L'impacte de la intel·ligència artificial en les polítiques públiques, la comunicació i les xarxes democràtiques.

Equidad algorítmica: perspectiva estadística

Carlos Castillo

upf. Universitat Pompeu Fabra Barcelona

Organitzat per:
Canòdrom - Ateneu d'Innovació Digital i Democràtica

¿Cómo descubrir la discriminación?

- Encuestas o entrevistas
- Denuncias
- **Análisis estadístico**:

    1. Determinar desigualdad en un resultado
    2. Controlar posibles variables explicativas
    3. …

Devah Pager and Hana Shepherd (2008). The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. Annual Review of Sociology.

Minimización vs no-discriminación

Tensión entre dos principios

Minimización de datos

> Un sistema no debería pedir datos que no son esenciales para su funcionamiento, justificadamente

Si no sabemos a qué grupo pertenecen las personas que recibieron un trato desventajoso, no podemos establecer que esas personas han sido discriminadas

Discriminación directa e indirecta

Aproximación basada en datos

Dada una **base de datos de decisiones anteriores** y un conjunto de **grupos potencialmente discriminados**

Encontrar **situaciones** (individuos) y **prácticas** (patrones) discriminatorias

S. Ruggieri, D. Pedreschi and F. Turini (2010). DCUBE: Discrimination discovery in databases. In SIGMOD, pp. 1127-1130.

Decisiones discriminatorias

De las decisiones anteriores, podemos aprender que:
género = fem. ⇒ crédito = no

P(género=fem, credit=no) / P(género=fem) > θ

Esto puede ser evidencia de discriminación

género	**sit. lab**	**crédito**
masc	trabaja	sí
masc	paro	sí
masc	trabaja	sí
fem	paro	no
fem	trabaja	no
fem	trabaja	sí
...	...	...

Hajian, S., Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. IEEE TKDE 25(7), 1445-1459.

Decisiones indirectamente discrim.

Si descubrimos que

cód. post = 8002 ⇒ crédito = no

… y sabemos …

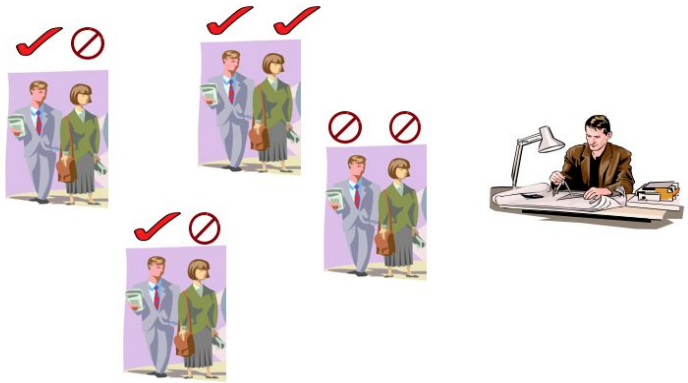
cód. post = 8002 ⇒
origen = extranjero

Esto puede ser evidencia de **discriminación indirecta**

origen	cód. post	crédito
nacional	8001	sí
nacional	8001	sí
nacional	8001	sí
extranjero	8002	no
extranjero	8002	no
extranjero	8002	sí
...	...	...

Experimentos ("*situational testing*")

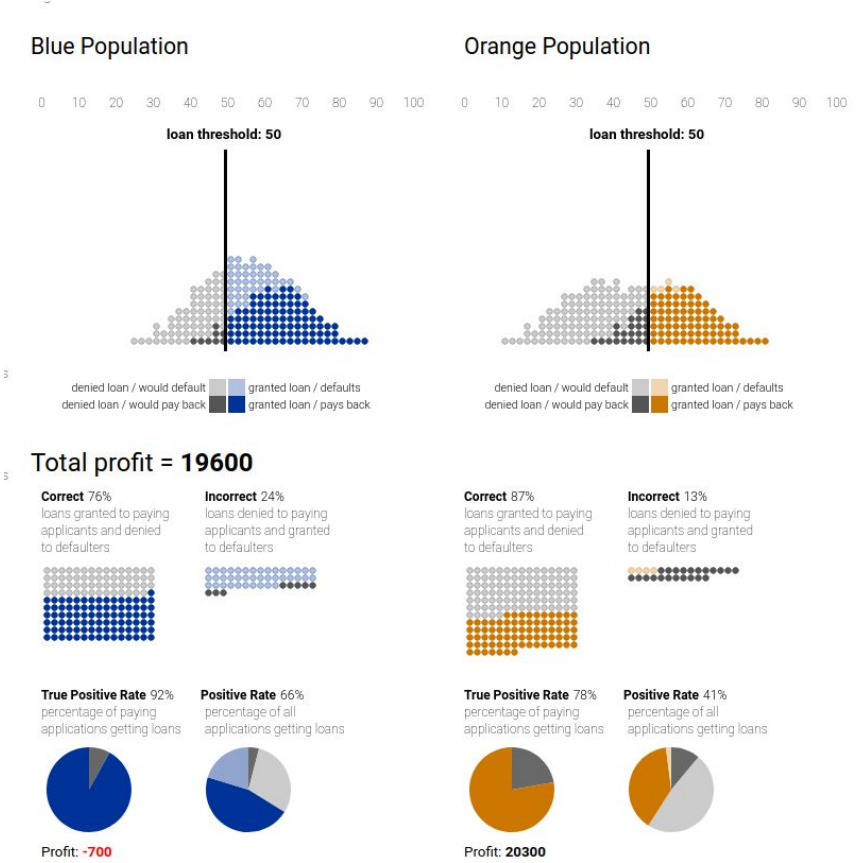
- Experimentos controlados
- Parejas correspondientes (iguales características excepto atributo protegido) postulan a un trabajo



Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. American economic review, 94(4), 991-1013.

Discriminación en el cálculo de riesgo

Simulador de crédito



https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Evaluación de riesgo

Actuarial -- estructurada

hecha en base a un instrumento estadístico

Clínica -- no estructurada

hecha en base a experiencia profesional

En muchos casos se usa una combinación de ambos

Confianza: dos extremos

Sesgo hacia la automatización / *automation bias*:

"Todo lo que dice el método actuarial es correcto"

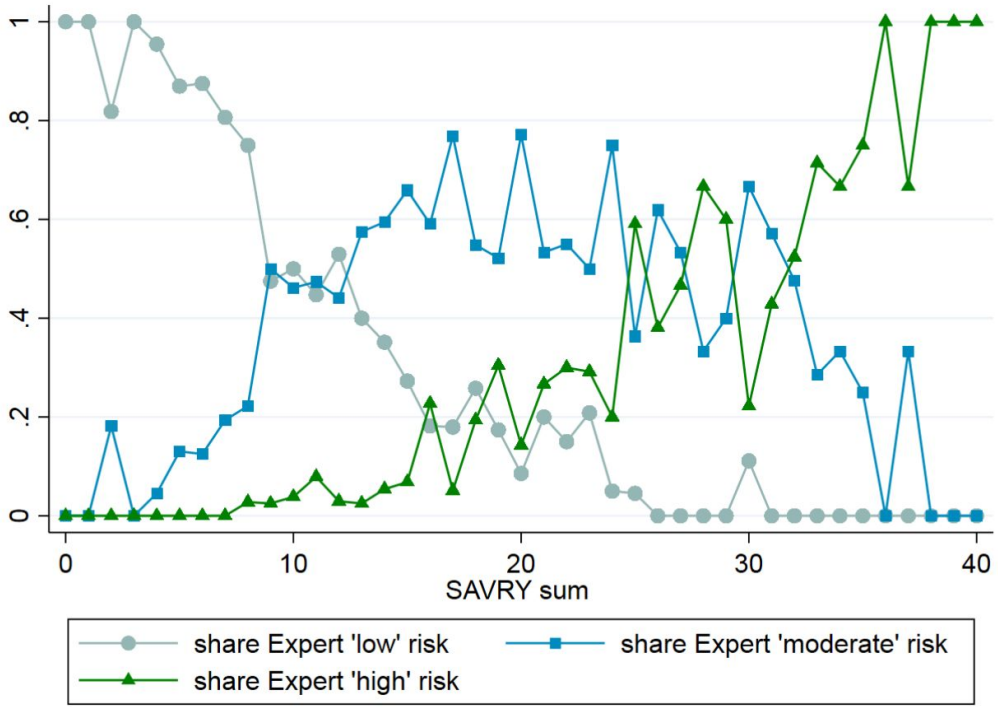
Aversión algorítmica / *algorithmic aversion*:

"Todo lo que dice el método actuarial es incorrecto"

Ejemplo de confianza acotada

SAVRY -- Riesgo de violencia en jóvenes

Él/la experto/a no confía plenamente en la herramienta, pero tampoco la ignora



Marius Miron, Songül Tolan, Emilia Gomez, Carlos Castillo: "Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia". Proc. ICAIL. ACM Press.

Discriminación en clasificación

Based on: Solon Barocas, Moritz Hardt, Arvind Narayanan: Fairness in Machine Learning (work in progress). Chapter 02. Classification.

Independencia y separabilidad

Métodos de inteligencia combinada (actuarial + clínica) que hacen predicciones sobre una persona (¿será buen estudiante?, ¿desarollará una infección?, ¿se quedará sin techo?, ¿sufrirá maltrato en el hogar?, ¿pagará el crédito?, ¿cometerá otro delito?, …)

Independencia: la predicción no depende del grupo

Separabilidad: la predicción depende del grupo exactamente debido al resultado de cada grupo

Criterio de **independencia**

Solo examina la predicción

Ejemplo:

"Grupo protegido" = "persona racializada"

"Beneficio" = "obtener una beca"

	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

Intuitivamente, si

a/n_1, el **riesgo** que tienen las **personas racializadas** de no obtener la beca

es mucho mayor que

c/n_2, el **riesgo** que tienen las **personas no racializadas** de no obtener la beca

… entonces la herramienta no cumple con el criterio de **independencia**.

D. Pedreschi, S. Ruggieri, F. Turini: A Study of Top-K Measures for Discrimination Discovery. SAC 2012.

Objeción al criterio de **independencia**

¿Y si los grupos son diferentes? Por ejemplo, las mujeres reinciden menos que los hombres. Si un sistema es independiente del género, predecirá una tasa de reincidencia más alta que la que las mujeres realmente tienen

Separabilidad

Este es un criterio de **paridad en el acierto y el error**:

- Verdaderos positivos: personas que dijimos que cometerían un nuevo delito y lo cometieron
- Falsos positivos: personas que dijimos que cometerían un nuevo delito y no lo cometieron

Para que exista separabilidad:

- Las tasas de verdaderos positivos de cada grupo deben ser iguales
- Las tasas de falsos positivos de cada grupo deben ser iguales

Discriminación en ranking

Equidad en un buscador

1. Presencia suficiente de personas de grupos protegidos

2. Tratamiento consistente de las personas

3. Representación apropiada de las personas

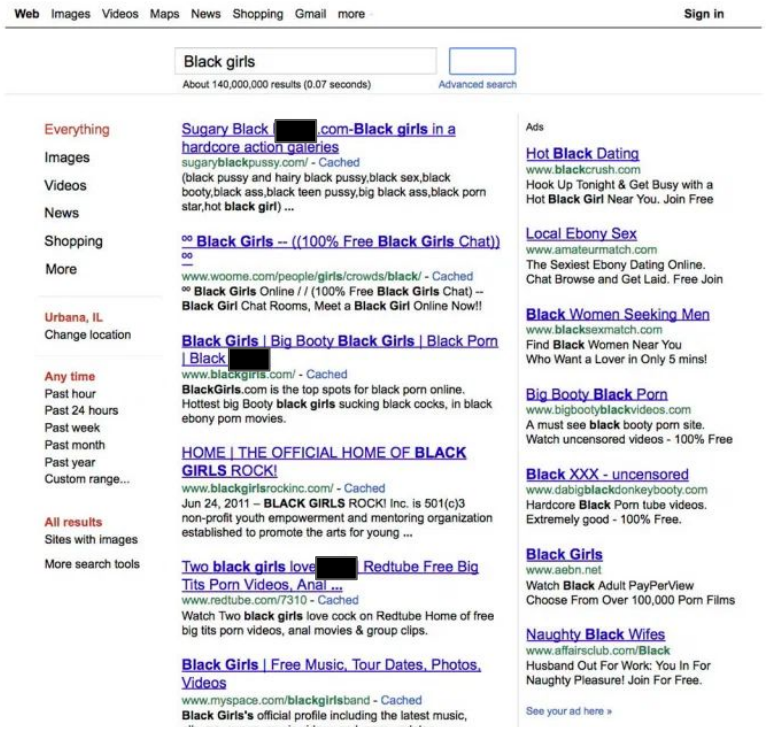
Castillo, C. (2019, January). Fairness and Transparency in Ranking. In ACM SIGIR Forum (Vol. 52, No. 1, pp. 64-71). ACM.

≃ "P-fairness"

Representación inapropiada

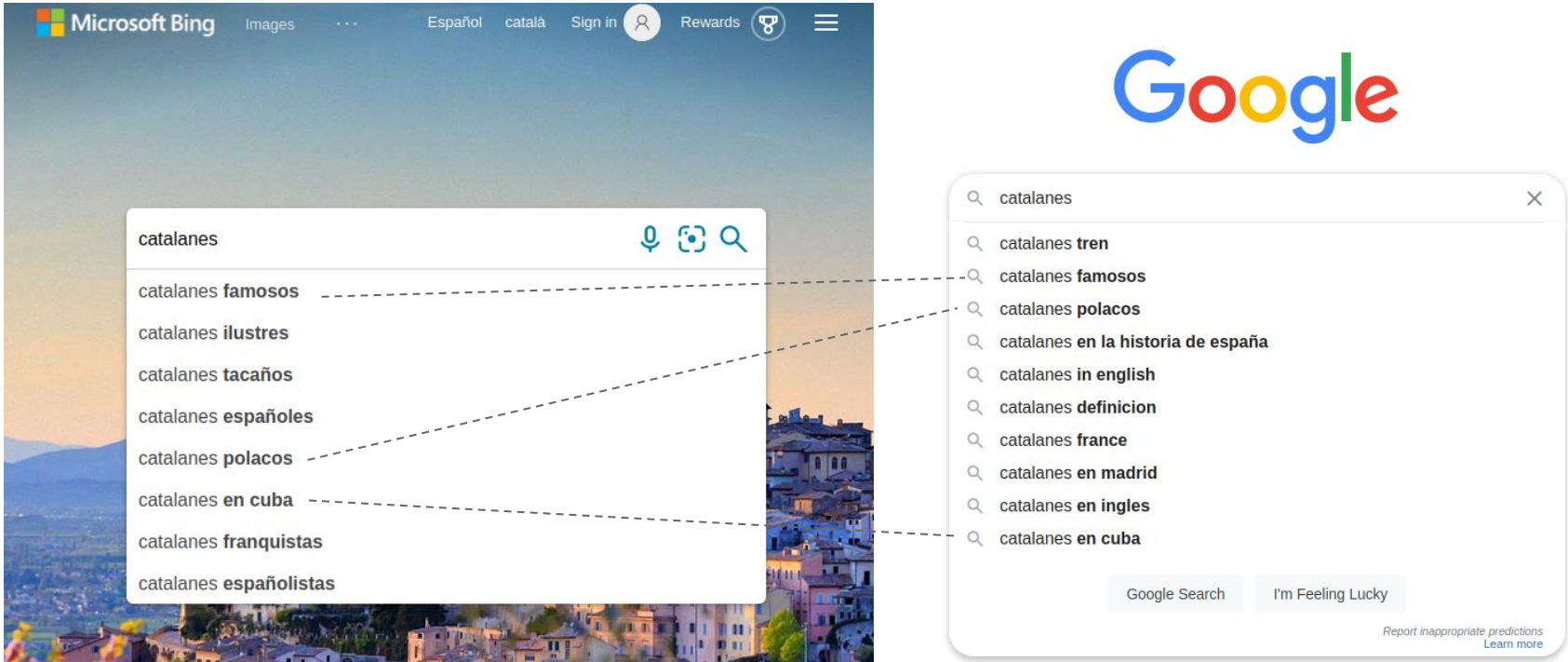
Búsqueda "black girls"

- Sexualized search results
  Google ca. 2013, "black women" but in general "(race) women"



Noble, S. U. (2018). Algorithms of Oppression: How search engines reinforce racism. NYU Press.
Crawford, K. (2017). The Trouble with Bias. Keynote at NIPS.

Representational harms (cont.)



Olteanu, A., Diaz, F., & Kazai, G. (2020). When Are Search Completion Suggestions Problematic? *Proc. of CSCW*.

¿Es esta una presencia suficiente?

	Position										top 10 male	top 10 female	top 40 male	top 40 female
	1	2	3	4	5	6	7	8	9	10				
Economista	f	m	m	m	m	m	m	m	m	m	90%	10%	73%	27%
Analista	f	m	f	f	f	f	f	m	f	f	20%	80%	43%	57%
Editor/a	m	m	m	m	m	m	f	m	m	m	90%	10%	73%	27%

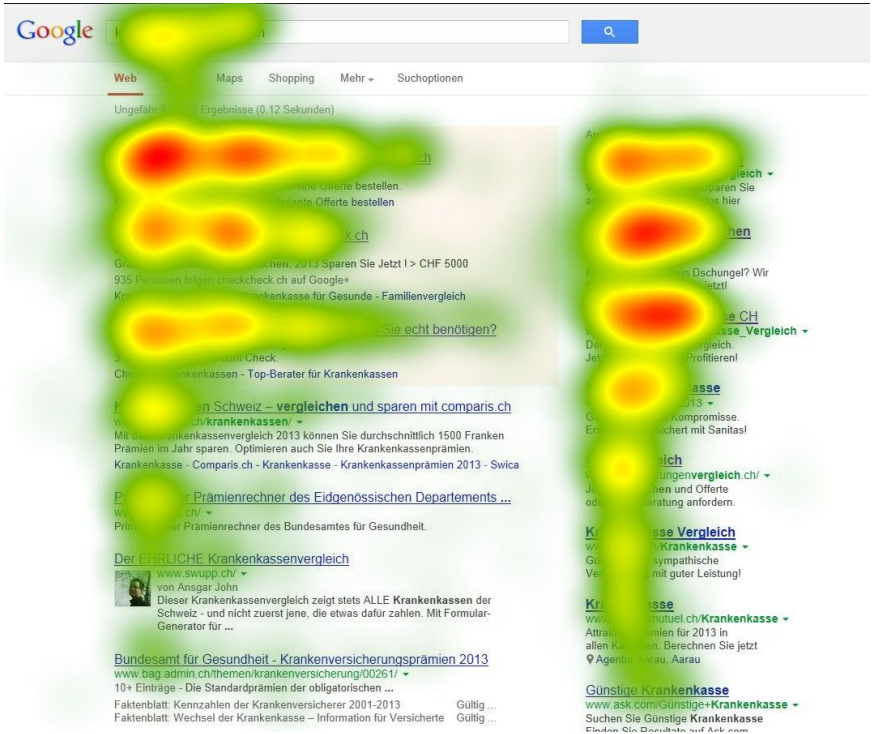
Primeros 10 resultados en un buscador de empleo

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., & Baeza-Yates, R. (2017). FA*IR: A fair top-k ranking algorithm. In Proc. of the ACM on Conference on Information and Knowledge Management (pp. 1569-1578). ACM.

Presencia = exposición

Cada posición en un ranking tiene una cierta visibilidad

Un ranking es paritario si da la misma visibilidad a distintos grupos



Singh, A., & Joachims, T. (2018). Fairness of Exposure in Rankings. In Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2219-2228). ACM.

Ejemplo: búsqueda de empleo

SPAIN			FRANCE			UNITED KINGDOM		
QUERY	K=16		QUERY	K=16		QUERY	K=15	
	LINKEDIN	VIADEO		LINKEDIN	VIADEO		LINKEDIN	VIADEO
	P	P		P	P		P	P
abogado			avocat			lawyer		0,20
arquitecto		0,30	architecte	0,80	0,60	architect	0,70	0,30
bombero		0,20	pompier		0,70	firefighter	0,40	
cartero	0,30	0,20	mailman		0,50	postman	0,20	0,20
científico	0,10	0,30	scientifique	0,70	0,80	scientist	0,50	0,60
cirujano	0,40	0,70	chirurgien		0,50	surgeon		0,30
cocinero	0,10	0,50	cuisinier	0,40	0,80	chef	0,40	0,40
consultor	0,50		consultant	0,20	0,40	consultant	0,60	0,30
dentista	0,90	0,50	dentiste		0,50	dentist	0,50	0,60
desarrollador	0,10	0,30	développeur	0,40	0,40	developer	0,60	0,40
diseñador	0,20	0,40	designer	0,50		designer	0,70	
economista	0,30	0,60	économiste	0,40	0,90	economist	0,60	0,30
AVERAGE	0,26	0,35	AVERAGE	0,40	0,59	AVERAGE	0,51	0,41

Gran diferencia en proporción de mujeres en empleos, países, buscadores

(Además: el tratamiento del masculino neutro en francés y castellano es inconsistente)

Sara Galindo: "Evaluating potential biases in commercial people search engines". MSc Thesis, UPF, July 2019.
Data: https://github.com/sgalinma/job-search-discrimination-data

Conclusiones

Conclusiones

Existen herramientas estadísticas para medir discriminación

Estas herramientas requieren datos de:

- Entradas: qué datos se utilizan
- Grupos: a qué grupo(s) pertenece cada persona
- Predicciones: qué predice el sistema
- Resultados: qué sucede realmente

License

Creative Commons BY-SA-4.0