Justicia social, algoritmos y aprendizaje automático en la era del 'Big Data'

Javier Sánchez-Monedero
sanchez-monederoj@cardiff.ac.uk

10 octubre 2018

Cardiff University, UK

Data Justice Lab

erc

datajusticelab.org

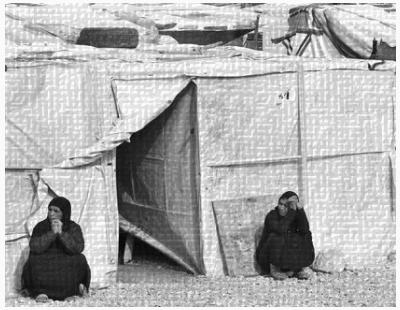
"Nos va a pillar la distopía justo cuando se llevan los pantalones tobilleros y qué bochorno, amigas", @Doble_Malta

Momento

Capitalismo de vigilancia

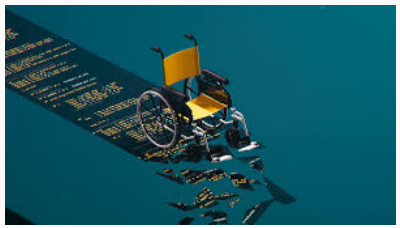
- Big data y el **capitalismo de vigilancia** como forma de gobierno Dencik et al. [2016]
- **Normalización de la recolección de datos y la cultura de la vigilancia**.
- Respuesta pública y de la sociedad civil: resignación, realismo de la vigilancia Dencik [2018].
- Respuesta desde el activismo: privacidad, seguridad, activismo de datos, análisis cualitativos sobre relaciones de poder.

La sociedad convertida en datos



¿Refugiado o terrorista? IBM puede tener la respuesta. Defense One

Cuando tu jefe es un algoritmo. Financial Times



¿Qué pasa cuando un algoritmo te corta la ayuda social?. The Verge.

Big Data

El **Big Data** como fenómeno cultural, tecnológico y académico Boyd and Crawford [2012]:

- Tecnología: recolectar y analizar grandes conjuntos de datos
- Análisis: identificar patrones para extraer conclusiones sociales, médicas, económicas...
- Mitología: los grandes datos ofrecen una forma superior de inteligencia y conocimiento

Big Data

El **Big Data** como fenómeno cultural, tecnológico y académico Boyd and Crawford [2012]:

- Tecnología: recolectar y analizar grandes conjuntos de datos
- Análisis: identificar patrones para extraer conclusiones sociales, médicas, económicas...
- Mitología: los grandes datos ofrecen una forma superior de inteligencia y conocimiento

Cuestiones críticas:

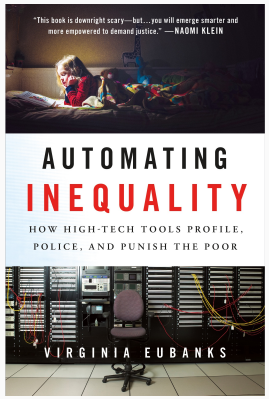
- Cambio en la definición de conocimiento
- Más grande no tiene por qué ser mejor: Ej. sociólogos de Twitter
- Porque se pueda hacer, no significa que se deba

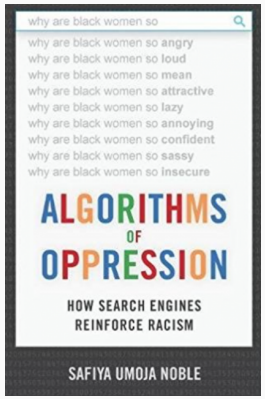
Respuestas

Más extendida: debate sobre eficiencia vs seguridad, privacidad individual y protección de datos.

Temas emergentes en torno al procesamiento de datos:

- Política predictiva (y especulativa)



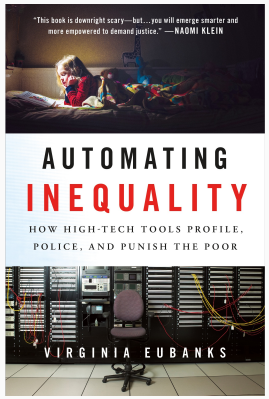


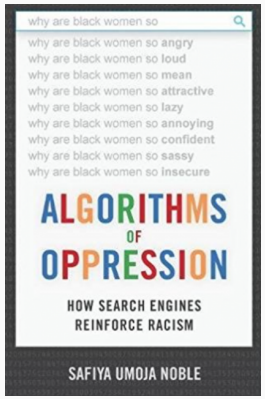
Respuestas

Más extendida: debate sobre eficiencia vs seguridad, privacidad individual y protección de datos.

Temas emergentes en torno al procesamiento de datos:

- Política predictiva (y especulativa)
- Clasificación y orden social de las personas



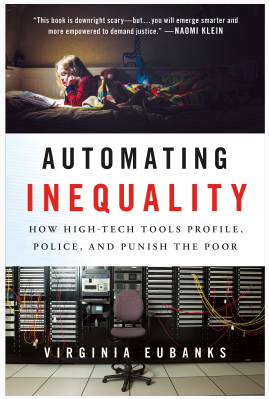


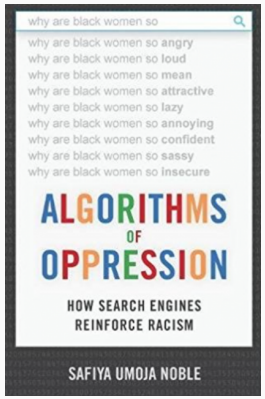
Respuestas

Más extendida: debate sobre eficiencia vs seguridad, privacidad individual y protección de datos.

Temas emergentes en torno al procesamiento de datos:

- Política predictiva (y especulativa)
- Clasificación y orden social de las personas
- Relaciones asimétricas de poder (creadores de perfiles vs sujetos que producen datos)



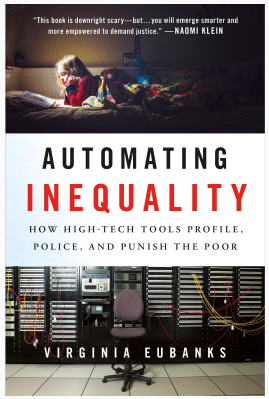


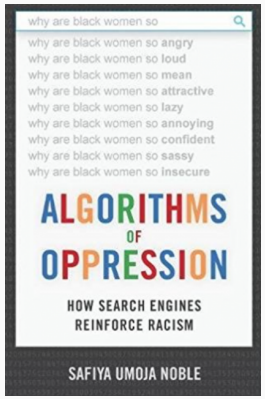
Respuestas

Más extendida: debate sobre eficiencia vs seguridad, privacidad individual y protección de datos.

Temas emergentes en torno al procesamiento de datos:

- Política predictiva (y especulativa)
- Clasificación y orden social de las personas
- Relaciones asimétricas de poder (creadores de perfiles vs sujetos que producen datos)
- Procesos de discriminación Guerrero Martín [2018]



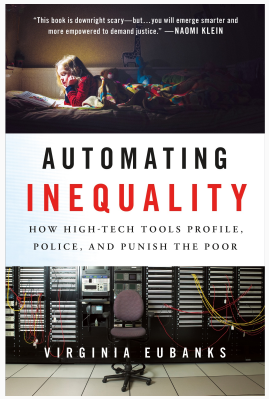


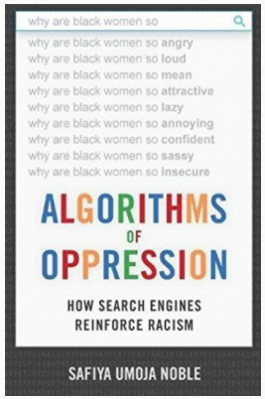
Respuestas

Más extendida: debate sobre eficiencia vs seguridad, privacidad individual y protección de datos.

Temas emergentes en torno al procesamiento de datos:

- Política predictiva (y especulativa)
- Clasificación y orden social de las personas
- Relaciones asimétricas de poder (creadores de perfiles vs sujetos que producen datos)
- Procesos de discriminación Guerrero Martín [2018]
- Procesos de exclusión: del "consumo luego existo" al "genero datos, luego existo" Lerman [2013]





Respuestas emergentes

Respuestas emergentes

- **Soluciones técnicas** a sesgos, discriminación, auditoría y transparencia de los sistemas de Big Data e IA.
- **Ética de datos**: solucionismo tecnológico, formación en ética, guías y principios

Respuestas emergentes

Respuestas emergentes

- **Soluciones técnicas** a sesgos, discriminación, auditoría y transparencia de los sistemas de Big Data e IA.
- **Ética de datos**: solucionismo tecnológico, formación en ética, guías y principios

¿Neutralización (¡despolicitación!) de los problemas?

Aprendizaje automático

Machine Learning

what society thinks I do

what my friends think I do

what my parents think I do

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i$$

$$\alpha_i \geq 0, \forall i$$

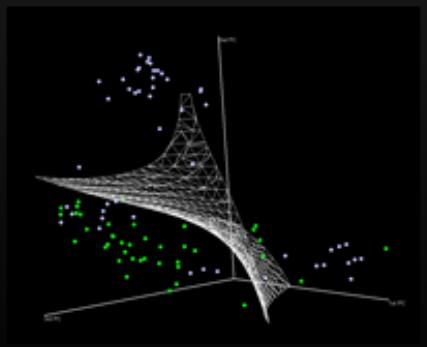
$$\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\nabla \hat{g}(\theta_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t).$$

$$\theta_{t+1} = \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t)$$

$$\mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] = \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t).$$

what other programmers think I do



what I think I do

```
>>> from sklearn import svm
```

what I really do


MACHINE LEARNING
ENTRAR Y SALIR. UNA AVENTURA DE 20 MINUTOS
imgflip.com

Resumen rápido del 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.

Resumen rápido del 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:

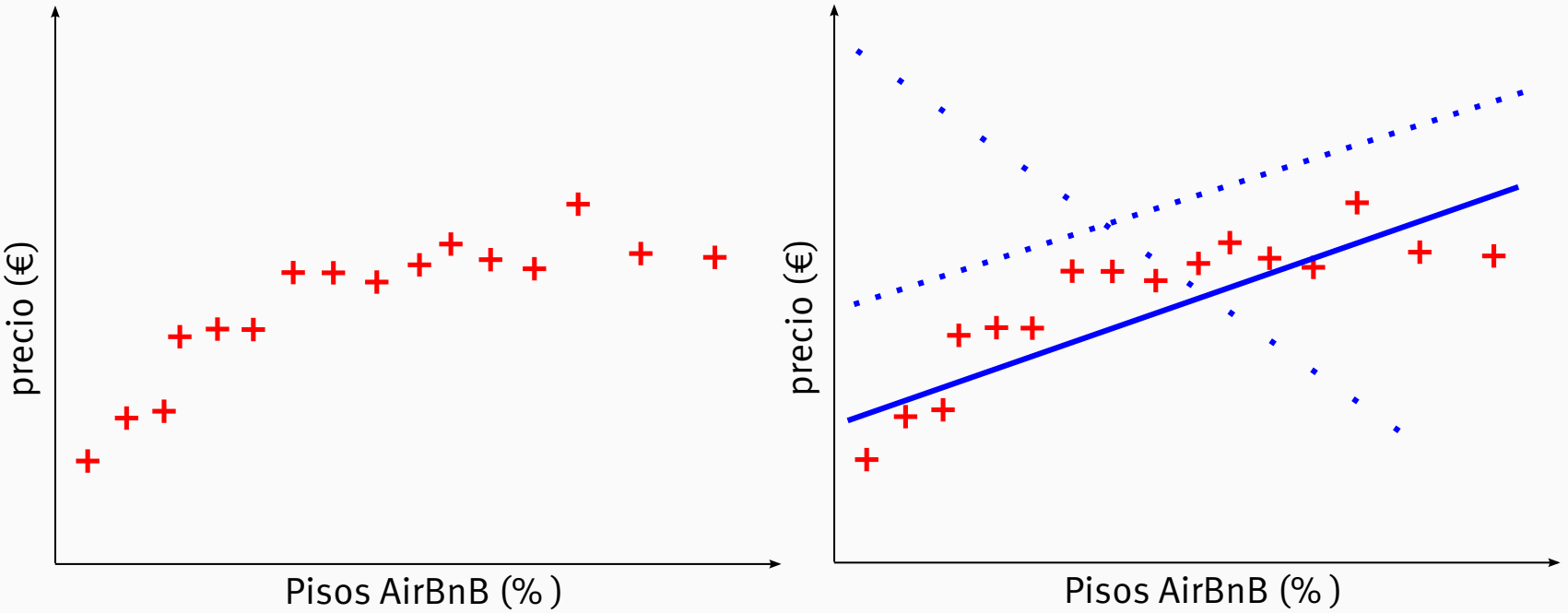
Aprender de los ejemplos:

```
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 que es justo, no discrimina, está libre de sesgos humanos, etc. NOTA: además el código es por lo general difícil o imposible de auditar

¿Qué es cambiar el modelo?



Modelo de predicción:

$$y^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \varepsilon^{(i)}$$

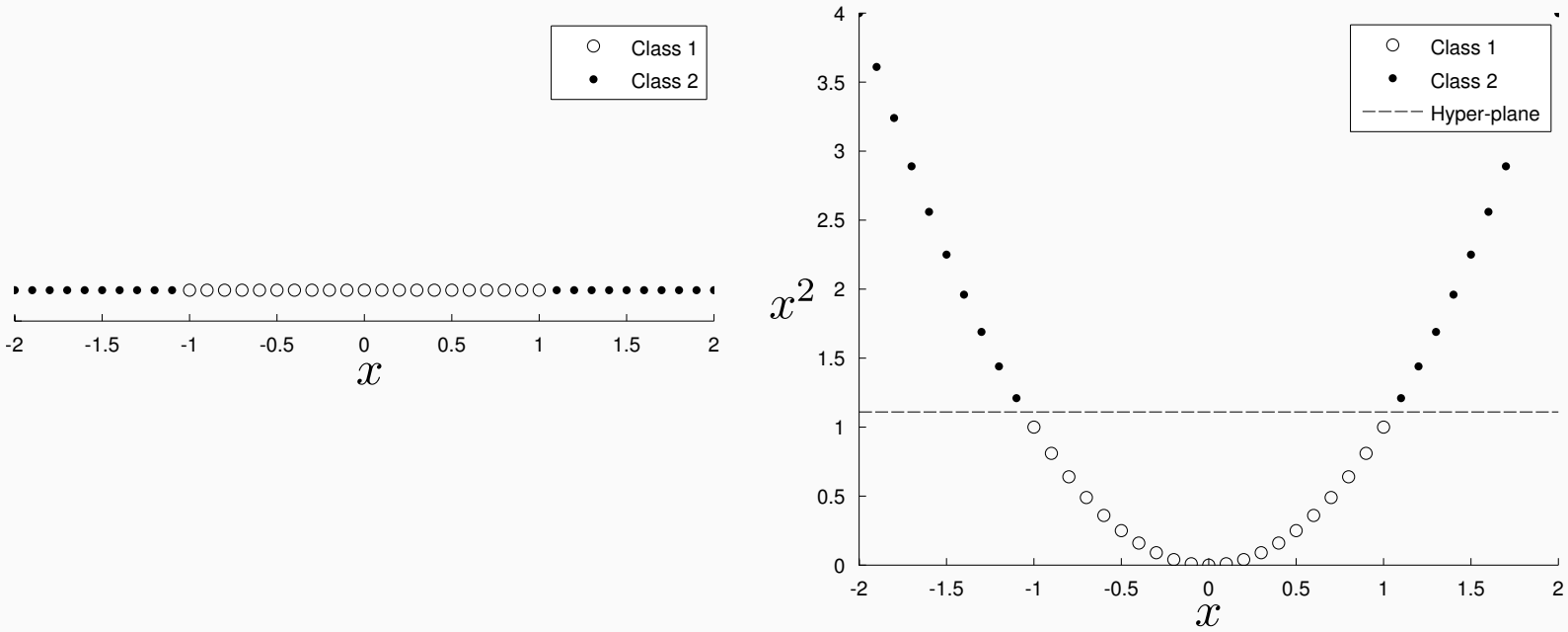
Error:

$$\varepsilon^{(i)} = \hat{y}^{(i)} - y^{(i)}$$

Función objetivo (encontrar β_0 y β_1 que minimicen el error):

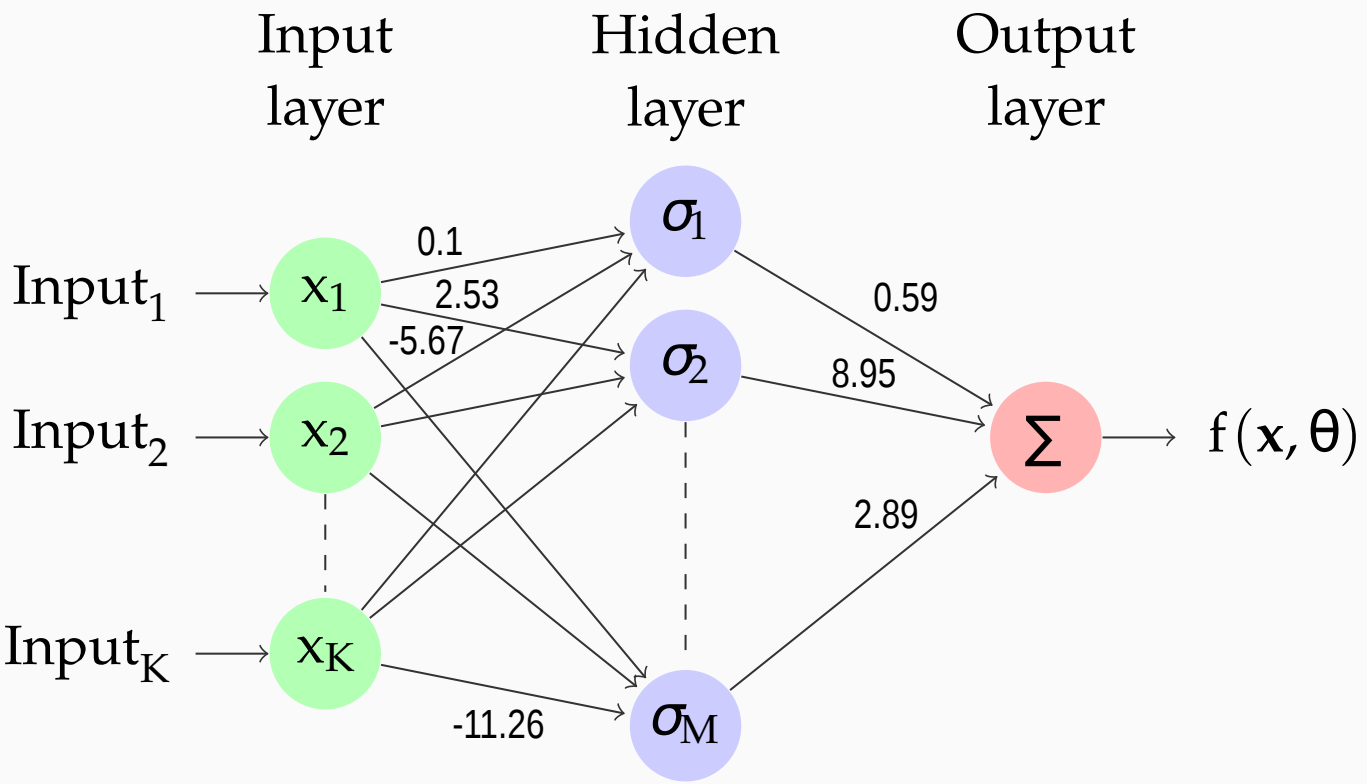
$$F = \frac{1}{n} \sum_{i=1}^{n} (\varepsilon^{(i)})^2$$

Ejemplo básico de transformación de datos



Muchos métodos de AM *aprenden* transformaciones no lineales para facilitar la separación de los ejemplos de cada clase.

La caja negra



¿Cómo evaluamos un clasificador?

	etiqueta real (observación)	
etiqueta predicha	TP (true positive) Acierto	FP (false positive) Error tipo I
	FN (false negative) Error tipo II	TN (true negative) Acierto

¿Cómo evaluamos un clasificador?

	etiqueta real (observación)	
etiqueta predicha	TP (true positive) Acierto 50	FP (false positive) Error tipo I 20
	FN (false negative) Error tipo II 5	TN (true negative) Acierto 100

Rendimiento global o precisión:

$$\text{Precisión} = \frac{TP + TN}{\text{total}} = (50 + 100)/175 = 0{,}86$$

¿Cómo evaluamos un clasificador?

	etiqueta real (observación)	
etiqueta predicha	TP (true positive) Acierto 50	FP (false positive) Error tipo I 20
	FN (false negative) Error tipo II 5	TN (true negative) Acierto 100

Rendimiento global o precisión:

$$\text{Precisión} = \frac{TP + TN}{\text{total}} = (50 + 100)/175 = 0{,}86$$

¿Cuántas veces acierta la clase positiva?:

$$\text{Ratio Verdaderos Positivos} = TP/\text{casos positivos} = 50/55 = 0{,}91$$

¿Cómo evaluamos un clasificador?

	etiqueta real (observación)	
etiqueta predicha	TP (true positive) Acierto 50	FP (false positive) Error tipo I 20
	FN (false negative) Error tipo II 5	TN (true negative) Acierto 100

Rendimiento global o precisión:

$$\text{Precisión} = \frac{TP + TN}{\text{total}} = (50 + 100)/175 = 0{,}86$$

¿Cuántas veces acierta la clase positiva?:

$$\text{Ratio Verdaderos Positivos} = TP/\text{casos positivos} = 50/55 = 0{,}91$$

¿Cuántas veces dice que hay un caso positivo y falla?

$$\text{Ratio Falsos Positivos} = FP/\text{casos negativos} = 20/120 = 0{,}17$$

¿Cómo evaluamos un clasificador?

	etiqueta real (observación)	
etiqueta predicha	TP (true positive) Acierto 50	FP (false positive) Error tipo I 20
	FN (false negative) Error tipo II 5	TN (true negative) Acierto 100

Rendimiento global o precisión:

$$\text{Precisión} = \frac{TP + TN}{\text{total}} = (50 + 100)/175 = 0{,}86$$

¿Cuántas veces acierta la clase positiva?:

$$\text{Ratio Verdaderos Positivos} = TP/\text{casos positivos} = 50/55 = 0{,}91$$

¿Cuántas veces dice que hay un caso positivo y falla?

$$\text{Ratio Falsos Positivos} = FP/\text{casos negativos} = 20/120 = 0{,}17$$

¿Cuántas veces dice que hay un caso negativo y falla?

$$\text{Ratio Falsos Negativos} = FN/FN + TP = 5/55 = 0{,}09$$

Algunas tareas de aprendizaje máquina

Si lo pensamos, todas estas tareas están en nuestras vidas:

- Predicción (clasificación/regresión)
- Motores de búsqueda y recomendación
- Aprendizaje de reglas de asociación
- Ordenación, ranking, puntuación, etc.
- Agrupamiento, redes generativas, etc.
- Procesamiento de lenguaje natural
- Algunos métodos de visualización
- Etiquetado semántico, comprensión de escenas, etc. sobre contenido multimedia
- ...

Estado del arte

Generated story about image
Model: Romantic Novels

"He was a shirtless man in the back of his mind, and I let out a curse as he leaned over to kiss me on the shoulder.

He wanted to strangle me, considering the beatiful boy I'd become wearing his boxers."

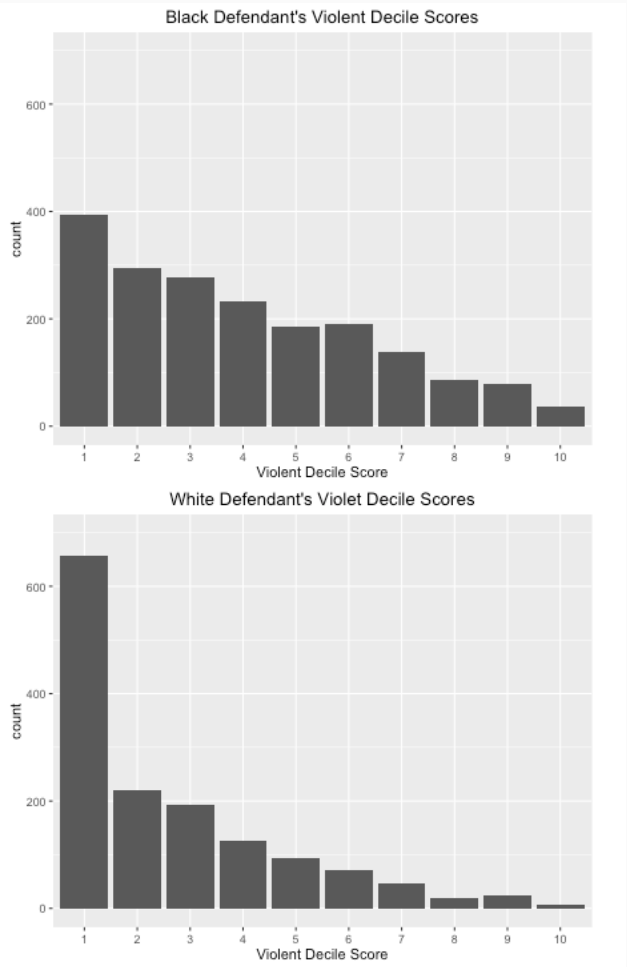
Fuente Generating Stories about Images. Código en
`https://github.com/ryankiros/neural-storyteller`

Las máquinas reproducen prejuicios

Cómo aprenden a discriminar las máquinas

Algunas causas de discriminación (ver Barocas, Solon; Selbst, Andrew D [2016]):

- Muestra sesgada



Source Jeff Larson [2016]

Cómo aprenden a discriminar las máquinas

Algunas causas de discriminación (ver Barocas, Solon; Selbst, Andrew D [2016]):

- Muestra sesgada
- Muestra contaminada

Aprender a predecir decisiones sobre contratación, préstamos, etc.

Cómo aprenden a discriminar las máquinas

Algunas causas de discriminación (ver Barocas, Solon; Selbst, Andrew D [2016]):

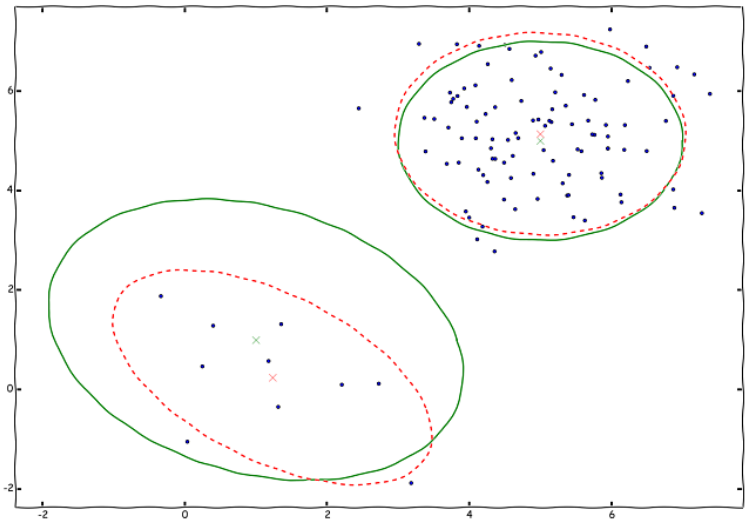
- Muestra sesgada
- Muestra contaminada
- Variables limitadas

¿Se han recogido con la misma fiabilidad todas las variables para todos los grupos? Ejemplo: "Los migrantes van más a urgencias"

Cómo aprenden a discriminar las máquinas

Algunas causas de discriminación (ver Barocas, Solon; Selbst, Andrew D [2016]):

- Muestra sesgada
- Muestra contaminada
- Variables limitadas
- Diferentes tamaños muestrales



Fuente How big data is unfair

Cómo aprenden a discriminar las máquinas

Algunas causas de discriminación (ver Barocas, Solon; Selbst, Andrew D [2016]):

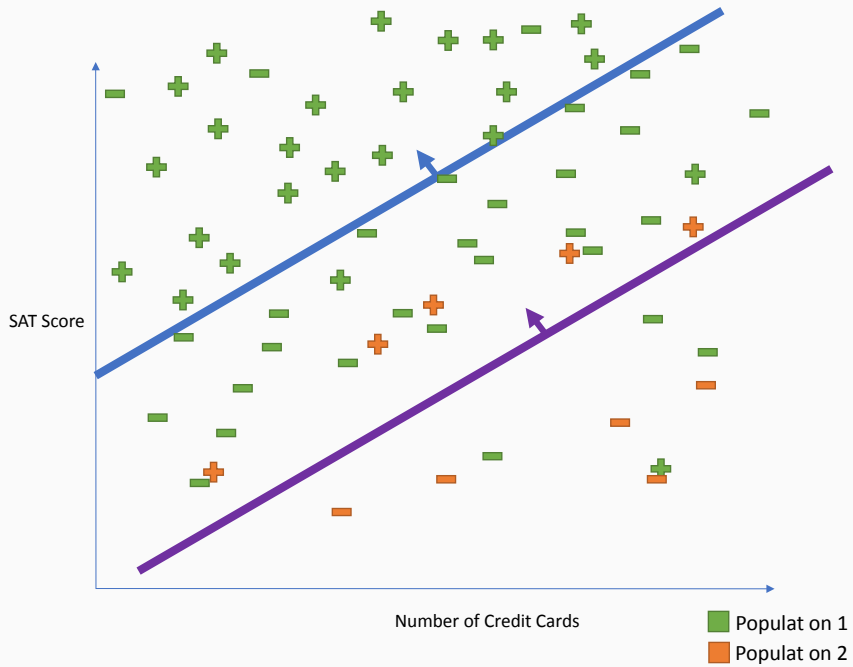
- Muestra sesgada
- Muestra contaminada
- Variables limitadas
- Diferentes tamaños muestrales
- Variables proxy

{Código postal, salario} está correlado con colectivos racializados

Cómo aprenden a discriminar las máquinas

Algunas causas de discriminación (ver Barocas, Solon; Selbst, Andrew D [2016]):

- Muestra sesgada
- Muestra contaminada
- Variables limitadas
- Diferentes tamaños muestrales
- Variables proxy
- Diferente comportamiento de las variables para cada (sub)grupo



Fuente Roth [2018]

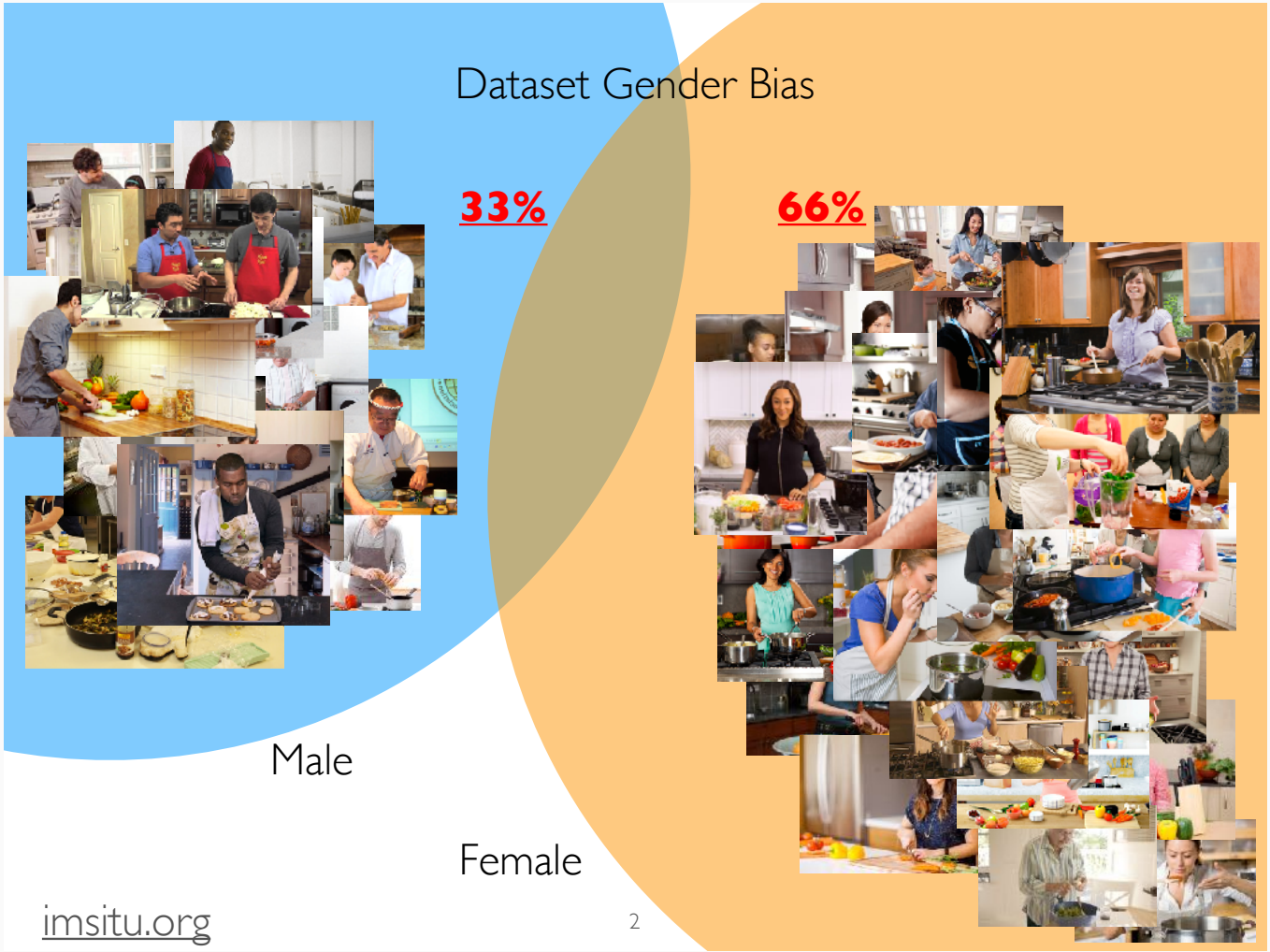
Reproducción y amplificación de prejuicios

Fuente Lum and Isaac [2016]

Los **bucles de retroalimentación** pueden reproducir y amplificar los prejuicios Barocas and Hardt [2017], Ensign et al. [2017], ejemplo PredPol:

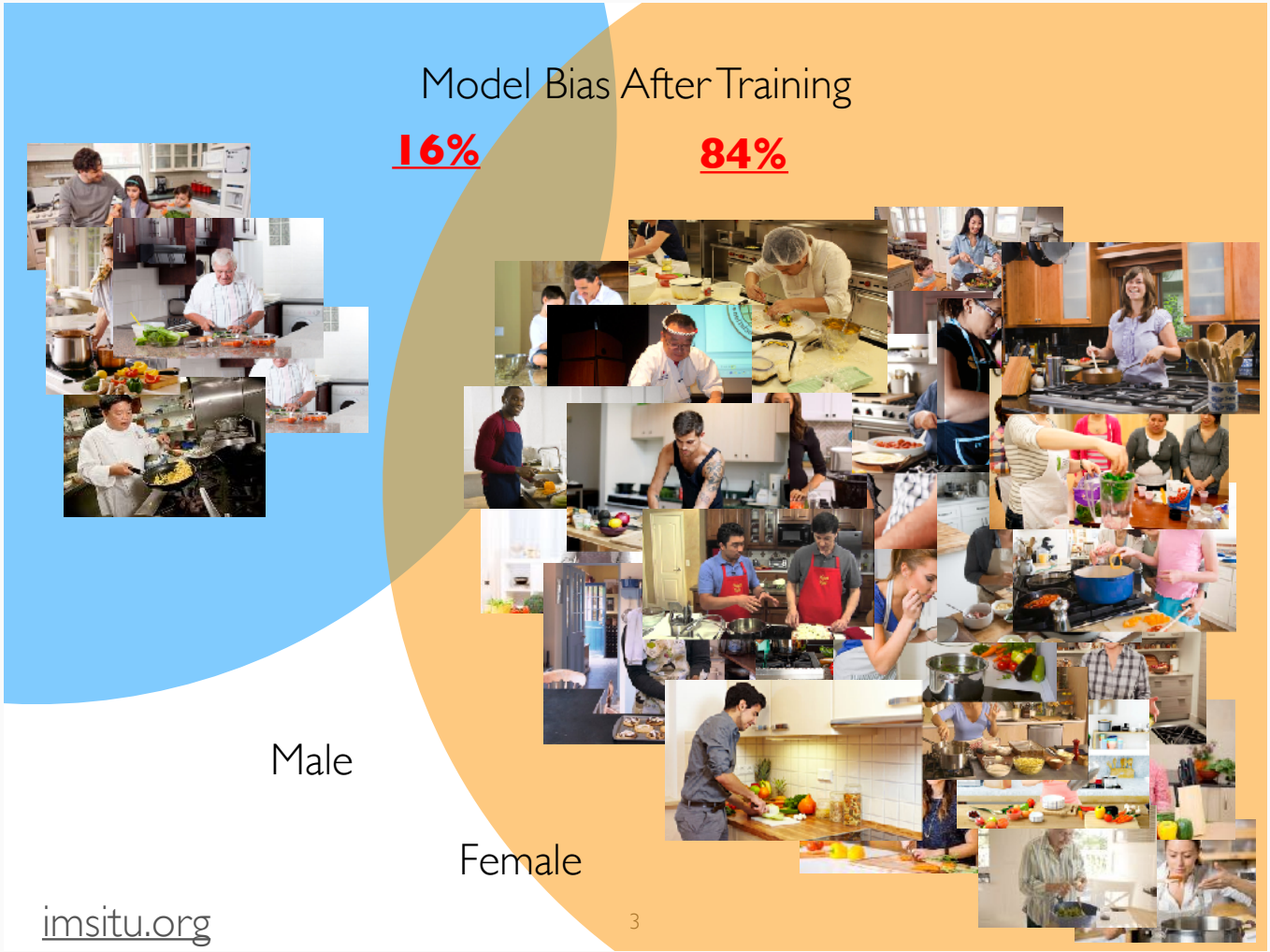
- La predicción de crimen en un área enviará recursos policiales a ese área
- Los eventos encontrados se añaden a la base de datos
- Es menos probable que se observen eventos que contradigan las predicciones

Amplificación de prejuicios i



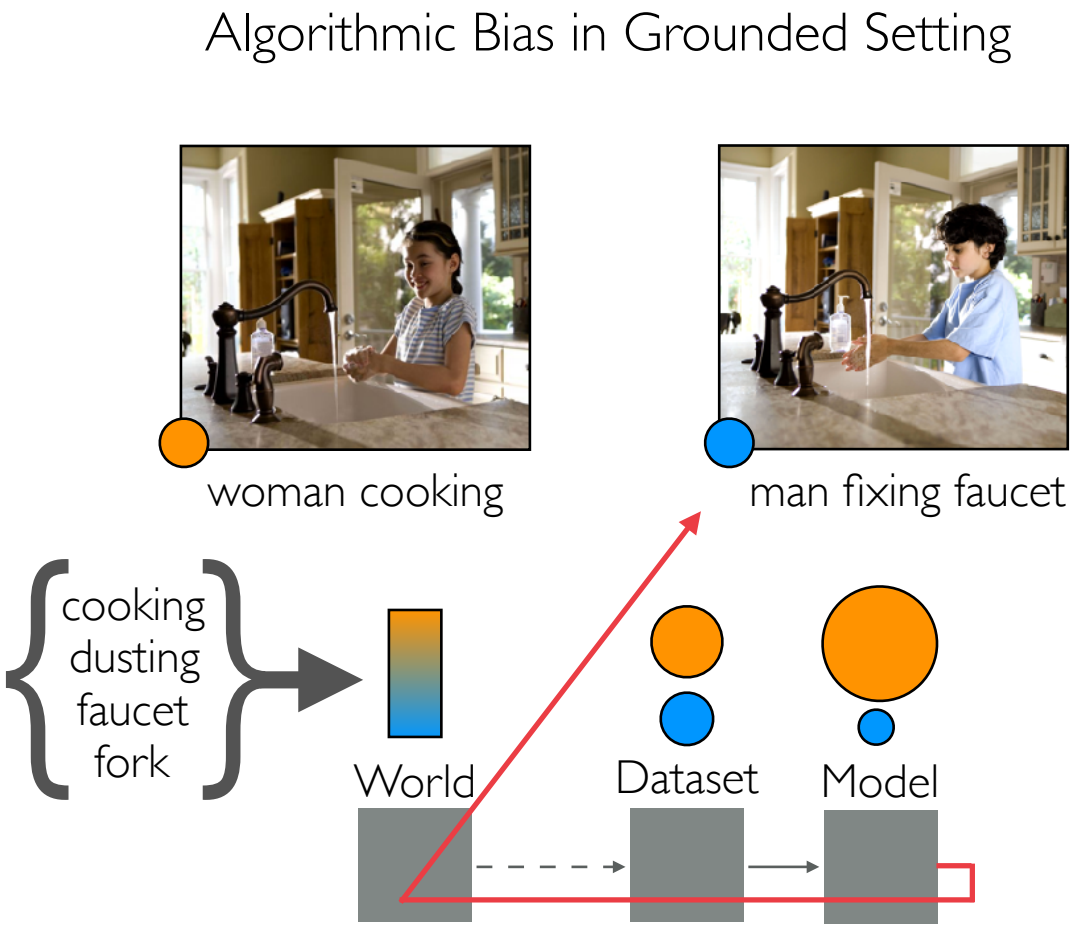
Fuente Zhao et al. [2017]

Amplificación de prejuicios ii



Fuente Zhao et al. [2017]

Amplificación de prejuicios iii



Fuente Zhao et al. [2017]

¿Cómo medimos la discriminación?

Cómo evaluar la *imparcialidad*:

- Interpretación de los modelos (Lectura recomendada Lipton [2017])

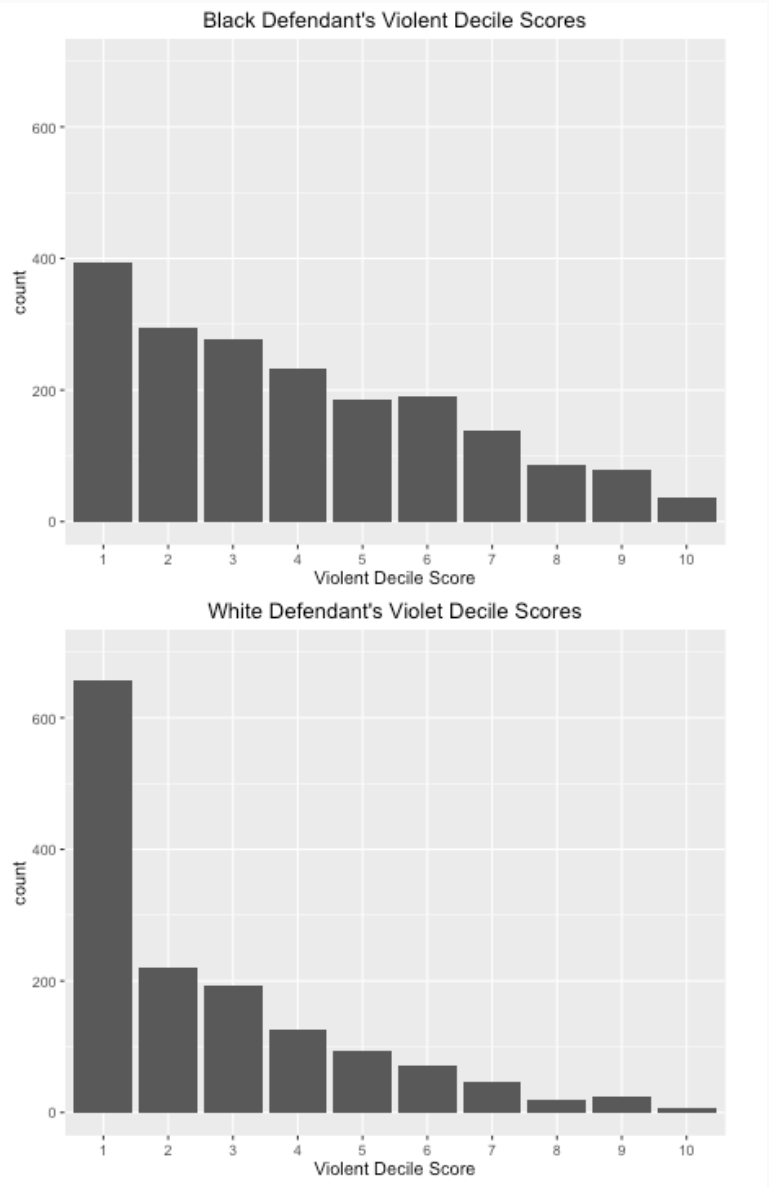
Risk of Violent Recidivism Logistic Model

	Dependent variable:
	Score (Low vs Medium and High)
Female	-0.729*** (0.127)
Age: Greater than 45	-1.742*** (0.184)
Age: Less than 25	3.146*** (0.115)
Black	0.659*** (0.108)
Asian	-0.985 (0.705)
Hispanic	-0.064 (0.191)
Native American	0.448 (1.035)
Other	-0.205 (0.225)
Number of Priors	0.138*** (0.012)
Misdemeanor	-0.164* (0.098)
Two Year Recidivism	0.934*** (0.115)
Constant	-2.243*** (0.113)
Observations	4,020
Akaike Inf. Crit.	3,022.779
*Note: *p<0.1; **p<0.05; ***p<0.01*	

¿Cómo medimos la discriminación?

Cómo evaluar la *imparcialidad*:

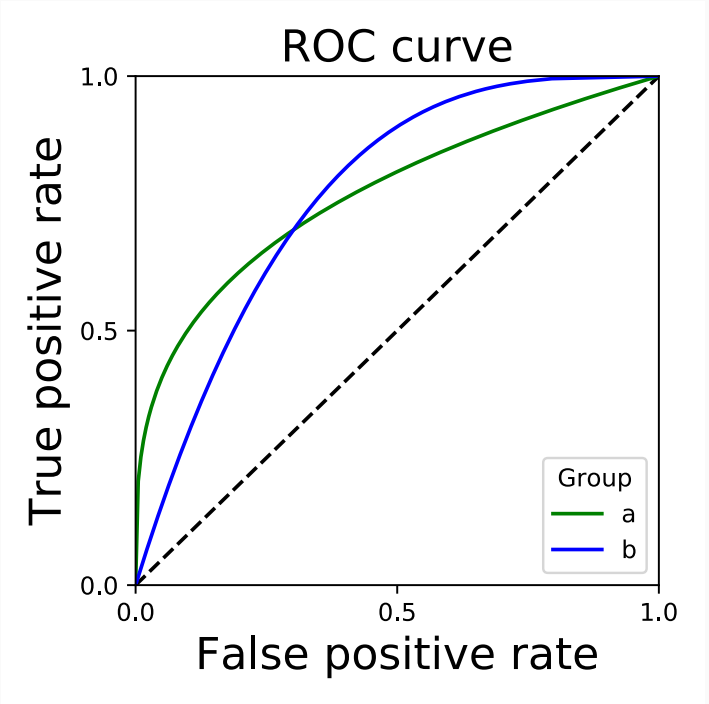
- Interpretación de los modelos (Lectura recomendada Lipton [2017])
- Análisis de los datos



¿Cómo medimos la discriminación?

Cómo evaluar la *imparcialidad*:

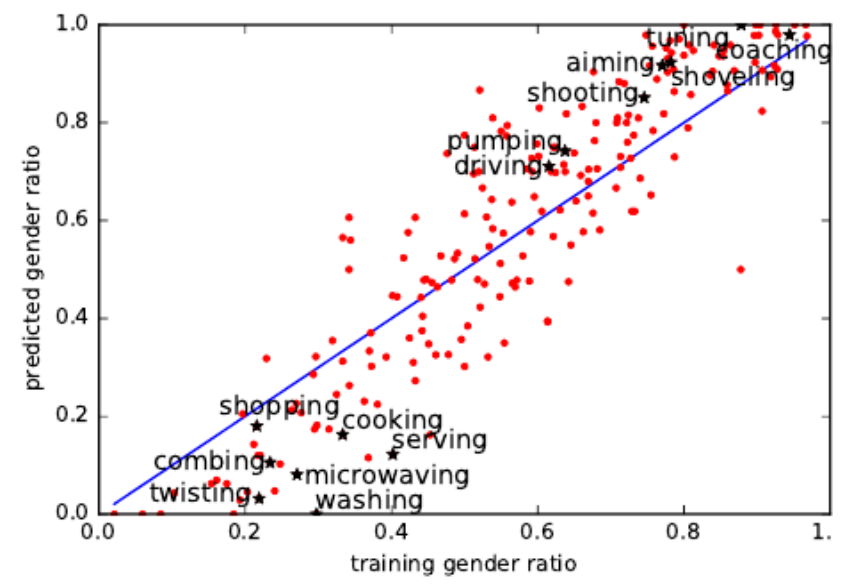
- Interpretación de los modelos (Lectura recomendada Lipton [2017])
- Análisis de los datos
- Análisis del comportamiento del modelo



¿Cómo medimos la discriminación?

Cómo evaluar la *imparcialidad*:

- Interpretación de los modelos (Lectura recomendada Lipton [2017])
- Análisis de los datos
- Análisis del comportamiento del modelo
- Evaluación del rendimiento del modelo respecto a los colectivos



¿Cómo medimos la discriminación?

Cómo evaluar la *imparcialidad*:

- Interpretación de los modelos (Lectura recomendada Lipton [2017])
- Análisis de los datos
- Análisis del comportamiento del modelo
- Evaluación del rendimiento del modelo respecto a los colectivos
- Descubrimiento de subgrupos (Zhang and Neill [2016])

¿Cómo medimos la discriminación?

Cómo evaluar la *imparcialidad*:

- Interpretación de los modelos (Lectura recomendada Lipton [2017])
- Análisis de los datos
- Análisis del comportamiento del modelo
- Evaluación del rendimiento del modelo respecto a los colectivos
- Descubrimiento de subgrupos (Zhang and Neill [2016])

pero al final... necesitamos un criterio

(Aaron Roth: "Weakly Meritocratic Fairness")

La discriminación no tiene un abordaje único

De el seminario de Barocas and Hardt [2017] en NIPS 2017, la discriminación:

- Es específica del dominio y depende del potencial impacto en las comunidades (marginalizadas)
- Es específica de cada variable. Ej. género binaro vs. no binario

Propuesta formal en la comunidad científica y tecnológica i

Tenemos las siguientes variables aleatorias en el mismo espacio probabilístico (Barocas and Hardt [2017]):

- X variables que describen a un individuo
- A atributo sensible (género, "raza"...)
- Y variable objetivo
- $C = f(X, A)$ predictor que estima Y

Función de verosimilitud de Y dados X y el atributo de grupo A:

$$P(Y|X, = x, A = a).$$

Propuesta formal en la comunidad científica y tecnológica ii

Muchas propuestas que se están haciendo (por ejemplo en FATML y FAT*ML) tratan de conseguir la independencia de C respecto a A (paridad estadística y otras cuestiones):

$$P(C = c|X, = x, A = a) \approx P(C = c|X, = x, A = b)$$

$$\frac{P(C = c|X, = x, A = a)}{P(C = c|X, = x, A = b)} > 0{,}8$$

Para más definiciones de requisitos para la ecuanimidad ver Barocas and Hardt [2017] y Roth [2018].

Algunas soluciones para 'arreglar' los clasificadores

Pre-procesamiento. Ej. ajustado o calibrado de variables Zemel et al. [2013]

Algunas soluciones para 'arreglar' los clasificadores

Pre-procesamiento. Ej. ajustado o calibrado de variables Zemel et al. [2013]

Post-procesamiento. Ej. calibrar umbrales de clasificación Hardt et al. [2016]

Algunas soluciones para 'arreglar' los clasificadores

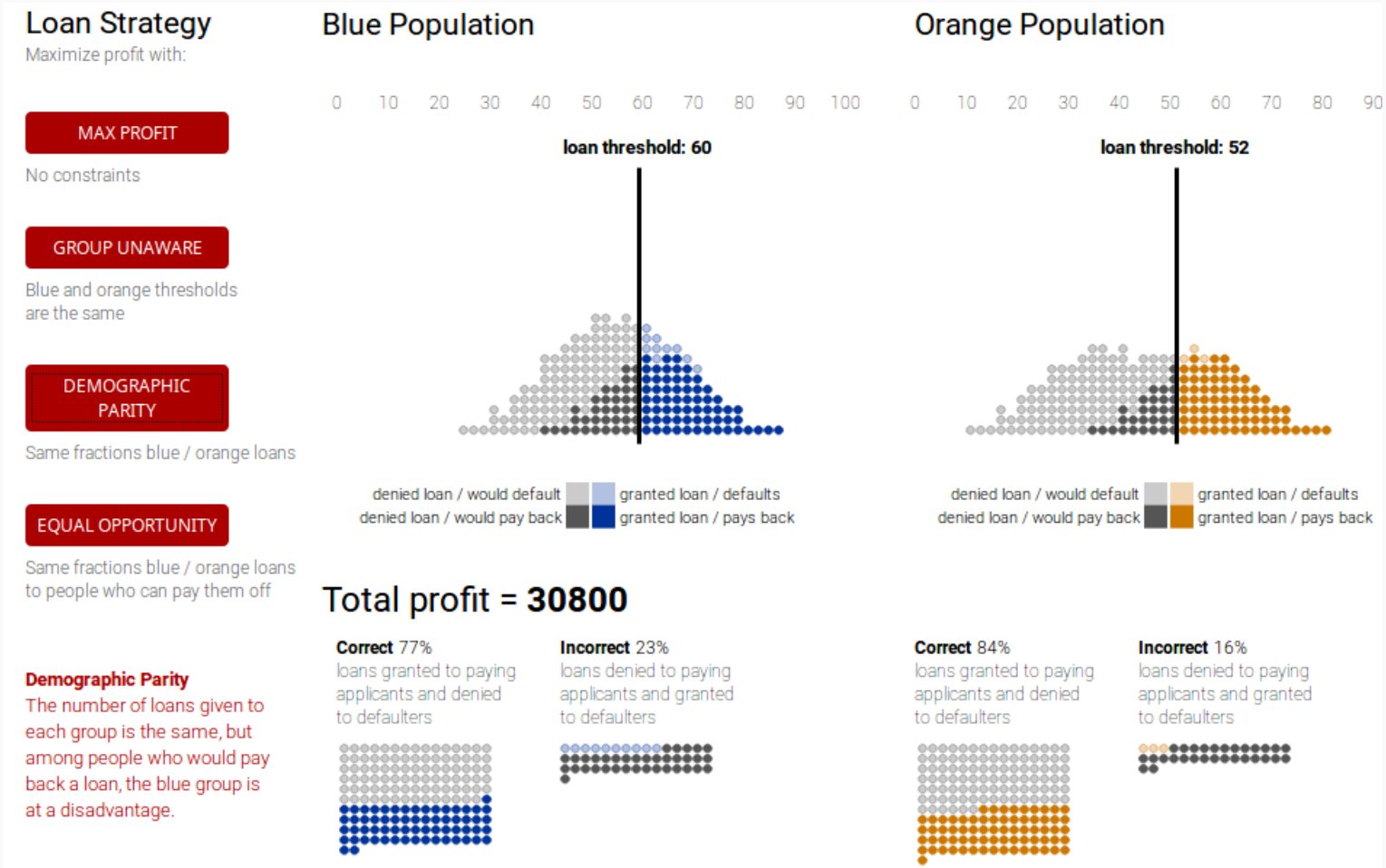
Pre-procesamiento. Ej. ajustado o calibrado de variables Zemel et al. [2013]

Post-procesamiento. Ej. calibrar umbrales de clasificación Hardt et al. [2016]

Algoritmo de aprendizaje. Ej. término de regularización en la función de error

Función de error = error etiquetas + (error grupo A - error grupo B) Y muchas más...

Calibrar umbrales



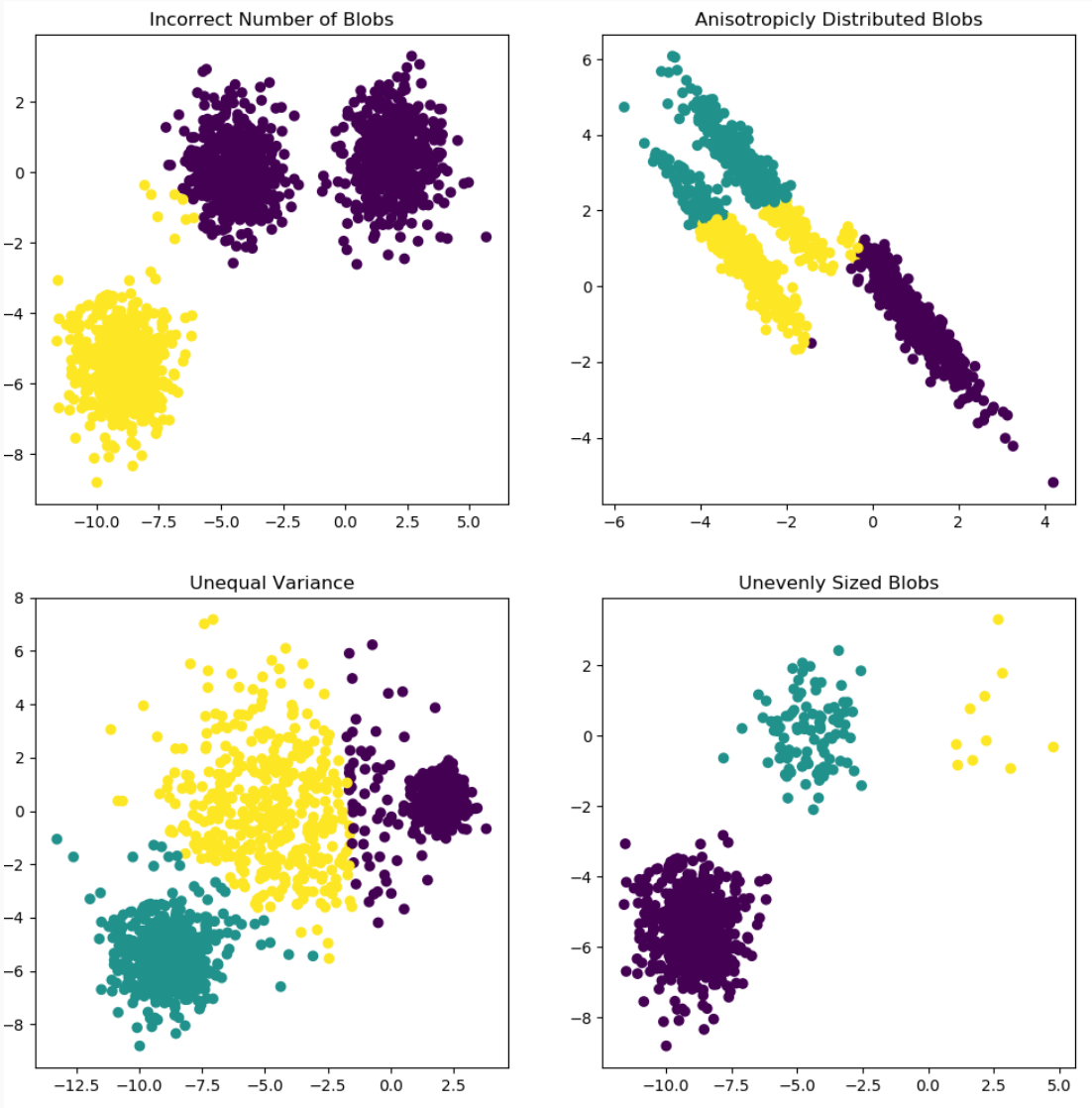
Fuente `http://research.google.com/bigpicture/attacking-discrimination-in-ml/`

Asunciones de algunos métodos

Al aplicar técnicas de aprendizaje automático hay que tomar algunas precauciones y siempre contrastar con los expertos:

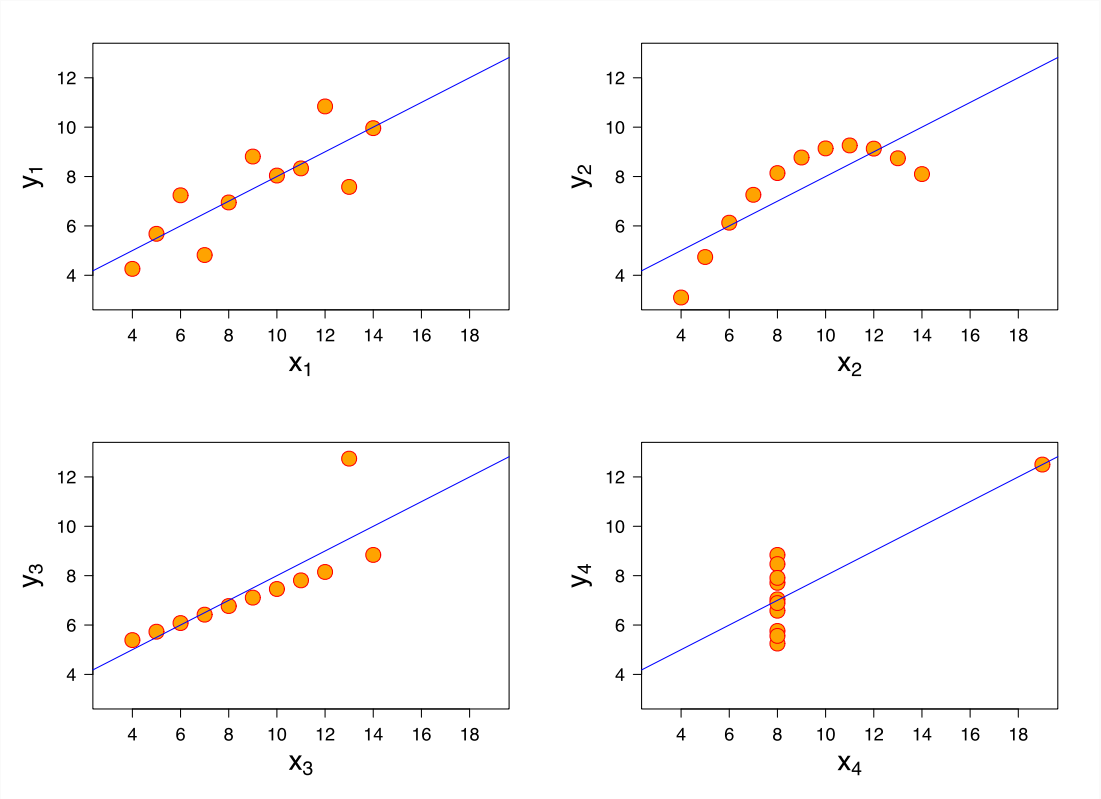
- **Función de error**: ¿Qué estamos optimizando realmente? Gürses et al. [2018]
- **Asunción de linealidad**, Ej., Modelo lineal generalizado, K-medias
- **Independencia** e **interacción** entre variables.
- ...

Asunciones del K-means



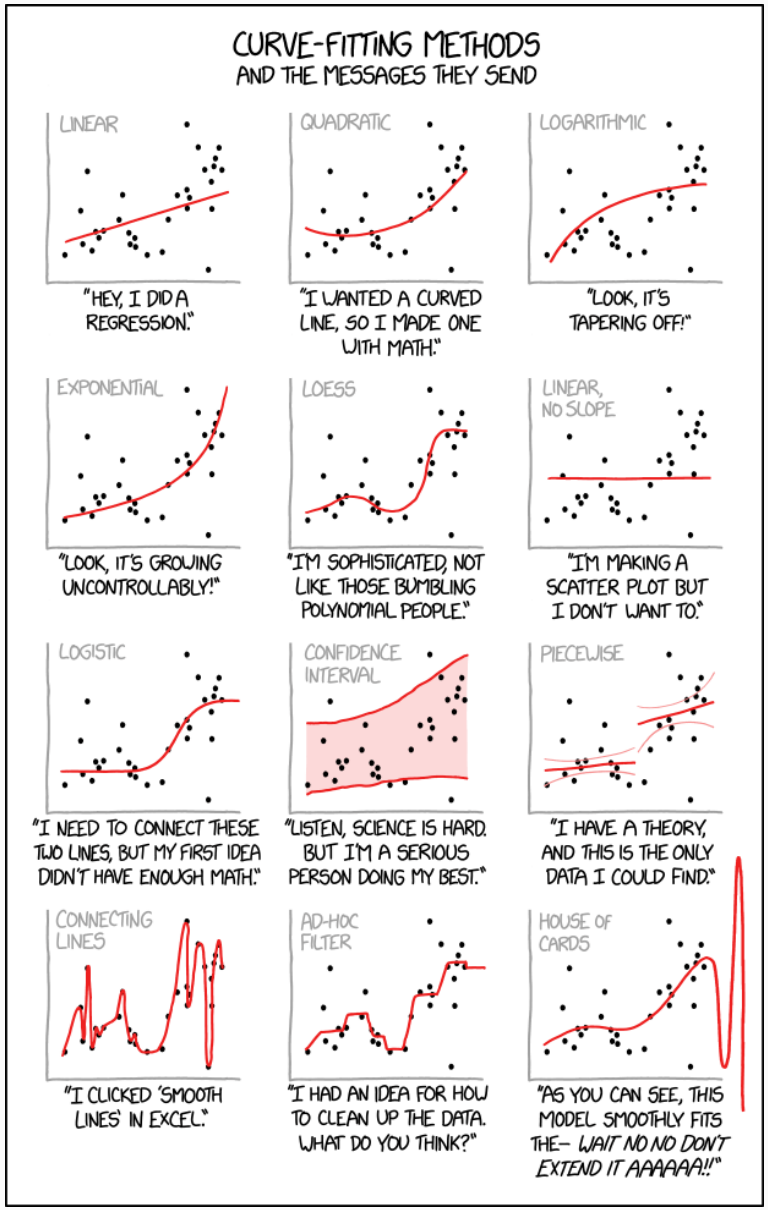
Fuente Documentation of scikit-learn

Datos distintos e igual modelo y descriptores estadísticos



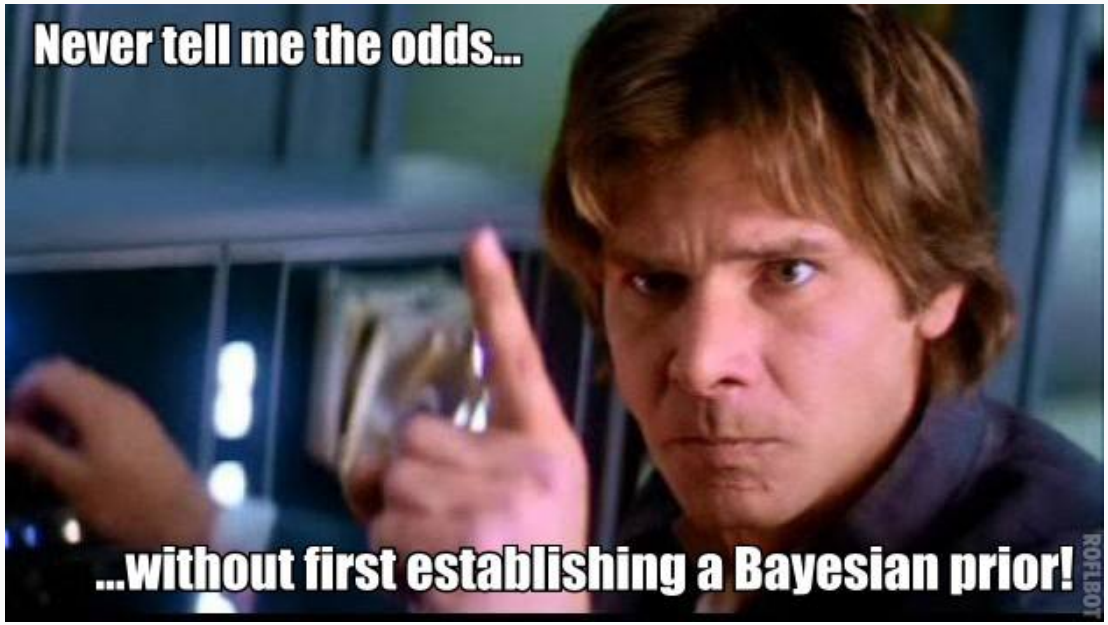
Fuente Wikipedia

Más precauciones



Fuente https://xkcd.com/

Consideraciones en la evaluación

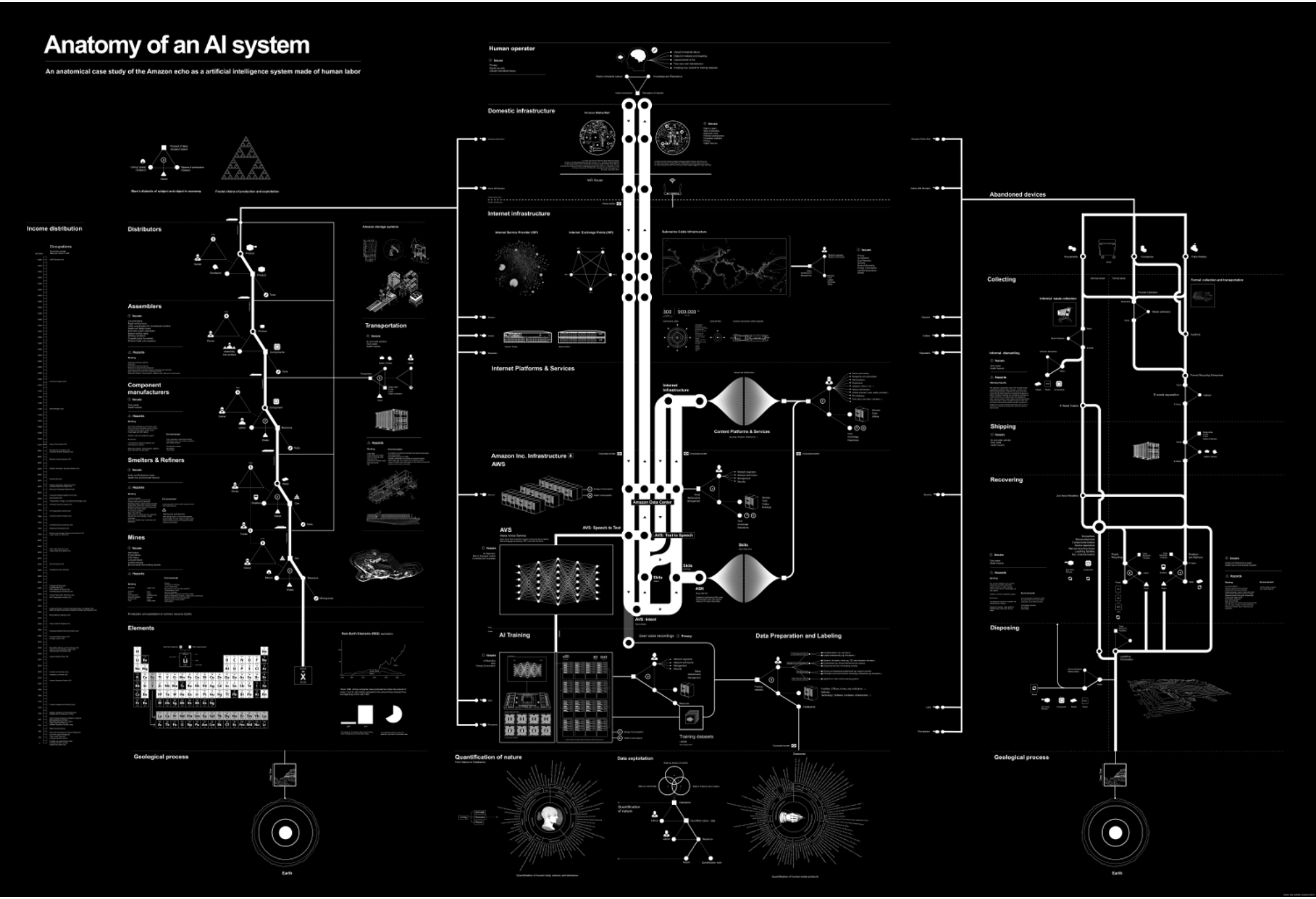


Fuente Han Solo and Bayesian Priors

Para evaluar un método en el contexto, seamos Bayesianos (atención a la Falacia de la frecuencia base).

Anatomy of an AI system

An anatomical case study of the Amazon echo as a artificial intelligence system made of human labor



Conclusiones y debate

Más preguntas

- Tenemos situaciones (técnicas) en las que “todo el mundo tiene razón”

Más preguntas

- Tenemos situaciones (técnicas) en las que "todo el mundo tiene razón"
- El aprendizaje estadístico (o aprendizaje máquina) siempre va a ser conservativo por definición

Más preguntas

- Tenemos situaciones (técnicas) en las que "todo el mundo tiene razón"
- El aprendizaje estadístico (o aprendizaje máquina) siempre va a ser conservativo por definición
- ¿Es imprescindible tratar diferente a los (sub) grupos? (con sus implicaciones respecto a privacidad)

Más preguntas

- Tenemos situaciones (técnicas) en las que "todo el mundo tiene razón"
- El aprendizaje estadístico (o aprendizaje máquina) siempre va a ser conservativo por definición
- ¿Es imprescindible tratar diferente a los (sub) grupos? (con sus implicaciones respecto a privacidad)
- ¿Puede haber sistemas de reconocimiento facial justos?

Más preguntas

- Tenemos situaciones (técnicas) en las que "todo el mundo tiene razón"
- El aprendizaje estadístico (o aprendizaje máquina) siempre va a ser conservativo por definición
- ¿Es imprescindible tratar diferente a los (sub) grupos? (con sus implicaciones respecto a privacidad)
- ¿Puede haber sistemas de reconocimiento facial justos?
- ¿Se debe o no construir?¿Cómo hacerlo?

Más preguntas

- Tenemos situaciones (técnicas) en las que "todo el mundo tiene razón"
- El aprendizaje estadístico (o aprendizaje máquina) siempre va a ser conservativo por definición
- ¿Es imprescindible tratar diferente a los (sub) grupos? (con sus implicaciones respecto a privacidad)
- ¿Puede haber sistemas de reconocimiento facial justos?
- ¿Se debe o no construir?¿Cómo hacerlo?
- Pertenencia a grupo no binarias

Más preguntas

- Tenemos situaciones (técnicas) en las que "todo el mundo tiene razón"
- El aprendizaje estadístico (o aprendizaje máquina) siempre va a ser conservativo por definición
- ¿Es imprescindible tratar diferente a los (sub) grupos? (con sus implicaciones respecto a privacidad)
- ¿Puede haber sistemas de reconocimiento facial justos?
- ¿Se debe o no construir?¿Cómo hacerlo?
- Pertenencia a grupo no binarias
- ...

¿Cómo abordar solucionar?

Respuesta de arriba a abajo: guías éticas, códigos demonológicos, etc.
Ethics [2016]

Respuesta de abajo (contexto) a arriba:
¡Nada Sobre Nosotras/os, Sin Nosotras/os! Costanza-Chock [2018]

Para saber más

Una pequeña selección:

https://datajusticelab.org/

https://www.fatconference.org/

https://callingbullshit.org

https://ainowinstitute.org/

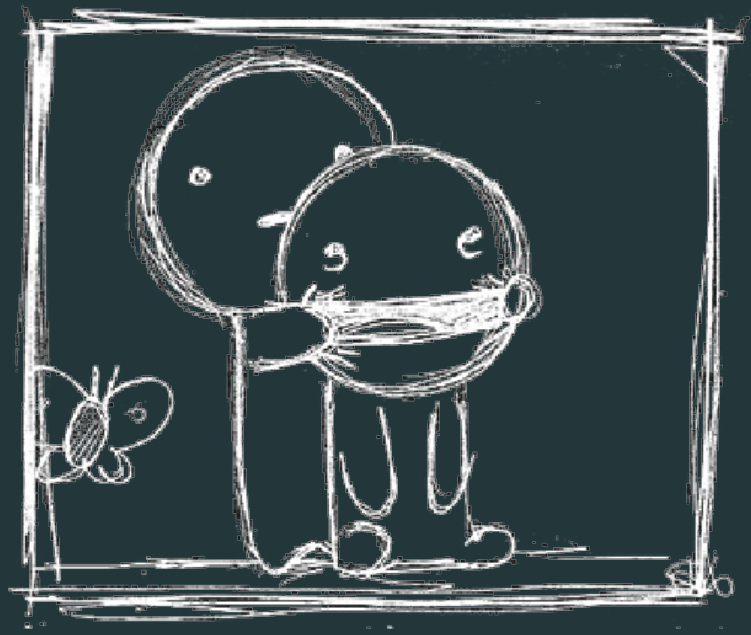
http://designjustice.org

https://morethancode.cc/

Fairness in Machine Learning. NIPS 2017 Tutorial

Just an Engineer: On the Politics of AI (Video)

¿Preguntas?¡Gracias!



Referencias i

S. Barocas and M. Hardt. Fairness in Machine Learning. NIPS 2017 Tutorial, 2017. URL http://fairml.how/.

Barocas, Solon; Selbst, Andrew D. Big Data's Disparate Impact. *California Law Review*, 2016. URL https://dx.doi.org/10.2139/ssrn.2477899.

D. Boyd and K. Crawford. CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679, June 2012. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2012.678878. URL http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878.

S. Costanza-Chock. Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*, July 2018. doi: 10.21428/96c8d426. URL https://jods.mitpress.mit.edu/pub/costanza-chock.

L. Dencik. Surveillance Realism and the Politics of Imagination: Is There No Alternative? *Krisis, Journal for Contemporary Philosophy*, 1, 2018. ISSN 1875-7103. URL http://krisis.eu/surveillance-realism-and-the-politics-of-imagination-is-there-no-alternative/.

L. Dencik, A. Hintz, and J. Cable. Towards data justice? The ambiguity of anti-surveillance resistance in political activism. *Big Data & Society*, 3(2):205395171667967, Dec. 2016. ISSN 2053-9517, 2053-9517. doi: 10.1177/2053951716679678. URL http://journals.sagepub.com/doi/10.1177/2053951716679678.

D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian. Runaway Feedback Loops in Predictive Policing. *arXiv:1706.09847 [cs, stat]*, June 2017. URL http://arxiv.org/abs/1706.09847. arXiv: 1706.09847.

A. Ethics. Code of Ethics, June 2016. URL https://ethics.acm.org/code-of-ethics/.

Referencias ii

D. Guerrero Martín. Apuntes de filosofía política en la era del «big data», 2018. URL http://www.mientrastanto.org/boletin-170/ensayo/apuntes-de-filosofia-politica-en-la-era-del-big-data.

S. Gürses, R. Overdorf, and E. Balsa. POTs: the revolution will not be optimized? page 2, 2018. URL https://petsymposium.org/2018/files/hotpets/3-gurses.pdf.

M. Hardt, E. Price, , and N. Srebro. Equality of Opportunity in Supervised Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3315–3323. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

J. A. Jeff Larson. How We Analyzed the COMPAS Recidivism Algorithm, May 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

J. Lerman. Big Data and Its Exclusions. *SSRN Electronic Journal*, 2013. ISSN 1556-5068. doi: 10.2139/ssrn.2293765. URL http://www.ssrn.com/abstract=2293765.

Z. C. Lipton. The Doctor Just Won't Accept That! *arXiv:1711.08037 [stat]*, Nov. 2017. URL http://arxiv.org/abs/1711.08037. arXiv: 1711.08037.

K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, Oct. 2016. ISSN 17409705. doi: 10.1111/j.1740-9713.2016.00960.x. URL http://doi.wiley.com/10.1111/j.1740-9713.2016.00960.x.

A. Roth. Course in (un)fairness in machine learning, 2018. URL http://cis.upenn.edu/~aaroth/FairnessZurich.pptx.

Referencias iii

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/zemel13.html.

Z. Zhang and D. B. Neill. Identifying Significant Predictive Bias in Classifiers. *arXiv:1611.08292 [cs, stat]*, Nov. 2016. URL http://arxiv.org/abs/1611.08292. arXiv: 1611.08292.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2941–2951, July 2017. URL http://arxiv.org/abs/1707.09457. arXiv: 1707.09457.