

Equidad, Rendición de Cuentas, Transparencia y Ética en el Aprendizaje Automático

**Departament de Matemàtiques i el Màster Universitari Investigació
Aplicada en Estudis Feministes, de Gènere i Ciutadania. 9/02/2024**

Javier Sánchez Monedero (Universidad de Córdoba)

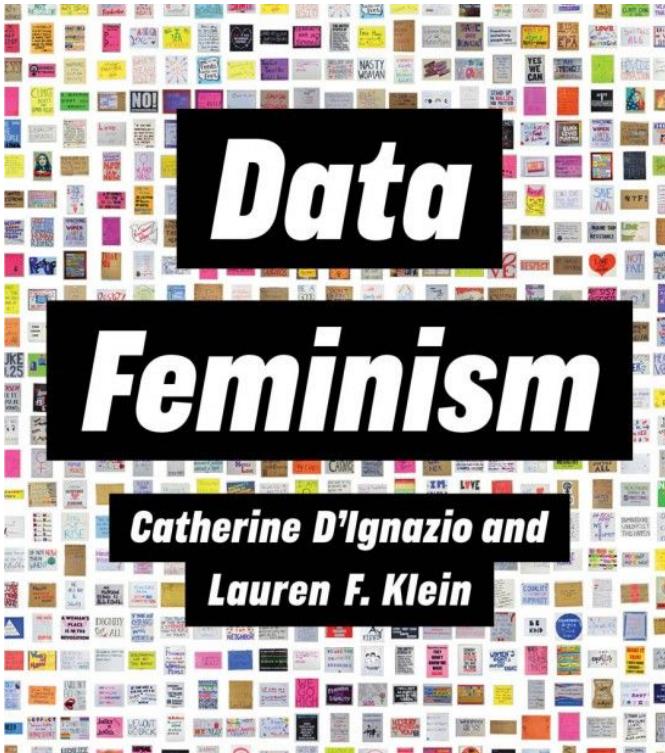
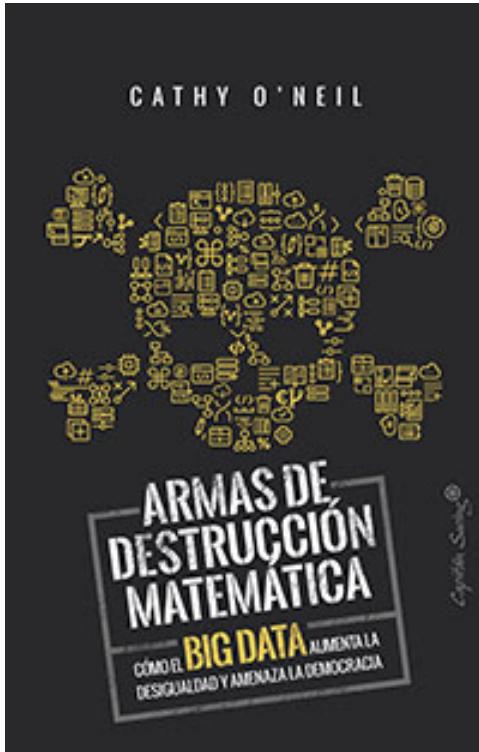
Objetivos

- Introducción y motivación a FATE en inteligencia artificial
- Cuantificando y mitigando sesgos: [FairLearn](#)



Introducción y motivación a FATE

¿Por dónde empezar? Libros



FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

CONTENTS

PREFACE

ACKNOWLEDGMENTS

- 1 [INTRODUCTION](#) [PDF](#)

- 2 [WHEN IS AUTOMATED DECISION MAKING LEGITIMATE?](#) [PDF](#)

We explore what makes automated decision making a matter of normative concern, situated in bureaucratic decision making and its mechanical application of formalized rules.

- 3 [CLASSIFICATION](#) [PDF](#)

We introduce formal non-discrimination criteria in a decision-theoretic setting, establish their relationships, and illustrate their limitations.

- 4 [RELATIVE NOTIONS OF FAIRNESS](#) [PDF](#)

We introduce formal non-discrimination criteria in a decision-theoretic setting, establish their relationships, and illustrate their limitations.

- 5 [CAUSALITY](#) [PDF](#)

We dive into the rich technical repertoire of causal inference and how it helps articulate and address shortcomings of the classification paradigm, while raising new conceptual and normative questions.

¿Por dónde empezar? En vídeo

- Documental [Coded Bias](#)
- TED Talk de Joy Buolamwini
[How I'm fighting bias in algorithms](#)

CODED BIAS

THERE IS NO ALGORITHM FOR TRUTH



FATE:

- **Fairness:** imparcialidad/ecuanimidad
- **Accountability:** rendición de cuentas
- **Transparency:** transparencia
- **Ethics:** ética

facctconference.org

facctconference.org/network



Objetivos del seminario

Discriminación en **sistemas/modelos** que toman decisiones trascendentales

- Esto no considera otras formas de discriminación o injusticia
- Las cuestiones de discriminación/igualdad necesitan de otro tipo de intervenciones no técnicas (ver libros recomendados)

La discriminación no es un concepto general, depende:

- Dominio del problema
- Grupo social

Grupos protegidos

Clases protegidas (no en todos los contextos):

- EEUU: “raza”, color, sexo, religión, ciudadanía, embarazo, edad...
- España: género, embarazo, “raza” (ley igualdad de trato), embarazo...

La definición de grupos protegidos va más allá e incluye las categorías **no binarias** y la **interseccionalidad**

There's No Scientific Basis for Race—It's a Made-Up Label. National Geographic. 2018, March 12.



Ley integral igualdad de trato y no discriminación

Artículo 23 Ley 15/2022, de 12 de julio:

Artículo 23. *Inteligencia Artificial y mecanismos de toma de decisión automatizados.*

1. En el marco de la Estrategia Nacional de Inteligencia Artificial, de la Carta de Derechos Digitales y de las iniciativas europeas en torno a la Inteligencia Artificial, las administraciones públicas favorecerán la puesta en marcha de mecanismos para que los algoritmos involucrados en la toma de decisiones que se utilicen en las administraciones públicas tengan en cuenta criterios de minimización de sesgos, transparencia y rendición de cuentas, siempre que sea factible técnicamente. En estos mecanismos se incluirán su diseño y datos de entrenamiento, y abordarán su potencial impacto discriminatorio. Para lograr este fin, se promoverá la realización de evaluaciones de impacto que determinen el posible sesgo discriminatorio.

2. Las administraciones públicas, en el marco de sus competencias en el ámbito de los algoritmos involucrados en procesos de toma de decisiones, priorizarán la transparencia en el diseño y la implementación y la capacidad de interpretación de las decisiones adoptadas por los mismos.

3. Las administraciones públicas y las empresas promoverán el uso de una Inteligencia Artificial ética, confiable y respetuosa con los derechos fundamentales, siguiendo especialmente las recomendaciones de la Unión Europea en este sentido.

4. Se promoverá un sello de calidad de los algoritmos.

Las personas también tienen sesgos



Diferencias (O'Neil 2016):

- Sistematización
- Escala
- Nuevos grupos "digitales" discriminados

Breve introducción al aprendizaje máquina

Programación tradicional

Reglas explícitas:

```
si email contiene Viagra  
    entonces marcarlo como  
es-spam;  
si email contiene ...;  
si email contiene ...;
```

Ejemplos de Jason's Machine Learning 101

Programas de aprendizaje automático:

Aprender de los ejemplos:
intentar clasificar algunos emails;
cambiar el modelo para
minimizar errores;
repetir;
...y luego utilizar el modelo aprendido para clasificar.

Como nadie está programando explícitamente a menudo se asume(asumía) que es justo, no discrimina, está libre de sesgos humanos, es efectivo, etc.

Tareas y funciones de pérdida

Aprender de los datos es aprender a "conectar" la entrada con la salida

Entrada	Salida
datos tabulares	riesgo fallo hepático [1,40)
imagen dermatoscópica	dianóstico melanoma [0,1]
comentario de código	código programación (texto más probable)
artículo científico	resumen (texto más probable)

Casos: PNL + Visión Artificial

Algorithmic Bias in Grounded Setting



Casos: reconocimiento facial

Análisis interseccional del rendimiento en reconocimiento facial de Amazon Rekognition. La menor tasa de acierto se da para las mujeres de piel oscura.

August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

98.7% **68.6%** **100%** **92.9%**



**DARKER
MALES**



**DARKER
FEMALES**



**LIGHTER
MALES**



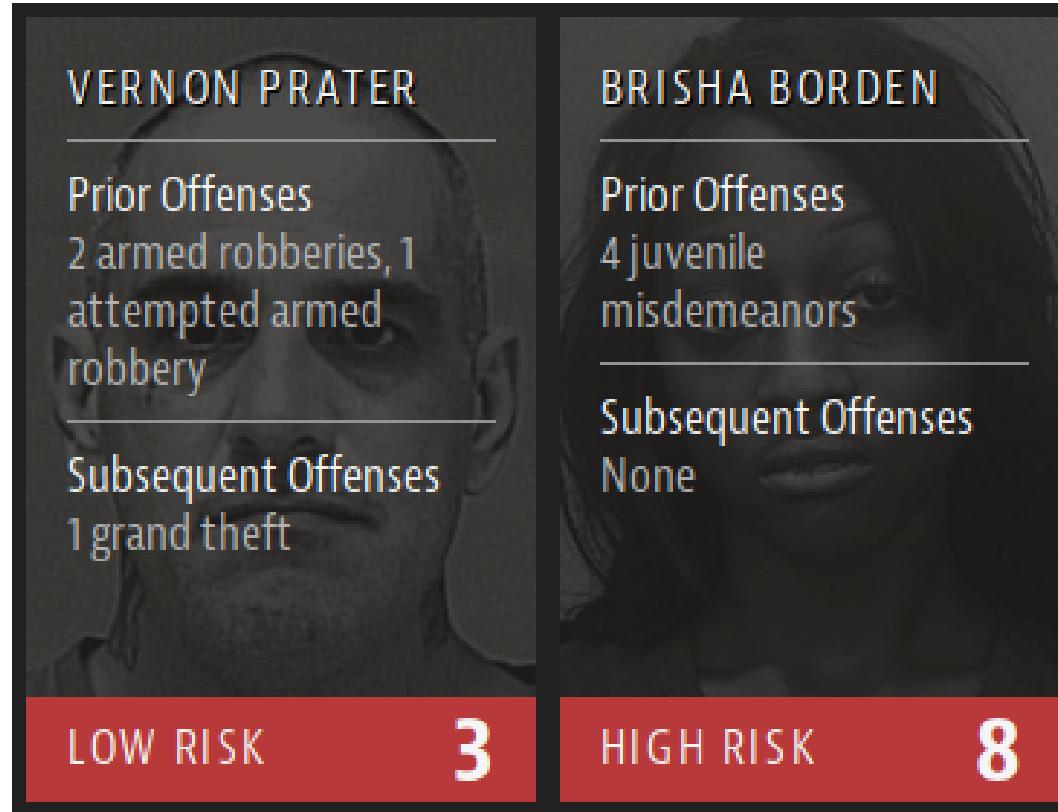
**LIGHTER
FEMALES**

Amazon Rekognition Performance on Gender Classification

Casos: justicia

- **COMPAS** (*Correctional Offender Management Profiling for Alternative Sanctions*): herramienta para calcular puntuaciones de riesgo de reincidencia de una persona en espera de juicio
- Utiliza ML para entrenar un modelo de estimación de **riesgo a partir de los registros históricos**
- **Variables de entrada:** historial criminal, tipo de cargos, género, grupo étnico, edad, preguntas sobre el entorno...
- **Variable dependiente:** grado de riesgo los grados altos van a prisión preventiva

Casos: justicia



Angwin, J., & Larson, J. (2016, May 23). [Machine Bias](#). ProPublica.

A1. ¿Cómo cuantificarías el sesgo en los problemas anteriores?

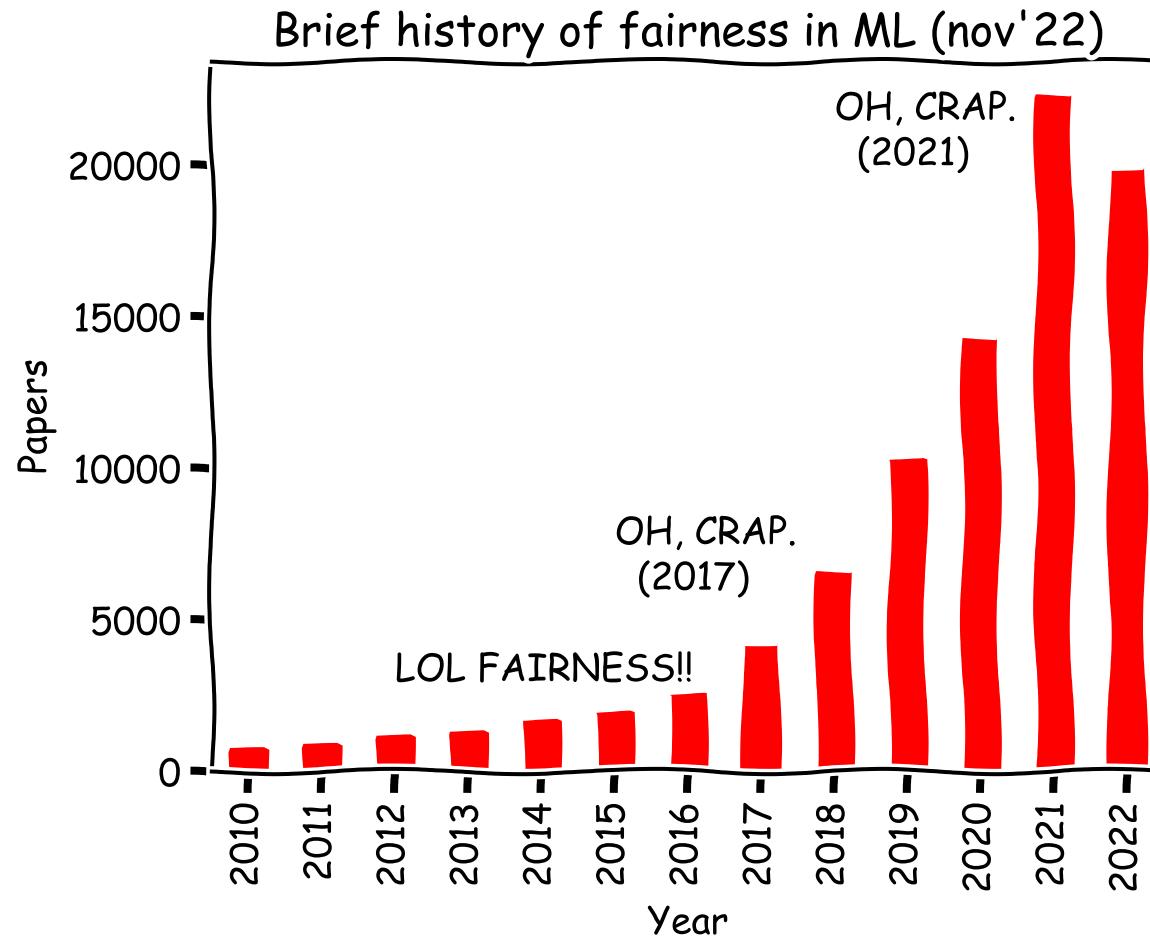


- **Reconocimiento facial:** el modelo tiene menos precisión identificando mujeres con piel oscura
- **Justicia:** el modelo sobreestima el riesgo de reincidencia de afroamericanos
- **Procesamiento lenguaje natural:** el sistema reproduce estereotipos de género asociados a profesiones

Cuantificando y mitigando sesgos

¿Cómo medir y mitigar el sesgo?

Ecuanimidad sin hacer nada (*unawareness*)



Análisis exploratorio

- Comprobar distribución (prevalecia/prior) etiqueta de clase
- Comprobar distribución (prevalecia/prior) etiqueta de clase por grupos
- Comprobar:
 - Visual
 - Estadística descriptiva
 - Contraste de hipótesis

Un ejemplo excelente lo podeis ver en Straw, I., & Wu, H. (2022).

Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics*, 29(1), e100457.
<https://doi.org/10.1136/bmjhci-2021-100457>

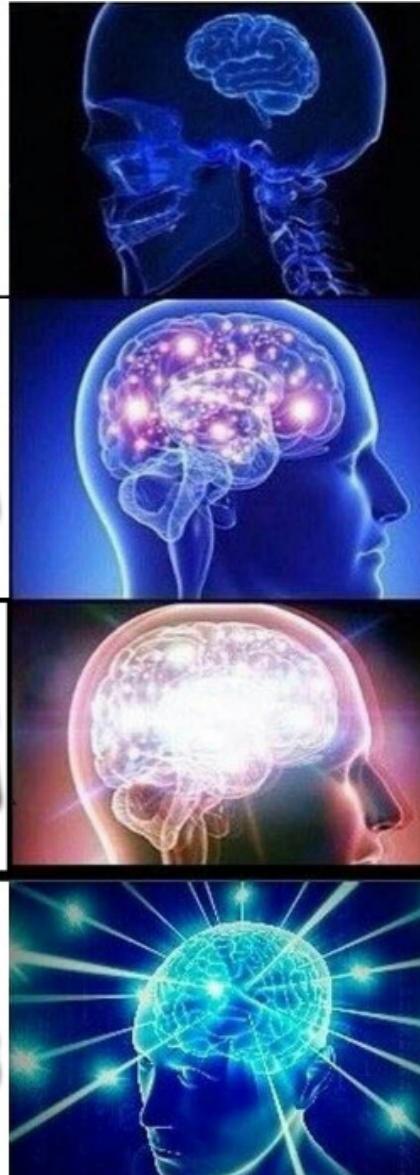
**MIRAR
LOS DATOS**

**MIRAR EL
HISTOGRAMA**

**ESTADÍSTICA
DESCRIPTIVA**

**CONTRASTE
DE HIPÓTESIS**

imgflip.com



El "zoo" de las métricas de ecuanimidad

	notion	use of Y	condition
group fairness	Demographic Parity	-	equal acceptance rate across groups
	Conditional Demographic Parity	-*	equal acceptance rate across groups in any strata
	Equal Accuracy	✓	equal accuracy across groups
	error parity	✓	equal FPR and FNR across groups
individual fairness	Equality of Odds	✓	equal precision across groups
	Predictive Parity	✓	
	FTU/Blindness	-	no explicit use of sensitive attributes
causality-based fairness	Fairness Through Awareness	-*	similar people are given similar decisions
	Counterfactual Fairness	-	an individual would have been given the same decision if she had had different values in sensitive attributes
	path-specific Counterfactual Fairness	-	same as above, but keeping fixed some specific attributes

* there are exceptions to these cases where Y is actually employed, e.g. CDP conditioning on Y becomes Equality of Odds, and there are notions of individual fairness that use a similarity metric defined on the target space ([Berk et al., 2017](#)).

A3. Caso judicial

- Supongamos test genérico (con o sin técnicas estadísticas) de estimación de riesgo de reincidencia. ¿Qué errores debemos minimizar?
- Respecto a la clase: ¿qué metricas nos interesan?
- ¿Y si el test implica pérdida de libertad?

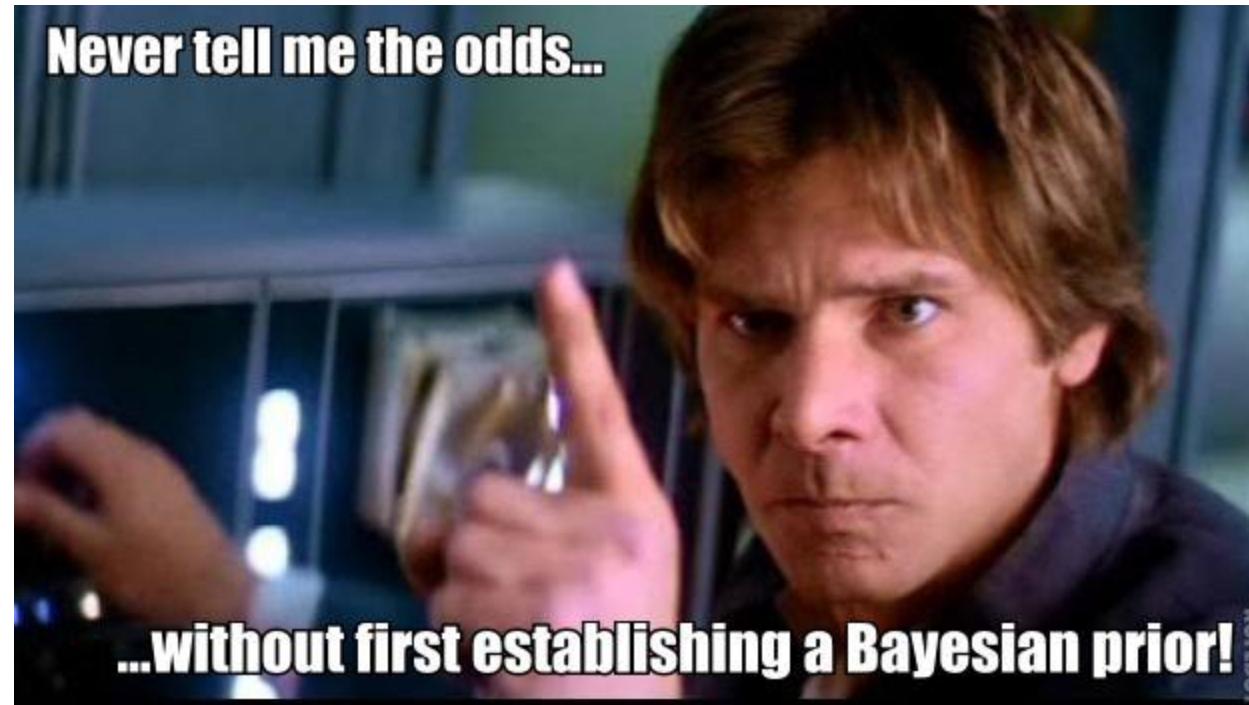
Casos: COMPAS

- **ProPublica:** el sistema discrimina porque sobreestima el riesgo para las personas afroamericanas (falsos positivos diferentes para los grupos: 44,8 % vs 23,4 %)
- **Northpointe:** el sistema no discrimina porque clasifica el riesgo alto por igual (verdaderos positivos similares para todos los grupos étnicos: 63 % vs 59 %)

Larson, J., & Angwin, J. (2016, May 23). [How We Analyzed the COMPAS Recidivism Algorithm](#). ProPublica.

Casos: COMPAS

¿Cómo pueden ser compatibles las definiciones matemáticas de equanimidad de ProPublica y Northpointe?

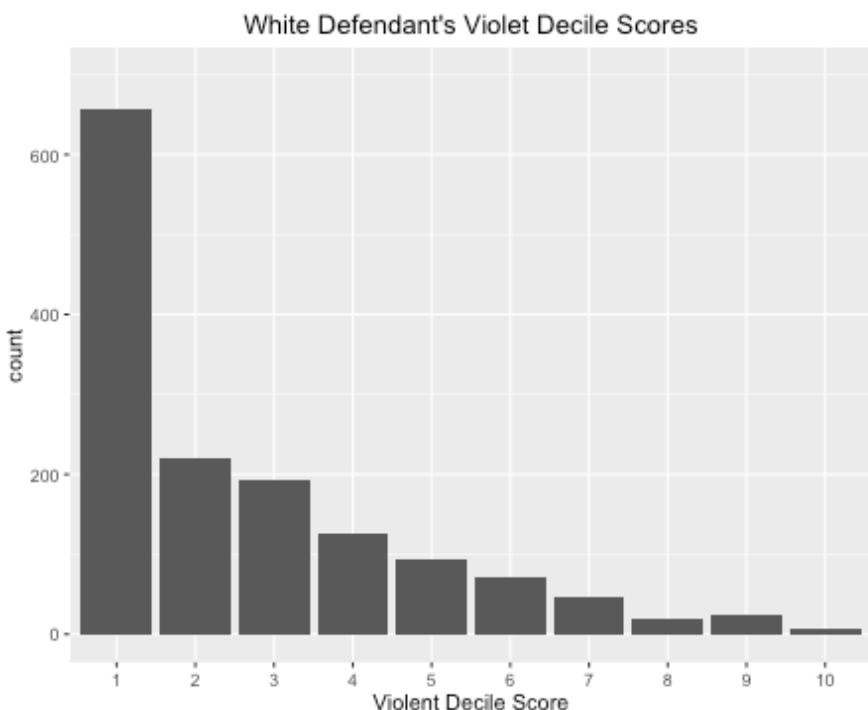
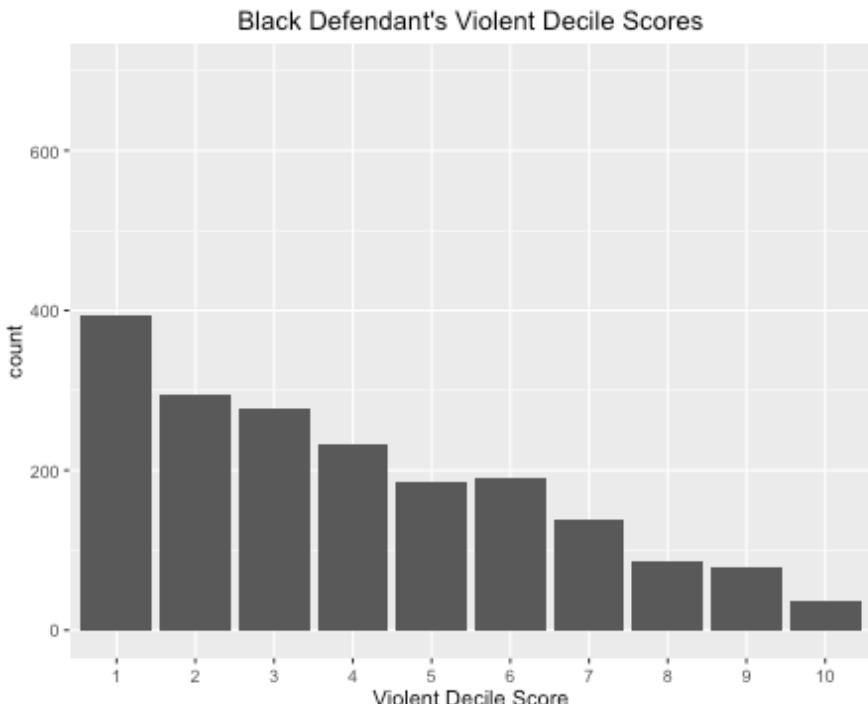


Source [Han Solo and Bayesian Priors](#)

Casos: COMPAS

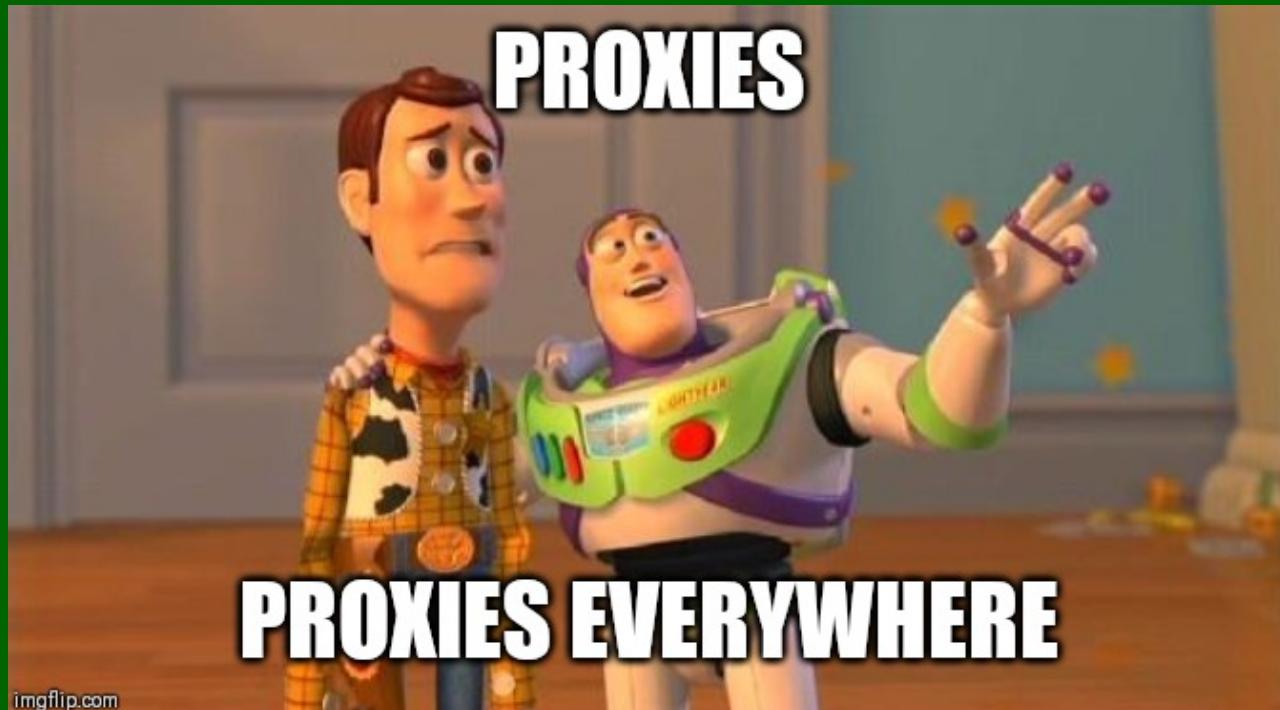
Es matemáticamente compatible porque la prevalencia/frecuencia base/probabilidad a priori de los dos grupos es diferente (ver Chouldechova (2017)).

A. Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 2017. <https://doi.org/10.1089/big.2016.0047>



A4. ¿Cómo podemos mitigar?

- Ya tenemos una medida del sesgo estadístico
- ¿Cómo podríamos mitigar?
- Pero antes: ¿tiene sentido una intervención estadística/algorítmica?



Técnicas de mitigación de sesgos

UNDERSTANDING BIAS

Socio-technical causes of bias

- Data generation
- Data collection
- Institutional bias

Bias manifestation in data

- Sensitive features & causal inferences
 - Data representativeness
 - Data modalities

Fairness definition

- Similarity-based
- Causal reasoning
- Predicted outcome
- Predicted & actual outcome
- Predicted probabilities & actual outcome

MITIGATING BIAS

Pre-processing

- Instance class modification
- Instance selection
- Instance weighting

In-processing

- Classification model adaptation
- Regularization / Loss function s.t. constraints
 - Latent fair classes

Post-processing

- Confidence/probability score corrections
- Promoting/demoting boundary decisions
- Wrapping a fair classifier on top of a black-box baselearner

Fuente Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., ... Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1356.

<https://doi.org/10.1002/widm.1356>

Herramientas ML para mitigación y explicabilidad



<https://fairlearn.org/>

Otras:

<https://ai-fairness-360.org/>

<https://pair-code.github.io/what-if-tool/>

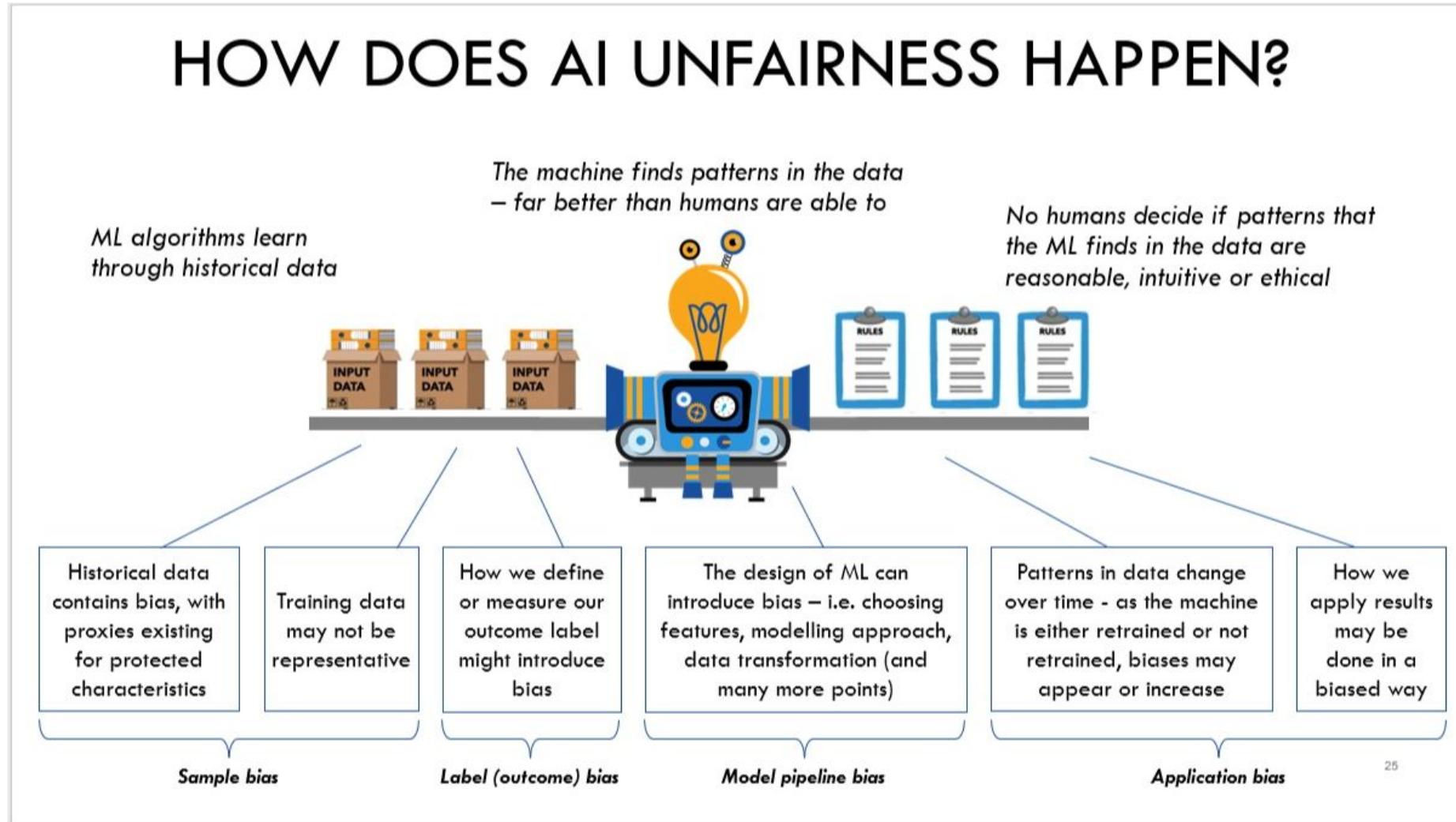
Cuaderno Jupyter con FairLearn y COMPAS

- Datos de reincidencia general de COMPAS
- Experimentos simplificados de ProPublica

<https://github.com/javism/seminarioUJI>

Resumen y Conclusiones

Recap: Fuentes de sesgo



Resumen

- El paso de prototipos de investigación a aplicaciones reales de la inteligencia artificial ha motivado la aparición de muchas áreas
- No solo FATE: IA robusta, privacidad en IA (aprendizaje federado, cifrado homeomórfico...), interacción persona-máquina (HCI)...
- Áreas implicadas según contexto: ética, derecho, política **¡¡Sistemas sociotécnicos!!**
- Regulaciones (IA Act, GDPR, Ley Rider, AESIA...) y estándares (IEEE,ISO)
- Oportunidades de aprendizaje y comprender mejor los problemas y los conceptos de estadística.

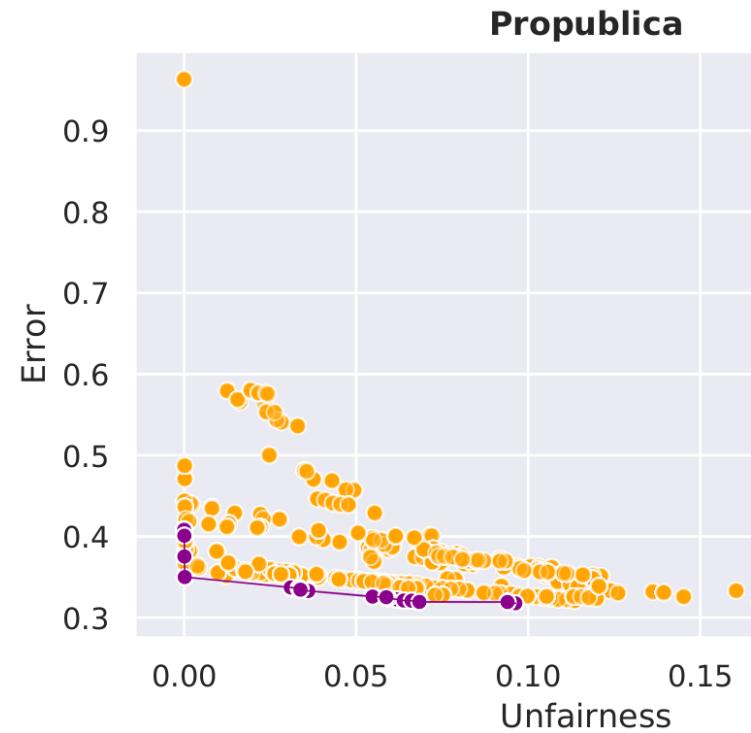
Trabajos relacionados de AYRNA

Explorar límites de precisión vs ecuanimidad

Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *Int J Intel Sys*, 36(4), 1619–1643. <https://doi.org/10.1002/int.22354>

Gender-Equity model for Liver Allocation

El grupo AYRNA, IMIBIC y otros centros trabajan en alternativas al MELD que no discriminan por género como estimador de riesgo de mortalidad en trasplantes hepáticos. <https://gema-transplant.com/>



Trabajos relacionados de AYRNA

Desarrollo Ley Rider

Guía práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral. *Ministerio de Trabajo y Economía Social. Gobierno de España.* 2022.

<https://prensa.mites.gob.es/WebPrensa/noticias/laboral/detalle/4125>

Proyecto AlgoRace

Proyecto AlgoRace. Investigación sobre discriminación racial e inteligencia artificial. 2021-2024. <https://algorace.org/>

Referencias (I)

- O'Neil, C (2018). Armas de destrucción matemática. Capitán Swing.
<https://capitanswing.com/libros/armas-de-destruccion-matematica/>
- Catherine D'Ignazio and Lauren F. Klein (2020). Data Feminism. MIT Press.
<https://mitpress.mit.edu/9780262044004/>
- Solon Barocas and Moritz Hardt and Arvind Narayanan (2019). *Fairness and Machine Learning: Limitations and Opportunities*. <http://www.fairmlbook.org>
- Moritz Hardt (2020). *Fairness and Machine Learning* ([Part 1](#), [Part 2](#)) (MLSS 2020)
- Zhao, J. et. al (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. <https://www.aclweb.org/anthology/D17-1319>
- Buolamwini (2019). [Response: Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces](#).

Referencias (II)

- Verna, E. C., & Lai, J. C. (2020). Time for Action to Address the Persistent Sex-Based Disparity in Liver Transplant Access. *JAMA Surgery*, 155(7), 545–547. <https://doi.org/10.1001/jamasurg.2020.1126>
- Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics*, 29(1), e100457. <https://doi.org/10.1136/bmjhci-2021-100457>
- Castelnovo, A., Crupi, R., Greco, G. et al. A clarification of the nuances in the fairness metrics landscape. *Sci Rep* 12, 4209 (2022). <https://doi.org/10.1038/s41598-022-07939-1>
- Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., ... Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1356. <https://doi.org/10.1002/widm.1356>
- A. Valdivia, C. Hyde-Vaamonde, J. García-Marcos. Judging the algorithm: A case study on the risk assessment tool for gender-based violence implemented in the Basque country. <https://arxiv.org/abs/2203.03723>