Privacidad y aspectos éticos en la ciencia de datos

Lorena Etcheverry (lorenae@fing.edu.uy)
Instituto de Computación, FING, UdelaR

Privacidad y anonimización

https://andertoons.com/privacy/cartoon/8229/i-like-the-privacy-but-it-does-make-it-hard-to-see

La privacidad desde un enfoque conceptual

- La privacidad de datos se refiere al derecho y la capacidad de un individuo para controlar la recopilación, uso y divulgación de su información personal.

- Implica garantizar que la información personal de las personas se maneje de manera segura y responsable, minimizando el riesgo de acceso no autorizado o uso indebido.

Conceptos vinculados a la privacidad de datos

Minimización de Datos:

La minimización de datos es el principio de recopilar y retener solo la información personal necesaria para cumplir con un propósito específico. Cuantos menos datos personales se recopilen y almacenen, menor es el riesgo potencial para la privacidad.

Consentimiento Informado:

El consentimiento informado es el proceso mediante el cual las personas otorgan su aprobación para que sus datos personales sean recopilados, procesados o compartidos. Debe ser un proceso transparente en el que se explique claramente cómo se utilizarán los datos y cuáles serán los posibles impactos.

Transparencia:

La transparencia se refiere a la obligación de las organizaciones de informar a las personas sobre cómo se recopilan, utilizan y comparten sus datos. Una comunicación clara y comprensible sobre las prácticas de privacidad es esencial para establecer la confianza.

La privacidad desde un enfoque técnico

Desidentificación:

no debería ser posible volver a identificar a ninguna persona

Confidencialidad o secreto:

los datos divulgados no deben revelar información confidencial relacionada con ningún individuo específico

RIESGOS

divulgación de identidad

divulgación de atributos

La privacidad como una propiedad de los datos

Pseudonimización:

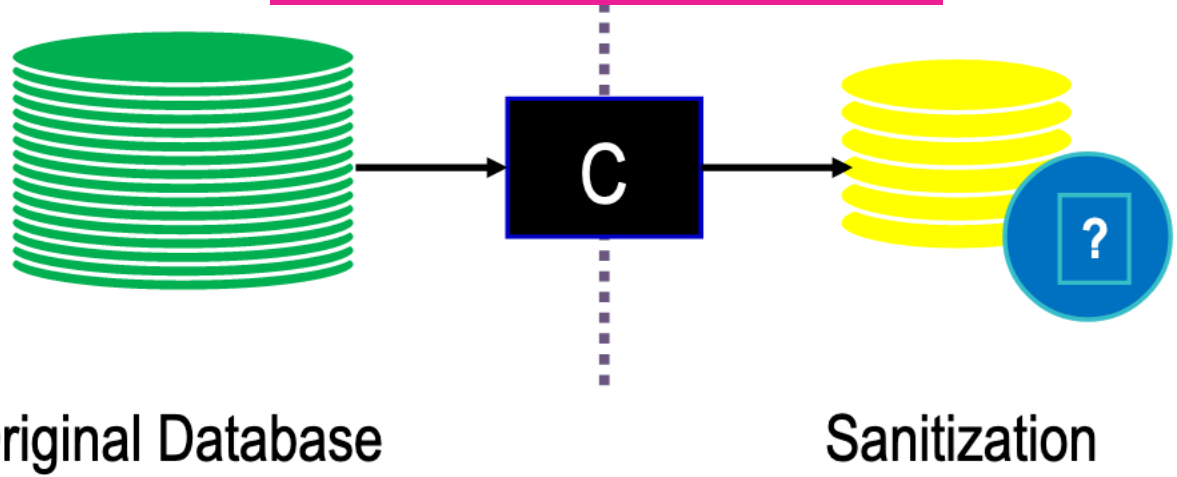
La pseudonimización implica reemplazar los identificadores directos de los datos con identificadores ficticios o pseudónimos. Aunque se conserva la posibilidad de realizar análisis, es más difícil vincular los datos con individuos reales sin información adicional.

Anonimización:

La anonimización es el proceso de eliminar o modificar ciertos atributos en los datos personales para que la información no pueda asociarse directamente con un individuo en particular. El objetivo es proteger la privacidad al tiempo que se permite el análisis de datos de manera agregada.

Pseudonimización

Anonimización



La anonimización busca esconder a cada individuo dentro de un grupo

Modelos de privacidad

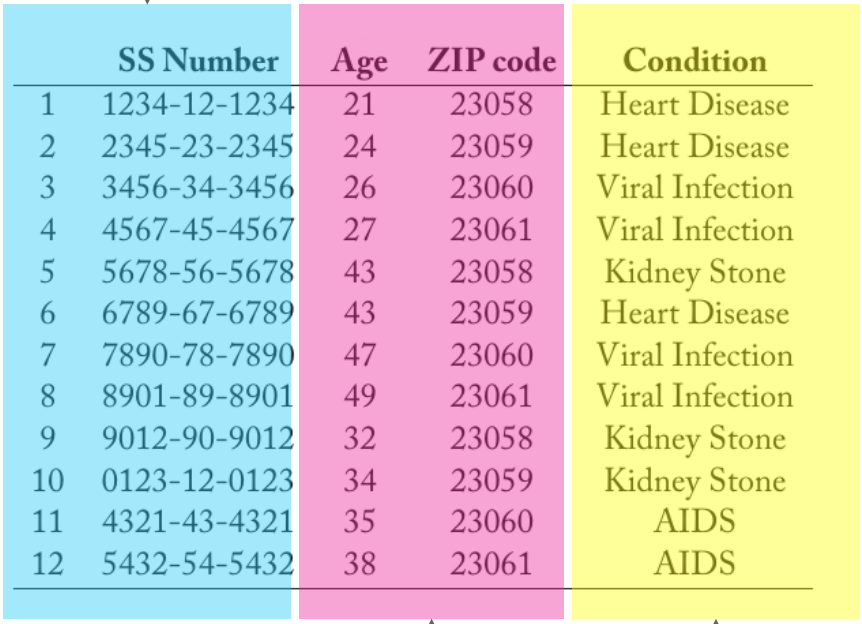
- Establecen propiedades que los datos deben cumplir para reducir ciertos riesgos.
- Diferentes algoritmos para transformar los datos y satisfacer las propiedades

- Modelos: k-anonymity, t-closeness, l-diversity

Clasificación de atributos

Cuatro clases de atributos:

(1) identificadores
(2) cuasi-identificadores,
(3) sensibles,
(4) todos los demás.



	SS Number	Age	ZIP code	Condition
1	1234-12-1234	21	23058	Heart Disease
2	2345-23-2345	24	23059	Heart Disease
3	3456-34-3456	26	23060	Viral Infection
4	4567-45-4567	27	23061	Viral Infection
5	5678-56-5678	43	23058	Kidney Stone
6	6789-67-6789	43	23059	Heart Disease
7	7890-78-7890	47	23060	Viral Infection
8	8901-89-8901	49	23061	Viral Infection
9	9012-90-9012	32	23058	Kidney Stone
10	0123-12-0123	34	23059	Kidney Stone
11	4321-43-4321	35	23060	AIDS
12	5432-54-5432	38	23061	AIDS

k-anonymity [Sweeney1998]

Modelo de ataque: re-identificación de los registros, asume que con los cuasi-identificadores alcanza para identificar un registro

Propiedad: que al menos k registros compartan los mismos valores en los cuasi-identificadores

Se puede lograr con supresión y generalización

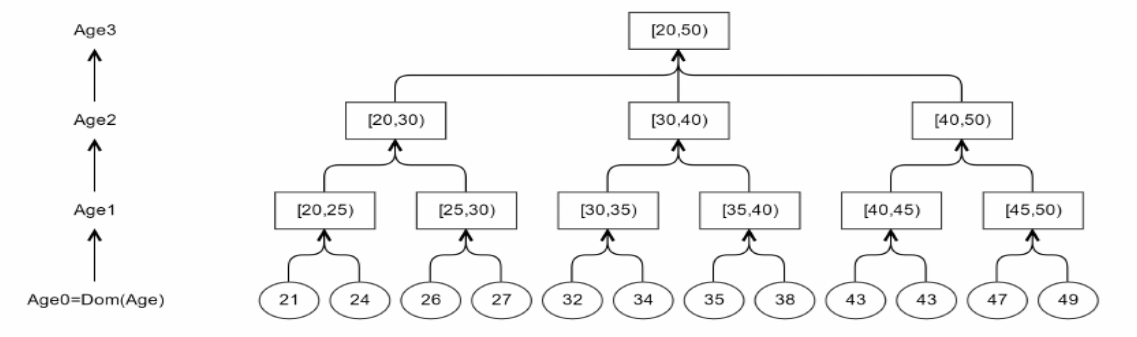
Supresión

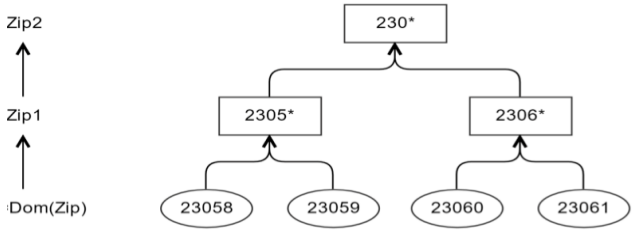
Generalización

	SS Number	Age	ZIP code	Condition
1	1234-12-1234	21	23058	Heart Disease
2	2345-23-2345	24	23059	Heart Disease
3	3456-34-3456	26	23060	Viral Infection
4	4567-45-4567	27	23061	Viral Infection
5	5678-56-5678	43	23058	Kidney Stone
6	6789-67-6789	43	23059	Heart Disease
7	7890-78-7890	47	23060	Viral Infection
8	8901-89-8901	49	23061	Viral Infection
9	9012-90-9012	32	23058	Kidney Stone
10	0123-12-0123	34	23059	Kidney Stone
11	4321-43-4321	35	23060	AIDS
12	5432-54-5432	38	23061	AIDS

	SS Number	Age	ZIP code	Condition
1	*	[20-30]	230**	Heart Disease
2	*	[20-30]	230**	Heart Disease
3	*	[20-30]	230**	Viral Infection
4	*	[20-30]	230**	Viral Infection
5	*	[40-50]	230**	Kidney Stone
6	*	[40-50]	230**	Heart Disease
7	*	[40-50]	230**	Viral Infection
8	*	[40-50]	230**	Viral Infection
9	*	[30-40]	230**	Kidney Stone
10	*	[30-40]	230**	Kidney Stone
11	*	[30-40]	230**	AIDS
12	*	[30-40]	230**	AIDS

Árboles de Generalización





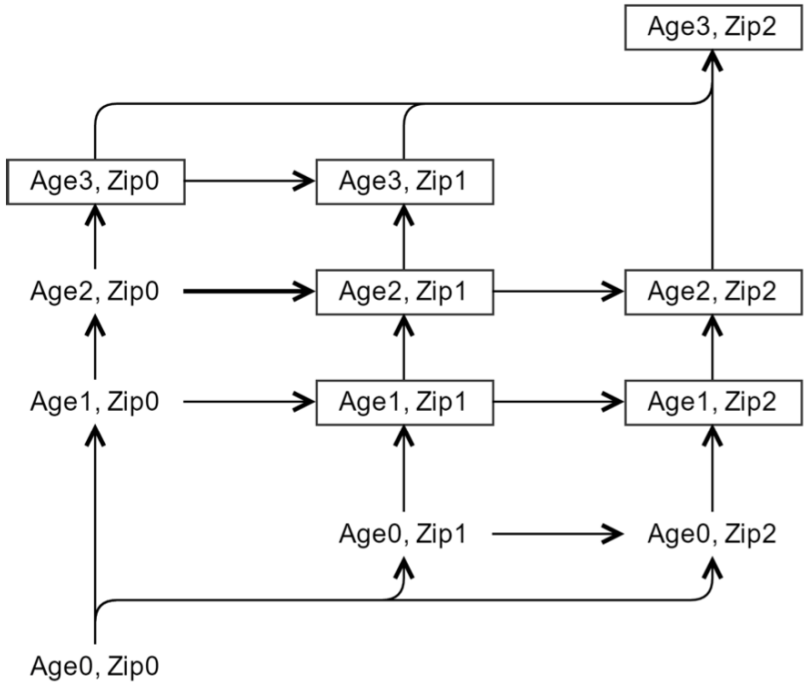
Los niveles y agrupamientos deben de tener sentido para el análisis (expertos de dominio)

Ej: ¿cuales son los rangos de edad que tienen sentido?

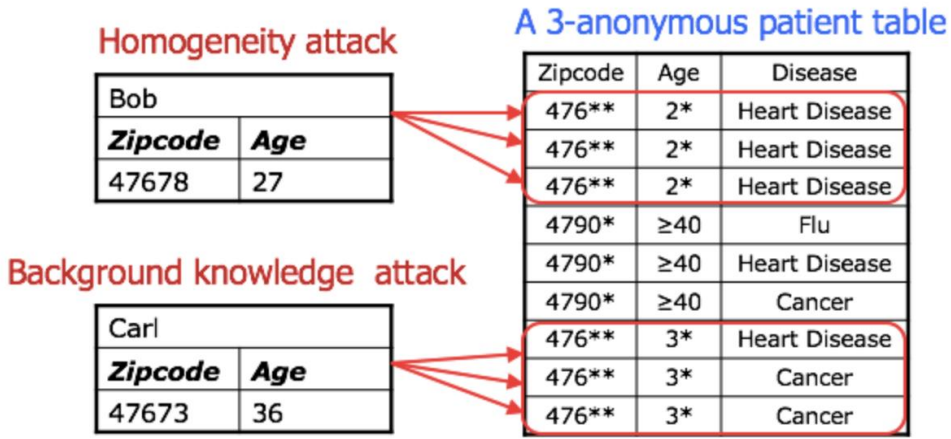
¿Cómo alcanzar k-anonymity?

El problema: determinar un nivel de cada árbol de generalización que satisfaga k-anonymity y minimice la pérdida de la información.

Es un problema NP-Hard



l-diversity[Machanavajjhala et al., 2007]



Cada grupo de registros satisface l-diversity si los valores del atributo sensible son lo *suficientemente variables* dentro de cada grupo.
La noción más sencilla, al menos l valores diferentes.

t-closeness [Li et al. 2007]

Ataque por similaridad: puedo inferir el rango de sueldo y el tipo de enfermedad de Bob

t-closeness requiere que la distribución del atributo sensible dentro de cada grupo sea similar a su distribución en el dataset

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Resumiendo

Diferentes modelos de privacidad por anonimización

Cada uno apunta a mitigar distintos riesgos

Múltiples técnicas y algoritmos para transformar los datos y alcanzar esos modelos

Pero la anonimización no alcanza para asegurar la privacidad

El caso del premio Netflix (2007)

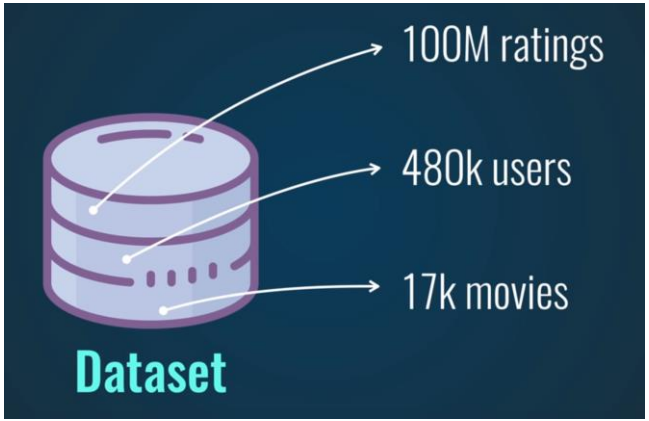
Desafío: predecir el rating de contenidos

Netflix libera 1/10 de su base de datos

Una fracción de las preferencias por usuario (centenas)

Identificadores de usuario anonimizados

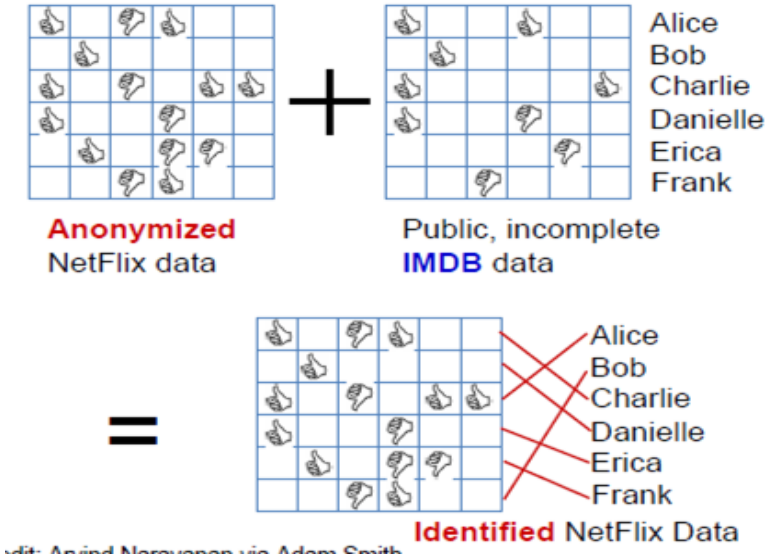
Preferencias perturbadas



Kearns, M., & Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Linkage attack

Combinan con IMDB
Algoritmo de similitud ScoreboardH
Con 8 películas + fechas con error < 14 días el 99% identificados (1 solo match)
Preferencias políticas o sexuales se pueden deducir fácilmente de los datos.



Narayanan, A., & Shmatikov, V. (2008, May). **Robust de-anonymization of large sparse datasets**. In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE.

La privacidad como una propiedad de las consultas

No tiene que ver con “re-identificación”.

Es completamente general, independiente de:

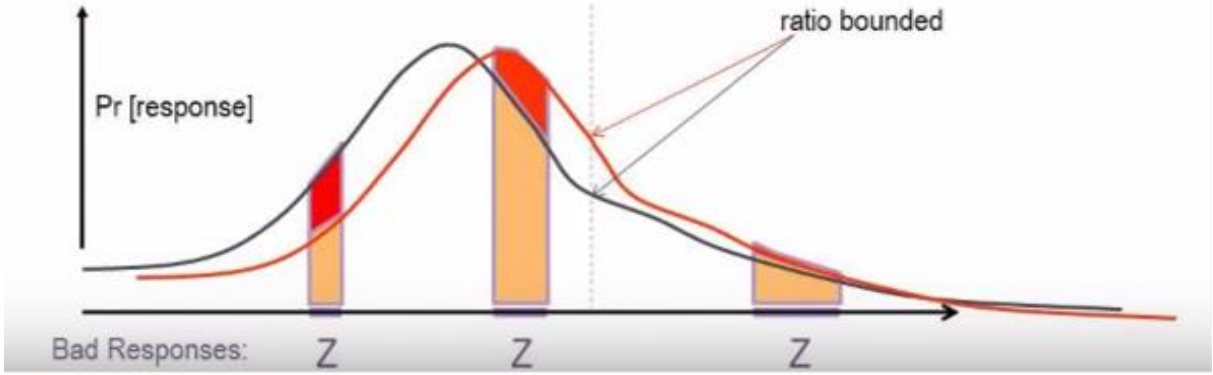
- La naturaleza de los atributos
- El tipo de conclusiones que saquemos
- Lo que sepa el atacante por otros medios (info auxiliar)

Hay una definición matemática y formal atrás

Differential Privacy [Dwork et al. 06]

M gives ϵ-differential privacy if for all pairs of databases x, x' differing in one row, and all subsets C of possible outputs

$$\Pr[M(x) \in C] \leq e^{\epsilon} \Pr[M(x') \in C]$$



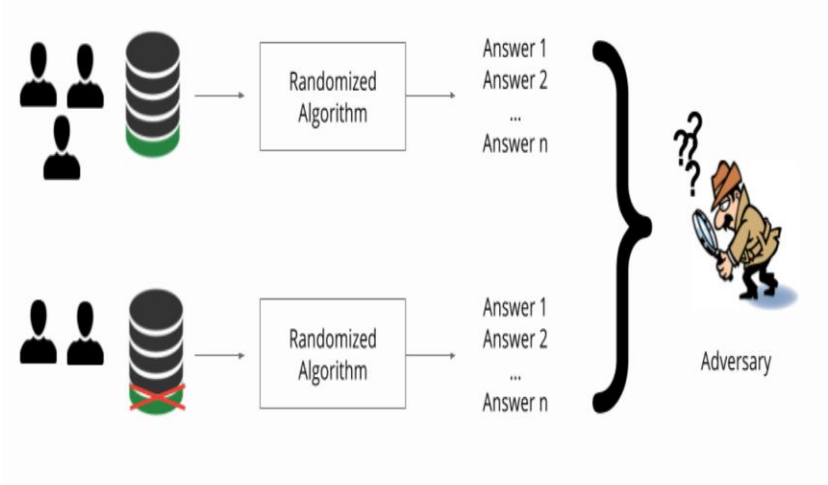
Differential Privacy y Machine Learning

- Existen dos enfoques principales:
    - **differential privacy global**, que agrega ruido a la salida del modelo
    - **differential privacy local**, que agrega ruido a cada dato individual antes de enviarlo al algoritmo
- **Desafíos:** trade-off entre precisión y privacidad, el overhead computacional, y la elección del presupuesto de privacidad

Differentially private stochastic gradient descent (DP-SGD) y Private Aggregation of Teacher Ensembles (PATE) son algunos enfoques relevantes.

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., ... & Thakurta, A. G. (2023). How to dp-fy ml: A practical guide to machine learning with differential privacy. Journal of Artificial Intelligence Research, 77, 1113-1201

Differential Privacy y Machine Learning



Privacy and machine learning: two unexpected allies?,Nicolas Papernot and Ian Goodfellow, 2018

Privacidad y Machine Learning

La tendencia actual es el diseño de algoritmos que incorporen la privacidad, fundamentalmente en las etapas de entrenamiento

Ji, Zhanglong, Zachary C. Lipton, and Charles Elkan. "Differential privacy and machine learning: a survey and review." arXiv preprint arXiv:1412.7584 (2014).

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kinal Talwar, and Li Zhang. Deep learning with differential privacy. ACM CCS 2016

Papernot, Nicolas, et al. "Semi-supervised knowledge transfer for deep learning from private training data." arXiv preprint arXiv:1610.05755 (2016).

Ryffel, Theo, et al. "A generic framework for privacy preserving deep learning." arXiv preprint arXiv:1811.04017 (2018).

Muchas propuestas

Algunas herramientas

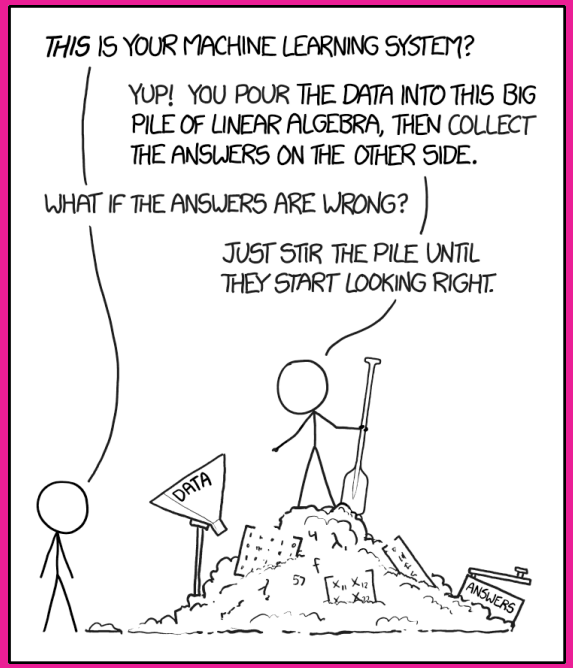
PySyft is a Python library for secure and private Deep Learning. PySyft decouples private data from model training, using Federated Learning, Differential Privacy, and Encrypted Computation (like Multi-Party Computation (MPC) and Homomorphic Encryption (HE)) within the main Deep Learning frameworks like PyTorch and TensorFlow. Join the movement on Slack.

TensorFlow Privacy

This repository contains the source code for TensorFlow Privacy, a Python library that includes implementations of TensorFlow optimizers for training machine learning models with differential privacy. The library comes with tutorials and analysis tools for computing the privacy guarantees provided.

The TensorFlow Privacy library is under continual development, always welcoming contributions. In particular, we always welcome help towards resolving the issues currently open.

Aspectos éticos: equidad y discriminación en la ciencia de datos

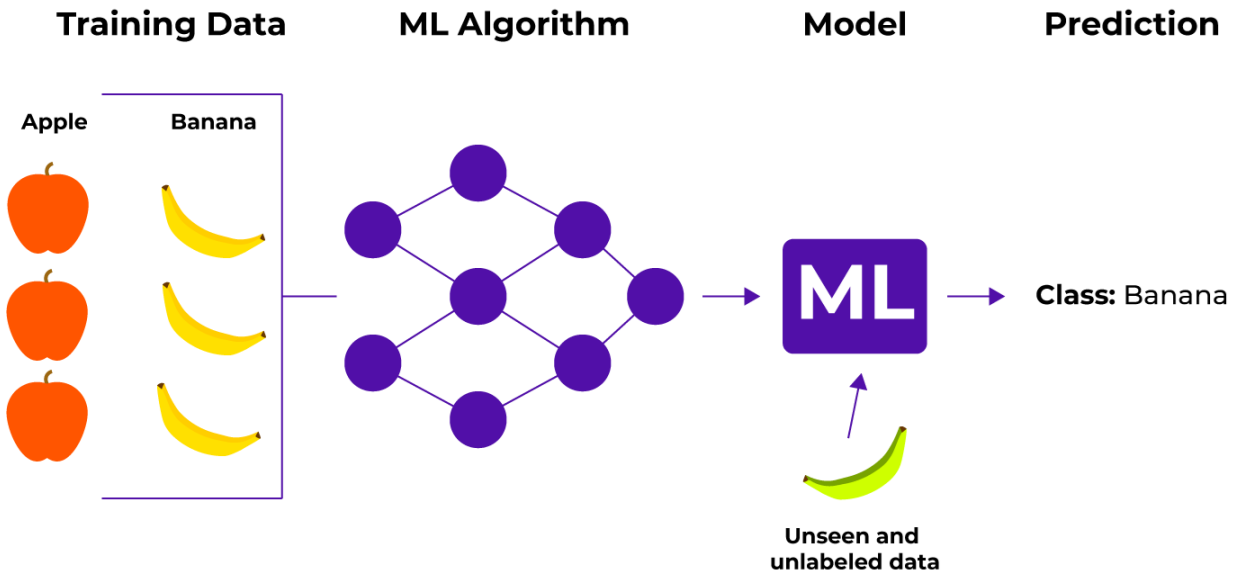


https://xkcd.com/1838/

Las promesas de la Ciencia de Datos y la Inteligencia Artificial

Mejorar la vida de las personas, (ej. sistemas recomendadores)
Acelerar los descubrimientos científicos, (ej., medicina),
Transformar la sociedad, (ej. *open government*),
Optimizar los negocios (ej. marketing personalizado), entre otras

Para esto se utilizan técnicas de análisis de datos, en particular algoritmos de Aprendizaje Automático (a.k.a Inteligencia Artificial)



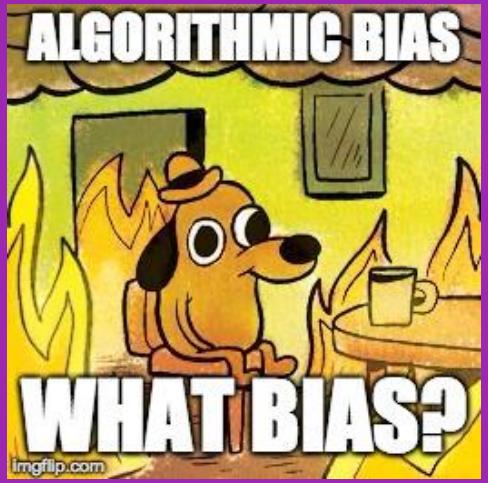
Aprendizaje Supervisado

Los algoritmos aprenden de los **datos** y construyen una representación de la **realidad** a partir de ellos.

Los **algoritmos** son programados por **humanos**, que les trasladan su forma de ver el mundo.

¿QUÉ REALIDAD?
¿CUÁN “FIEL” ES ESTA REPRESENTACIÓN?

I. Sobre los sesgos



¿De qué hablamos cuando hablamos de sesgo?

Sesgo

El **sesgo** es un peso desproporcionado a favor o en contra de una cosa, persona o grupo en comparación con otra, generalmente de una manera que se considera injusta.

Los sesgos se pueden aprender observando contextos culturales. Las personas pueden desarrollar sesgos hacia o en contra de un individuo, un grupo étnico, una identidad sexual o de género, una nación, una religión, una clase social, un partido político, paradigmas e ideologías teóricas dentro de los dominios académicos o una especie.[1] Sesgo significa unilateral, carece de un punto de vista neutral o no tiene una mente abierta. El sesgo puede venir en muchas formas y está relacionado con el prejuicio y la intuición.[2]

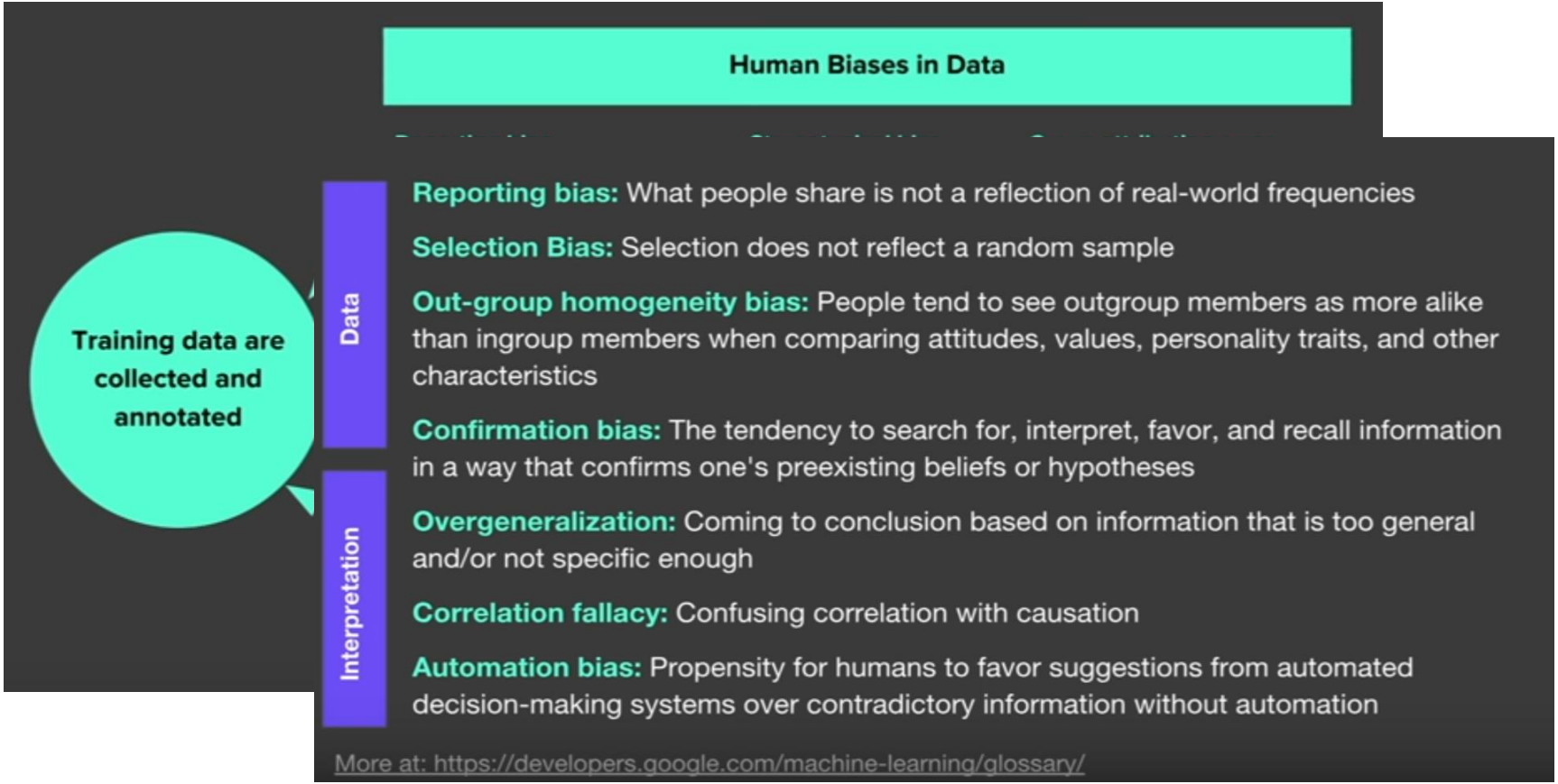
En ciencia e ingeniería, un sesgo es un error sistemático. El sesgo estadístico resulta de un muestreo injusto de una población, o de un proceso de estimación que no da resultados precisos en promedio.

Pero hay varios tipos de sesgos

Sesgo estadístico: la diferencia entre la esperanza del estimador y el valor numérico del parámetro que estima.

Sesgo cognitivo: una característica en particular de un sujeto, que incide en el procesamiento de la información y que forma lo que se conoce como prejuicio cognitivo (la clase de distorsión que afecta el modo de percibir la realidad).

Sesgo algorítmico: errores sistemáticos y repetidos que crean resultados injustos, como dar privilegios a un grupo de usuarios por encima de otros de forma arbitraria.



Bias in AI, Margaret Mitchell, 2019

¿Cuándo se produce el sesgo algorítmico?

A lo largo de todo el proceso! En particular:

- Al definir el problema
- En la colecta y preparación de los datos
- En el desarrollo de los modelos
- En la interpretación de los resultados

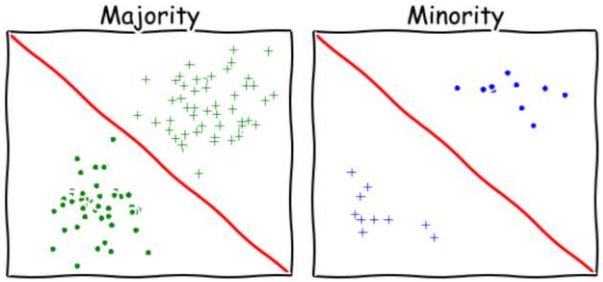
Big Data's Disparate Impact, Barocas & Selbst, 2016
How big data is unfair, Hardt, M., Medium, 2014

Diferencias culturales

Ejemplo: intentar identificar nombres reales de falsos.

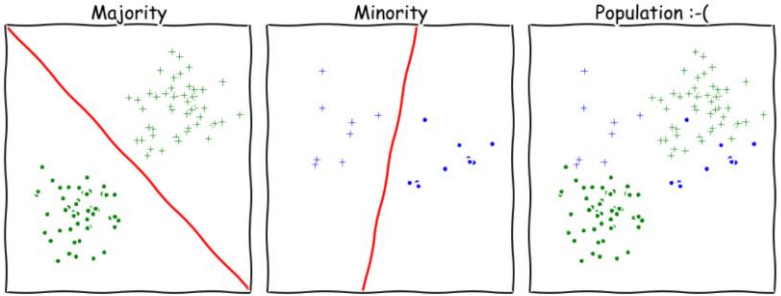
La población blanca en Estados Unidos tiende a usar nombres relativamente comunes, mientras que los nombres étnicos presentan más diversidad.

Un patrón estadístico que aplica para la mayoría puede ser inválido en un grupo minoritario



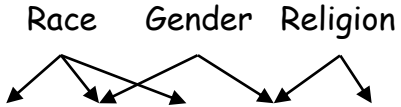
Positively labeled examples are on opposite sides of the classifier for the two groups.

La búsqueda de equidad agrega **complejidad** a los modelos



Even if two groups of the population admit simple classifiers, the whole population may not.

Los datos como reflejo de la sociedad

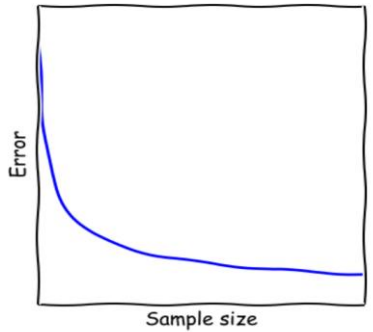


No alcanza con excluir los atributos protegidos (ej: raza) de los datos de entrenamiento, estos pueden/suelen estar codificados de forma redundante.

Podemos encontrar correlación entre atributos protegidos y la salida: ¿siempre es sesgo? ¿qué umbral es preocupante?

Disparidad en el tamaño de las muestras

Menos datos sobre minorías → más error en los clasificadores



The error of a classifier often decreases as the inverse square root of the sample size. Four times as many samples means halving the error rate.

How big data is unfair, Hardt, M., Medium, 2014

¿Cómo se percibe el sesgo algorítmico?

Un ejemplo: los contenidos ofrecidos por buscadores y plataformas

User bias: diferentes usuarios reciben contenido o resultados diferentes basado en atributos que deberían ser protegidos, como por ejemplo género, sexo, etnia o religión

Content bias: algún aspecto está desproporcionadamente representado en los resultados que recibe el usuario (ej: resultados de una consulta, news feed)

On Measuring Bias in Online Information, Pitoura et al., SIGMOD Record, 2017

La herramienta AdFisher (Carnegie Mellon) simuló personas buscando trabajo que no diferían en el comportamiento de navegación, las preferencias o las características demográficas, excepto en género.

Un experimento mostró que Google presentaba anuncios para ofertas de servicio de orientación profesional para ejecutivos de "$ 200k +" **1,852 veces para el grupo masculino y solo 318 veces para el grupo femenino.**

Otro experimento, en julio de 2014, mostró una tendencia similar pero no fue estadísticamente significativa.

Information Flow Experiments
https://www.cs.cmu.edu/~mtschant/ife/

Forbes Billionaires Innovation Leadership Money Business Small Business Lifestyle Lists

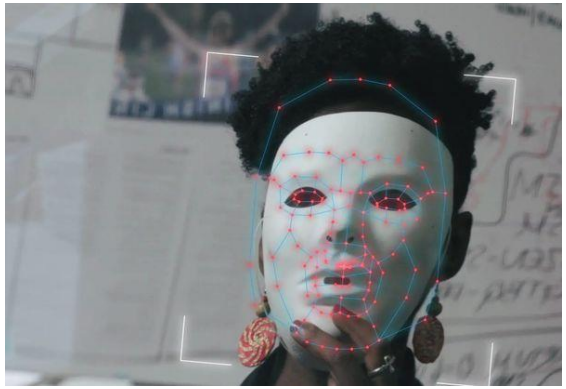
EDITORS' PICK | 1,927 views | Jun 12, 2020, 09:26pm EDT

IBM, Microsoft And Amazon Not Letting Police Use Their Facial Recognition Technology

Larry Magid Senior Contributor
Consumer Tech

In the wake of protests around the death of George Floyd, IBM, Microsoft and Amazon are now denying police departments access to their facial recognition technology.

Some have hailed facial recognition technology as a great boon to law enforcement while others say the technology not only violates privacy rights, but is prone to errors with females and people of color.



INDEPENDENT News InFact Politics Voices Indy/Life Sport Business Video Culture

News › UK › Home News

Metropolitan Police's facial recognition technology 98% inaccurate, figures show

'Intrinsically Orwellian' systems must be scrapped, campaigners say as biometrics commissioner brands them 'not yet fit for use'

Sesgo en sistemas predictivos

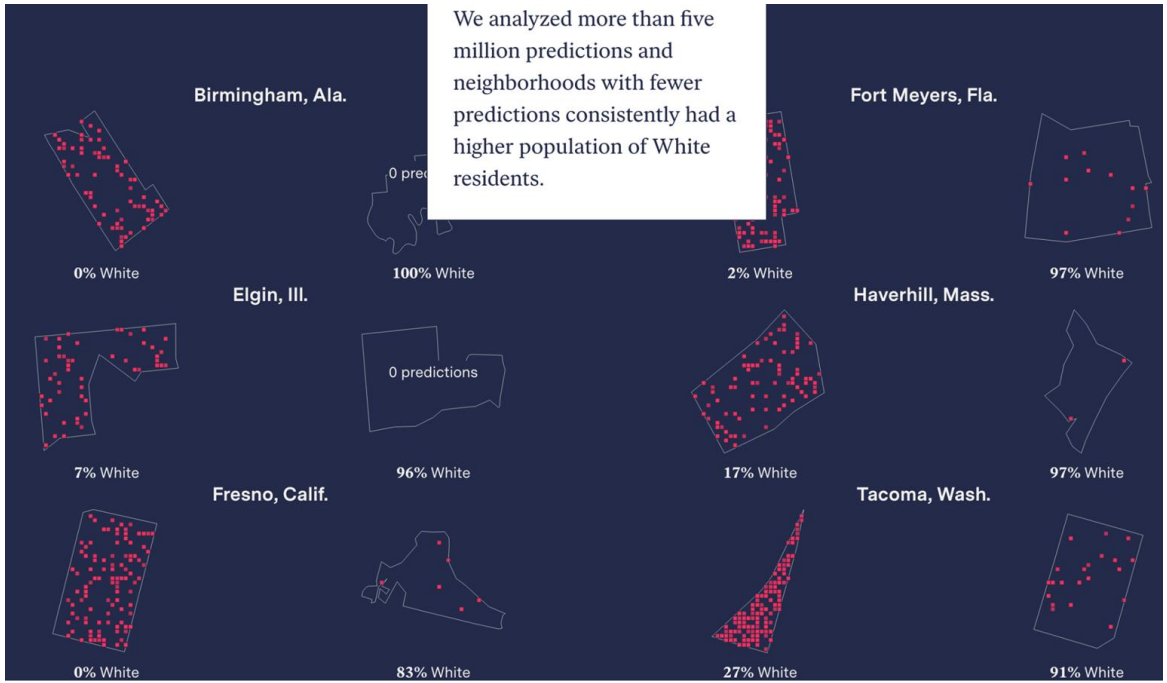
The Markup

Prediction: Bias

Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them

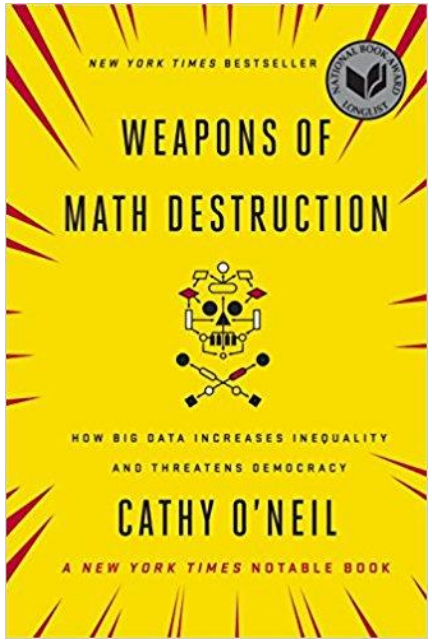
Millions of crime predictions left on an unsecured server show PredPol mostly avoided Whiter neighborhoods, targeted Black and Latino neighborhoods

By Aaron Sankin, Dhruv Mehrotra for Gizmodo, Surya Mattu, and Annie Gilbertson



https://themarkup.org/prediction-bias/2021/12/02/crime-prediction-software-promised-to-be-free-of-biases-new-data-shows-it-perpetuates-them

La Ciencia de Datos no es **justa** por defecto



Los modelos y algoritmos pueden ser peligrosos cuando:

- Son opacos para los sujetos que son analizados (personas!)
- Son perjudiciales para sus intereses
- Son ejecutados a gran escala
- No tienen feedback o procesos de ajuste

Sobretodo cuando son procesos de toma de decisiones que involucran **personas**

Decisiones de negocio vs políticas públicas

https://weaponsofmathdestructionbook.com/

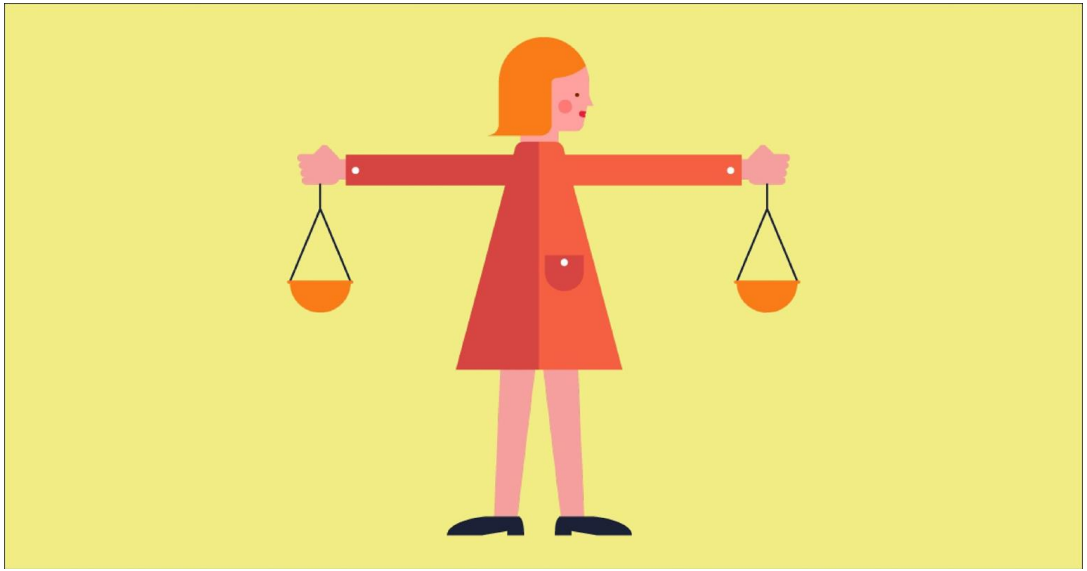
Weapons of Math Destruction | Cathy O'Neil | Talks at Google 2017
https://www.youtube.com/watch?v=TQHs8SA1qpk

Image by Robin Higgins from Pixabay

1. Tomar conciencia del problema

Gerd Altmann from Pixabay

2. Medir, evaluar, ajustar



Existen múltiples enfoques para evaluar discriminación y equidad (fairness)

Las diferencias en los resultados deben ser explicables, en su mayoría, por los atributos no-protegidos.

Al menos dos marcos para medir la discriminación:

- Fairness individual: personas similares, salidas similares
- Fairness grupal: tratar a grupos diferentes de manera similar

A Survey on Bias and Fairness in Machine Learning, MEHRABI et al, 2021
A survey on measuring indirect discrimination in machine learning, Zliobaite, I, Arxiv.org, 2015.

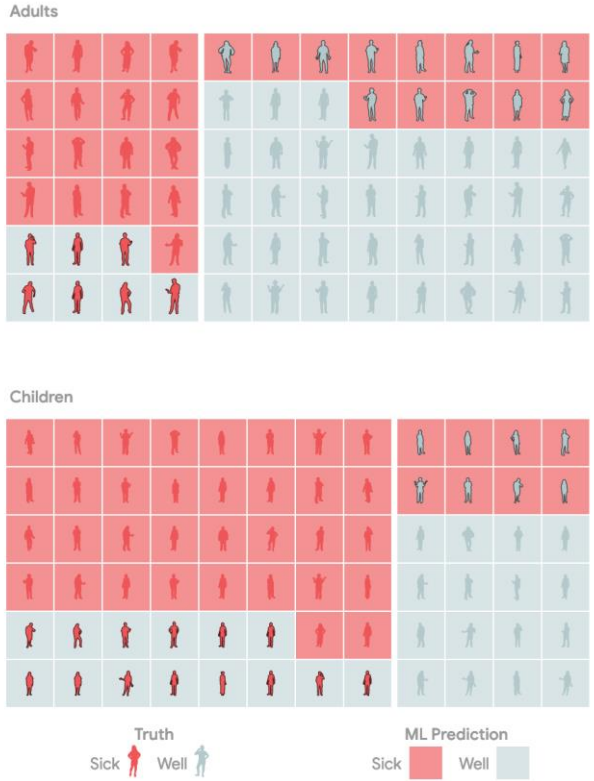
Fairness: definiciones y métricas

Def 1(probabilidades igualadas): los grupos protegidos y no protegidos deben tener iguales tasas de verdaderos positivos y de falsos positivos [Hardt et al, 2016]

Def 2 (igualdad de oportunidades): los grupos protegidos y no protegidos deben tener iguales tasas de verdaderos positivos [Hardt et al, 2016].

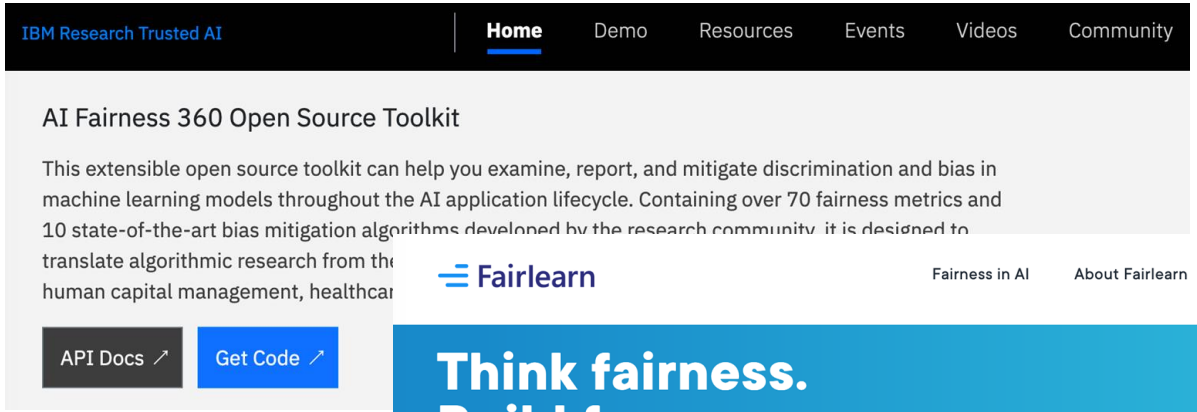
Def 3 (paridad demográfica o estadística): la probabilidad de un resultado positivo debería ser independiente de la pertenencia a un grupo protegido [Dwork et al, 2012].

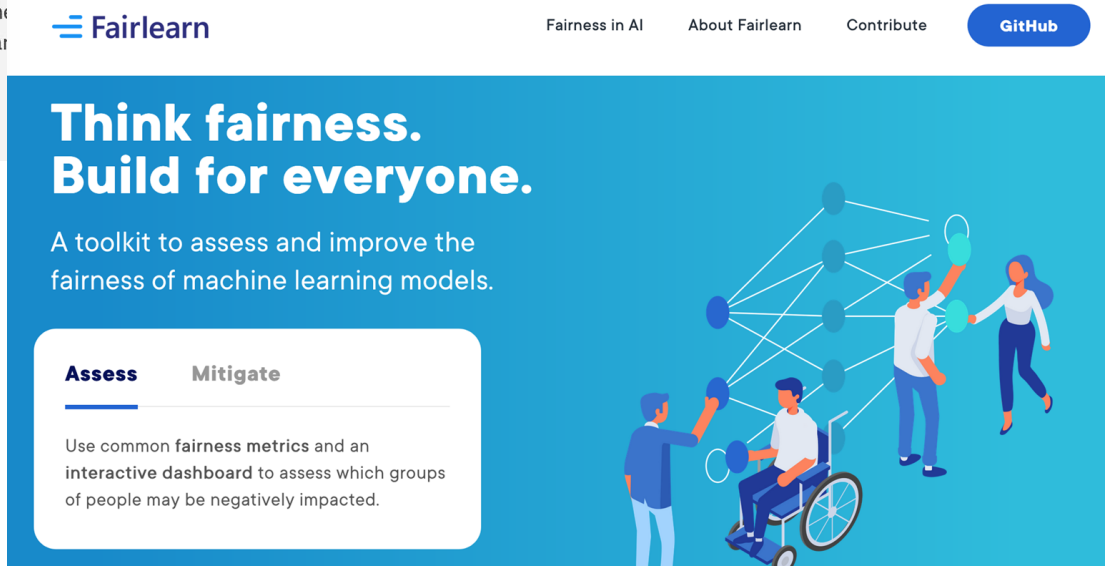
Def 4 (Fairness through awareness): individuos similares, salidas similares [Dwork et al, 2012].



https://pair.withgoogle.com/explorables/measuring-fairness/

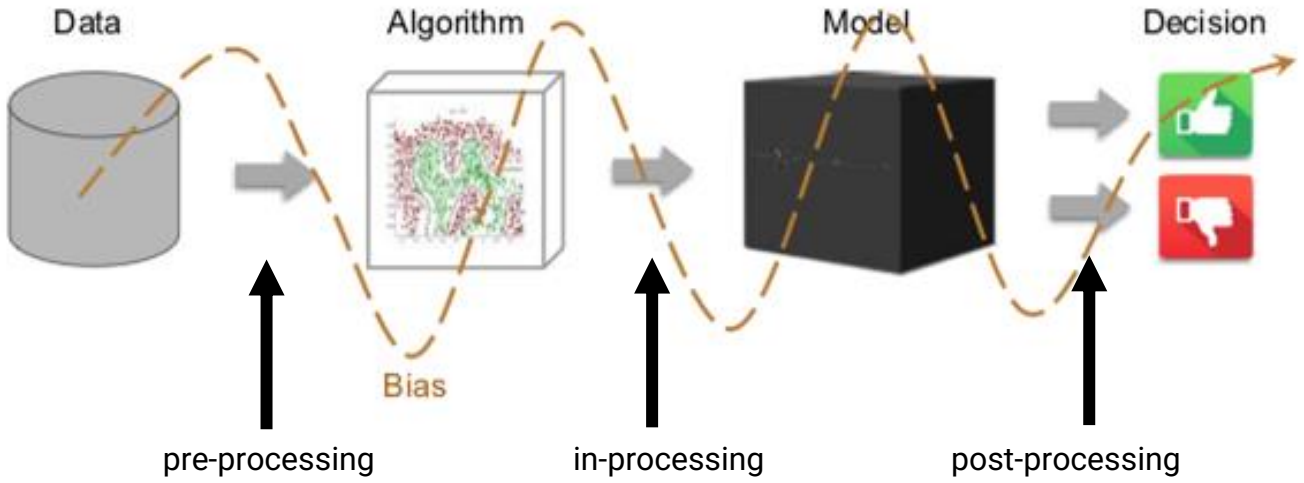
Algunas herramientas para medir fairness





3. Adoptar enfoques que mitiguen el sesgo

"human-in-the-loop"



Algorithmic Bias from discrimination discovery to fairness-aware data mining, Hajian et al, KDD 2016

Data, Responsibly: Fairness, Neutrality and Transparency in Data Analysis, Stoyanovich et al, EDBT 2016

Fair machine learning

Objetivo: desarrollar sistemas de toma de decisiones que no discriminen, preservando la calidad de la decisión.

Equidad  **Utilidad**

Pasos:
(1) Definir restricciones anti-discriminatorias y de equidad
(2) Transformar datos/algoritmos/modelos para satisfacer las restricciones
(3) Evaluar la utilidad de los datos/modelos

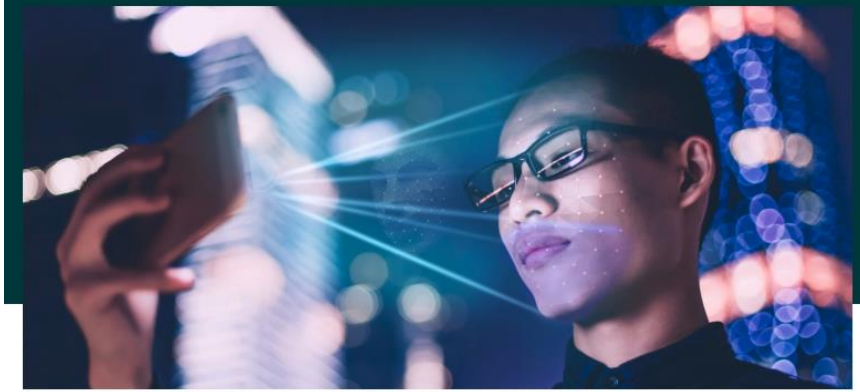
Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), 1-38.

4. Promover la diversidad

Study finds diversity in data science teams is key in reducing algorithmic bias

Kyle Wiggers
@Kyle_L_Wiggers

December 9, 2020 1:10 PM



Applause is expanding its platform to help reduce AI bias by sourcing training data at scale
Image Credit: Applause

https://venturebeat.com/2020/12/09/columbia-researchers-find-white-men-are-the-worst-at-reducing-ai-bias/

5. Legislar y buscar transparencia

INDEPENDENT

News InFact Politics Voices Indy/Life Sport Business Video Culture

News › UK › Home News

Metropolitan Police's facial recognition technology 98% inaccurate, figures show

'Intrinsically Orwellian' systems must be scrapped, campaigners say as biometrics commissioner brands them 'not yet fit for use'

The Official Website of the City of New York

NYC

E

SHARE

Email

Print

Mayor de Blasio Announces First-In-Nation Task Force To Examine Automated Decision Systems Used By The City

May 16, 2018

NEW YORK— Today, Mayor de Blasio announced the creation of the Automated Decision Systems Task Force which will explore how New York City uses algorithms. The task force, the first of its kind in the U.S., will work to develop a process for reviewing "automated decision systems," commonly known as algorithms, through the lens of equity, fairness and accountability.

"In terms of governance, technical development and deployment is running ahead of legislation and these new biometrics urgently need a legislative framework, as already exists for DNA and fingerprints."

OASI OBSERVATORY OF ALGORITHMS WITH SOCIAL IMPACT

What are Algorithms? ›

Auditing Algorithms ›

The OASI Register ›

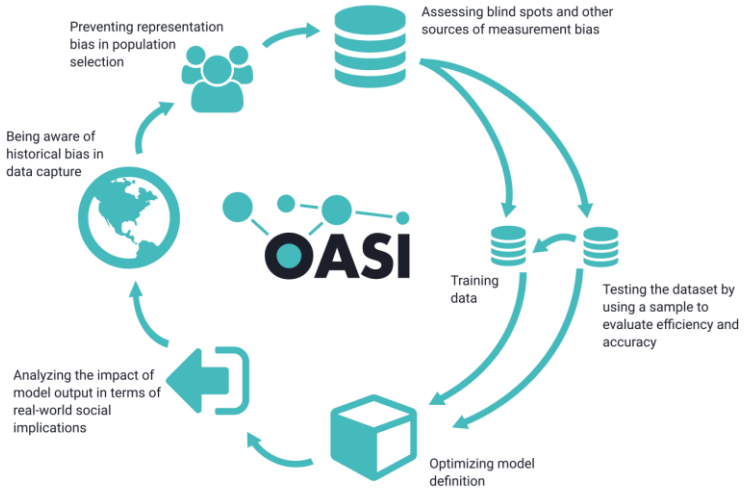
Methodology and categories ›

Social Impacts of Algorithms ›

Tell us about an Algorithm ›

Hide fields | Filter | Group | Sort | ...

	Algorithm	Implemented by	Location	Developed by	Domain	Aim
1	Nutriscore | Algorithm to rank nutritional content of various foods		France, Germany	N/A	product safety	ran
2	SyRI | Detecting welfare fraud	Dutch Government	Netherlands	N/A	social services	pre
3	LS/CMI | Risk assessment for parole	Massachusetts P...	Massachusetts (...	MultiHealth Syste...	justice and democratic ...	pre
4	CHINOOK | Algorithm to augment and replace human decision-making in immigration ...	Canadian Govern...	Canada	N/A	justice and democratic ..., policing and security	aut



https://eticasfoundation.org/oasi/

Resumiendo

Los algoritmos y los modelos no son ni objetivos ni neutrales.

Debemos aprender a lidiar con el sesgo, y a mitigarlo durante todas las etapas del proceso de análisis y desarrollo. Hay varios enfoques posibles y herramientas disponibles.

Debemos incluir el feedback y el ajuste en el ciclo de vida de los modelos.

La legislación debe hacer cumplir la transparencia y la responsabilidad en los procesos de toma de decisiones.

Un área de investigación muy activa, mucho trabajo por hacer!

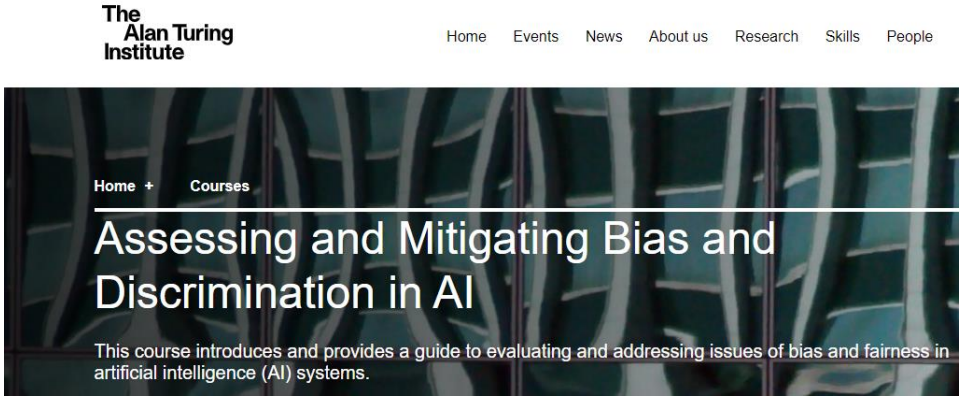
Recursos útiles para profundizar sobre estos temas

https://www.fatml.org/

https://dataresponsibly.github.io/

https://www.ajlunited.org/

https://aplusalliance.org/global-fair/



https://www.turing.ac.uk/courses/assessing-and-mitigating-bias-and-discrimination-ai