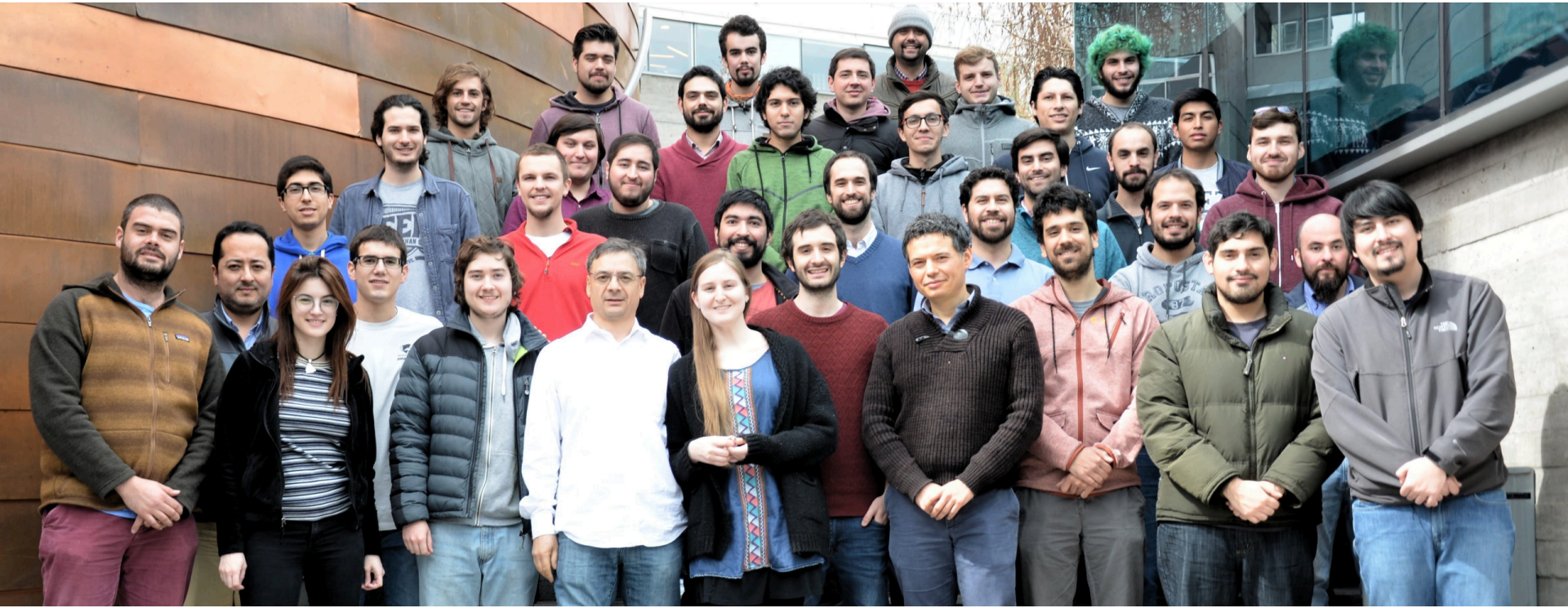
¿Por qué Necesitamos Sistemas de Inteligencia Artificial Justos, Explicables y Transparentes?

Denis Parra
PUC Chile & IA Lab UC & IMFD

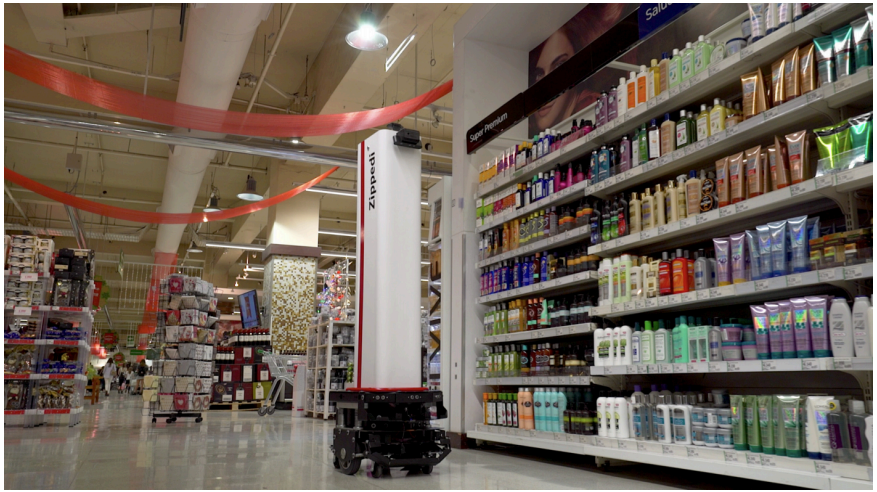
Pequeña Biografía

- 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/

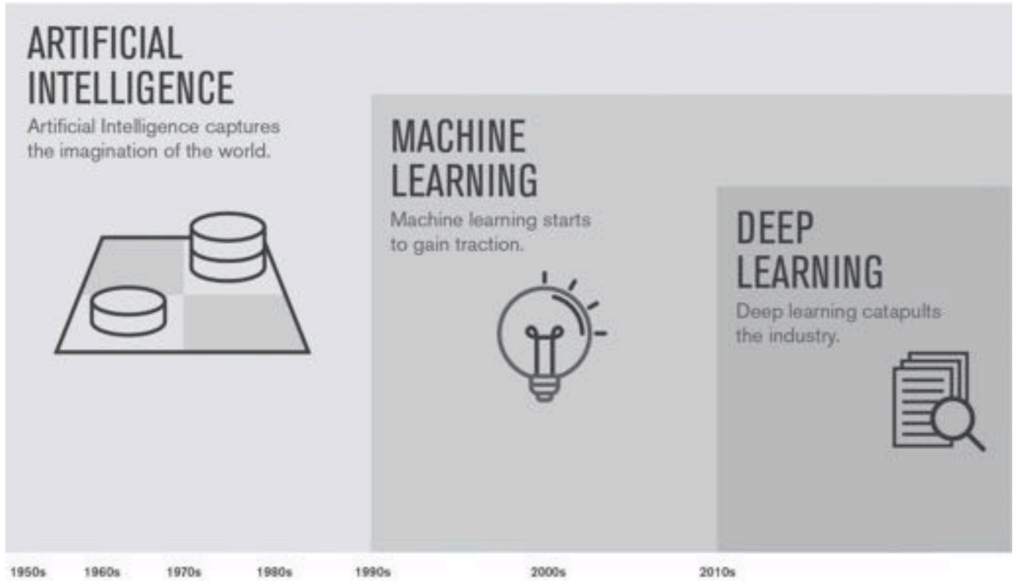


Zippedi: robots para el retail

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

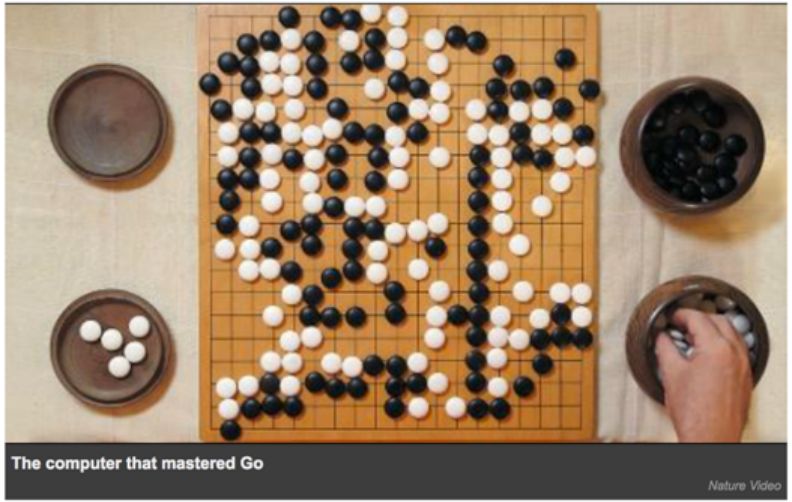
Google AI algorithm masters ancient game of Go

Deep-learning software defeats human professional for first time.

Elizabeth Gibney

27 January 2016

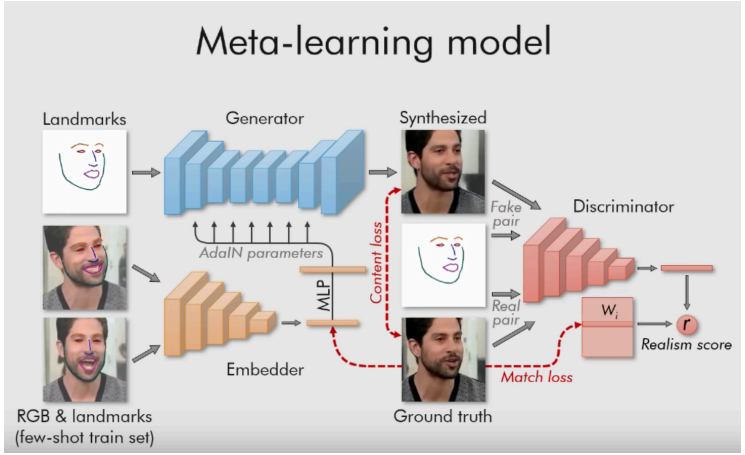
PDF | Rights & Permissions

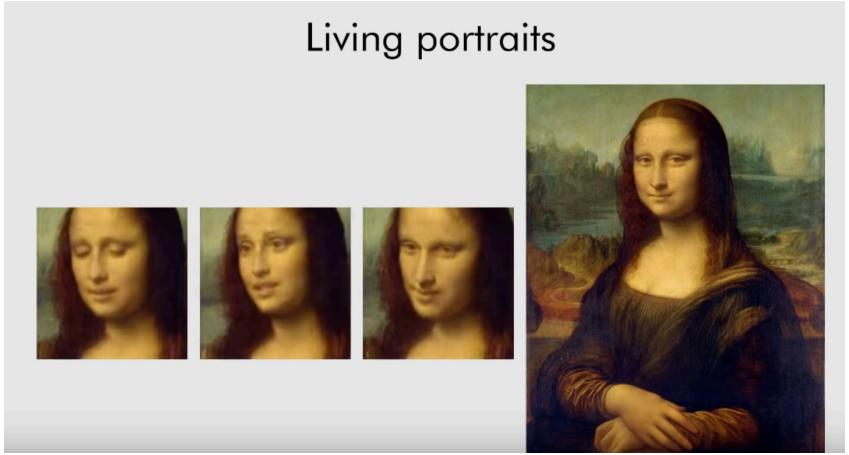


The computer that mastered Go

Nature Video

¡Retratos vivos!





Pero hay algunos problemas …

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

Bryan Menegus and Kate Conger
3/19/18 12:51pm • Filed to: UBER

750 16

Photo: Eric Risberg, (AP)

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

- ProPublica realizó un estudio sobre su efectividad

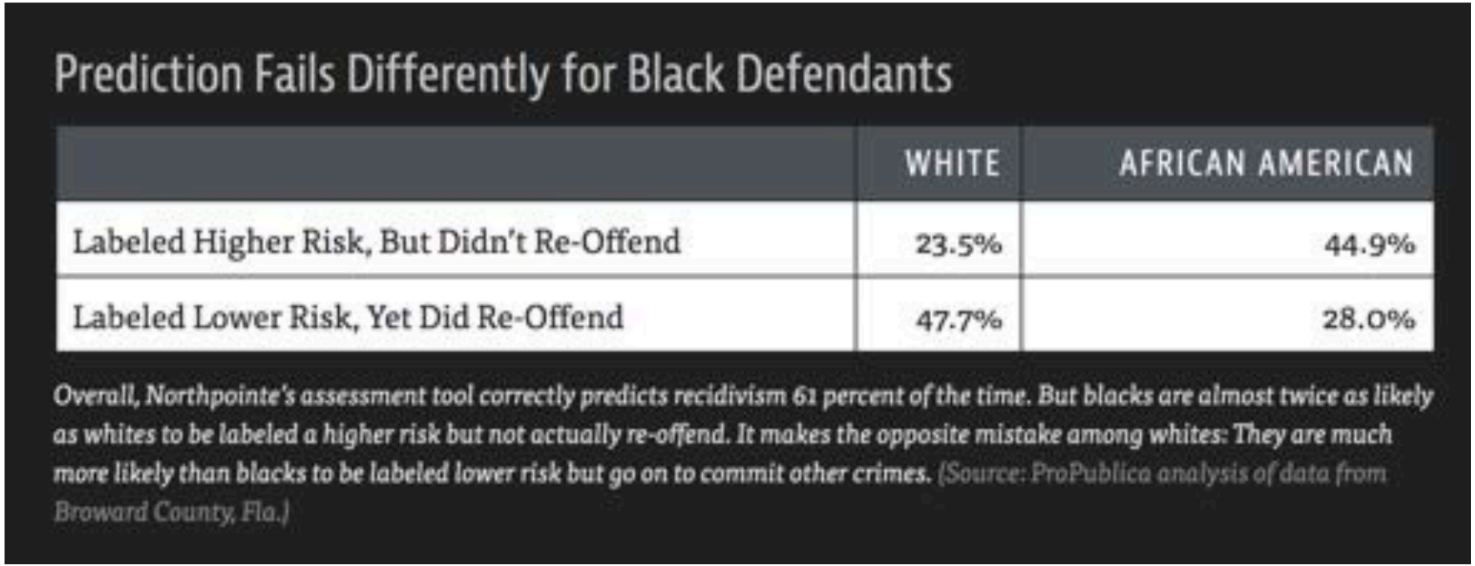
There's software used across the country to predict future criminals. And it's biased against blacks.

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

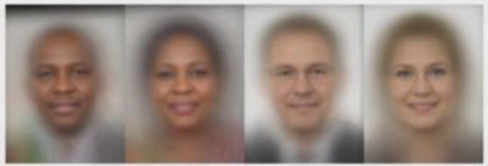
Other case: Gender Shades

- 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.

When we analyze the results by intersectional subgroups - darker males, darker females, lighter males, lighter females - we see that all companies perform worst on darker females.

IBM and Microsoft perform best on lighter males. Face++ performs best on darker males.

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



http://gendershades.org/overview.html

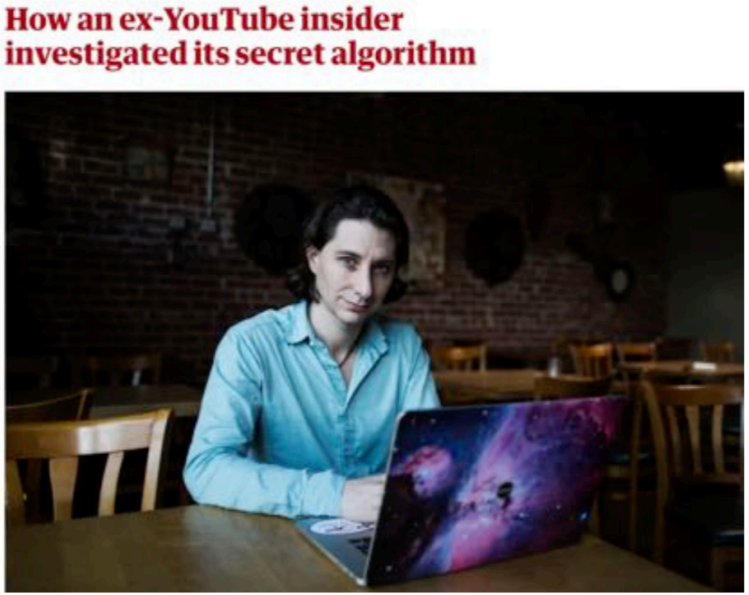
https://www.media.mit.edu/projects/gender-shades/overview/

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

Recomendador de YouTube

- Guillaume Chaslot

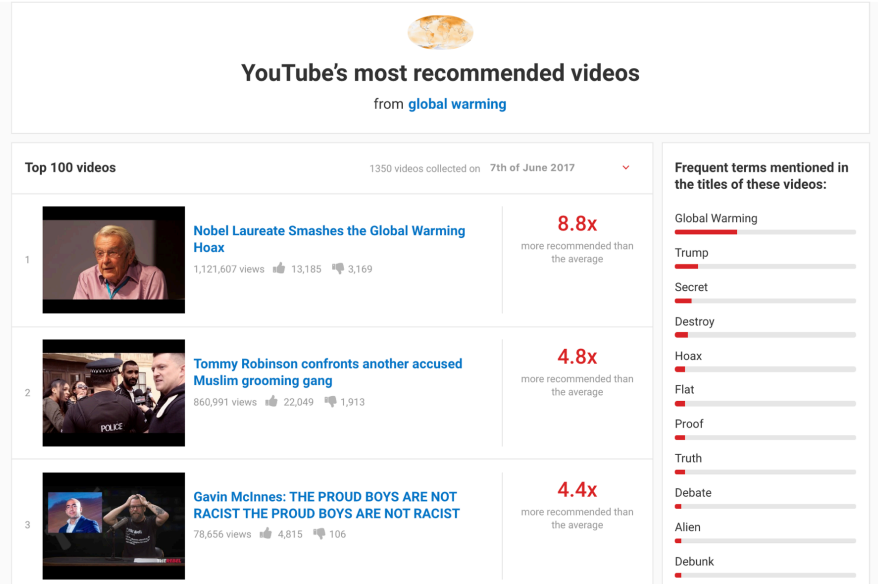
- After resigning from YouTube, he created a system to estimate what was being recommended



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

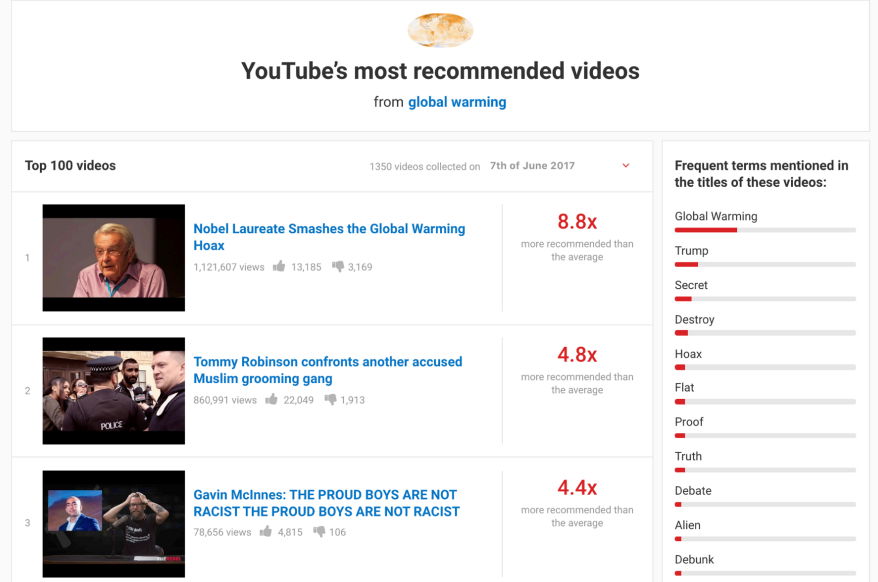
Recomendador de YouTube

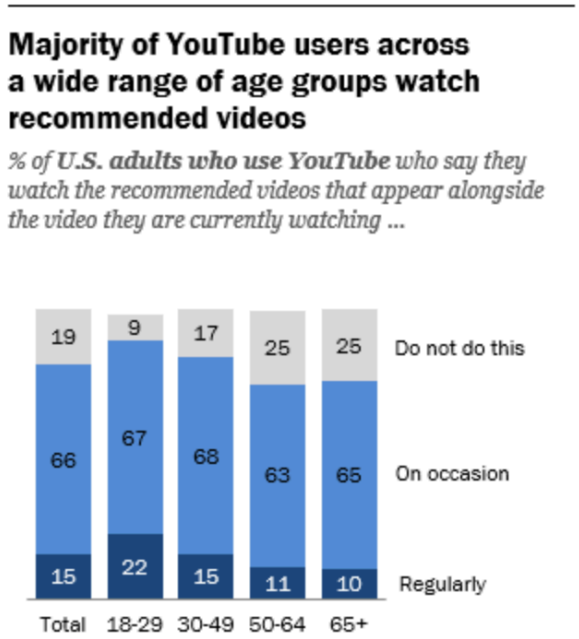
https://algotransparency.org



Recomendador de YouTube

https://algotransparency.org

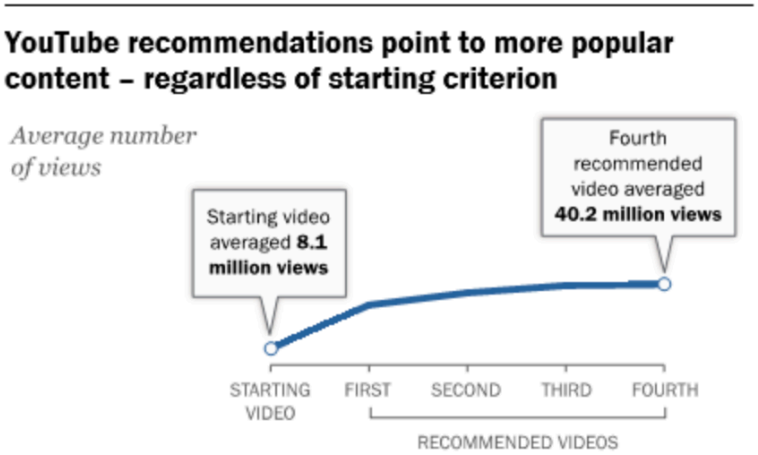


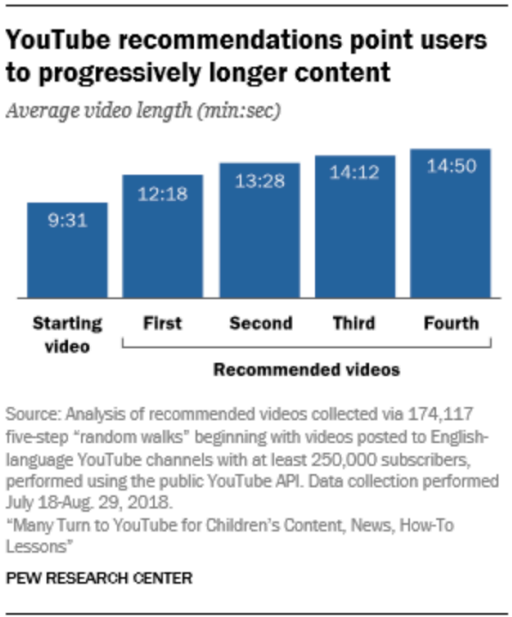


https://www.pewinternet.org/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons

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.

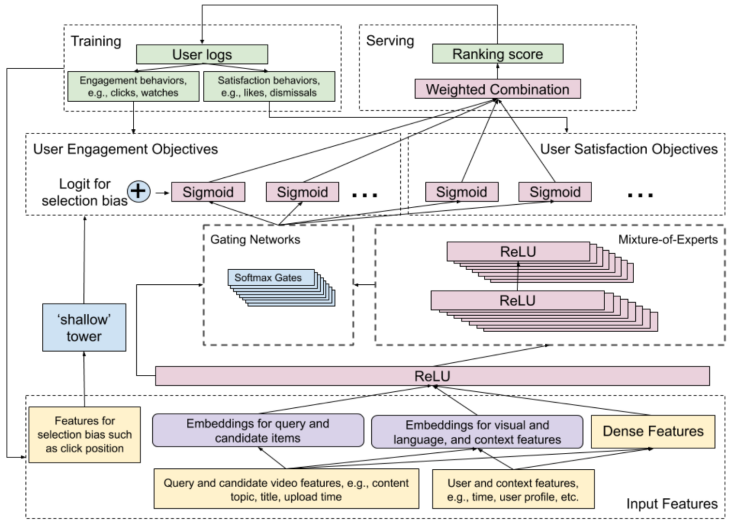


Figure 1: Model architecture of our proposed ranking system. It consumes user logs as training data, builds Multi-gate Mixture-of-Experts layers to predict two categories of user behaviors, i.e., engagement and satisfaction. It corrects ranking selection bias with a side-tower. On top, multiple predictions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

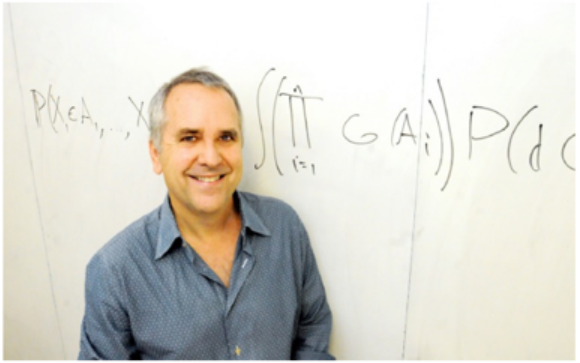


Photo credit: Peg Skorpinski

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

FAT definitions

- **Fairness**: The property of being fair or equitable

  vs. **Bias**: inclination towards something; predisposition, partiality, prejudice, preference, predilection, discrimination.

- Accountability:

- Transparency

FAT definitions

- **Fairness**: The property of being fair or equitable

  vs. **Bias**: inclination towards something; predisposition, partiality, prejudice, preference, predilection, discrimination.

- According to Friedman and Nissembaum (1994) a computer system is biased “if it systematically and unfairly discriminate[s] against certain individuals or groups of individuals in favor of others.”

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Transactions on Information Systems 14, 3 (1996), 330–347.

FAT definitions

- **Fairness**: The property of being fair or equitable

  vs. **Bias**: inclination towards something; predisposition, partiality, prejudice, preference, predilection, discrimination.

- According to Friedman and Nissembaum (1994) a computer system is biased "if it systematically and unfairly discriminate[s] against certain individuals or groups of individuals in favor of others."
  - "… a system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or a group of individuals on grounds that are unreasonable or inappropriate."

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Transactions on Information Systems 14, 3 (1996), 330–347.

FAT Definitions

- Fairness

- **(Algorithmic) Accountability**: To be accountable means to be subject to giving an account or having the obligation to report, explain or justify something -> explainable AI (**XAI**).

- Transparency

FAT Definitions

- Fairness
- Accountability
- **(Algorithmic) Transparency:** is the principle that the factors that influence the decisions made by algorithms should be visible, or transparent, to the people who use, regulate, and are affected by systems that employ those algorithms.

Important Distinction

- Algorithmic accountability vs algorithmic transparency: Some people use it interchangeably, but a system can be accountable (provide explanations, justifications) without necessarily being transparent (completely opening the complexity of a black-box)
- From the DARPA XAI Program

Image from Zhang et al. (2019) Tutorial on ExplainAble Recommendation and Search

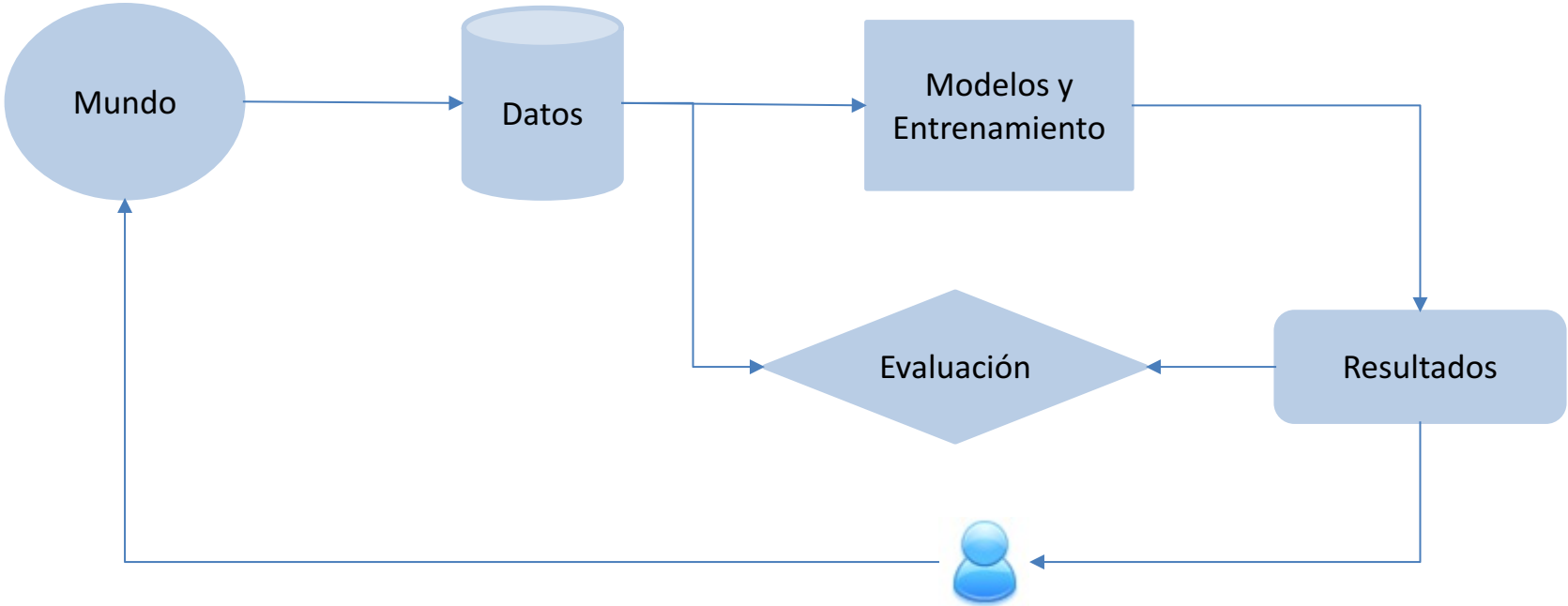
Other relevant terms

- Interpretability, in the context of AI/ML:
  - “the degree to which a human can understand the cause of a decision” (T. Miller, et al. AI 2018)
  - “the degree to which a human can consistently predict the model’s result” (B. Kim, et al. NIPS 2016)
  - “the ability to explain or to present in understandable terms to a human” (Doshi-Velez and Kim, 2017)

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

… y cómo me afecta esto a mi?

- Yo estoy estudiando Ingeniería, Computación, algoritmos, quizás voy a emprender y crear mi empresa … pero no me interesan estos temas.

Law: What happened in May 25th,

- The EU General Data Protection Regulation (GDPR) becomes enforceable.

And why do we care in this room ?

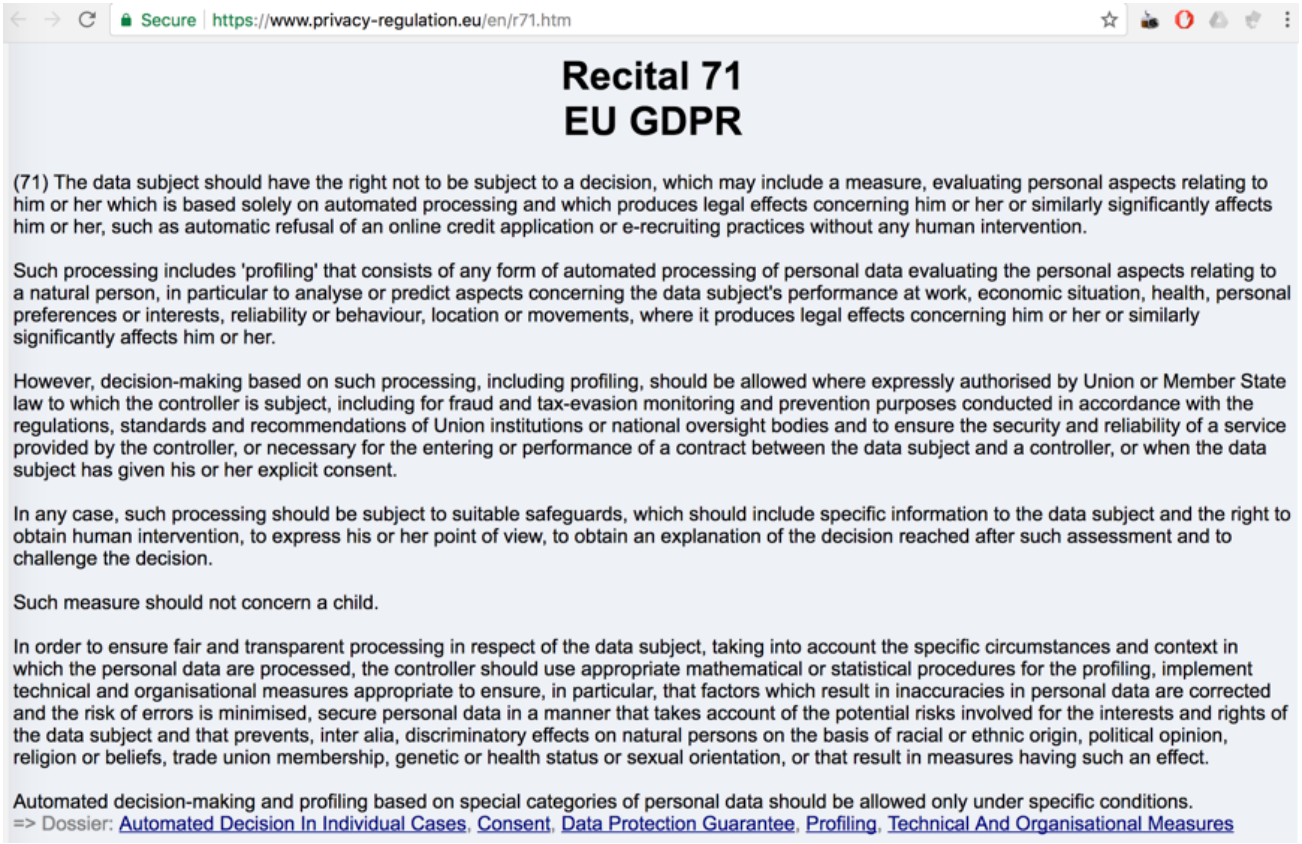
- The GDPR **not only applies to organisations located within the EU** but it will also apply to **organisations located outside of the EU** if they offer goods or services to, or monitor the behaviour of, EU data subjects.
- **It applies to all companies processing and holding the personal data** of data subjects residing in the European Union, regardless of the company's location.

Which is the effect on my current

Right to explanation

- Article 15 “**Right of access by the data subject**”
- Article 22 “**Automated individual decision-making, including profiling**”
- Recital 71 (linked to **art. 22**)

Recital 71



Recital 71
EU GDPR

(71) The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.

However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent.

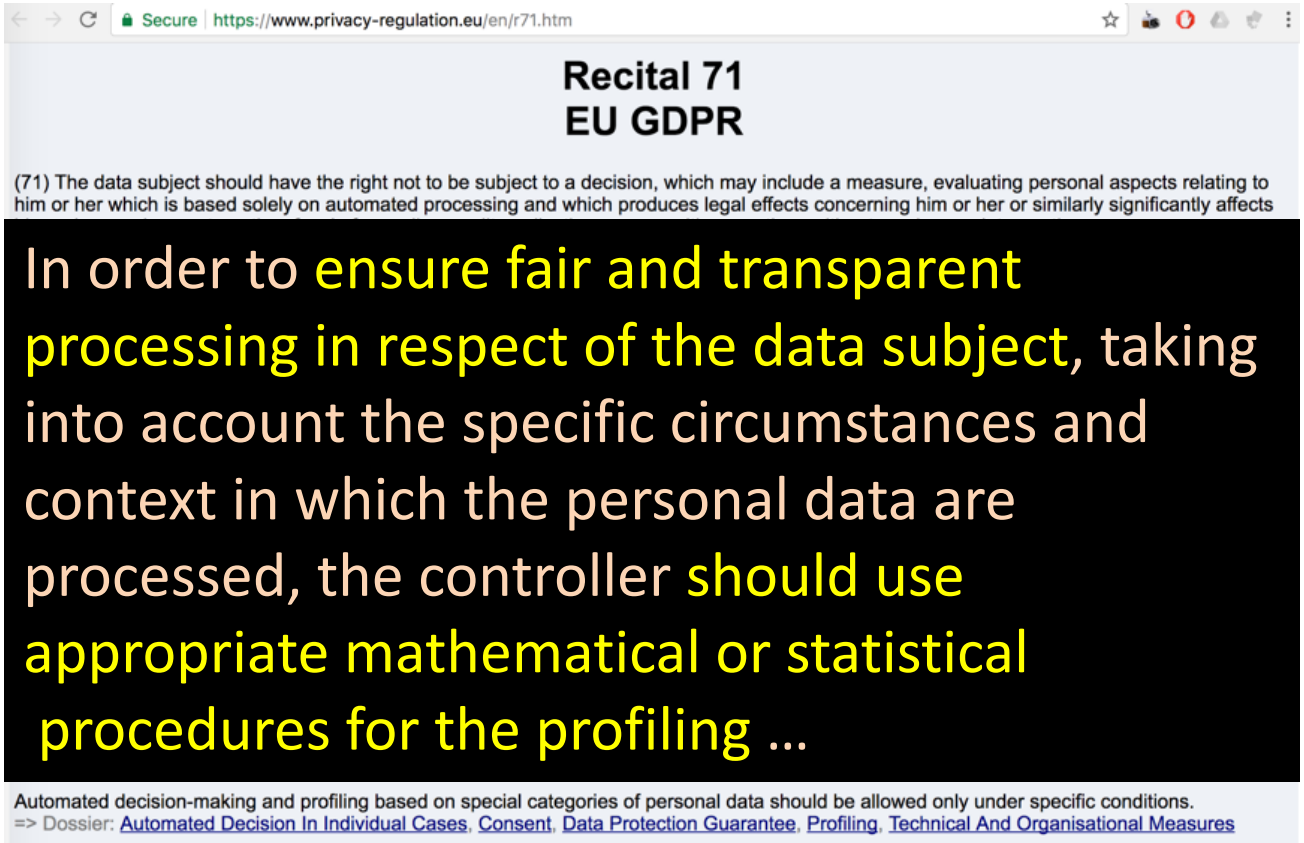
In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

Such measure should not concern a child.

In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect.

Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.
=> Dossier: Automated Decision In Individual Cases, Consent, Data Protection Guarantee, Profiling, Technical And Organisational Measures

Recital 71


Secure https://www.privacy-regulation.eu/en/r71.htm
Recital 71
EU GDPR
(71) The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects
In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling …
Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.
=> Dossier: Automated Decision In Individual Cases, Consent, Data Protection Guarantee, Profiling, Technical And Organisational Measures

Human Interpretability in ML

- https://arxiv.org/abs/1606.08813

arXiv.org > stat > arXiv:1606.08813

Search or Article ID | All papers

(Help | Advanced search)

Statistics > Machine Learning

European Union regulations on algorithmic decision-making and a "right to explanation"

Bryce Goodman, Seth Flaxman

(Submitted on 28 Jun 2016 (v1), last revised 31 Aug 2016 (this version, v3))

We summarize the potential impact that the European Union's new General Data Protection Regulation will have on the routine use of machine learning algorithms. Slated to take effect as law across the EU in 2018, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which "significantly affect" users. The law will also effectively create a "right to explanation," whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation.

Comments: presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY
Subjects: **Machine Learning (stat.ML)**; Computers and Society (cs.CY); Learning (cs.LG)
Cite as: arXiv:1606.08813 [stat.ML]
(or arXiv:1606.08813v3 [stat.ML] for this version)

Download:
- PDF
- Other formats

(license)

Current browse context:
stat.ML
< prev | next >
new | recent | 1606

Change to browse by:
cs
cs.CY
cs.LG
stat

References & Citations
- NASA ADS

Bookmark (what is this?)

Other Initiatives

The first bill to examine 'algorithmic bias' in government agencies has just passed in New York City

BI Zoë Bernard
Business Insider December 19, 2017

- New York City has passed the algorithmic accountability bill, which will assign a task force to examine the way that city government agencies use algorithms.
- Algorithmic bias is a critical issue in the justice system, which often relies on algorithmic risk assessments to inform criminal sentencing in federal court.
- The bill is the first of its kind to be passed in the nation, and will attempt to provide transparency in the way that the government uses algorithms.

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

Council Home | Legislation | Calendar | City Council | Committees

Details | Reports

File #:	Int 1696-2017 Version: A	Name:	Automated decision systems used by agencies.
Type:	Introduction	Status:	Enacted
		Committee:	Committee on Technology
On agenda:	8/24/2017		
Enactment date:	1/11/2018	Law number:	2018/049
Title:	A Local Law in relation to automated decision systems used by agencies		
Sponsors:	James Vacca, Helen K. Rosenthal, Corey D. Johnson, Rafael Salamanca, Jr., Vincent J. Gentile, Robert E. Cornegy, Jr., Jumaane D. Williams, Ben Kallos, Carlos Menchaca		
Council Member Sponsors:	9		
Summary:	This bill would require the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public and how agencies may address instances where people are harmed by agency automated decision systems.		
Indexes:	Oversight		
Attachments:	1. Summary of Int. No. 1696-A, 2. Summary of Int. No. 1696, 3. Int. No. 1696, 4. August 24, 2017 - Stated Meeting Agenda with Links to Files, 5. Committee Report 10/16/17, 6. Hearing Testimony 10/16/17, 7. Hearing Transcript 10/16/17, 8. Proposed Int. No. 1696-A - 12/12/17, 9. Committee Report 12/7/17, 10. Hearing Transcript 12/7/17, 11. December 11, 2017 - Stated Meeting Agenda with Links to Files, 12. Hearing Transcript - Stated Meeting 12-11-17, 13. Int. No. 1696-A (FINAL), 14. Fiscal Impact Statement, 15. Legislative Documents - Letter to the Mayor, 16. Local Law 49, 17. Minutes of the Stated Meeting - December 11, 2017		

History (13) | Text

13 records | Group | Export

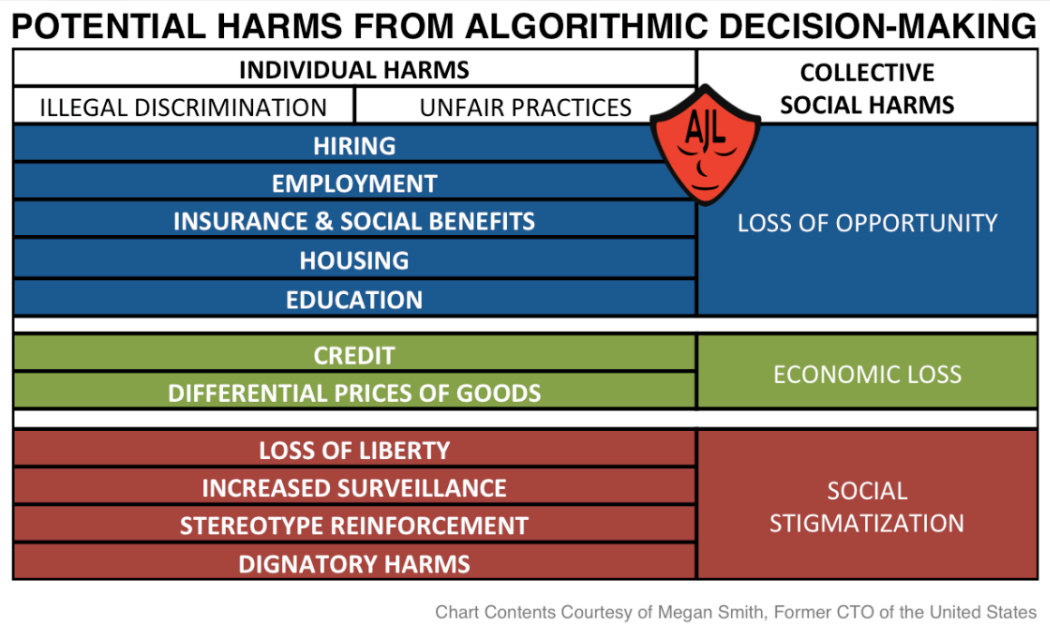
Date	Ver.	Prime Sponsor	Action By	Action	Result	Action Details	Meeting Details	Multimedia
1/17/2018	A	James Vacca	City Council	Returned Unsigned by Mayor		Action details	Meeting details	Not available
1/11/2018	A	James Vacca	Administration	City Charter Rule Adopted		Action details	Meeting details	Not available
12/18/2017	A	James Vacca	Mayor	Hearing Scheduled by Mayor		Action details	Meeting details	Not available
12/11/2017	A	James Vacca	City Council	Sent to Mayor by Council		Action details	Meeting details	Not available

This bill would require the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public and how agencies may address instances where people are harmed by agency automated decision systems.

Potential Harms on algorithmic

Automated systems are not inherently neutral. They reflect the priorities, preferences, and prejudices - the coded gaze - of those who have the power to mold artificial intelligence.

We risk losing the gains made with the civil rights movement and women's movement under the false assumption of machine neutrality. We must demand increased transparency and accountability.



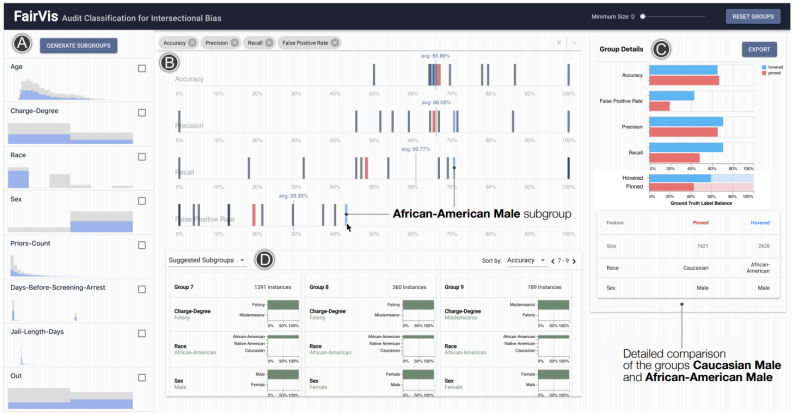
http://gendershades.org/overview.html

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*

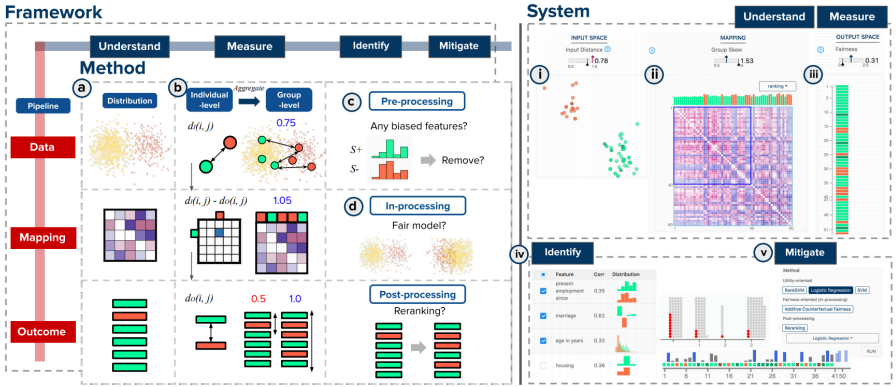
Georgia Institute of Technology



https://arxiv.org/abs/1904.05419

FairSight: Visual Analytics for Fairness in Decision Making

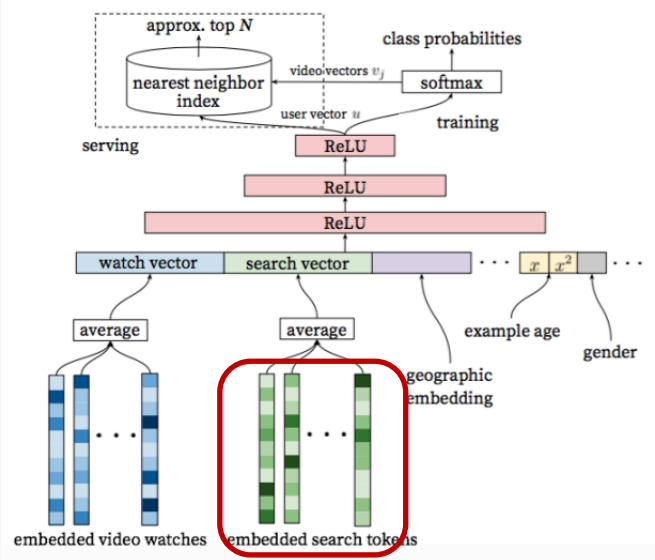
Yongsu Ahn, Yu-Ru Lin

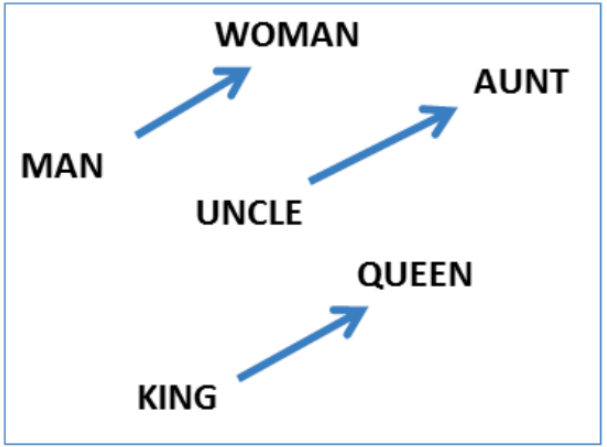


https://arxiv.org/abs/1908.00176

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

$p=0.3$

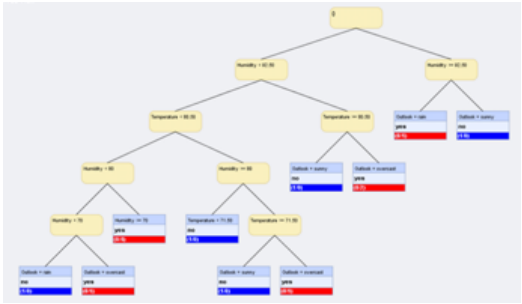
rank	gender
1	M
2	F
3	M
4	F
5	M
6	F
7	M
8	F
9	M
10	F

$p=0.5$

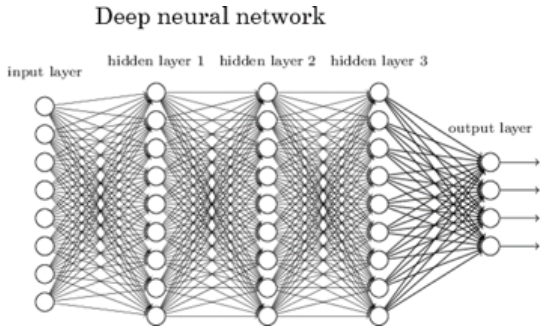
Yang, K., & Stoyanovich, J. (2017, June). Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (p. 22). ACM.

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

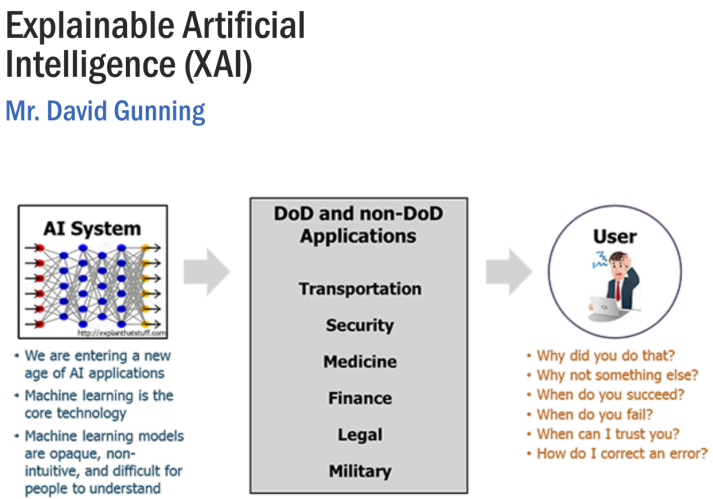
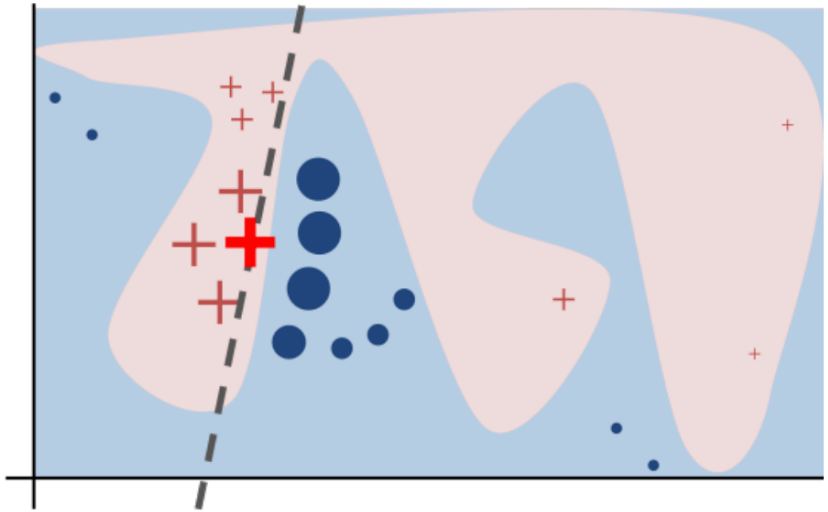


Figure 1. The Need for Explainable AI

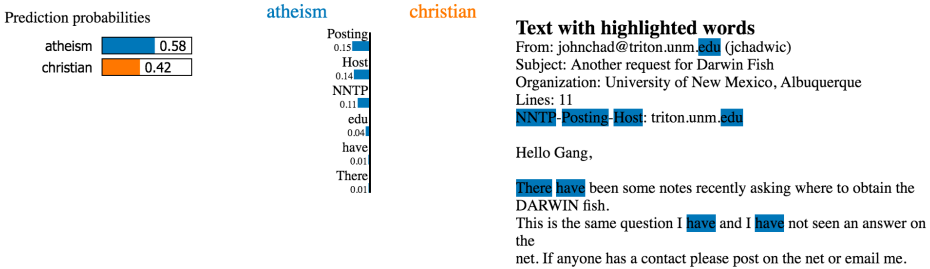
Mr. David Gunning
Information Innovation Office (I2O)
Program Manager

LIME

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

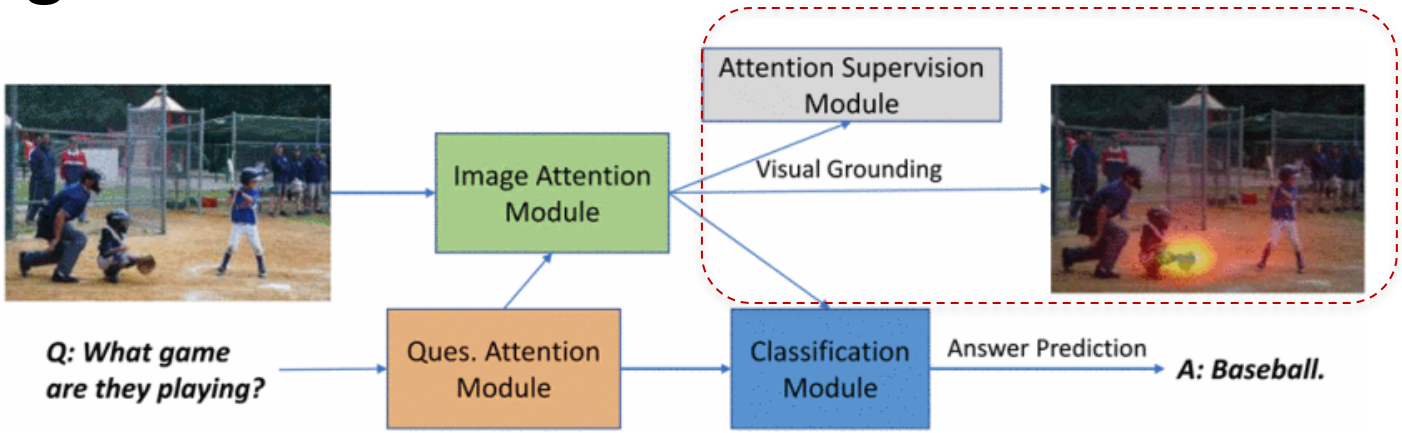


$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$



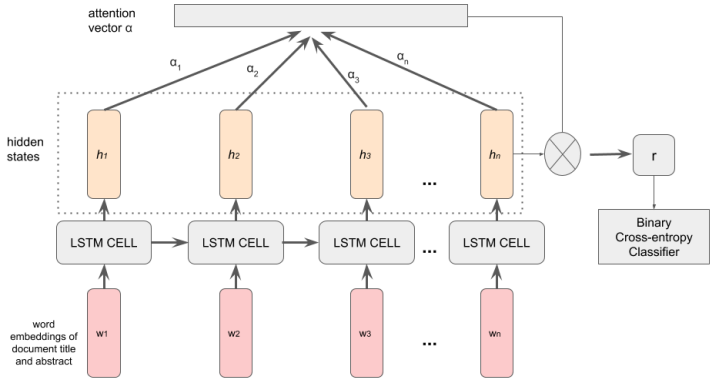
Investigación IMFD

- 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