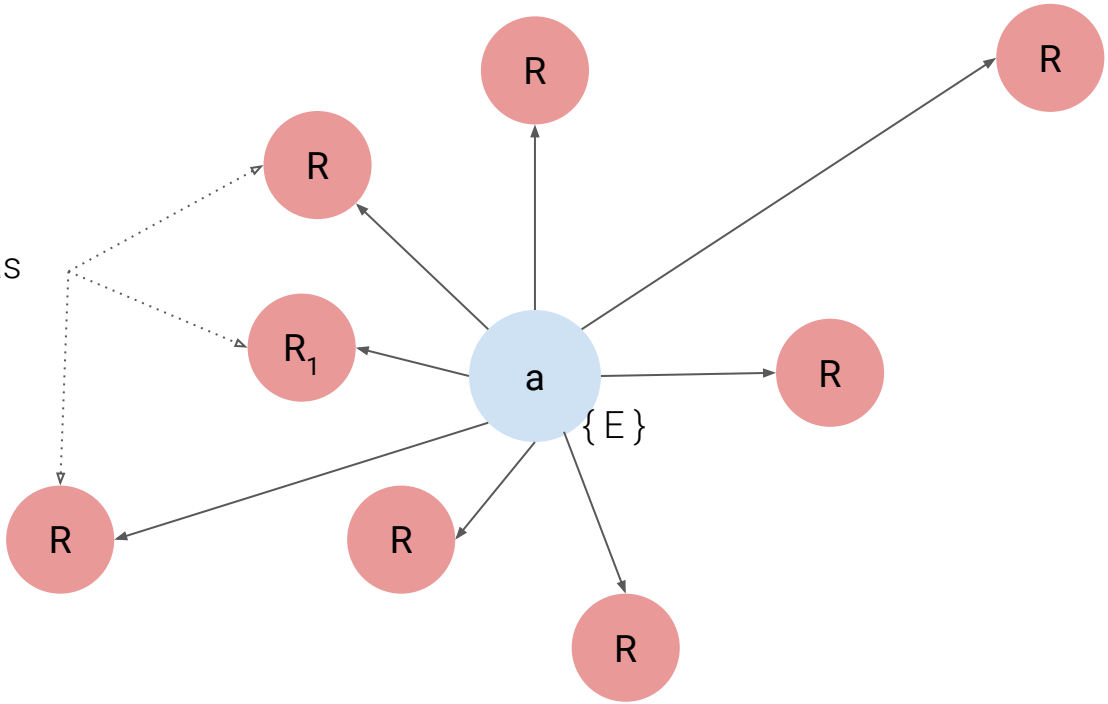
Anomalía



- Prasada y Dilinghan

Explicación Social. Selección

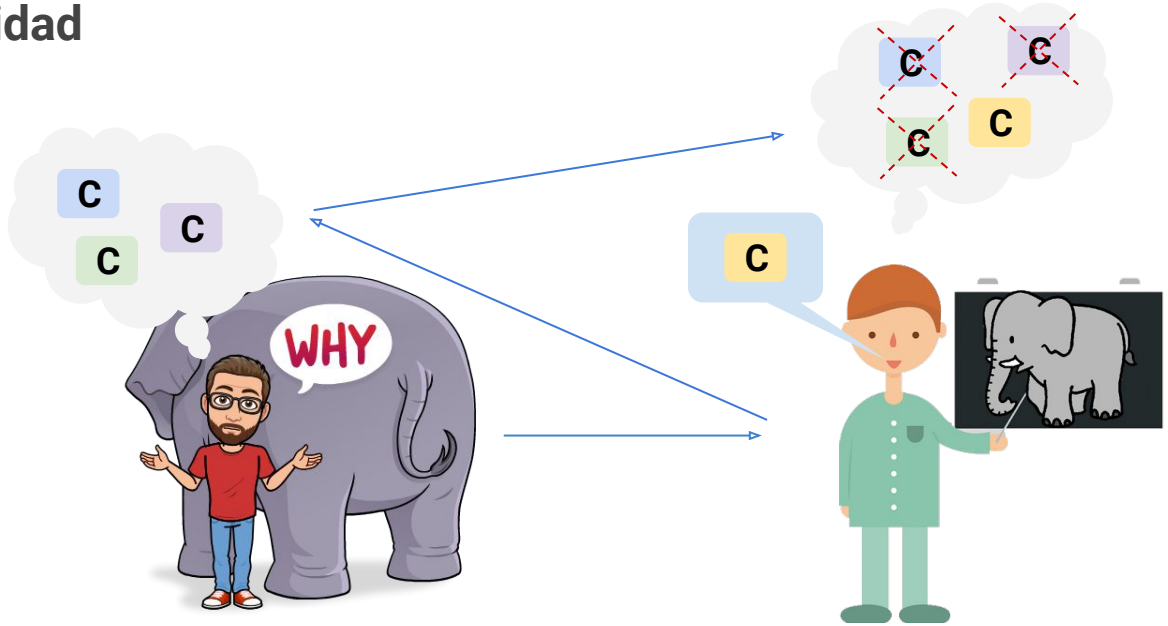
Clases de referencias



- Hesslow <- Lipton le apoya

Explicación Social. Selección

Anormalidad



- Hilton y Slugoski. Modelo de condiciones anómalas

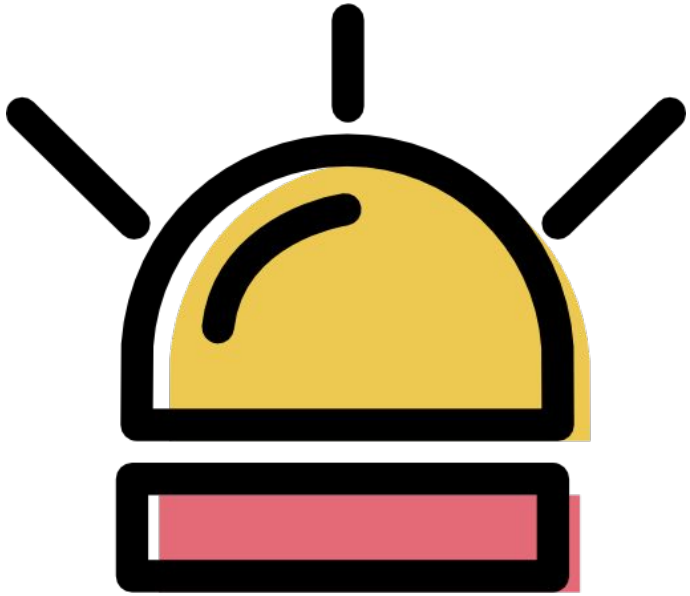
Explicación Social. Evaluación

- **Probabilidad**. No siempre la más probable o verdadera es la mejor (Hilton): Juzgan más por su utilidad o relevancia

- **Simplicidad**

- **Generalización**

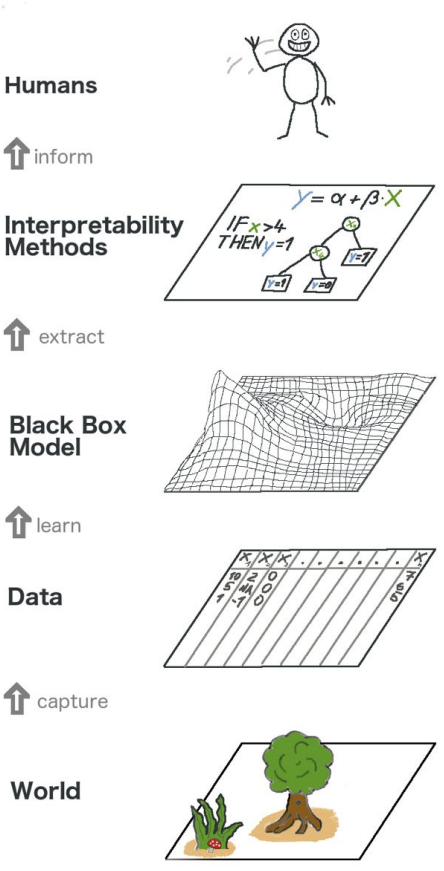
- **Coherencia** con creencias previas (*Teoría para la coherencia explicativa* - Thagard)



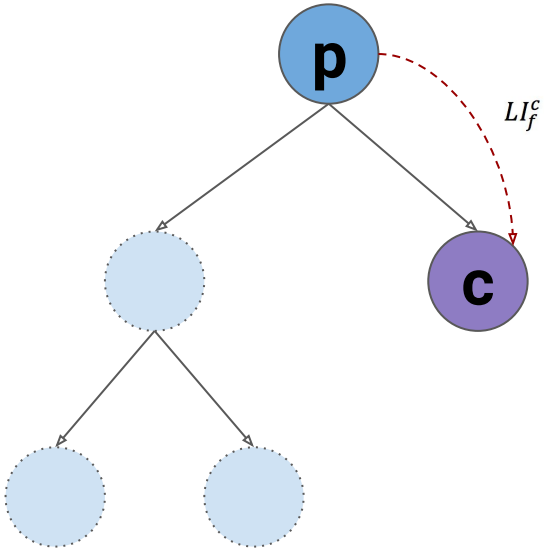
TU TURNO

Interpretación agnóstica de modelos

- Flexibilidad del modelo
- Flexibilidad de la explicación
- Flexibilidad en la representación



Contribución de características



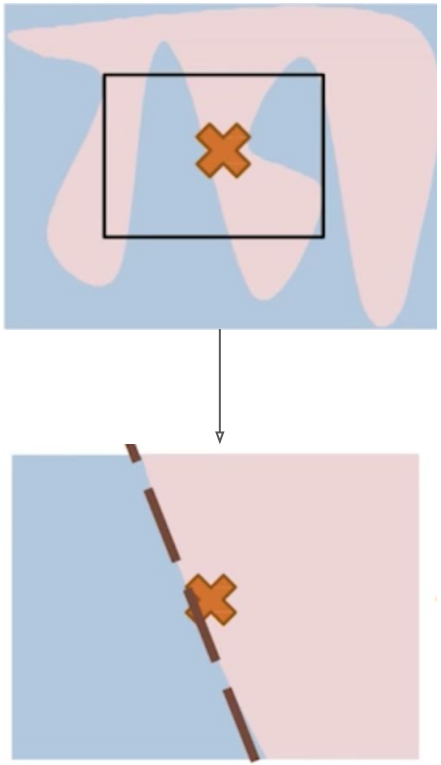
El número de instancias que pasa por p que pertenece a la clase C_k

$$LI_f^c = \begin{cases} Y_{mean}^c - Y_{mean}^p, & \text{si la divisón en el padre se realiza sobre la característica } f \\ 0, & \text{en cualquier otro caso} \end{cases}$$

$$FC_i^f = \frac{1}{T} \sum_{t=1}^{T} FC_{i,t}^f .$$

L.I.M.E

Local Interpretable Model-agnostic Explanations



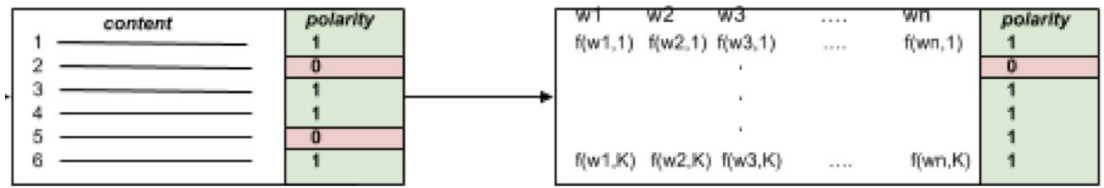
1. Se centra en una instancia o conjunto de ellas de la misma clase y crea **datos fake** a partir de ellas.
2. **Calcula la distancia** entre los datos fake y los reales.
3. Usa el modelo *black-box* para **crear las predicciones** del nuevo dataset
4. Con distintos modelos conoce las ***m* características más importantes (Lasso)**
5. Crea un **regresor lineal** para sacar los coeficientes
6. Crea una **explicación** con esos coeficientes

Detección automática de ciberbullying

Datos

formspring

2000 tuplas con dataset compensado

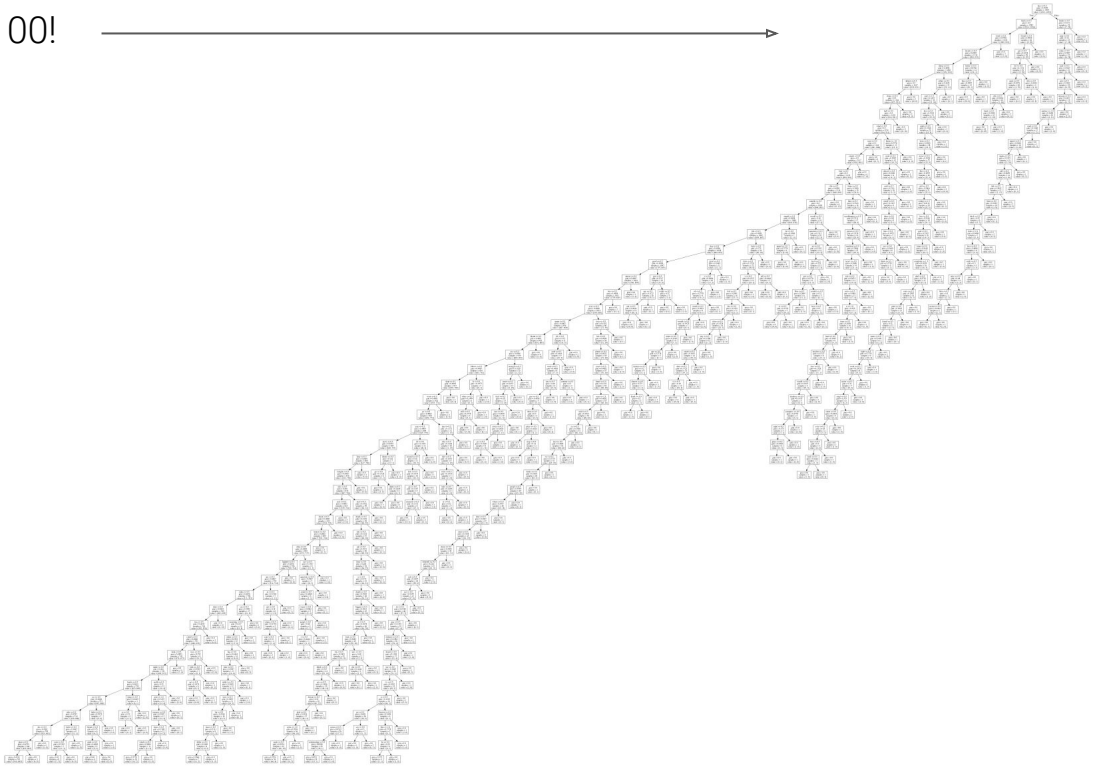


Modelos

	Score (%)	Falsos Positivos (%)
SVM	69,39	12,97%
RF	72,74	13,8%
GBM	72,78	13,11%

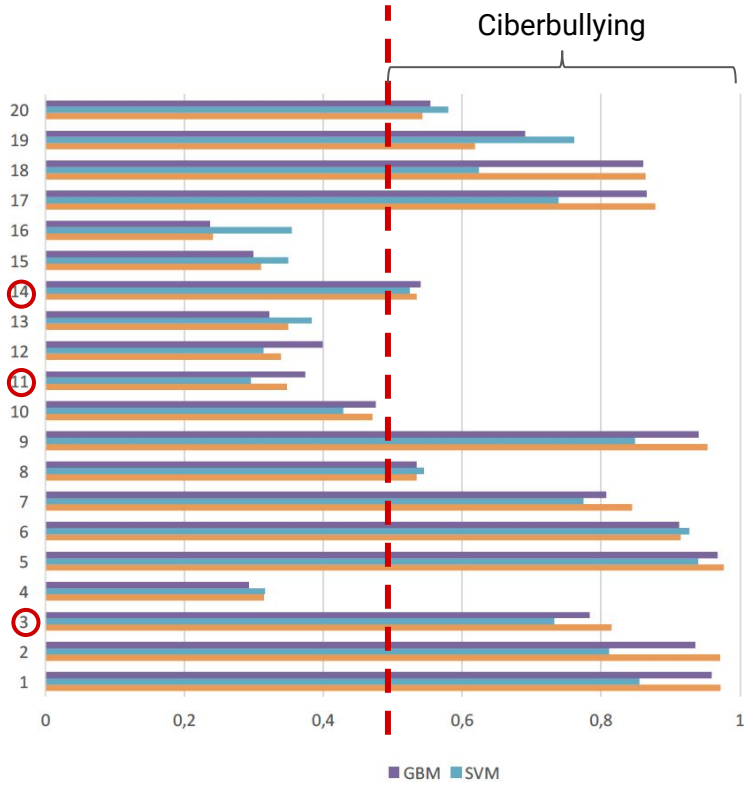
Interpretación

Primer árbol de RF de 100!

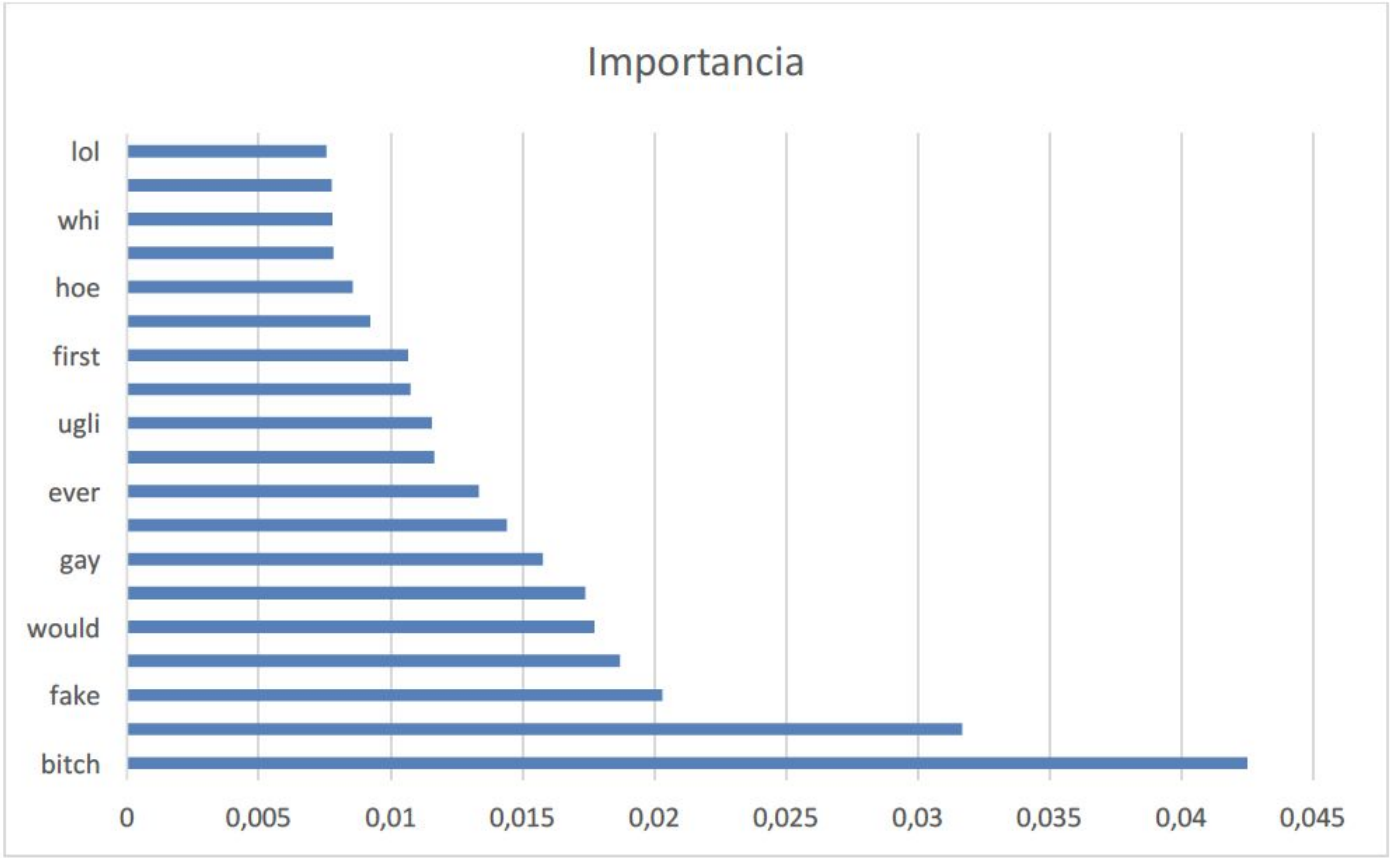


Interpretación

id	Comentario	¿Ciberbullying?
1	You're ugly and a sack of shit	Sí
2	Today I see you with your boyfriend... doesnt matter. you're still a fake bitch	Sí
3	I need to see you :)	No
4	Do you like the new movie?	No
5	Suck my big dick	Sí
6	Hello mr gay fuck	Sí
7	I want to fuck u	Sí
8	When is your brithday?	Sí
9	I now you are an asskisser, always sucking dicks	Sí
10	I'm trying help you	No
11	Do you have a favorite article of clothing?	No
12	Where do you wish you were right now?	No
13	Good look tomorrow in the exam!	No
14	You're so cute	No
15	Do you have any favorite football player?	No
16	"How would you describe yourself?	No
17	you are gay	Sí
18	Go to hell, piece of shit	Sí
19	Stop eating, fat!	Sí
20	You are nasty	Sí

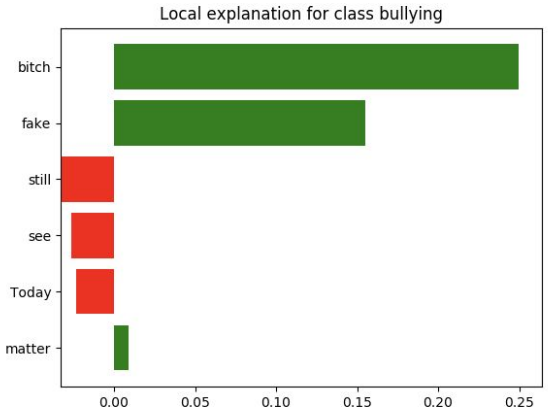


Interpretación

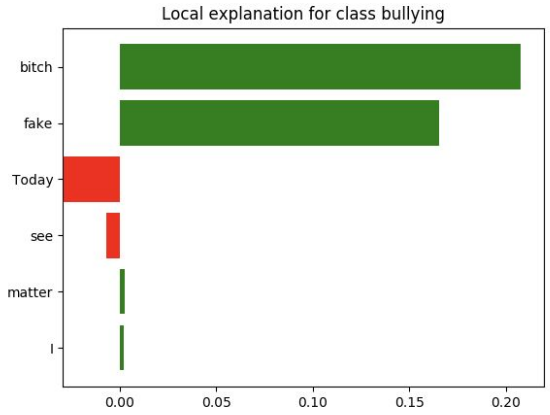


Explicación

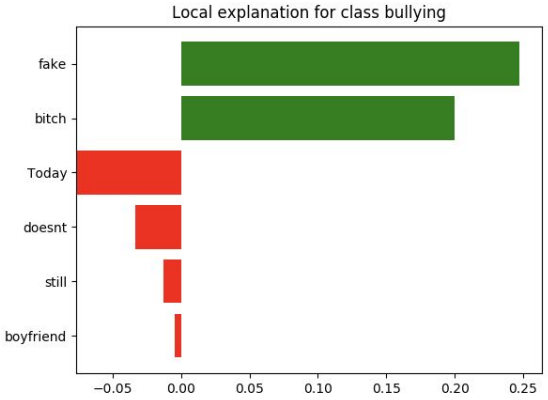
“Today I see you with your boyfriend... doesn’t matter. you're still a fake bitch.”



SVM



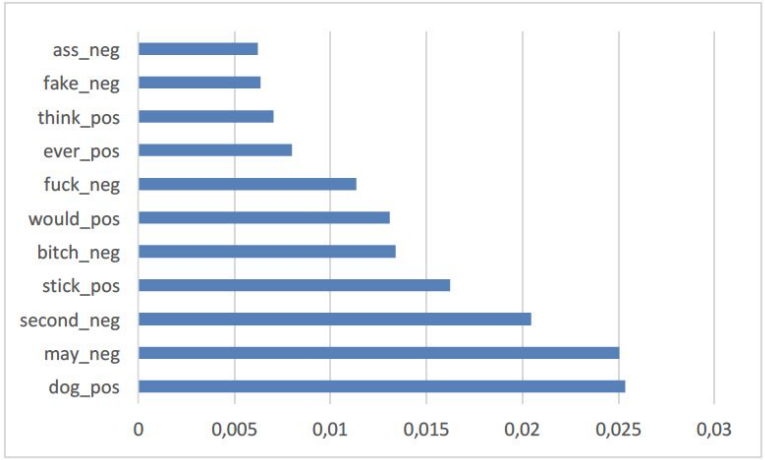
RF



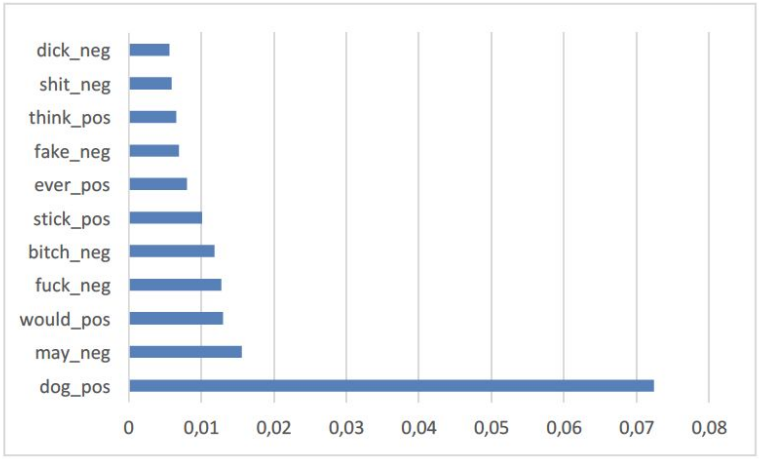
GBM

Explicación

“Today I see you with your boyfriend... doesn’t matter. you're still a fake bitch.”



RF

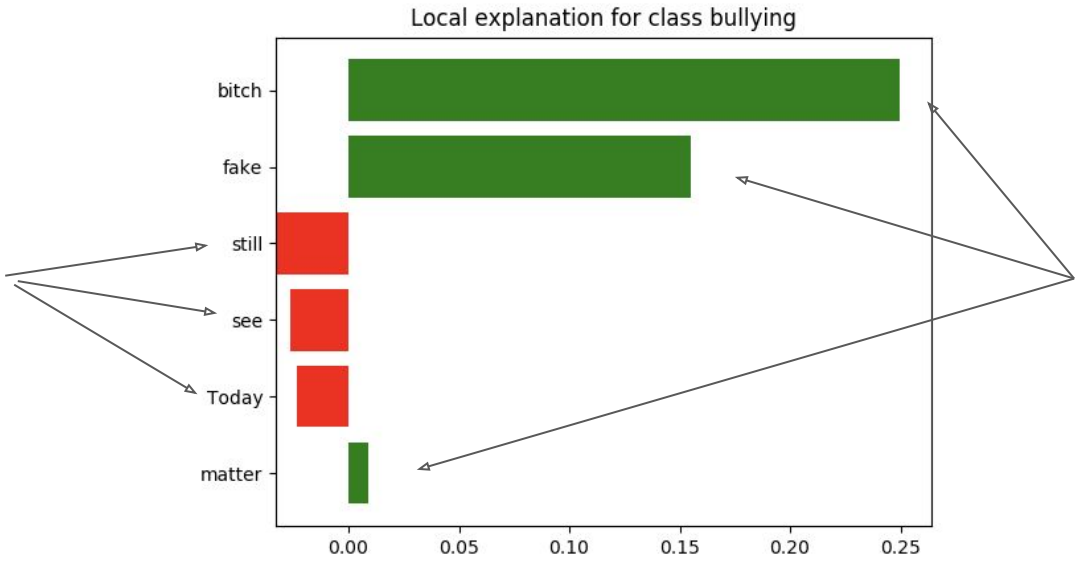


GBM

Explicación

“Today I see you with your boyfriend... doesn’t matter. you're still a fake bitch.”

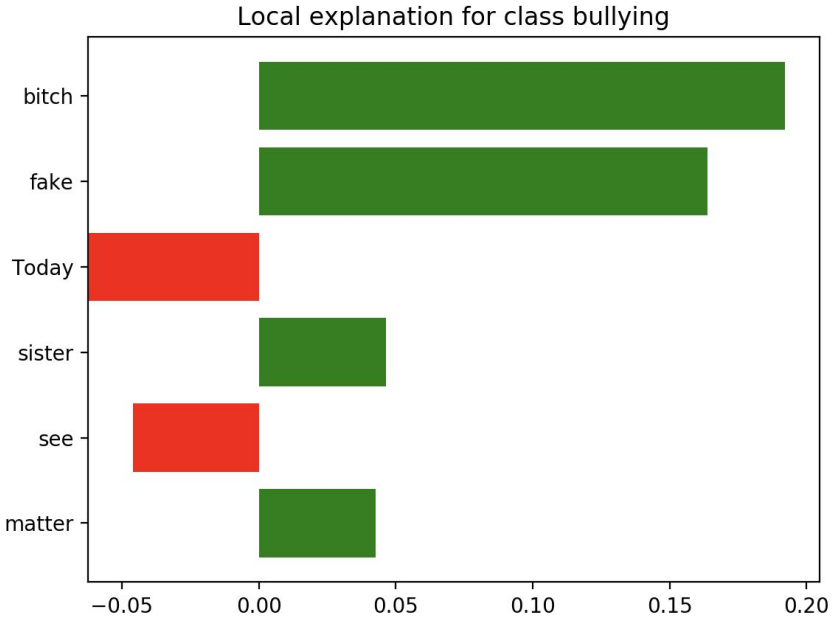
t-propiedades
(con. fácticas)



k-propiedades
(con. de principios)

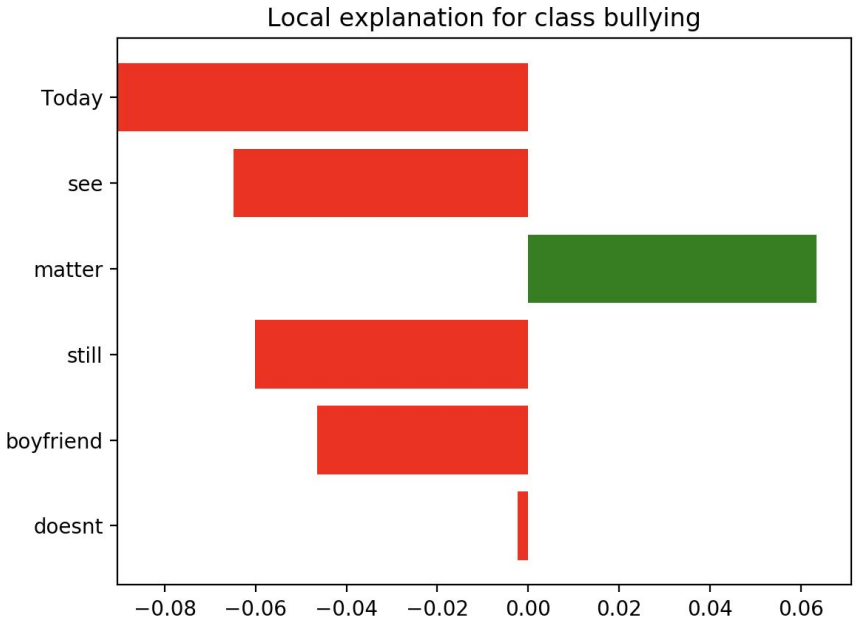
Explicación

“Today I see you with your **sister**... doesn’t matter. you're still a fake bitch.”



Explicación

“Today I see you with your boyfriend... doesn’t matter. you're still a **beautiful princess**.”



Conclusiones...

¡GRACIAS!

Sergio Jiménez Barrio