Aprendizaje de máquina interpretable

DAVID AUGUSTO CÁRDENAS PEÑA
Universidad Tecnológica de Pereira
Ingeniero Electrónico
Magister en Automatización Industrial
Doctor en Ingeniería

~~Casos de éxito~~ Fracasos

IA en el sector financiero

- Empleadores, propietarios y compañías de seguros usan historial crediticio para tomar decisiones.
- El problema: El historial y los puntajes crediticios no son neutrales respecto a los grupos étnicos.
- La consecuencia: Satisfacción de necesidades por productos y servicios desestabilizadores y de alto costo (paga diario, cambiadores de cheques…).
- El estado: Las leyes obligan a los bancos a satisfacer las necesidades crediticias de todas las comunidades a las que sirven, de acuerdo con la seguridad y la privacidad.
- La causa: Durante décadas, los bancos han negado sistemáticamente préstamos en localidades de negritudes y latinos; y no ubican sucursales en áreas no blancas y de bajos ingresos.

https://www.wsj.com/articles/https://www.theguardian.com/commentisfree/2015/oct/13/your-credit-score-is-racist-heres-why

S.E.S. Hospital Universitario de Caldas
Queremos devolver sonrisas

ACREDITACION EN SALUD

automática Grupo I+D

Universidad Tecnológica de Pereira

~~Casos de éxito~~ Fracasos

The New York Times

Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler
June 13, 2017

Give this article 230

Sally Deng

IA en la justicia penal:

- COMPAS retrasó la libertad condicional de Glenn Rodríguez debido a errores en la entrada del algoritmo.
- El caso “El estado contra Chubbs” es referencia en EEUU para negar a los acusados el acceso a pruebas de secretos comerciales.
- El secreto comercial ha obstruido a las defensas evaluar si una aplicación de reconocimiento facial está sobre ajustado para ciertos grupos raciales y no para otros.
- “La raíz del problema es que las tecnologías de justicia penal automatizada son en gran parte de propiedad privada y se venden con fines de lucro.”

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

~~Casos de éxito~~ Fracasos

IA en el sector salud:

- OPTUM jerarquiza pacientes crónicos para recibir atención médica adicional.
- El problema: Los pacientes negros tenían menos probabilidades que los pacientes blancos de obtener ayuda médica adicional, a pesar de estar más enfermos.
- La causa: “OPTUM consideró el gasto médico para clasificar pacientes, el cual era menor en pacientes negros que para los pacientes blancos con condiciones médicas similares.”

https://www.wsj.com/articles/researchers-find-racial-bias-in-hospital-algorithm-11571941096

S.E.S Hospital Universitario de Caldas
Queremos devolver sonrisas

ACREDITACION EN SALUD

automática Grupo I+D

Universidad Tecnológica de Pereira

Interpretación

- Interpretable ML (iML)
- La causa y el efecto pueden ser determinados.

Explicación

- Explainable AI (XAI)
- Se tiene el significado de un parámetro y su importancia en el desempeño.

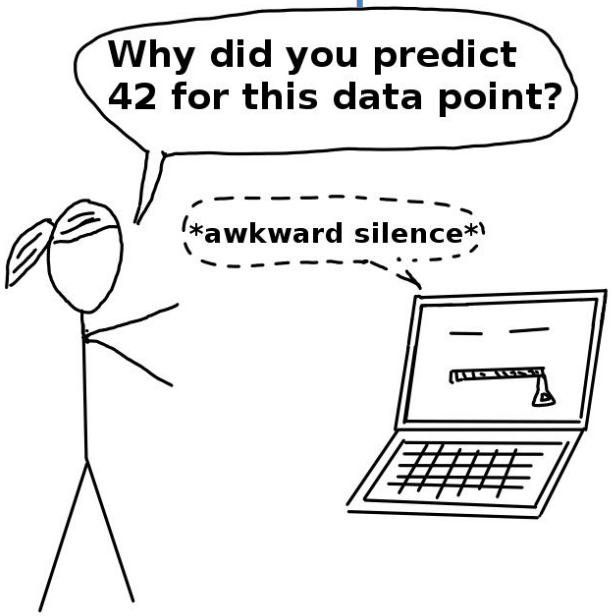
Interpretación

- Interpretable ML (iML)
- La causa y el efecto pueden ser determinados.

Explicación

- Explainable AI (XAI)
- Se tiene el significado de un parámetro y su importancia en el desempeño.

Ambos tratan de resolver el problema de la caja negra



“El aprendizaje de máquina interpretable

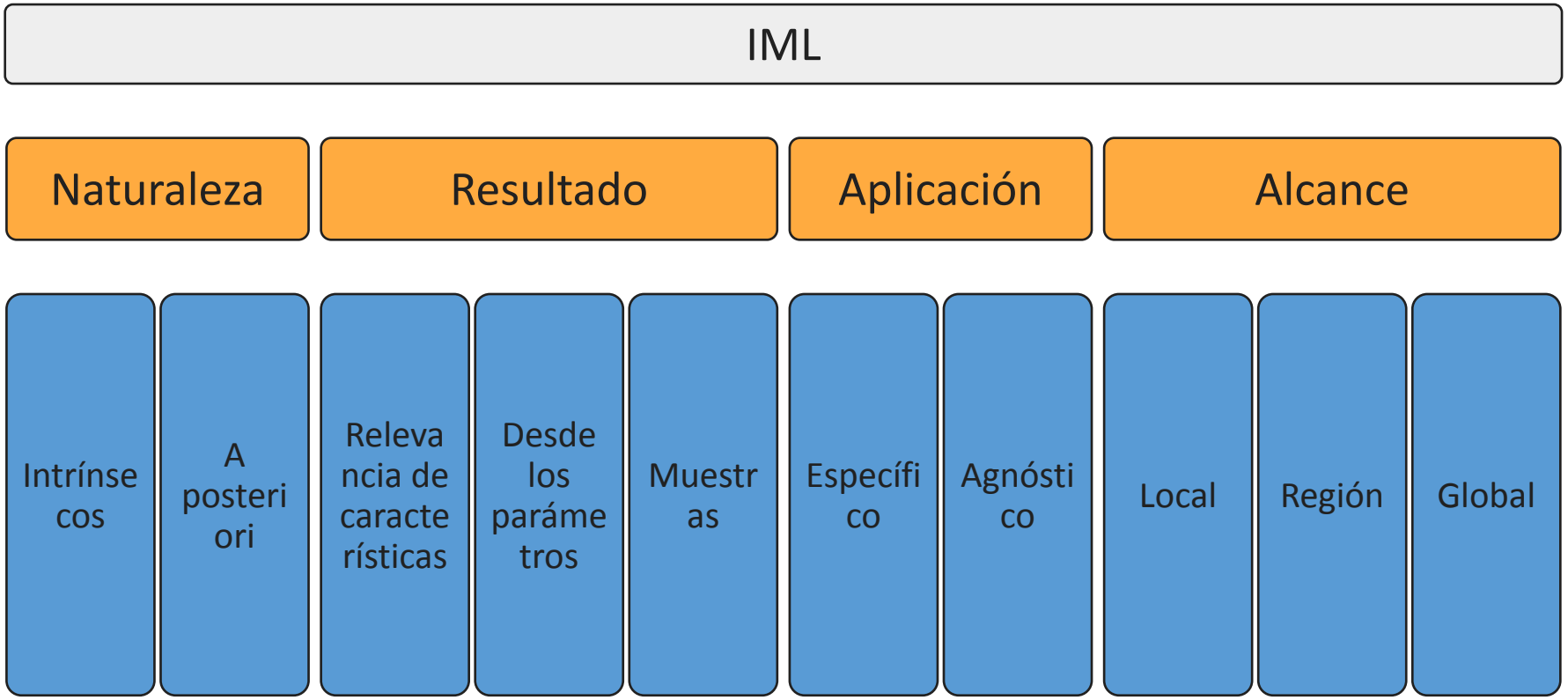
se refiere a métodos y modelos que hacen que el comportamiento y las predicciones de los sistemas de aprendizaje automático sean **comprensibles para los humanos**.”

Christoph Molnar

PhD Ludwig-Maximilians-University

Interpretable Machine Learning: A Guide for Making Black Box Models Explainable

Taxonomía del iML



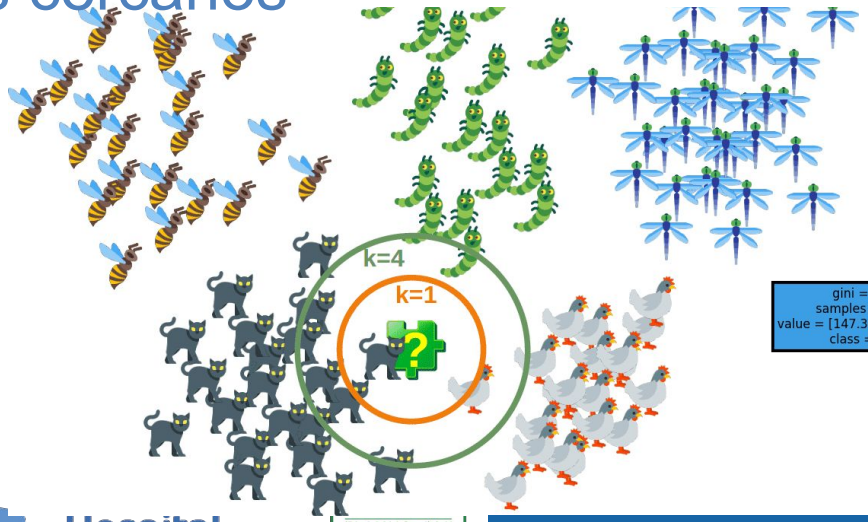
Intrínsecamente interpretables

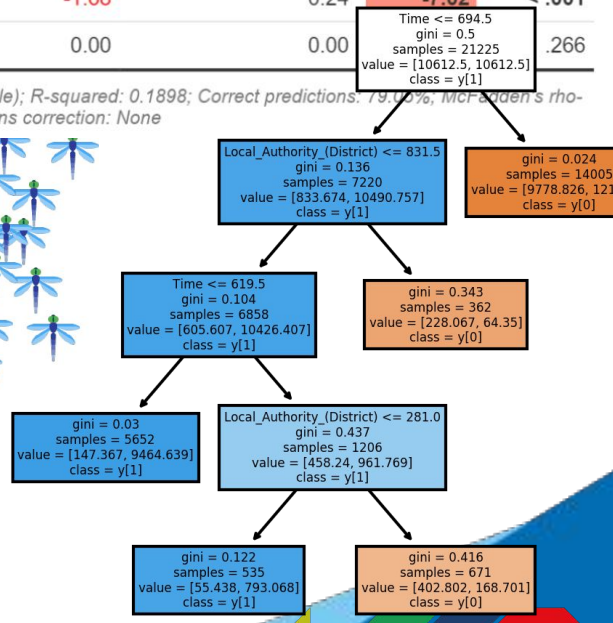
- Regresor lineal
- Regresor logístico
- Árboles de decisión
- Naive Bayes
- Vecinos más cercanos

Binary Logit: Churn

	Estimate	Standard Error	z	p
(Intercept)	-1.41	0.16	-8.73	< .001
Senior Citizen: Yes	0.41	0.11	3.60	< .001
Tenure	-0.03	0.00	-11.38	< .001
Internet Service: DSL	0.92	0.21	4.39	< .001
Internet Service: Fiber optic	1.82	0.32	5.66	< .001
Contract: One year	-0.88	0.14	-6.25	< .001
Contract: Two year	-1.68	0.24	-7.02	< .001
Monthly Charges	0.00	0.00		.266

n = 3,522 cases used in estimation (Training sample); R-squared: 0.1898; Correct predictions: 79.05%; McFadden's rho-squared: 0.2564; AIC: 3,065.1; multiple comparisons correction: None





S.E.S Hospital Universitario de Caldas
Queremos devolver sonrisas

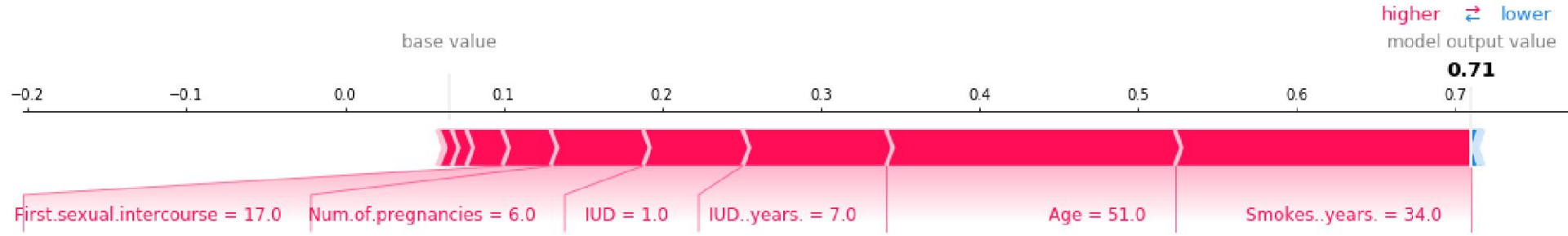
Ministerio de Salud y Protección Social
ACREDITACION EN SALUD

automática Grupo I+D

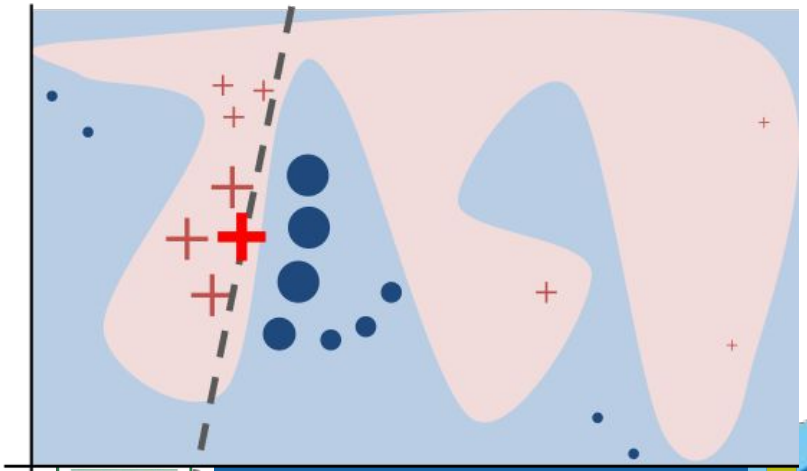
Universidad Tecnológica de Pereira

Agnósticos locales

- SHAP (SHapley Additive exPlanations)

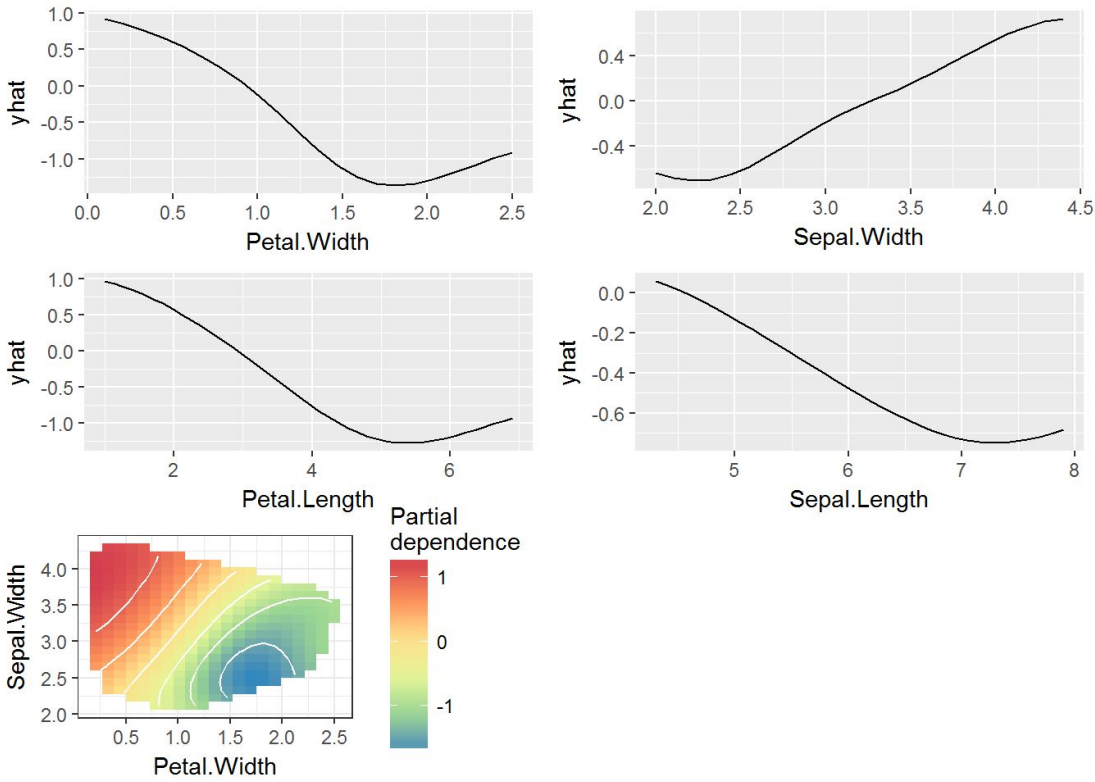


- Local interpretable model-agnostic explanations (LIME)



Agnósticos globales

- **Partial dependence plots**: Efecto marginal de una o dos variables en la predicción.
- **Feature Interaction:** "El todo es mayor que la suma de sus partes".
- **Permutation Feature Importance:** Incremento en el error al cambiar el valor de un predictor.

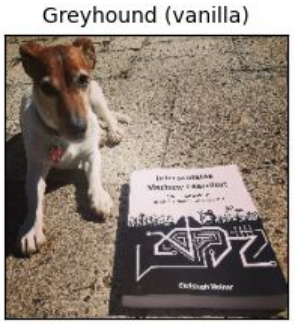


PDP for one and two predictors

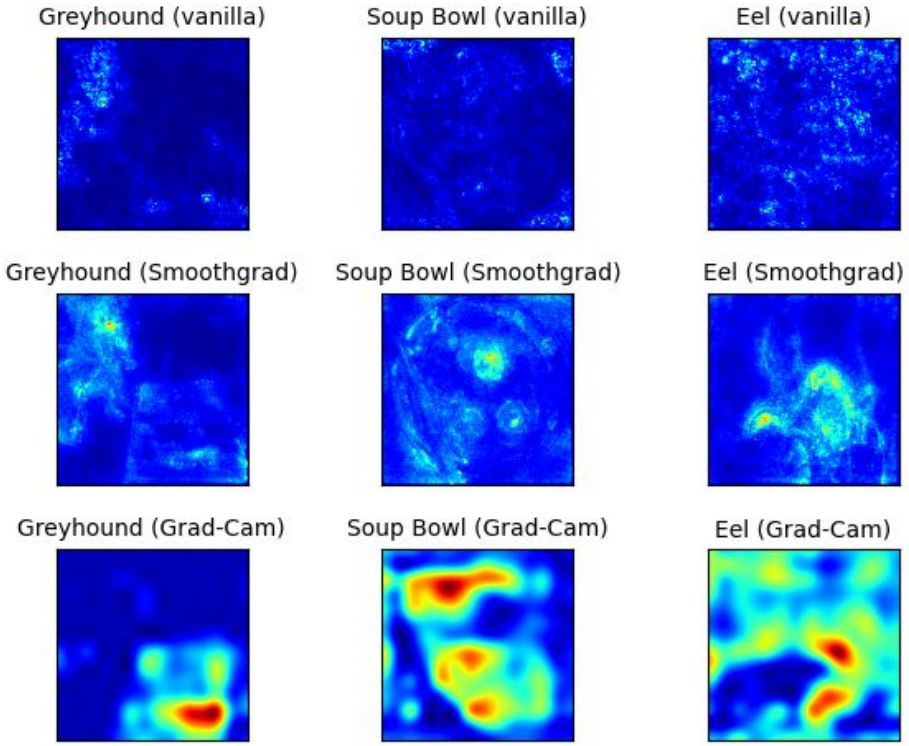
Específicos para DNN

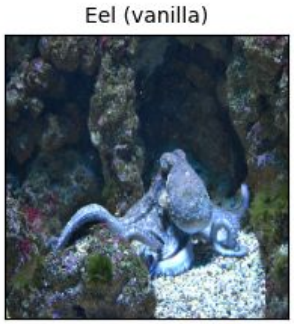
- **Pixel Attribution (Saliency Maps):** Resalta pixeles relevantes para obtener la etiqueta

50%



35%





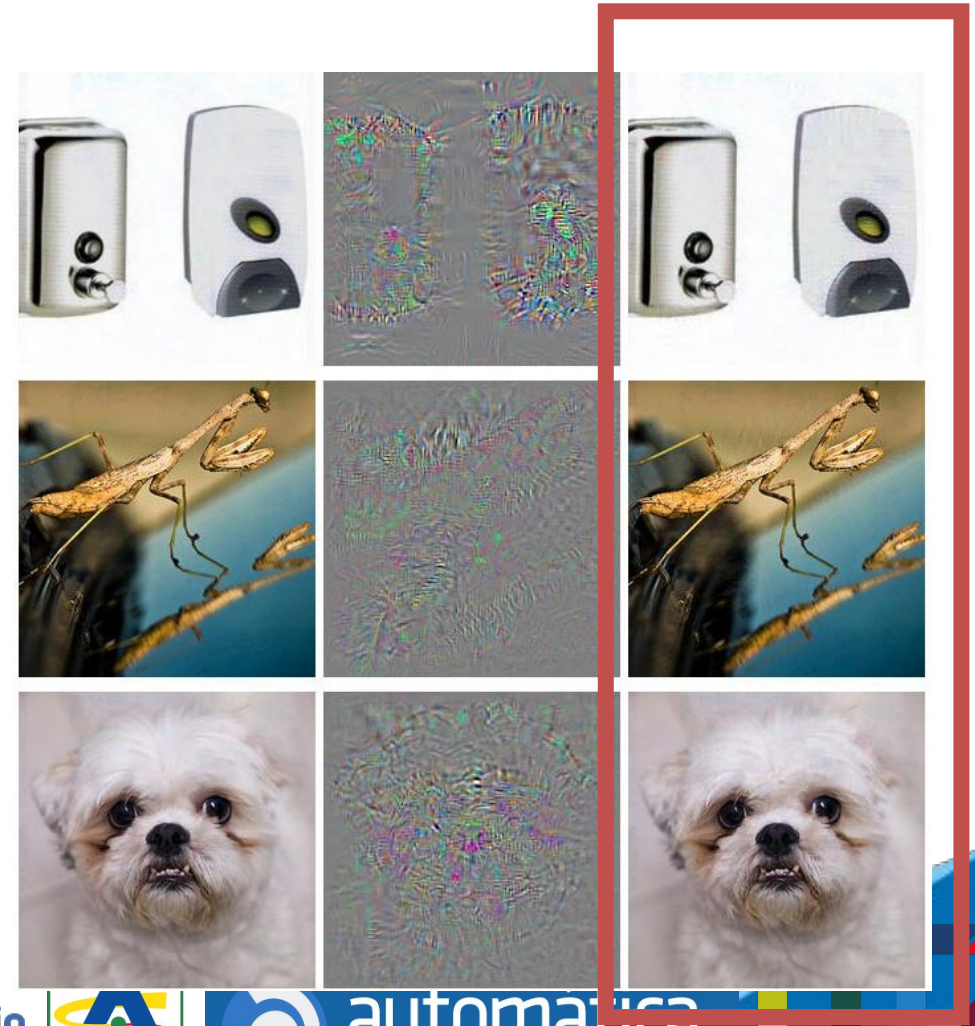
70%

Específicos para DNN - Adversarial Ex.

- **Ejemplos adversarios:** Minimiza el costo en una categoría deseada respecto de una perturbación:

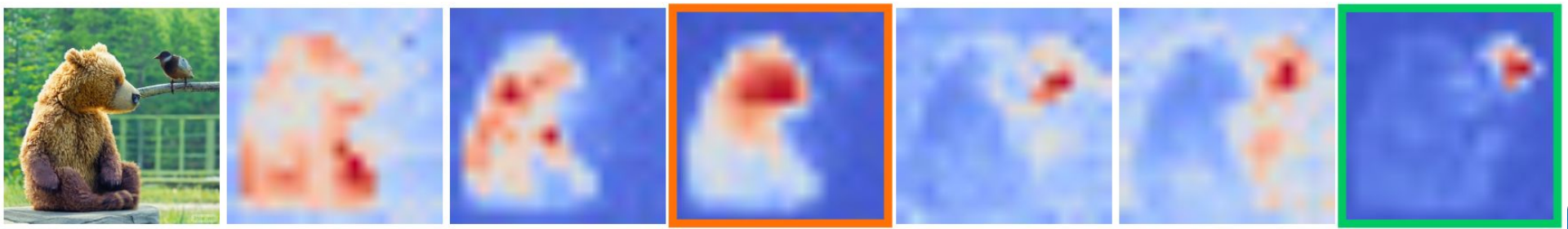
$$loss(\hat{f}\,(x+r), l) + c \cdot |r|$$

- Responde: ¿Qué tanto hay que cambiar la entrada para obtener una nueva salida?



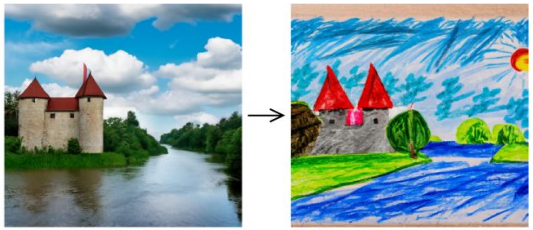
Específicos para DNN - Detecting Concepts

- Conceptos como abstracciones (color, objeto, idea…)
- Detecta conceptos incrustados en el espacio *latente* aprendido por la red.
- Puede generar interpretaciones que no están limitadas por el espacio de características de una red neuronal.
- *Prompt-to-Prompt Image Editing with Cross Attention Control* (Aug, 2022):

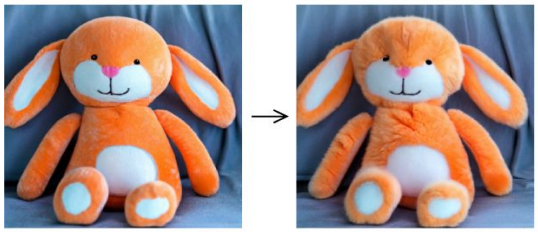


Específicos para DNN - Detecting Concepts

- *Prompt-to-Prompt Image Editing with Cross Attention Control* (Aug, 2022):



"Children drawing of a castle next to a river."



"My fluffy bunny doll."

"A photo of a butterfly on..."

"...on the river." "...on a fruit"

"A black bear is walking in the grass."

real image reconstructed

"...next to red flowers." "...when snow comes down." "while another black bear is watching." "Oil painting of..."

El conocimiento es de todos

Minciencias

Desde el proyecto “Desarrollo de una herramienta de seguimiento de aguja y segmentación de estructuras nerviosas en imágenes de ultrasonido”...

MUCHAS GRACIAS!

S.E.S. Hospital Universitario de Caldas

Queremos devolver sonrisas

ACREDITACION EN SALUD

automática Grupo I+D

Universidad Tecnológica de Pereira