

# Sesgos en los modelos



# ¡¿Estoy generando desigualdad con mis modelos?!

Hablemos sobre sesgos en la inteligencia artificial y cómo evitarlo"



¡Hola!



**Maris Botero**

in/marisbotero

**Sharon Camacho**

in/sharoncamachog

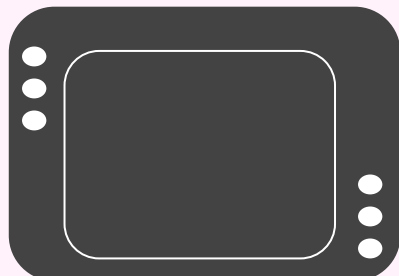
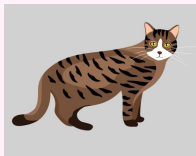
01

**Empecemos**





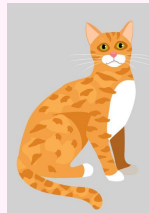
# Modelos



**PERRO**



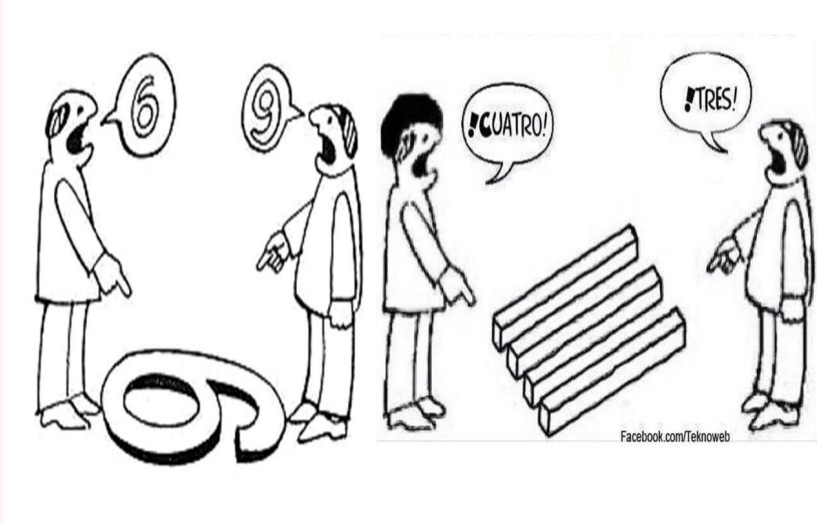
**GATO**



**GATO**



# ¿Qué son los sesgos?



# ¿Qué son los sesgos?

Estadística:  
“la diferencia  
entre el valor  
esperado de un  
estimador y su  
estimado”

Mayor sesgo



Menor sesgo



Menor varianza Mayor varianza



# ¿Qué son los sesgos?



recopilación de datos

psicología


ciencias sociales

“prejuicio”



## Sesgo Algorítmico:

“cualquier fenómeno  
que implica una  
*influencia excesiva de  
condiciones pasadas  
-irrelevantes- o  
decisiones actuales*”





# Sistema dual de pensamiento



## Sistema automático

Los seres humanos somos predeciblemente irracionales

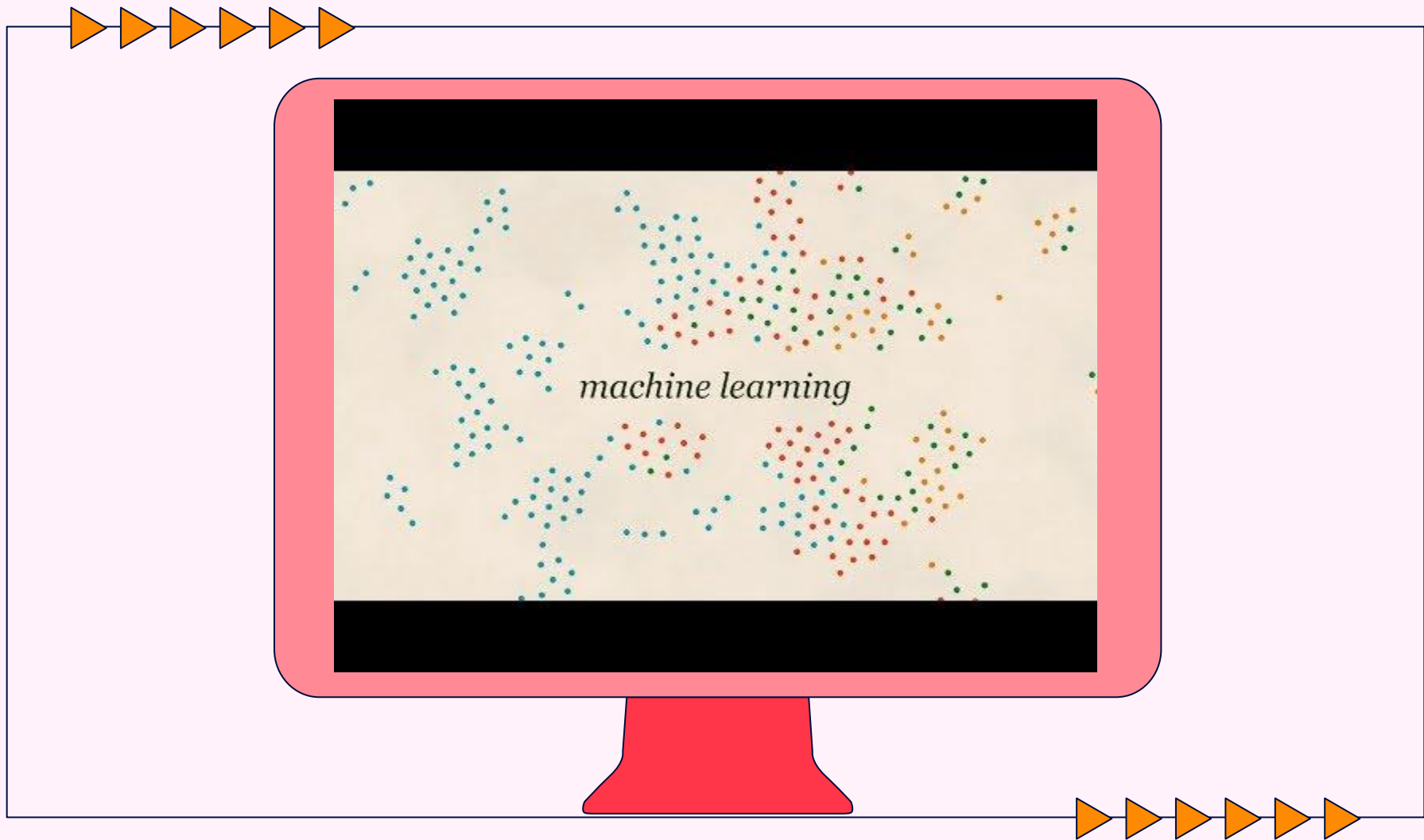


## Sistema reflexivo

Implica tener una relación más consciente con nosotros mismos y con todo lo que nos rodea

**BIAS**

**BIAS EVERYWHERE**

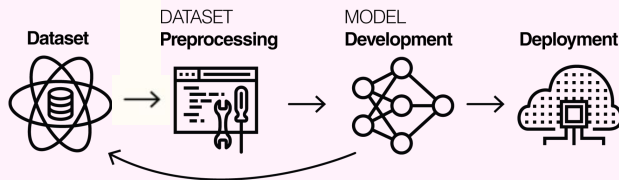


# ¿Cómo afectan los sesgos inconscientes en la analítica?

## ¿Para qué hacemos analítica?

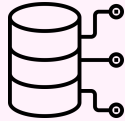
- Para descubrir, interpretar, y comunicar patrones y tendencias de los datos.
- Para tomar mejores decisiones.

## ¿Qué pasa cuando no hay datos?





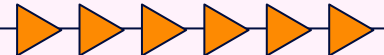
# TIPOS



**Datos**



**Interpretabilidad**



# Datos

## Selección

No aleatoria

## Exclusión

Limpieza

## Muestreo

## Medición

## Prejuicio\*

## Espacial





# Interpretabilidad



## Observador

Ver lo que queremos ver

## Falacia de la correlación



Correlación  $\neq$  Causación

## Sobregeneralización

Generalizar sobre una data  
limitada

## Auto-Bias

Prevalece la decisión del AI sobre la  
del persona



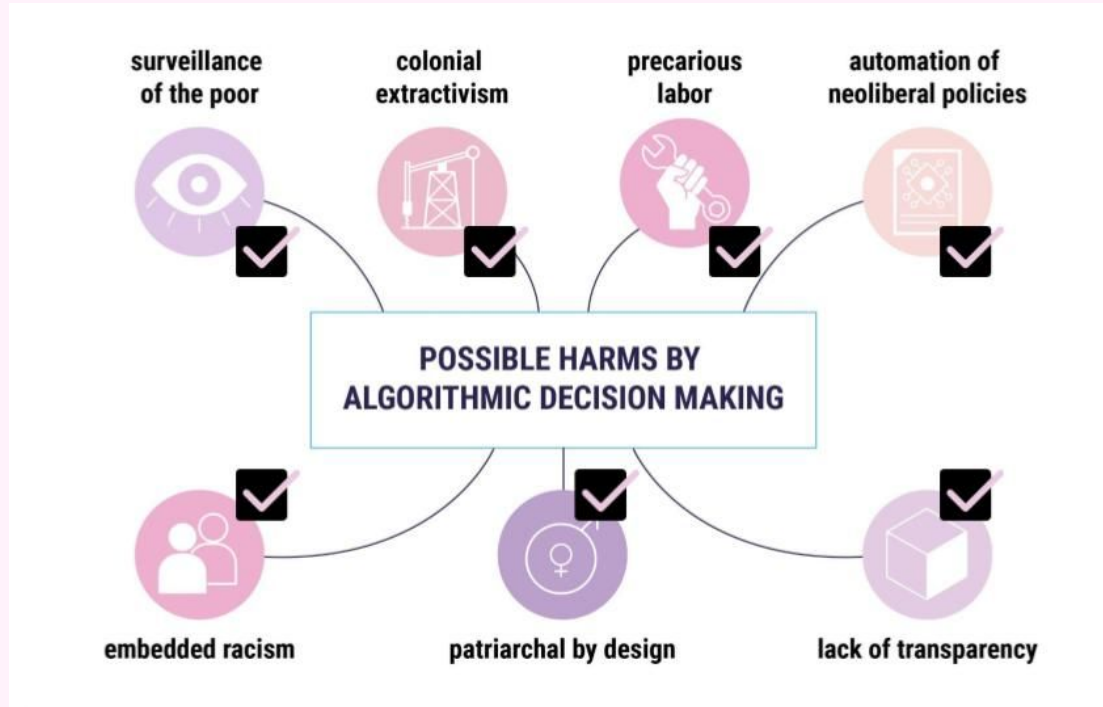


**¿Y esto por qué es importante?**





# Posibles daños a partir de estos sesgos



Fuente: <https://notmy.ai/2021/05/03/case-study-plataforma-tecnologica-de-intervencion-social-argentina-and-brazil/>



# Algunos ejemplos

## **Amazon prescinde de una inteligencia artificial de reclutamiento por discriminar a las mujeres**

El sistema había sido entrenado con los perfiles de los solicitantes de empleo de los últimos 10 años

**‘Unseen bias causes loan applications by women entrepreneurs to be delayed or rejected more often’**



# El algoritmo que ‘adivina’ los delitos futuros falla tanto como un humano

El sofisticado programa COMPAS analiza la posibilidad de reincidir de un millón de convictos reales

## MIT Researcher Exposing Bias in Facial Recognition Tech Triggers Amazon's Wrath

By Matt O'Brien | April 8, 2019



Select photo



X The photo you want to upload does not meet our criteria because:

- Subject eyes are closed

Please refer to the technical requirements.  
You have 9 attempts left.

Check the photo [requirements](#).

Read more about [common photo problems and how to resolve them](#).

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

Reference number: 20161206-81

Filename: Untitled.jpg



FACEBOOK

# Disculpas de Facebook por etiquetar por error a hombres negros como 'primates'



# Más ejemplos

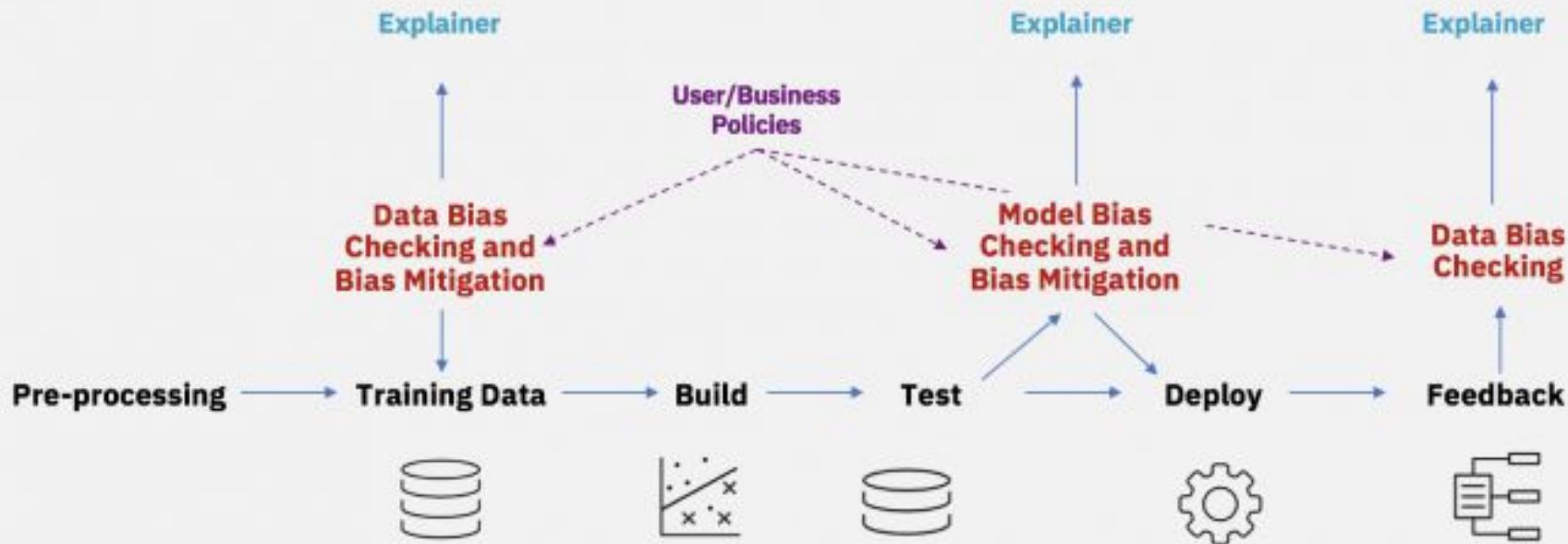
- Sistemas de reconocimiento de voz
- Sistemas de traducción automática y en general del procesamiento del lenguaje natural
- Chatbots



¿Qué podemos hacer?











# Algunas recomendaciones



## Data

¡Revisa tus datos!

## Equipo

Diversidad

## Validación

Constante

## Funcionamiento

Para todos

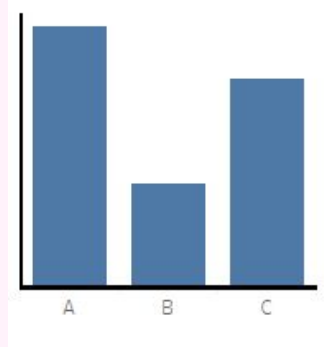


# Mitigar los sesgos

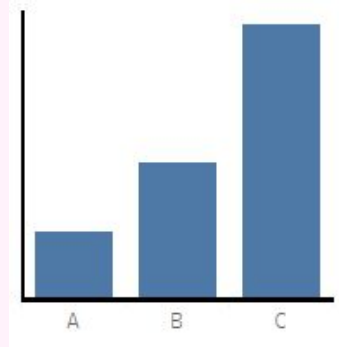
- Antes del entrenamiento:  
Evaluar sesgos en el dataset y equilibrar
- Durante: optimizar métricas de evaluación de sesgos
- Después: Evaluar sesgos e informar



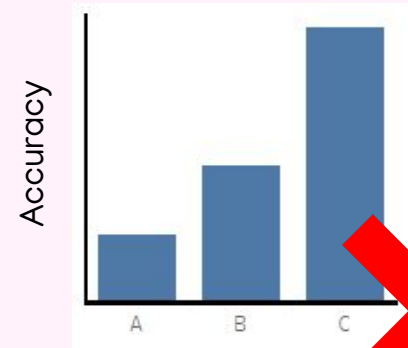
Frecuencia en  
la realidad



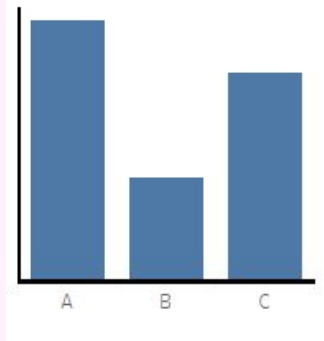
Frecuencia en  
el dataset



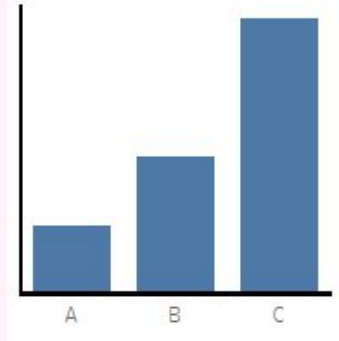
Accuracy del  
Modelo



Frecuencia en  
la realidad

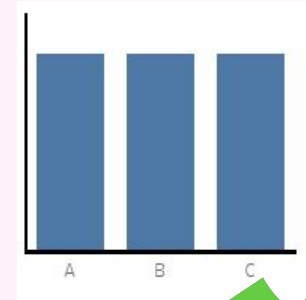


Frecuencia en  
el dataset



Accuracy del  
Modelo

Accuracy







# Inconsciencia temporal



## El Yo futuro

Nuestro Yo presente pospone  
las cosas pensando que  
nuestro Yo futuro lo va a  
resolver





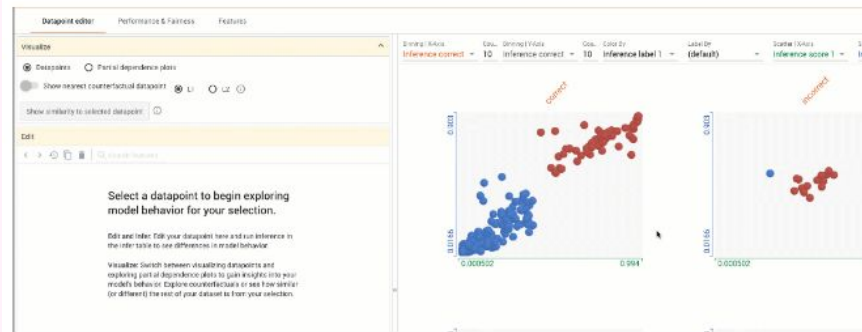
**MUCHA GENTE PEQUEÑA  
EN LUGARES PEQUEÑOS,  
HACIENDO COSAS PEQUEÑAS,  
PUEDEN CAMBIAR EL MUNDO.**

Eduardo Galeano



Visually probe the behavior of **trained machine learning models**, with minimal coding.

GET STARTED



## AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

[Python API Docs](#)

[Get Python Code](#)

[Get R Code](#)



# Let's code



```
31     self.file = None
32     self.fingerprints = set()
33     self.logdups = True
34     self.debug = debug
35     self.logger = logging.getLogger(__name__)
36     if path:
37         self.file = open(os.path.join(path, 'requests.log'),
38                         'a')
39         self.file.seek(0)
40         self.fingerprints.update(x.request() for x in self.requests)
41
42     @classmethod
43     def from_settings(cls, settings):
44         debug = settings.getbool('SUPPRESS_LOG_MESSAGES')
45         return cls(job_dir(settings), debug)
46
47     def request_seen(self, request):
48         fp = self.request_fingerprint(request)
49         if fp in self.fingerprints:
50             return True
51         self.fingerprints.add(fp)
52         if self.file:
53             self.file.write(fp + os.linesep)
54
55     def request_fingerprint(self, request):
56         return request_fingerprint(request)
```

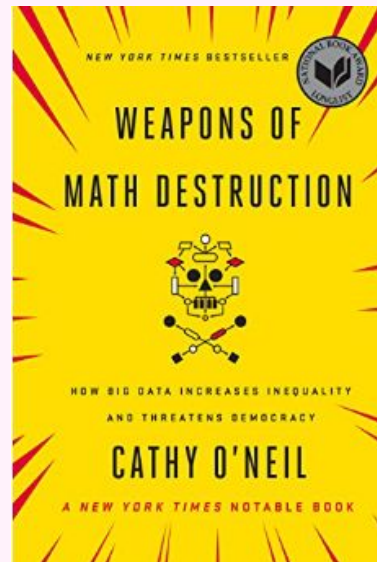
Notebook





# ★ Recursos:

- <https://huggingface.co/spaces/society-ethics/disaggregators>
- <https://huggingface.co/blog/evaluating-llm-bias>
- <https://huggingface.co/datasets/MilaNLProc/honest>
- <https://towardsdatascience.com/analysing-fairness-in-machine-learning-with-python-96a9ab0d0705>
- Guía sobre la implementación de un aprendizaje automático inclusivo: AutoML
- What-if-tool
- AI Fairness
- Cómo usar what-if-tool
- Estructura Latente aprendida
- Adversarial Learning
- The Algorithmic Justice League





**Muchas gracias**

