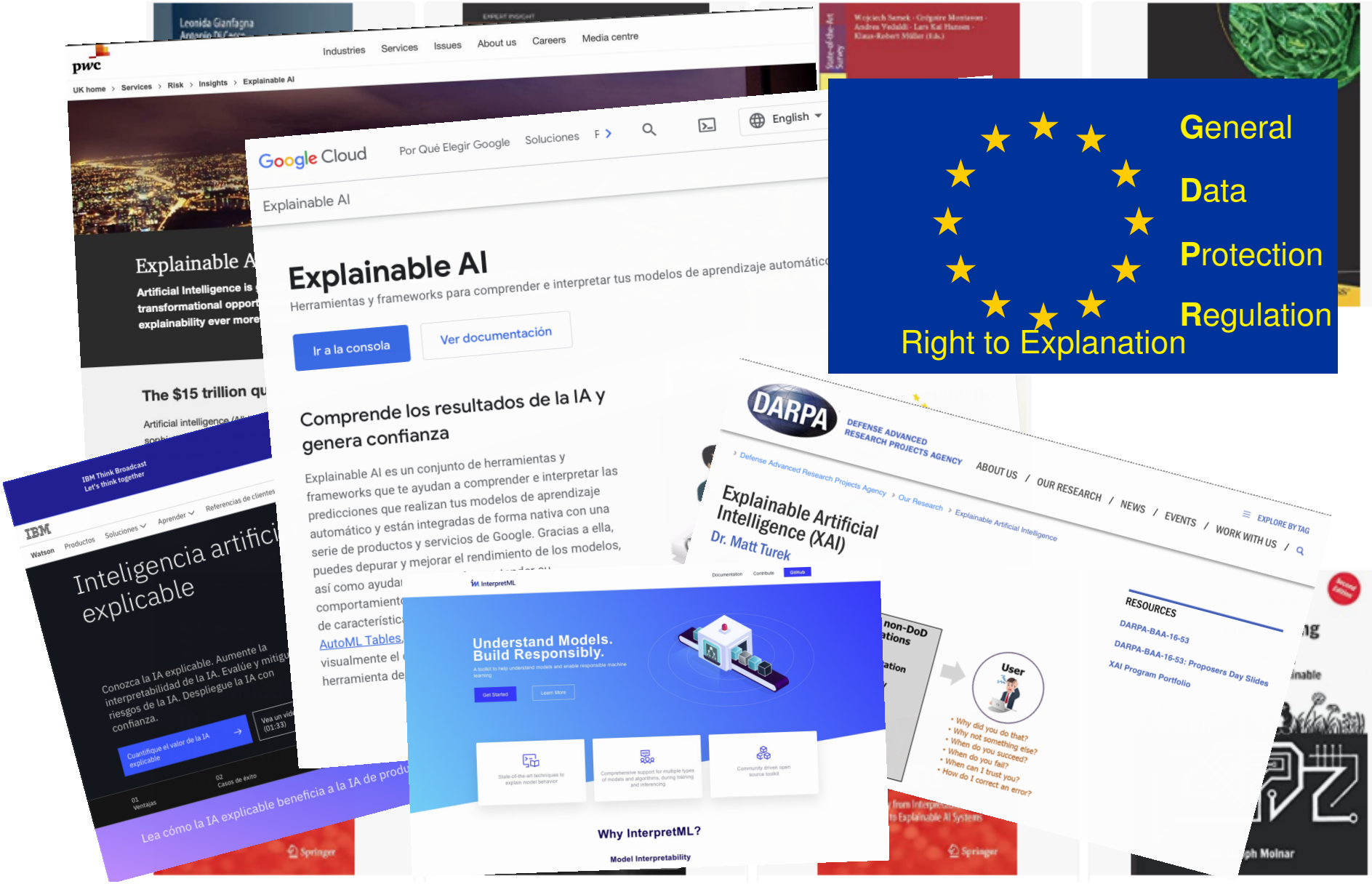
uc3m | Universidad Carlos III de Madrid

Máquinas Discriminativas Profundas (*Deep Learning*)

http://www.tsc.uc3m.es/~mlazaro/Docencia/DL.html

Marcelino Lázaro

Universidad Carlos III de Madrid


pwc
Industries Services Issues About us Careers Media centre
UK home › Services › Risk › Insights › Explainable AI
Explainable A
Artificial Intelligence is
transformational opport
explainability ever more
The $15 trillion qu
Artificial intelligence
Google Cloud
Por Qué Elegir Google Soluciones
English
Explainable AI
Explainable AI
Herramientas y frameworks para comprender e interpretar tus modelos de aprendizaje automátic
Ir a la consola
Ver documentación
Comprende los resultados de la IA y genera confianza
Explainable AI es un conjunto de herramientas y frameworks que te ayudan a comprender e interpretar las predicciones que realizan tus modelos de aprendizaje automático y están integradas de forma nativa con una serie de productos y servicios de Google. Gracias a ella, puedes depurar y mejorar el rendimiento de los modelos, así como ayudar
comportamient
de característic
AutoML Tables,
visualmente el
herramienta de
General
Data
Protection
Regulation
Right to Explanation
IBM Think Broadcast
Let's think together
IBM
Watson Productos Soluciones Aprender Referencias de clientes
Inteligencia artifici
explicable
Conozca la IA explicable. Aumente la interpretabilidad de la IA. Evalúe y mitigu riesgos de la IA. Despliegue la IA con confianza.
Cuantifique el valor de la IA explicable
Vea un vid (01:33)
01 Ventajas
02 Casos de éxito
Lea cómo la IA explicable beneficia a la IA de produ
DARPA
DEFENSE ADVANCED RESEARCH PROJECTS AGENCY
ABOUT US / OUR RESEARCH / NEWS / EVENTS / WORK WITH US /
Defense Advanced Research Projects Agency › Our Research › Explainable Artificial Intelligence
EXPLORE BY TAG
Explainable Artificial Intelligence (XAI)
Dr. Matt Turek
Documentation Contribute GitHub
InterpretML
Understand Models. Build Responsibly.
A toolkit to help understand models and enable responsible machine learning
Get Started Learn More
Why InterpretML?
Model Interpretability
User
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
RESOURCES
DARPA-BAA-16-53
DARPA-BAA-16-53: Proposers Day Slides
XAI Program Portfolio
Springer
ph Molnar

La IA en la sociedad actual

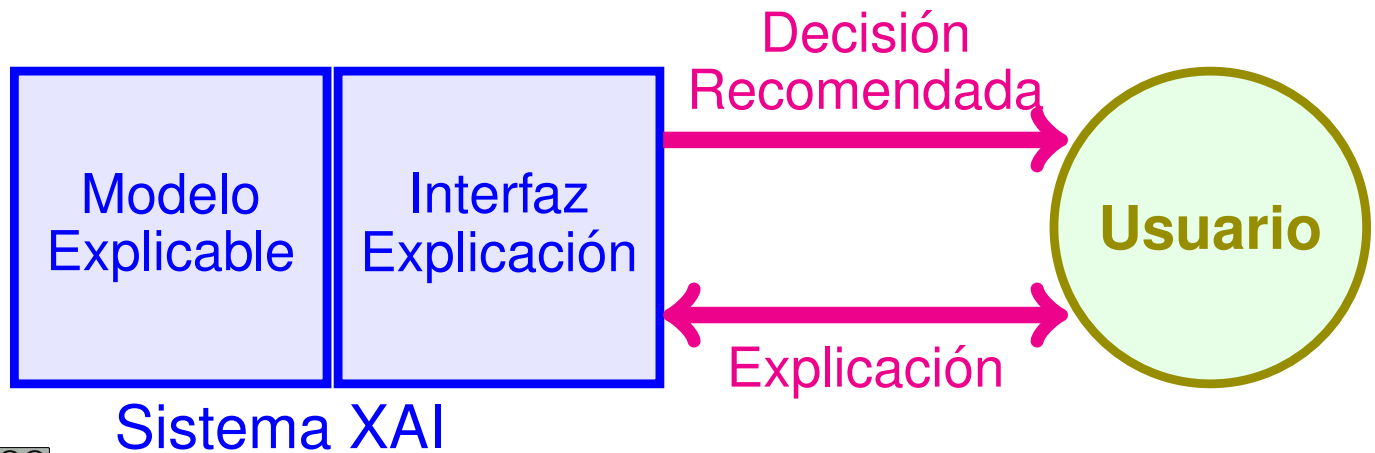
- La IA está ya teniendo un significante impacto en la sociedad
  - Inversión
    - ★ 12.000 millones de dólares en 2017
    - ★ Más de 52.000 millones de dólares en 2021
  - Beneficios
    - ★ 480.000 millones en 2017
    - ★ Más de 2.500.000 millones en 2021
    - ★ Estima de más de 15.000.000 millones en 2030
- La IA ya toma decisiones en nuestra vida diaria
  - Recomendaciones de productos (Netflix, Amazon,...)
  - Publicidad personalizada (Google, Facebook,...)
- ¿Decisiones sobre aspectos vitales para el ser humano?
  - La necesidad de entender y/o explicar las decisiones de la IA parece evidente

Contexto - El problema de la "Caja Negra"

Sistema que puede proporcionar excelentes prestaciones

Pero el usuario NO sabe cómo funciona !!!

- Respuestas a preguntas con implicaciones
  - ¿Esta máquina o equipo fallará en breve?
  - ¿Es un tumor cancerígeno lo que aparece en la imagen?
  - ¿Devolverá el cliente el préstamo?
  - ¿Reincidirá este recluso?
  - ¿Disparará el dron/robot a un potencial enemigo?



Interés en explicar las decisiones de una máquina

- Explicabilidad
  - Barrera que limita el uso del aprendizaje máquina (ML), o de la Inteligencia Artificial (AI) en general en numerosos sectores
  - Esfuerzo para definir explicabilidad en este ámbito
    - ★ Tendencia hacia la **Inteligencia Artificial Responsable**
- Decisiones de AI que afectan a los humanos (medicina, ley, defensa)
  - Necesidad de comprensión de cómo se han construido las decisiones
  - Reticencia a adoptar técnicas que no sean interpretables, transparentes y fiables
- La interpretabilidad puede reducir reticencias
  - Asegura la imparcialidad de la toma de decisiones
  - Facilita la provisión de robustez
    - ★ Resalta potenciales perturbaciones adversarias que pueden cambiar la decisión
  - Garantiza que sólo variables relevantes afectan a la decisión
    - ★ Existe una causalidad subyacente (fiable)
- Normativa
  - Regulaciones propias de algunos campos
  - UE : Regulación de las decisiones basadas en algoritmos
    - ★ En vigor desde abril de 2018
    - ★ Incluye el "**Derecho a Explicación**"

Índice de contenidos

- Introducción al problema de la explicabilidad
  - Definiciones
  - Objetivos
  - Impacto social
  - Uso en el diseño de métodos de aprendizaje
- Tipos de técnicas de explicación
  - Taxonomía general
- Algunos modelos de explicación
  - LIME (Local Interpretable Model-Agnostic Explanations)
  - SHAP (SHapley Additive exPlanations)
  - Explicaciones Contrafactuales
  - LRP (Layerwise Relevant Propagation)
  - Visualización de características (Activation Maximization, AM)
- Evaluación

Inteligencia Artificial Explicable
(XAI, *eXplainable Artificial Intelligence*)

Explicabilidad

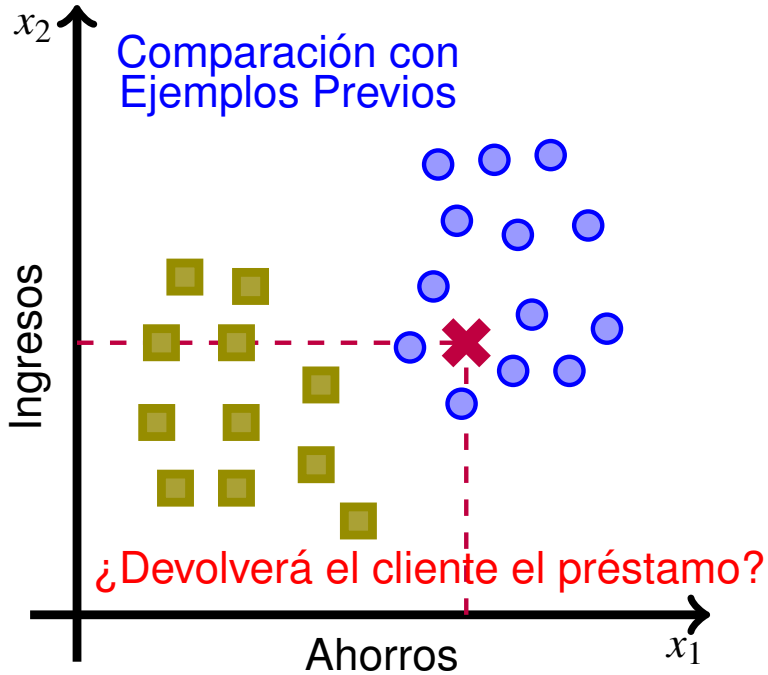
- Terminología variable y confusa en la literatura
  - Explicabilidad, inteligibilidad, comprensibilidad, interpretabilidad, transparencia,...
    - ★ Understandability, intelligibility, comprehensibility, intepretability, explainability, transparency,...
- Diferentes definiciones de Inteligencia Artificial Explicable (XAI)
  - D. Gunning: "*XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners*"
  - DARPA: "XAI aims to produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust and effectively manage the emerging generation of artificially intelligent patterns"
- Énfasis: comprensión y confianza (*understanding and trust*)
  - Impulso para la utilización de la IA en nuevos dominios

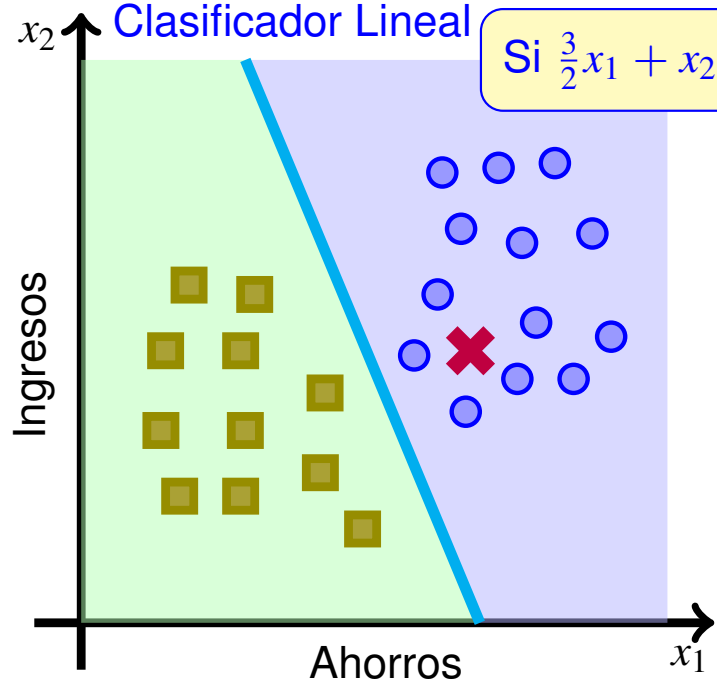
Dominios de aplicación

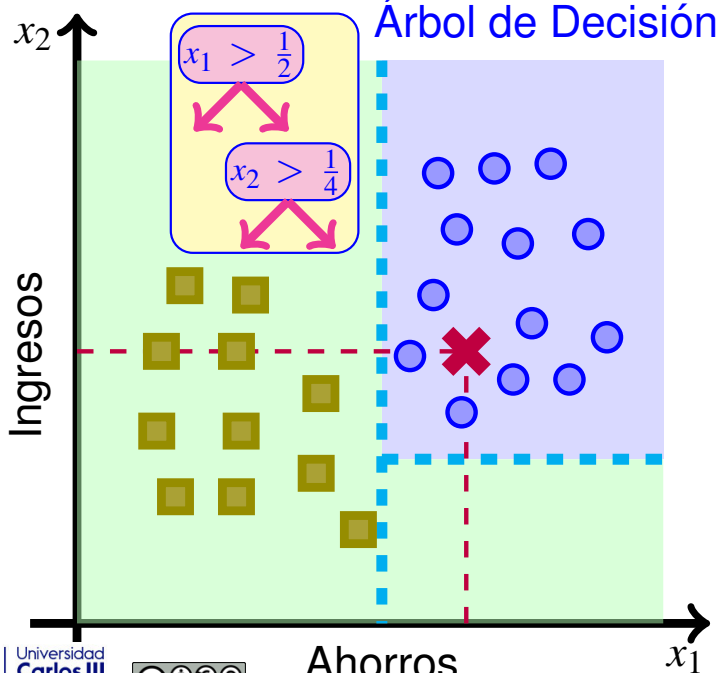
- Transporte
  - Vehículos autónomos
- Salud
  - Diagnosis médica
  - Detección de causalidad de enfermedades
    - ★ Análisis estadístico (hábitos de vida, consumo, etc.)
    - ★ Análisis genético
- Judicial
  - Análisis de riesgo de reincidencia
  - Costes asociados con crimen y encarcelación
- Económico
  - Múltiples aplicaciones
    - ★ Credit Scoring, detección de fraude, rotación de clientes, etc.
- Militar
  - Inicio e impulso de la XAI (Proyectos DARPA)
  - Sistemas autónomos para operaciones militares
    - ★ Dilemas éticos y legales
  - Otras aplicaciones
    - ★ Ciberseguridad, reconocimiento de imágenes (inteligencia), formación, asistencia en la toma de decisiones, etc.

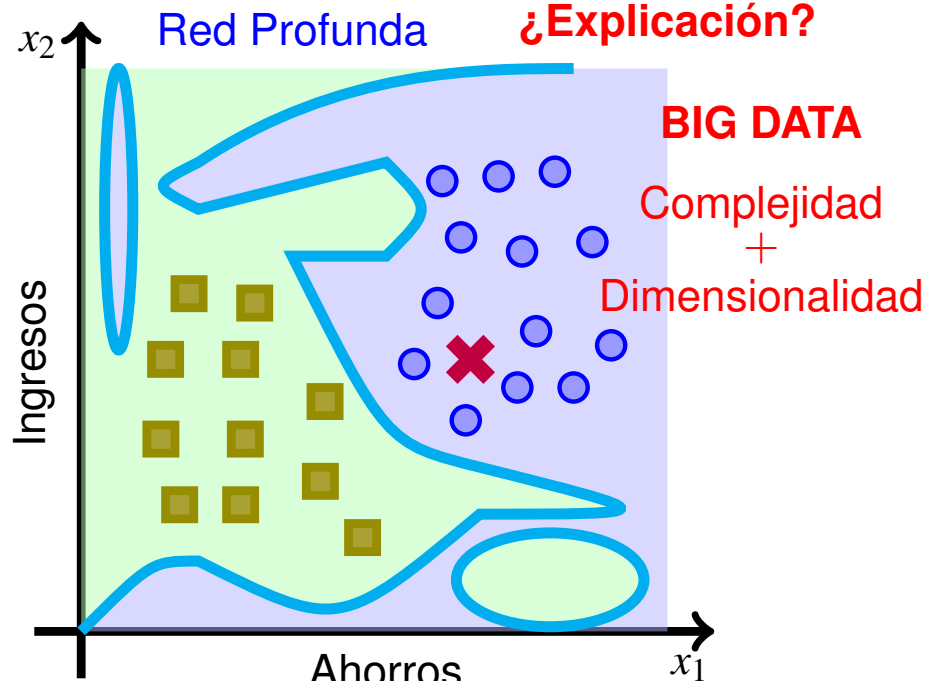
Dificultades para explicar el aprendizaje máquina

- Gran variedad de arquitecturas y reglas de decisión
  - Clasificador lineal, árboles de decisión
    - ★ Reglas de decisión claras y fáciles de interpretar
  - **Red neuronal profunda**
    - ★ Modelo extremo de tipo "*caja negra*" : difícil de interpretar
    - ★ Multiplicidad de buenos modelos (para los mismos datos)
- Concepto dependiente de la audiencia
  - Hacer claro o fácil de entender el funcionamiento de una máquina a una determinada audiencia
- Compromiso entre prestaciones y explicabilidad
  - Tradicionalmente considerado estático
    - ★ Los modelos más precisos usualmente no son fácilmente explicables
  - Tendencia de investigación: objetivo dinámico

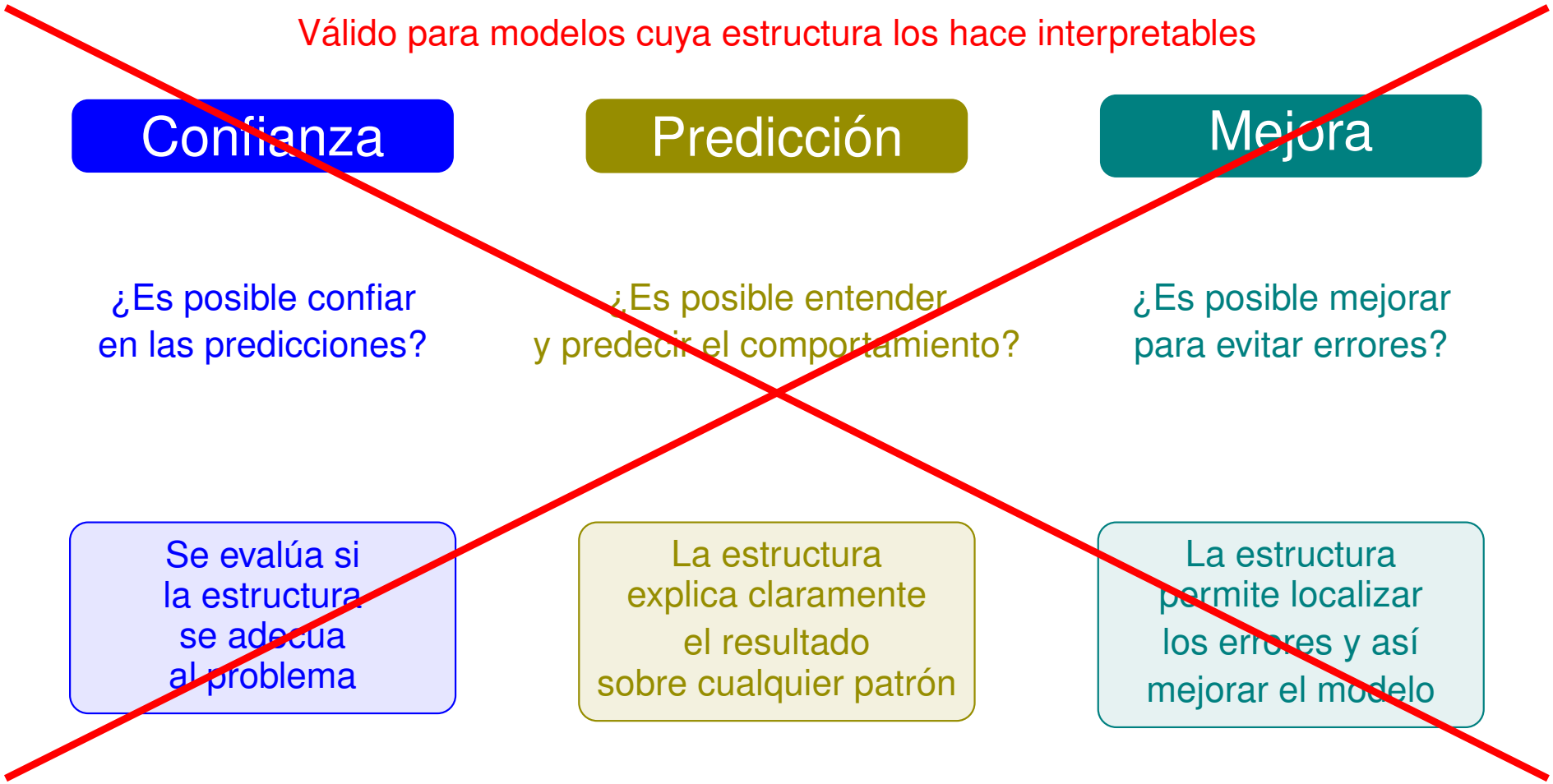

x_2
Comparación con
Ejemplos Previos
Ingresos
¿Devolverá el cliente el préstamo?
Ahorros
x_1


x_2
Clasificador Lineal
Si $\frac{3}{2}x_1 + x_2 \leqslant 2$
Ingresos
Ahorros
x_1


x_2
Árbol de Decisión
$x_1 > \frac{1}{2}$
$x_2 > \frac{1}{4}$
Ingresos
Ahorros
x_1


x_2
Red Profunda
¿Explicación?
BIG DATA
Complejidad
+
Dimensionalidad
Ingresos
Ahorros
x_1

Estructura del Clasificador



BIG DATA: Complejidad + Dimensiones

Ingresos

Ahorros

Edad

Estado Civil

Cantidad Préstamo

Profesión

Formación Académica

Préstamos Anteriores

Hijos

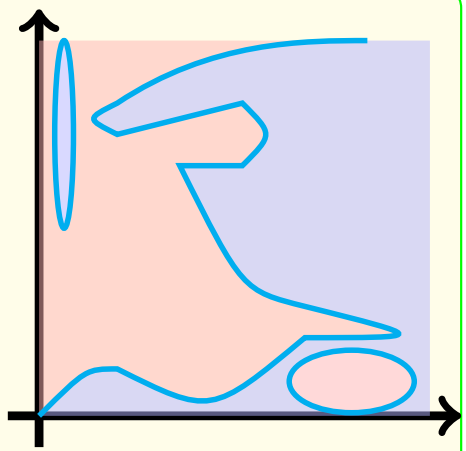
· · ·

Otras aplicaciones:

- ⋆ Texto: cientos de miles
- ⋆ Imágenes: miles/millones

ALTA DIMENSIONALIDAD

+



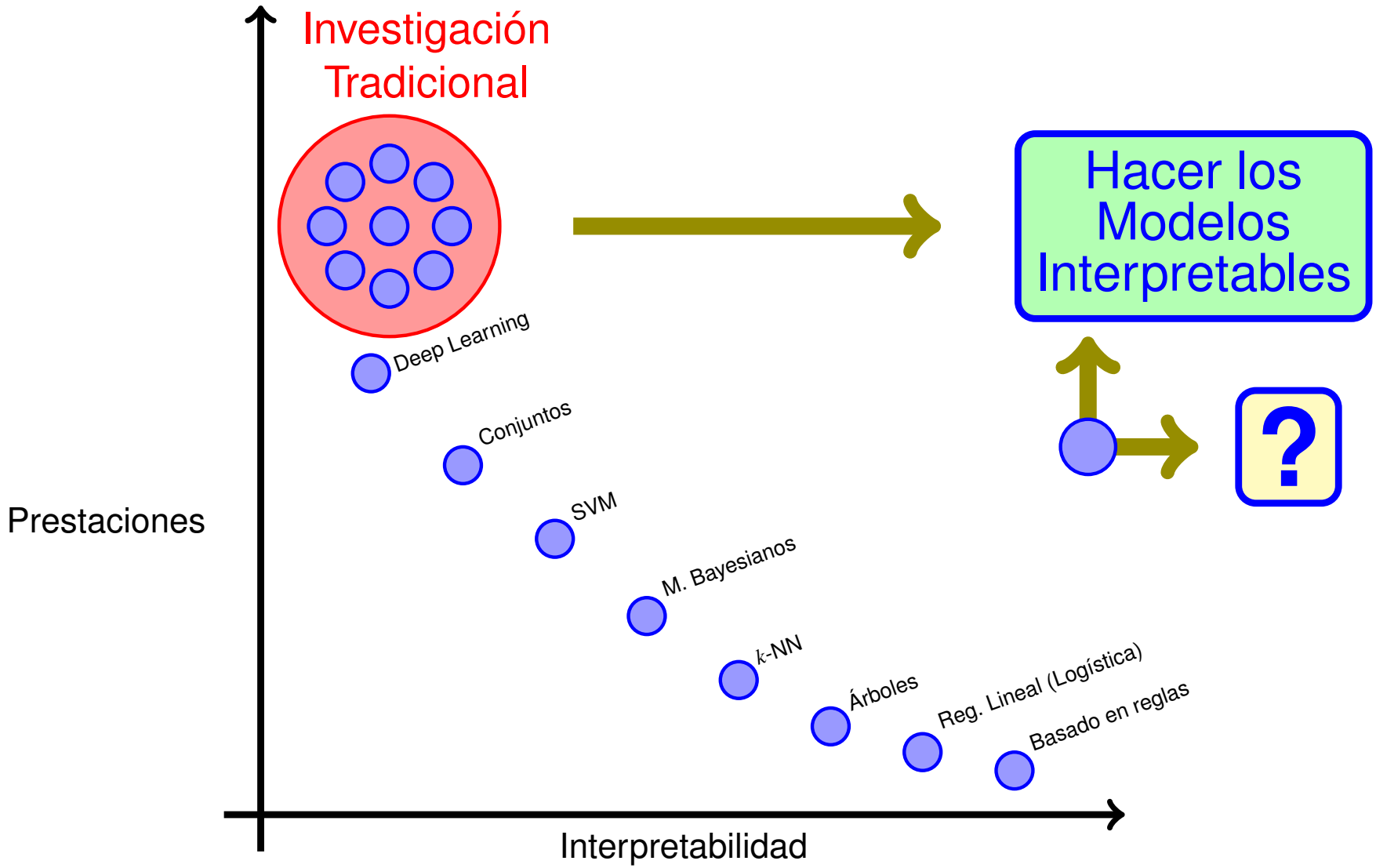
FRONTERAS COMPLEJAS

=

Caja Negra

Explaining Black-Box Machine Learning Predictions (Sameer Singh)
https://www.youtube.com/watch?v=TBJqgvXYhfo

Prestaciones vs Interpretabilidad



Objetivos de la XAI

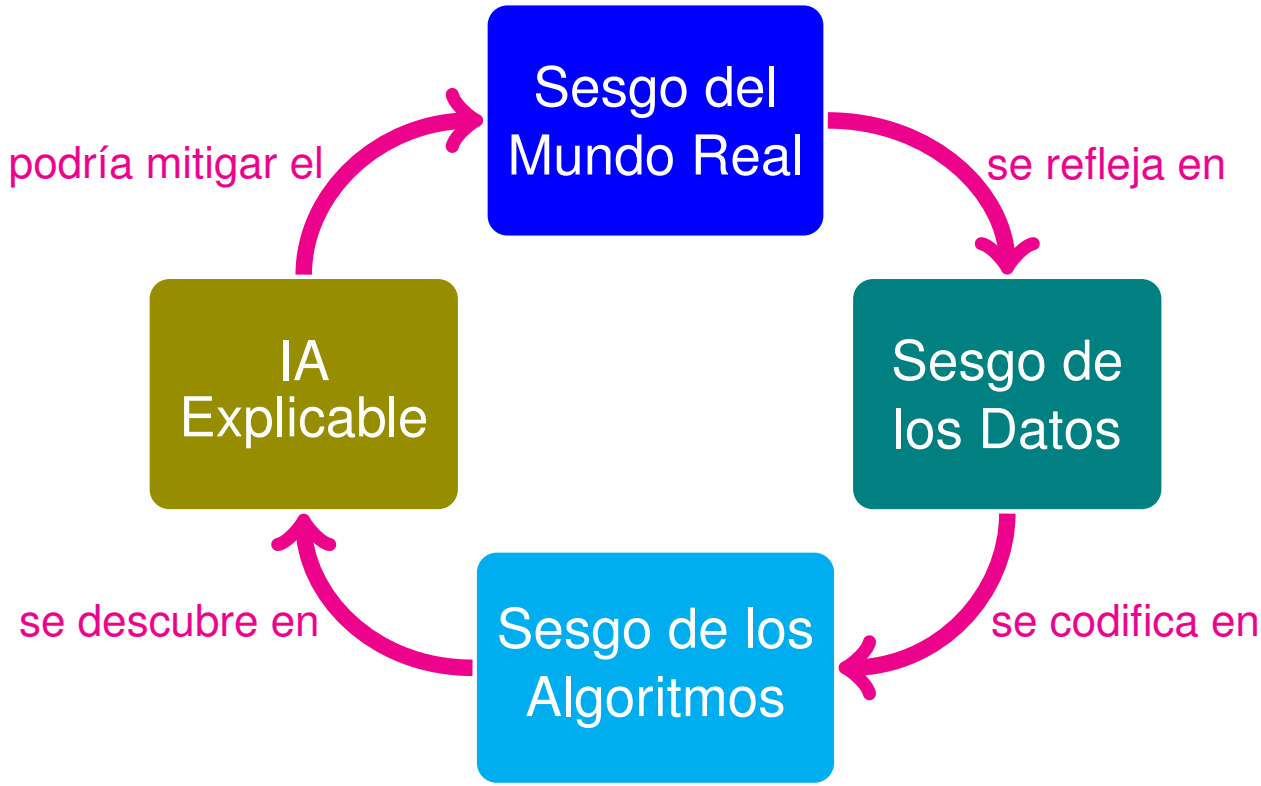
- Inteligibilidad (*Understandability*)
  - La función subyacente de un modelo IA debe ser entendible para los humanos
    - ★ Ejemplo - Autocodificador: se entiende su función (aunque no se conozca el detalle de cómo se codifican y decodifican las entradas)
- Comprensibilidad (*Comprehensibility*)
  - Representación del conocimiento aprendido de forma comprensible para los humanos
    - ★ Ejemplo: análisis de factores
- Interpretabilidad (*Interpretability*)
  - También denominada 'Explicabilidad"
  - Capacidad para explicar las predicciones del modelo
    - ★ Concepto subjetivo
    - ★ Depende del contexto y de la audiencia
- Transparencia (*Transparency*)
  - Estructura y algoritmo comprensibles
    - ★ ¿Por qué el modelo funciona del modo en el que lo hace?
  - Distintos grados de comprensión para un modelo

Impacto social

- Seguridad (*Safety*)
  - Prevenir decisiones (maliciosas o no) que dañen a los humanos
    - ★ Ejemplos: conducción autónoma, tratamientos oncológicos, etc.
    - ★ Identificación de estados de fallo en el sistema
- Verificabilidad (*Verifiability*)
  - Garantizar que se cumplen ciertas propiedades
    - ★ Técnicas que garantizan que un modelo de IA es correcto
    - ★ Cómputo de cotas para la salida del modelo (resistencia y seguridad)
- Responsabilidad (*Accountability*)
  - Reconocer y atribuir la responsabilidad en la toma de decisiones
    - ★ Aspecto básico para generación de confianza
    - ★ Relacionada con la transparencia del modelo
    - ★ Las organizaciones son responsables del funcionamiento de sus aplicaciones de IA
- Igualdad y Sesgo (*Fairness and Bias*)
  - Capacidad para descubrir algoritmos no igualitarios o no éticos
    - ★ Algoritmos pueden implicitamente codificar prejuicios de género, religiosos, raciales, etc.

XAI: facilita la adopción de aplicaciones en la sociedad

Sesgo en IA : *Fairness*



NY Times : https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

Glenn Rodríguez: libertad condicional denegada

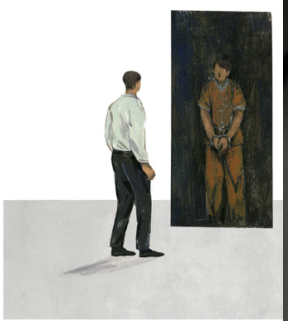
COMPAS: Correctional Offender Management Profiling for Alternative Sanctions
- ⋆ Programa propietario (Northpointe)
- ⋆ Error tipográfico: “Risk Score” erróneo

The New York Times

When a Computer Program Keeps You in Jail

By Rebecca Wexler
June 13, 2017

Give this article 230



Sally Deng

ProPublica

Two Petty Theft Arrests

VERNON PRATER	BRISHA BORDEN
Prior Offenses: 2 armed robberies, 1 attempted armed robbery	Prior Offenses: 4 juvenile misdemeanors
Subsequent Offenses: 1 grand theft	Subsequent Offenses: None
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests

DYLAN FUGETT	BERNARD PARKER
Prior Offense: 1 attempted burglary	Prior Offense: 1 resisting arrest without violence
Subsequent Offenses: 3 drug possessions	Subsequent Offenses: None
LOW RISK 3	HIGH RISK 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Two Shoplifting Arrests

JAMES RIVELLI	ROBERT CANNON
Prior Offenses: 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking	Prior Offense: 1 petty theft
Subsequent Offenses: 1 grand theft	Subsequent Offenses: None
LOW RISK 3	MEDIUM RISK 6

After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted $1,000 worth of tools from a Home Depot.

Two DUI Arrests

GREGORY LUGO	MALLORY WILLIAMS
Prior Offenses: 3 DUIs, 1 battery	Prior Offenses: 2 misdemeanors
Subsequent Offenses: 1 domestic violence battery	Subsequent Offenses: None
LOW RISK 1	MEDIUM RISK 6

Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

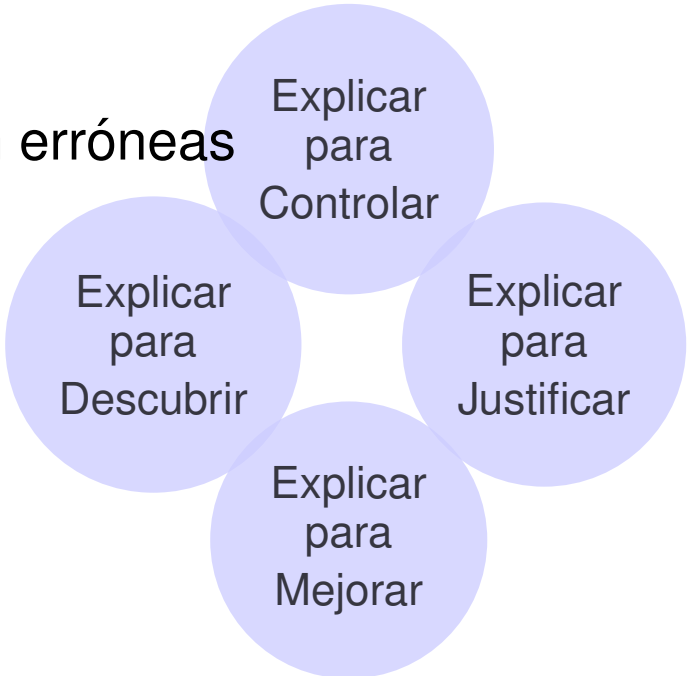
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

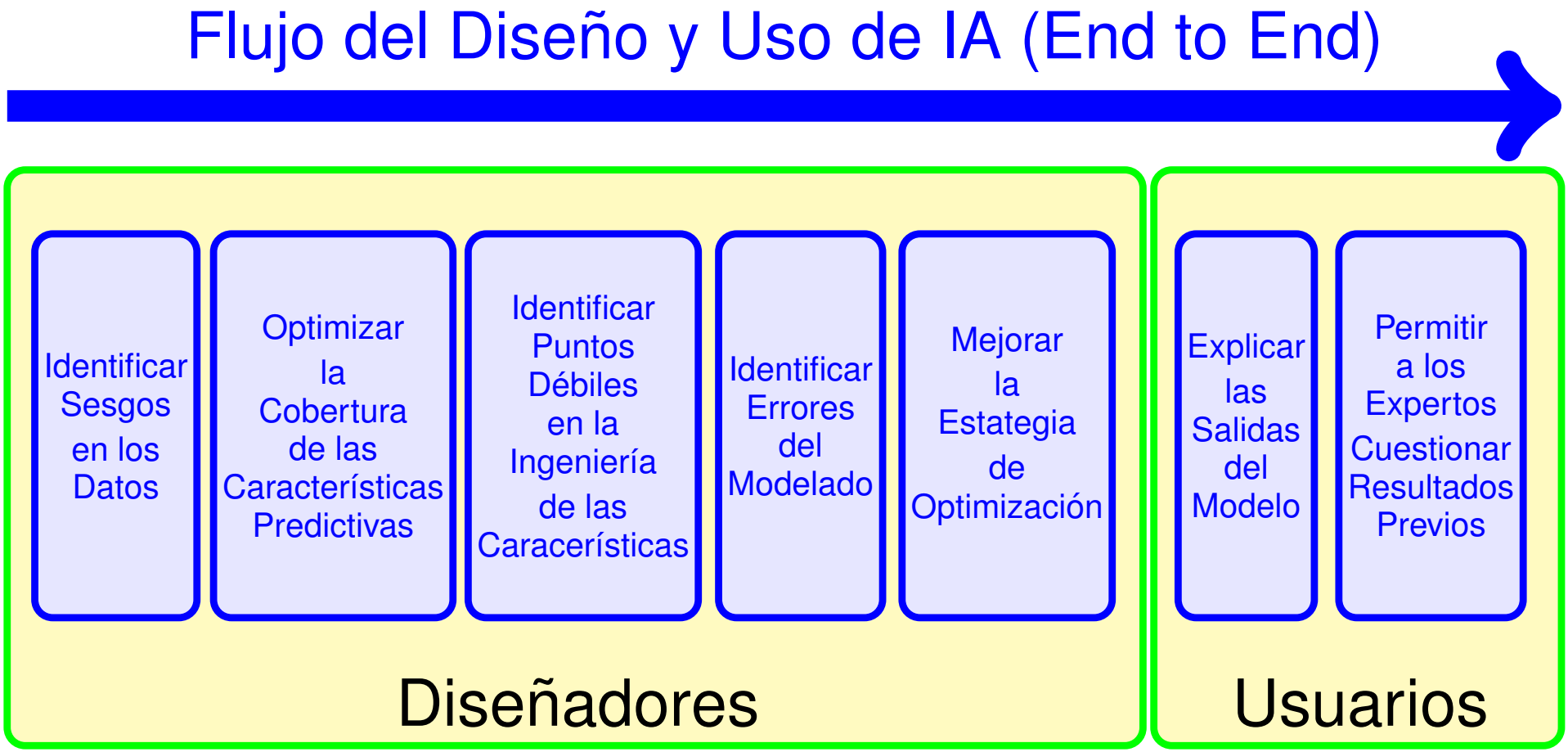
ProPublica : https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Necesidad de la explicación

- Explicar para Justificar
  - Garantía de que las decisiones no son erróneas
    - ★ Evitar sesgos o discriminaciones
  - Cumplimiento de la legislación
    - ★ Derecho a la explicación (GDPR)
- Explicar para Controlar
  - Evitar funcionamientos indeseados
    - ★ Detección de vulnerabilidades
    - ★ *Debugging* para mejora de control
- Explicar para Mejorar
  - Sostenimiento de un proceso de mejora continua
    - ★ Fundamento para establecer iteración hombre-máquina
- Explicar para Descubrir
  - Ayuda para descubrir nuevos hechos o situaciones
    - ★ Ejemplo: AlphaGo Zero: explicación de estrategias
    - ★ Futuro: ¿Nuevas estrategias en biología, química o física?

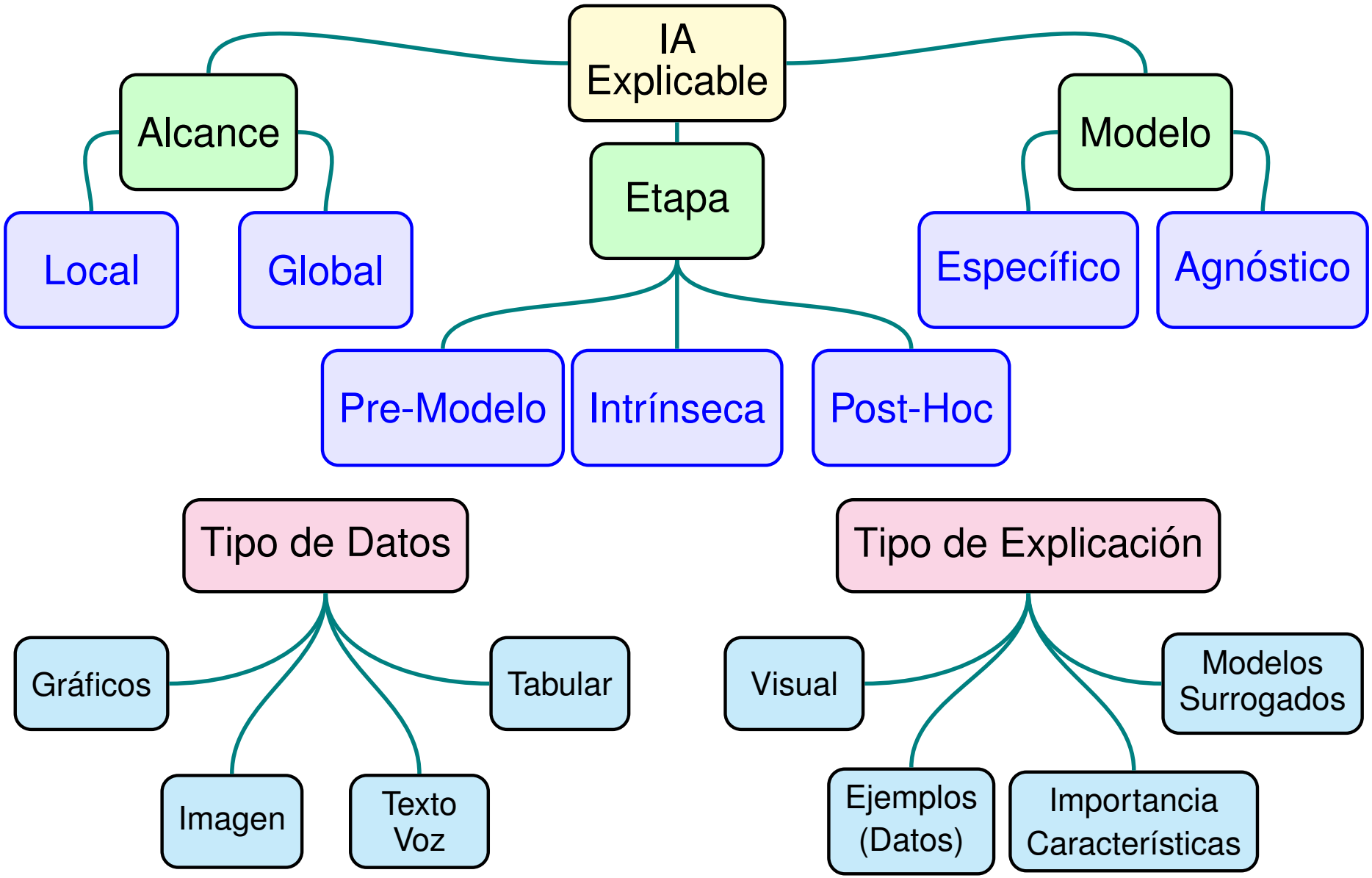


La XAI en el diseño y uso de modelos de aprendizaje

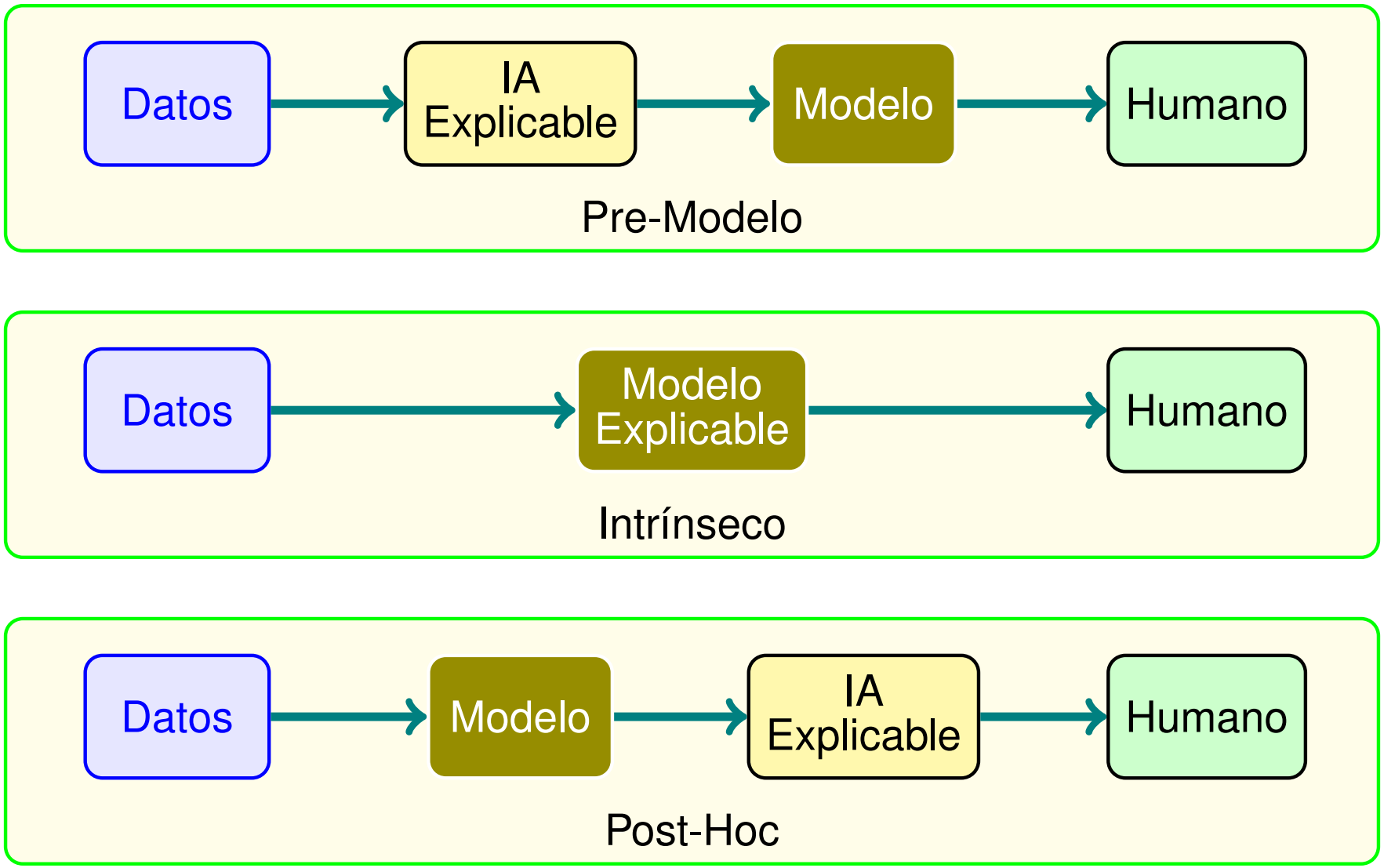


Tipos de técnicas de explicación (XAI, *eXplainable Artificial Intelligence*)

Taxonomías



XAI: categorías por etapa



XAI: categorías por etapa

- Técnicas Pre-Modelo
  - Independientes del modelo
  - Se aplican a los datos
    - ★ Técnicas de visualización de datos
  - Se suelen emplear como paso previo a la selección del modelo
  - Propiedades deseables:
    - ★ Descubrimiento de características significativas intuitivas
    - ★ Reducción del número de características (*sparsity*)
- Técnicas Intrínsecas
  - Uso de modelos autoexplicables
    - ★ La estructura del modelo facilita la explicación
      - Lineales (generalizados), árboles, agrupamiento (*clustering*),EBMs,...
  - Generalmente penalizan las prestaciones
- Técnicas Post-Hoc
  - Conjunto de técnicas aplicables a modelos tipo “Caja Negra”
    - ★ Modelos con elevadas prestaciones
  - Varias aproximaciones metodológicas
    - ★ Relevancia visual (características), modelos surrogados, basados en ejemplos,...

XAI: categorías por alcance

- Métodos Globales
  - Explicación de las predicciones del modelo “Top-Down”
    - ★ Cómo la estructura y parámetros conducen a las predicciones
    - ★ Mapeo “entrada-características” + mapeo “características-salida”
    - ★ Interesante en aplicaciones donde el efecto global es más relevante que las explicaciones para las múltiples idiosincradias posibles
      - Ej: Hábitos de consumo de drogas, tendencias climáticas, etc.
  - Diferentes aproximaciones
    - ★ Maximización de activación: síntesis de entradas preferidas
      - Varios niveles: neurona, capa, etc.
    - ★ Particionamiento recursivo
      - Construcción de un árbol interpretativo global
- Métodos Locales
  - Explica cómo un patrón específico se mapea en su salida
    - ★ Comprensión de cómo el modelo llegó a su predicción
    - ★ Contribución de las características para la predicción
  - Aproximación del modelo en la región de interés usando un modelo más simple
    - ★ En muchos casos modelos lineales
    - ★ Imagen: píxeles que contribuyen a la predicción
    - ★ Gradientes se utilizan de varios modos
      - Asignación de importancia a las variables
      - Cómo modificar un patrón para cambiar la decisión

XAI: categorías por modelo

- Técnicas de modelo específico
  - Aplicables sólo a un modelo específico
    - ★ Ejemplos:
      - *Tree SHAP* para árboles de decisión
      - *Grad-CAM* (*Gradient-weighted Class Activation Mapping*) para CNNs
  - Se explotan las características específicas de la arquitectura para generar la explicación
    - ★ Potencial ventaja en la explicación para esas arquitecturas
- Técnicas agnósticas sobre el modelo
  - Aplicables a cualquier tipo de modelo
    - ★ Se aplican asumiendo modelos “Caja Negra”
    - ★ Información: patrones de entrada y decisiones del modelo
  - Habitualmente usan modelos surrogados o aproximaciones
    - ★ Riesgo de degradación en la información proporcionada

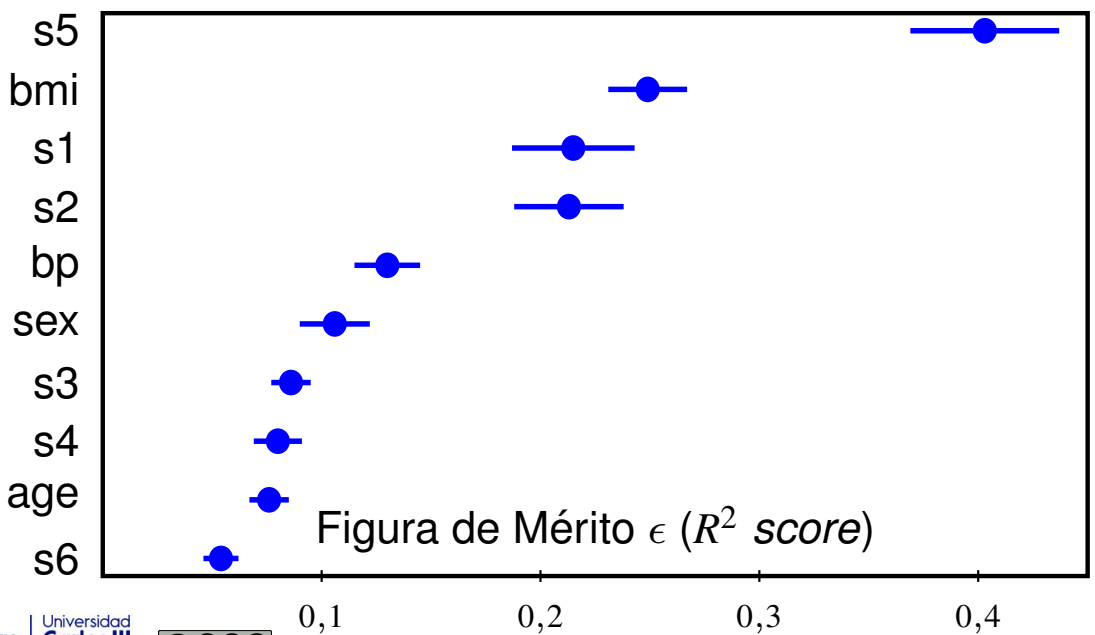
Tipos de explicación

- Explicación Global
  - ¿Cómo funciona el modelo?
- Explicación Local
  - ¿Por qué llega a esa predicción para esa entrada?
- Explicación Contrastiva
  - ¿Por qué X y no Y?
- Explicación tipo *What-If*
  - ¿Cuáles son las relaciones entre una predicción y las características?
- Explicación Contrafactual
  - ¿Cómo llegar a una predicción dada (deseada)?
- Explicación basada en ejemplos
  - Resalta ejemplos particulares de interés entre los datos

Algunos modelos de explicación

PFI (*Permutation Feature Importance*)

- Medida de la importancia relativa de las características
  - Diferencia en figura de mérito si se permutan los valores de una característica
    - ★ Figuras de mérito: MSE, R^2 *score* (coef. correlación), P_e, AUC,...
- Procedimiento para el cálculo de la importancia de una característica
  - Estima de figura de mérito original: ϵ_{orig}
  - Para cada característica:
    - ★ Permutación aleatoria de valores de esa característica sobre los patrones
    - ★ Medida de la nueva figura de mérito: ϵ_{perm}
    - ★ Importancia de la característica i : $FI_i = \epsilon_{perm} - \epsilon_{orig}$ (o $FI_i = \frac{\epsilon_{perm}}{\epsilon_{orig}}$)
    - ★ Se promedian varias realizaciones
  - Se ordenan las características por orden decreciente de FI_i



Diabetes Dataset

age : age in years
sex : gender (male/female)
bmi : body mass index
bp : average blood pressure
s1 : tc, total serum cholesterol
s2 : ldl, low-density lipoproteins
s3 : hdl, high-density lipoproteins
s4 : tch, total cholesterol / HDL
s5 : ltg, possibly log of serum triglycerides level
s6 : glu, blood sugar level

Target : quantitative measure of disease progression (one year after baseline)

PDPs (*Partial Dependence Plots*)

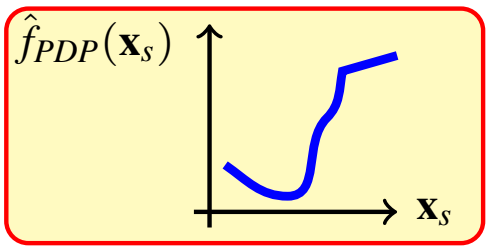
- Método global y agnóstico
  - Representa la contribución de características individuales en el modelo
- Formulación matemática
  - $\mathbf{x}_s$: características de interés
  - $\mathbf{x}_c$: vector con el resto de características

$$f_{PDP}(\mathbf{x}_s) = E_{\mathbf{X}_c}\left[f(\mathbf{x}_s, \mathbf{x}_c)\right] = \int_{\mathbf{X}_c} f(\mathbf{x}_s, \mathbf{x}_c)\, f_{\mathbf{X}_c}(\mathbf{x}_c)\, d\mathbf{x}_c$$

$f(\mathbf{x}_s, \mathbf{x}_c)$: salida de la red

- Estima muestral
  - Se reemplaza el valor de una característica por los valores dentro de su rango
  - Se promedian los valores de la predicción sobre los datos disponibles

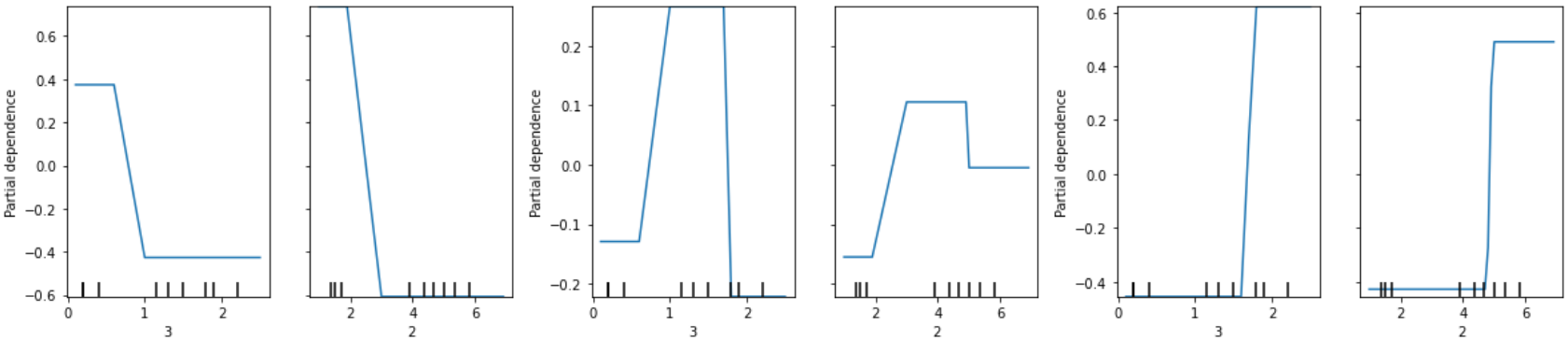
$$\hat{f}_{PDP}(\mathbf{x}_s) = \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}_s, \mathbf{x}_c^{(n)})$$

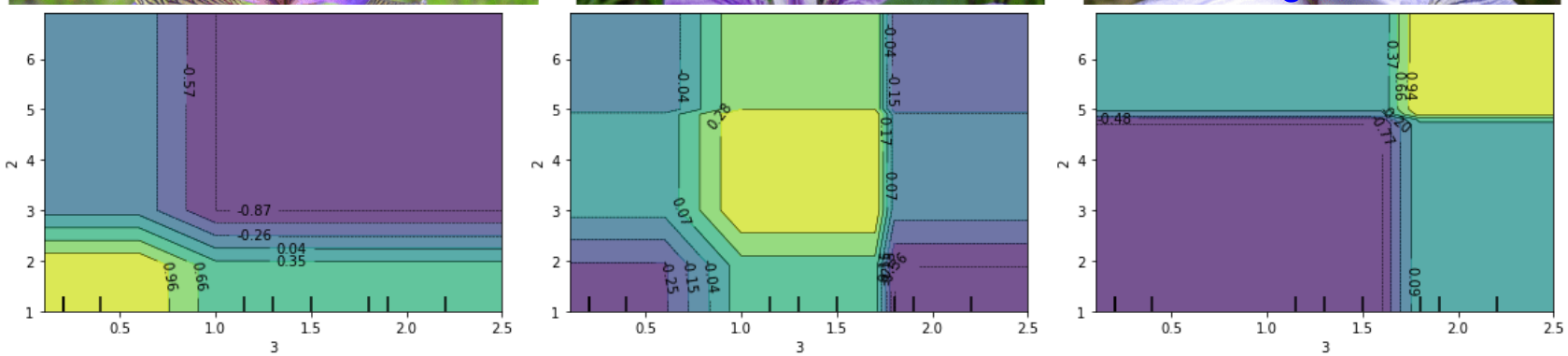


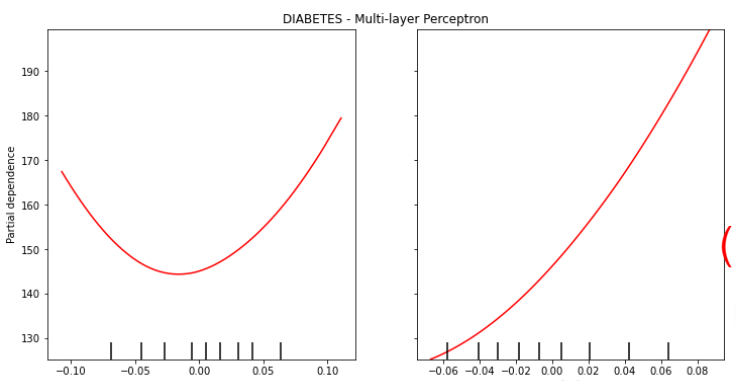
- Ventajas
  - Representación sencilla, fácil de entender y de calcular
- Inconvenientes
  - Información sobre características individuales
  - Capacidad de visualización conjunta limitada a 2 características

PDPs : Ejemplos

IRIS X=[longitud del sépalo, ancho del sépalo, longitud del pétalo, ancho del pétalo]





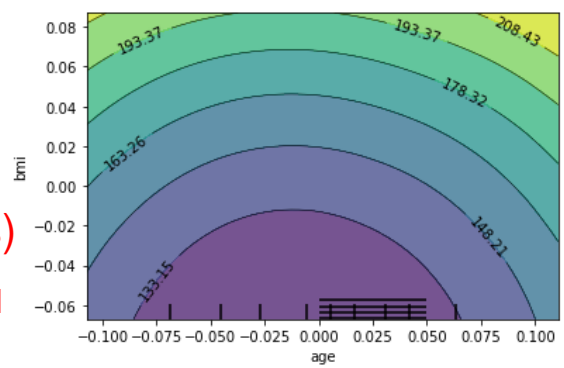


Diabetes Dataset

(Evolución en 1 año)

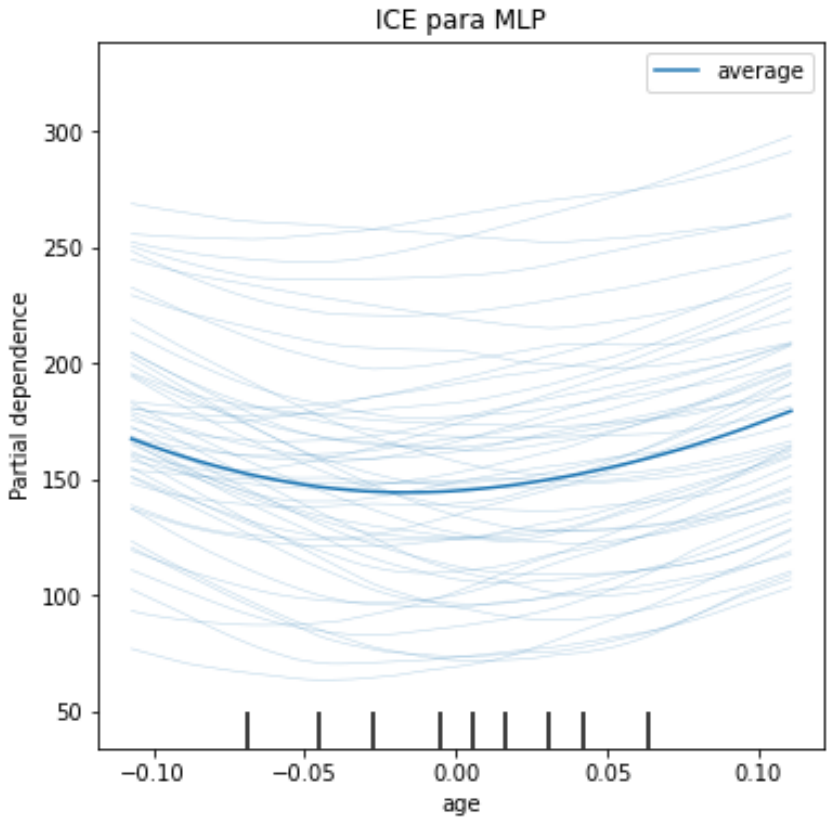
(10 variables fisiológicas)

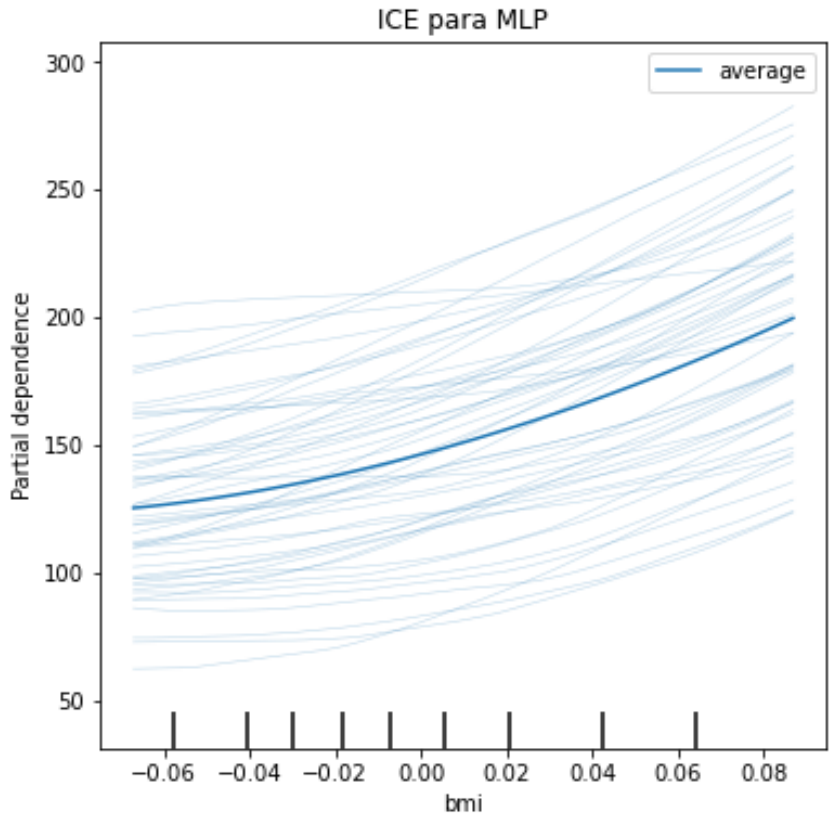
Edad / Índice de Masa Corporal



ICEs (*Individual Conditional Expectation*)

- Equivalente a PDPs para patrones individuales
  - Un PDP es el promedio de las líneas representadas en un ICE





SHAP (*SHapley Additive exPlanations*)

- Modelo local y agnóstico
  - Permite explicar cualquier tipo de modelo
  - Explicaciones locales sobre la decisión para un cierto patrón
  - Se han realizado particularizaciones a modelos específicos
    - ★ Tree SHAP (árboles), Deep SHAP,...
- Basado en el concepto de “Valores de Shapley” (Teoría de Juegos)
  - Distribución del efecto marginal de un jugador en una coalición
    - ★ Propuesto por Lloyd Shapley (Premio Nobel de Economía)
- Adaptación a explicabilidad de métodos de aprendizaje
  - Estima del valor de Shapley para cada característica
    - ★ Se considera la predicción (*score*) como la ganancia de un juego cooperativo
    - ★ Contribuciones: ponderación de la diferencia en la predicción cuando faltan esas características
    - ★ Se tienen en cuenta las distintas permutaciones
- Inconveniente intrínseco
  - Complejidad exponencial con el número de características
    - ★ En la mayoría de problemas prácticos no es posible la implementación completa
    - ★ Se recurre a aproximaciones
      - Se consideran subconjuntos de las posibles permutaciones

SHAP - Valores de Shapley

Coalición	Puntos
A	1000
B	1200
C	1800
AB	2500
AC	2100
BC	2000
ABC	3500

Permutaciones	A	B	C
ABC	1000	AB-A = 1500	ABC-AB = 1000
ACB	1000	ABC-AC = 1400	AC-A = 1100
BAC	AB-B = 1300	1200	ABC-AB = 1000
BCA	ABC-BC = 1500	1200	BC-B = 800
CAB	AC-C = 300	ABC-AC = 1400	1800
CBA	ABC-BC = 1500	BC-C = 200	1800
Promedios	1100	1150	1250

- Contribución de A, B y C en la coalición ABC
  - No es igual a las puntuaciones cuando operan individualmente
    - Las interacciones son más complejas (1000 + 1200 + 1800 $\neq$ 3500)
- Contribuciones Marginales (Coalición ABC)
  - Contribución de A
    - 1000 puntos
  - Contribución de B
    - Contribución de AB (2500) - Contribución de A (1000) = 1500 puntos ($\neq$ 1200)
  - Contribución de C
    - Contribución de ABC (3500) - Contribución de AB (2500) = 1000 puntos ($\neq$ 1800)
- El orden importa !!!
  - Ejemplo: ABC vs ACB (Contribución de A = 1000 puntos)
  - Contribución de C : Contribución de AC (2100) - Contribución de A (1000) = 1100 ($\neq$ 1000)
  - Contribución de B : Contribución de ABC (3500) - Contribución de AC (2100) = 1400 ($\neq$ 1500)

SHAP - Cálculo de los valores de Shapley

$$\phi_i(f) = \sum_{z \subseteq \{1,2,\cdots,M\} \setminus \{i\}} \frac{|z|!(M - |z| - 1)!}{M!} [f(z \cup i) - f(z)]$$

$\phi_i(f) \equiv$ valor de Shapley para la característica i

$f \equiv$ función de ganancia (XAI: modelo (caja negra))

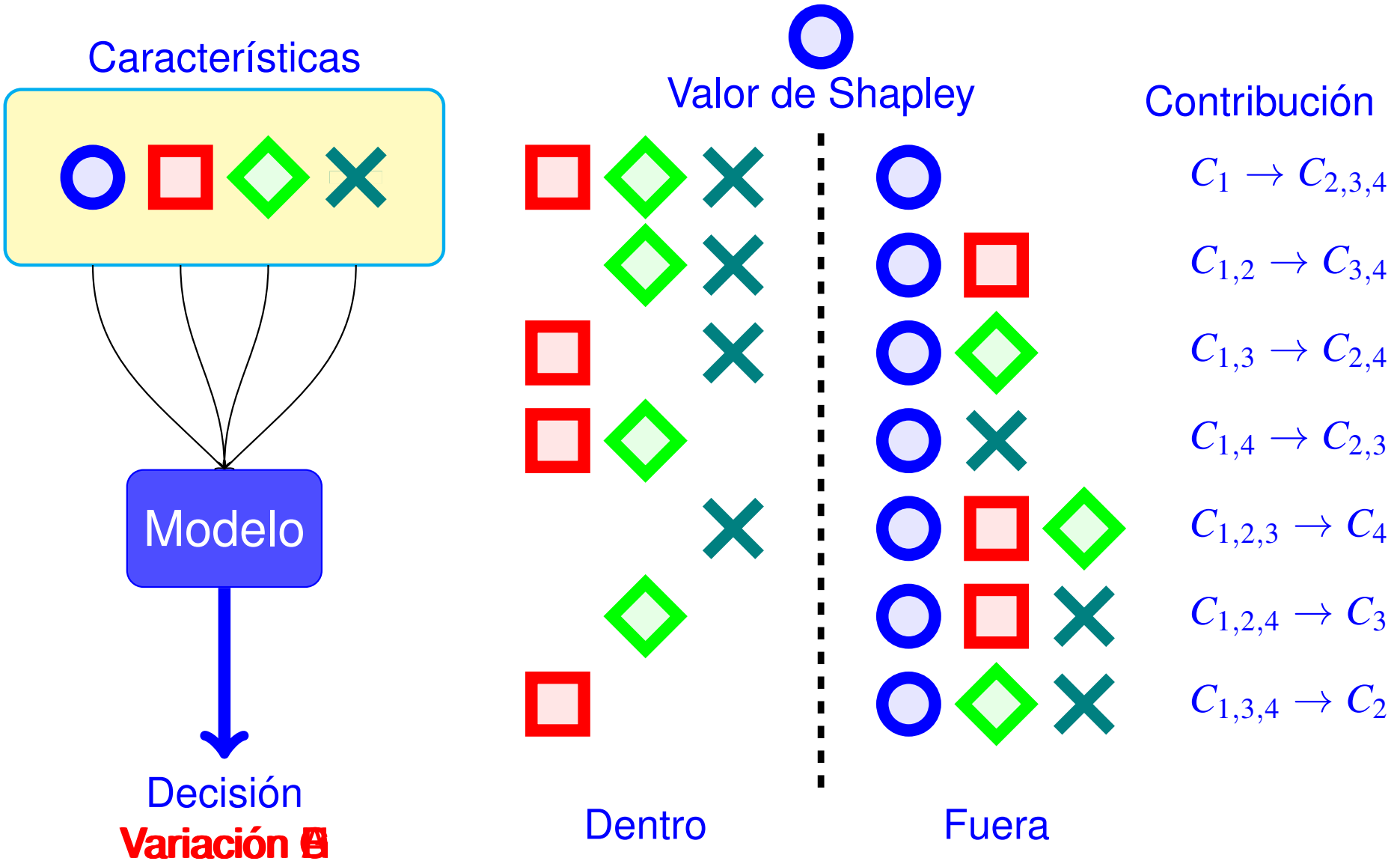
$z \equiv$ permutación que no contiene i

$f(z) \equiv$ salida para la permutación

$f(z \cup i) \equiv$ salida para la permutación incluyendo i

- Propiedades de los valores de Shapley
  - Eficiencia
    - ★ La ganancia total se distribuye como la suma de los valores
  - Simetría
    - ★ Valores iguales si cuando se unen individualmente a cada coalición generan la misma ganancia
  - Linealidad
    - ★ Sobre la función de ganancia ($\phi_{i,j}(f) = \phi_i(f) + \phi_j(f)$)

SHAP para XAI - Contribuciones de las características



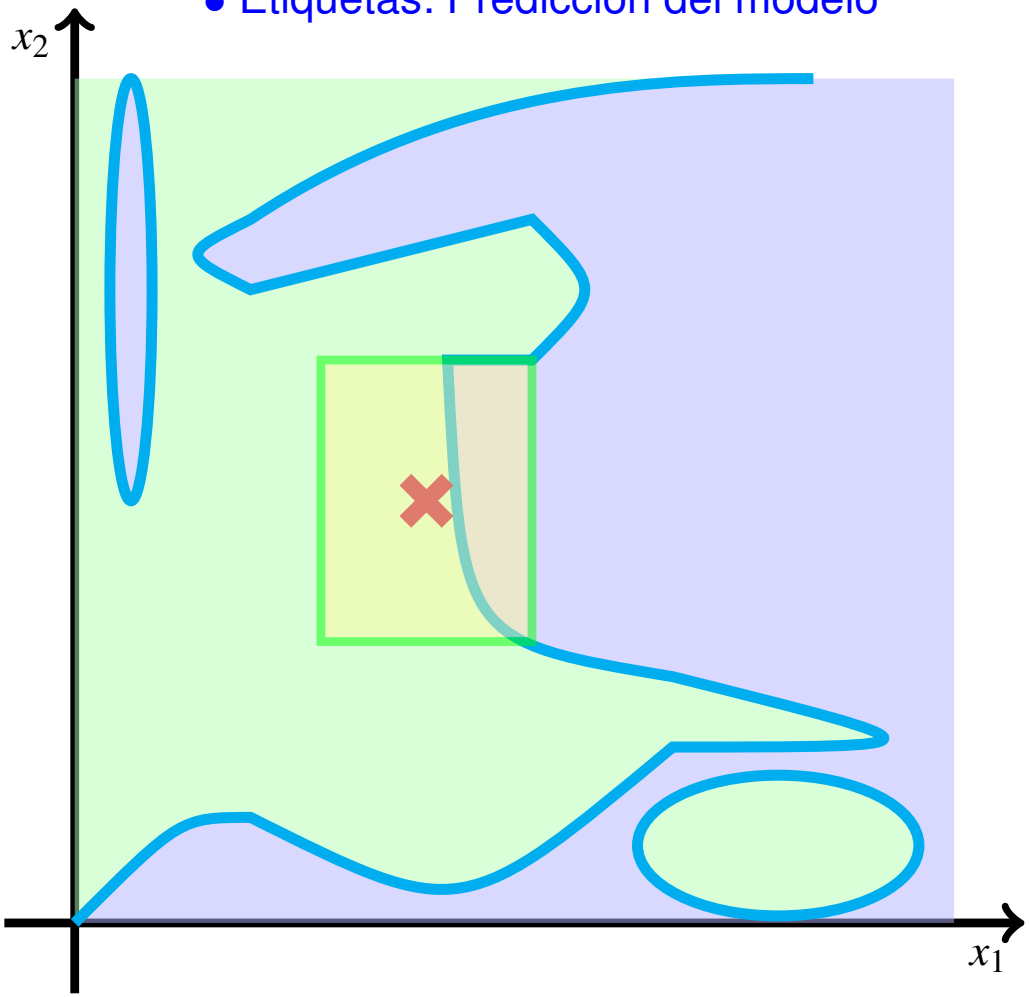
LIME (*Local Interpretable Model-Agnostic Explanations*)

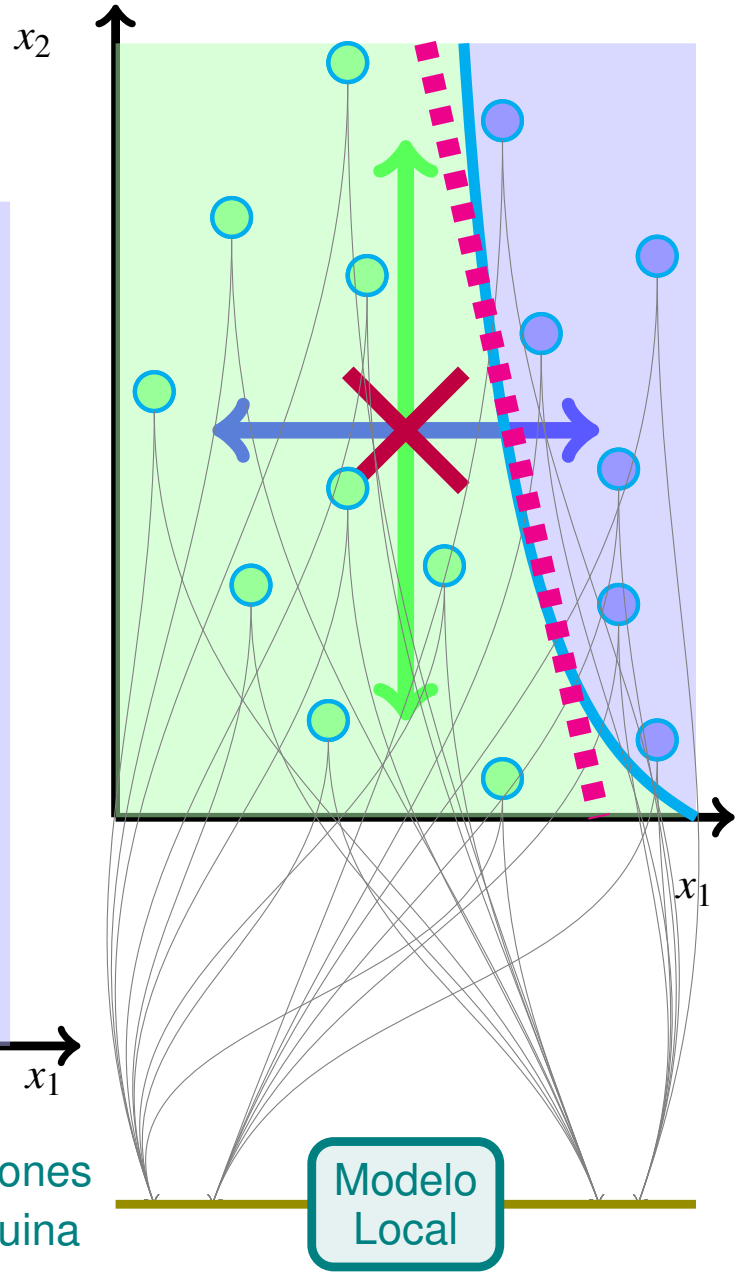
- Modelo local y agnóstico
  - Permite explicar cualquier tipo de modelo
  - Explicaciones locales sobre la decisión para un cierto patrón
- Funcionamiento básico
  - Se generan perturbaciones locales sobre el patrón
    - ★ Conjunto de patrones sintéticos “similares”
  - Se evalúan las decisiones del modelo para esas perturbaciones
  - Se entrena un modelo local para ajustar el nuevo conjunto de entrenamiento generado
    - ★ Primer modelo utilizado: Modelo Lineal
      - Se pueden utilizar otros modelos explicables
    - ★ Entrenamiento: Patrones (perturbaciones) + decisiones de la máquina
  - Se utiliza el modelo local para explicar localmente la decisión
    - ★ Análisis de sensibilidad
    - ★ Influencia relativa de las variables
    - ★ Etc.

LIME

Nuevo DATASET:

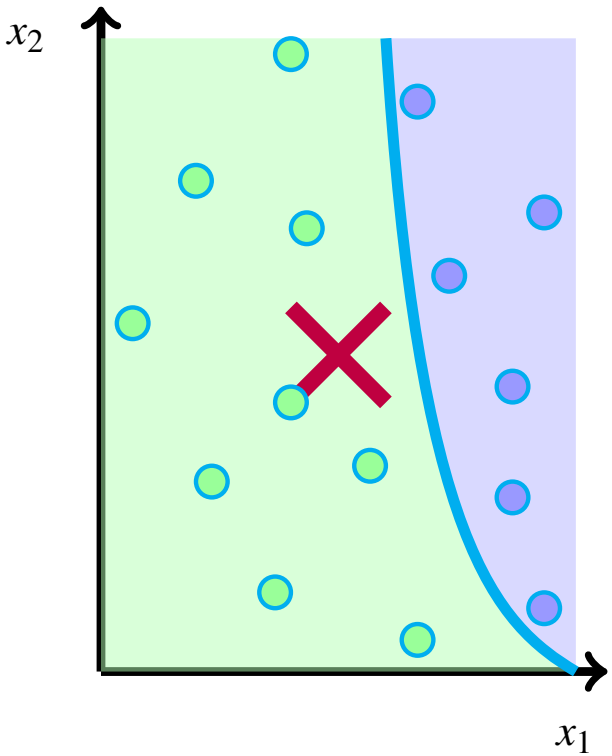
- Características: Nuevos patrones
- Etiquetas: Predicción del modelo





Modelo Local: Ajustado a las perturbaciones del patrón y a las decisiones de la máquina

LIME: Ajuste del modelo surrogado



Ajuste Suavidad

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$f \equiv$ Salida de la máquina

$g \equiv$ Modelo local, $G \equiv$ Conjunto de posibles modelos

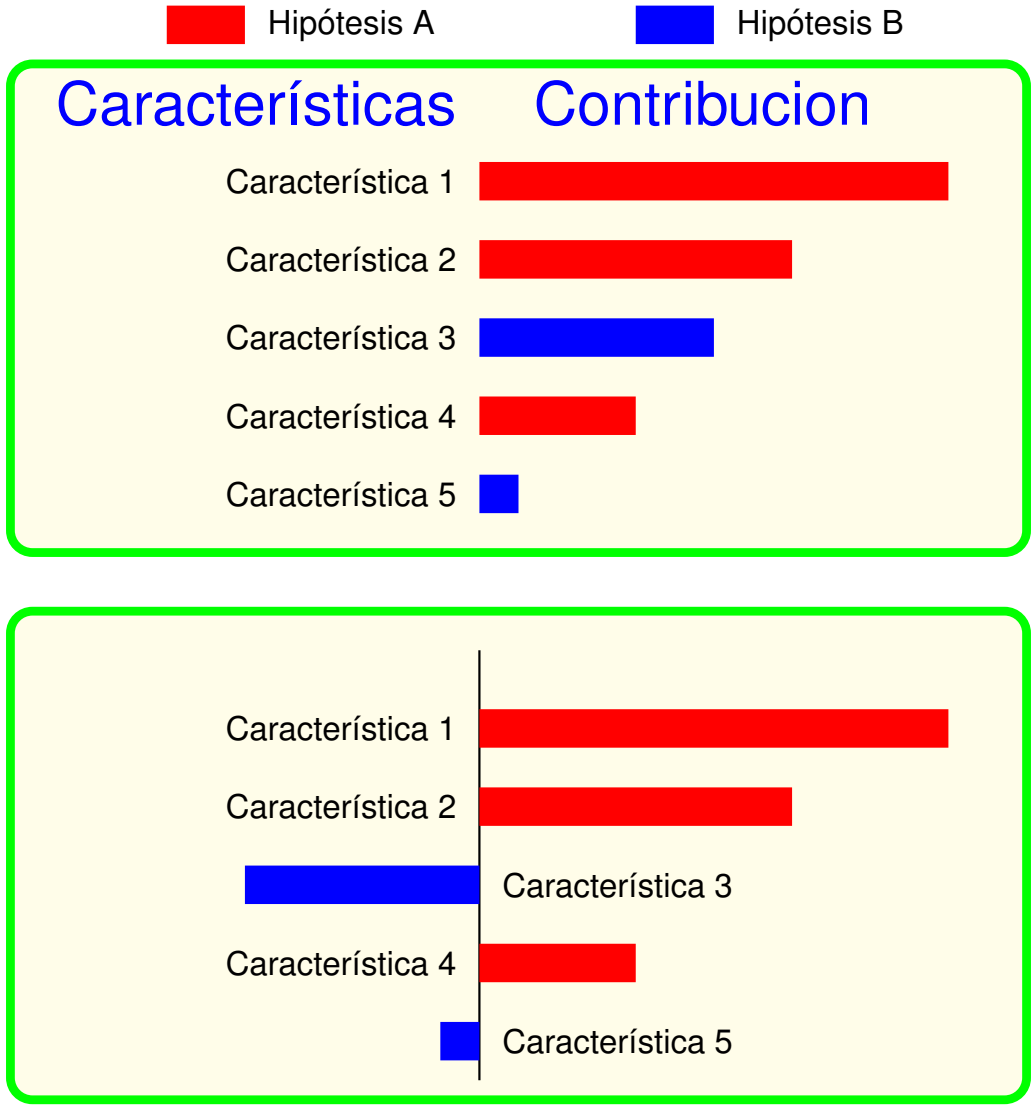
$\pi_x \equiv$ Medida de distancia al patrón

(p.e., evaluada mediante un kernel gausiano)

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

$\Omega(g) \equiv$ se prefieren soluciones dispersas (facilitan la explicación)

Representación de importancia de características



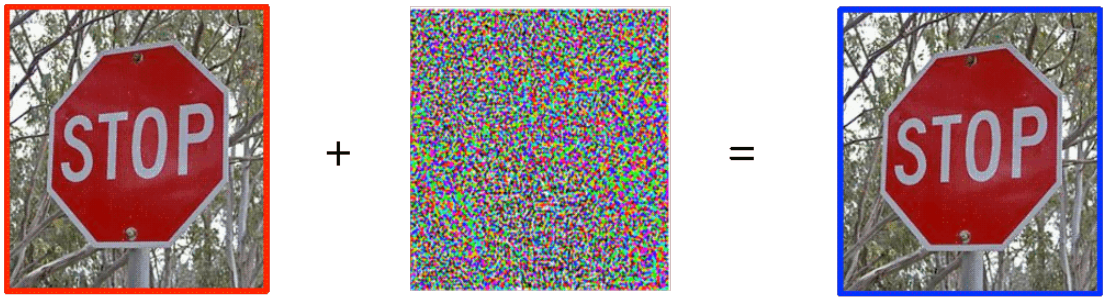
LIME: Algunos resultados

- En [Ribeiro1016] se presenta el método
  - https://dl.acm.org/doi/10.1145/2939672.2939778
- Se presentan varios resultados
  - Análisis de algunas predicciones de Google's Inception
    - ★ Imagen: Labrador + Guitarra eléctrica + Guitarra acústica
      - Explicación visual de las decisiones con mayor probabilidad
  - Experimento Husky vs Lobo
    - ★ Reconocimiento de perros de raza Husky vs Lobos
      - Identificación de una debilidad del clasificador
    - ★ También en la presentación https://www.youtube.com/watch?v=TBJqgvXYhfo
- LIME se ha utilizado desde entonces en múltiples problemas

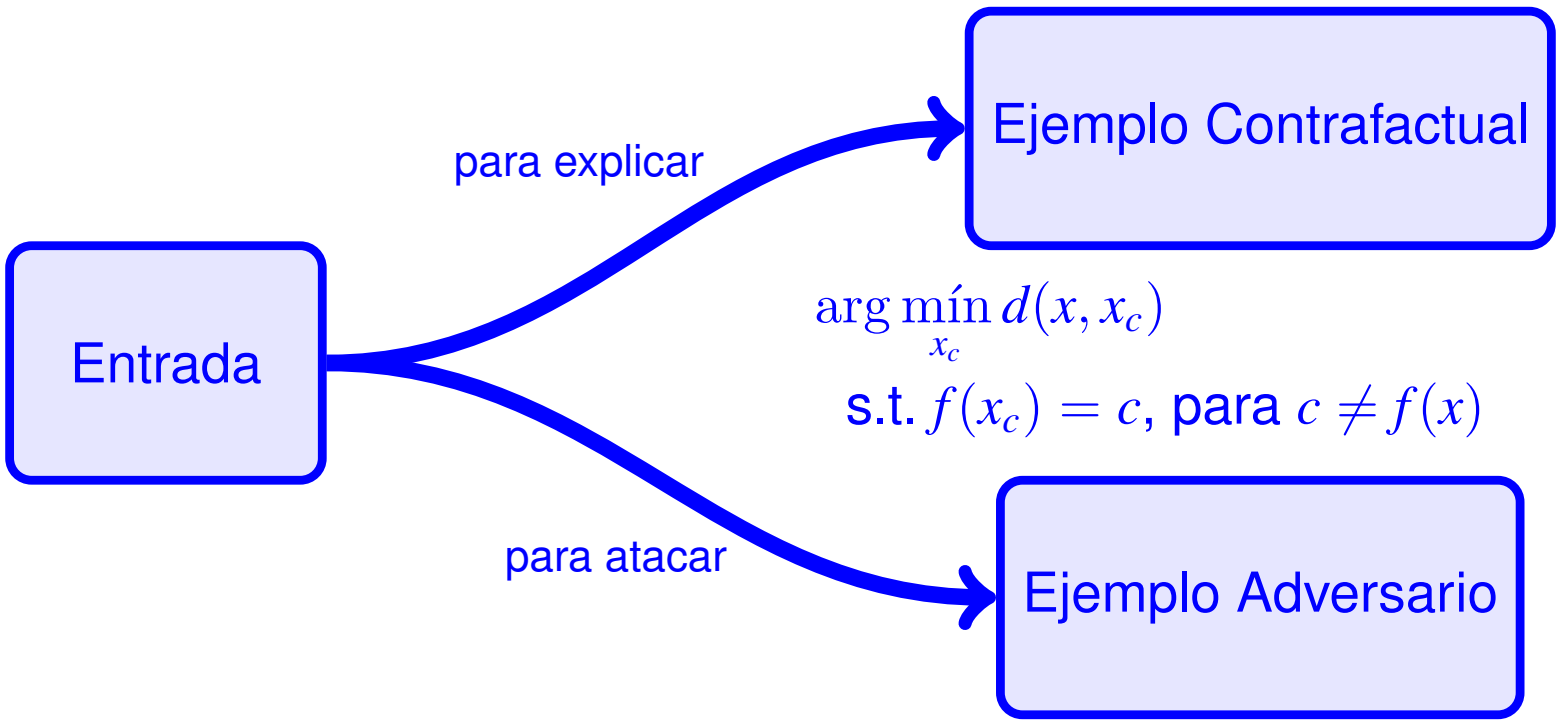
Explicaciones Contrafactuales

- Explicación de tipo local (realizada para un patrón específico)
  - Puede ser tanto agnóstica como de modelo específico
    - ★ La metodología es diferente en ambos casos
- Explicación basada en ejemplos (datos)
  - Se presentan ejemplos similares pero con distinta decisión
    - ★ Se suelen presentar varios (por el Efecto Rashomon)
- Definición de ejemplo contrafactual
  - Ejemplo con el menor cambio en las características de entrada para el que cambia la decisión con respecto al ejemplo de referencia
- Técnica relacionada con el aprendizaje adversario
  - Generación de ejemplos similar a la de ataques adversarios

Ejemplos “contrafactuales” y ataques “adversarios”

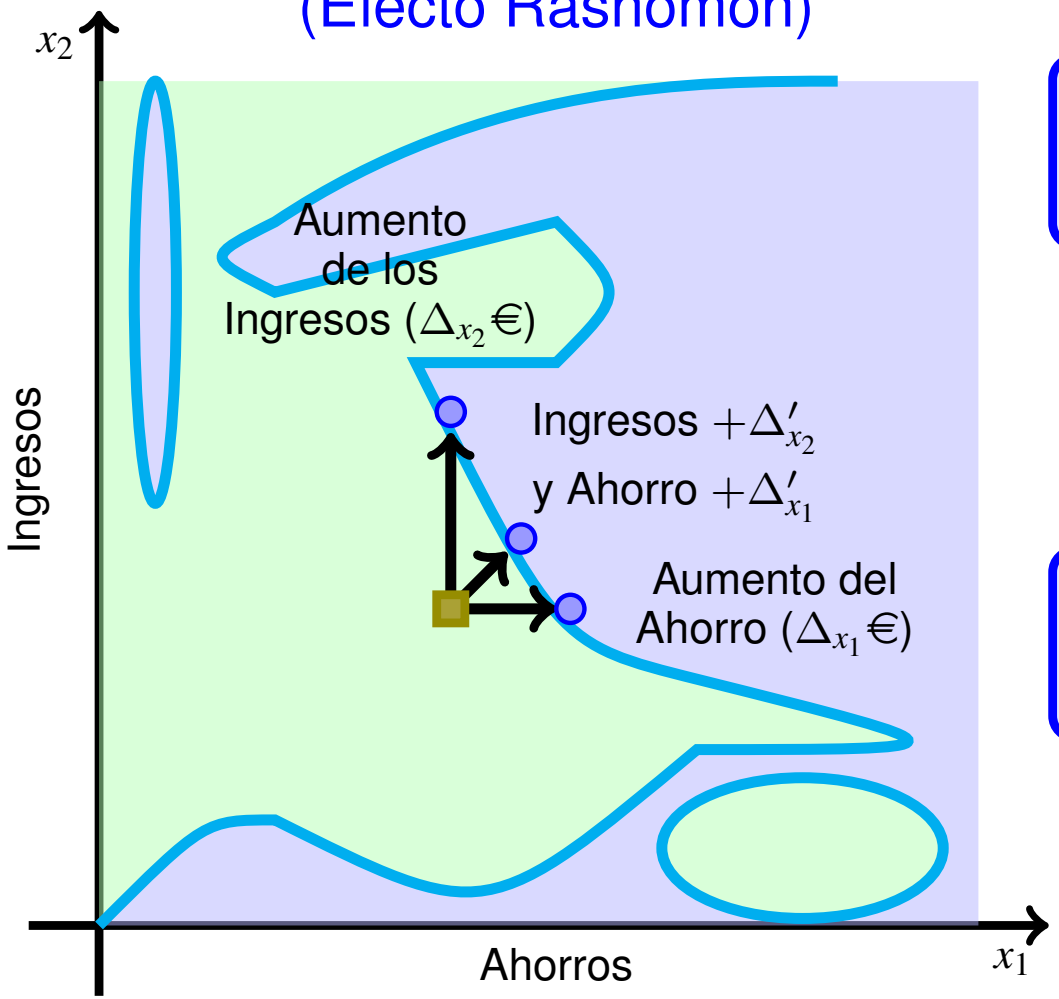


Fuente: Jiefeng Chen, Xi Wu https://deep.ghost.io/robust-attribution/



Generación de ejemplos contrafactuales

No hay un único ejemplo bueno
(Efecto Rashomon)



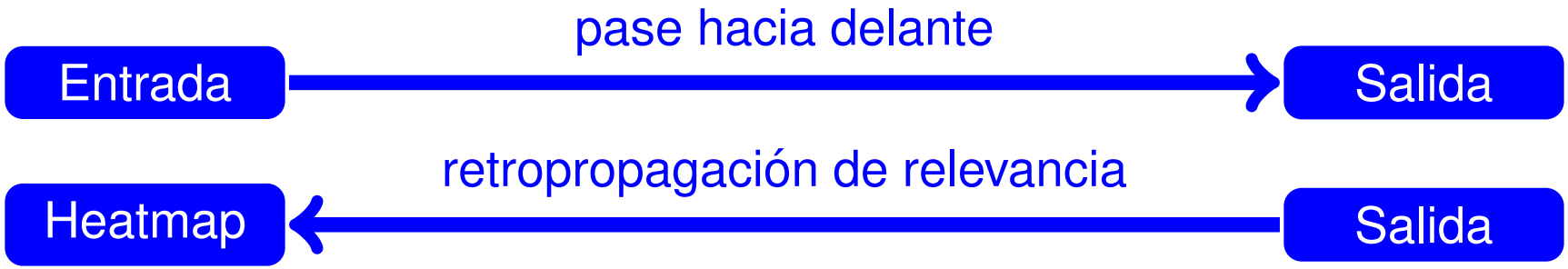
Modelo Específico (White Box)

- ⋆ Uso de información del modelo
- ⋆ Gradientes
  (sobre variables inmutables)

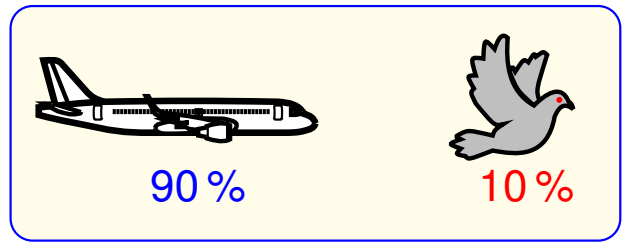
Agnóstico (Black Box)

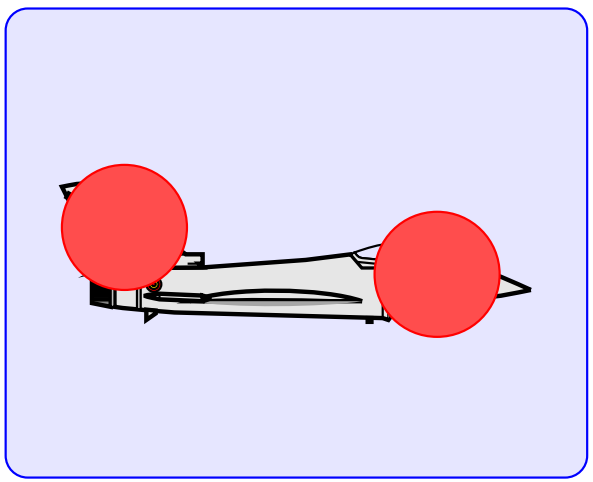
- ⋆ Evaluación repetida del modelo
  (variaciones de características)

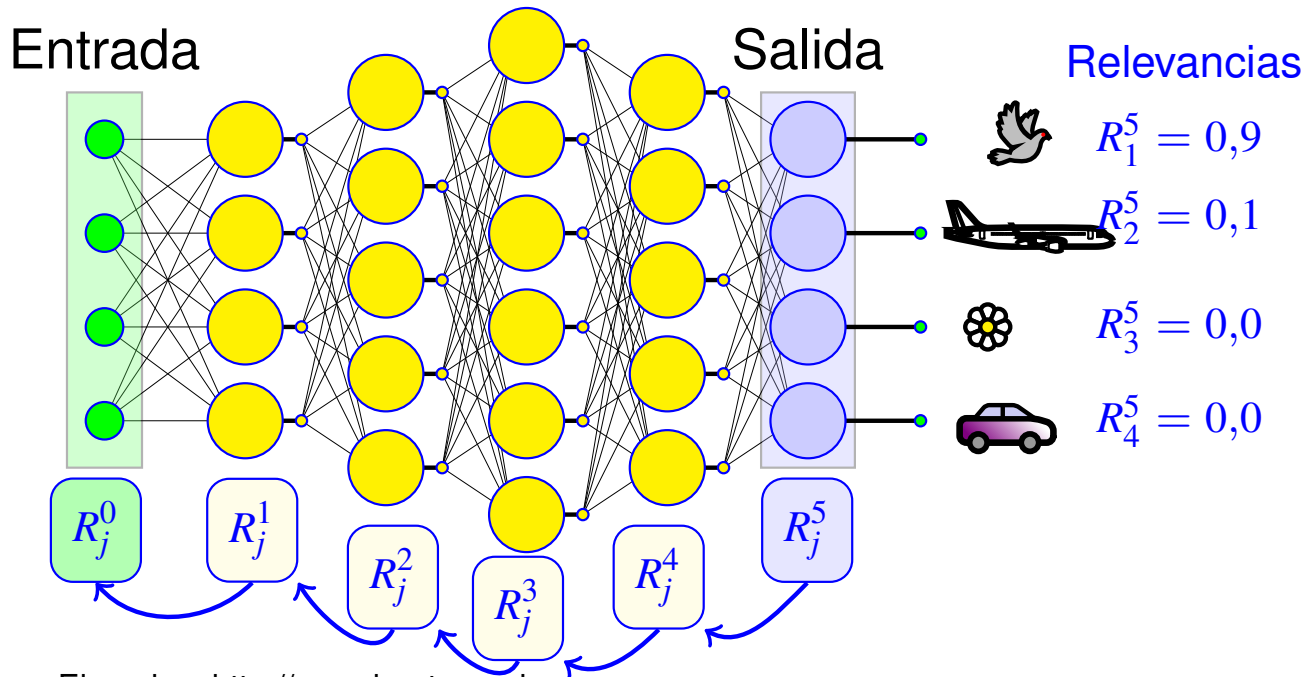
LRP (*Layerwise Relevance Propagation*)



$$R_j^{(\ell)} = \sum_j \frac{z_{i,j}}{\sum_k z_{k,j}} R_j^{(\ell+1)} \text{ con } z_{i,j} = x_i^{(\ell)} w_{i,j}^{(\ell,\ell+1)}$$







Ejemplos: http://www.heatmapping.org

AM (*Activation Maximization*)

- También conocidas como *Visualización de Características*
- Búsqueda de representaciones prototípicas de un concepto
  - Muy utilizada en aplicaciones de procesado de imagen
    - ★ Las habilidades cognitivas humanas favorecen la comprensión de datos visuales
- Análisis de qué patrones estimulan las representaciones internas de la red a varios niveles
  - Neurona
  - Canal
  - Capa
  - Probabilidades de cada clase

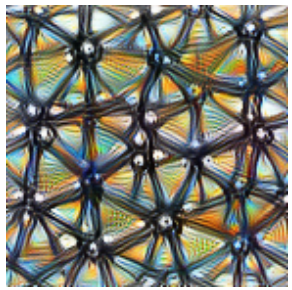
$$\arg\max_{\mathbf{X}} a_{i,j}(\mathbf{X}, \theta)$$

- Visualización de las representaciones internas
  - Valoración de si los conceptos aprendidos son (humanamente) interpretables
- Los patrones se suelen buscar mediante optimización
  - Se suele partir de ejemplos
    - ★ Generan diversidad de patrones
  - La optimización aisla las causas de comportamiento de las simples correlaciones

Visualización de características (CNNs)

Olah et al. 2017 https://distill.pub/2017/feature-visualization/

Neurona

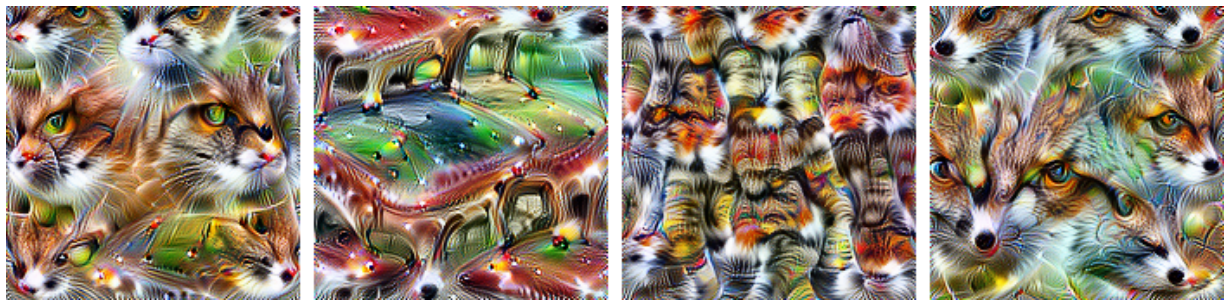


Canal

Capa

Clase

Patrones



Optimización (con diversidad)

Otras técnicas de explicabilidad

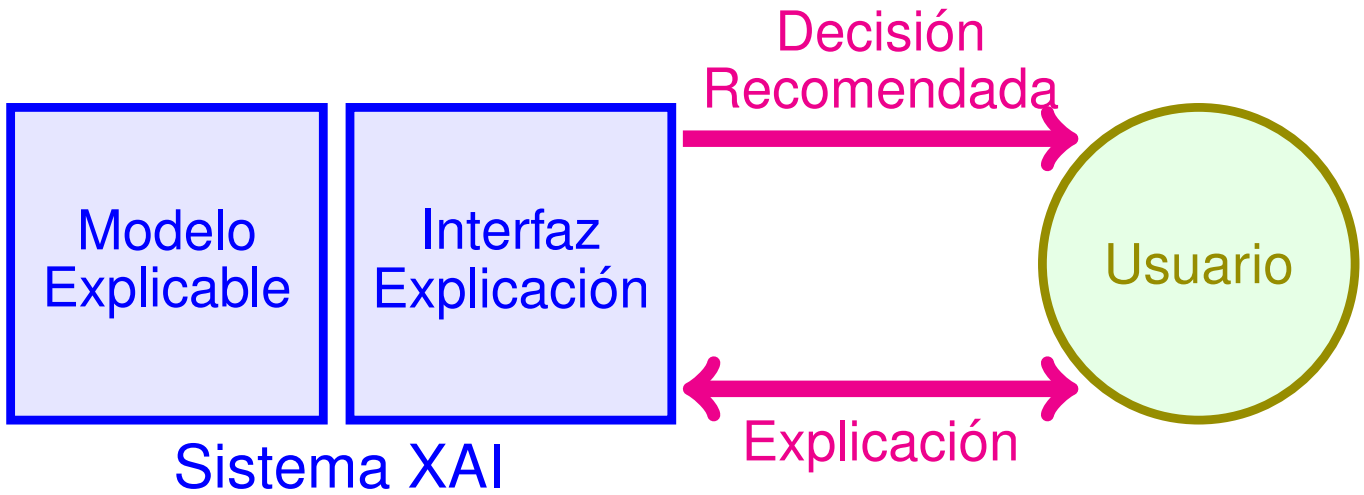
- Análisis de importancia de características
  - LOCO (Leave One Covariate Out)
    - ★ Diferencia de la decisión sin una variable
  - Anchors
    - ★ Búsqueda local de condiciones suficientes (tipo If-Then)
  - DeepLIFT
    - ★ Explica las diferencias con respecto a una referencia (de entrada)
  - XRAI
    - ★ Segmenta la imagen y evalúa la relevancia de cada región
- Métodos basados en ejemplos
  - ProtoAttend
    - ★ Búsqueda de ejemplos prototipo en base a relaciones entre muestras y representaciones codificadas
  - MMD Critic
    - ★ Búsqueda de ejemplos prototipo y criticismos (ejemplos que no están bien representados)
  - Influence Functions
    - ★ Búsqueda de los ejemplos en el conjunto de entrenamiento con más responsabilidad en la decisión
  - Representer Point Selection (similar conceptualmente al anterior)
- Métodos basados en modelos surrogados
  - Destilación
    - ★ Transferencia de conocimiento de un modelo complejo a uno explicable

Otras técnicas de explicabilidad (II)

- Técnicas de extracción de conceptos
  - TCAV (Testing with Concept Activation Vectors)
    - ★ Interpretación del estado interno de la red en términos de conceptos humanos
      - Identificación de CAVs (Concept Activation Vectors)
  - DeepR (Deep Record)
    - ★ Extracción de características de historiales médicos para predecir futuros riesgos
      - Cada informe se codifica como una secuencia indexada temporalmente
  - ACE (Automated Concept-based Explanation)
    - ★ Método global de explicación (explica cada clase globalmente)
      - Para cada clase se extraen conceptos y cuantifica su importancia
- Técnicas de modelos inherentemente explicables
  - EBM (Explainable Boosting Machine)
    - ★ Modelo aditivo generalizado con árboles por características individuales
      - Obtención de gráficas explicativas
  - DLN (Deep Lattice Networks)
    - ★ Modelos monotónicos respecto a un subconjunto de características (indicadas por el usuario)
  - Modelos Bayesianos (múltiples desarrollos)
    - ★ Explicaciones a partir de medidas/relaciones probabilísticas

Evaluación de las Explicaciones

Entorno de Evaluación



Medida de la Efectividad de la Explicación

Satisfacción de Usuario

- ⋆ Claridad de la explicación
- ⋆ Utilidad de la explicación (Calificación del Usuario)

Modelo Mental

- ⋆ Comprensión de las decisiones
- ⋆ Comprensión del modelo
- ⋆ Análisis de debilidades

Prestaciones

- ⋆ ¿Mejoran con la explicación?
- ⋆ ¿Corrección de errores?
- ⋆ ¿Entrenamiento continuo?

Evaluación de la Confianza

- ⋆ ¿Cómo medirla?
- ⋆ Básico para su desarrollo

Trabajo de Investigación Necesario!!!

Algunas Referencias Bibliográficas

Revisiones

Adadi, A. and Berrada, M. (2018).

Peeking inside the black-box: A survey on explainable artificial intelligence (XAI).
IEEE Access, 6:52138–52160.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020).

Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI.
Information Fusion, 58:82–115.

Goodman, B. and Flaxman, S. (2017).

European Union regulations on algorithmic decision making and a “Right to Explanation”.
AI Magazine, 30(3):50–57.

Rudin, C. (2019).

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.
Nature Machine Intelligence, 1:206–215.

Métodos

Chen, Z., Tan, S., Nori, H., Inkpen, K., Lou, Y., and Caruana, R. (2022).

Using explainable boosting machines (EBMs) to detect common flaws in data.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 534–551.

Ghorbani, A., Wexler, J., Zou, J., and Kim, B. (2019).

Towards automatic concept-based explanations.
In *International Conference on Neural Information Systems (NIPS)*, pages 9277–9286.

Kapishnikov, A., Bolukbasi, T., Viégas, F., and Terry, M. (2019).

XRAI: Better attributions through regions.
In *International Conference on Computer Vision (ICCV)*, pages 4947–4956.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018).

Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV).
In *International Conference on Machine Learning*.

Métodos (II)

Lundberg, S. M. and Lee, S.-I. (2017).

A unified approach to interpreting model predictions.
In *International Conference on Neural Information Processing Systems (NIPS)*, pages 4768–4777.

Nguyen, P., Tran, T., Wickramashinghe, N., and Venkatesh, S. (2017).

Deepr: A convolutional net for medical records.
IEEE Journal of Biomedical and Health Informatics, 21(1):22–30.

Nori, H., Caruana, R., Bu, Z., Shen, J. H., and Kulkarni, J. (2021).

Accuracy, interpretability, and differential privacy via explainable boosting.
In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8227–8237. PMLR.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).

"Why should I trust you?" Explaining the predictions of any classifier.
In *International Conference on Knowledge Discovery and Data (KDD)*, pages 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2018).

Anchors: High-precision model-agnostic explanations.
In *AAAI Conference on Artificial Intelligence*, volume 32, pages 1527–1535.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017).

Learning important features through propagating activation differences.
In *International Conference on Machine Learning (ICML)*, volume 70, pages 3145–3153.

Wachter, S., Mittelstadt, B., and Russell, C. (2017).

Counterfactual explanations without opening the black box: Automated decisions and the GDPR.
arXiv:1711.00399 [cs.AI].

You, S., Ding, D., Canini, K., Pfeifer, J., and Gupta, M. R. (2017).

Deep lattice networks and partial monotonic functions.
In *International Conference on Neural Information Systems (NIPS)*, pages 2985–2993.