Privacidad y aspectos éticos en la ciencia de datos

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

Privacidad

¿Qué es la privacidad?

desidentificación:

no debería ser posible volver a identificar a ninguna persona

divulgación de identidad

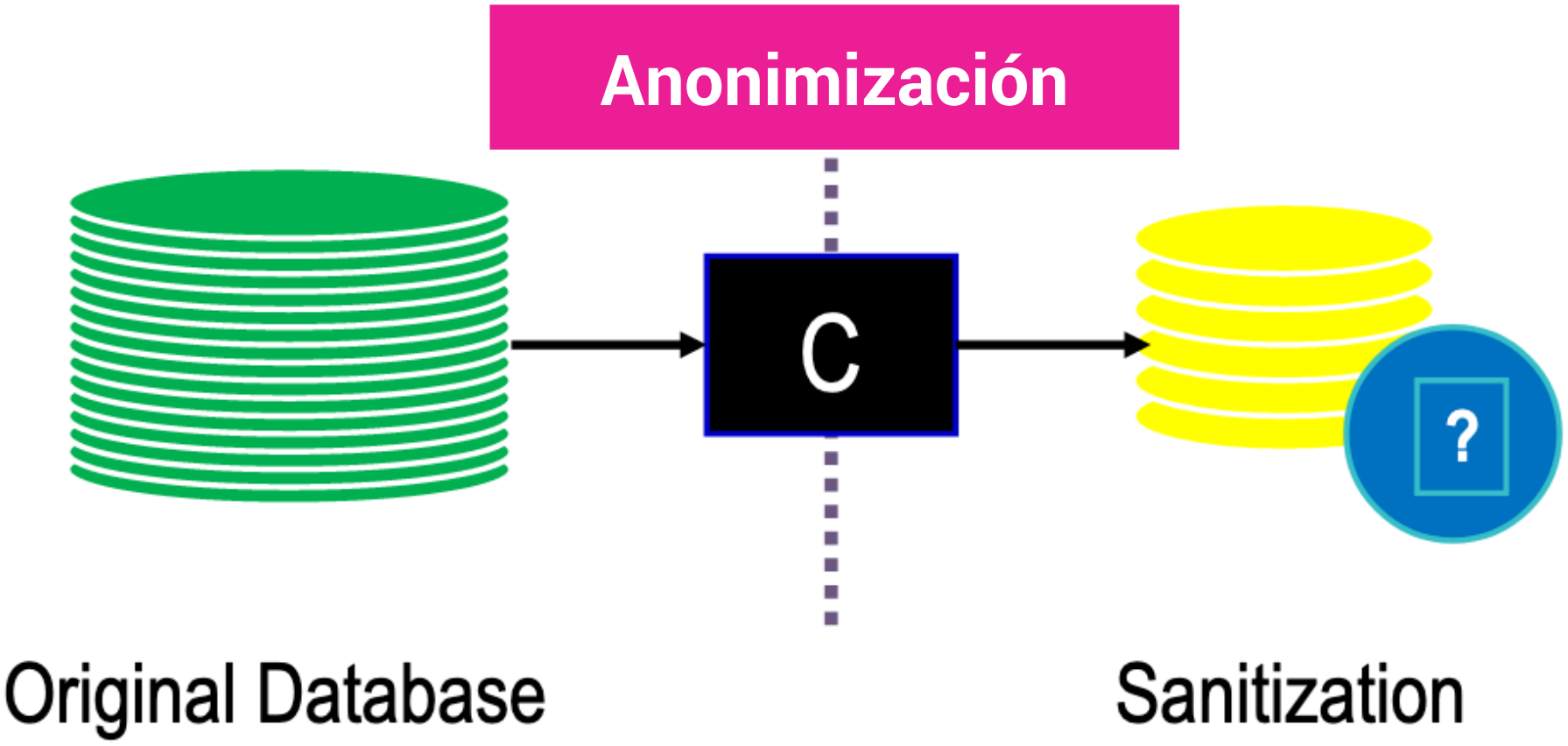
confidencialidad o secreto:

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

divulgación de atributos

RIESGOS

La privacidad como una propiedad de los datos



La anonimización busca esconder a cada individuo dentro de un grupo

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

Definiciones

Sea r1 un registro de la base original O con atributos sensibles en S.

Sea C el contexto de información sobre r1 manejado por el atacante.

Sea T la base anonimizada construida a partir de O usando el método f

En ese contexto, el problema de la privacidad se reduce a asegurar dos propiedades:

Desidentificación:$\forall r2 \in T : Pr[r1 \in f\text{-}1(r2) \mid C(r1)] < \varepsilon$.

Confidencialidad: $\forall x \in S \; \forall v \in Dom(x) : Pr[r1.x=v \mid C(r1)] < \varepsilon$

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

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

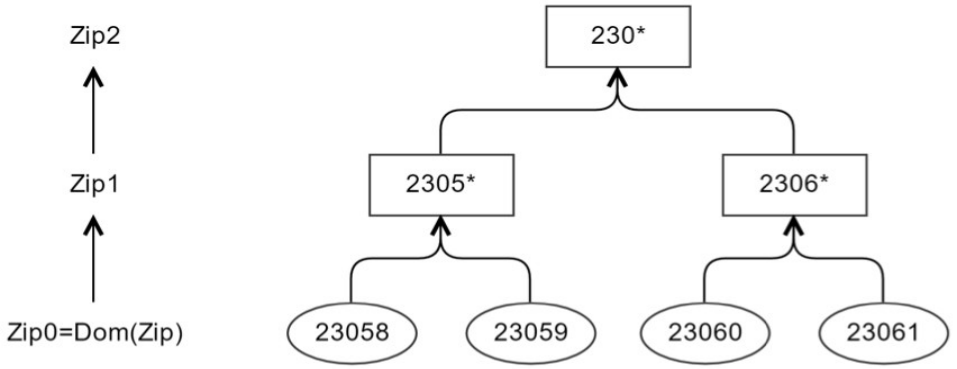
	SS Number	Age	ZIP code	Condition
1	1234-12-1234	21	23058	Heart Disease
2	2345-23-2345	24	23059	Heart Disease
3	3456-34-3456	26	23060	Viral Infection
4	4567-45-4567	27	23061	Viral Infection
5	5678-56-5678	43	23058	Kidney Stone
6	6789-67-6789	43	23059	Heart Disease
7	7890-78-7890	47	23060	Viral Infection
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

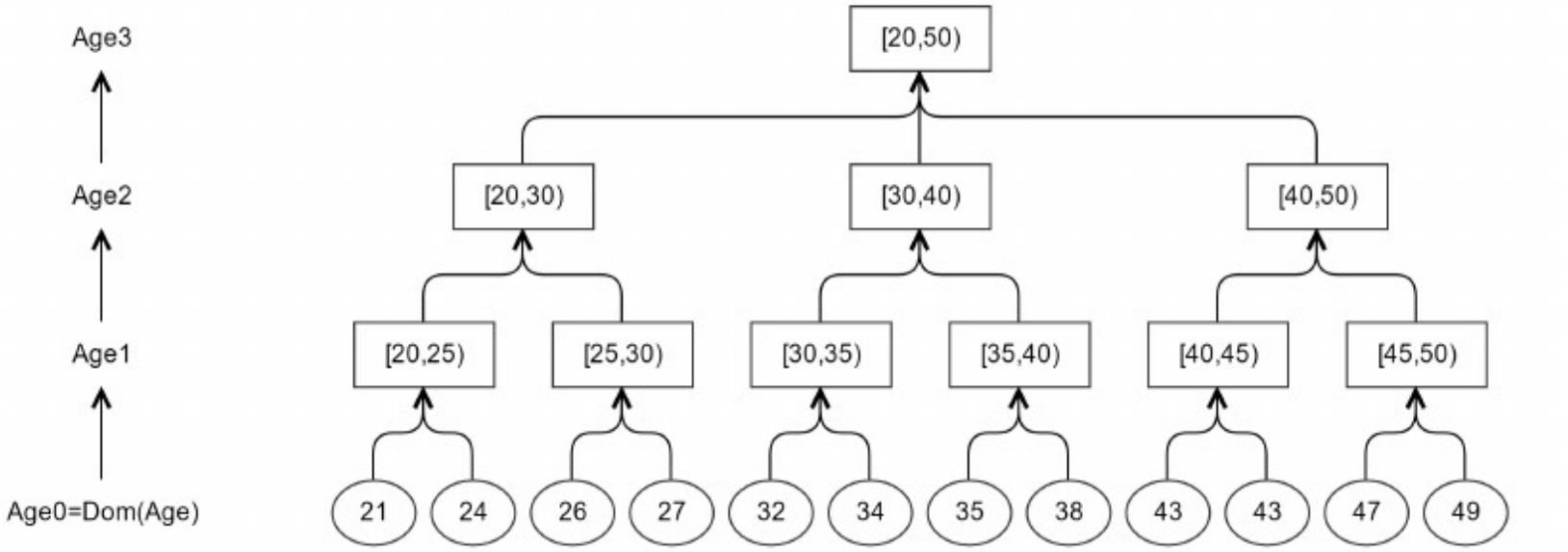
	SS Number	Age	ZIP code	Condition
1	*	[20-30]	230**	Heart Disease
2	*	[20-30]	230**	Heart Disease
3	*	[20-30]	230**	Viral Infection
4	*	[20-30]	230**	Viral Infection
5	*	[40-50]	230**	Kidney Stone
6	*	[40-50]	230**	Heart Disease
7	*	[40-50]	230**	Viral Infection
8	*	[40-50]	230**	Viral Infection
9	*	[30-40]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

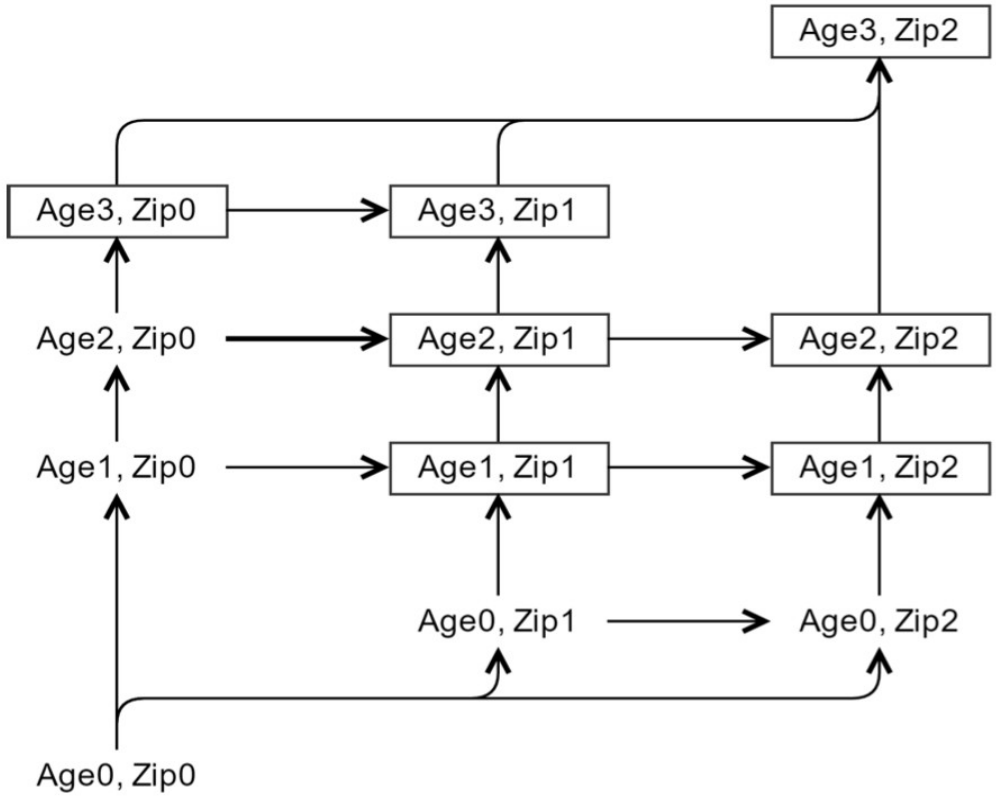




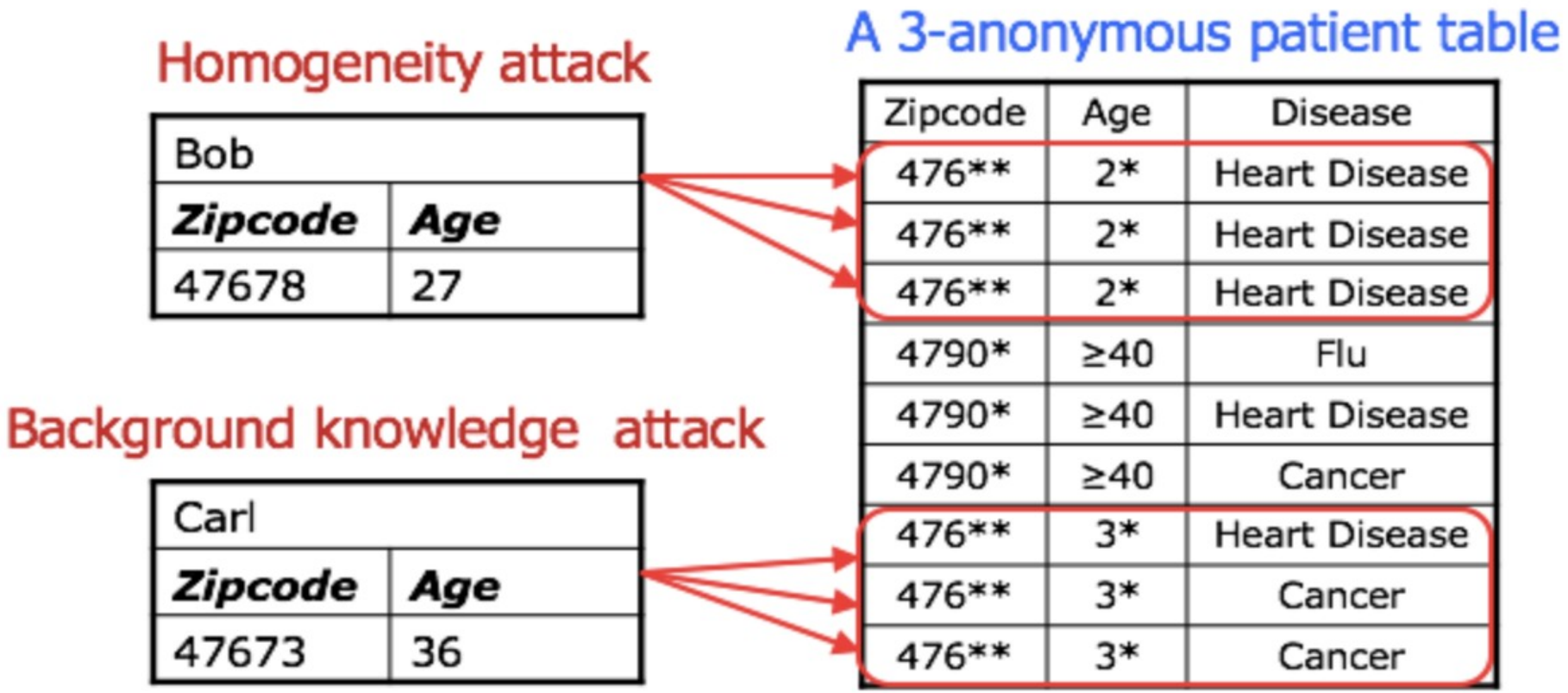
¿Cómo alcanzar k-anonymity?

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]

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

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

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

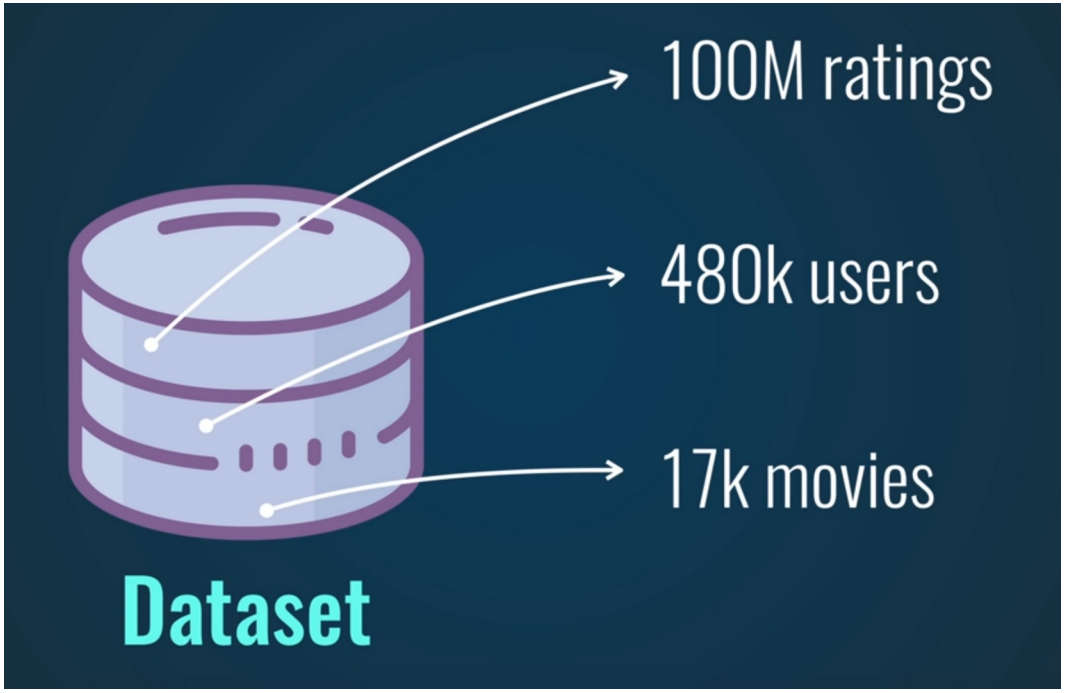
Desafío: predecir el rating de contenidos

Netflix libera 1/10 de su base de datos

Una fracción de las preferencias por usuario (centenas)

Identificadores de usuario anonimizados

Preferencias perturbadas



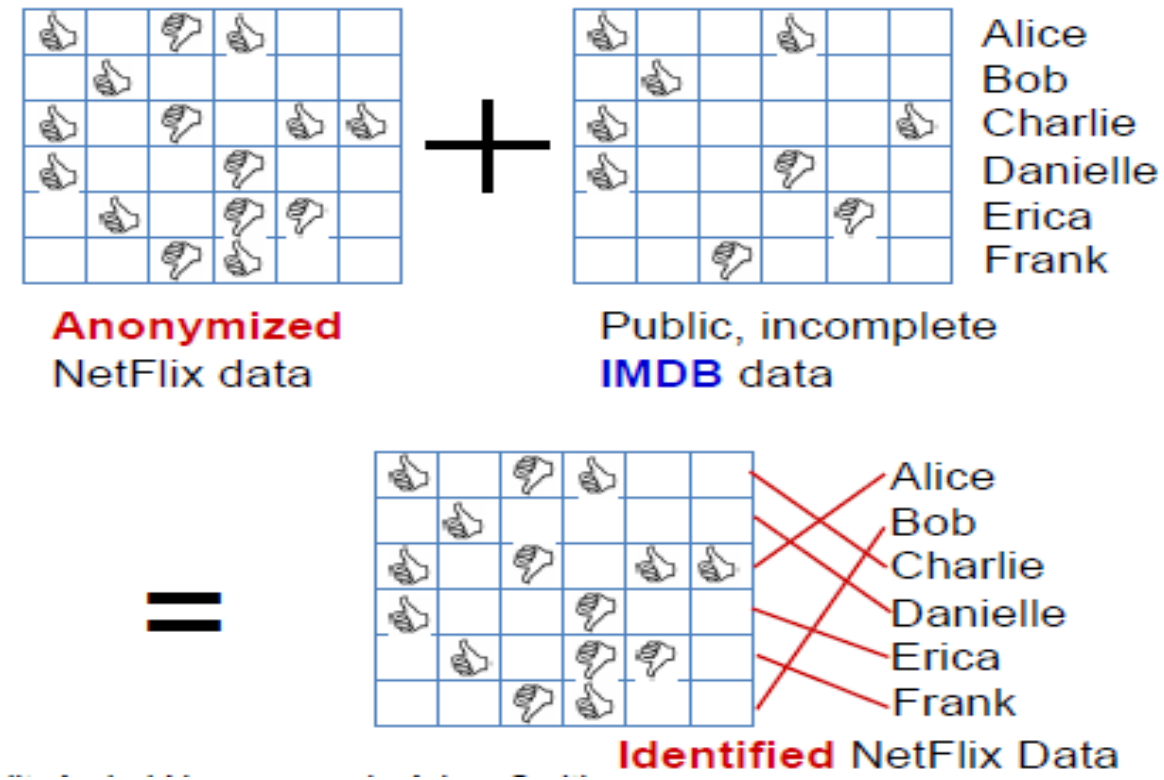
Linkage attack

Combinan con IMDB

Algoritmo de similitud ScoreboardH

Con 8 películas + fechas con error < 14 días 99% identificados (1 solo match)

Preferencias políticas o sexuales se pueden deducir fácilmente de los datos.



Robust De-anonymization of Large Sparse Datasets - Narayanan & Shmatikov

La privacidad como una propiedad de las consultas

No tiene que ver con “re-identificación”.

Es completamente general, independiente de:

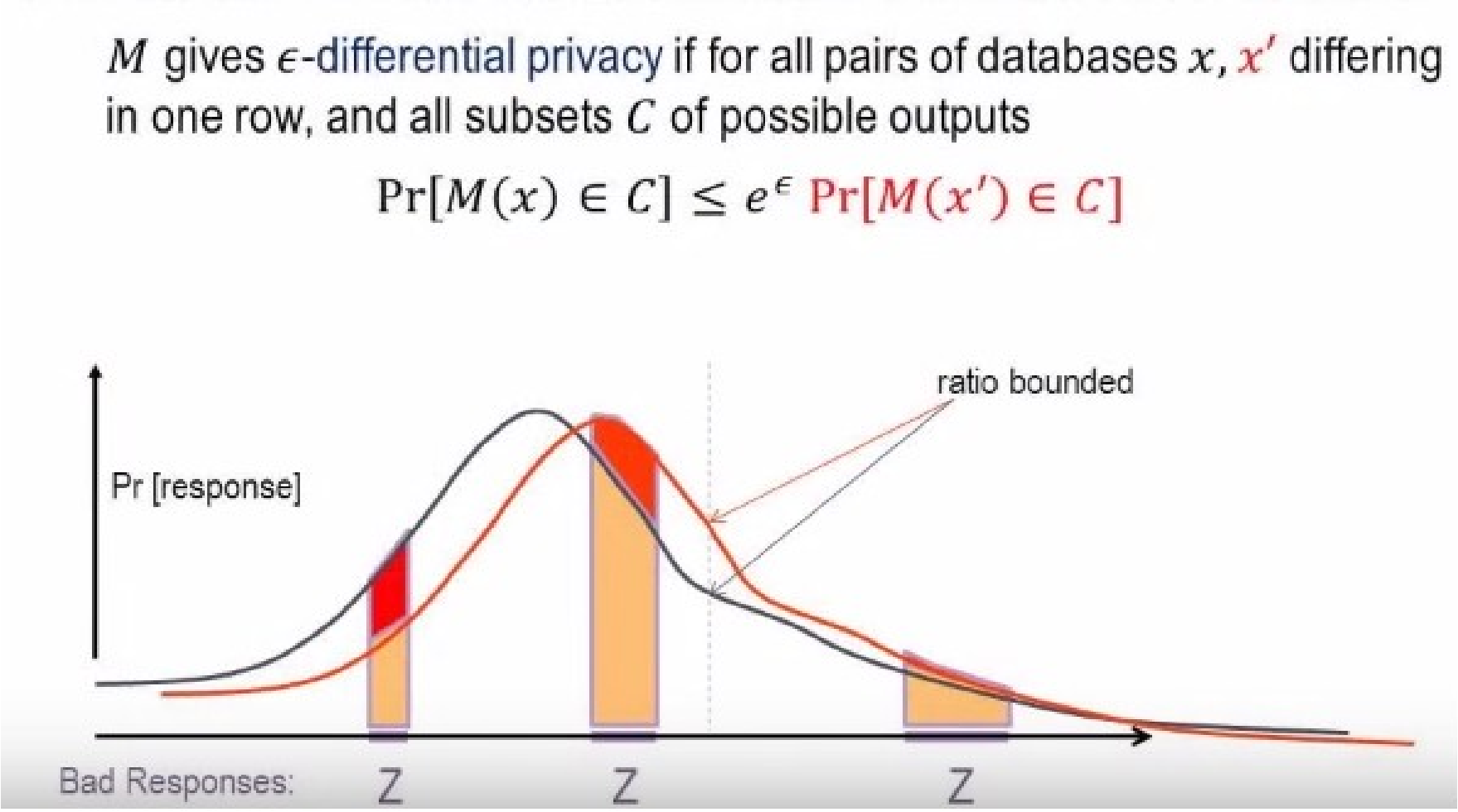
- La naturaleza de los atributos
- El tipo de conclusiones que saquemos
- Lo que sepa el atacante por otros medios (info auxiliar)

Hay una definición matemática y formal atrás

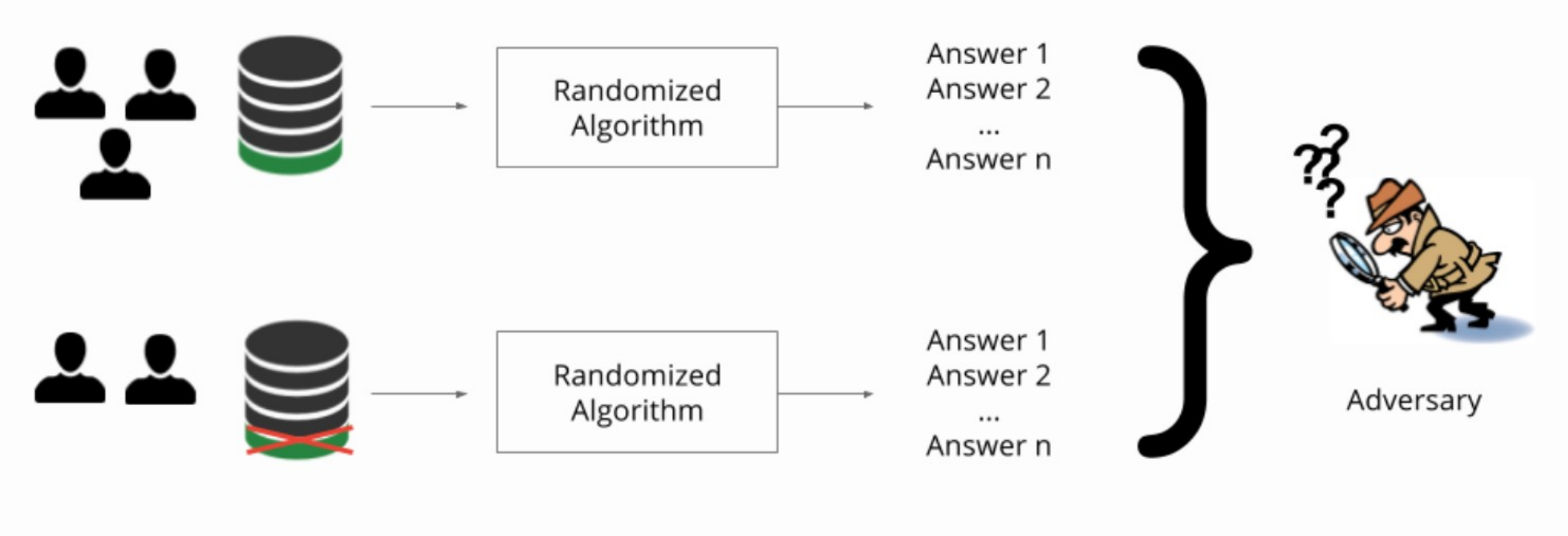
Differential Privacy [Dwork et al. 06]

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



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

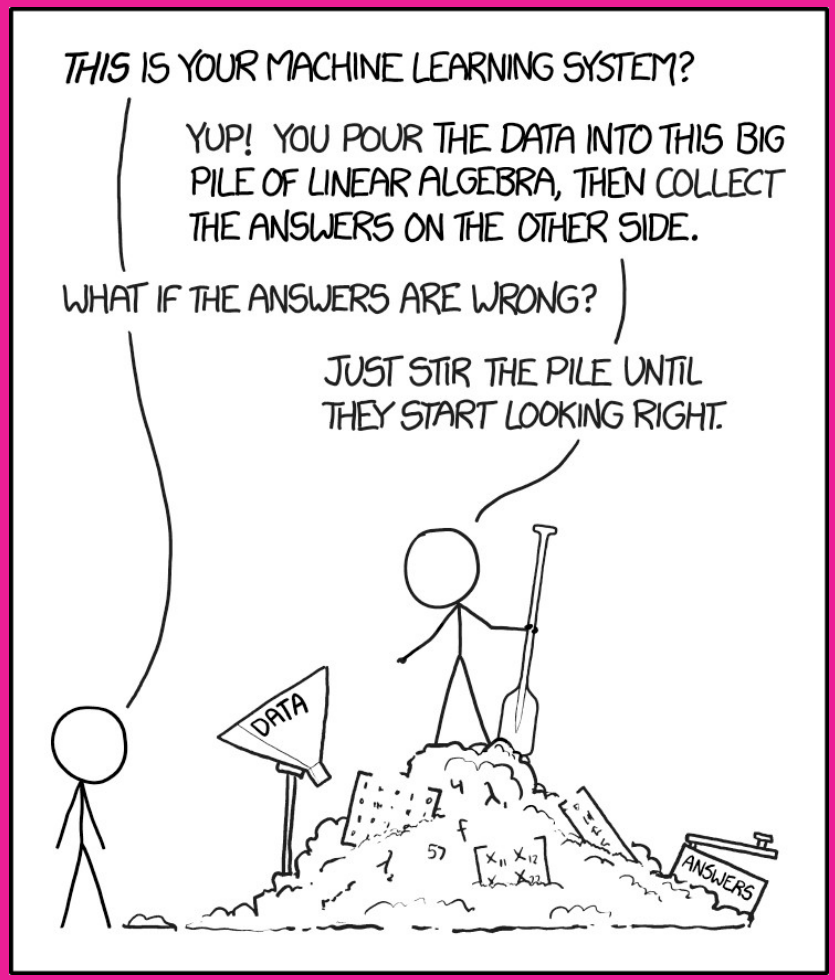
PySyft is a Python library for secure and private Deep Learning. PySyft decouples private data from model training, using Federated Learning, Differential Privacy, and Encrypted Computation (like Multi-Party Computation (MPC) and Homomorphic Encryption (HE)) within the main Deep Learning frameworks like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

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

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

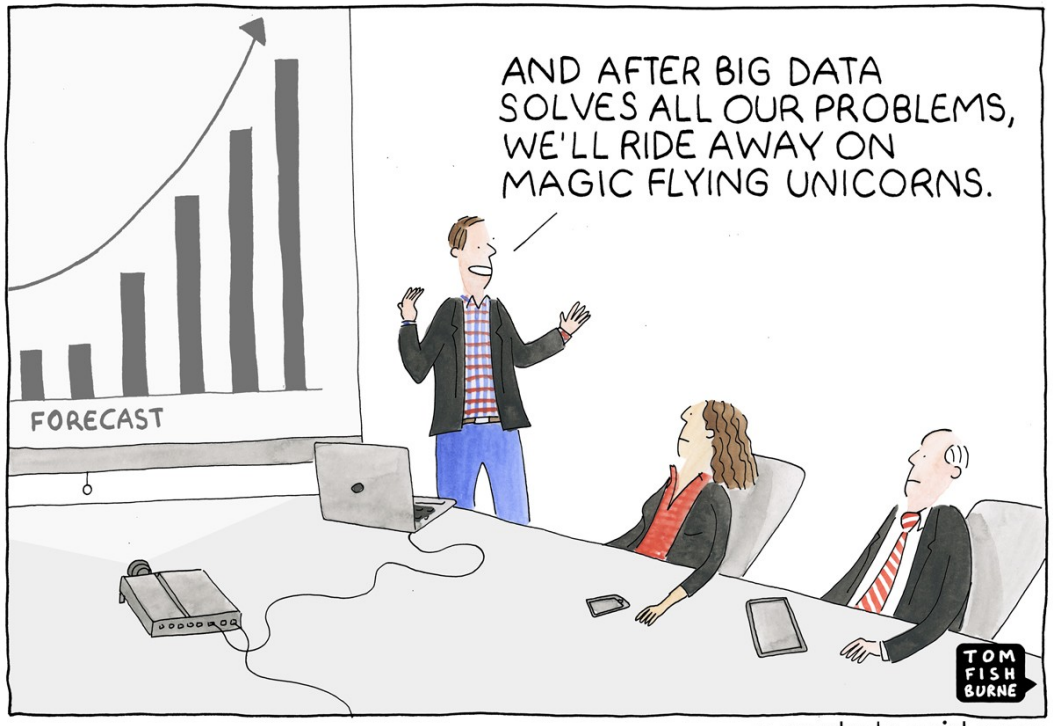
Larry Magid Senior Contributor ⓘ
Consumer Tech

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.

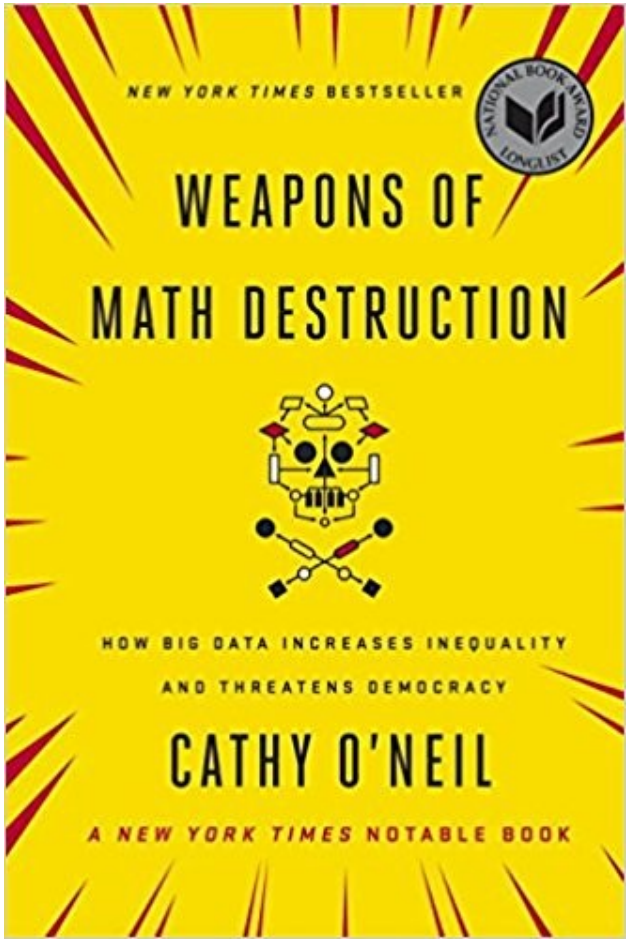
Some have hailed facial recognition technology as a great boon to law enforcement while others say the technology not only violates privacy rights, but is prone to errors with females and people of color.

Las promesas de la ciencia de datos y el big data

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)



Pero la Ciencia de Datos no es justa por defecto



Los modelos/algoritmos pueden ser peligrosos cuando:

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

Procesos de toma de decisiones que involucran personas

Decisiones de negocio vs **políticas públicas**

https://weaponsofmathdestructionbook.com/

¿Qué es el sesgo?

"una inclinación o prejuicio hacia o contra una persona o grupo, de una forma que resulta *injusta*"

"una concentración o interés particular en un área o tema"

User bias: diferentes usuarios reciben contenido diferente 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

Verificación del sesgo online

The Washington Post
Democracy Dies in Darkness

The Intersect

Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

By Julia Carpenter July 6, 2015

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 mostraba 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.

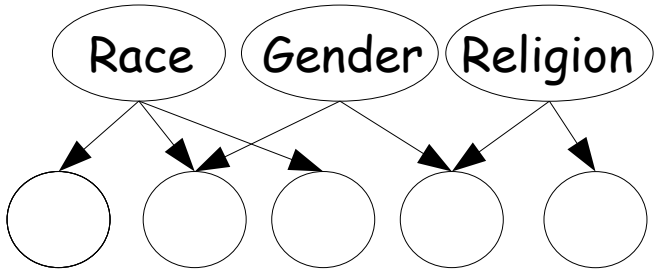
https://www.cs.cmu.edu/~mtschant/ife/

Algunas fuentes de sesgo

Los datos como reflejo de la sociedad

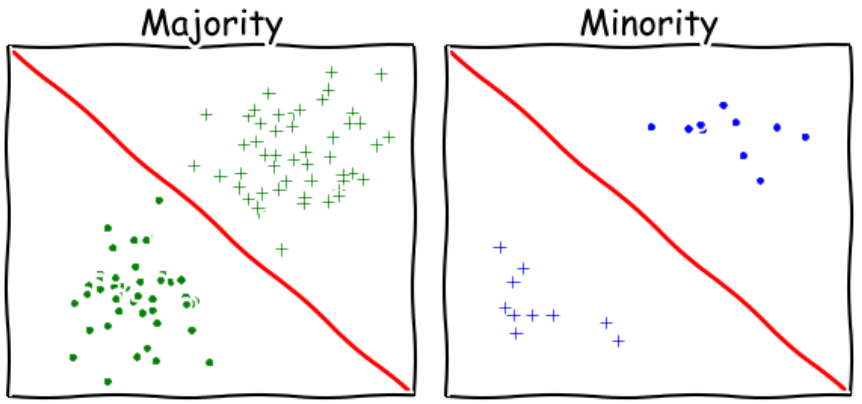
Atributos protegidos codificados de forma redundante

Correlación entre atributos protegidos y la salida (¿que umbral es preocupante?)



Diferencias culturales

Patrones que aplican al grupo mayoritario no aplican al minoritario (ej: nombres en la comunidad afroamericana)



Calidad de datos (correctitud y completitud)

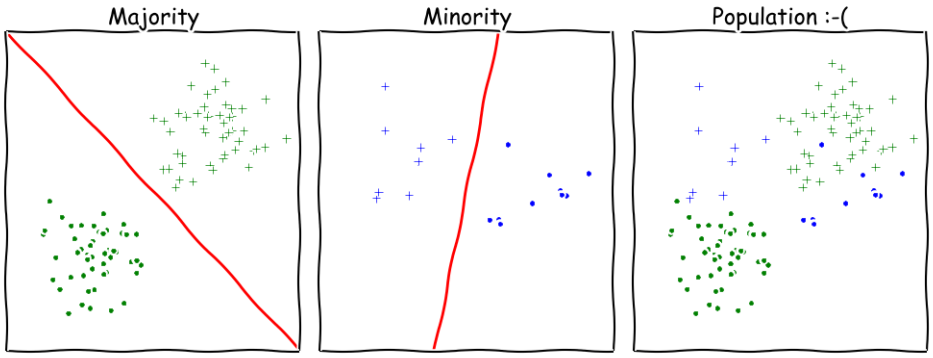
Garbage in, Garbage out (GIGO)

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

Algunas fuentes de sesgo (II)

Disparidad en el tamaño de las muestras

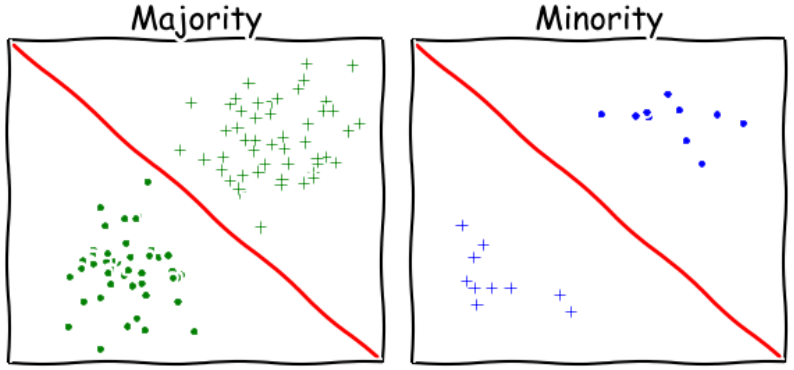
Menos datos sobre minorías → más error en los clasificadores



Diferencias culturales

Patrones que aplican al grupo mayoritario no aplican al minoritario (ej: nombres en la comunidad afroamericana)

La búsqueda de equidad puede agregar **complejidad**



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

Sesgo en sistemas predictivos

NewScientist

NEWS & TECHNOLOGY 4 October 2017, updated 27 April 2018

Biased policing is made worse by errors in pre-crime algorithms

PredPol es un sistema utilizado por la policía en USA para predecir crimen.

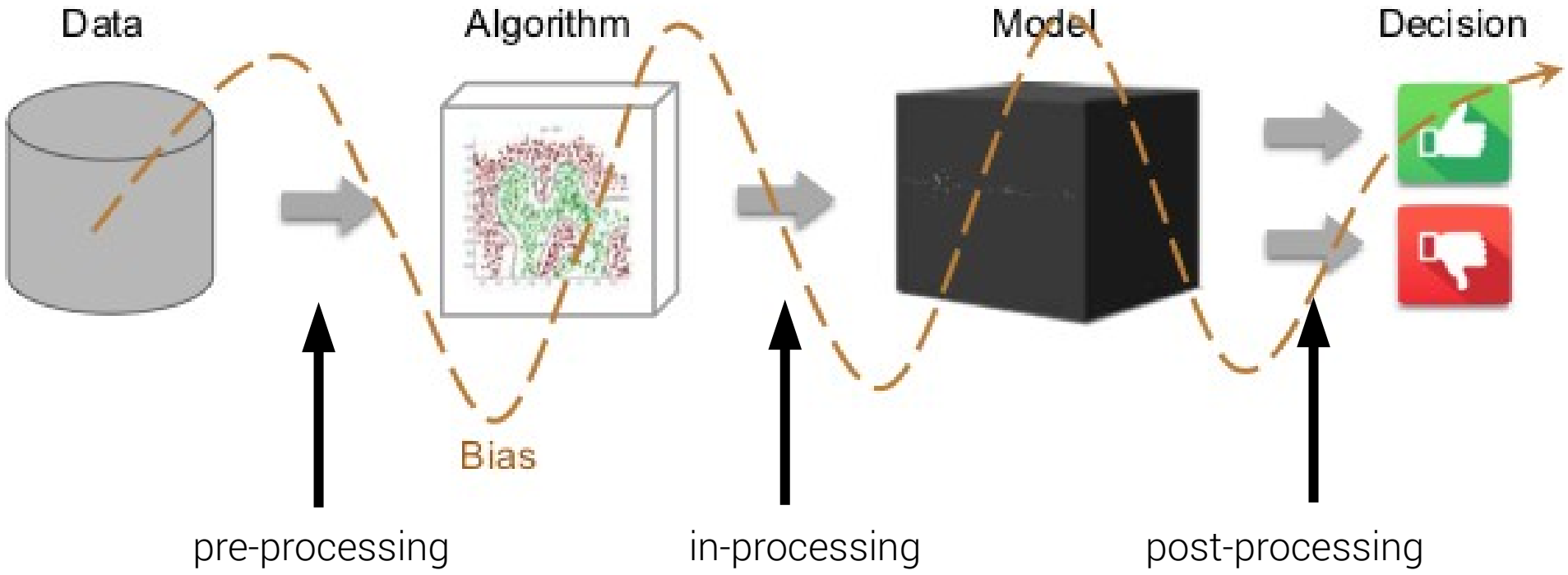
El sistema aprende de los reportes de la policía: más policía, más reportes

Feedback loop: arrestos en un área indican una buena chance de predecir crímenes en esa área en el futuro.

https://www.newscientist.com/article/mg23631464-300-biased-policing-is-made-worse-by-errors-in-pre-crime-algorithms/

Runaway Feedback Loops in Predictive Policing, D Ensign et al. arXiv preprint, 2017

¿Qué podemos hacer?



Diferentes enfoques. Área de investigación muy activa

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

Fairness-aware data mining

Objetivo: desarrollar un sistema de toma de decisiones que no discrimine, 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

Evaluando discriminación y equidad

Las predicciones para personas con atributos no-protegidos **similares** deben ser **similares**

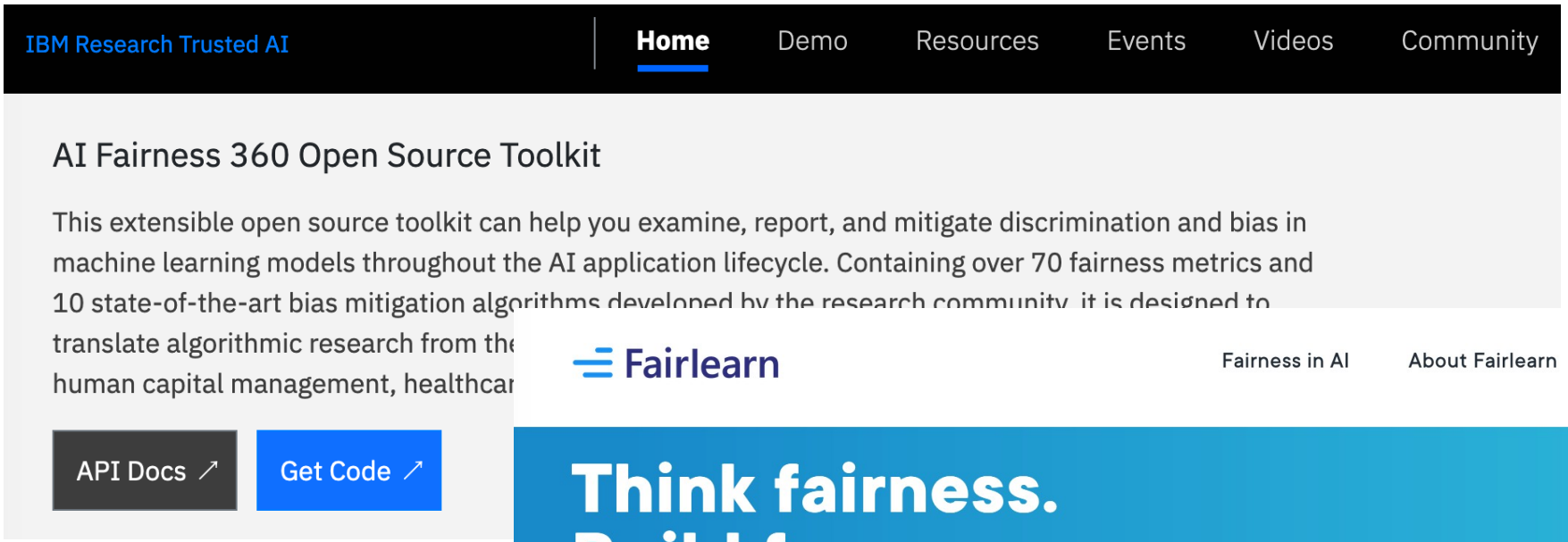
Las diferencias deben ser explicables, en su mayoría, por los atributos no-protegidos.

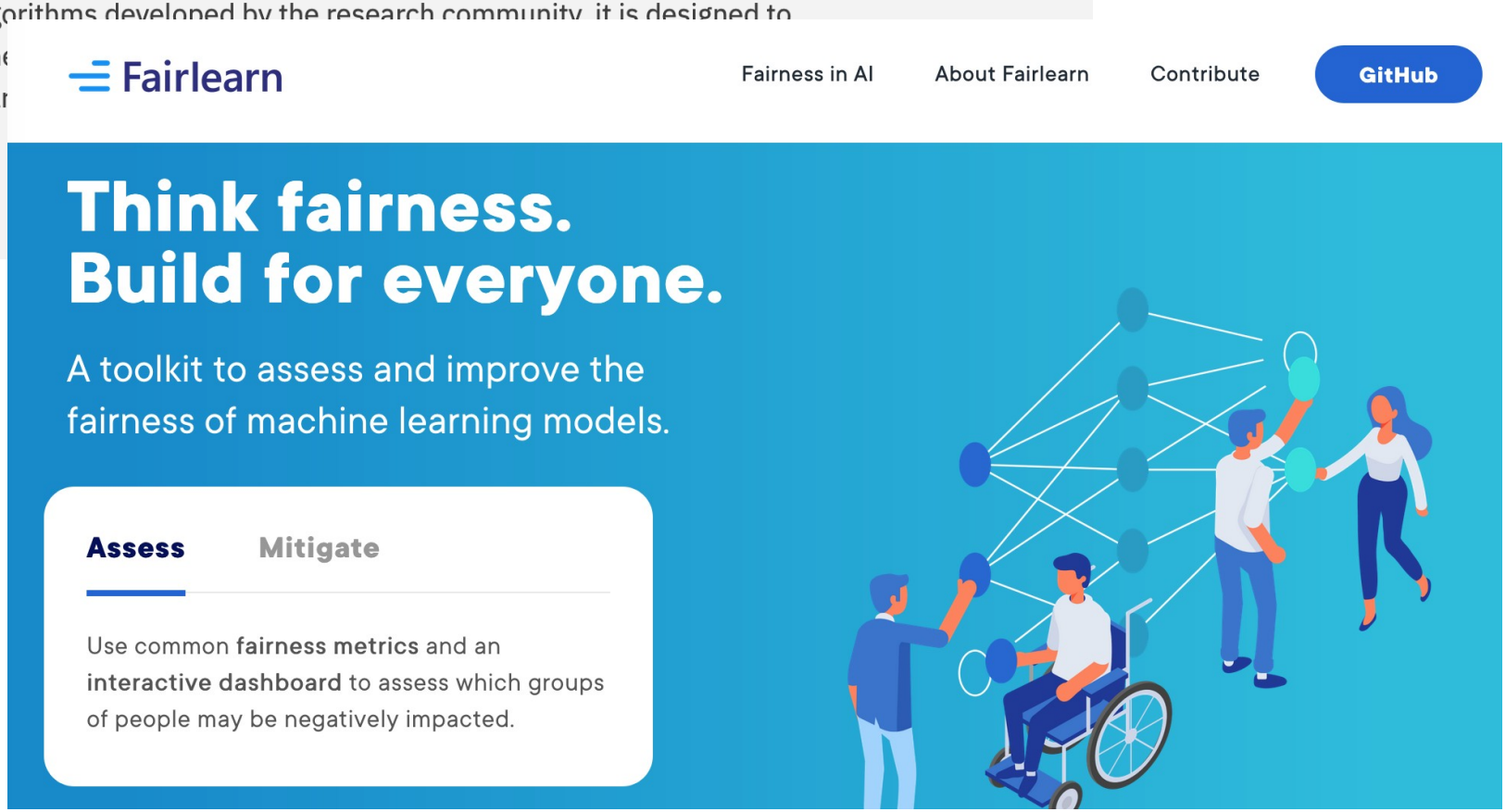
Dos marcos para medir la discriminación:

- Discriminación a nivel individual: consistencia o **equidad individual** (personas similares, salidas similares)
- Discriminación a nivel grupal: **paridad estadística**

Fairness through Awareness, Hardt M Dwork C Pitassi T, ITCS 2012
A survey on measuring indirect discrimination in machine learning, Zliobaite, I, Arxiv.org, 2015.

Algunas herramientas





Legislación y gobernanza

"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."

https://www.independent.co.uk/news/uk/home-news/met-police-facial-recognition-success-south-wales-trial-home-office-false-positive-a8345036.html

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.

http://www1.nyc.gov/office-of-the-mayor/news/251-18/mayor-de-blasio-first-in-nation-task-force-examine-automated-decision-systems-used-by

Legislación y gobernanza (II)

Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise

Berkeley Technology Law Journal, Vol. 34, 2019

45 Pages • Posted: 22 Mar 2018 • Last revised: 17 Apr 2019

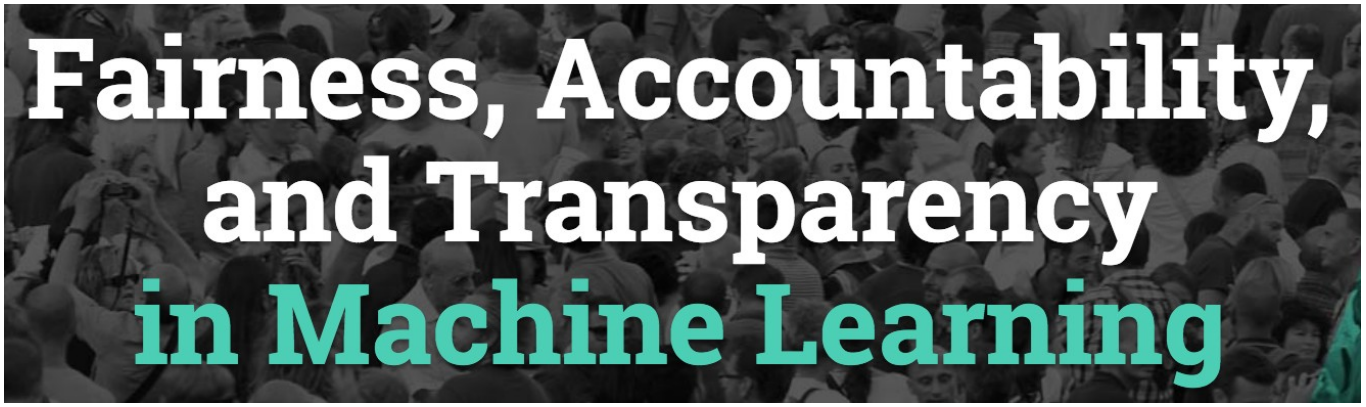
Bryan Casey
Stanford Law School

Ashkon Farhangi
Stanford University - Stanford Codex Center

Roland Vogl
Stanford Law School

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3143325

Recursos útiles



https://www.fatml.org/

data RESPONSIBLY

https://dataresponsibly.github.io/

25 years and counting

Human Rights
Data Analysis
Group hrdag.org

https://hrdag.org/

ALGORITHMIC JUSTICE LEAGUE

https://www.ajlunited.org/

Resumiendo

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

Debemos aprender a lidiar con el sesgo durante todas las etapas del proceso de análisis. 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!