

Spain AI

Descubriendo la IA en **Burgos**

22 de Diciembre de 2022



Inteligencia Artificial y Justicia Algorítmica

Adrián Arnaiz Rodríguez

Estudiante de Doctorado ELLIS
en Transparencia y Justicia en Toma
de Decisiones Algorítmicas



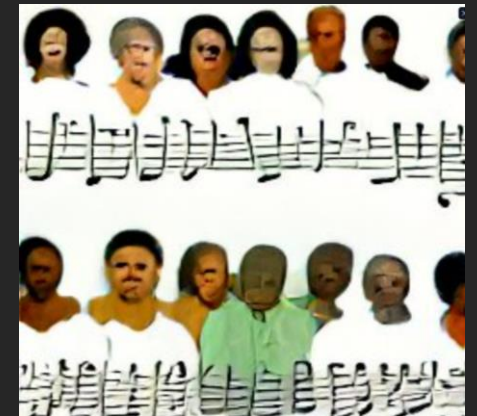
Directores de Tesis: Nuria Oliver (ELLIS) Francisco Escolano (UA) Manuel Gómez Rodríguez (MPI-SWS)

¿Dónde hay discriminación y diversidad?

Decisiones



En las dinámicas sociales



No somos perfectos

¿Por qué utilizar tecnología?



Decisiones no óptimas



Peticiones de asilo de refugiados

"Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires"
Daniel Chen, Tobias J. Moskowitz, Kelly Shue. 2016.

Decisiones no óptimas

Peticiones de asilo de refugiados



"Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires"
Daniel Chen, Tobias J. Moskowitz, Kelly Shue. 2016.

Decisiones no óptimas

Peticiones de asilo de refugiados



"Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires"
Daniel Chen, Tobias J. Moskowitz, Kelly Shue. 2016.

Decisiones no óptimas

Peticiones de asilo de refugiados

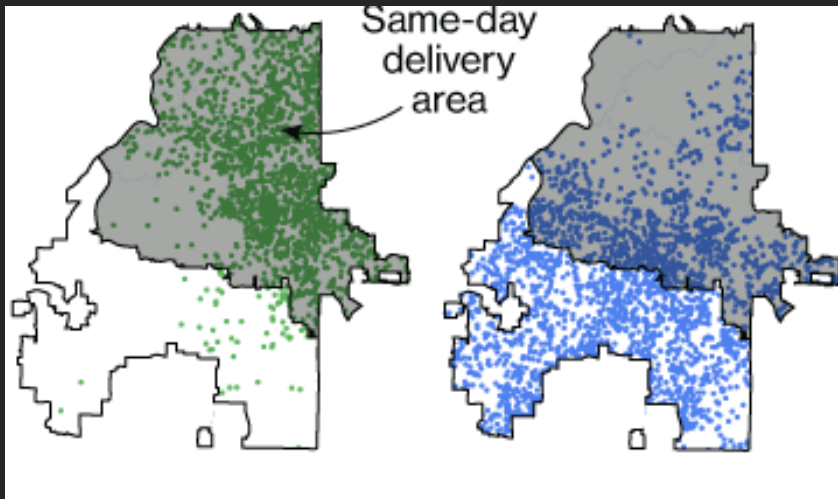


"Decision-Making Under the Gambler's Fallacy: Evidence from Asylum Judges, Loan Officers, and Baseball Umpires"
Daniel Chen, Tobias J. Moskowitz, Kelly Shue. 2016.

Decisiones sesgadas no intencionadas

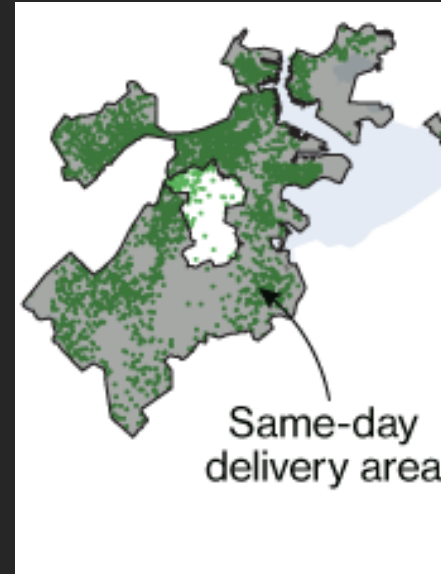
Atlanta

Residentes
blancos

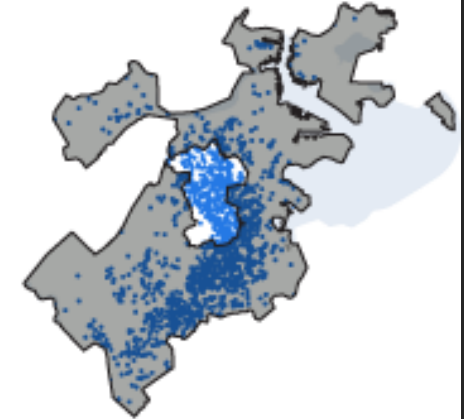


Chicago

Residentes
blancos



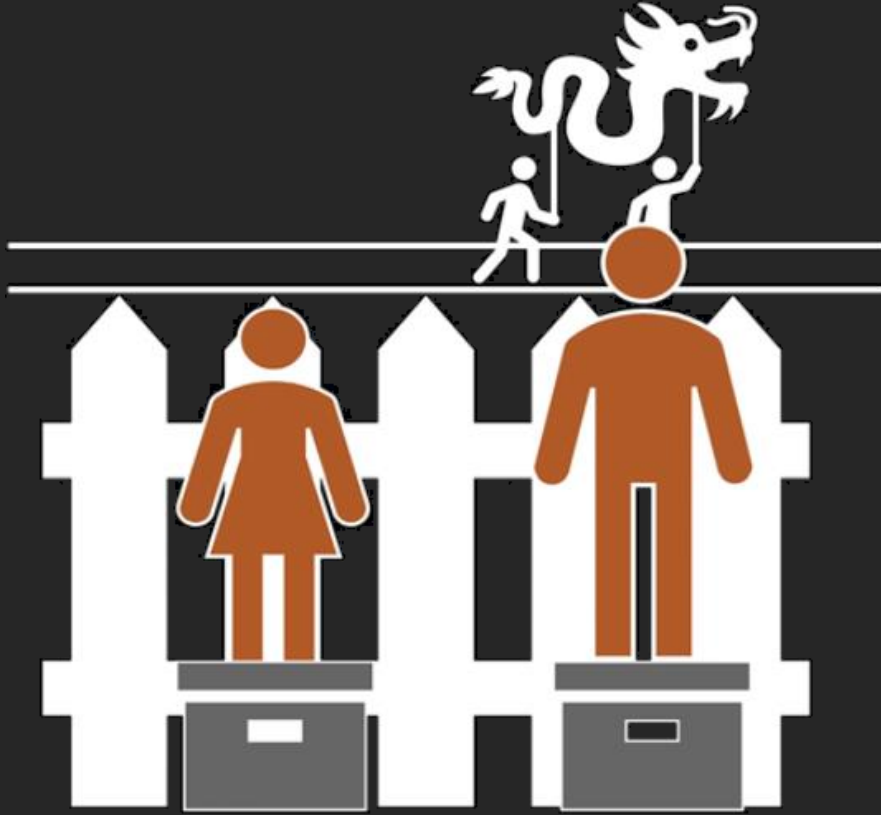
Residentes
negros



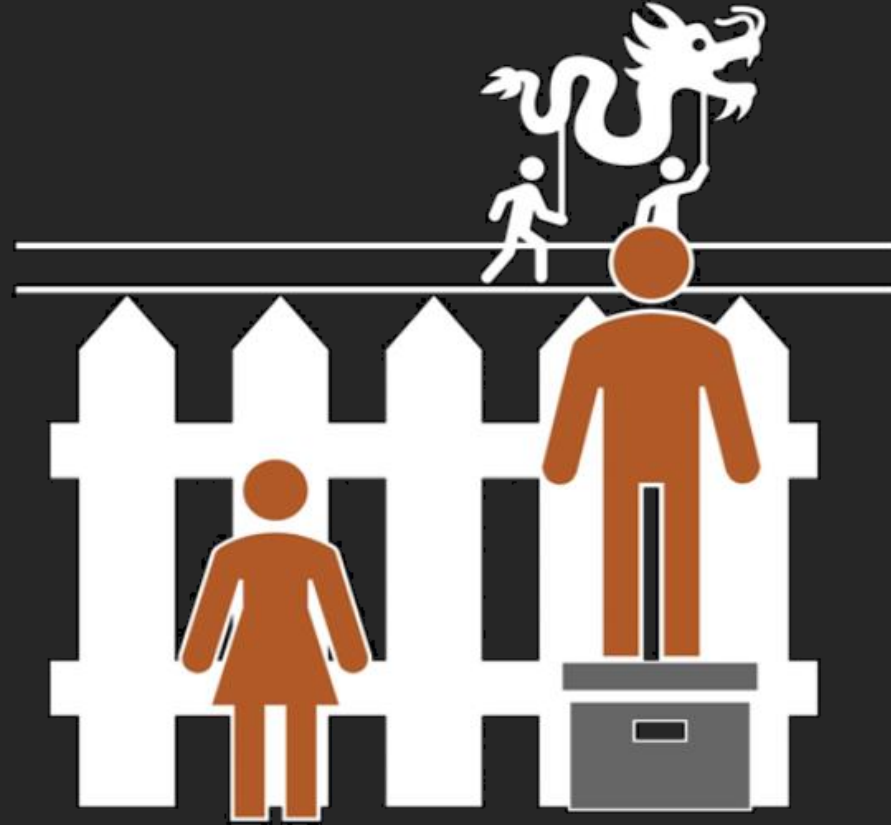
"Amazon Doesn't Consider the Race of Its Customers. Should It?"

David Ingold, Spencer Soper. 2016

Tratamiento dispar VS Impacto dispar



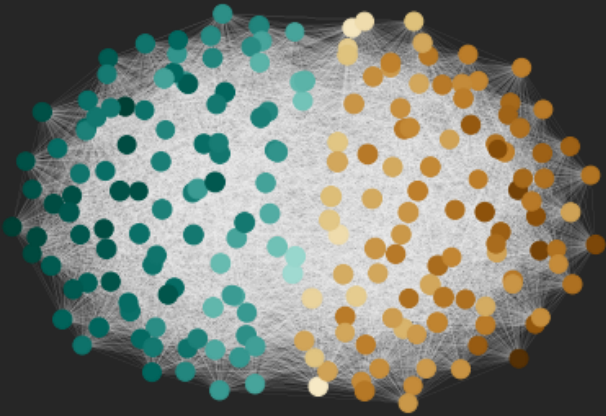
Disparate Impact



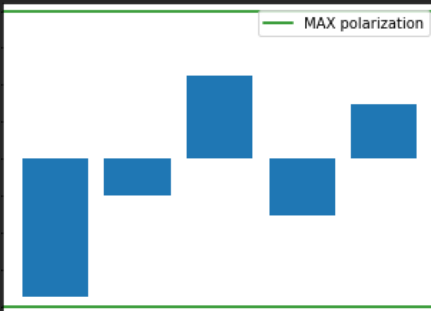
Disparate Treatment

Polarización cuando la mezcla es diversidad

Cámara de Representantes de los Estados Unidos

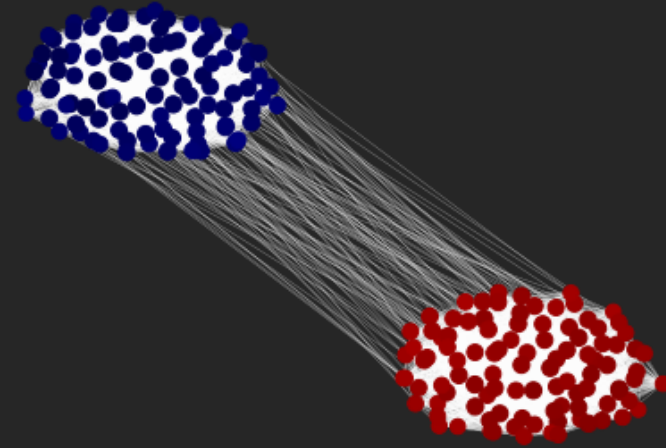
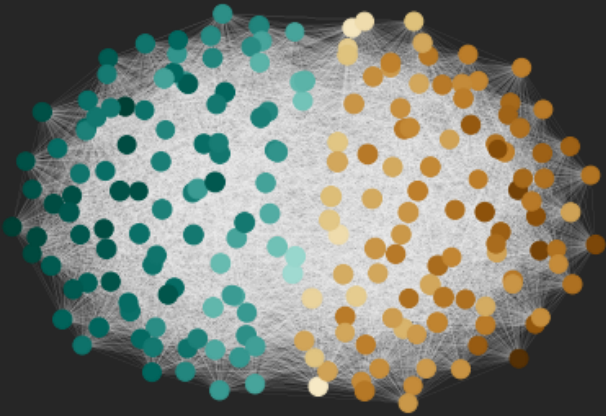


1973

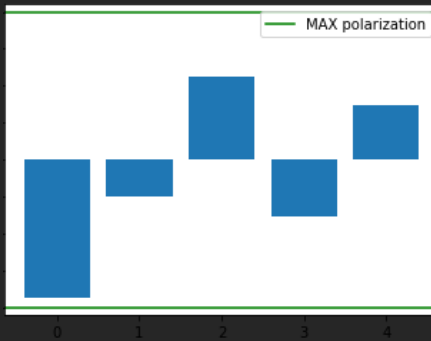


Polarización cuando la mezcla es diversidad

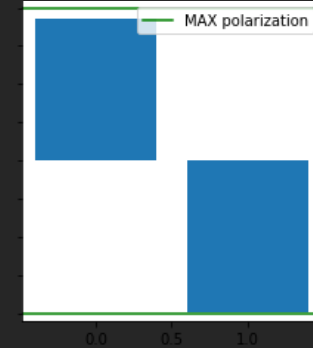
Cámara de Representantes de los Estados Unidos



1973



2016



**La tecnología por sí
misma es óptima**

~~La tecnología por sí
misma es óptima~~

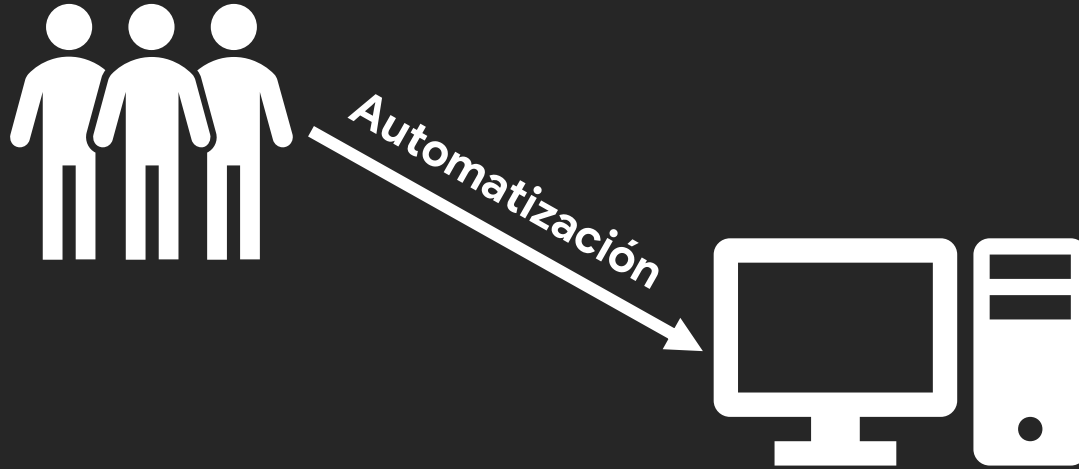
~~La tecnología por sí
misma es óptima~~

Usar la tecnología como un
oráculo tampoco lo es

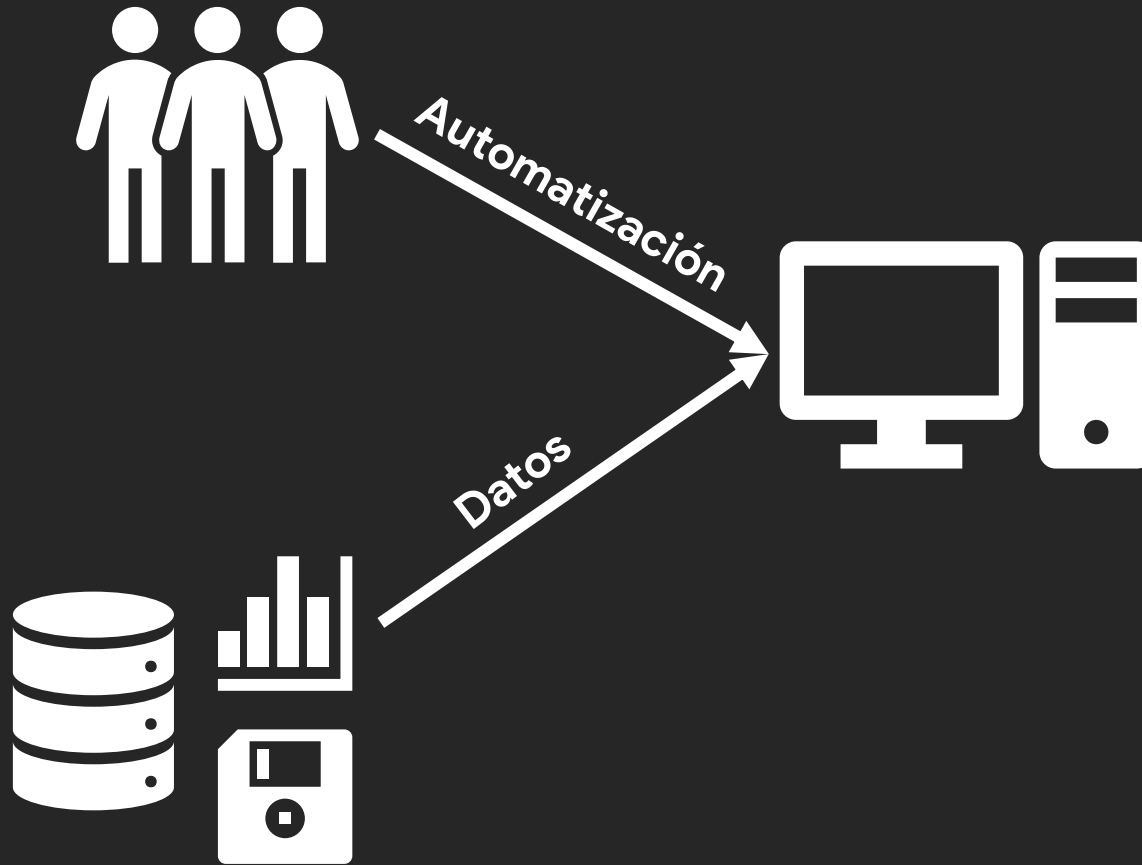
Transmitimos los mismos problemas



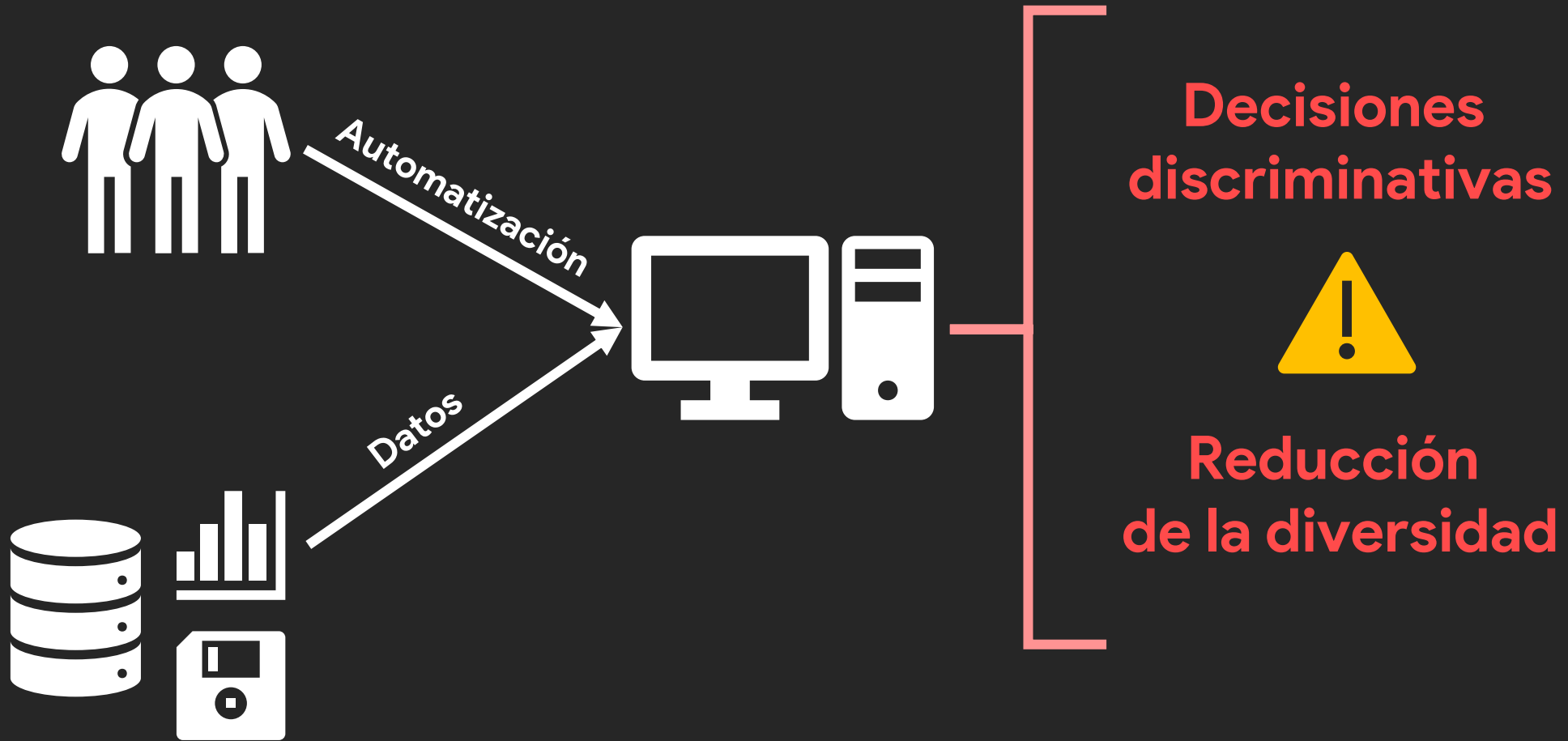
Transmitimos los mismos problemas



Transmitimos los mismos problemas



Transmitimos los mismos problemas



Transmitimos los mismos problemas



Personas

Transmitimos los mismos problemas

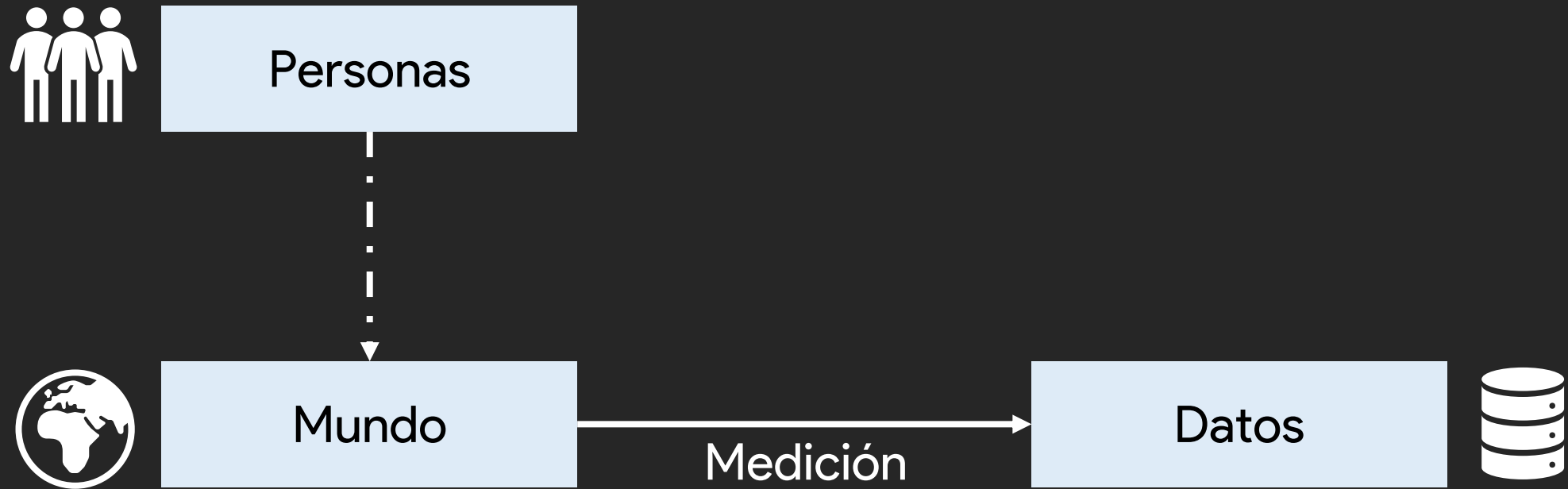


Personas

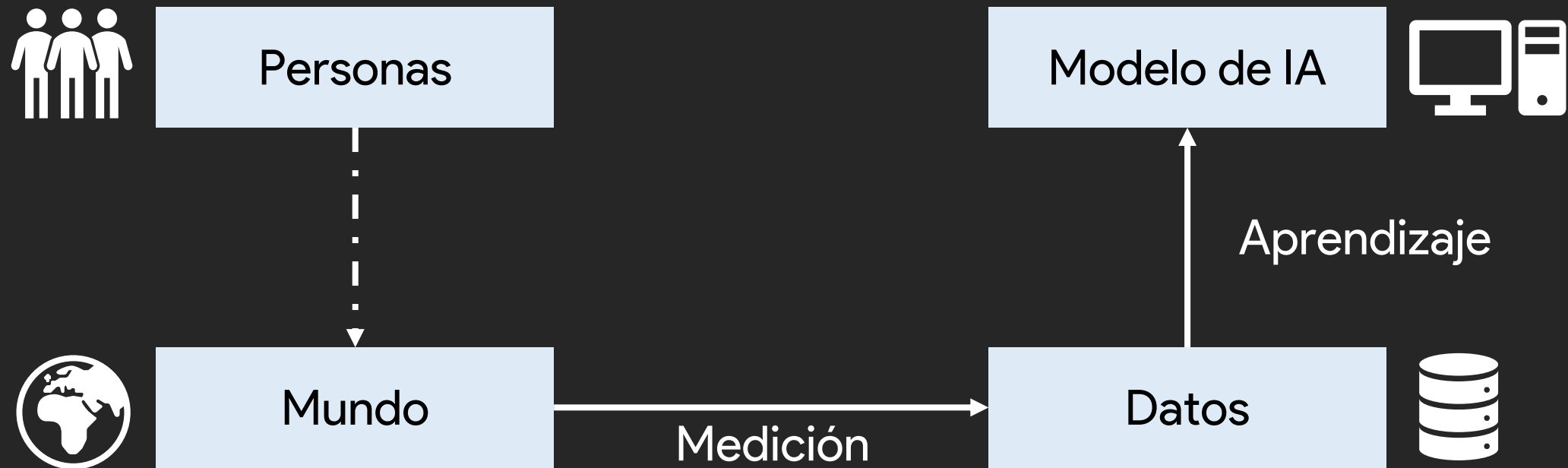


Mundo

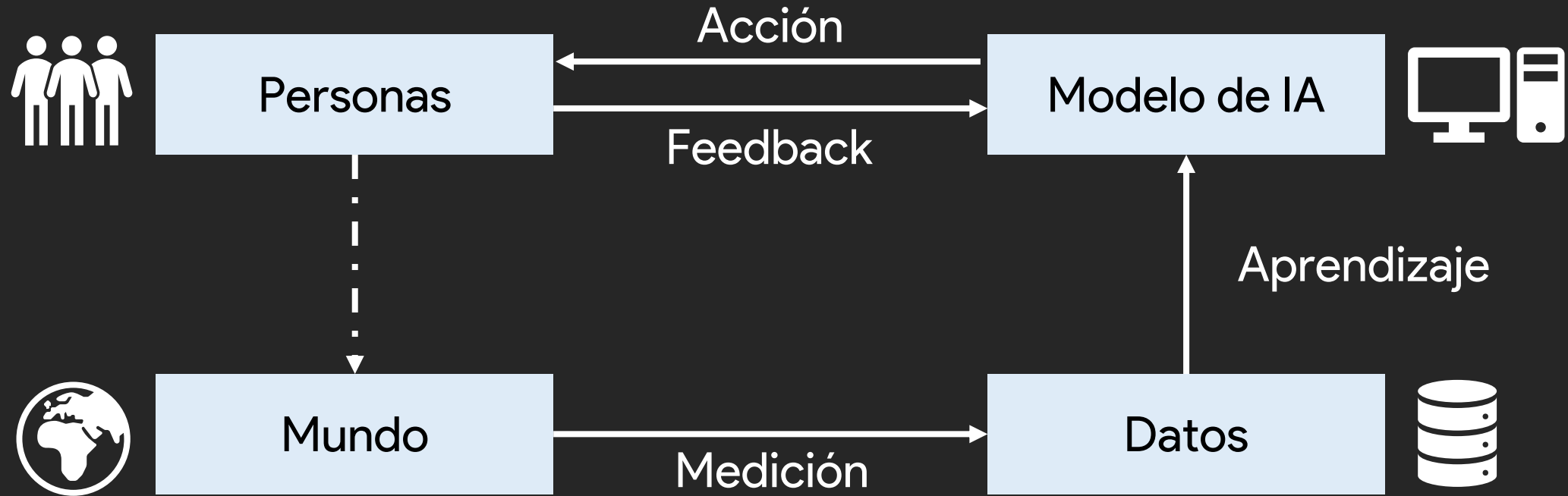
Transmitimos los mismos problemas



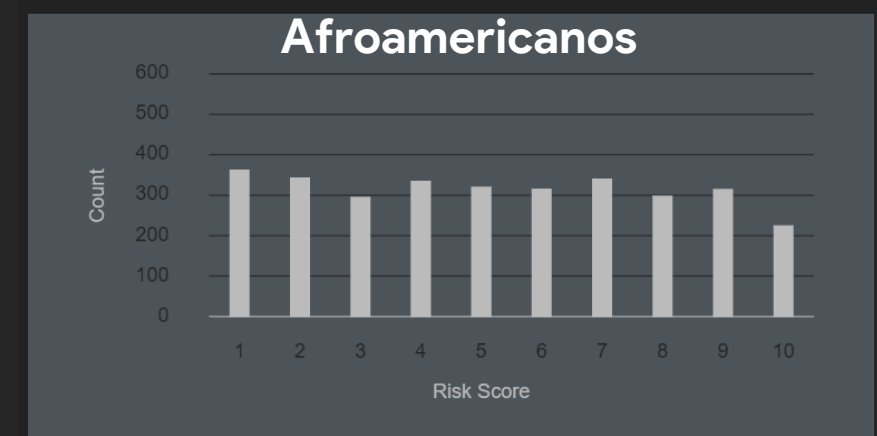
Transmitimos los mismos problemas



Transmitimos los mismos problemas



Probabilidad de reincidencia en crimen



Probabilidad de reincidencia en crimen



VERNON PRATER

Delitos anteriores
2 robos a mano armada, 1 intento de robo a mano armada

Ofensas Subsiguientes
1 hurto mayor

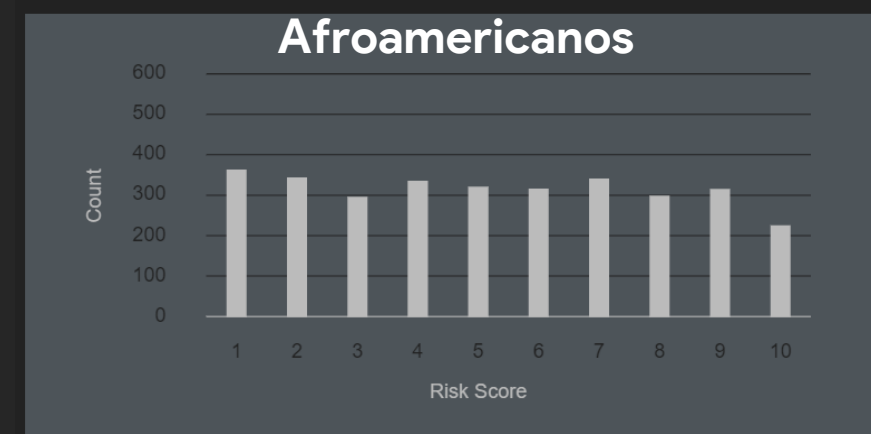
BAJORIESGO 3

BRISHA BORDEN

Delitos anteriores
4 delitos menores menores

Ofensas Subsiguientes
Ninguna

ALTORIESGO 8



Probabilidad de reincidencia en crimen



VERNON PRATER

Delitos anteriores
2 robos a mano armada, 1 intento de robo a mano armada

Ofensas Subsiguientes
1 hurto mayor

BAJORIESGO

3

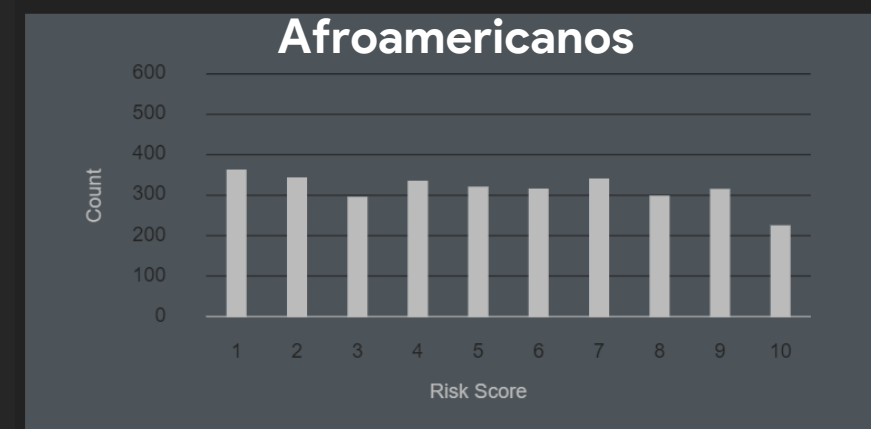
BRISHA BORDEN

Delitos anteriores
4 delitos menores menores

Ofensas Subsiguientes
Ninguna

ALTORIESGO

8



	Blancos	Afroamericanos
Predichos como alto riesgo , pero no reincidieron	23.5%	44.9%

Probabilidad de reincidencia en crimen



VERNON PRATER

Delitos anteriores
2 robos a mano armada, 1 intento de robo a mano armada

Ofensas Subsiguientes
1 hurto mayor

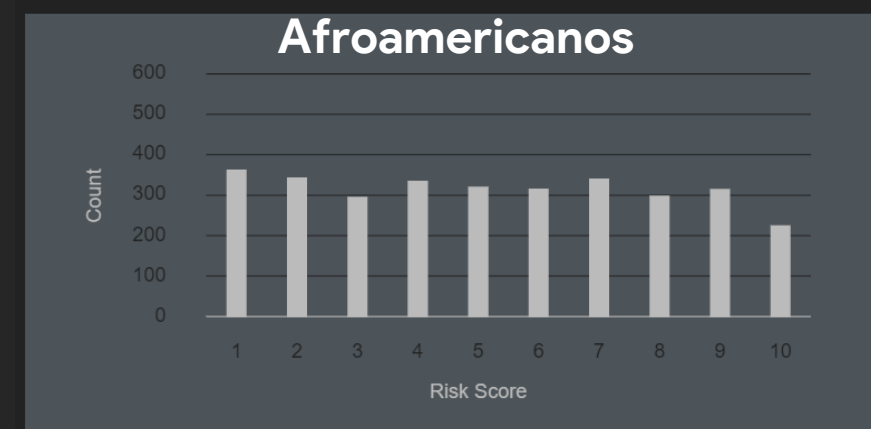
BAJORIESGO 3

BRISHA BORDEN

Delitos anteriores
4 delitos menores menores

Ofensas Subsiguientes
Ninguna

ALTORIESGO 8



	Blancos	Afroamericanos
Predichos como alto riesgo , pero no reincidieron	23.5%	44.9%
Predichos como bajo riesgo , pero sí reincidieron	47.7%	28.0%

"Machine bias" Angwin. Angwin J, Larson J, Mattu S, Kirchner L. 2016.

No es la excepción...



**The
Guardian**
For **200** years

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to
review résumés in an effort to automate the search process

No es la excepción...

**The
Guardian**
For 200 years

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

SCIENTIFIC
AMERICAN.

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

No es la excepción...



The Guardian
For 200 years

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process


SCIENTIFIC
AMERICAN.

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

Forbes

Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules

 **Jonathan Keane** Contributor ©
Consumer Tech
Freelance technology journalist covering the gig economy.

[Follow](#)

No es la excepción...

The Guardian
For 200 years

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process


SCIENTIFIC AMERICAN

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

Forbes

Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules

 **Jonathan Keane** Contributor
Consumer Tech
Freelance technology journalist covering the gig economy.

[Follow](#)

Razones personales



Aceptación o cancelación
del pedido

No es la excepción...

The Guardian
For 200 years

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process


SCIENTIFIC AMERICAN

Racial Bias Found in a Major Health Care Risk Algorithm

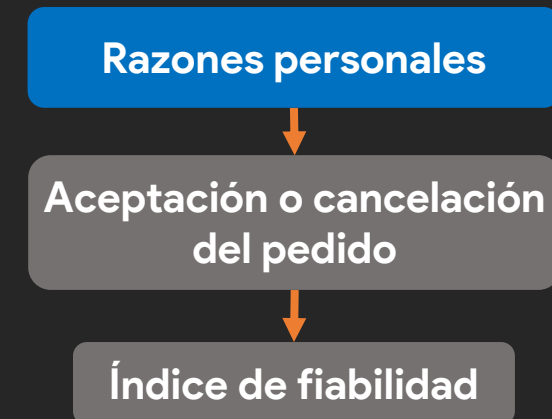
Black patients lose out on critical care when systems equate health needs with costs

Forbes

Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules

 **Jonathan Keane** Contributor
Consumer Tech
Freelance technology journalist covering the gig economy.

[Follow](#)



No es la excepción...

The Guardian
For 200 years

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process


SCIENTIFIC AMERICAN

Racial Bias Found in a Major Health Care Risk Algorithm

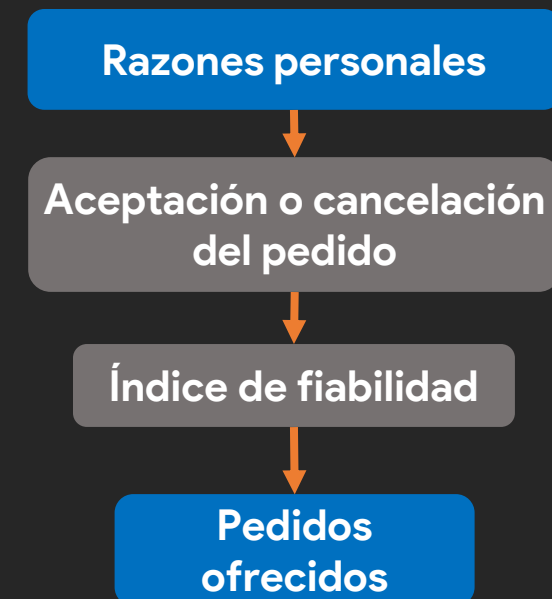
Black patients lose out on critical care when systems equate health needs with costs

Forbes

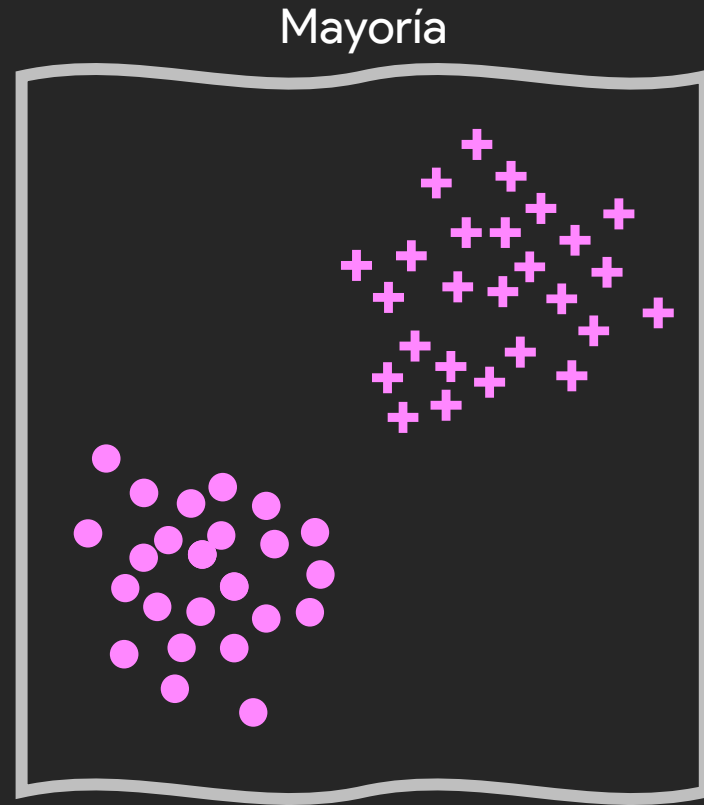
Deliveroo Rating Algorithm Was Unfair To Riders, Italian Court Rules

 **Jonathan Keane** Contributor
Consumer Tech
Freelance technology journalist covering the gig economy.

[Follow](#)

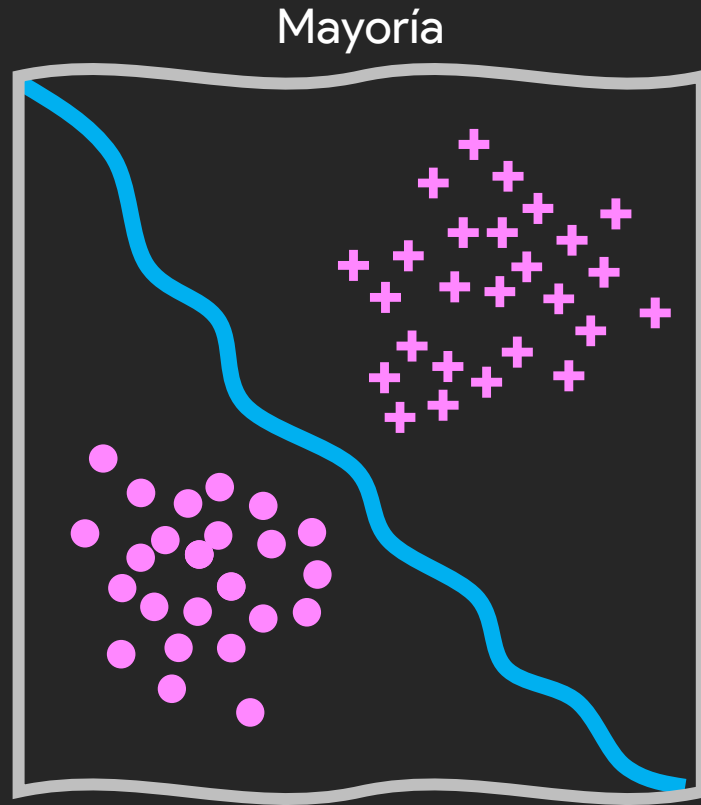


¿Por qué la IA aprende sesgada?



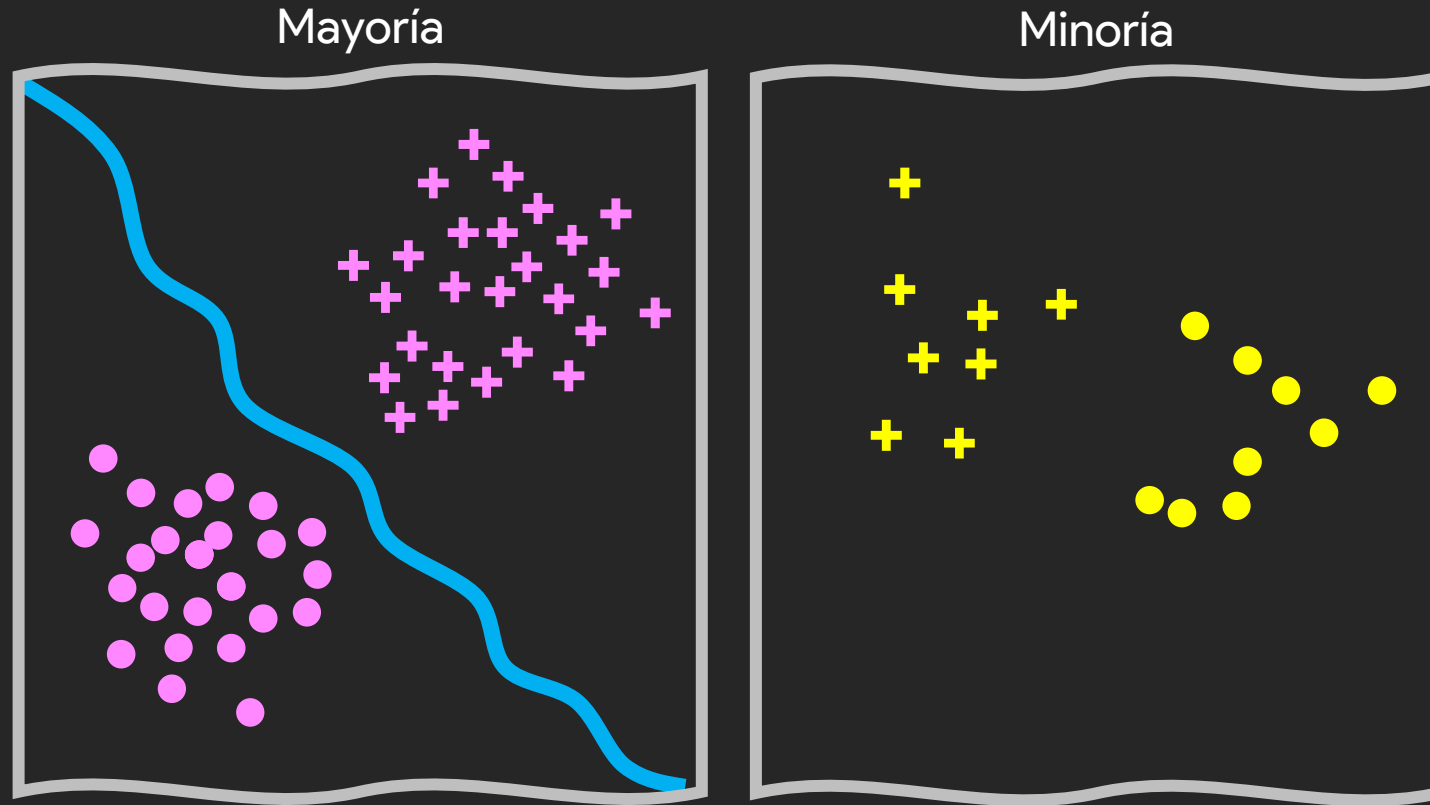
Moritz Hardt. [How big data is unfair.](#)

¿Por qué la IA aprende sesgada?



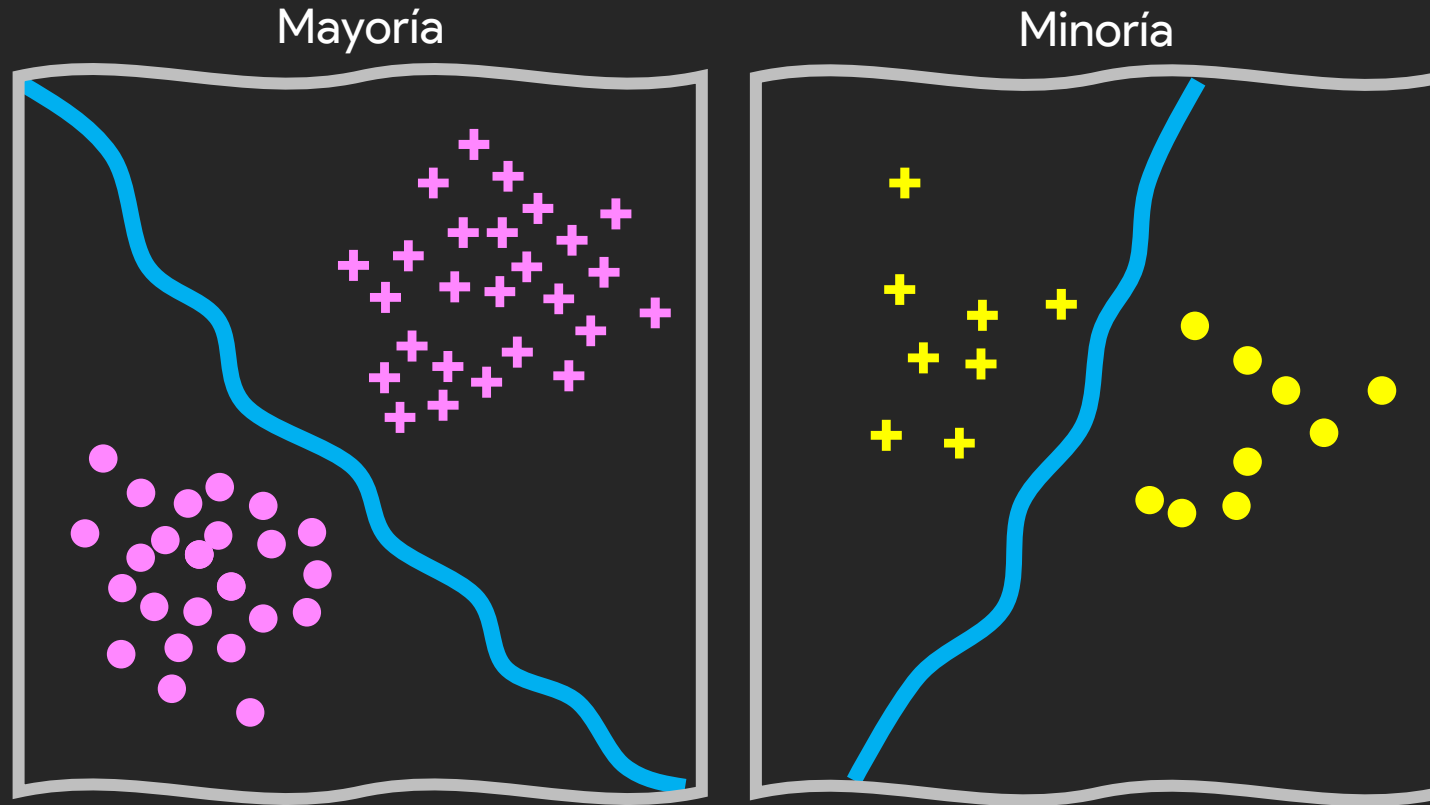
Moritz Hardt. How big data is unfair.

¿Por qué la IA aprende sesgada?



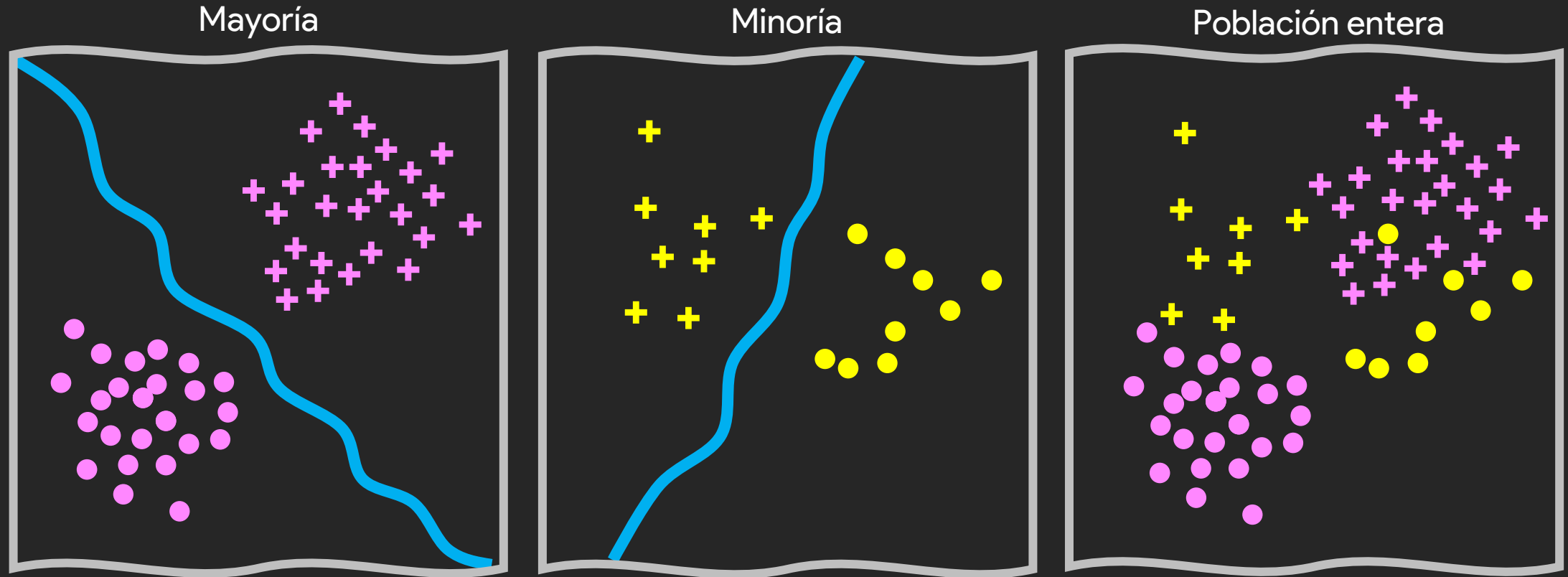
Moritz Hardt. [How big data is unfair.](#)

¿Por qué la IA aprende sesgada?



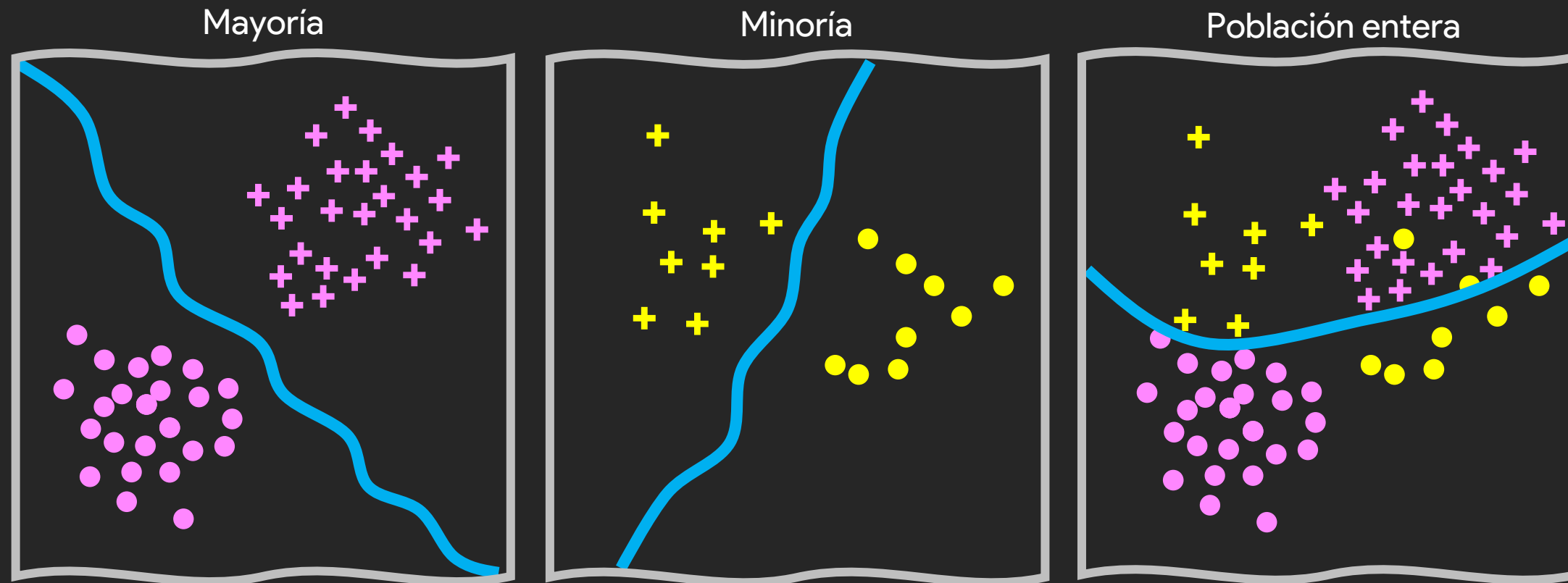
Moritz Hardt. [How big data is unfair.](#)

¿Por qué la IA aprende sesgada?



Moritz Hardt. How big data is unfair.

¿Por qué la IA aprende sesgada?



Moritz Hardt. How big data is unfair.

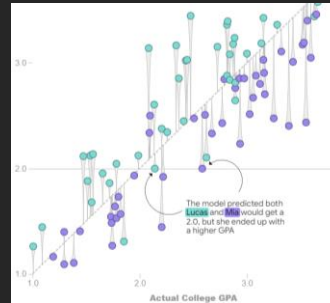
Visualizaciones interactivas

Webs interactivas* para entender distintos orígenes del sesgo en diferentes tareas.
Explicaciones interactivas de distintas métricas de justicia en IA. **para todos los públicos*

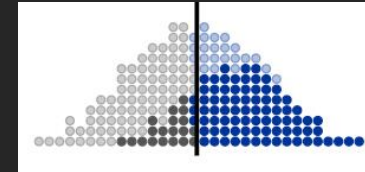
Measuring Fairness
by PAIR Google



Hidden Bias
by PAIR Google



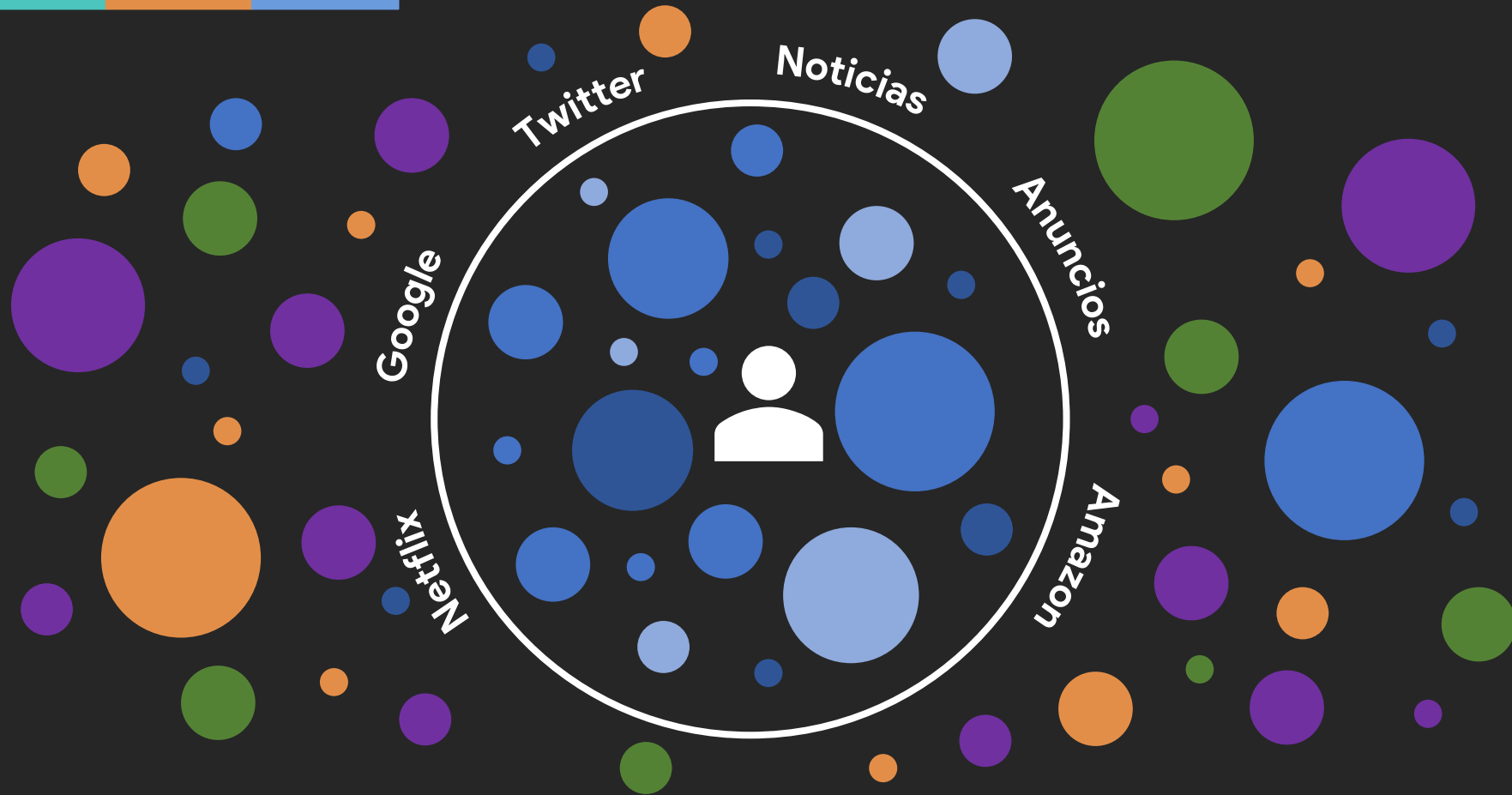
Attacking discrimination with
smarter machine learning
by Google



Mas visualizaciones y demos
interactivas
by PAIR Google

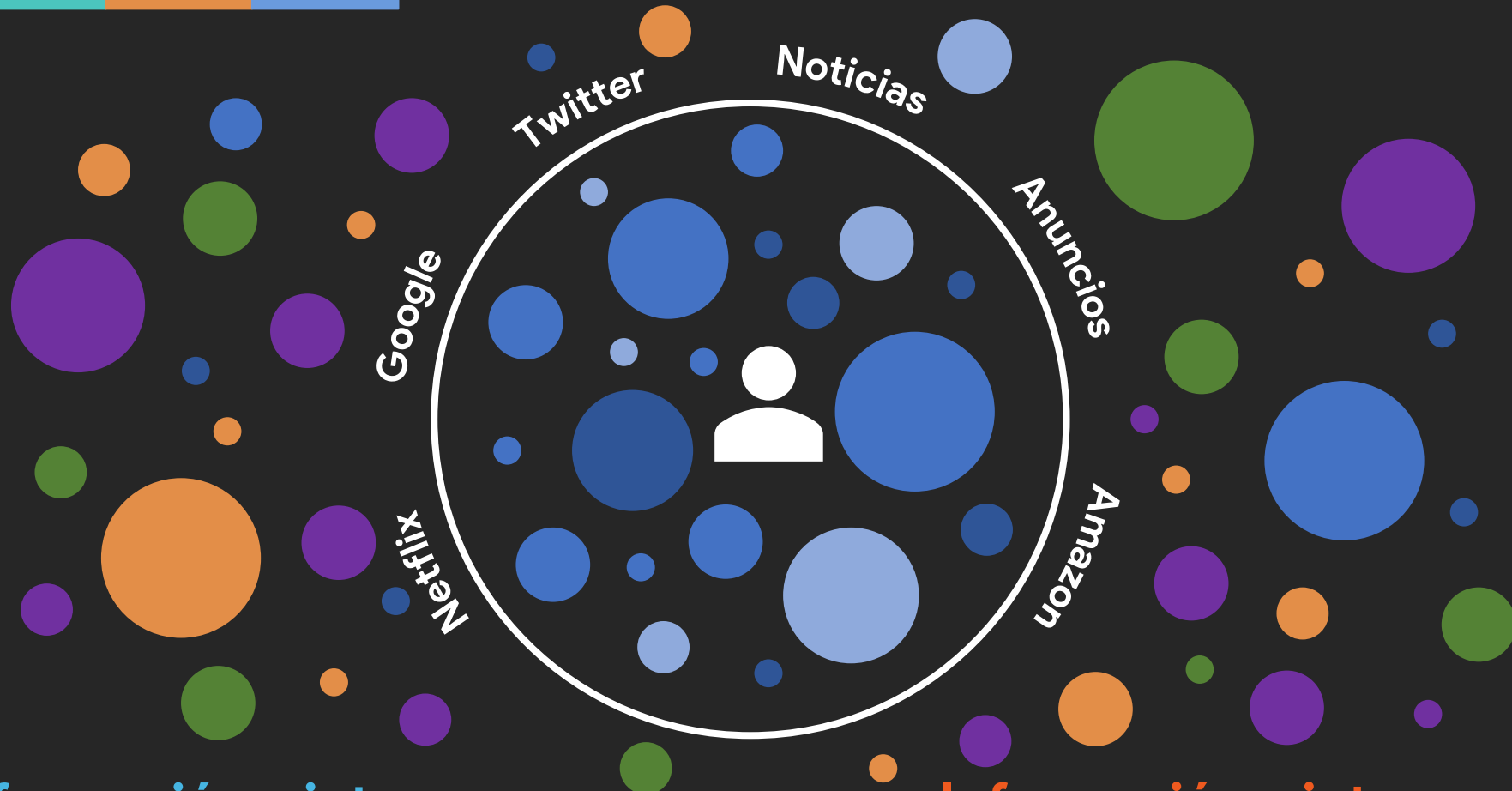
What-IF tool
by Google

Filtros Burbuja: Polarización



"The filter bubble: What the Internet is hiding from you". Eli Pariser. ([TED Talk](#))

Filtros Burbuja: Polarización

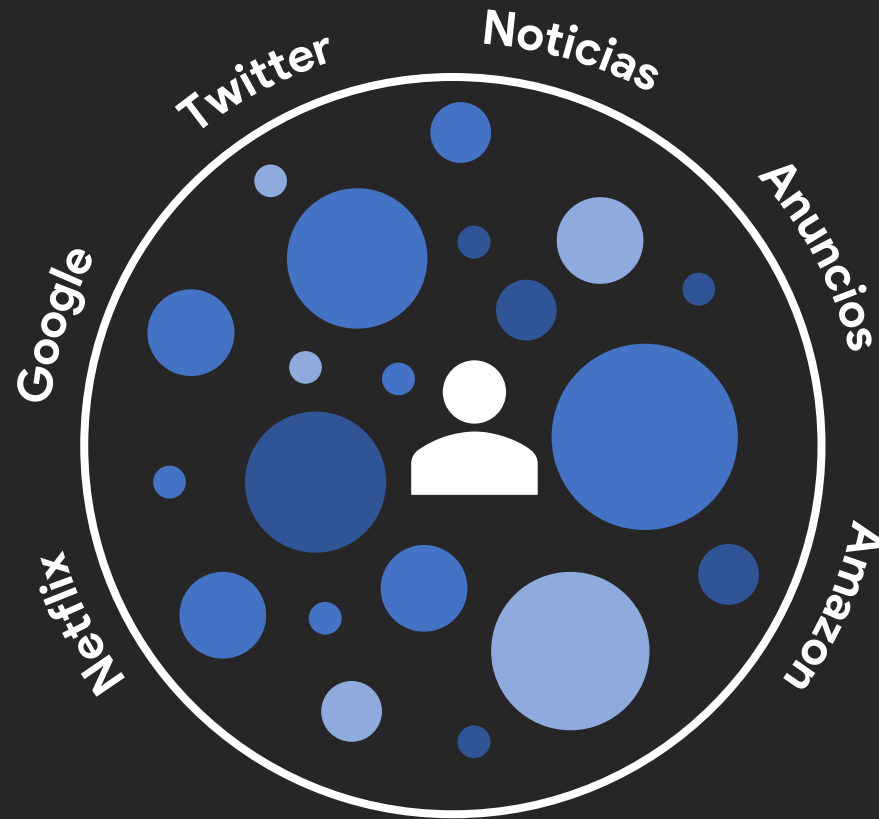


Misma información e intereses
Mismas ideas y opiniones
Gente similar

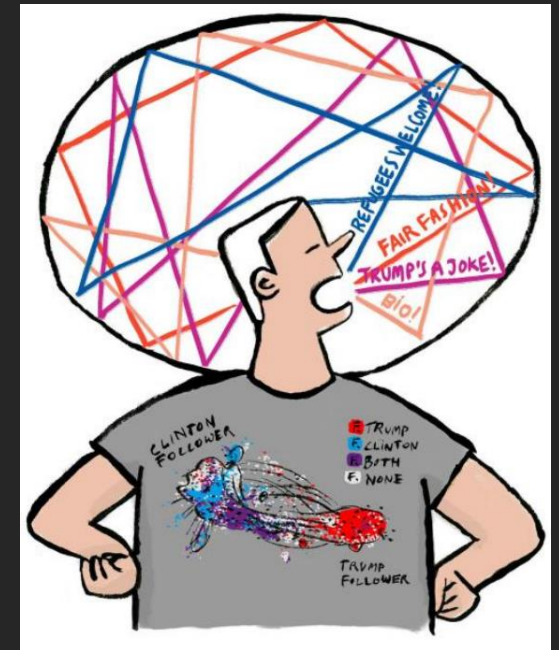
Información e intereses variados
Diversas ideas y opiniones
Amplio espectro de gente

"The filter bubble: What the Internet is hiding from you". Eli Pariser. (TED Talk)

Filtros Burbuja: Polarización



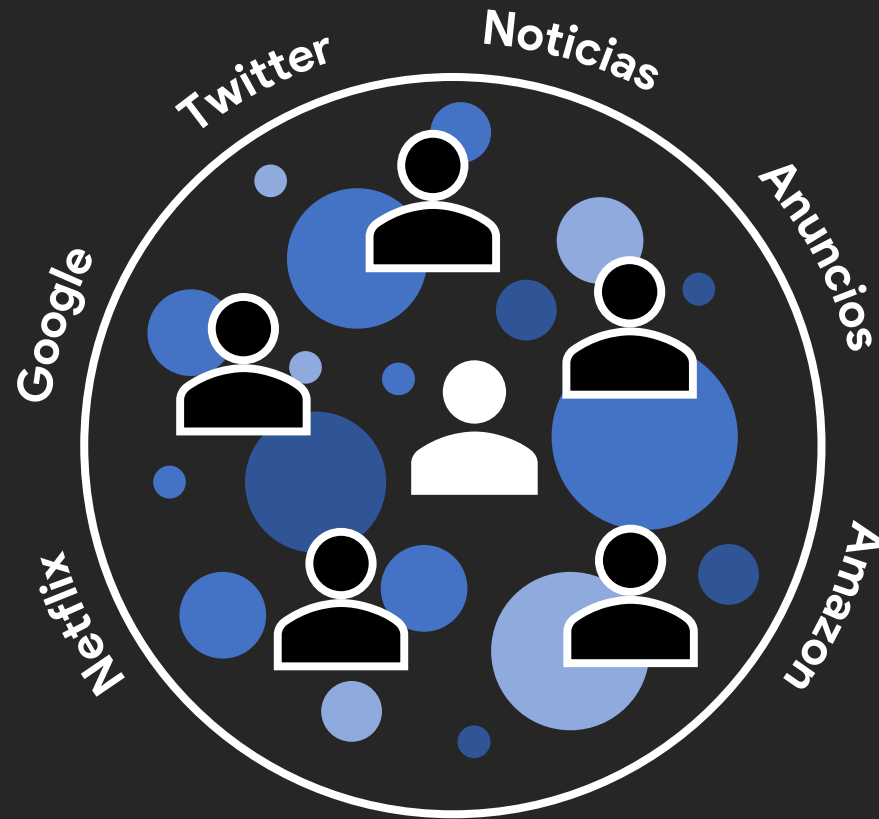
Misma información e intereses
Mismas ideas y opiniones
Gente similar



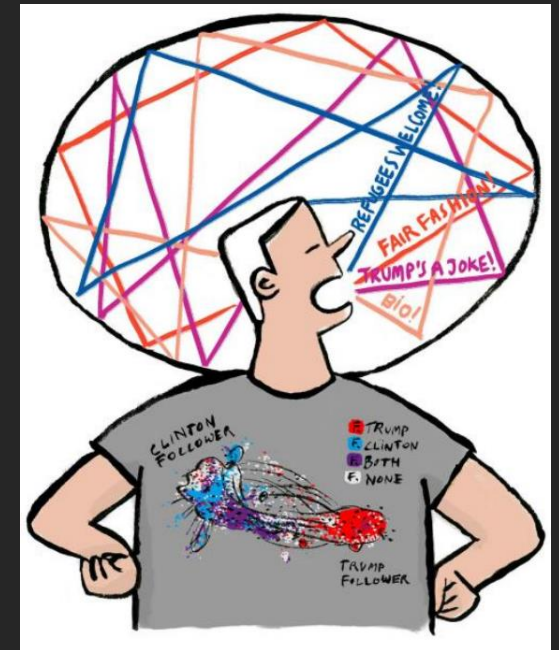
"Echo chamber". Goethe-Institut Schweden. Alex Klobouk. 2018.
[\[Link\]](#)

"The filter bubble: What the Internet is hiding from you". Eli Pariser. [\(TED Talk\)](#)

Filtros Burbuja: Polarización



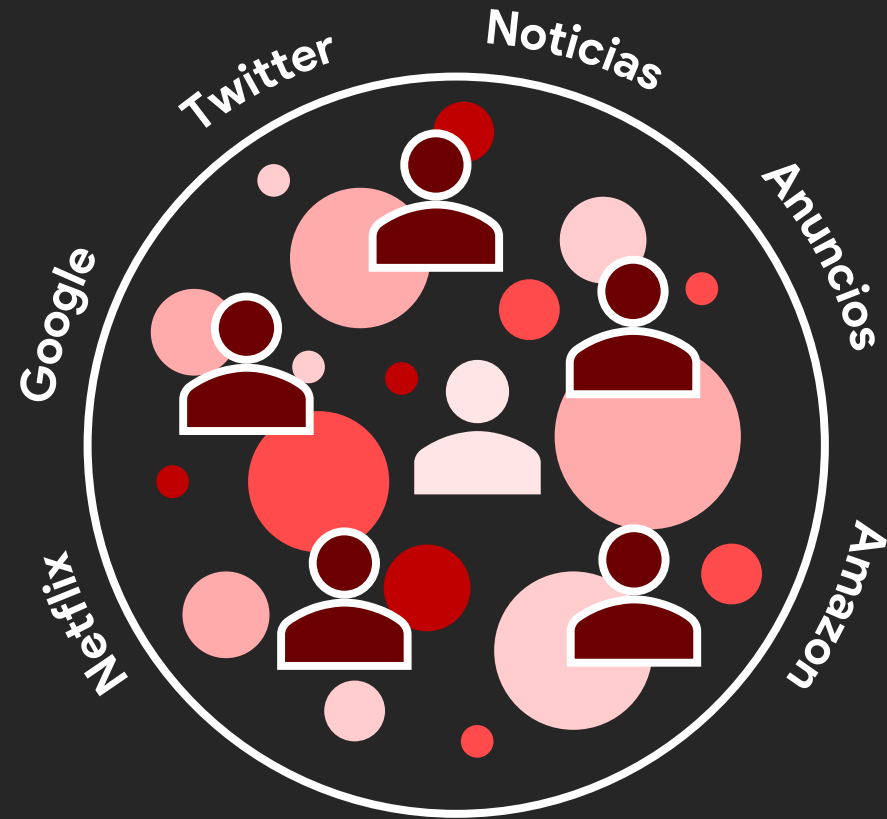
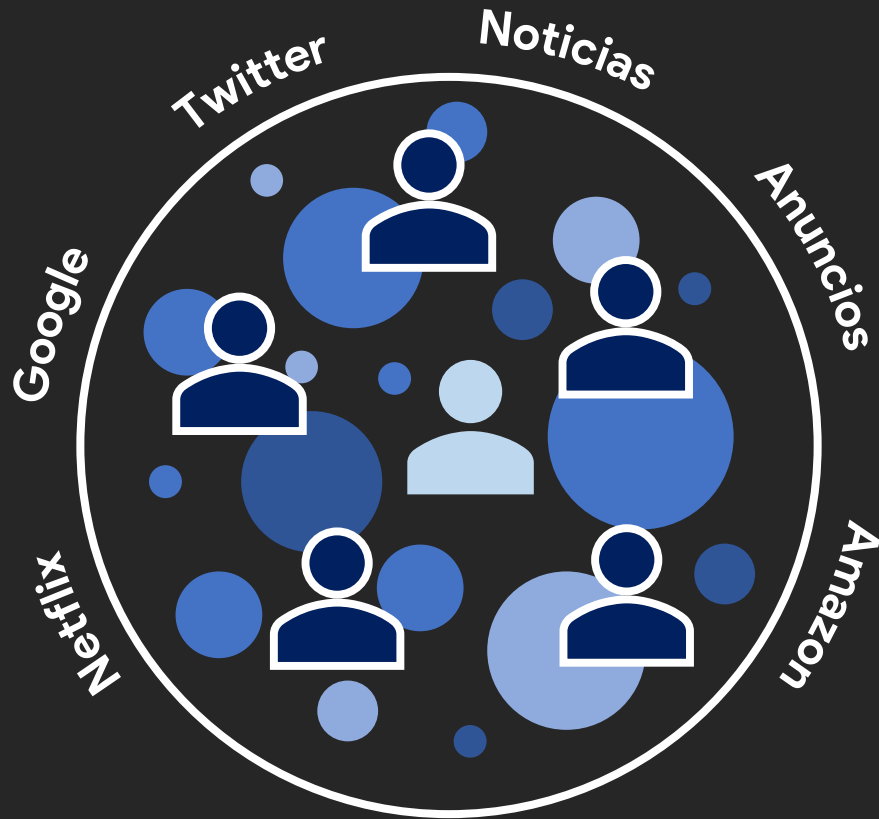
Misma información e intereses
Mismas ideas y opiniones
Gente similar



"Echo chamber". Goethe-Institut Schweden. Alex Klobouk. 2018.
[\[Link\]](#)

"The filter bubble: What the Internet is hiding from you". Eli Pariser. [\(TED Talk\)](#)

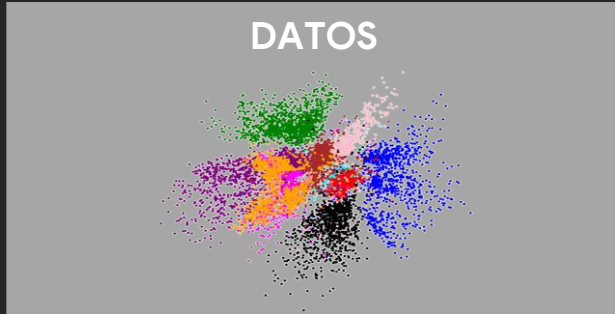
Filtros Burbuja: Polarización



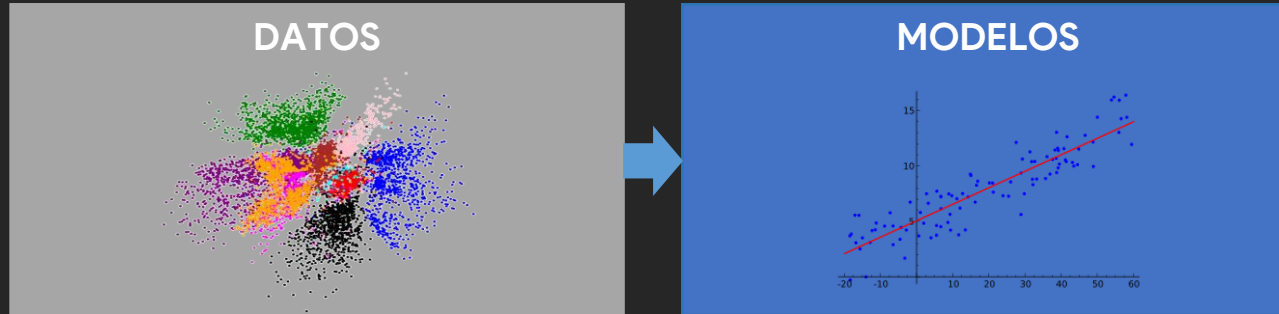
"The filter bubble: What the Internet is hiding from you". Eli Pariser. ([TED Talk](#))

IA + Humanos
para
reducir la discriminación
y
fomentar la diversidad

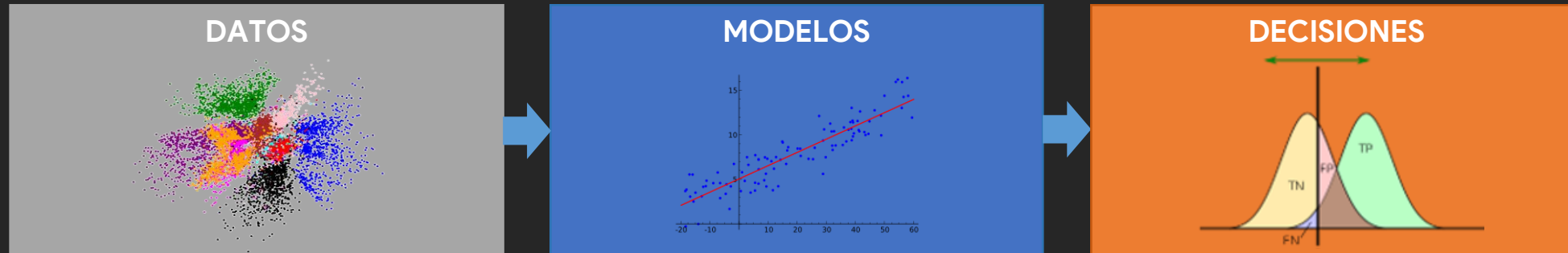
Fases en Algorithmic Fairness



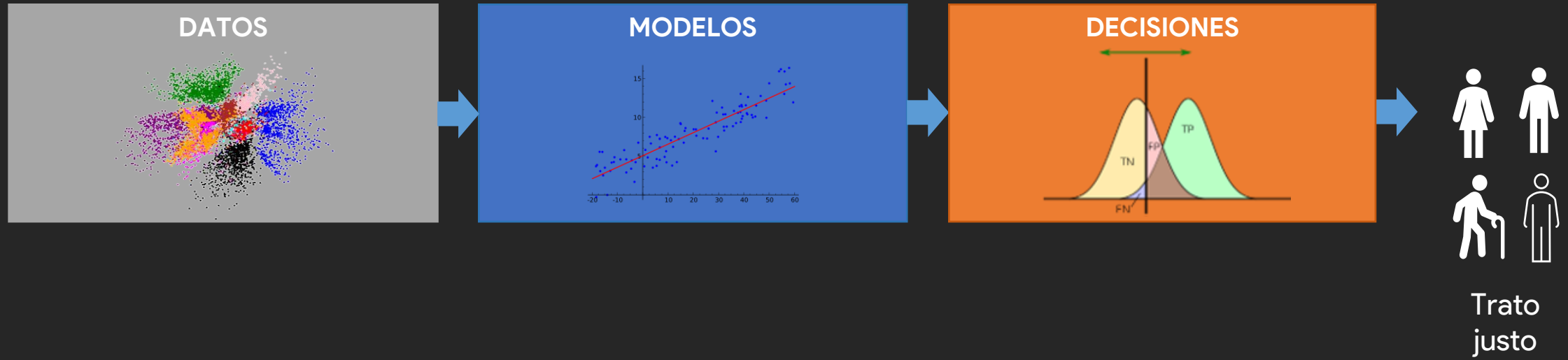
Fases en Algorithmic Fairness



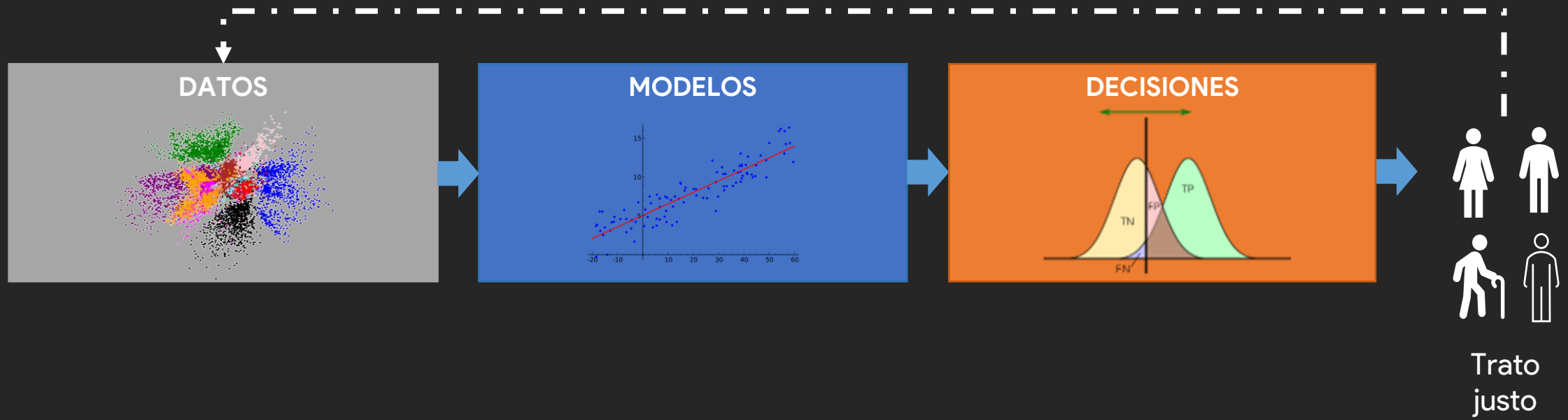
Fases en Algorithmic Fairness



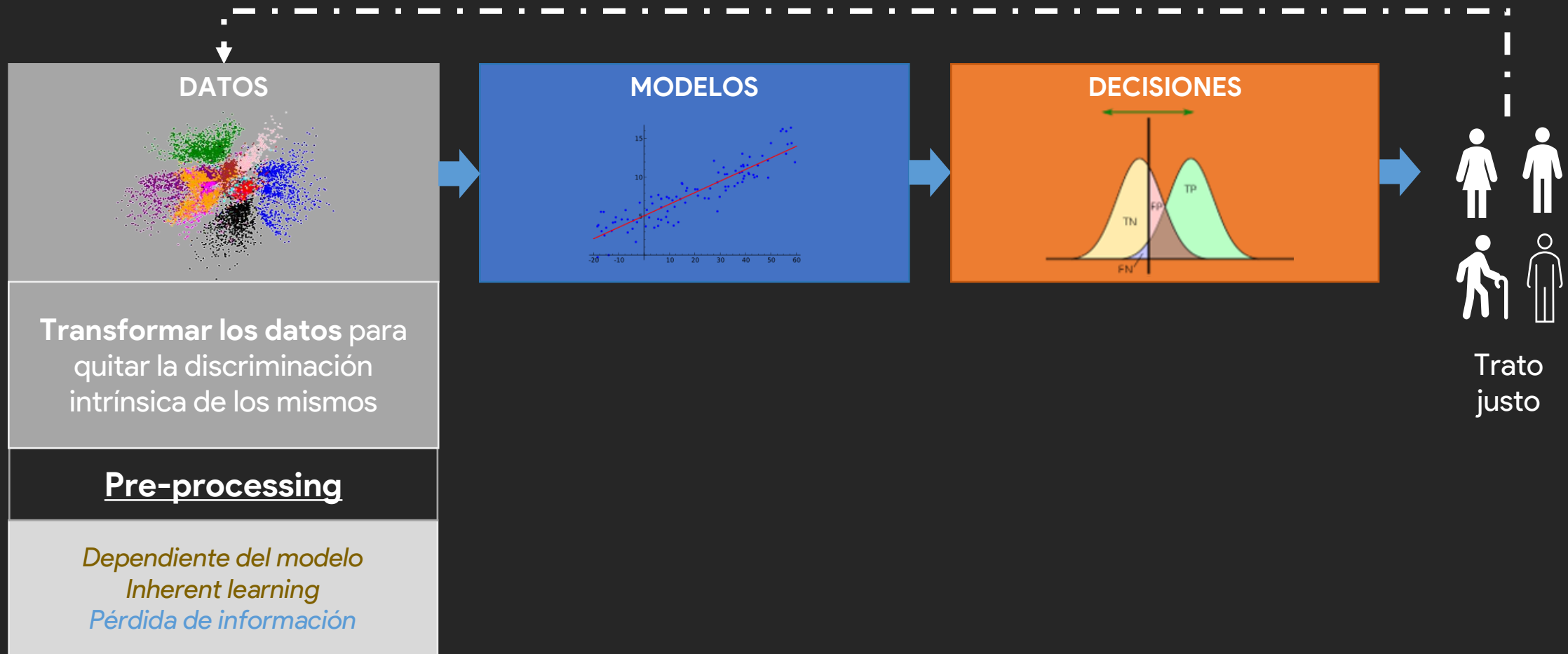
Fases en Algorithmic Fairness



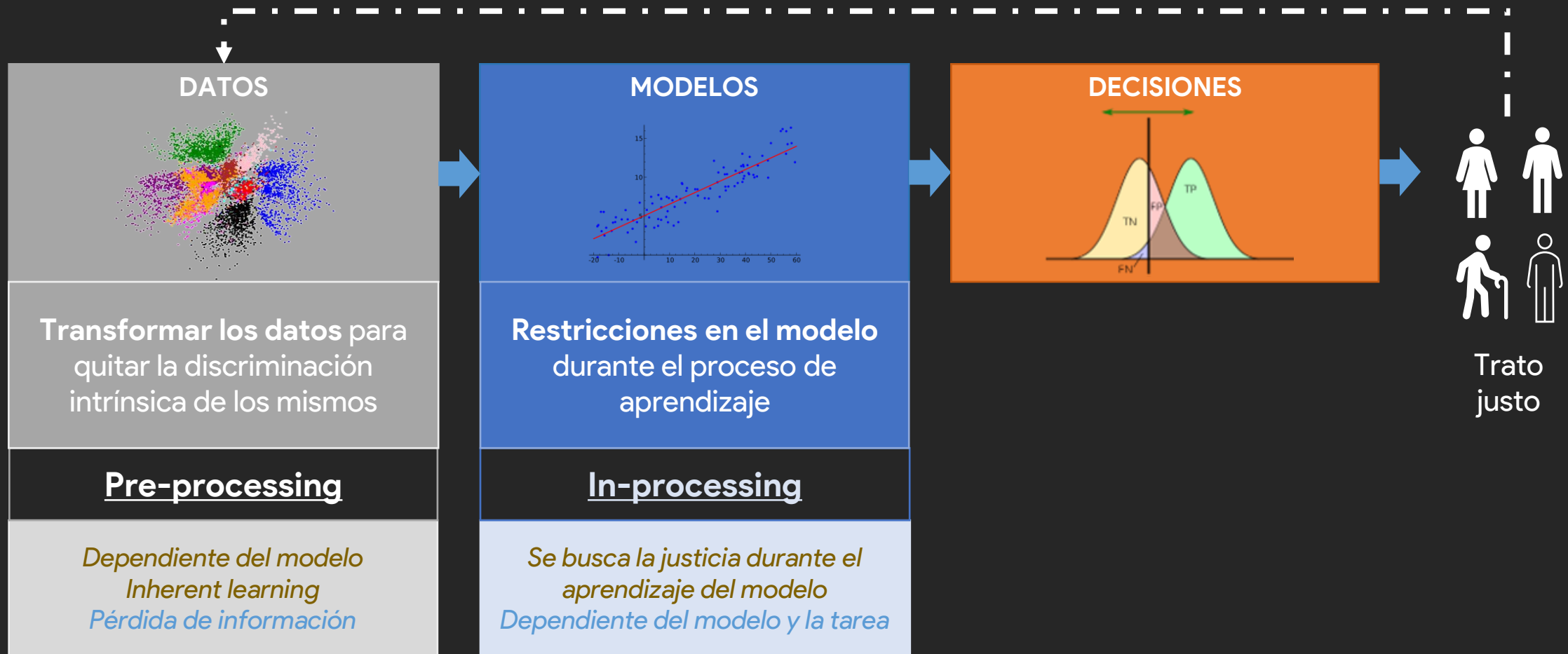
Fases en Algorithmic Fairness



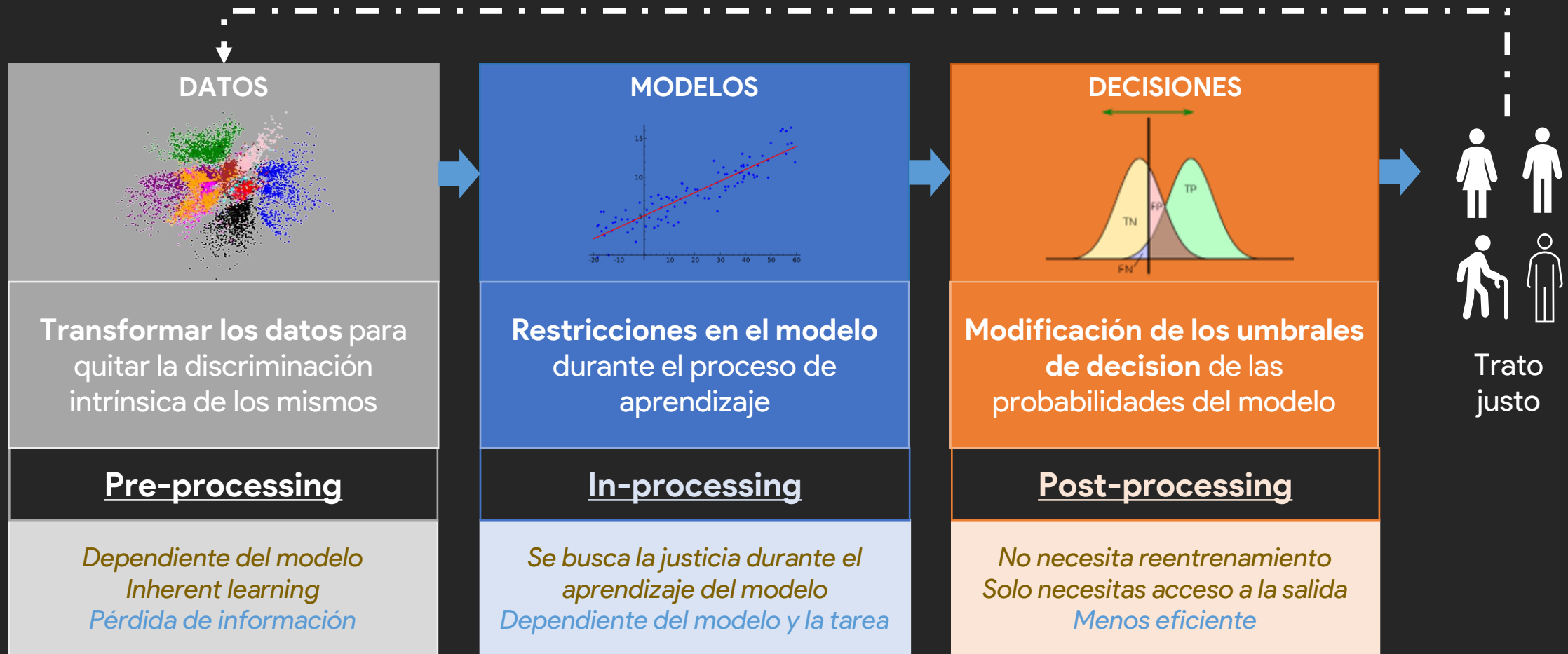
Fases en Algorithmic Fairness



Fases en Algorithmic Fairness



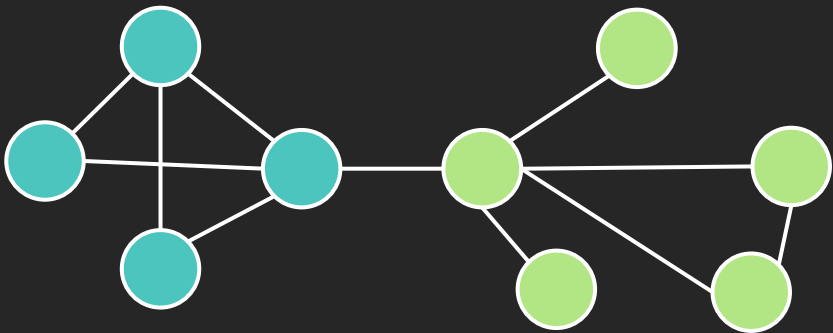
Fases en Algorithmic Fairness



Fairness en Grafos y GNNs

Utilizar redes/grafos para tomar decisiones → **La estructura importa**

Las relaciones de la red/grafio se basan en los atributos protegidos → **Homofilia**

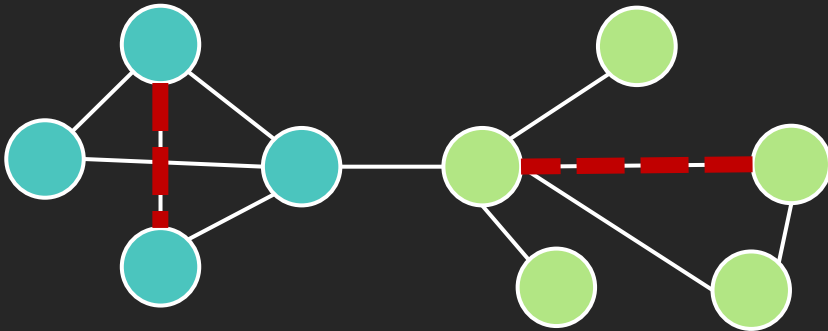


Fairness en Grafos y GNNs

Utilizar redes/grafos para tomar decisiones → **La estructura importa**

Las relaciones de la red/grafio se basan en los atributos protegidos → **Homofilia**

- Más relaciones **homofílicas** que las que debería haber

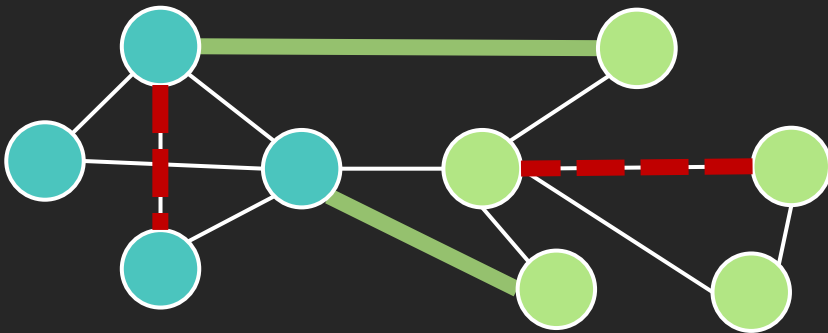


Fairness en Grafos y GNNs

Utilizar redes/grafos para tomar decisiones → **La estructura importa**

Las relaciones de la red/grafio se basan en los atributos protegidos → **Homofilia**

- Más relaciones **homofílicas** que las que debería haber
- Menos relaciones **heterofílicas** que las que hubiera habido en un mundo no sesgado

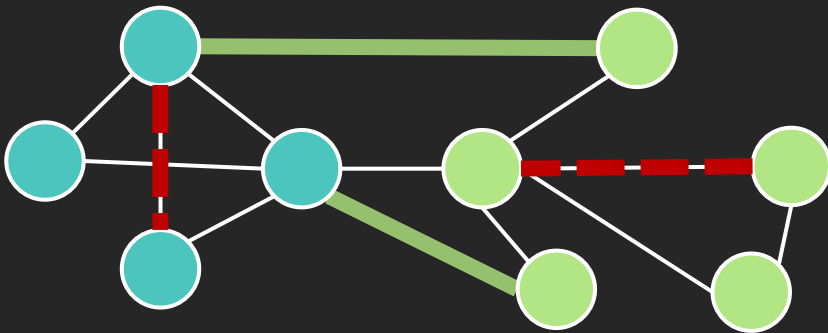


Fairness en Grafos y GNNs

Utilizar redes/grafos para tomar decisiones → **La estructura importa**

Las relaciones de la red/grafio se basan en los atributos protegidos → **Homofilia**

- Más relaciones **homofílicas** que las que debería haber
- Menos relaciones **heterofílicas** que las que hubiera habido en un mundo no sesgado
- Decisiones sesgadas aunque todas las personas tengan las mismas características

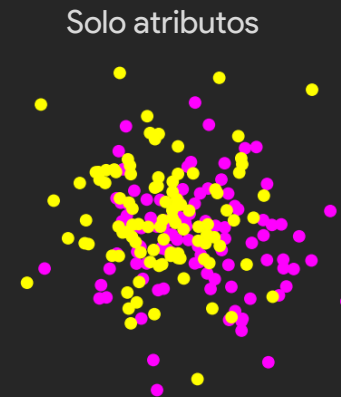
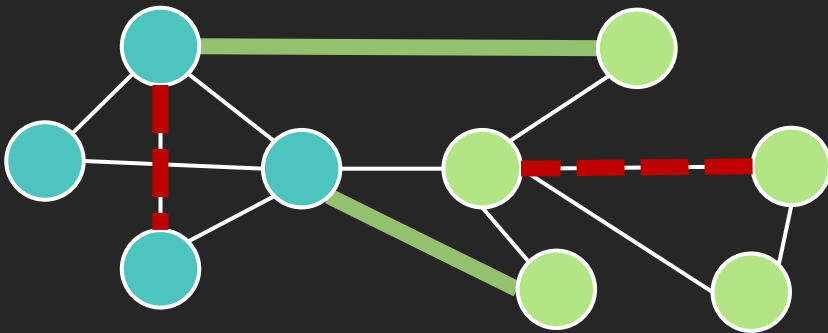


Fairness en Grafos y GNNs

Utilizar redes/grafos para tomar decisiones → **La estructura importa**

Las relaciones de la red/grafio se basan en los atributos protegidos → **Homofilia**

- Más relaciones **homofílicas** que las que debería haber
- Menos relaciones **heterofílicas** que las que hubiera habido en un mundo no sesgado
- Decisiones sesgadas aunque todas las personas tengan las mismas características

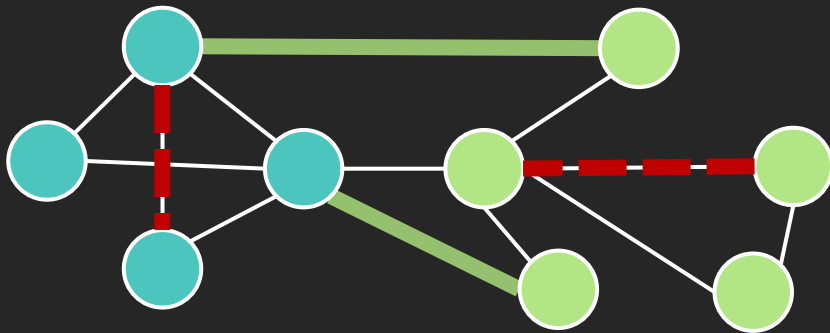


Fairness en Grafos y GNNs

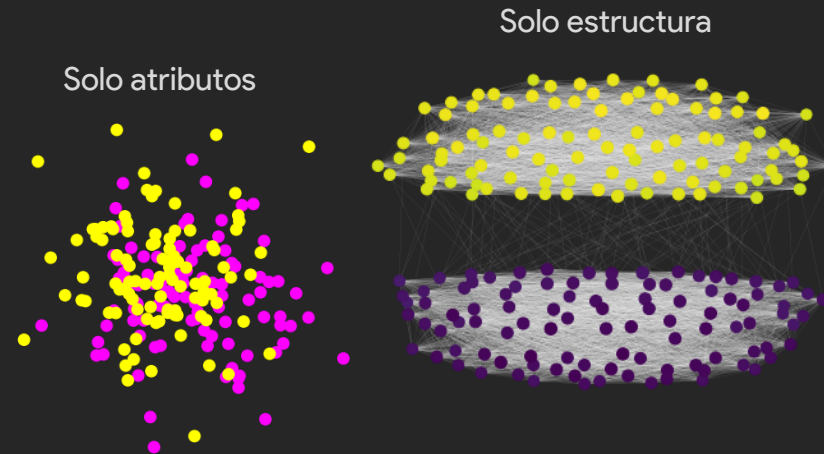
Utilizar redes/grafos para tomar decisiones → **La estructura importa**

Las relaciones de la red/grafio se basan en los atributos protegidos → **Homofilia**

- Más relaciones **homofílicas** que las que debería haber
- Menos relaciones **heterofílicas** que las que hubiera habido en un mundo no sesgado



- Decisiones sesgadas aunque todas las personas tengan las mismas características

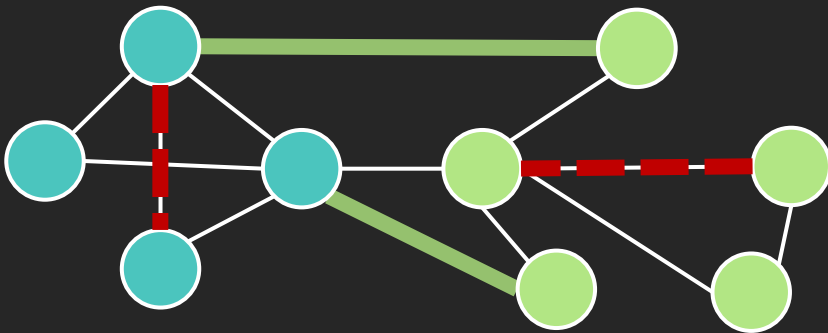


Fairness en Grafos y GNNs

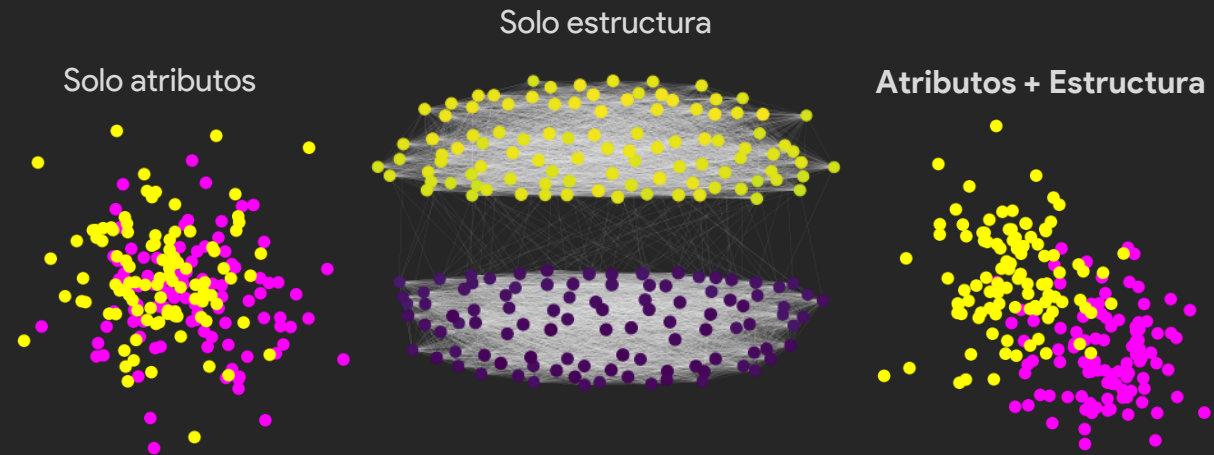
Utilizar redes/grafos para tomar decisiones → **La estructura importa**

Las relaciones de la red/grafio se basan en los atributos protegidos → **Homofilia**

- Más relaciones **homofílicas** que las que debería haber
- Menos relaciones **heterofílicas** que las que hubiera habido en un mundo no sesgado



- Decisiones sesgadas aunque todas las personas tengan las mismas características

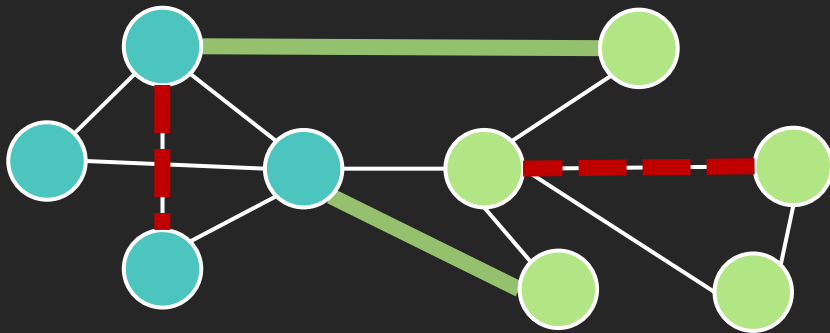


Fairness en Grafos y GNNs

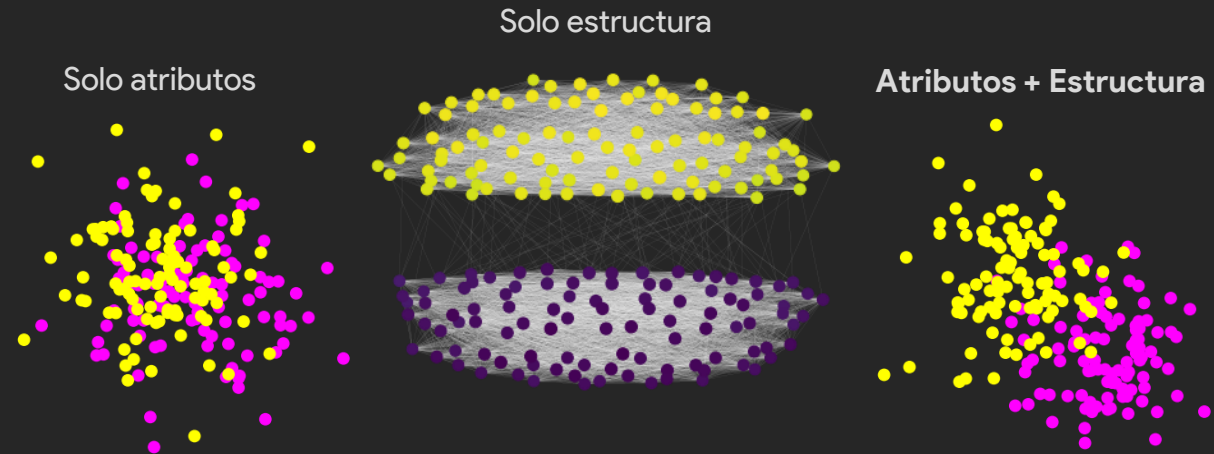
Utilizar redes/grafos para tomar decisiones → **La estructura importa**

Las relaciones de la red/grafo se basan en los atributos protegidos → **Homofilia**

- Más relaciones **homofílicas** que las que debería haber
- Menos relaciones **heterofílicas** que las que hubiera habido en un mundo no sesgado

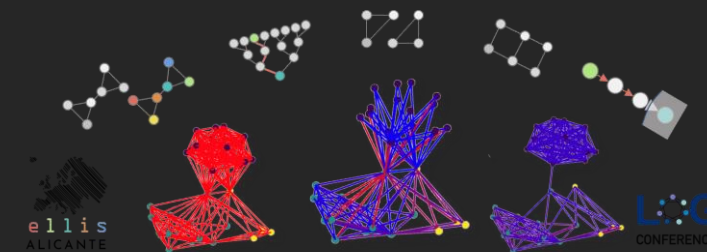


- Decisiones sesgadas aunque todas las personas tengan las mismas características



Solución: Graph Rewiring

Tutorial on Graph Rewiring

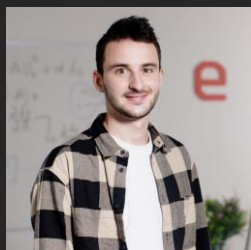


Graph Rewiring: from Theory to Applications in Fairness

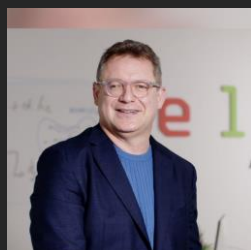
<https://ellisalicante.org/tutorials/GraphRewiring>

Graph Rewiring usando redes neuronales, teoría de grafos y geometría

Presentado en Learning on Graph Conference 2022 → Contenido de 3h de teoría, código y discusión con expertos



Adrián
Arnaiz Rodríguez
ELLIS Alicante



Francisco Escolano
ELLIS Alicante




Edwin Hancock
University of York



Nuria Oliver
e l l i s
ALICANTE unit



Petar Veličković




Marinka Zitnik




Francesco Fabbri




Francesco Di Giovanni


Áreas de investigación en ELLIS Alicante



Inteligencia
Artificial que
nos entienda

Modelización computacional del comportamiento humano usando IA

Modelar y predecir automáticamente el comportamiento humano individual y agregado a partir de datos.

¿Cómo la IA puede ayudar a y contribuir al **Bien Social** usando el modelado de comportamiento agregado?

Áreas de investigación en ELLIS Alicante



Inteligencia Artificial que nos entienda

Modelización computacional del comportamiento humano usando IA

Modelar y predecir automáticamente el comportamiento humano individual y agregado a partir de datos.

¿Cómo la IA puede ayudar a y contribuir al **Bien Social** usando el modelado de comportamiento agregado?



IA que interactúe con nosotros

Desarrollo de nuevos sistemas inteligentes e interactivos

Construcción de **interfaces de usuario inteligentes** que interactúen con los seres humanos. Investigación sobre servicios móviles sensibles al contexto, novedosas aplicaciones móviles para ayudar a las personas, computación persuasiva y asistentes personales.

Áreas de investigación en ELLIS Alicante



Inteligencia Artificial que nos entienda

Modelización computacional del comportamiento humano usando IA

Modelar y predecir automáticamente el comportamiento humano individual y agregado a partir de datos.

¿Cómo la IA puede ayudar a y contribuir al **Bien Social** usando el modelado de comportamiento agregado?



IA que interactúe con nosotros

Desarrollo de nuevos sistemas inteligentes e interactivos

Construcción de **interfaces de usuario inteligentes** que interactúen con los seres humanos. Investigación sobre servicios móviles sensibles al contexto, novedosas aplicaciones móviles para ayudar a las personas, computación persuasiva y asistentes personales.



Inteligencia Artificial en la que confiemos

Retos éticos, desafíos y limitaciones de los sistemas de inteligencia artificial

Retos éticos de IA, sus riesgos y las posibles consecuencias negativas.

- Discriminación algorítmica
- Falta de transparencia y veracidad
- Manipulación subliminal del comportamiento humano
- Privacidad
- Fragilidad
- Impacto social de los sistemas ampliamente usados como RRSS o las aplicaciones móviles



El instituto de Inteligencia Artificial centrada en las personas

The Institute for Humanity-Centric Artificial Intelligence

<https://ellisalicante.org>

Adrián Arnaiz Rodríguez

<https://adrian-arnaiz.netlify.app/>