¿Por qué Necesitamos Sistemas de Inteligencia Artificial Justos, Explicables y Transparentes?

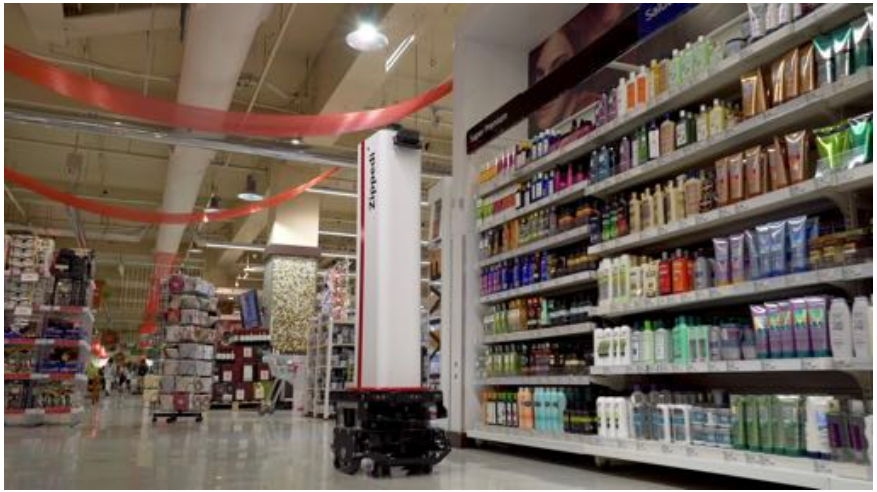
Denis Parra
PUC Chile & IA Lab UC & IMFD

Pequeña Biografía

- Ingeniero Civil en Informática UACh (2004)
- PhD in Information Science and Technology (U. Pittsburgh, 2013)
- Profesor Asociado DCC UC, miembro IALab UC
- Investigador adjunto del IMFD

IA Lab UC http://ialab.ing.puc.cl/

IA Lab UC http://ialab.ing.puc.cl/

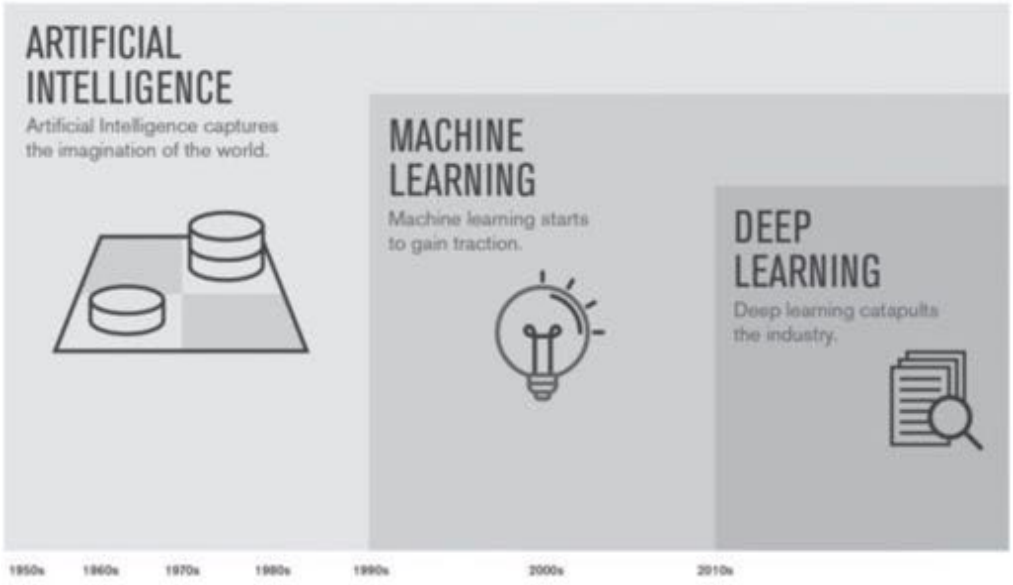


Zippedi: robots para el retail

NotCo: Industria de alimentos

Estamos viviendo días increíbles

- La tecnología nos muestra resultados que parecen de ciencia ficción



Procesamiento de Lenguaje Natural

- IBM Watson vence a los campeones de Jeopardy. << ... With all of its processing CPU power, Watson can scan two million pages of data in three seconds.>> E. Nyberg, CMU professor
- Implicancias: Aplicaciones en medicina

http://www.aaai.org/Magazine/Watson/watson.php

Vehículos Autónomos

Venciendo a los humanos en Go

Google AI algorithm masters ancient game of Go

Deep-learning software defeats human professional for first time.

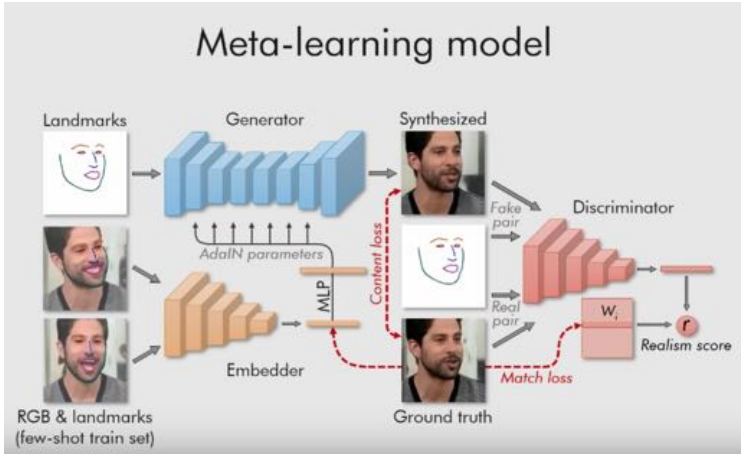
Elizabeth Gibney

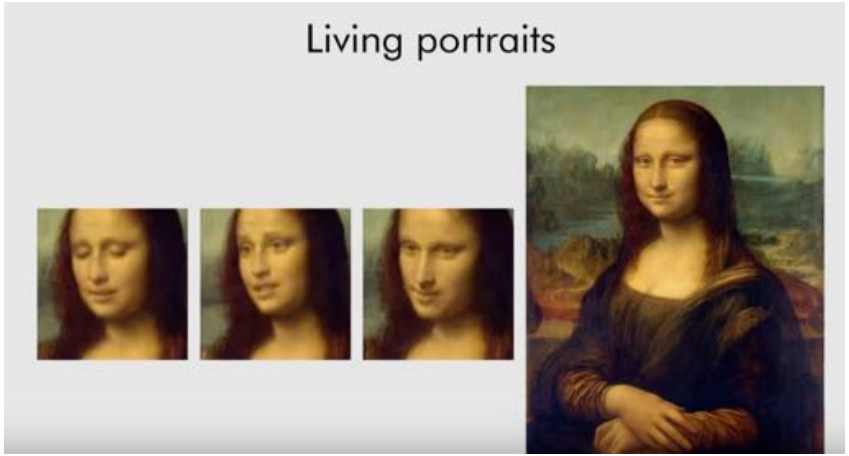
27 January 2016

PDF | Rights & Permissions

The computer that mastered Go

¡Retratos vivos!





Pero hay algunos problemas …

Uber Self-Driving Car Struck and Killed Arizona Woman While in Autonomous Mode

Last night a woman was struck by an autonomous Uber vehicle in Tempe, Arizona. She later died of her injuries in the hospital.

The deadly collision—reported by ABC15 and later confirmed to Gizmodo by Uber and Tempe police—took place around 10PM at the intersection of Mill Avenue and Curry Road, both of which are multi-lane roads. Autonomous vehicle developers often test drive at night, during storms, and other challenging conditions to help their vehicles learn to navigate in a variety of

Machine Bias

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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

Sistema COMPAS

- Se usa en EEUU para predecir reincidencia

Machine Bias

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

- ProPublica realizó un estudio sobre su efectividad

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

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Sistema COMPAS

- ProPoblica indica que cuando COMPAS se equivoca, falla en contra de afroamericanos.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Sistemas de Reconocimiento Facial

Other case: Gender Shades

- A Project by Joy Buolamwini, researcher at MIT Media Lab
- Examination of facial-analysis software shows error rate of 0.8 percent for light-skinned men, 34.7 percent for dark-skinned women.

http://gendershades.org/overview.html

https://www.media.mit.edu/projects/gender-shades/overview/

https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212

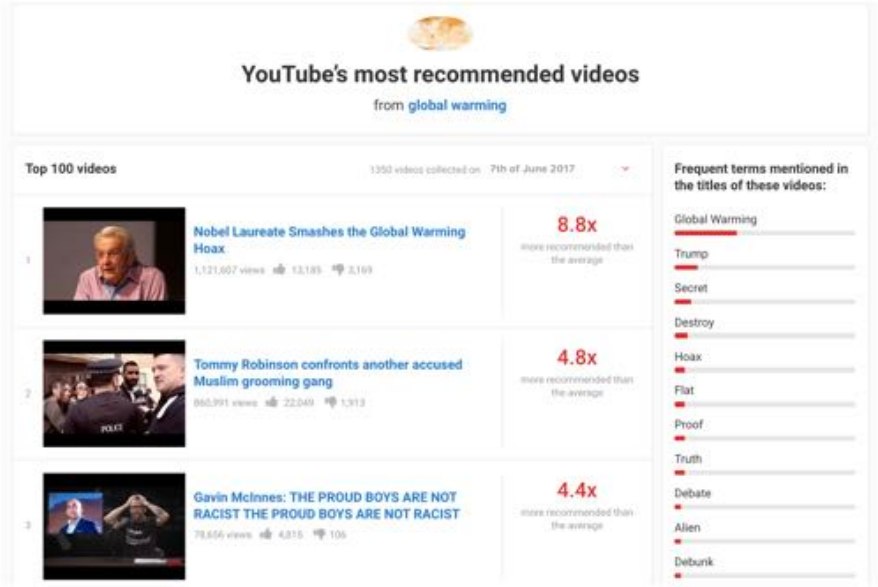
Recomendador de YouTube

- Guillaume Chaslot
- After resigning from YouTube, he created a system to estimate what was being recommended

https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot

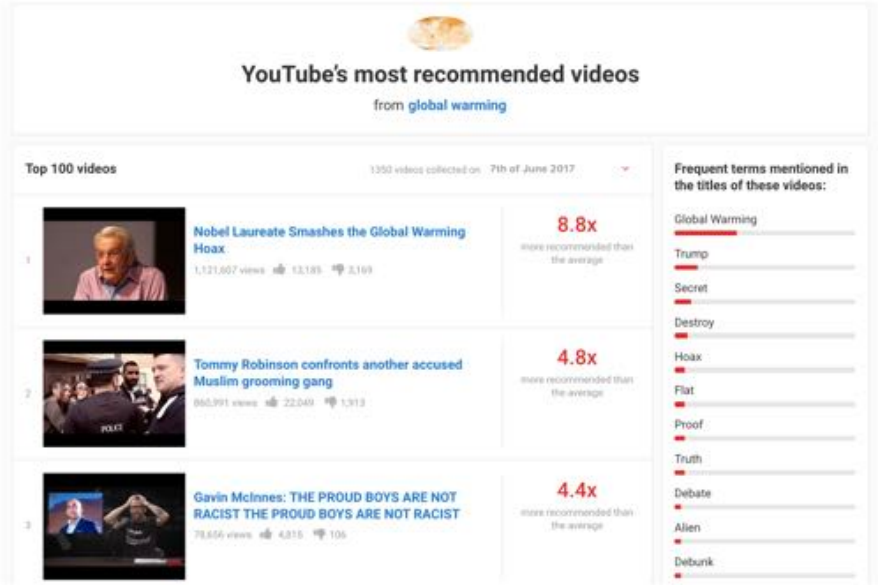
Recomendador de YouTube

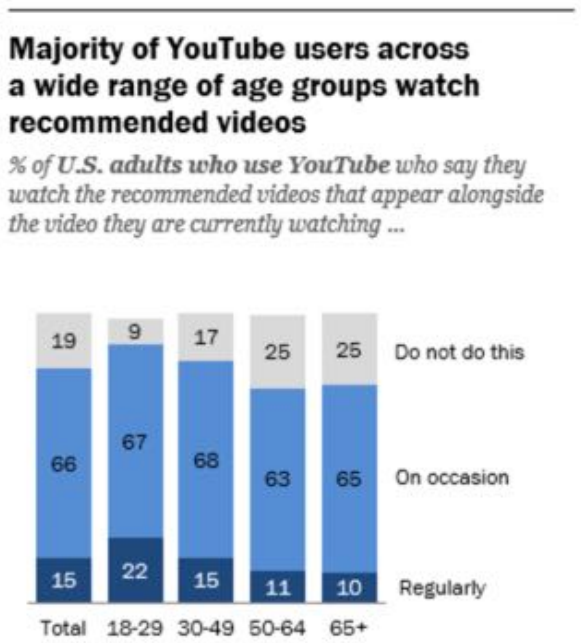
https://algotransparency.org



Recomendador de YouTube

https://algotransparency.org

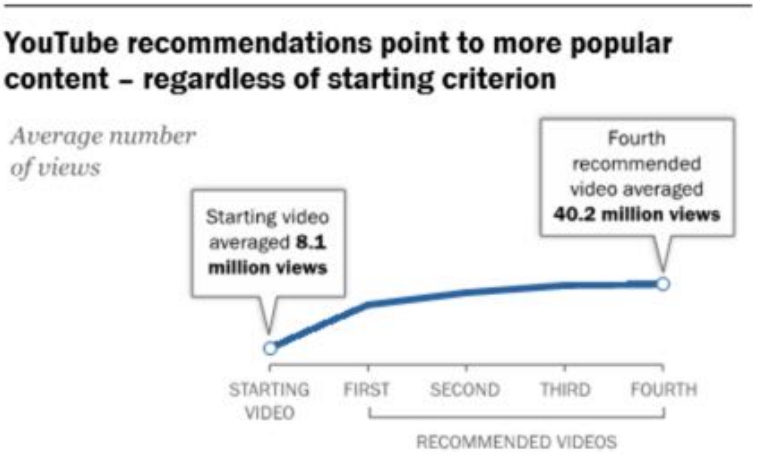


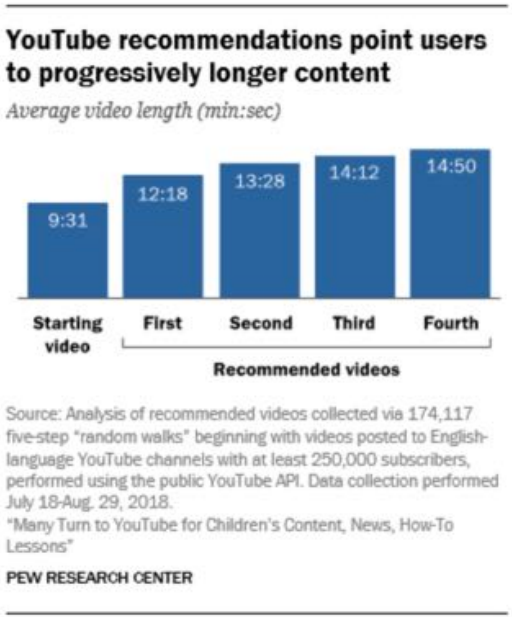


https://www.pewinternet.org/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons

Recomendador de YouTube

- YouTube recomienda contenido más Popular y de mayor duración.





Recomendador de YouTube

- Nuevo Sistema recomendador: Presentado in RecSys 2019: agrega multitask learning
- Aún no aborda el problema de calidad y fake news.

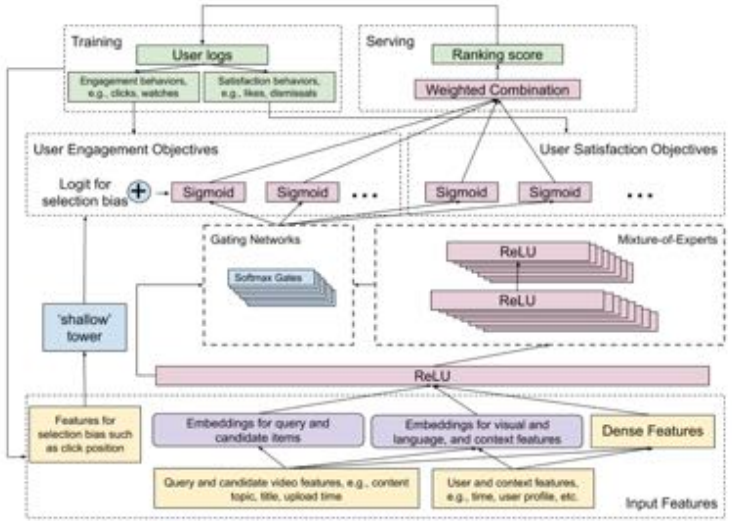


Figure 1: Model architecture of our proposed ranking system. It consumes user logs as training data, builds Multi-gate Mixture-of-Experts layers to predict two categories of user behaviors, i.e., engagement and satisfaction. It corrects ranking selection bias with a side-tower. On top, multiple predictions are combined into a final ranking score.

Algunos expertos sugieren calma….

We need to realize that the current public dialog on AI—which focuses on a narrow subset of industry and a narrow subset of academia—risks blinding us to the challenges and opportunities that are presented by the full scope of AI, IA and II.

Artificial Intelligence—The Revolution Hasn't Happened Yet

Just as early buildings and bridges sometimes fell to the ground — in unforeseen ways and with tragic consequences — (before there was civil engineering)

…

many of our early societal-scale inference-and-decision-making systems are already exposing serious conceptual flaws.

https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7

JET IA / FAT AI

- Justo
- Explicable
- Transparente

- **F**airness
- **A**ccountability
- **T**ransparency

JET IA / FAT AI

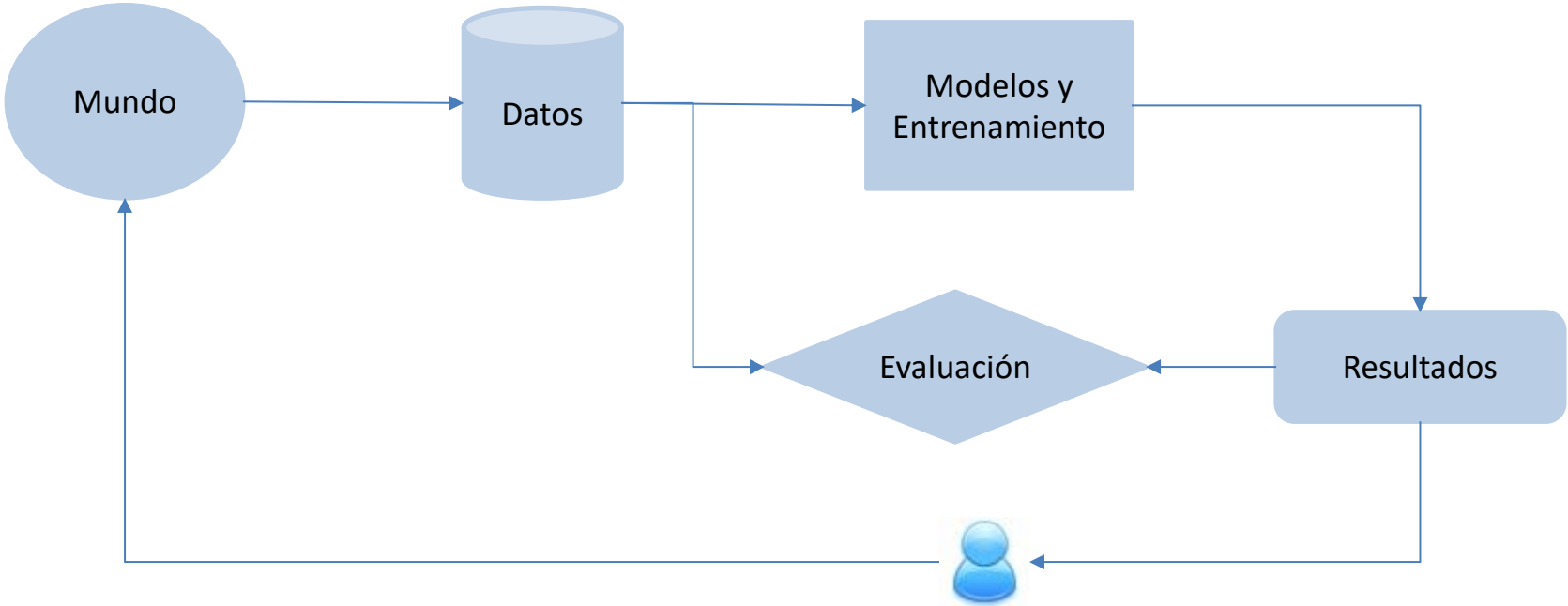
- Justo (no sesgado, ecuánime)
- Explicable (responsable de decisiones)
- Transparente (a qué nivel? Interpretable)

- **F**airness
- **A**ccountability
- **T**ransparency

The FAT* Conference

- https://fatconference.org
- A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

¿De dónde proviene el Sesgo?



From tutorial by Diaz, Ekstrand & Burke (SIGIR and RecSys 2019): https://fair-ia.ekstrandom.net/sigir2019

Fairness: Iniciativas de Regulación

- GDPR (Privacidad y derecho a explicación)

Fairness: Visualización

FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning

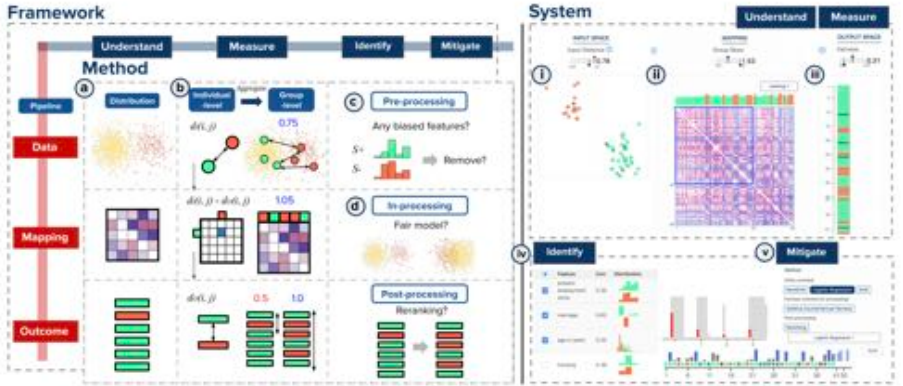
Ángel Alexander Cabrera Will Epperson Fred Hohman Minsuk Kahng
Jamie Morgenstern Duen Horng (Polo) Chau*

Georgia Institute of Technology

https://arxiv.org/abs/1904.05419

FairSight: Visual Analytics for Fairness in Decision Making

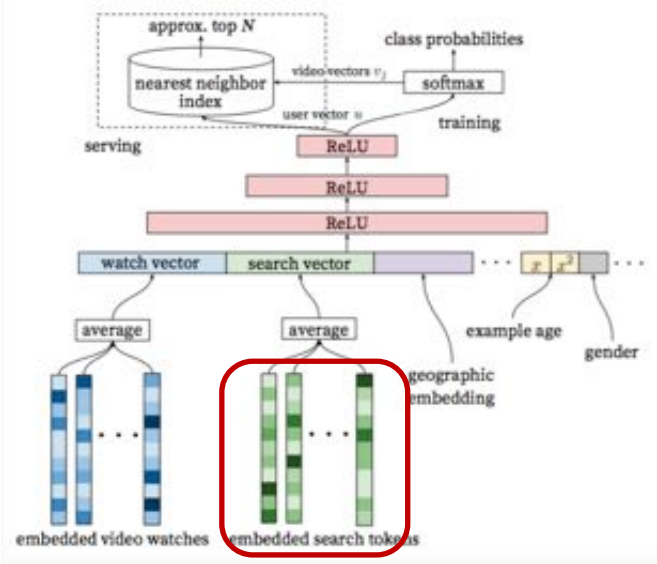
Yongsu Ahn, Yu-Ru Lin

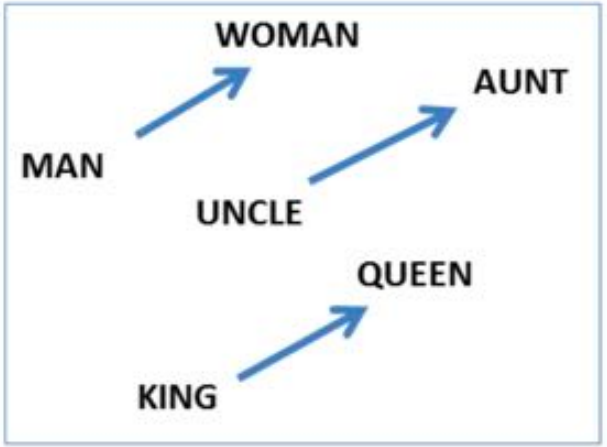


https://arxiv.org/abs/1908.00176

Fairness: Modelos de Lenguaje

- Bolukbasi et al. (2016) : ‘man’ - ‘computer programmer’ + ‘woman’ en word2vec -> ‘homemaker’





https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

Fairness: Modelos de Lenguaje

Figure 5: Words most associated with women (left) and men (right), estimated with *Pointwise Mutual Information*. Font size is inversely proportional to PMI rank. Color encodes frequency (the darker, the more frequent).

Wagner, C., Graells-Garrido, E., Garcia, D., & Menczer, F. (2016). Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science*, *5*(1), 5.

Fairness: Ranking

- From Tutorial on Algorithmic Bias in Rankings (Carlos Castillo, 2019)

1. Rank protected and unprotected separately

2. For each position:
   - Pick protected with probability *p*
   - Pick nonprotected with probability *1-p*

Continue until exhausting both lists

rank	gender
1	M
2	M
3	M
4	M
5	M
6	F
7	F
8	F
9	F
10	F

p=0

rank	gender
1	M
2	M
3	F
4	M
5	M
6	F
7	M
8	F
9	F
10	F

p=0.3

rank	gender
1	M
2	F
3	M
4	F
5	M
6	F
7	M
8	F
9	M
10	F

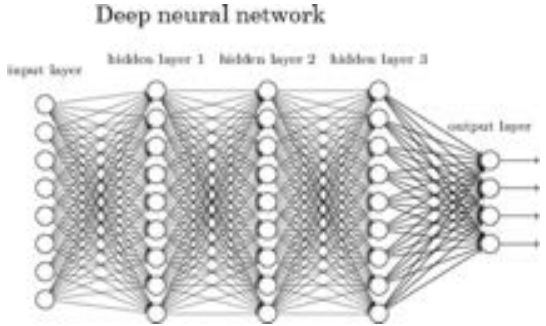
p=0.5

Yang, K., & Stoyanovich, J. (2017, June). Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (p. 22). ACM.

Explicabilidad

- ¿Cómo explicamos modelos de AI?
- De Decision Trees a Deep Neural Networks

Explainable decision model, explicit variables, not very accurate



Black-box decision model, latent variables, accurate

DARPA XAI

- Programa liderado por David Gunning

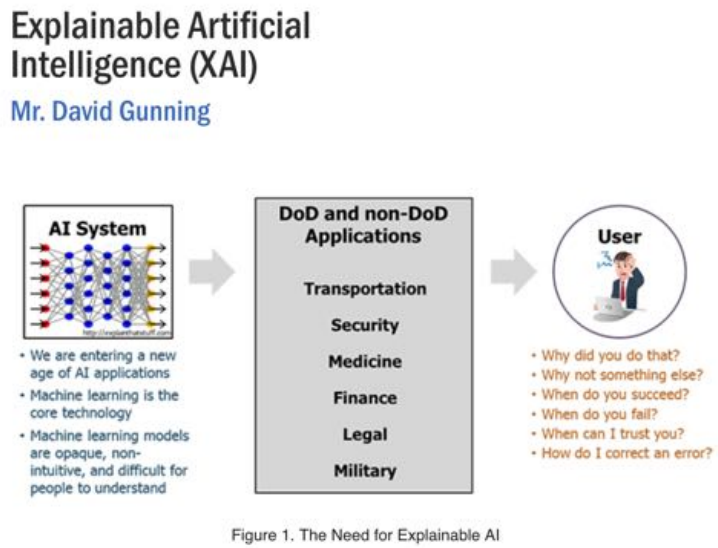
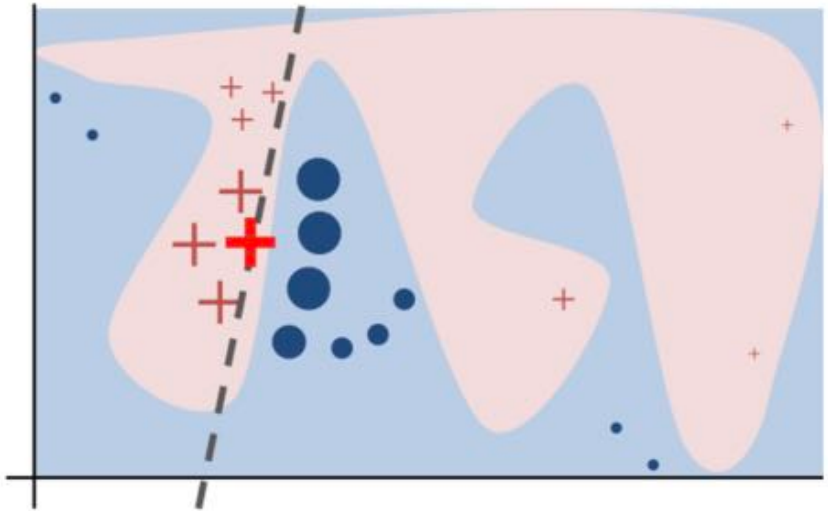


Figure 1. The Need for Explainable AI

Mr. David Gunning
Information Innovation Office (I2O)
Program Manager

LIME

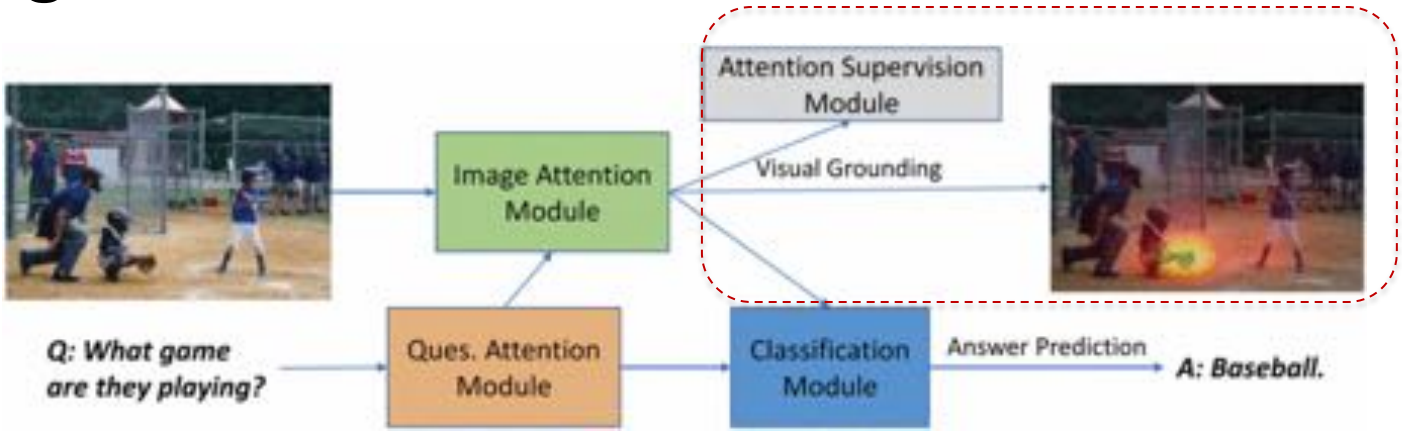
- LIME: Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. KDD 2016.



$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

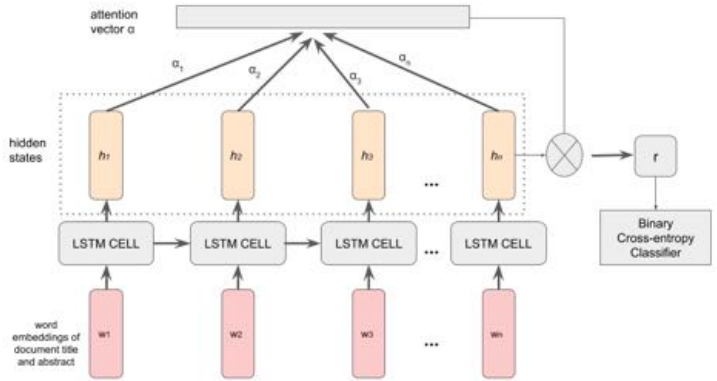
Investigación IMFD

- Alvaro Soto: Modelo interpretable de QA sobre imágenes



Investigación IMFD

- Valdivieso, Cavallo, Parra (VisXAI 2019): visualización de modelos de atención.



A meta analysis of birth origin effects on reproduction in diverse captive environments

Prediction: Not Relevant (NR)

Ground truth: Not Relevant (NR)

Title: a meta analysis of birth origin effects on reproduction in diverse captive environments
Abstract: successfully establishing captive breeding programs is priority across diverse industries to address food security demand for ethical laboratory research animals and prevent extinction differences in reproductive success due to birth origin may threaten the long term sustainability of captive breeding our meta analysis examining effect sizes from species of invertebrates fish birds and mammals shows that overall captive born animals have decreased odds of reproductive success in captivity compared to their wild born counterparts the largest effects are seen in commercial aquaculture relative to conservation or laboratory settings and offspring survival and offspring quality were the most sensitive traits although somewhat weaker trend reproductive success in conservation and laboratory research breeding programs is also in negative direction for captive born animals our study provides the foundation for future investigation of non genetic and genetic drivers of change

Conclusiones

- Los sistemas actuales de IA tienen problemas de sesgo y se hace imprescindible investigar, implementar y aplicar:
  - Implicancias legales.
  - Métodos para detectar y prevenir sesgos.
  - Métodos para apoyar la interacción humano-IA en la toma de decisiones.

Referencias

- https://sites.google.com/view/ears-tutorial/
- https://fair-ia.ekstrandom.net/sigir2019
- http://denisparra.github.io/pdfs/RecSysFAT-LARS2019_small.pdf

Denis Parra
Profesor Asociado
Pontificia Universidad Católica de Chile
dparra@ing.puc.cl