

TECH CONFERENCE

DotNet 2020

#DotNet2020

En busca de una IA inclusiva



ORGANIZATION

plain concepts 

PLATINUM SPONSORS



COLLABORATORS



Thank you!

Adivina adivinanza...

Un padre y un hijo tienen un accidente grave, el padre muere y el hijo necesita una compleja operación de emergencia.

Al llegar al hospital, llaman a una eminencia médica, pero cuando llega y ve al paciente dice: no puedo operarlo, es mi hijo. *¿Quién es?*



Sara San Lu s

AI Software Developer Engineer

@sara_sanluis
sluis@plainconcepts.com



Daniela Sol s

AI Team Lead

@Danysolism
dsolis@plainconcepts.com

Casos reales

REDES SOCIALES • Microsoft ha pedido disculpas y la ha desactivado

Una inteligencia artificial se vuelve
racista, antisemita y
menos de un día en

Español ▾



Turco ▾

ella es doctora



o bir doktor
o bir hemşire

Amazon built an AI tool to hire people but had to shut it down
because it was discriminating against women

Isobel Asher Hamilton, Business Insider 10 Oct 2018 11:48h.



led two

black people's faces with the word 'gorillas'

Molly Mulshine Jul 1 2015 10:55 PM

≡ EL PAÍS

TECNOLOGÍA

MÓVILES REDES SOCIALES BANCO DE PRUEBAS RETINA MERISTATION

Facebook intercambia información de los usuarios con
entidades bancarias en EE UU

La red social asegura que lo que pretende es potenciar el uso de Messenger

doctor
enfermera



La pregunta no es qué puede hacer la IA, sino *qué debería hacer*

Principios éticos

INTELIGENCIA
ARTIFICIAL



Justicia



Inclusión



Seguridad



Privacidad



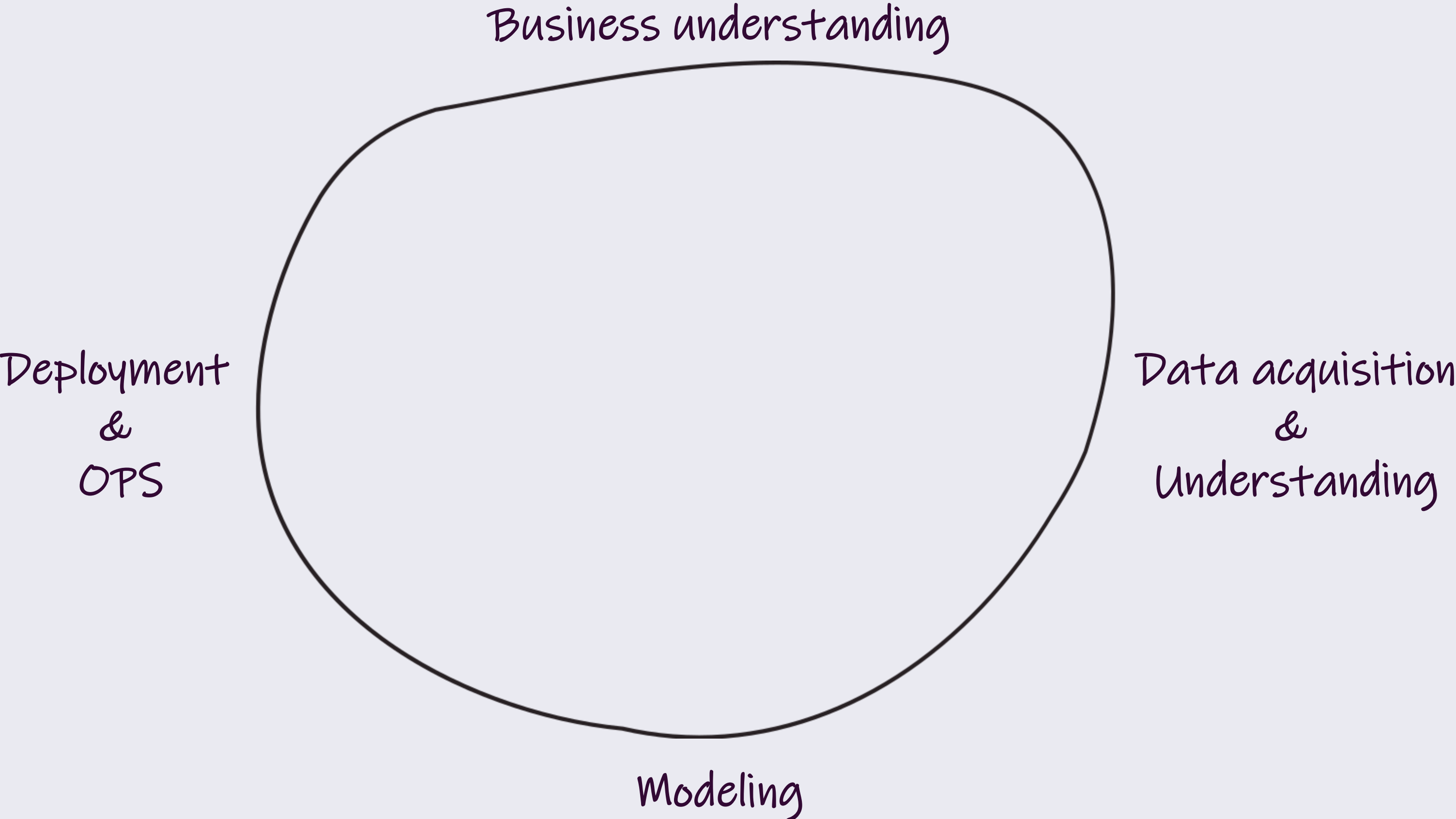
Responsabilidad

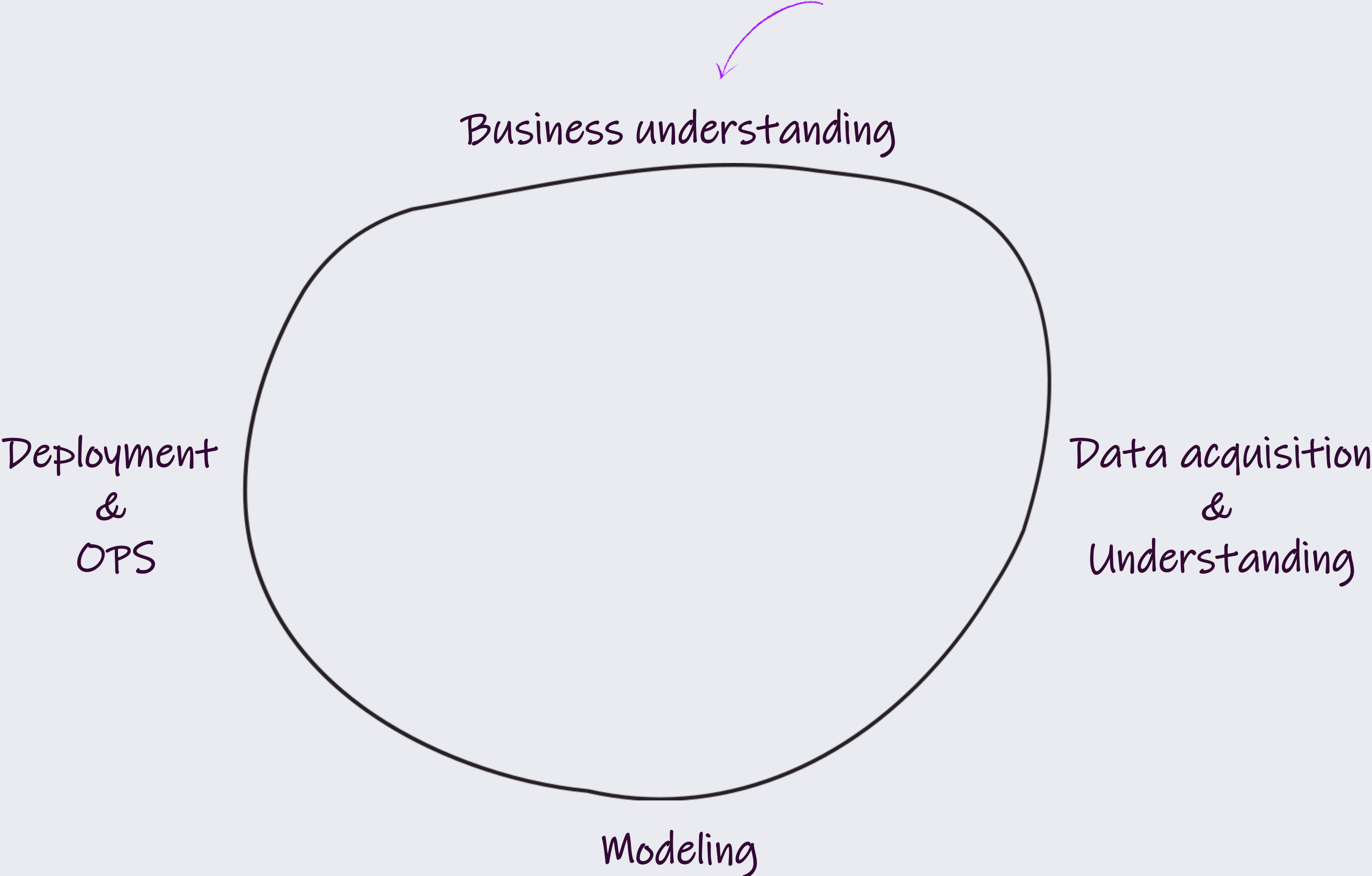


Transparencia

Machine Learning

WORKFLOW



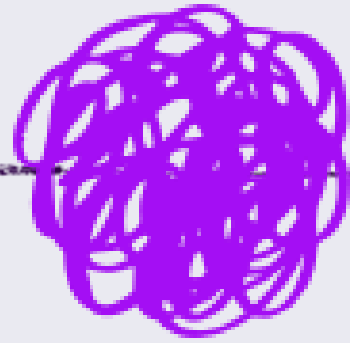




Severidad



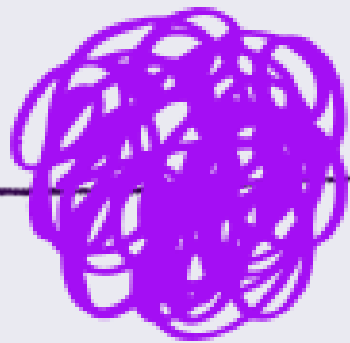
Escala



Probabilidad



Frecuencia



Facilidad de actualización



Mal uso

Análisis de riesgos

- Riesgo de lesión
- Negación de futuros servicios
- Erosión de estructuras sociales y democráticas
- Incumplimiento de los derechos humanos



Mitigación de daños

Primer grado

La compañía *causa* el daño

Mitigar y remediar

Plataforma

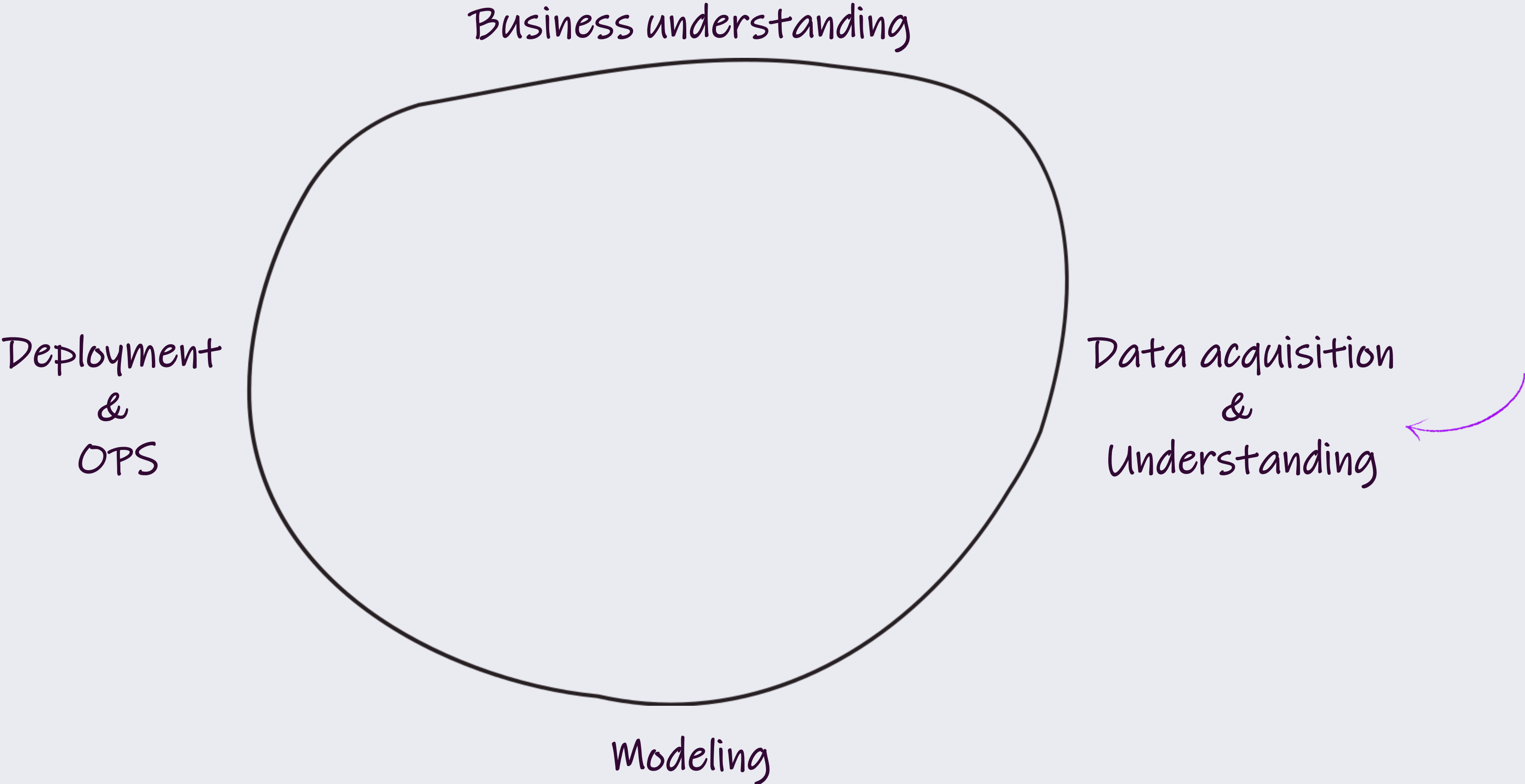
La compañía *contribuye* al daño

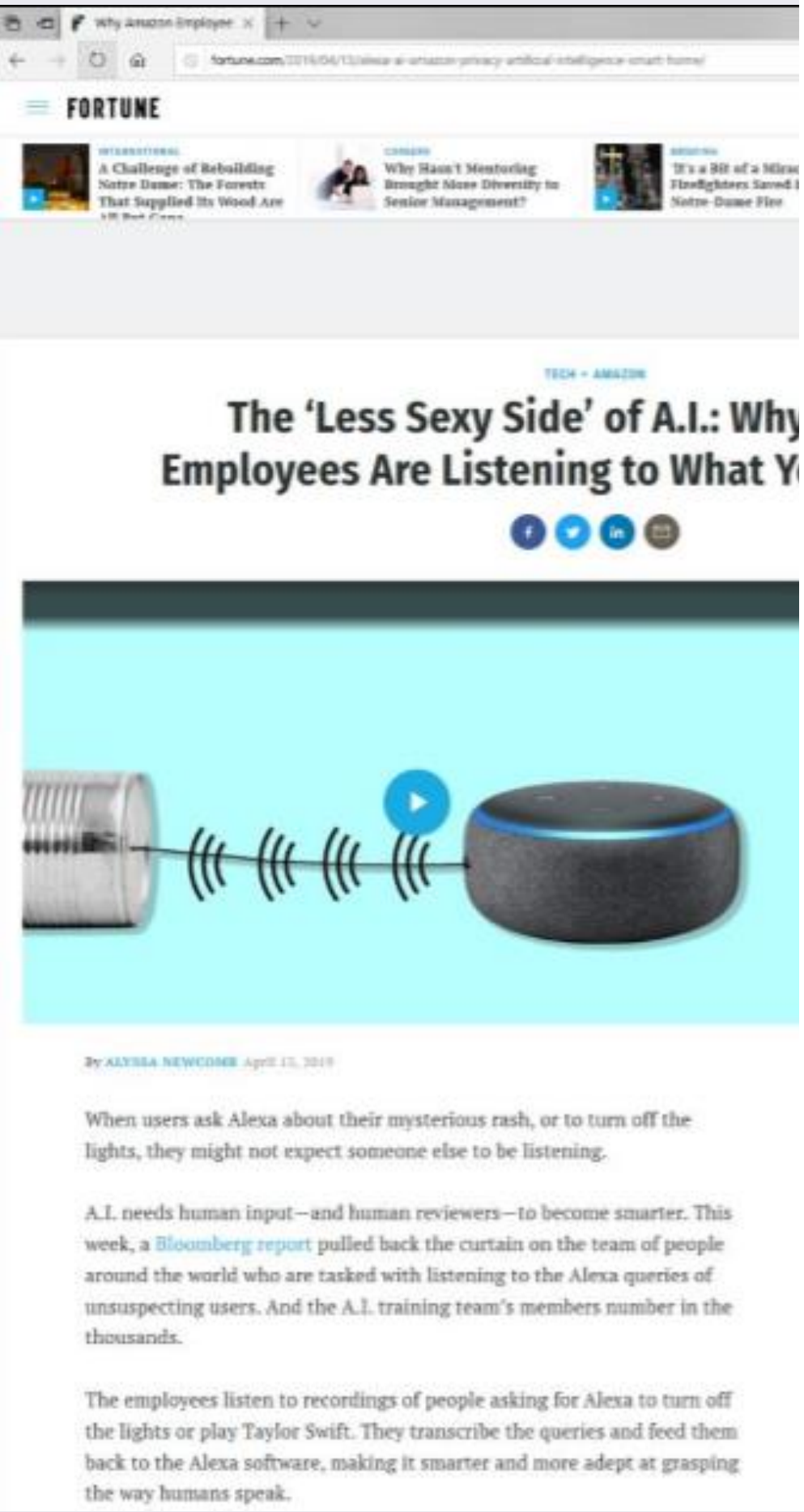
Cooperación en mitigación y remedio

Ecosistema

La compañía está *directamente ligada* al daño mediante una relación de negocio

Influenciar o terminar la relación





The New York Times

Opinion | [THE PRIVACY PROJECT](#)

Insurers Want to Know How Many Steps You Took Today

TECHNOLOGY.ORG

SCIENCE & TECHNOLOGY NEWS DEVELOPMENTS COMPE

Advancement of artificial intelligence opens health data privacy to attack

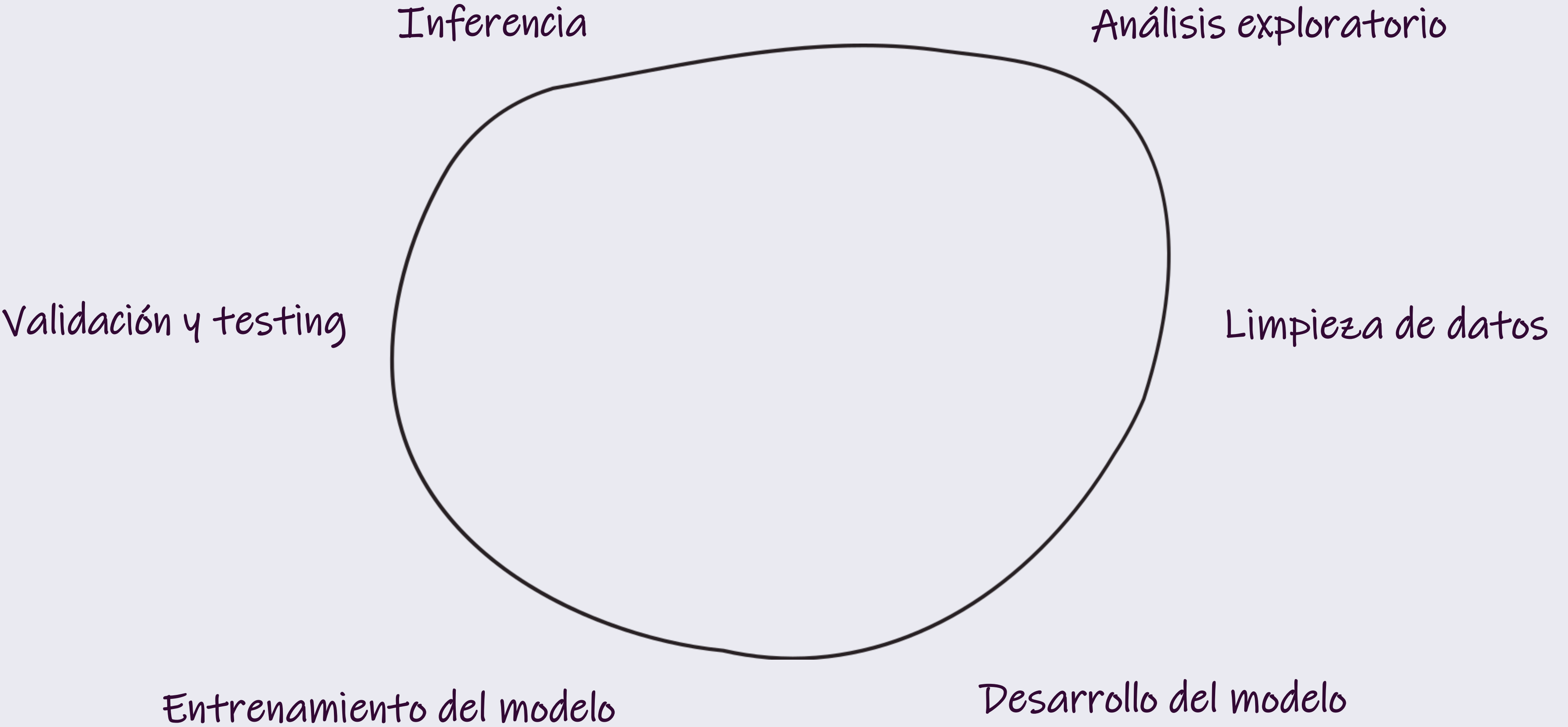
Posted December 27, 2018

Me gusta 0 Share

Advances in artificial intelligence have created new threats to the privacy of health data, a new UC Berkeley study shows.

The study, led by professor Anil Aswani of the Industrial Engineering & Operations Research Department (IEOR) in the College of Engineering and his team, suggests current laws and regulations are nowhere near sufficient to keep an individual's health status private in the face of AI development. The research was [released on JAMA Network Open](#).



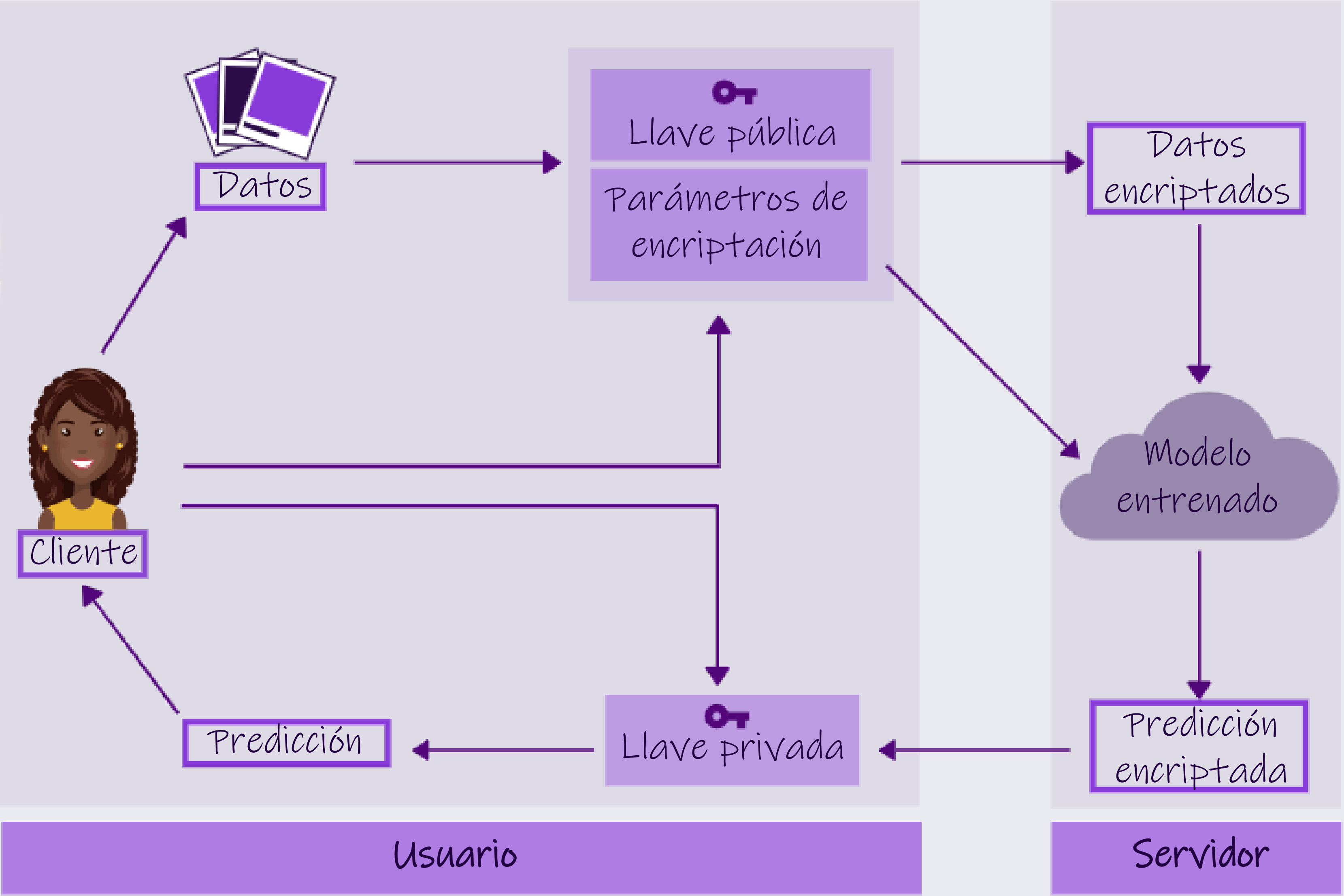


- Privacidad y seguridad de los conjuntos de datos
- Detección de sesgos durante el análisis exploratorio
- Protección de variables sensibles

Encriptación homomórfica

Tipo especial de encriptación

Permite realizar *operaciones* con datos encriptados



Differential Privacy

-  No se debería de saber nada de un individuo que no se pueda saber sin mirar el conjunto de datos. T Dalenius (1977)
- Garantiza que no se pueda hacer *inferencias* sobre individuos, si no sobre el *total*.
- Permite acceder a los datos sin comprometer la privacidad

Pseudo código

Entrada: X, eps

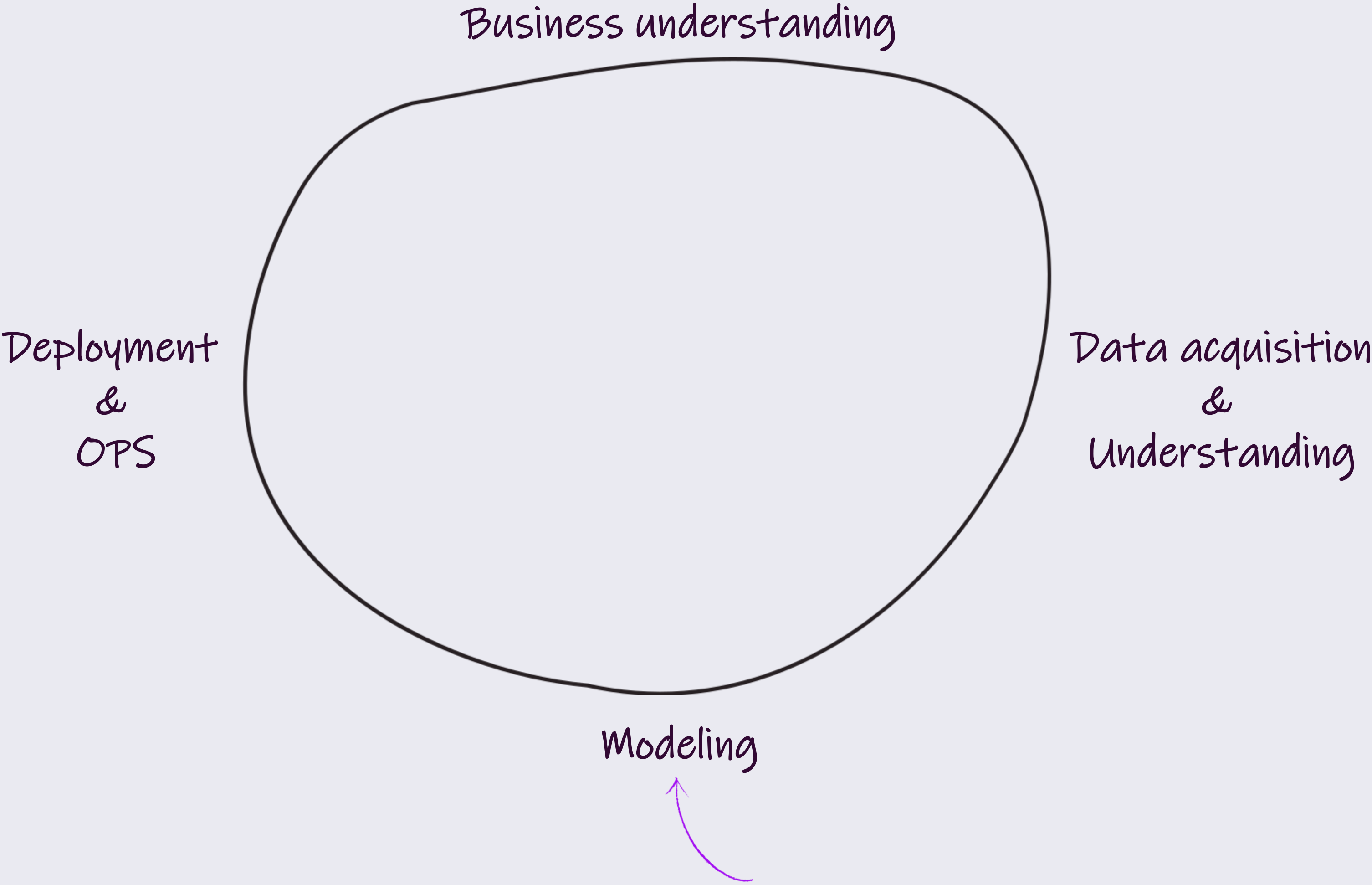
Calcular $F(X)$

Calcular la sensibilidad de F: S

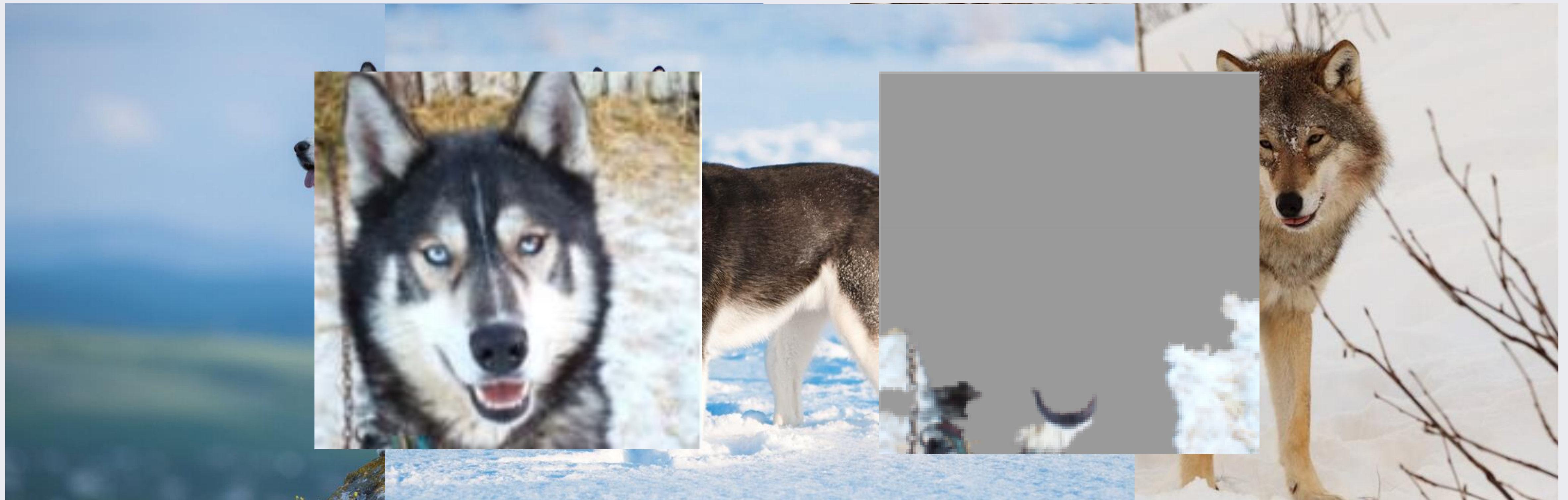
Extraer un ruido L de una distribución de Lapace con varianza:

$$2 \cdot (S/\text{eps})^2$$

Salida: $F(X) + L$



 Podemos fiarnos de nuestros modelos ?



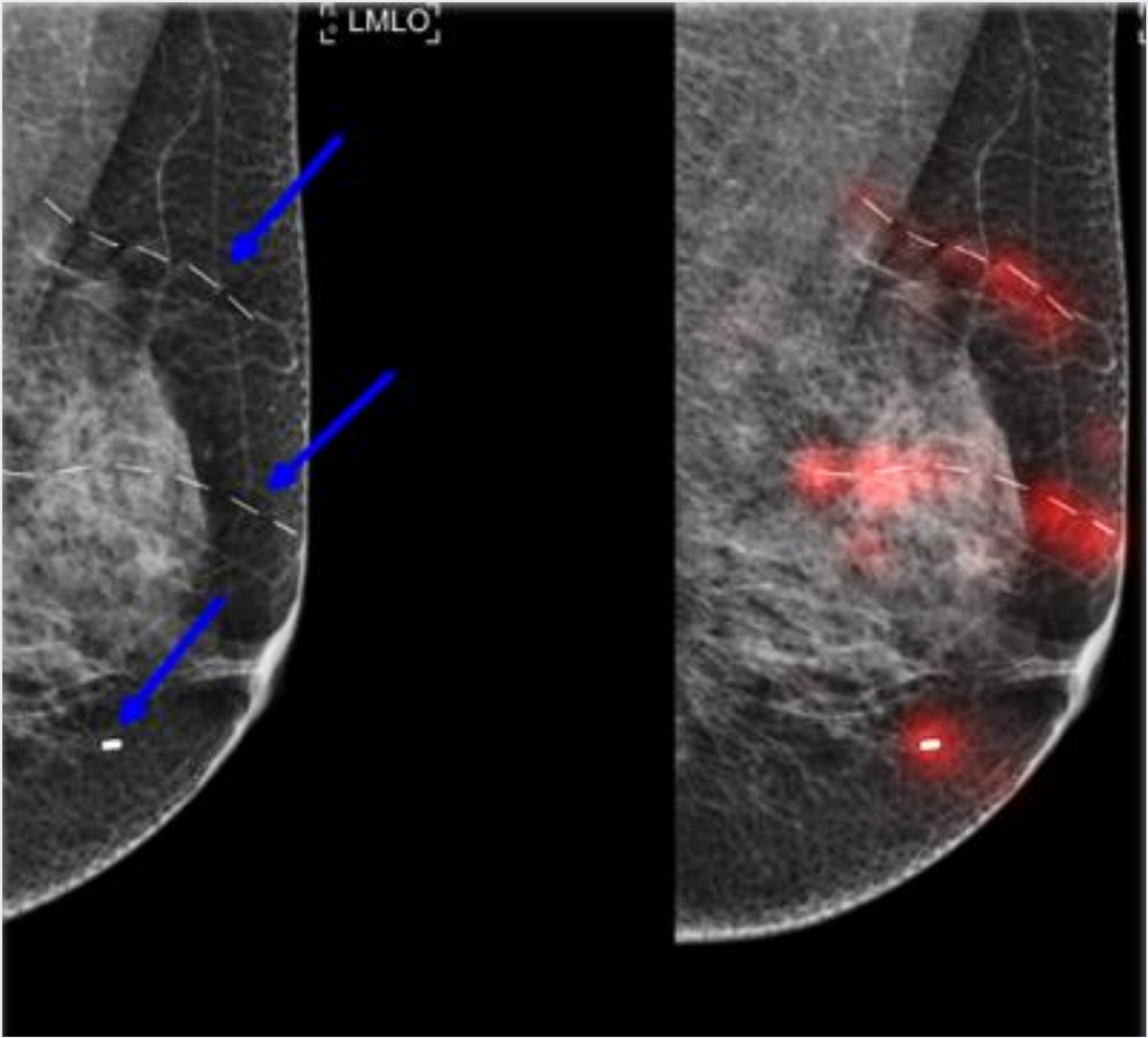
Husky El modelo se fija en si hay **nieve** para predecir que hay un **lobo** Lobo

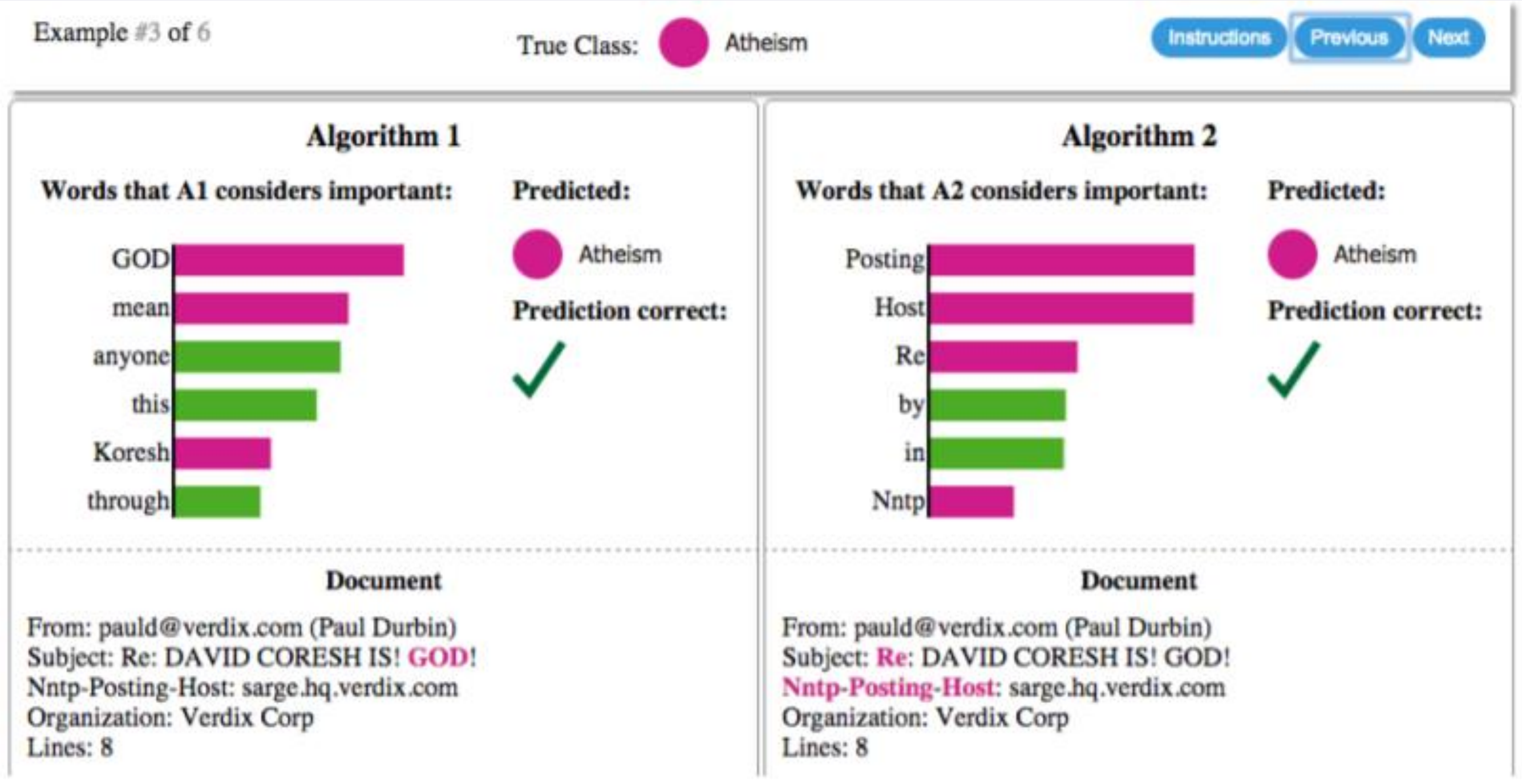
Beneficios

- Inspirar *confianza*
- *Mejora* de los modelos
- Detección de posibles *sesgos*
- Explicación de los *resultados* obtenidos por el modelo
- *Comparativa* de modelos

- Explicación global
¿Podemos fiarnos del *modelo*?
- Explicación local
¿Podemos fiarnos de una *predicción individual*?



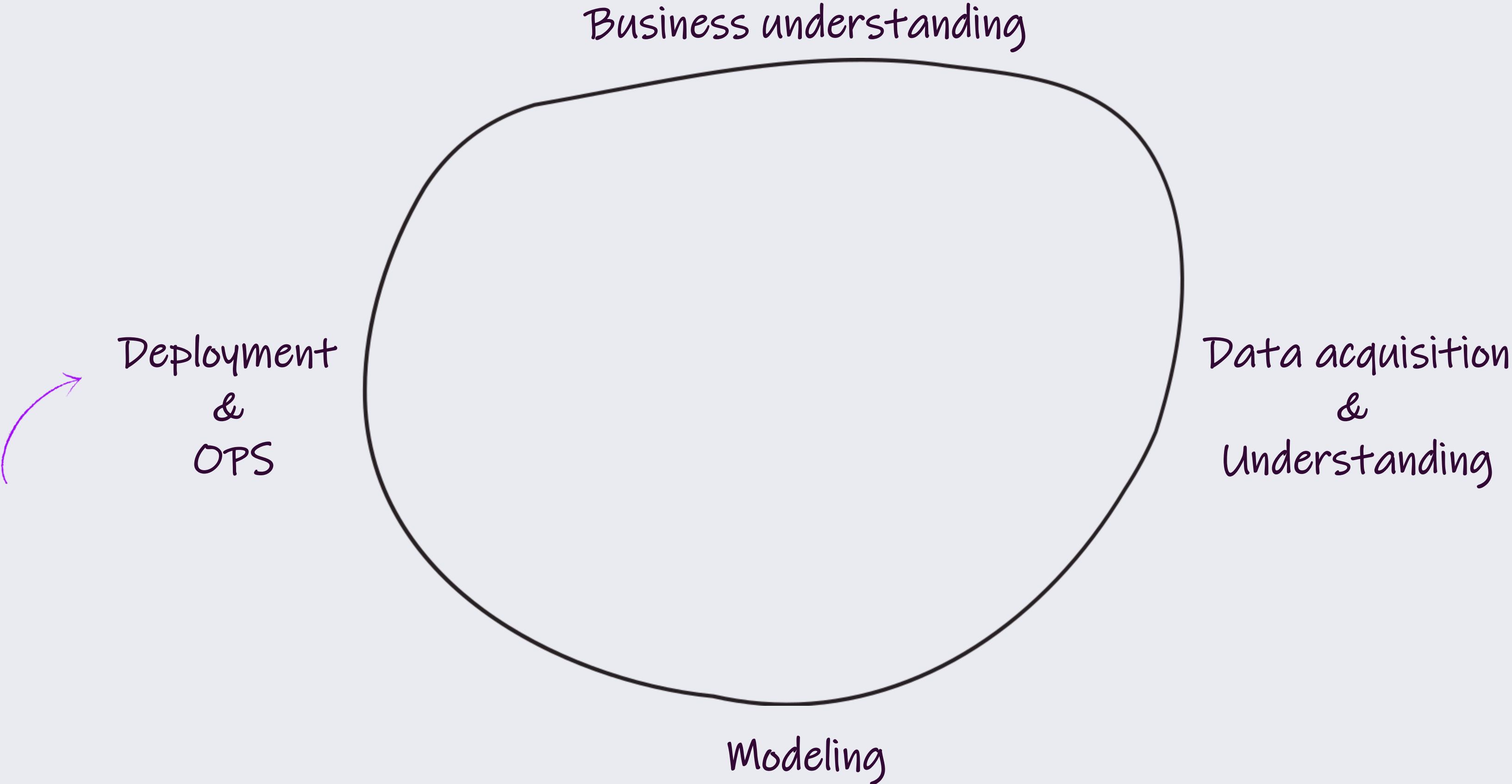




¿Cristianismo o ateísmo?

- Modelos simples
Modelos *entendibles* por los *humanos*
Regresión lineal, Naïve-Bayes, ...
- Modelos caja negra
Extracción de *explicaciones a posteriori*
LIME, análisis de sensibilidad, ...

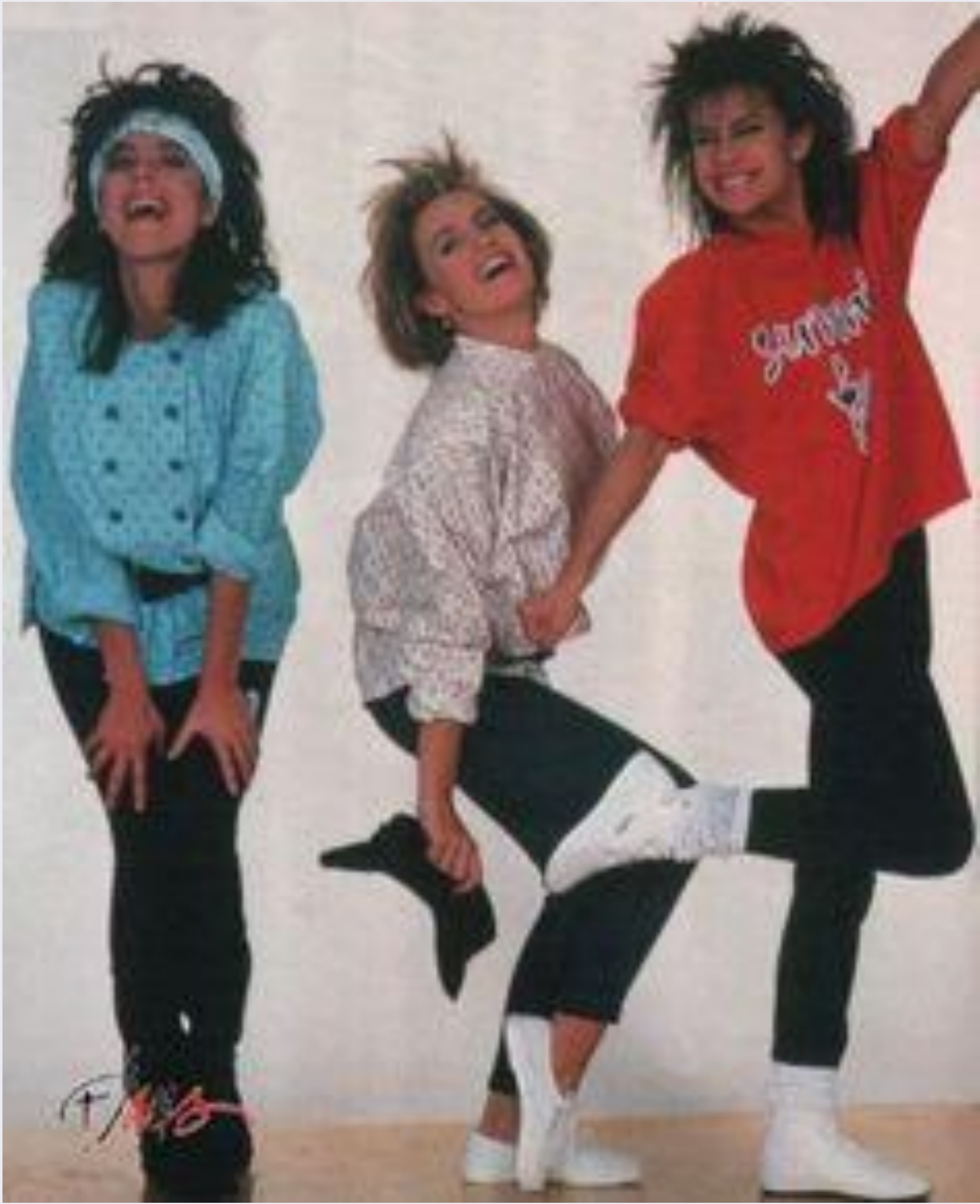


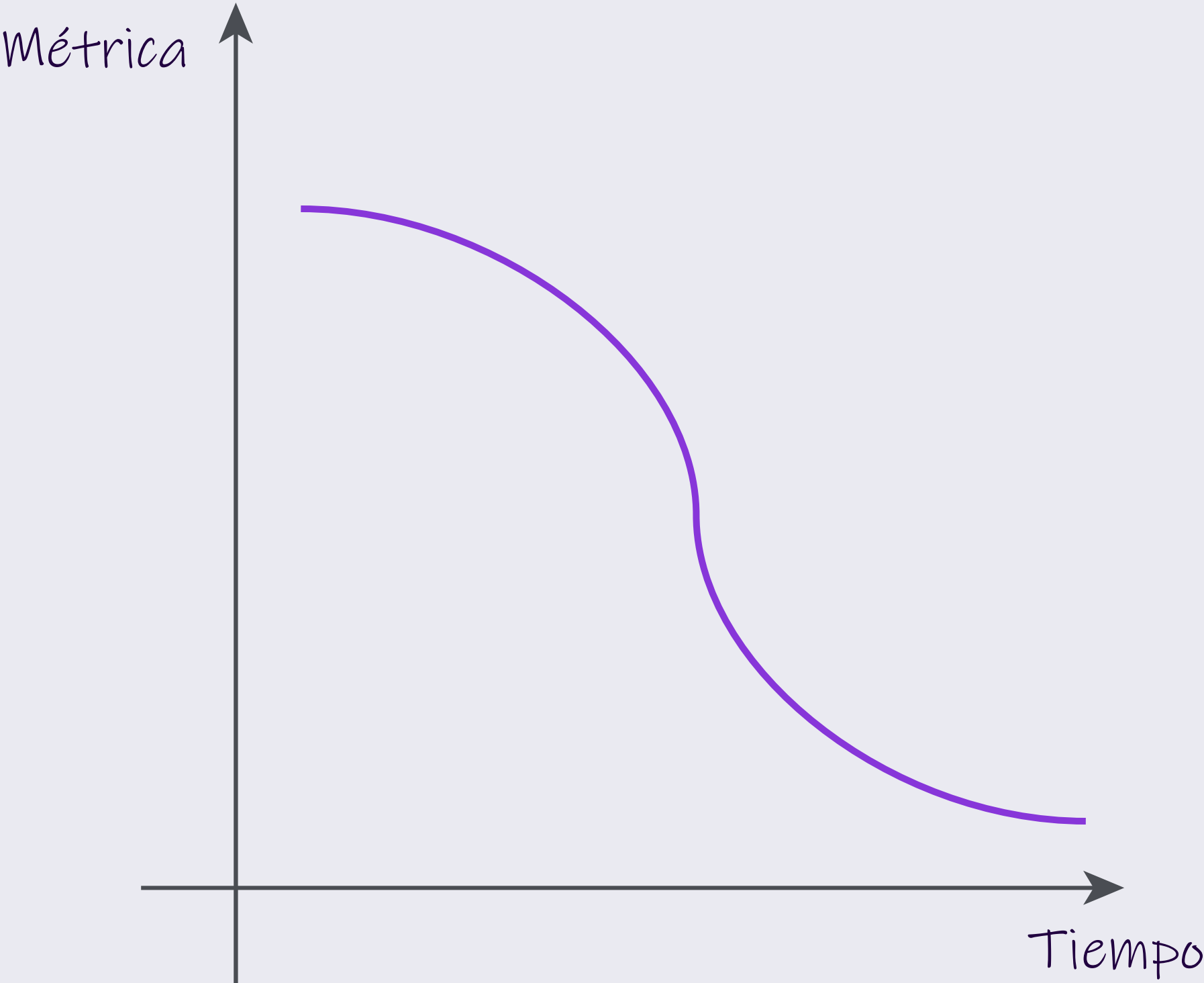




El mejor modelo, entrenado con *datos distintos* a los que realmente se enfrenta en el mundo real, *funcionará* en el mejor de los casos *de manera* poco confiable y en el peor de los casos *catastrófica*

David Talby





Data drift

- *Monitorización* de la performance del modelo
- Envío de correos con alertas
- Detección de la variable que lo está causando
- *Visualización* de lo que está ocurriendo

Questions & Answers



Thanks and ...
See you soon!

Thanks also to the sponsors.
Without whom this would not have been possible.

plain concepts 

