

Machine Teaching

Ética en Machine Learning

Si piensas en términos de “*machine learning*”, pones el foco en que la MÁQUINA está aprendiendo a hacer algo que tú quieres que haga...

En cambio, si piensas en términos de “*machine teaching*”, pones el foco en lo que **TÚ** estás enseñando a la máquina a hacer...

...LA RESPONSABILIDAD ES TUYA.

“A subtle difference in language, but a big difference in understanding...”



Las buenas intenciones...

Amazon AI recruiting tool
2015

Los empleadores han soñado durante mucho tiempo con aprovechar la tecnología para ampliar la red de contratación y reducir la dependencia de las opiniones subjetivas de los reclutadores humanos ... *"Todos querían este santo grial"*, dijo una de las personas. *"Ellos literalmente querían un motor al que le das 100 curriculums, que escupa los cinco primeros, y contratarlos"*.

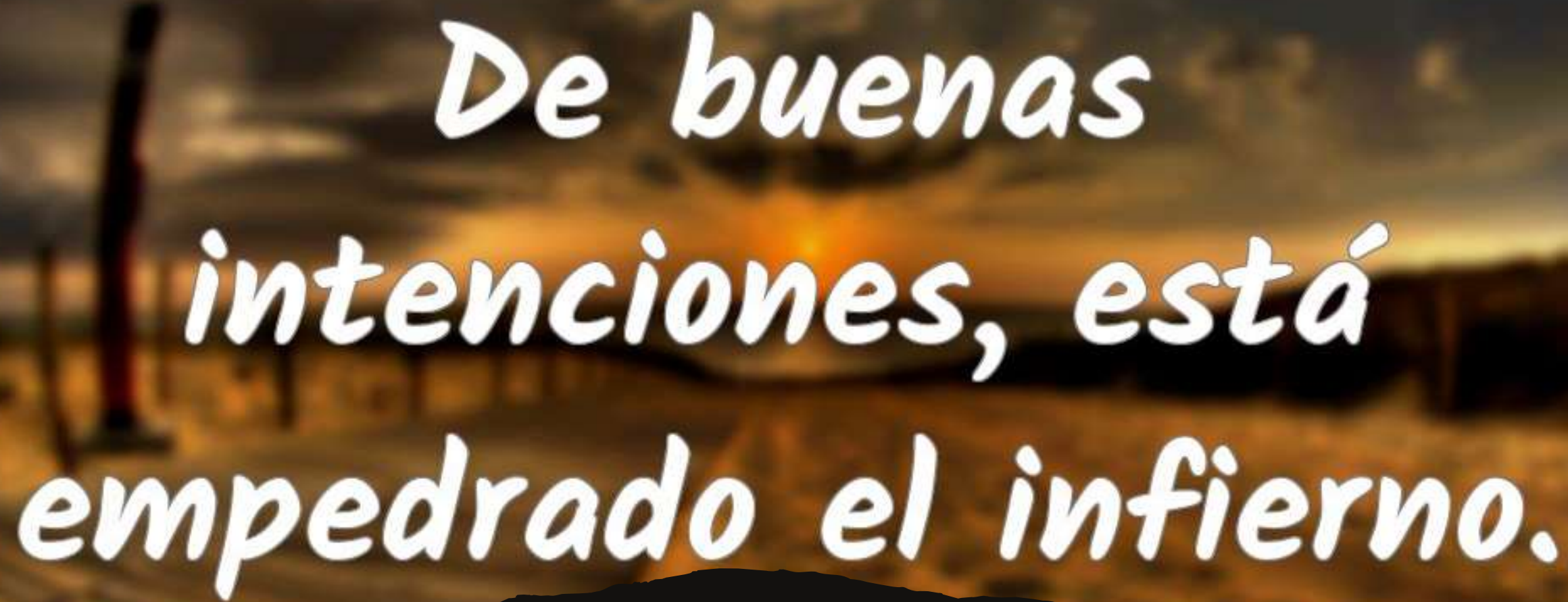
Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 11 de octubre de 2015

<https://www.youtube.com/watch?v=QvRZuHQBTPs>

Resultados sesgados... ¿quién sale perdiendo?

- Las candidatas descartadas por el algoritmo.
- Amazon,
 - Daño reputacional... ¿Cuántas empresas pueden sobrevivir a tal daño?
 - Pérdida de talento femenino.
- Los empleados de amazon que desarrollaron el algoritmo (el departamento fue desmantelado).
- La sociedad en su conjunto,
 - por el daño que a medio y largo plazo van a hacer los algoritmos mal diseñados...y
 - por perder los enormes beneficios de una IA bien diseñada.
- La inteligencia artificial ... ¿nuevo invierno de la IA?

Todo el mundo, vamos

The background of the image is a photograph of a sunset or sunrise. The sun is a bright, glowing orb in the center of the frame, partially obscured by dark, silhouetted clouds. The sky is a mix of orange, yellow, and dark grey. Below the horizon, there is a dark, silhouetted shape that appears to be a person standing on a beach or shore, looking out at the water. The overall mood is contemplative and dramatic.

*De buenas
intenciones, está
empedrado el infierno.*

Un toque de sabiduría popular...

Sesgos, discriminación, injusticia...

¿Nada nuevo bajo el sol?

¿Qué nos estamos jugando con el uso
[bien intencionado] de machine
learning?



“A recipe for negligence amplified”

- De un alto nivel de abstracción a un inédito nivel de distracción.

En lugar de programar "haz esto, luego esto, luego esto, luego ...", puedes decir "intenta obtener una buen Accuracy con estos datos"...

- Agravado por el “misticismo matemático”

La capa de matemáticas hace que las personas piensen aún menos en lo que están haciendo cuando eligen sus objetivos y dataset.

- Y la evolución de las herramientas...

A medida que las herramientas mejoren, las barreras de entrada serán tan bajas que cualquiera podrá montar un modelo de machine learning sin apenas saber lo que hace...

“Alas, there are no brains here but your own and the math is a tiny layer of objectivity in the middle of your subjectivity sandwich.”

*“When the wellbeing of our fellow humans is at stake, **thoughtlessness** is a hazard. ML/AI is a thoughtlessness enabler”.*

Forget the robots! Here’s how AI will get you

The real reason AI is more dangerous than traditional software

[Cassie Kozyrkov](#), Chief Decision Scientist at Google.

<https://towardsdatascience.com/forget-the-robots-heres-how-ai-will-get-you-b674c28d6a34>


Todo lo anterior, sumado a,

- El potencial alcance y relevancia del daño,

Un algoritmo puede tener impacto significativo sobre la vida de gran cantidad de personas, en relación con la educación, las oportunidades de empleo o el acceso a servicios financieros, entre otras áreas.

- Y la duración de los efectos en el tiempo.

Se puede extender mucho más allá del periodo de vida de una persona que toma decisiones sesgadas, pudiendo ser ilimitado si no se corrige...



***Por favor,
no seamos el
mono***

Ética en la IA

Definición de ética según la Real Academia Española

- Conjunto de normas morales que rigen la conducta de la persona en cualquier ámbito de la vida (*Ética profesional, cívica, deportiva*).
- Parte de la filosofía que trata del bien y del fundamento de sus valores.

Ethics guidelines for trustworthy AI

EU Commission.

Abril 2019

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

“Las presentes directrices pretenden fomentar la innovación responsable y sostenible en el campo de la IA en Europa.

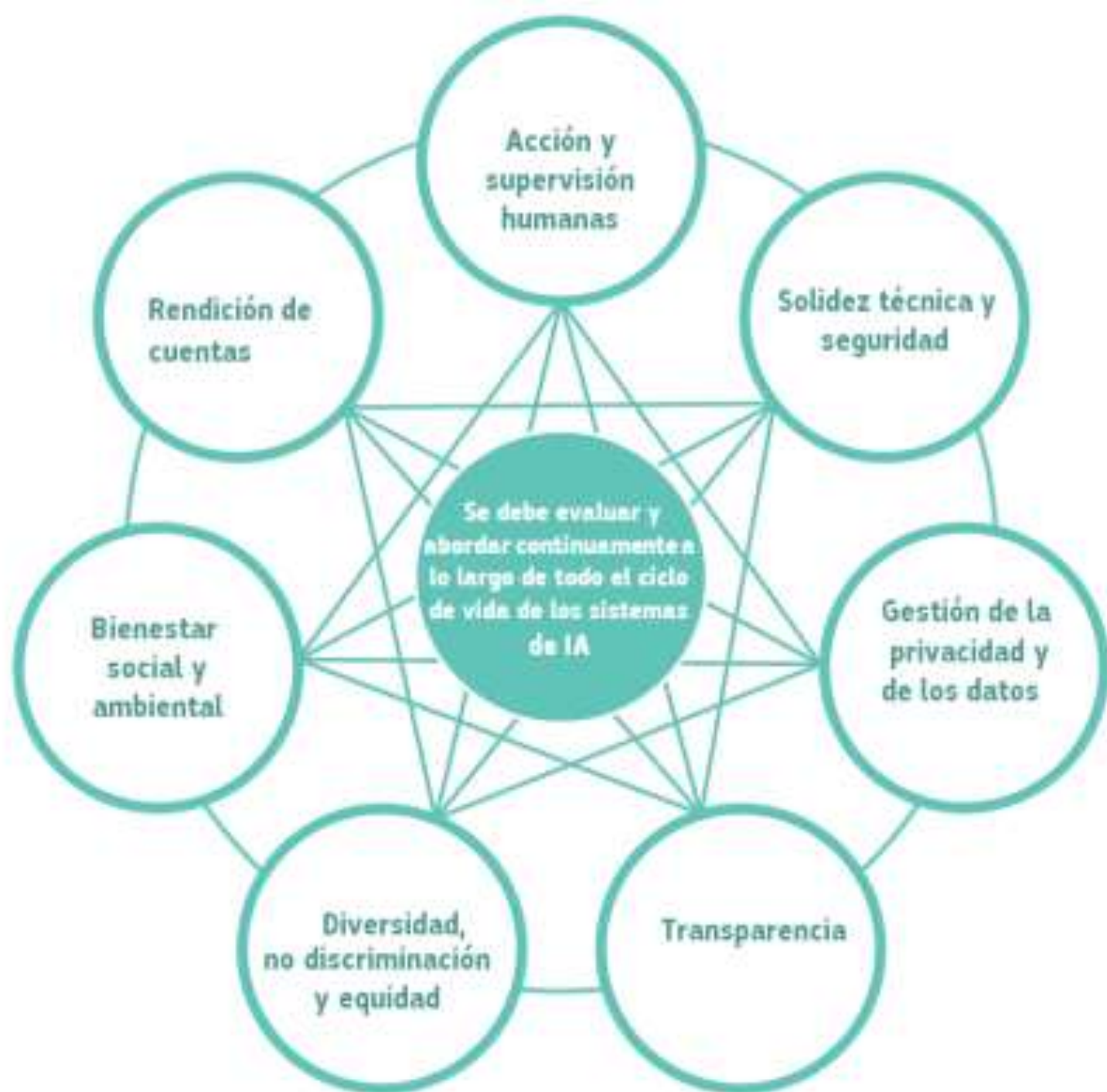
Su finalidad es convertir la ética en un pilar fundamental para desarrollar un enfoque único con respecto a la IA, que busque beneficiar, empoderar y proteger tanto la prosperidad humana a nivel individual como el bien común de la sociedad.”

A large, diverse crowd of people is shown from a slightly elevated perspective, looking down into the crowd. The people are of various ages and ethnicities, many wearing hats and sunglasses, suggesting a sunny outdoor event. A semi-transparent dark rectangle is overlaid on the lower half of the image, containing text in red and white.

La ética de la inteligencia artificial es un subcampo de la ética aplicada que estudia los problemas éticos que plantea el desarrollo, despliegue y utilización de la IA.

SIETE requisitos clave que los sistemas de IA deben cumplir para ser CONFIABLES.

Todos ellos de la misma importancia, que se apoyan unos sobre otros y deben ser implementados y evaluados a lo largo de todo el ciclo de vida de la IA.



1. Acción y supervisión humanas,
Incluidos los derechos fundamentales, la acción humana y la supervisión humana.

2. Solidez técnica y seguridad
Incluida la capacidad de resistencia a los ataques y la seguridad, un plan de repliegue y la seguridad general, precisión, fiabilidad y reproducibilidad.

3. Gestión de la privacidad y de los datos, Incluido el respeto de la privacidad, la calidad y la integridad de los datos, así como el acceso a estos.

Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Agencia Española de Protección de Datos, Febrero de 2020.

Objetivo : Abordar las dudas planteadas en el marco de protección de datos de carácter personal y señalar los aspectos más relevantes en la relación IA-RGPD que deben ser tenidos en cuenta desde el diseño y en la implementación de tratamientos que incluyan IA.

<https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf>

4. Transparencia, Incluidas la trazabilidad, la explicabilidad y la comunicación.

5. Diversidad, no discriminación y equidad, Incluida la ausencia de sesgos injustos, la accesibilidad y el diseño universal, así como la participación de las partes interesadas.

6. Bienestar social y ambiental
Incluida la sostenibilidad y el respeto del medio ambiente, el impacto social, la sociedad y la democracia.

7. Rendición de cuentas. Incluidas la auditabilidad, la minimización de efectos negativos y la notificación de estos, la búsqueda de equilibrios y las compensaciones.

DIVERSIDAD,
NO DISCRIMINACIÓN, Y
EQUIDAD



Las decisiones humanas están sesgadas...

Un estudio del año 2011 demostró que existían más probabilidades de que las juntas de libertad condicional liberasen a los convictos a primera hora de la mañana y cuando los jueces acababan de hacer una pausa para comer.

The authors of the peer-reviewed paper looked at more than 1,000 rulings made in 2009 by eight Israeli judges. They found that the likelihood of a favourable ruling peaked at the beginning of the day, steadily declining over time from a probability of about 65% to nearly zero, before spiking back up to about 65% after a break for a meal or snack.

Jonathan Levav, associate professor of business at Columbia University, who co-authored the paper, said: "You are anywhere between two and six times as likely to be released if you're one of the first three prisoners considered versus the last three prisoners considered."

<https://www.theguardian.com/law/2011/apr/11/judges-lenient-break>

Sesgos cognitivos

Patrones sistemáticos de desviación de la norma y / o racionalidad en el juicio.

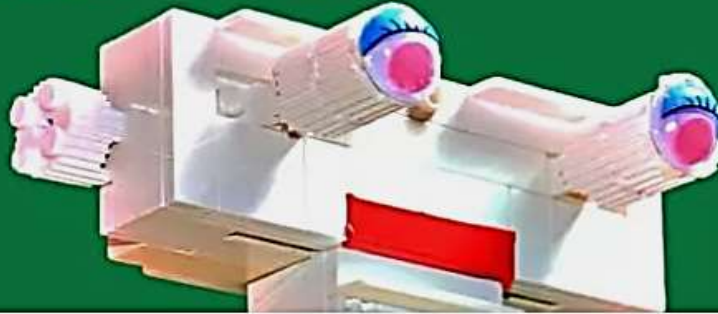
https://en.wikipedia.org/wiki/List_of_cognitive_biases



¿Y las decisiones de los algoritmos?

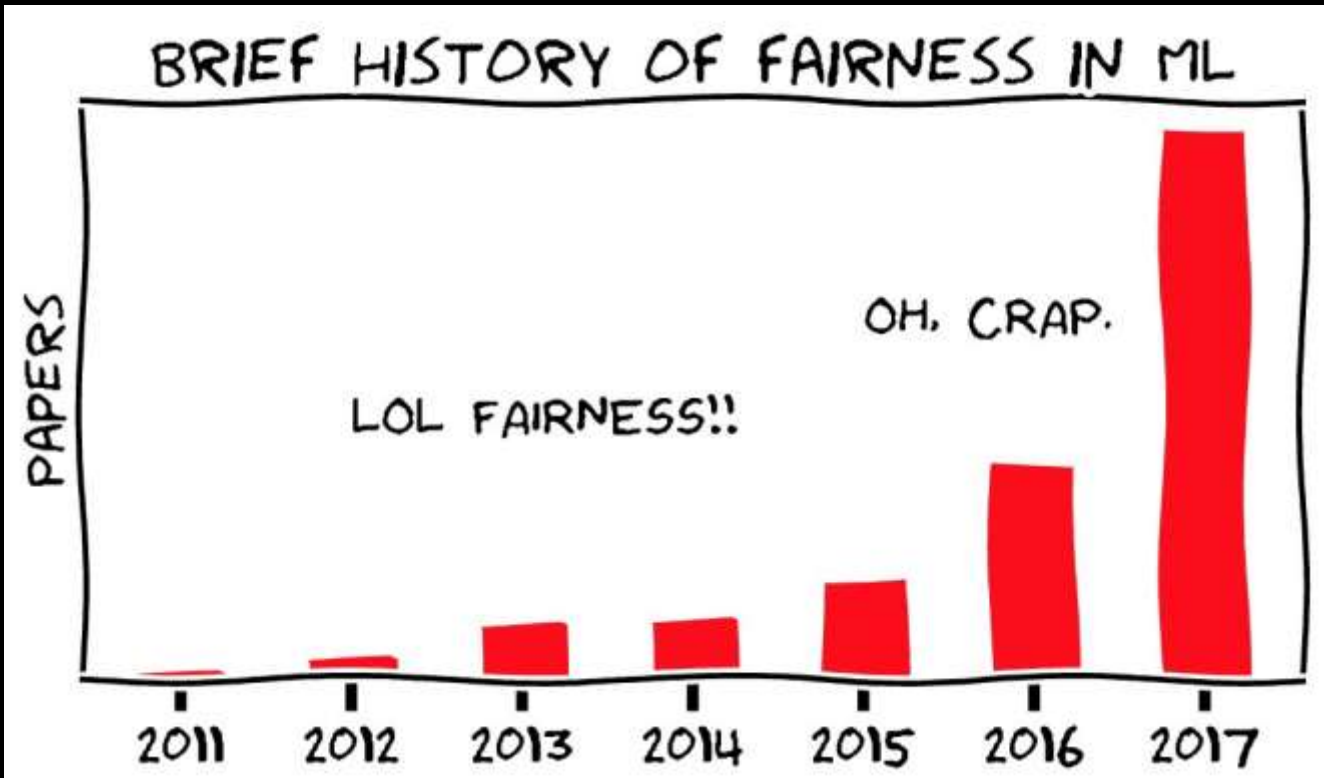
Algorithms Hire Better Than Humans

By Christina Boodée | December 1, 2015 | Informative



“Besides increasing retention rates, using algorithms in hiring can also eliminate bias from the process and increase workplace diversity. “From a human perspective, we like people who are like us,” said [Julie Moreland](#), senior VP of strategy and people science at PeopleMatters, the company that built the assessments for this study. “They’re not thinking about the job, they’re thinking ‘I can work with this person, I relate to them.”

Yup, you read that right. A [study](#) released last month by the National Bureau of Economic Research suggests that algorithms make better hiring decisions than humans do.



Y, según la Wikipedia, la gran mayoría de los papers son de los últimos tres años...

¿Qué entendemos por sesgo algorítmico?

El sesgo algorítmico (*algorithmic bias*) se refiere a las situaciones en las que un sistema informático refleja los sesgos de los seres humanos que lo desarrollaron...

...entendiendo como tales la diferentes maneras en que las experiencias pasadas de los seres humanos distorsionan su percepción y reacción a la información, especialmente en el contexto de tratar a los otros seres humanos injustamente.

[Cassie Kozyrkov](#). Chief Decision Scientist at Google.

"Pay attention to that man behind the curtain. AI bias and what you can do about it." [Towards Data Science](#), 25 de enero de 2019.

<https://medium.com/hackernoon/ai-bias-and-what-you-can-do-about-it-4a6ae48b338e>

¿Qué entendemos por equidad (Fairness) en machine learning?

El compromiso de garantizar una distribución justa e igualitaria de los beneficios y costes de los sistemas de inteligencia artificial , y asegurar que las personas y grupos no sufran sesgos injustos, discriminación ni estigmatización.

Si se pueden evitar los sesgos injustos, los sistemas de IA podrían incluso aumentar la equidad social.[...]

Ethics guidelines for trustworthy AI, European Commission.

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

“Most of the research on the topic of bias and fairness in ML/AI is about making sure that your system doesn’t have a disproportionate effect on some group of users relative to others.”

Cassie Kozyrkov. Chief Decision Scientist at Google.

“Pay attention to that man behind the curtain. AI bias and what you can do about it.” Towards Data Science, 25 de enero de 2019.

<https://medium.com/hackernoon/ai-bias-and-what-you-can-do-about-it-4a6ae48b338e>

Grupo vulnerable

Conjunto de personas que comparten una o varias características de vulnerabilidad.

Variables protegidas

Rasgos o características que han servido de base para el trato injusto y sistemáticamente adverso de los seres humanos en el pasado.

La Carta de los Derechos Fundamentales de la Unión Europea recoge en su artículo 21, relativo a la no discriminación, los motivos de discriminación siguientes, que pueden servir como punto de referencia, entre otros:

- El sexo,
- la raza,
- el color,
- los orígenes étnicos o sociales,
- las características genéticas,
- la lengua,
- la religión o las convicciones,
- las opiniones políticas o de cualquier otro tipo,
- la pertenencia a una minoría nacional,
- el patrimonio,
- el nacimiento,
- la discapacidad,
- la edad o
- la orientación sexual.

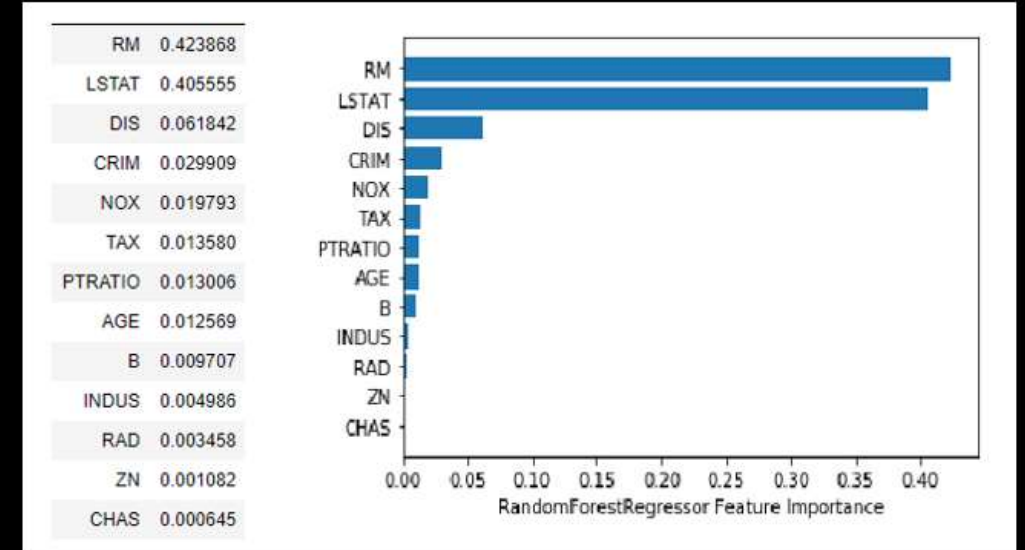
Ethics guidelines for trustworthy AI, Glosario de términos.

Variables materialmente y moralmente irrelevantes.

Boston dataset (Sklearn)

- **B**: Black proportion of population. (1970 US Census)
Muy poco predictiva
- **LSTAT**: Proportion of population that is lower status = $1/2 * (\text{proportion of adults without some high school education and proportion of male workers classified as laborers})$. (1970 US Census)
La segunda variable más predictiva
- Investigación de la variable “B” por M Carlisle en 2019.
Racist data destruction?

<https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8>



DEPRECATED.

En septiembre de 2021, Sklearn declara este dataset como deprecated por problemas éticos.

(<https://mobile.twitter.com/ogrisel/status/1442894248488046595>).

Se elimina de SKlearn para la versión 1.2



¿Podemos evitar que el resultado de los modelos de ML sea discriminatorio simplemente ignorando las variables protegidas?
(Unawareness)

Desgraciadamente, NO



DISPARATE TREATMENT

Utilizar la variable protegida
para hacer predicciones.

DISPARATE IMPACT

NO utilizar la variable
protegida,
pero a pesar de ello obtener
un resultado discriminatorio
a través de variables no
protegidas altamente
correlacionadas con la
variable protegida o proxies.

AMAZON

AI recruitment tool

Para entrenar su modelo de IA, Amazon había utilizado datos históricos de los últimos 10 años, que reflejan el predominio masculino en la industria tecnológica en general y en la plantilla de Amazon en particular.

El modelo de Amazon aprendió incorrectamente que los candidatos masculinos eran preferibles. Penalizó los currículos que incluían la palabra "women", como en "women's chess club captain".

Amazon dejó de usar el algoritmo con fines de reclutamiento.

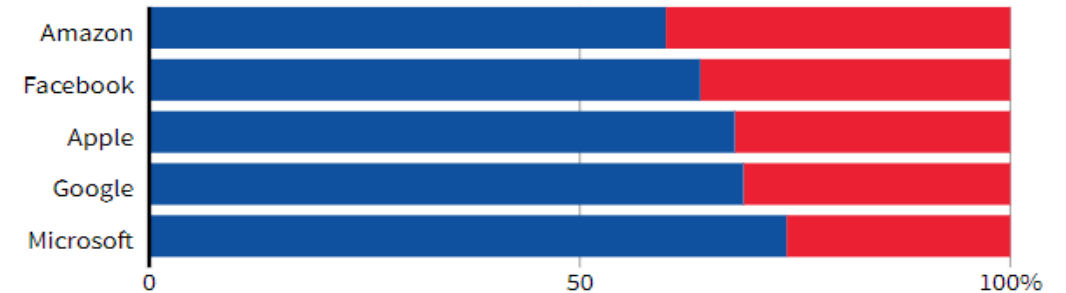
- Variable protegida:
Género del candidato (no utilizada por el modelo)
- Proxies:
Variables relativas a la educación y los hobbies del candidato.

Dominated by men

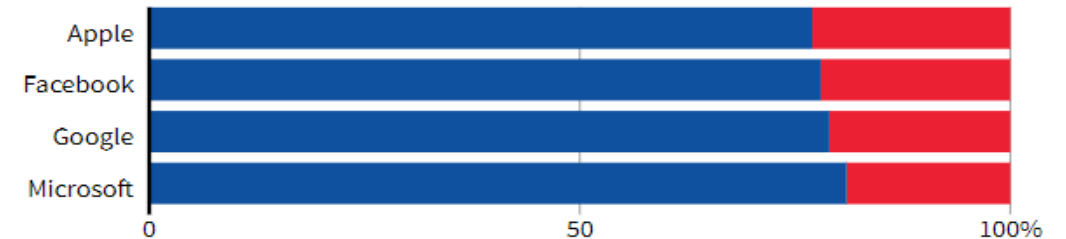
Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

GLOBAL HEADCOUNT

■ Male ■ Female



EMPLOYEES IN TECHNICAL ROLES



Los algoritmos son espejos y reflejan el mundo en que vivimos, no el mundo en que queremos vivir. Reflejan los sesgos de nuestras preguntas y los sesgos de nuestros datos.

[Rahul Bhargava, MIT](#)

<https://www.kdnuggets.com/2019/01/algorithms-arent-biased-we-are.html>



TIPOS DE DAÑOS

- 5 tipos principales de daños producidos por los sesgos algorítmicos.
- No son excluyentes, la mayoría de los ejemplos ilustran varios tipos de daños.

FATE: Fairness, Accountability, Transparency, and Ethics in AI.

[Microsoft Research](#)

Machine Learning and Fairness Webinar

<https://www.youtube.com/watch?v=7CH0xLWQLRw>

1. ALLOCATION

Decisiones de alto riesgo donde modelos de machine learning se utilizan para asignar oportunidades, recursos o información de forma que pueden conllevar impactos negativos significativos en la vida de las personas.



AMAZON AI recruitment tool

Se deniega a las candidatas de género femenino la oportunidad de acceder a un puesto de trabajo técnico (sesgo de género)



COMPAS

Se deniega a acusados de raza negra la oportunidad de acceder al beneficio de la libertad provisional (sesgo racial)

2. QUALITY OF SERVICE

Se produce cuando un sistema basado en machine learning no funciona tan bien para una persona como para otra.

AMAZON : Prime Free Same-Day Delivery

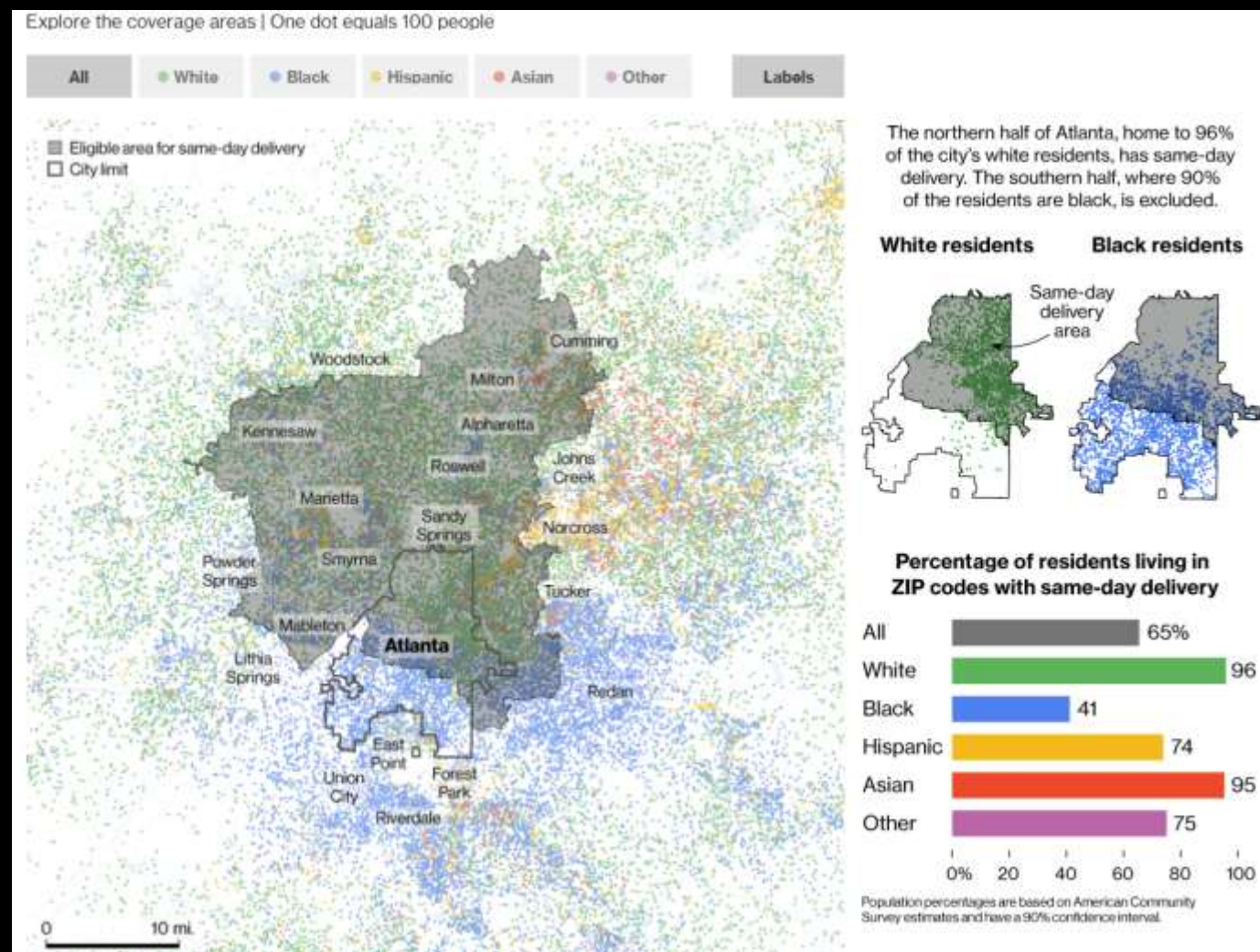
En general, en las ciudades donde el servicio el mismo día no se extiende todos los residentes, los excluidos son desproporcionadamente de raza negra.

- **Variable protegida:**
Demografía del código postal (no utilizada por el modelo)
- **Proxy:**
Concentración de Prime members en el distrito postal.
Los distritos con mayor concentración de suscriptores de Prime tienen mayor nivel de riqueza (\$99 membership fee).

Fuente: Amazon Doesn't Consider the Race of Its Customers. Should It?

[Bloomberg, abril 2016](#)

<https://www.bloomberg.com/graphics/2016-amazon-same-day/>





Open-source code for face detection

Joy Buolamwini

Graduate researcher at the MIT Media Lab and founder of the [Algorithmic Justice League](https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148) – an organization that aims to challenge the biases in decision-making software.

<https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148>

Diciembre 2016

3. STEREOTYPING

Cuando los sistemas basados en algoritmos de machine learning refuerzan estereotipos negativos asociados a grupos protegidos.

Ads related to latanya farrell ⓘ

[Latanya Farrell, Arrested?](#)
www.instantcheckmate.com/
1) Enter Name and State. 2) Access Full Background Checks Instantly.

[Latanya Farrell](#)
www.publicrecords.com/
Public Records Found For: **Latanya Farrell**. View Now.

Ads related to Jill Schneider ⓘ

[Jill Schneider Art](#)
www.posters2prints.com/
Custom Frame Prints and Canvases. Shop Now, SAVE Big + Free Shipping!

[We Found Jill Schneider](#)
www.intelius.com/
Current Phone, Address, Age & More. Instant & Accurate **Jill Schneider**
10,256 people +1'd this page
Reverse Lookup - Reverse Cell Phone Directory - Date Check - Property Records

[Located: Jill Schneider](#)
www.instantcheckmate.com/
Information found on **Jill Schneider** **Jill Schneider** found in database.

*Google ads, black names and white names, racial discrimination, and click advertising.
by Latanya Sweeney. Mayo 2013.*

The image displays two screenshots of the Google Translate interface. The top screenshot shows the English input 'He is a nurse' and 'She is a doctor' being translated into Turkish as 'O bir hemşire' and 'O bir doktor' respectively. The bottom screenshot shows the Turkish input 'O bir hemşire' and 'O bir doktor' being translated back into English as 'She is a nurse' and 'He is a doctor'. This illustrates how the model has learned to associate nursing with females and doctorhood with males, reinforcing gender stereotypes.

Google Translator
Caliskan, Bryson, Narayanan (2017)

4. DENIGRATION

Cuando el sistema en si mismo forma parte de un proceso que es activamente despectivo y ofensivo.



A tan sólo un día de su lanzamiento por Microsoft, el chat bot Tay comenzó a emitir comentarios e insultos racistas y xenófobos (marzo 2016)

5. OVER AND UNDER REPRESENTATION

Cuando un modelo de machine learning arroja resultados todavía más sesgados que la sobre o infra representación de determinados grupos en la realidad.



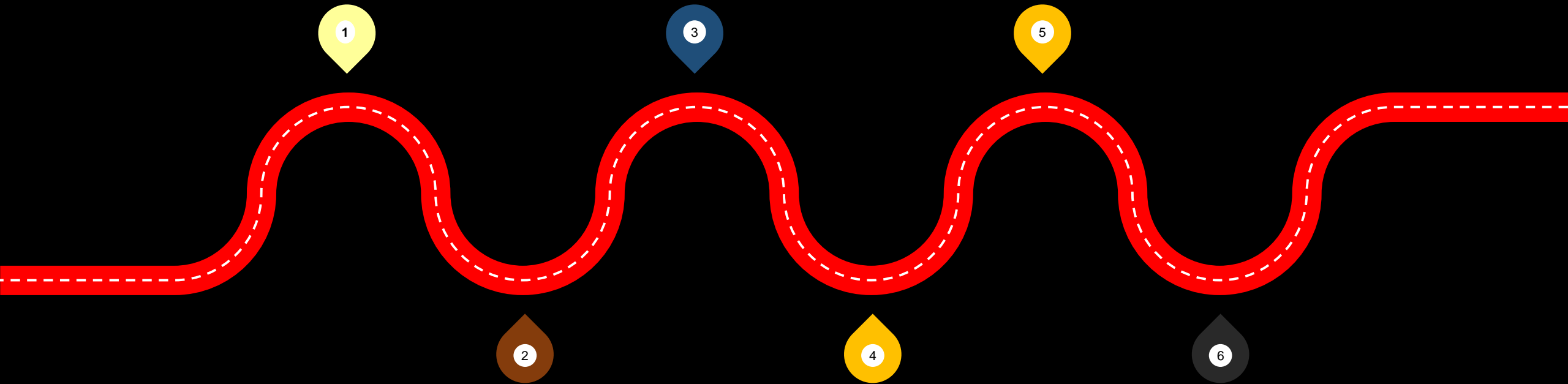
Google image search results for CEO
Kay, Matuszek, Munson (2015)

(Ya corregido por Google)

IA systems	Allocation	Quality of service	Stereotyping	Denigration	Over / Under representation
Hiring system does not rank women as highly as men for technical jobs	X		X		X
Gender classification software misclassifies darker skin women		X			
Machine translation system exhibits male/female gender stereotypes			X	X	
Photo management program labels image of black people as "gorillas".		X		X	
Image searchers for CEO yield only photos of white men on first page.			X		X

CAUSAS DE LOS SESGOS

La equidad es un problema que debe tenerse muy presente a lo largo de todo el proceso de desarrollo y productivización del modelo de machine learning.



Las decisiones que se toman en cada punto del proceso pueden introducir sesgos en el modelo.

ML Pipeline



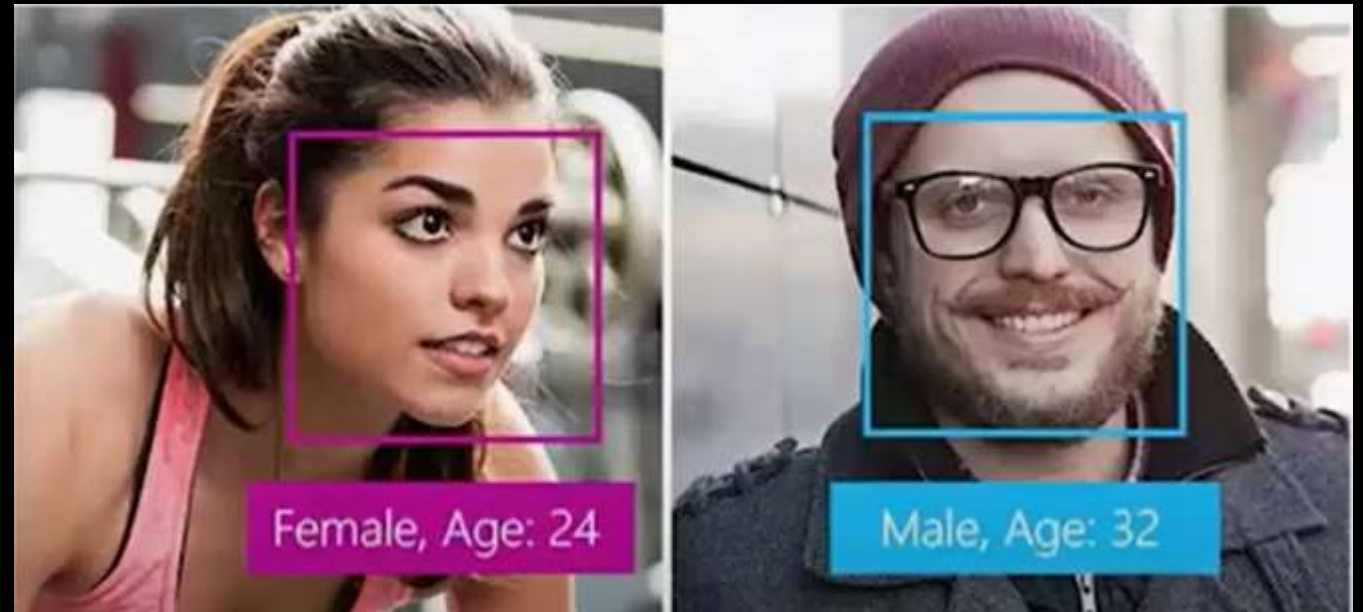
Definición del problema (TASK DEFINITION)

¿Cuál es el problema que estás tratando de resolver con machine learning? ¿Cuál el coste del error para las personas mal clasificadas?



2016, China.

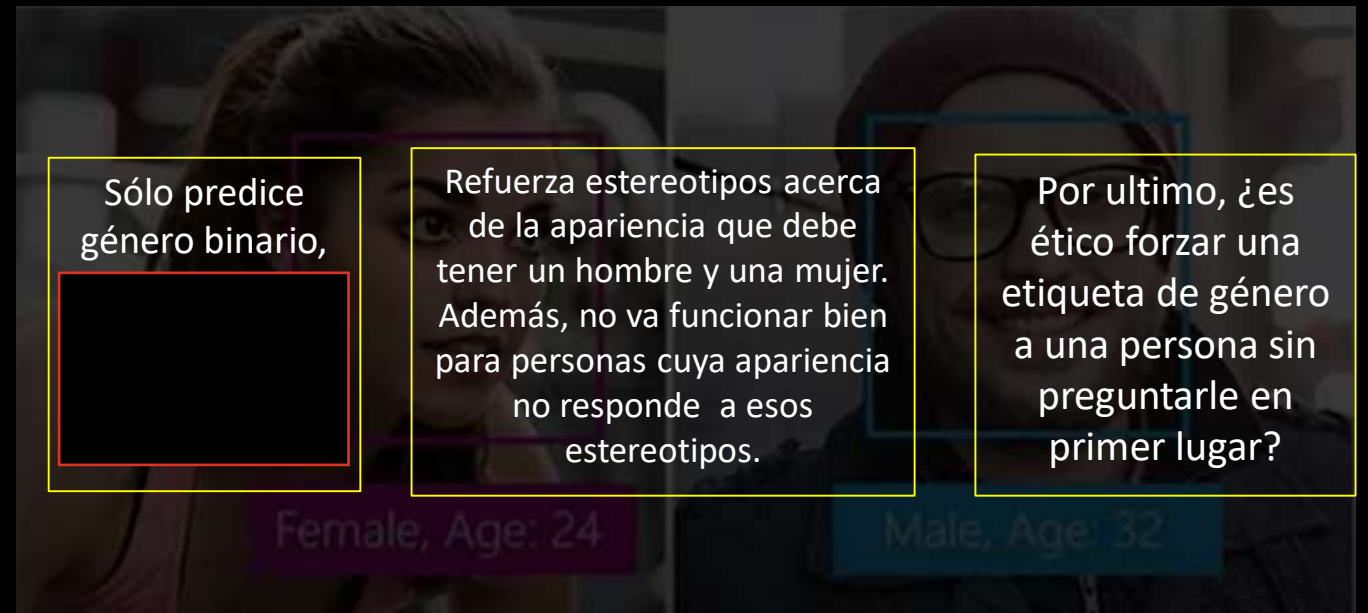
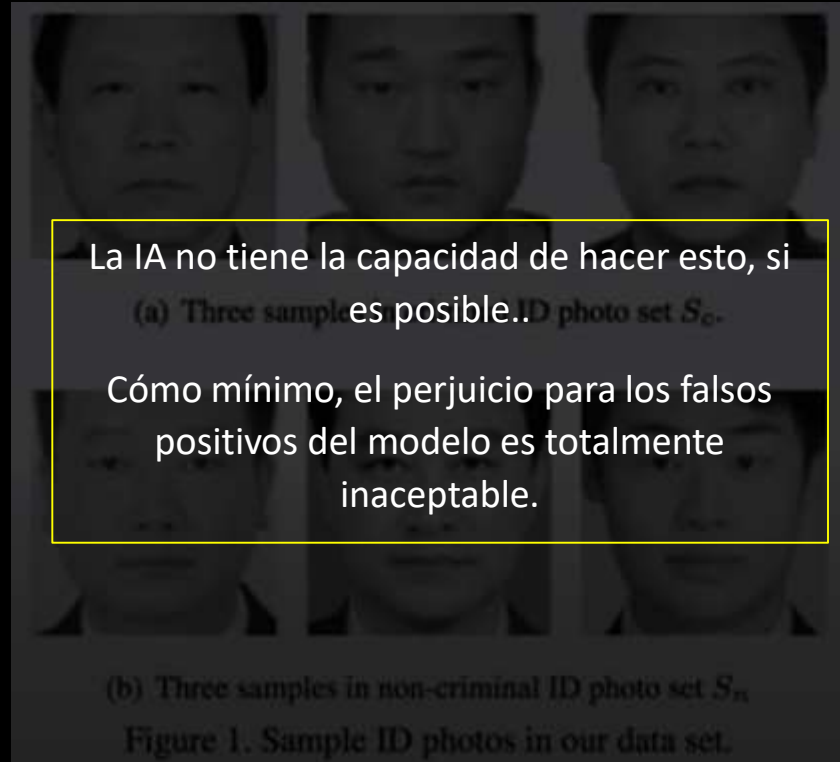
Modelo entrenado para predecir quién va a cometer un crimen en el futuro.



Modelo de clasificación de género

Definición del problema (TASK DEFINITION)

¿Cuál es el problema que estás tratando de resolver con machine learning? ¿Cuál el coste del error para las personas mal clasificadas?



ML Pipeline



Construcción del dataset (DATASET CONSTRUCTION)

Fase muy proclive a la introducción de sesgos.

1. Sesgos sociales o históricos (historical bias)

El dataset refleja los sesgos de nuestra sociedad.



Boston releases Street Bump app that automatically detects potholes while driving

By DAILY MAIL REPORTER

PUBLISHED: 19:37 EST, 20 July 2012 | UPDATED: 20:01 EST, 20 July 2012



View comment

The next time your car hits a pothole, a new technology could help you immediately tell someone who can do something about it.

Boston, smartphone App 2011. Los ciudadanos podían reportar los baches al pasar por la calle para que fueran reparados.

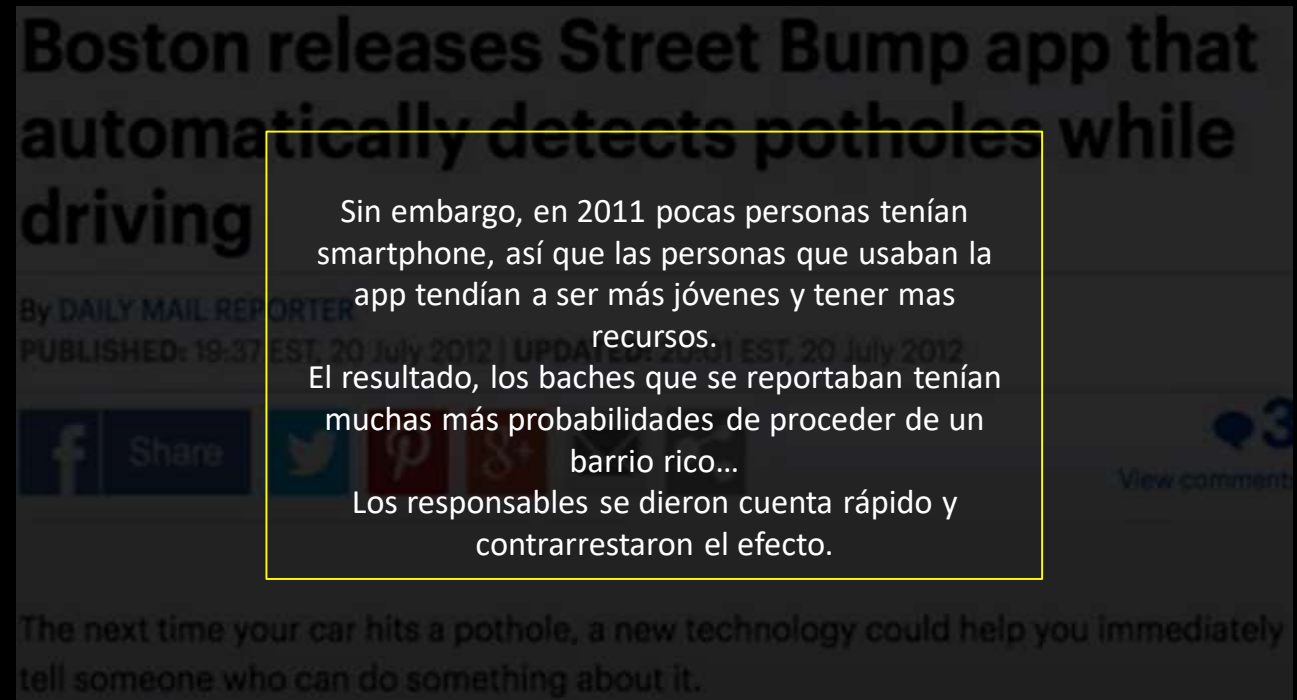
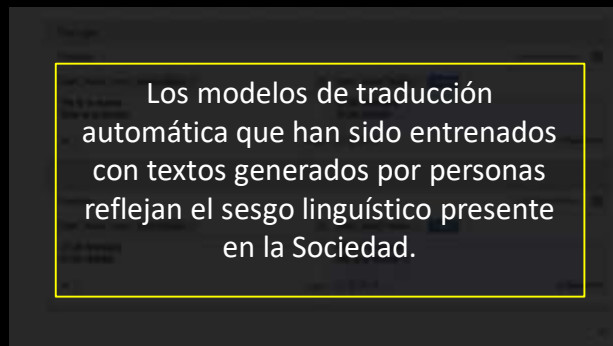
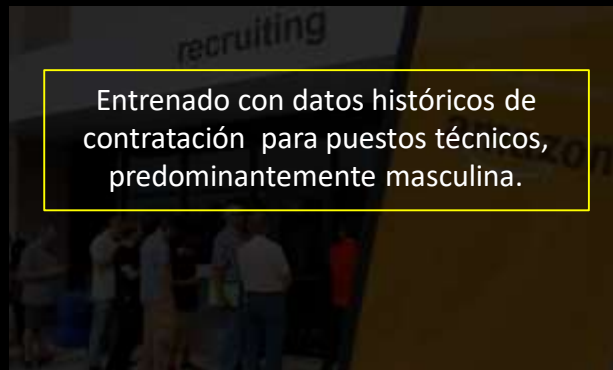
¡Parece muy buena idea!

Construcción del dataset (DATASET CONSTRUCTION)

Fase muy proclive a la introducción de sesgos.

1. Sesgos sociales o históricos (historical bias)

El dataset refleja los sesgos de nuestra sociedad.

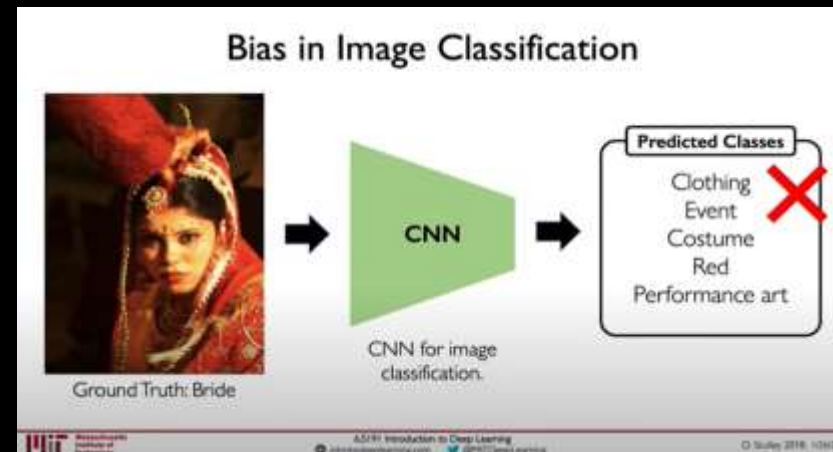
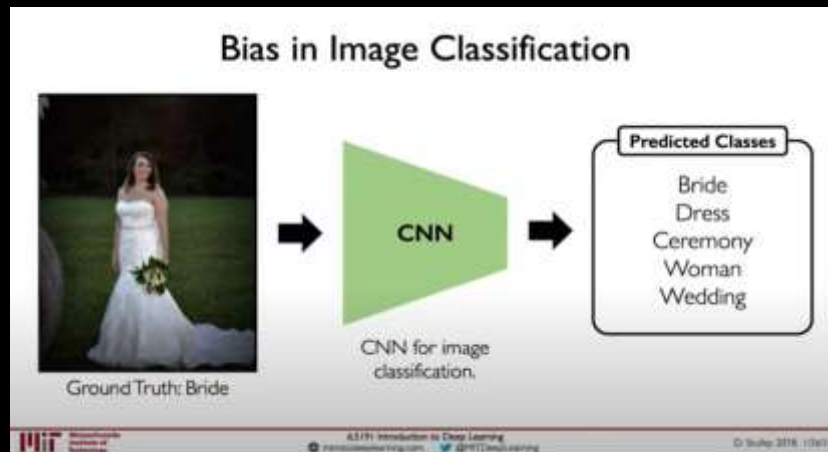
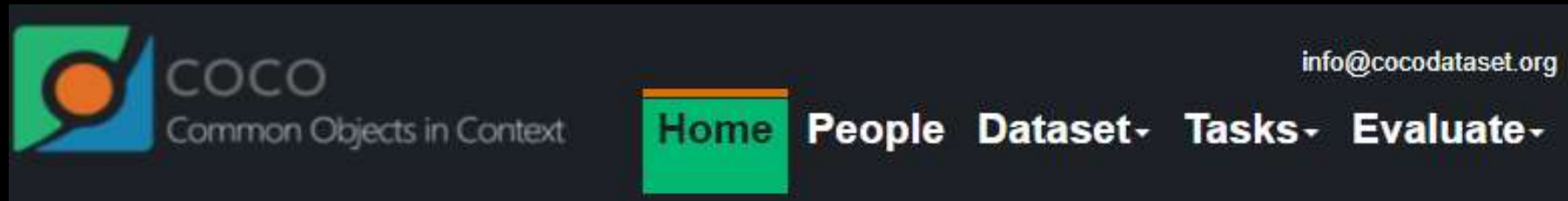


Construcción del dataset (DATASET CONSTRUCTION)

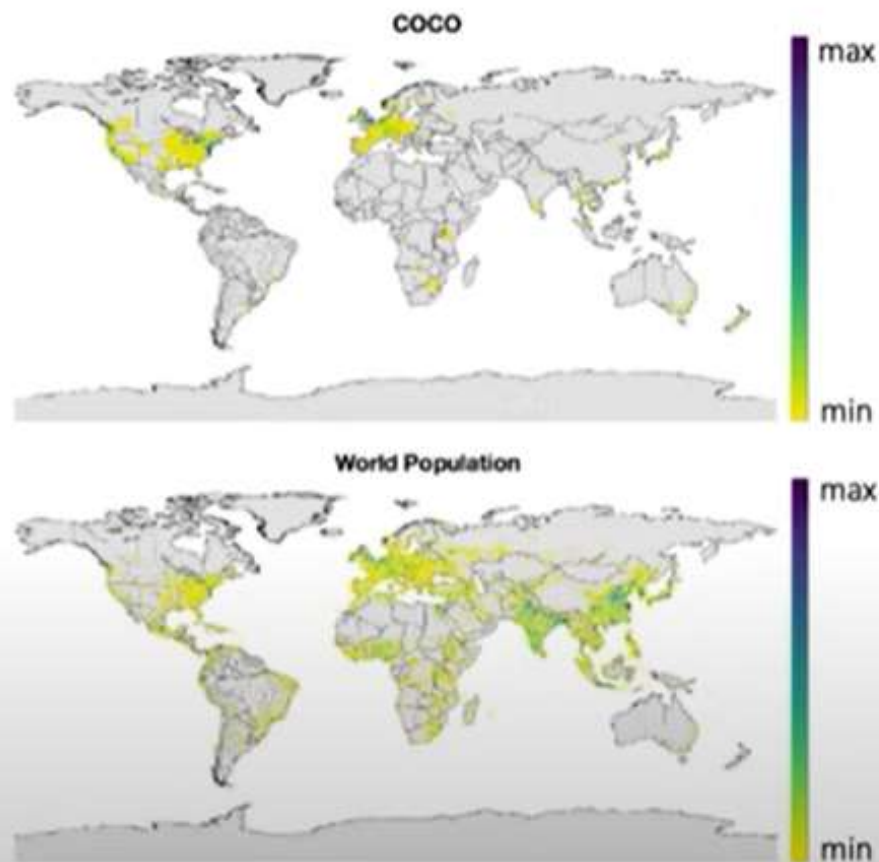
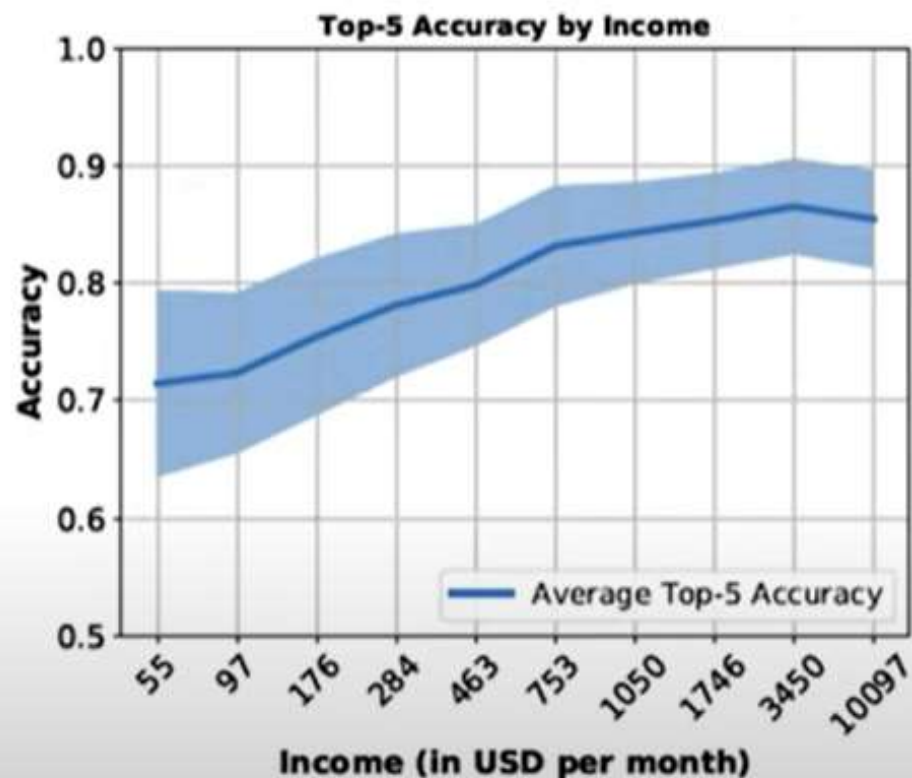
Fase muy proclive a la introducción de sesgos.

2. Sesgos de representación (Representation bias)

COCO is a large-scale object detection, segmentation, and captioning dataset.



Bias Correlation with Income and Geography



Data Cards

DATA COLLECTION METHOD(S)

Crowdsourced

SAMPLING METHOD(S)

Unsampled

LABELING METHOD(S)

Human Labels

Algorithmic Labels

DATA SOURCE(S)

- Contributions by global users of the [Crowdsourcing](#) app
- Vendor data collection efforts

GEOGRAPHIC DISTRIBUTION

83% India
2% Vietnam
2% Brazil
1% Israel
1% Nigeria
1% Thailand
1% Colombia
1% UAE
8% Others (each less than 1%)

LABEL TYPE(S)

Human Labels

Algorithmic Labels

Free-form text labels

Additional labels

DATA SELECTION

All images are opted-in for open-sourcing by Crowdsourcing app contributors

FILTERING CRITERIA

- PII: Name tags, Unblurred faces, etc.
- Inappropriate Content
- Unusable Imagery

LABELING PROCEDURE - HUMAN

Free-form labels are provided by users of the Crowdsourcing app. The user who has taken the picture provides the label.

ML Pipeline

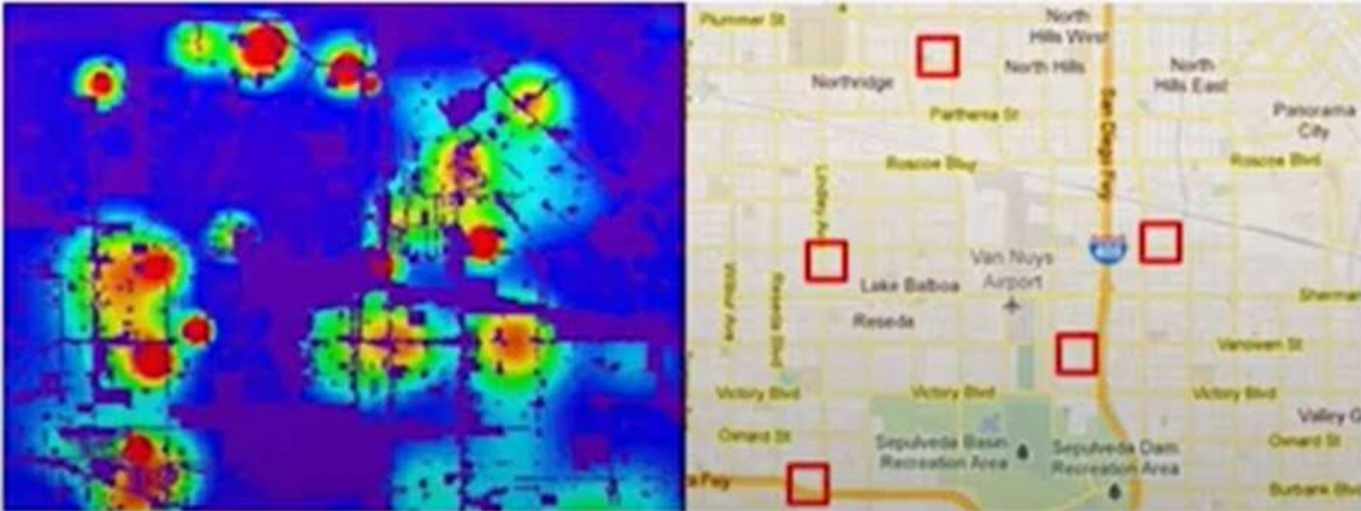


Selección del modelo de ML (Model definition)

Un modelo es necesariamente más simple que la realidad y por tanto requiere realizar **ASUNCIONES**.

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.



Modelos de predicción de criminalidad

Objetivo: predecir dónde se van a cometer los crímenes a partir de datos históricos de arrestos policiales.

ASUNCIÓN IMPLÍCITA:

El número de arrestos policiales que se ha llevado a cabo en un determinado lugar es una variable que predice acertadamente la cantidad de crímenes que se van a producir en ese lugar.

Definición del modelo de ML (Model definition)

Un modelo es necesariamente más simple que la realidad y por tanto requiere realizar **ASUNCIONES**.

Esta asunción no tiene en cuenta que las practicas de detención policial pueden estar sesgadas ni que puede haber un exceso de presencia policial en determinados barrios.

La polémica que rodea las prácticas de "parar y cachear" del Departamento de Policía de Nueva York (EEUU) demuestra por qué. Entre enero de 2004 y junio de 2012, la policía de la ciudad llevó a cabo 4,4 millones de altos bajo un programa que les permitía dar el alto, interrogar y registrar a personas en la calle en busca de armas y otros artículos de contrabando. Pero, de hecho, "el 88% de los 4,4 millones de altos y registros no condujo a nada, lo que significa que una gran mayoría de las personas a las que se les dio el alto no estaba haciendo nada malo", criticaba The New York Times en un editorial. Es más: "En el 83% de los casos, la persona en cuestión era negra o hispana, aunque los dos grupos sólo representaban a poco más de la mitad de la población".

<https://www.technologyreview.es/s/7950/unamonos-para-evitar-la-discriminacion-de-los-algoritmos-que-nos-gobiernan>
2017

Modelos de predicción de criminalidad

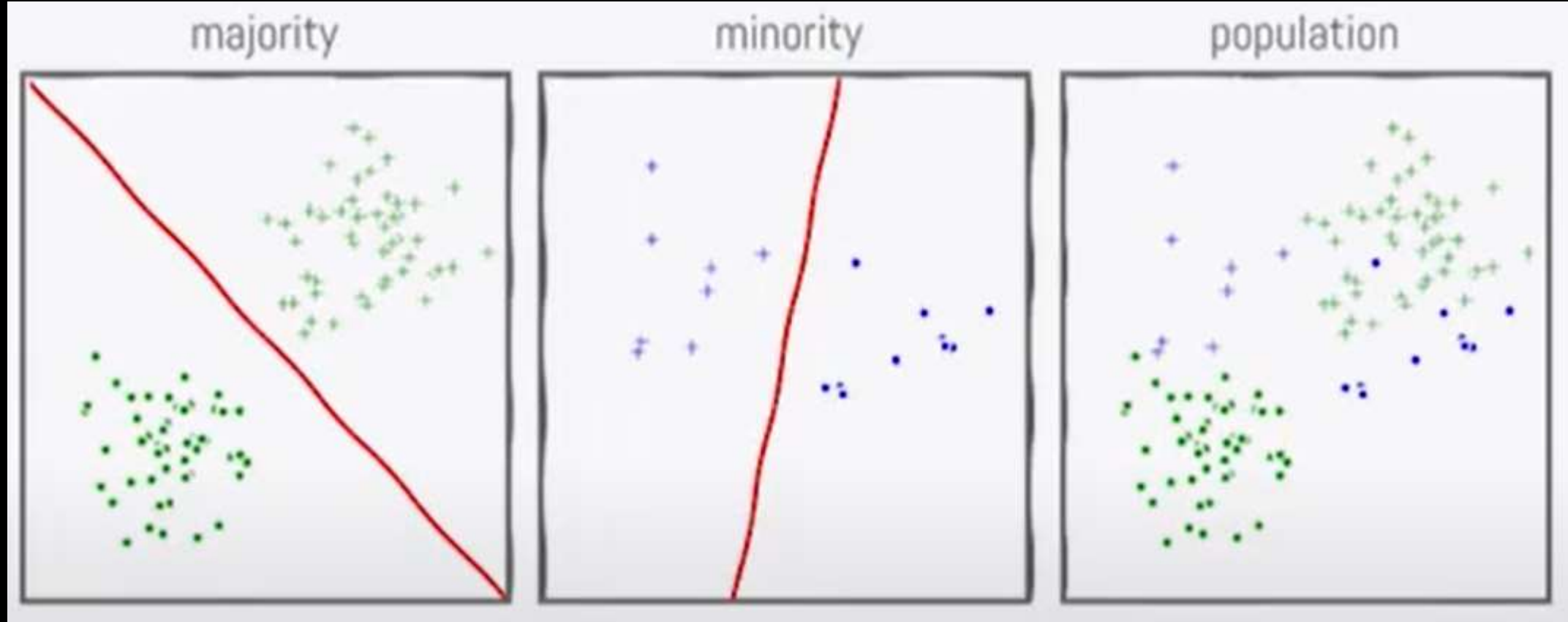
Objetivo: predecir dónde se van a cometer los crímenes a partir de datos históricos de arrestos policiales.

ASUNCIÓN IMPLÍCITA:

El número de arrestos policiales que se ha llevado a cabo en un determinado lugar es una variable que predice acertadamente la cantidad de crímenes que se van a producir en ese lugar.

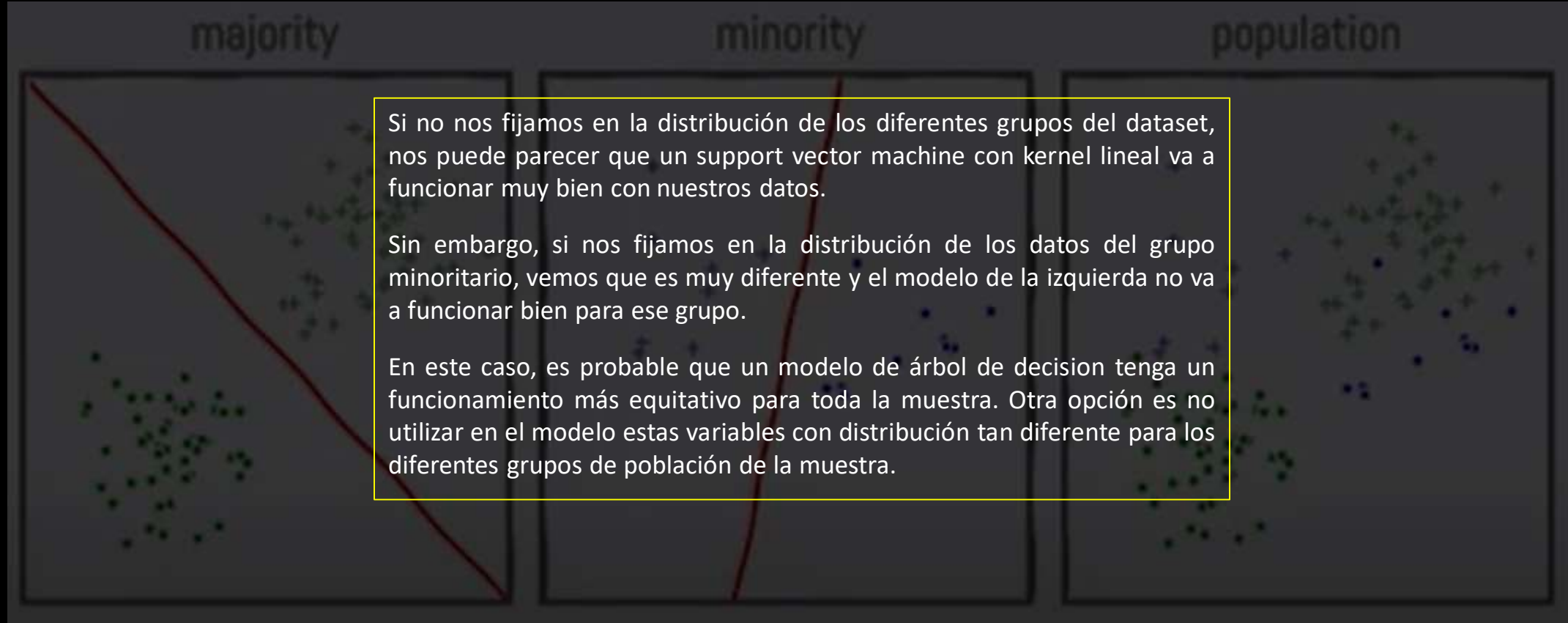
Definición del modelo de ML (Model definition)

Estructura del modelo de ML



Definición del modelo de ML (Model definition)

Asunciones implícitas en la estructura el modelo

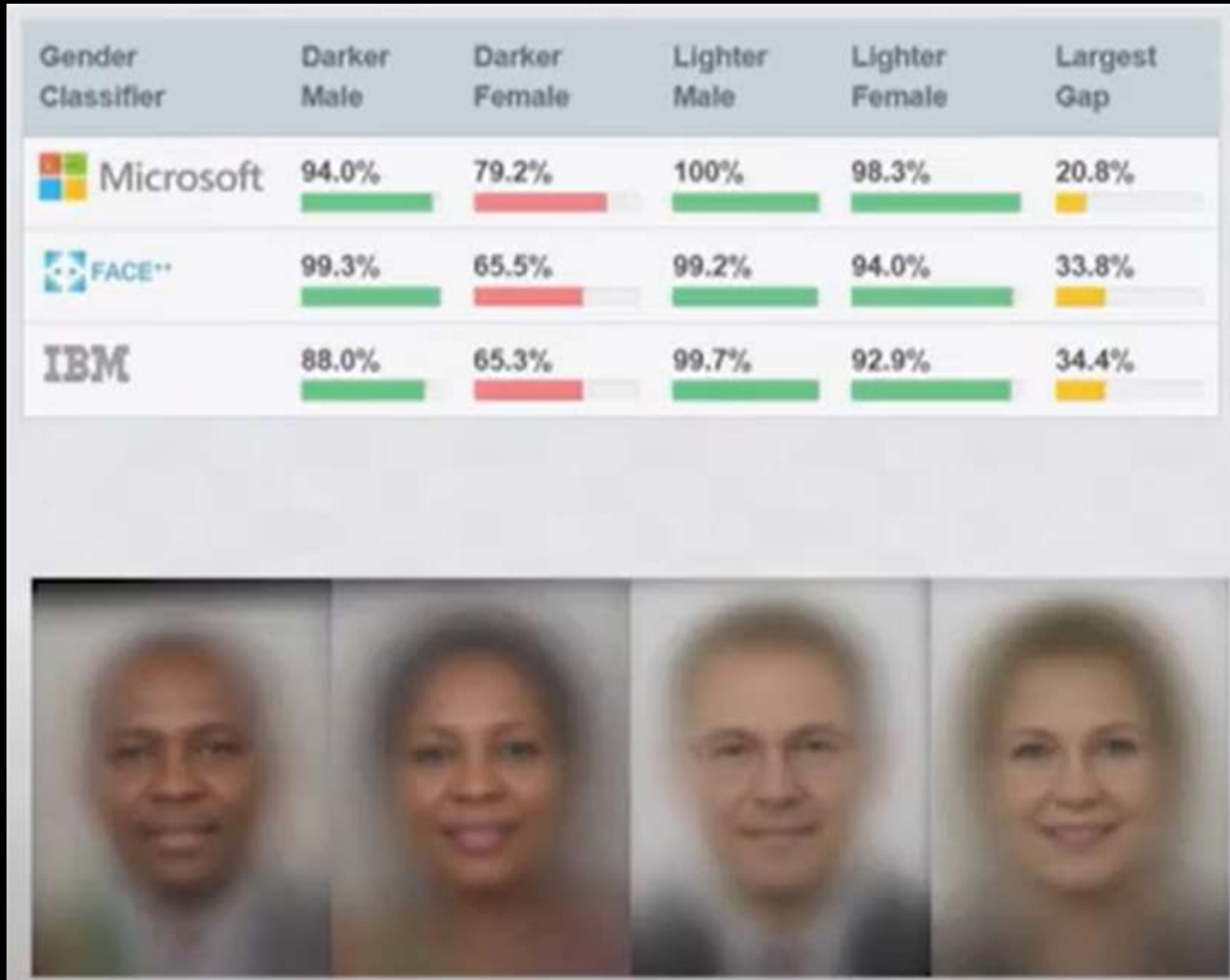


ML Pipeline



Testeo del modelo (Model testing)

Distribución de los datos utilizados para testear el modelo



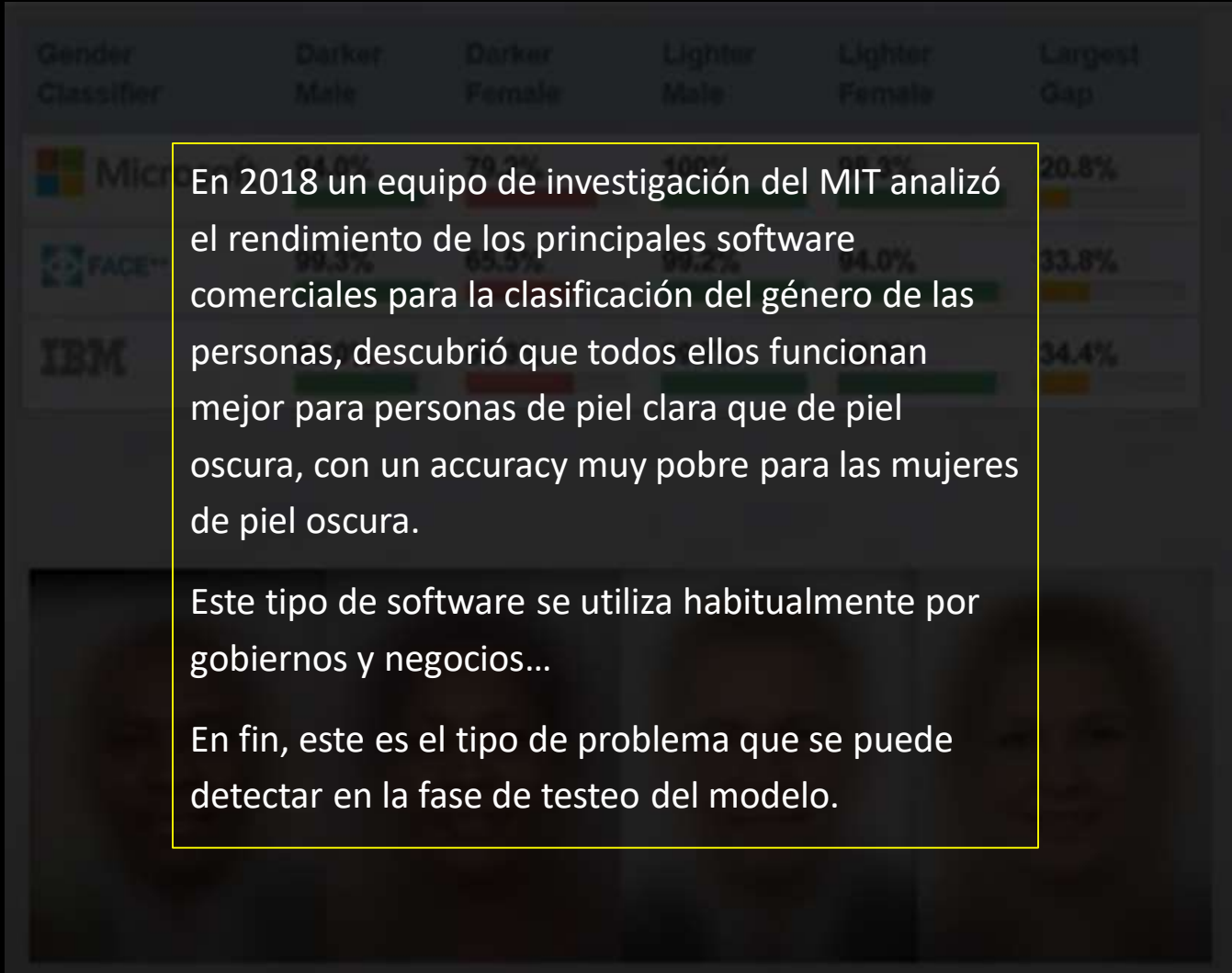
"Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification".

Joy Buolamwini, 2018.

joyab@mit.edu MIT Media Lab

Testeo del modelo (Model testing)

Para que los modelos funcionen bien para todos, deben testarse explícitamente para diferentes segmentos de población.



En 2018 un equipo de investigación del MIT analizó el rendimiento de los principales software comerciales para la clasificación del género de las personas, descubrió que todos ellos funcionan mejor para personas de piel clara que de piel oscura, con un accuracy muy pobre para las mujeres de piel oscura.

Este tipo de software se utiliza habitualmente por gobiernos y negocios...

En fin, este es el tipo de problema que se puede detectar en la fase de testeo del modelo.

"Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification".

Joy Buolamwini, 2018.

joyab@mit.edu MIT Media Lab

Testeo del modelo (Model testing)

MÉTRICAS DE EQUIDAD

- Más allá del accuracy del modelo: Métricas de equidad.
- Métricas de equidad de acierto vs error.
- El caso “Machine Bias” se analiza detenidamente en el taller.



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner,

ProPublica

May 23, 2016

Testeo del modelo (Model testing)

Matriz de confusión (I)

	Actual – Positive	Actual – Negative
Predicted – Positive	True Positive (TP) $PPV = \frac{TP}{TP+FP}$ $TPR = \frac{TP}{TP+FN}$	False Positive (FP) $FDR = \frac{FP}{TP+FP}$ $FPR = \frac{FP}{FP+TN}$
Predicted – Negative	False Negative (FN) $FOR = \frac{FN}{TN+FN}$ $FNR = \frac{FN}{TP+FN}$	True Negative (TN) $NPV = \frac{TN}{TN+FN}$ $TNR = \frac{TN}{TN+FP}$

Testeo del modelo (Model testing)

Matriz de confusión (II)

	Actual – Positive	Actual – Negative
Predicted – Positive	True Positive (TP) $PPV = \frac{TP}{TP+FP}$ $TPR = \frac{TP}{TP+FN}$	False Positive (FP) $FDR = \frac{FP}{TP+FP}$ $FPR = \frac{FP}{FP+TN}$
Predicted – Negative	False Negative (FN) $FOR = \frac{FN}{TN+FN}$ $FNR = \frac{FN}{TP+FN}$	True Negative (TN) $NPV = \frac{TN}{TN+FN}$ $TNR = \frac{TN}{TN+FP}$

Testeo del modelo (Model testing)

- **Positive predicted value (PPV)**: the fraction of positive cases which were correctly predicted out of all the positive predictions. It is usually referred to as **precision**, and represents the **probability** of a correct positive prediction. It is given by the following formula:

$$PPV = P(actual = + | prediction = +) = \frac{TP}{TP + FP}$$

- **False discovery rate (FDR)**: the fraction of positive predictions which were actually negative out of all the positive predictions. It represents the **probability** of an erroneous positive prediction, and it is given by the following formula:

$$FDR = P(actual = - | prediction = +) = \frac{FP}{TP + FP}$$

- **Negative predicted value (NPV)**: the fraction of negative cases which were correctly predicted out of all the negative predictions. It represents the **probability** of a correct negative prediction, and it is given by the following formula:

$$NPV = P(actual = - | prediction = -) = \frac{TN}{TN + FN}$$

- **False omission rate (FOR)**: the fraction of negative predictions which were actually positive out of all the negative predictions. It represents the **probability** of an erroneous negative prediction, and it is given by the following formula:

$$FOR = P(actual = + | prediction = -) = \frac{FN}{TN + FN}$$

- **True positive rate (TPR)**: the fraction of positive cases which were correctly predicted out of all the positive cases. It is usually referred to as sensitivity or recall, and it represents the **probability** of the positive subjects to be classified correctly as such. It is given by the following formula:

$$TPR = P(prediction = + | actual = +) = \frac{TP}{TP + FN}$$

- **False negative rate (FNR)**: the fraction of positive cases which were incorrectly predicted to be negative out of all the positive cases. It represents the **probability** of the positive subjects to be classified incorrectly as negative ones, and it is given by the formula:

$$FNR = P(prediction = - | actual = +) = \frac{FN}{TP + FN}$$

- **True negative rate (TNR)**: the fraction of negative cases which were correctly predicted out of all the negative cases. It represents the **probability** of the negative subjects to be classified correctly as such, and it is given by the formula:

$$TNR = P(prediction = - | actual = -) = \frac{TN}{TN + FP}$$

- **False positive rate (FPR)**: the fraction of negative cases which were incorrectly predicted to be positive out of all the negative cases. It represents the **probability** of the negative subjects to be classified incorrectly as positive ones, and it is given by the formula:

$$FPR = P(prediction = + | actual = -) = \frac{FP}{TN + FP}$$

Testeo del modelo (Model testing)

MÉTRICAS DE EQUIDAD

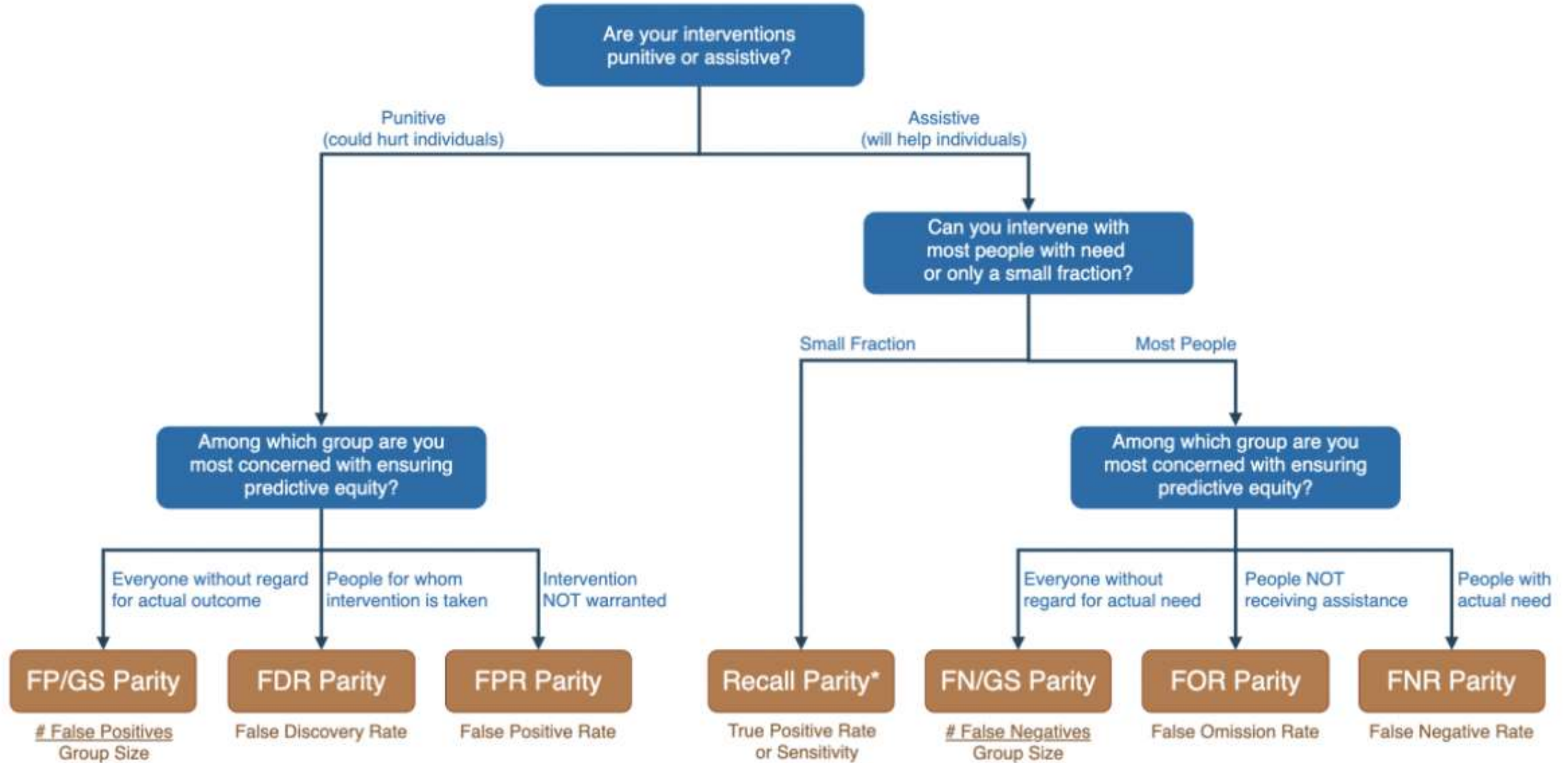
La idea detrás de las métricas de equidad es verificar que las métricas de rendimiento del modelo arrojan resultados similares para los diferentes grupos de población que componen la muestra.

Las más utilizadas (pero hay muchas más):

- ***Predictive parity***, también denominado ***outcome test***. Un clasificador satisface esta definición si los sujetos de los grupos protegidos y no protegidos presentan el mismo Positive Predictive Value (PPV o precisión), es decir, que tienen la misma probabilidad de ser clasificados correctamente como positivos.
- ***False positive error rate balance***, también denominada ***predictive equality***. Un clasificador satisface esta definición si los sujetos de los grupos protegidos y no protegidos presentan el mismo False Positive Rate (FPR), es decir, que los sujetos negativos de todos los grupos presentan la misma probabilidad de ser clasificados incorrectamente como positivos.
- ***False negative error rate balance***, también denominado ***equal opportunity***. Un clasificador satisface esta definición si los sujetos de los grupos protegidos y no protegidos presentan el mismo False Negative Rate (FNR), es decir, que los sujetos positivos de todos los grupos presentan la misma probabilidad de ser clasificados incorrectamente como negativos.

FAIRNESS TREE

(Zoomed in)



Testeo del modelo (Model testing)

MÉTRICAS DE EQUIDAD

Pero...hay un problema...

Para poder evaluar si los modelos están o no sesgados necesitamos MEDIR el rendimiento del modelo en relación con los diferentes grupos de población a los que aplica y, consecuentemente, su impacto sobre los grupos vulnerables.

Sin embargo, muchas veces los desarrolladores no tienen acceso a las variables protegidas.

Testeo del modelo (Model testing)

METRICAS DE EQUIDAD

SOLUCIONES

- Recabar esta información únicamente a los efectos de auditar los modelos.
- Usar machine learning para inferir los valores de las variables protegidas.

Muy problemático, información muy sensible que plantea cuestiones de privacidad (GDPR). Los usuarios se pueden oponer a facilitarla.

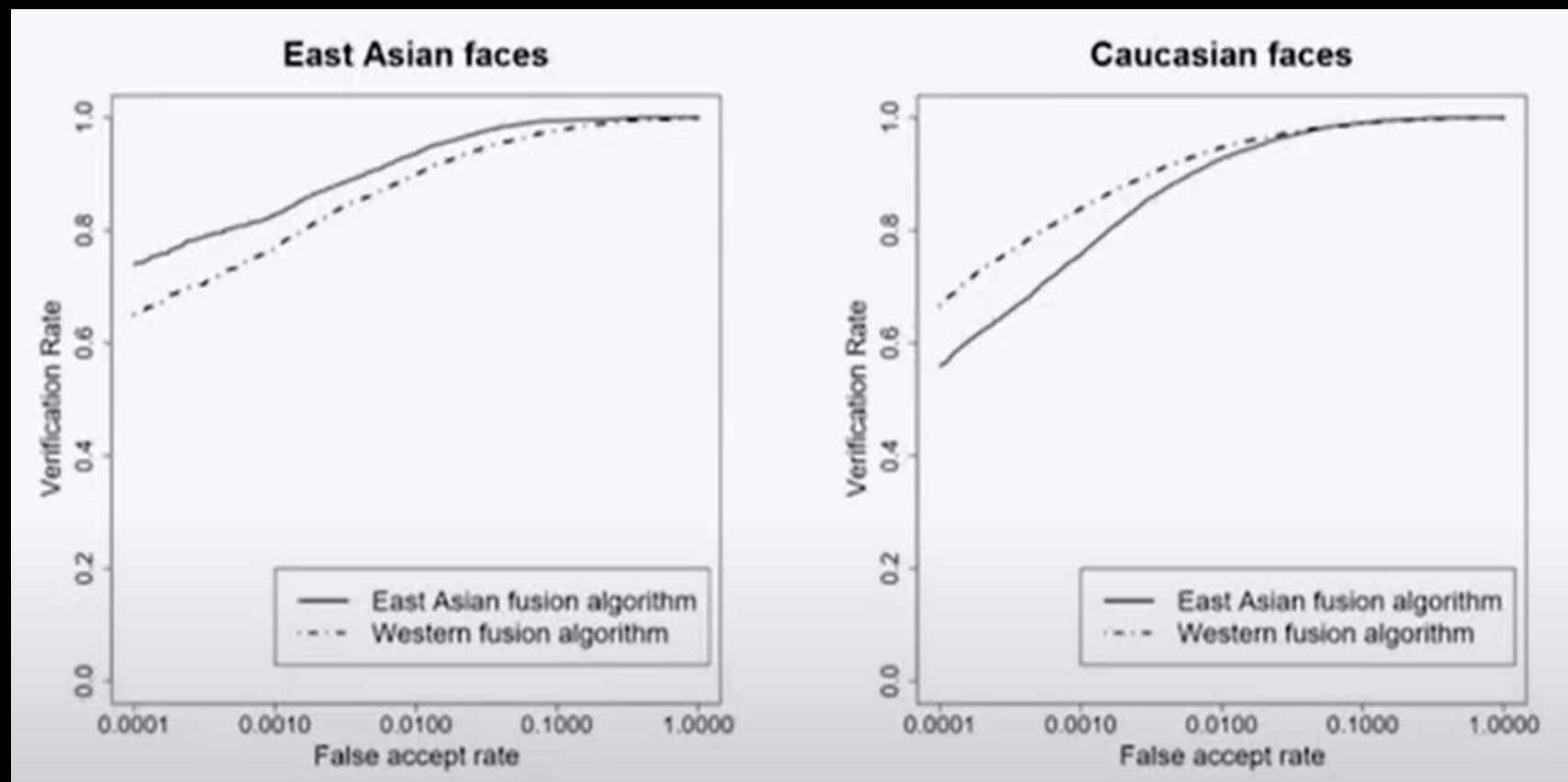
También problemático, puede introducir nuevos sesgos en el modelo. cambiando un problema por otro. Además, los usuarios pueden oponerse.

ML Pipeline



Puesta en producción del modelo (Deployment process)

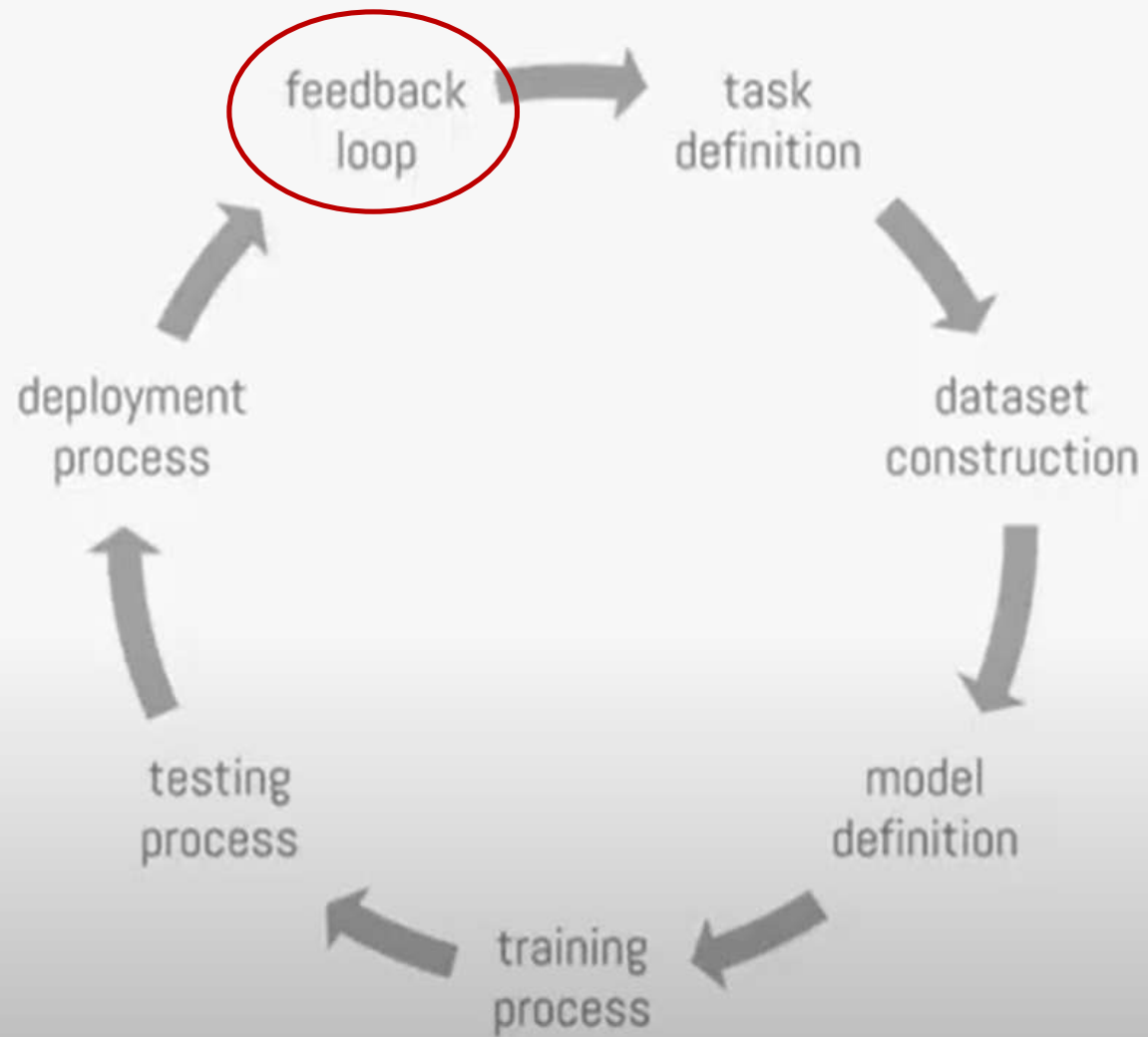
Los usuarios del modelo difieren de la tipología de la población para la que en un principio se definió el modelo o la población con la que se ha entrenado y/o testeado.



Investigación del año 2011

Los modelos de reconocimiento facial tienen un rendimiento significativamente inferior sobre la población de zonas geográficas diferentes de aquella donde se desarrolló el modelo que sobre la población de la misma zona geográfica.

ML Pipeline

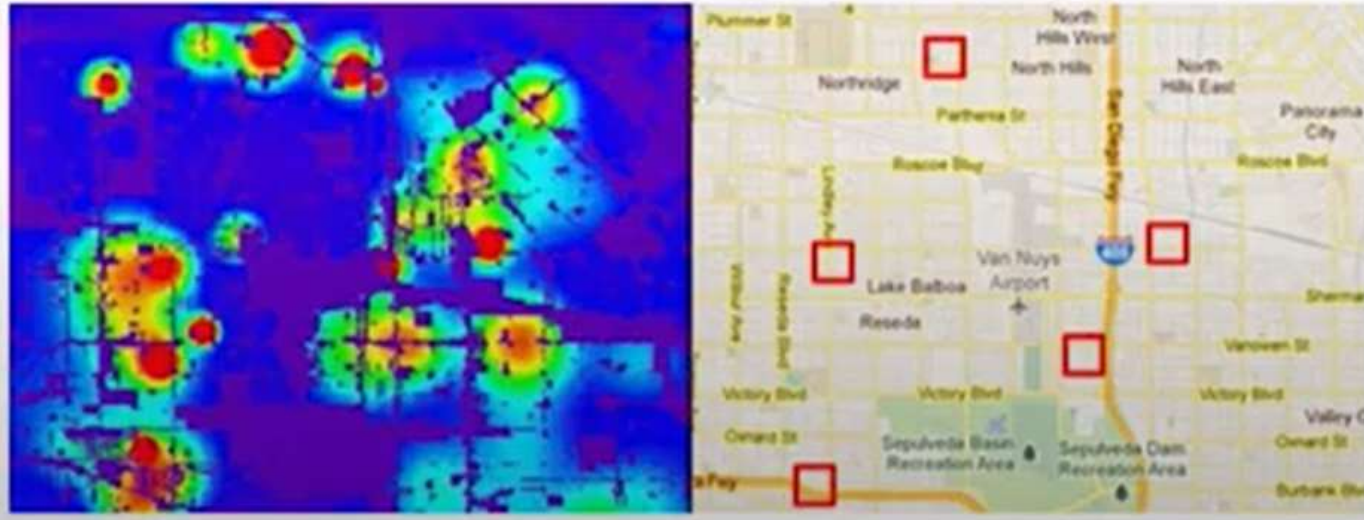


Retroalimentación del modelo (Feedback loop)

Los modelos de machine learning pueden generar una profecía autocumplida...

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.



Modelos de predicción de criminalidad

Retroalimentación del modelo (Feedback loop)

Los modelos de machine learning pueden generar una profecía autocumplida...

Como resultado del modelo, la jefatura de policía incrementará la cantidad de policías destinados a determinados barrios. La mayor presencia policial conducirá de forma natural a la detección de un mayor número de ilícitos y, por tanto, a un mayor número de detenciones realizadas en ese barrio con respecto a los demás barrios.

Todo ello llevará al modelo a predecir un mayor ratio de delitos en ese barrio y todavía más policías serán asignados al mismo...

A sensu contrario, menos policías serán destinados a los barrios menos conflictivos según el modelo, lo cual reducirá las tasas de delitos que la policía consigue detectar en esos barrios, lo que llevará al modelo a predecir un número todavía menor de delitos en esos barrios...

CONSTRUYENDO MODELOS EQUITATIVOS



LA OPORTUNIDAD...

“Los procesos de Big Data codifican el pasado. No inventan el futuro. Tenemos que incorporar explícitamente mejores valores en nuestros algoritmos, creando modelos de big data que sigan nuestro liderazgo ético. A veces eso significa poner la justicia por delante de las ganancias”.

Cathy O'Neil.

Catherine ("Cathy") Helen O'Neil es una [matemática estadounidense](#) y autora del blog [mathbabe.org](#) y varios libros sobre [ciencia de datos](#), entre los que se incluye [Armas de destrucción matemática](#).

https://es.wikipedia.org/wiki/Cathy_O%27Neil

Medidas para mitigar los sesgos

Al definir el problema

- **Piensa** bien si el problema que quieres resolver debe afrontarse con *machine learning*.
- **Define** claramente el problema y el resultado que buscas.
- **Investiga e identifica** los efectos indeseados y sesgos que ese tipo de modelo ha presentado antes.
- **Involucra** personas de muy diversos perfiles en el equipo, busca perspectivas diferentes.
- **Documenta** detalladamente los efectos indeseados y sesgos que preveas que puede presentar el futuro modelo.
- **Redefine** el problema y **abandona** el proyecto si los errores del modelo son demasiado costosos.

Al construir el dataset

- **Piensa** antes de recolectar los datos.
- **Identifica** los sesgos sociales presentes en tu fuente de datos.
- **Verifica** que tu fuente de datos es consistente con el entorno en que vas a implementar el modelo.
- **Identifica** los sesgos que puedan presentar las personas que recolectan los datos y la tecnología que usan para ello.
- **Asegúrate** de que los grupos protegidos tienen suficiente representación en el dataset.
- **Asegúrate** de que las personas que etiquetan el dataset no están introduciendo sesgos en el dataset.

Al seleccionar el modelo de ML

- **Define** claramente las asunciones presentes en el modelo.
- **Identifica** los sesgos sociales presentes en las asunciones.
- **Asegúrate** de que la estructura del modelo elegido no esté introduciendo sesgos.

En las fases de testeo, puesta en producción y en el feedback loop

- **Testea** explícitamente el modelo con diferentes segmentos de población para asegurarte de que funciona bien para todos.
- **Verifica** que el contexto de tus datos coincide con el contexto en que se va a implementar el modelo.
- **Utiliza** métricas de equidad para medir la equidad del modelo tanto en fase de testeo como una vez puesto en producción.
- **Define** qué métricas de equidad vas a optimizar para satisfacer el principio de equidad en tu caso concreto.
- **Involucra** personas de diferentes perfiles y perspectivas tanto en el testeo del modelo como en la auditoría del mismo una vez puesto en producción.
- **Monitoriza** el feedback de los usuarios en su interacción con el sistema, en particular las quejas de los usuarios.

Soluciones en el mercado

Aequitas

An open source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions

You can audit your risk assessment system for two types of biases:

- Biased actions or interventions that are not allocated in a way that's representative of the population.
- Biased outcomes through actions or interventions that are a result of your system being wrong about certain groups of people.

<http://www.datasciencepublicpolicy.org/projects/aequitas/>

Bias and Fairness Audit Report

Generated by Aequitas for [Large US City] Criminal Justice Project
January 29, 2018

Project Goal: Identify individuals likely to get booked/charged by police in the near future

Performance Metric: Accuracy (Precision) in the top 150 identified individuals

Bias Metrics Considered: Demographic Disparity, Impact Disparity, FPR Disparity, FNR Disparity, FOR Disparity, FDR Disparity




Reference Groups: Race/Ethnicity – White, Gender: Male, Age: None

Model Audited: #841 (Random Forest)

Model Performance: 73%



Aequitas has found that Model 841 is **BIASED**. The Bias is in the following attributes:

Group Variable	Group Value	Group Size	
gender	female	229	
	male	1,414	
marital_status	divorced	29	
	married	639	
	separated	9	
	single	823	
	unknown	142	
race	black	288	
	other	12	
	pacific_islander	36	
	unknown	65	
	white	1,235	

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 ↗](#)

Not sure what to do first? Start here!

Read More

Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

Try a Web Demo

Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

Watch Videos

Watch videos to learn more about AI Fairness 360.

Read a paper

Read a paper describing how we designed AI Fairness 360.

Use Tutorials

Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application

Ask a Question

Join our AIF360 Slack Channel to ask questions, make comments and tell stories about how you use the toolkit.

AI Fairness 360 (IBM)

Librería open-source que permite a los programadores de IA:

- Detectar y evaluar sesgos en modelos y conjuntos de datos con un conjunto completo de métricas.
- Mitigar los sesgos con la ayuda de 12 algoritmos empaquetados como Learning Fair Representations, Reject Option Classification, Disparate Impact Remover.

These are ten state-of-the-art bias mitigation algorithms that can address bias throughout AI systems. Add more!

Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.



Reweighting

Use to mitigate bias in training data. Modifies the weights of different training examples.



Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.



Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.



Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.



Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.



Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing

Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

Meta Fair Classifier

Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

<http://aif360.mybluemix.net/>

Otras herramientas

Watson OpenScale de IBM

Realiza la verificación y mitigación de sesgos en tiempo real cuando la IA toma sus decisiones.

What-If de Google

Con la herramienta What-If, puede probar el rendimiento en situaciones hipotéticas, analizar la importancia de diferentes características de datos y visualizar el comportamiento del modelo en múltiples modelos y subconjuntos de datos de entrada, y para diferentes métricas de equidad de ML.

Marcos de trabajo (Frameworks)

Tackling bias in artificial intelligence (and in humans)

<https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>

Minimizing bias will be critical if artificial intelligence is to reach its potential and increase people's trust in the systems.

Six potential ways forward for artificial-intelligence (AI) practitioners and business and policy leaders to consider



McKinsey
& Company



“Think about how your actions will affect people and do your best to give those you’ll affect a voice to guide you through your blind spots.”

[Cassie Kozyrkov](#), Chief Decision Scientist at Google.

¿Regulación?



- 8/04/2019 - Ethics guidelines for trustworthy AI
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- 19/02/2020 - Whitepaper on artificial intelligence. An European approach to excellence and trust.
https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
- **21/04/2021 - Proposal for a regulation of the European Parliament and of the Council: Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union. (AIA)**
<https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206>

<https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=celex:52021PC0206>

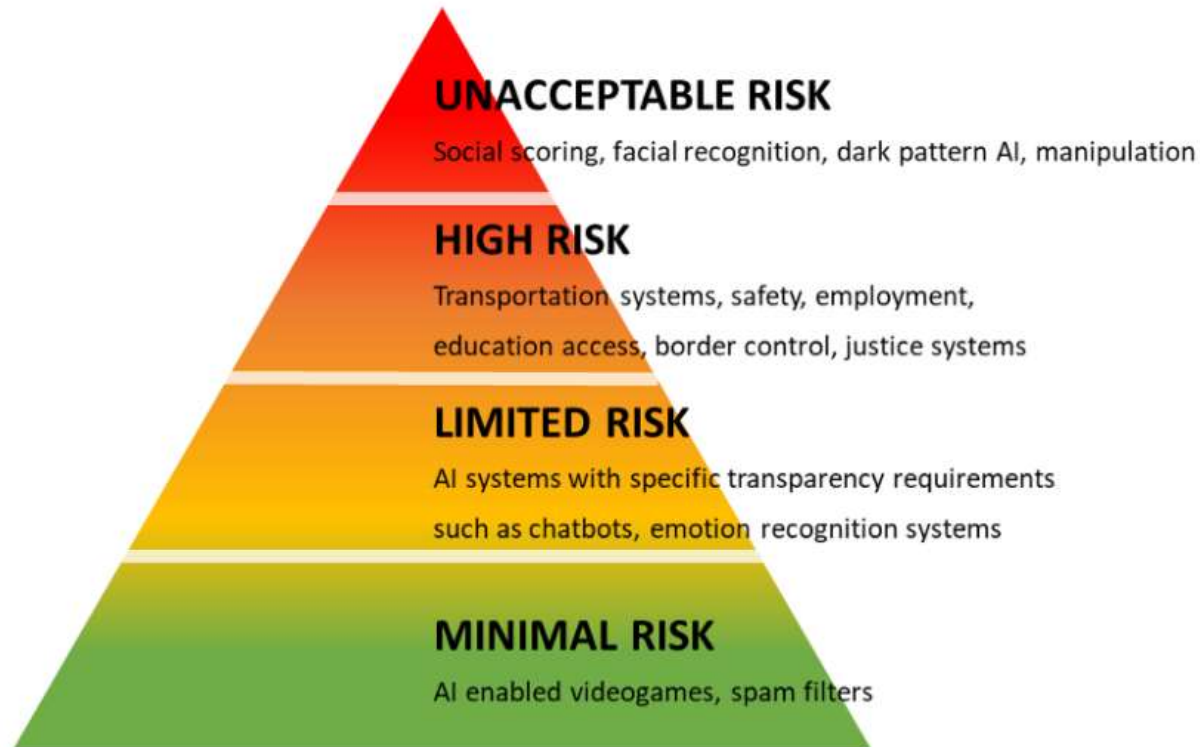
The AIA is currently going through a detailed legislative process which involve negotiations between the co-legislators, the European Parliament and the Council of the European Union ("Council"). The status in January 2023 is awaiting the European Economic and Social Committee decision.

[https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2021/0106\(OLP\)](https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2021/0106(OLP))

The Parliament is scheduled to vote on the draft AI Act by end of March 2023. Following this vote, discussions between the Member States, the Parliament and the Commission are expected to commence in April. If this timeline is met, the final AI Act should be adopted by the end of

EU COMMISSION. Proposal for AI regulation

La Comisión propone un enfoque basado en el riesgo, con cuatro niveles de riesgo:



Source: [Laying Down Harmonized Rules on Artificial Intelligence \(Artificial Intelligence Act\) and Amending Certain Union Legislative Acts](#)

Following the so-called 'risk pyramid', most uses of AI are expected to have no or low risks, while high-risk or harmful applications are considered to be less numerous but will be strictly regulated.

The draft regulation includes a series of obligations for AI applications that are considered to be 'risky', as they could have a direct impact on someone's personal or professional life (e.g. mortgage risk assessment or recruitment process). In these cases, the organisation using AI would need to ensure high quality of data, a detailed documentation that proves compliance with the existing regulations, transparency and human oversight on the process, as well as an high level of accuracy and cybersecurity.

The Commission also proposes to prohibit certain AI uses that are considered as incompatible with EU values. These include AI systems aimed at procuring harm or manipulating human behaviour as they are considered an 'unacceptable risk'. Similarly, 'Chinese-style' scoring systems are also banned, a senior Commission official told reporters.



- Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.
<https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673>
- 2/12/2020, Estrategia Nacional de Inteligencia Artificial, enumera una serie de acciones a desarrollar en los próximos años (plan de acción).
<https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/ENIA2B.pdf>
- **La Ley 15/2022, de 12 de julio (LA LEY 15917/2022), integral para la igualdad de trato y la no discriminación (BOE de 13 de julio)**, en su artículo 23, contiene la primera regulación positiva del uso de la inteligencia artificial por las administraciones públicas y las empresas en nuestro país.

<https://www.boe.es/buscar/doc.php?id=BOE-A-2022-11589>

La norma diseña unas grandes líneas de actuación de las administraciones públicas, con el objetivo de “favorecer”, “promover” y “priorizar” determinadas políticas y prácticas relacionadas con el uso de “algoritmos involucrados en la toma de decisiones”, ello “si es posible técnicamente”... (sic)

Artículo interesante sobre el tema:

<https://diariolaley.laleynext.es/dll/2022/07/13/la-ley-15-2022-introduce-la-primera-regulacion-positiva-de-la-inteligencia-artificial-en-espana>



¡GRACIAS!

Bibliografía

Europa y la IA

<https://carnegieendowment.org/2020/07/09/europe-and-ai-leading-lagging-behind-or-carving-its-own-way-pub-82236>

Pay attention to that man behind the curtain. AI bias and what you can do about it.

[Cassie Kozyrkov. Towards Data Science, 25 de enero de 2019.](#)

<https://medium.com/hackernoon/ai-bias-and-what-you-can-do-about-it-4a6ae48b338e>

What is bias?

[Cassie Kozyrkov. Towards Data Science, 24 de enero de 2019.](#)

<https://towardsdatascience.com/what-is-ai-bias-6606a3bcb814>

FATE Microsoft Research.

Machine Learning and Fairness Webinar

<https://www.youtube.com/watch?v=7CH0xLWQLRw>

Machine Learning Fairness: Lessons Learned (Google I/O'19)

<https://www.youtube.com/watch?v=6CwzDoE8J4M>

Machine Learning, Ethics, and Fairness

Dr. Solon Barocas in conversation with Prof. Foster Provost

Monday, April 15, 2019

<https://www.stern.nyu.edu/experience-stern/about/departments-centers-initiatives/centers-of-research/fubon-center-technology-business-and-innovation/initiatives/data-analytics/algorithmic-fairness-in-business>

Ethics of Machine Learning (Day 1)

https://www.youtube.com/watch?v=tdG_BtOCD4c

Ethics of Machine Learning (Day 2)

<https://www.youtube.com/watch?v=NSyfV76L9Tg>