



Fairness, Accountability and Transparency (FAT) in Recommender Systems

Denis Parra
PUC Chile, i-Health, CENIA, IMFQ

IIC3633

Actualización: Mayo 2024



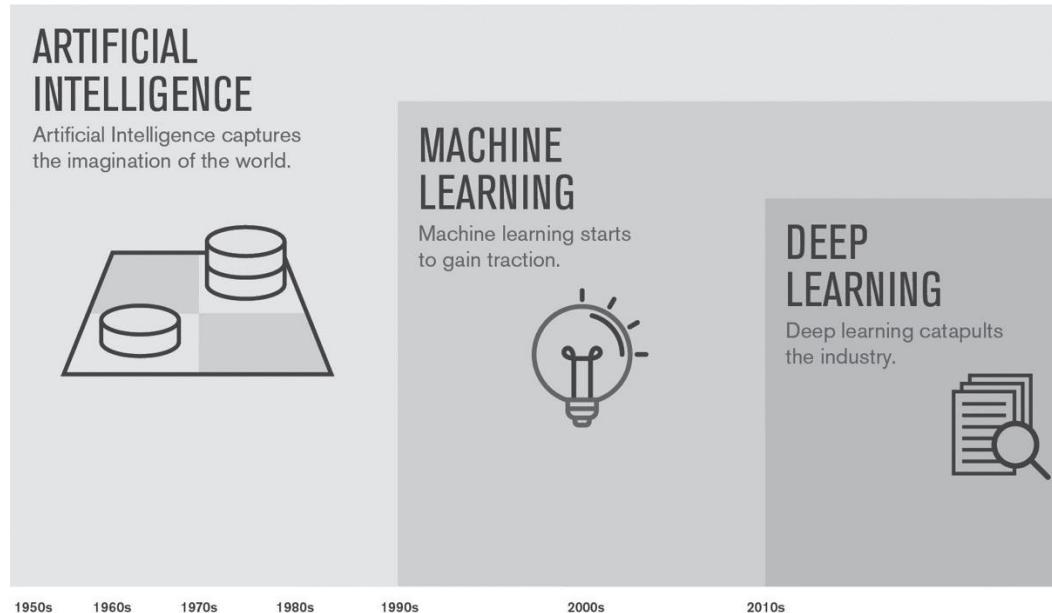
denisparra

We are living incredible days...

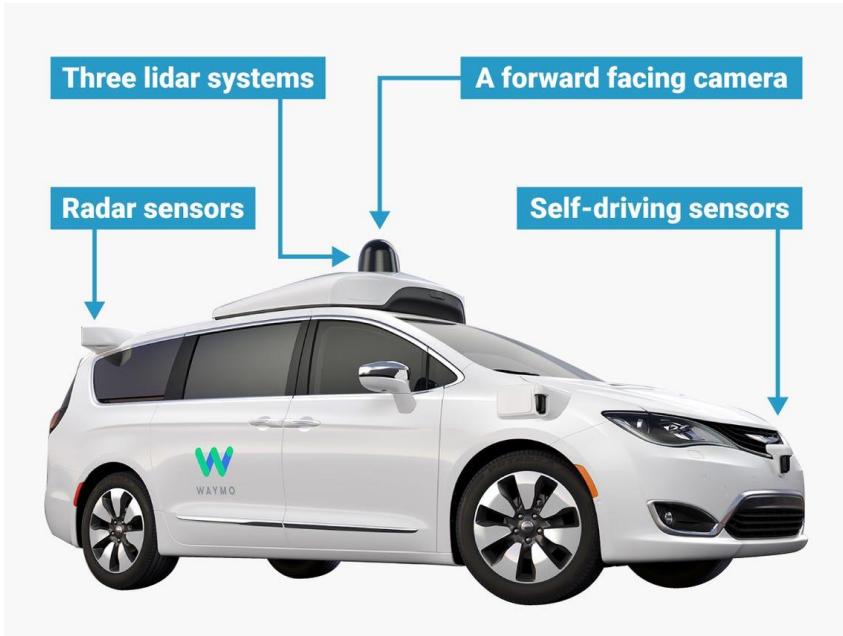
- Technology is showing results which resemble science fiction, specially in the area called Artificial Intelligence

We are living incredible days...

- Technology is showing results which resemble science fiction, specially in the area called Artificial Intelligence.



Vehículos Autónomos



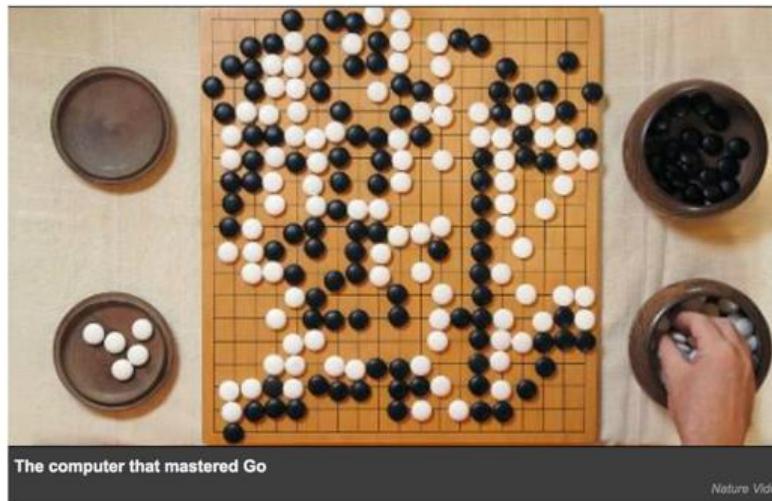
Mastering Go

Google AI algorithm masters ancient game of Go

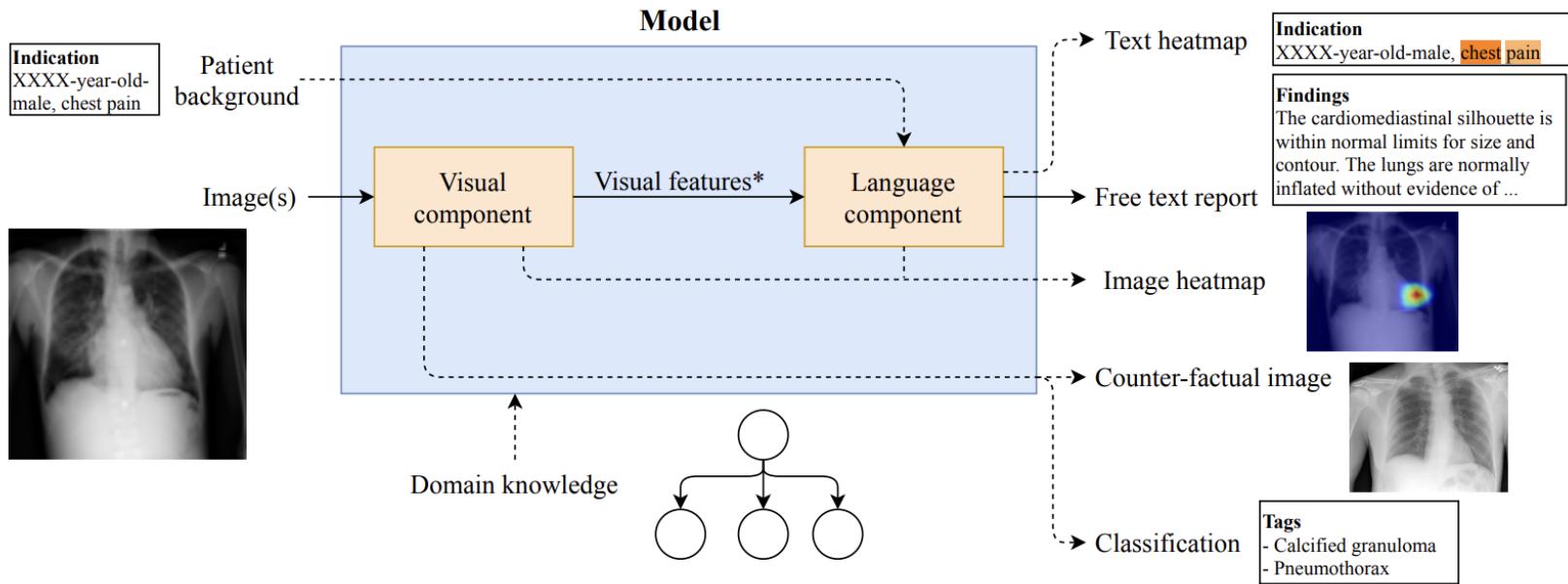
Deep-learning software defeats human professional for first time.

Elizabeth Gibney

27 January 2016



Aplicaciones para diagnóstico clínico



Modelos Generativos de Lenguaje y Visuales



But there are some problems



Machine Bias

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

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.

- The COMPAS system is used in the USA to predict recidivism

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>

But there are some problems



Machine Bias

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

By Julia Angwin, Jeff Larson, Kara Motes and Lauren Kirchner, ProPublica
May 25, 2016

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.

- When the COMPAS system correctly predicts recidivism, it does it similarly to black and white,
- But, when it fails to predict correctly:

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.)

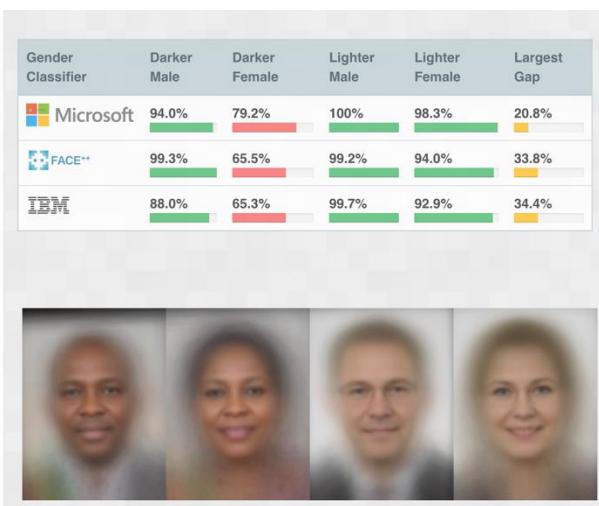
<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

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

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.



<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>



Some voices call for deeper discussion

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.

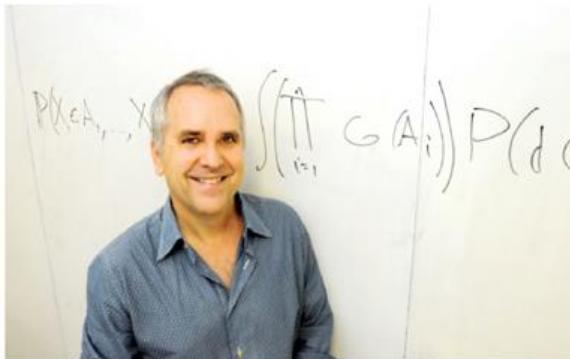


Photo credit: Peg Skorpinski

**Artificial Intelligence—The Revolution
Hasn't Happened Yet**

Thus, just as humans built buildings and bridges before there was civil engineering, humans are proceeding with the building of societal-scale, inference-and-decision-making systems that involve machines, humans and the environment.

Just as early buildings and bridges sometimes fell to the ground—in unforeseen ways and with tragic consequences—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>

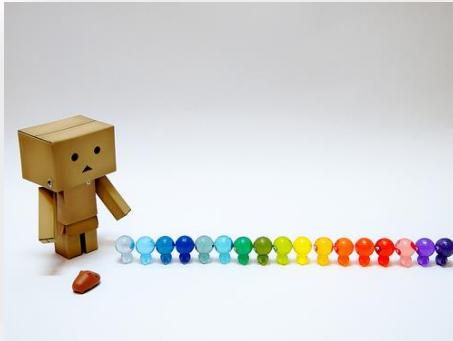
Can Recommender Systems be affected?

Can Recommender Systems be affected?

- Yes

Can Recommender Systems be affected?

- Yes, RecSys are socio-technical systems !
- RecSys help people on filtering noise, identifying relevant items from a large information space. They are usually optimized on accuracy and ranking metrics, not on fairness.

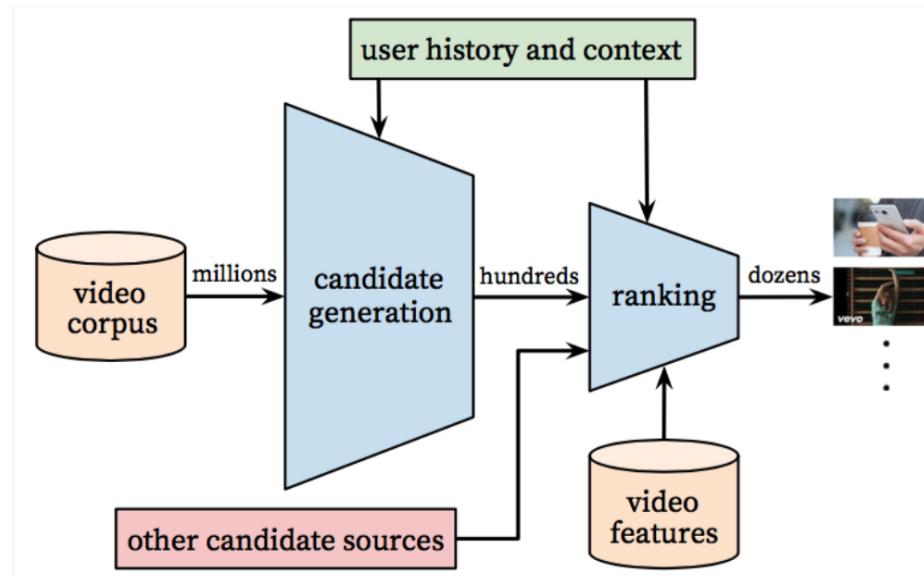


Can Recommender Systems be affected?

- Yes, RecSys are socio-technical systems !
- RecSys help people on filtering noise, identifying relevant items from a large information space. They are usually optimized on accuracy and ranking metrics, not fairness.
- The actual effects on user experience due to optimizing an accuracy/ranking metric are hard to predict.

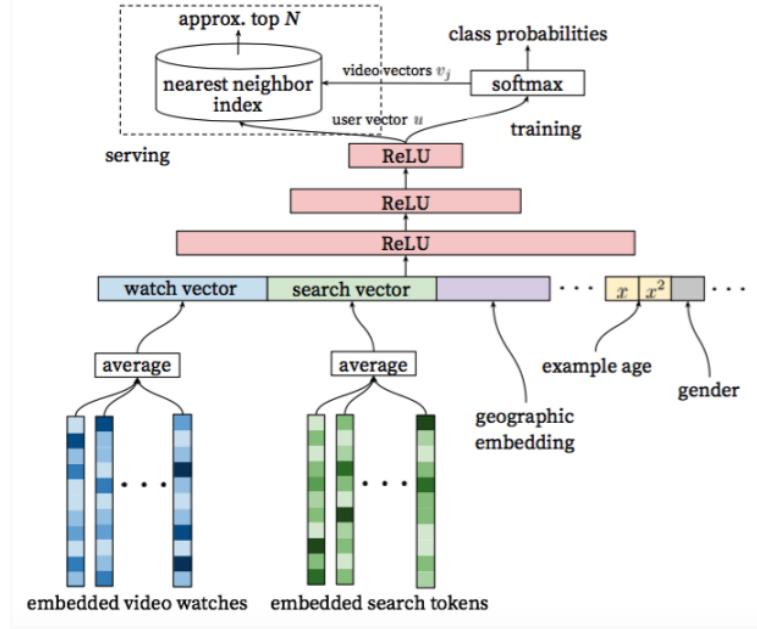
YouTube Deep Recommender System

- YouTube, ACM RecSys (2016)

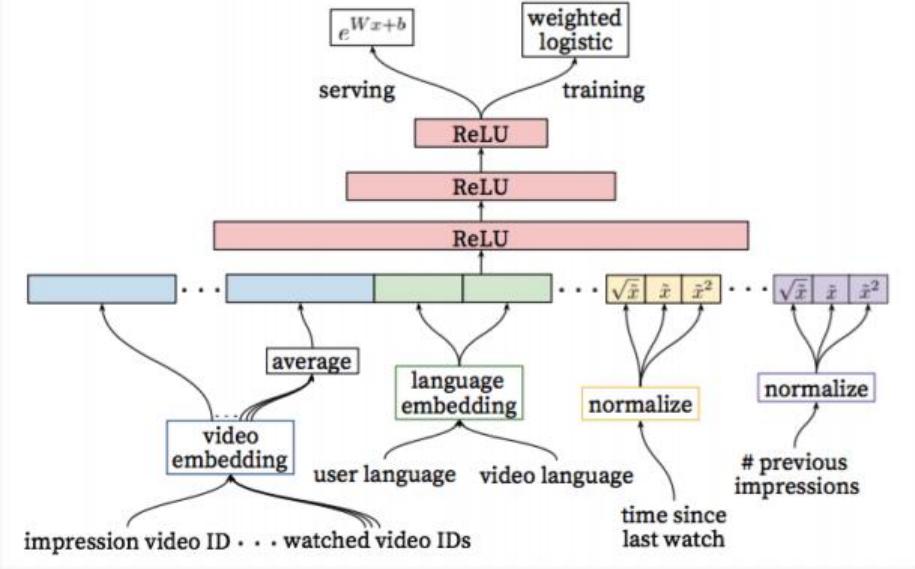


Covington, P., Adams, J., & Sargin, E. (2016, September). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 191-198). ACM.

Neural Networks



Candidate Generation



Ranking

What does YouTube RecSys try to learn ?

- Artificial Intelligence systems still do not decide what to learn: a human tells them the task(s).

What does YouTube try to learn ?

- Artificial Intelligence systems still do not decide what to learn: a human tells them
- In the case of YouTube, tasks are: 1) predict the next video watched, and 2) predict the time the user spent watching the next coming video.

What does YouTube try to learn ?

- Artificial Intelligence systems still do not decide what to learn: a human tells them
- In the case of YouTube, tasks are: 1) predict the next video watch, and 2) predict the time the user spent watching the next coming
- The system is never told to distinguish good from bad content quality (fake news, violence, etc.)

What Does YouTube RecSys recommends the most?

- Guillaume Chaslot
- He worked developing the first recommender system of YouTube.

How an ex-YouTube insider investigated its secret algorithm



What Does YouTube RecSys recommends the most?

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

How an ex-YouTube insider investigated its secret algorithm

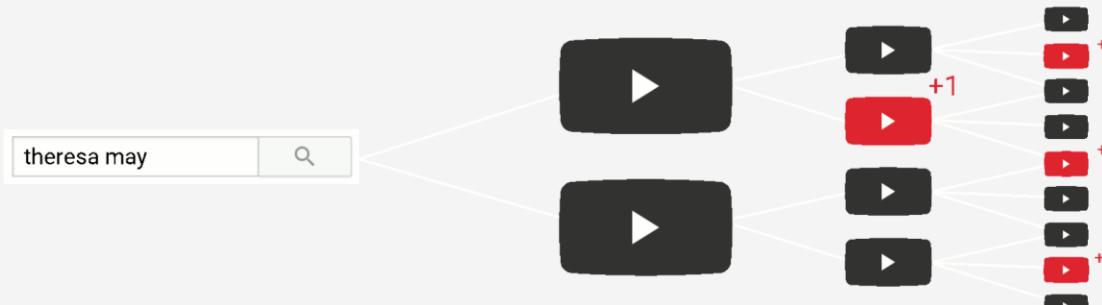


<https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot>

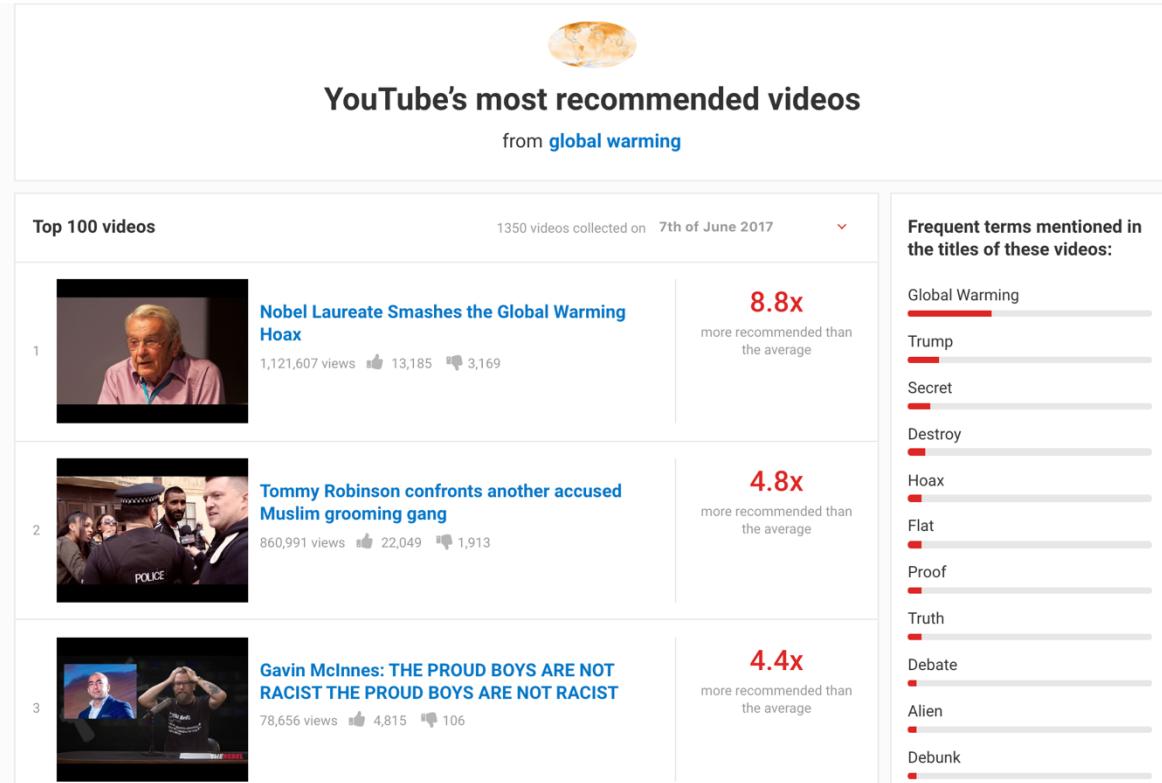
<https://algotransparency.org>

We created a robot to find out

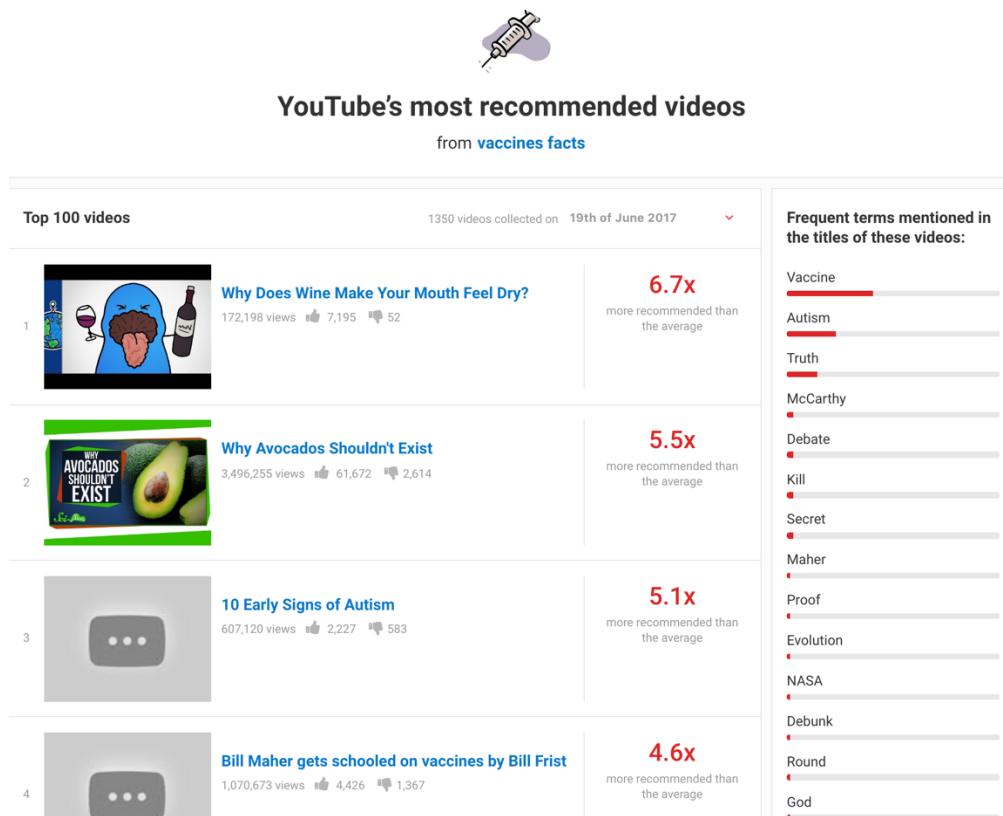
Using search terms, our robot crawls all the top recommended videos and keeps track of the ones which are recommended most often.



<https://algotransparency.org>



<https://algotransparency.org>



Do people consume YouTube Recommendations ?

Many Turn to YouTube for Children's Content, News, How-To Lessons

An analysis of videos suggested by the site's recommendation engine finds that users are directed toward progressively longer and more popular content

BY AARON SMITH, SKYE TOOR AND PATRICK VAN KESSEL



(MaaHoo Studio/Getty Images)

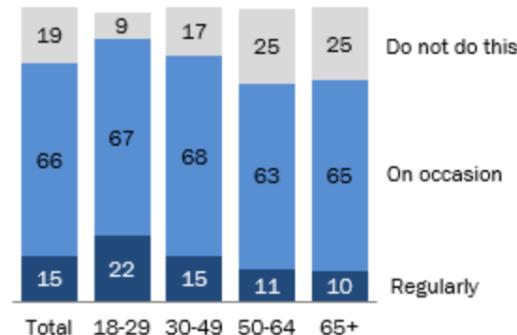
<https://www.pewinternet.org/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons>

People do follow recommendations, indeed.

- Study by Pew Research Center

Majority of YouTube users across a wide range of age groups watch recommended videos

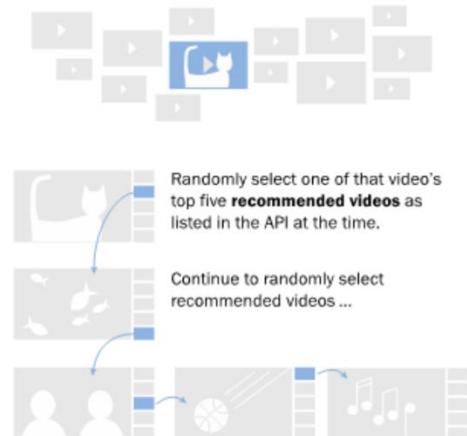
% of U.S. adults who use YouTube who say they watch the recommended videos that appear alongside the video they are currently watching ...



Methodology (Pew Research Internet)

How the Center conducted the ‘random walks’ analysis used in this report

Select a video at random from a list of 14,509 popular YouTube channels using the public YouTube API. This is the initial **starting video**.



... Until a total of five videos have been collected
(the starting video plus four recommended videos).

“Many Turn to YouTube for Children’s Content, News, How-To Lessons”

PEW RESEARCH CENTER

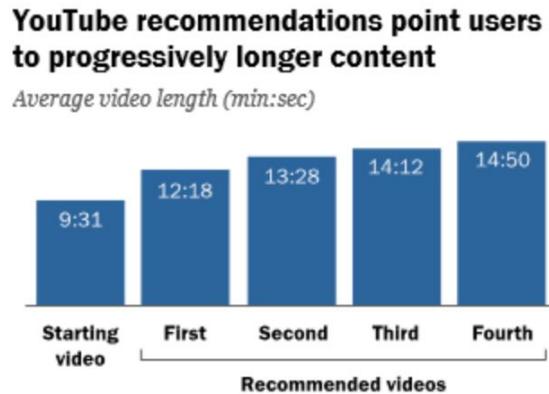
174,117 random walks
resulted in

696,468 total encounters
with

346,086 unique
recommended videos

Are there trends in terms of Video length ?

- Data from Pew Research Center



Source: Analysis of recommended videos collected via 174,117 five-step “random walks” beginning with videos posted to English-language YouTube channels with at least 250,000 subscribers, performed using the public YouTube API. Data collection performed July 18-Aug. 29, 2018.

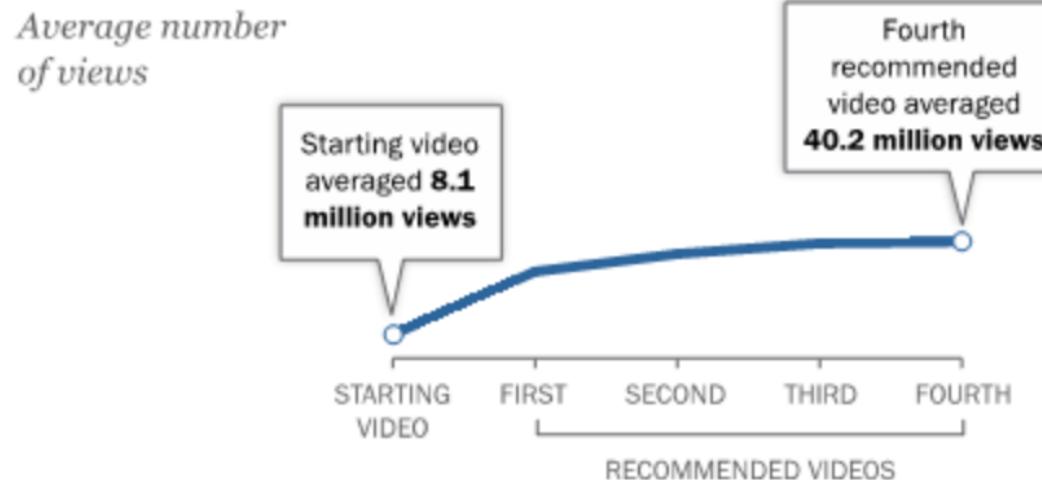
“Many Turn to YouTube for Children’s Content, News, How-To Lessons”

PEW RESEARCH CENTER

Are there trends in terms of Video Popularity ?

- Data from Pew Research Center

YouTube recommendations point to more popular content – regardless of starting criterion



New YouTube Recommender

- Presented in RecSys 2019, main change: multitask learning
- Still not addressing the issue of video quality / fake news

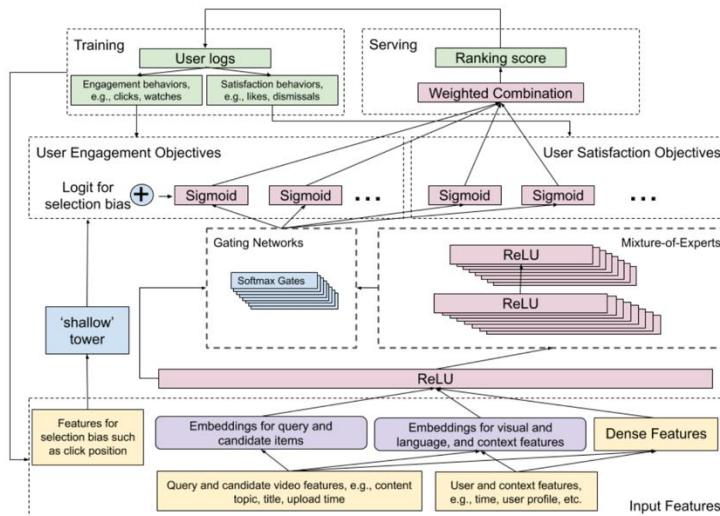


Figure 1: Model architecture of our proposed ranking system. It consumes user logs as training data, builds Multi-gate Mixture-of-Experts layers to predict two categories of user behaviors, i.e., engagement and satisfaction. It corrects ranking selection bias with a side-tower. On top, multiple predictions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

Best Paper RecSys 2021

- Matus Tomlein et al. 2021. *An Audit of Misinformation Filter Bubbles on YouTube: Bubble Bursting and Recent Behavior Changes*. In RecSys '21. DOI:
<https://doi.org/10.1145/3447548.3459900>

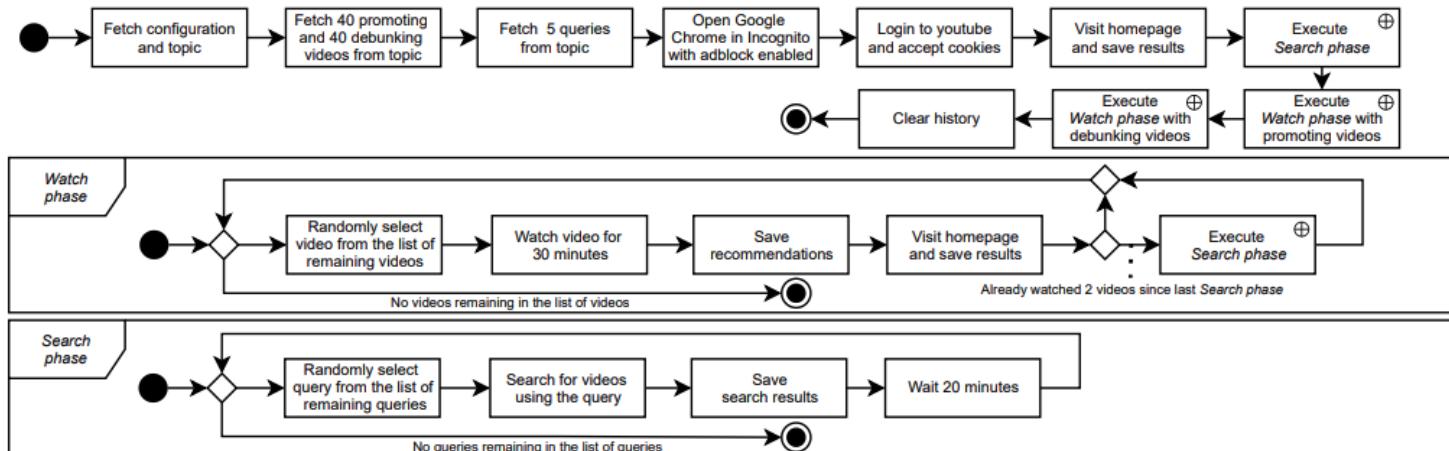


Figure 1: Agent scenario for creating and bursting misinformation filter bubbles

Should I care about this ?

- How can this affect my regular practice as professional developing or evaluating recommender systems ?

Law: What happened in May 25th, 2018 ?

- 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 practice ?

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

(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

(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 this purpose.

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, in particular, discriminatory effects such as unfavourable treatment 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](#)

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



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.

SIGN-IN

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

Council Home Legislation Calendar City Council Committees RSS Alerts

Details Reports

File #: Int 1696-2017 Version: A Type: Introduction Name: Automated decision systems used by agencies. Status: Enacted Committee: Committee on Technology

On agenda: 8/24/2017 Enactment date: 1/1/2018 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., Tamara 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. Letter to the Mayor, 10/16/17, 9. Letter to the Mayor, 10/16/17, 10. Letter to the Mayor, 10/16/17, 11. December 11, 2017 - Stated Meeting Agenda with Links to Files, 12. Hearing Testimony, 12/11/17, 13. Int. No. 1696-A (FINAL), 14. Fiscal Impact Statement, 15. Legislative Documents - Letter to the Mayor, 16. Local Law 16, 17. Minutes of the Stated Meeting - December 11, 2017

History (13) Text

13 records Draft Export

Date	Ver.	Prima 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 Decision Making

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.

POTENTIAL HARMS FROM ALGORITHMIC DECISION-MAKING

INDIVIDUAL HARMS	COLLECTIVE SOCIAL HARM
ILLEGAL DISCRIMINATION	UNFAIR PRACTICES
HIRING	A JL logo featuring a shield with a face and the letters 'JL' inside.
EMPLOYMENT	LOSS OF OPPORTUNITY
INSURANCE & SOCIAL BENEFITS	
HOUSING	
EDUCATION	
CREDIT	ECONOMIC LOSS
DIFFERENTIAL PRICES OF GOODS	
LOSS OF LIBERTY	SOCIAL STIGMATIZATION
INCREASED SURVEILLANCE	
STEREOTYPE REINFORCEMENT	
DIGNATORY HARMS	

Chart Contents Courtesy of Megan Smith, Former CTO of the United States

<http://gendershades.org/overview.html>

- How are researchers and practitioners addressing these issues ?

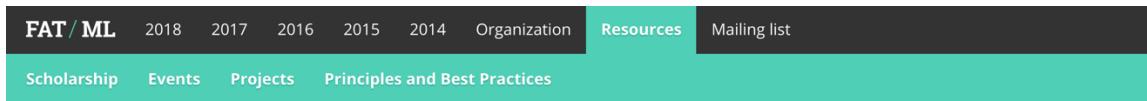
The FAT* Conference

- <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.



The FATML group

- The FATML group suggests best practices:
- <https://www.fatml.org/resources/principles-and-best-practices>



Principles for Accountable Algorithms and a Social Impact Statement for Algorithms

Principles for Accountable Algorithms

Automated decision making algorithms are now used throughout industry and government, underpinning many processes from dynamic pricing to employment practices to criminal sentencing. Given that such algorithmically informed decisions have the potential for significant societal impact, the goal of this document is to help developers and product managers design and implement algorithmic systems in publicly accountable ways. Accountability in this context includes an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms.

We begin by outlining five equally important guiding principles that follow from this premise:

Algorithms and the data that drive them are designed and created by people -- There is always a human ultimately responsible for decisions made or informed by an algorithm. "The algorithm did it" is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes.

Workshops & Tutorials

- Tutorial on Fairness & Discrimination in Retrieval & Recommendation: M. Ekstrand, F. Diaz, R. Burke (SIGIR & RecSys 2019) <https://boi.st/FairIPTutorial>
- Tutorial on ExplainAble Recommendation and Search: Y. Zhang, Q. Ai, J. Mao, X. Chen (SIGIR 2019)
- Tutorial on Fairness and Transparency in Ranking: Carlos Castillo (LA-Web 2019, DAB 2018)

How to Measure, Study and Prevent Bias in RecSys?

- Some definitions
- Explainability and transparency in RecSys
- Fairness in RecSys
- Open challenges
- Summary and conclusions

FAT definitions

- Fairness
- Accountability
- Transparency

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 Nissenbaum (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 Nissenbaum (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.

Ley Antidiscriminación en Chile

- Ley 20609 también llamada “Ley Zamudio”
- <https://www.bcn.cl/leychile/navegar?idNorma=1042092>

Promulgación: 12-JUL-2012

Publicación: 24-JUL-2012

Versión: Última Versión - 27-DIC-2019

Última modificación: 02-MAY-2019 - Ley 21155

Materias: Acción de No Discriminación Arbitraria, Constitución Política de la República, Código del Trabajo, Código Penal, Código de Procedimiento Civil, Estatuto Administrativo, Estatuto Administrativo para Funcionarios Municipales, Ley no. 18.883, Ley no. 20.609, Ley Antidiscriminación

Resumen: Establece medidas contra las discriminaciones arbitrarias y un procedimiento judicial para re establecer ... [ver más >>](#)

Ley 20609 (ley Zamudio) en Chile

- Para la aplicación de esta ley, se entiende por discriminación arbitraria toda distinción, exclusión o restricción que carezca de justificación razonable, efectuada por agentes del Estado o particulares, y que cause privación, perturbación o amenaza en el ejercicio legítimo de los derechos fundamentales establecidos en la Constitución Política de la República o en los tratados internacionales sobre derechos humanos ratificados por Chile y que se encuentren vigentes.
- Se establece que una discriminación es arbitraria cuando se funda en motivos tales como la raza o etnia, la nacionalidad, la situación socioeconómica, el idioma, la ideología u opinión política, la religión o creencia, la sindicación o participación en organizaciones gremiales o la falta de ellas, el sexo, la orientación sexual, la identidad de género, el estado civil, la edad, la filiación, la apariencia personal y la enfermedad o discapacidad.
- Se crea la acción judicial de no discriminación arbitraria y el procedimiento para su tramitación en los tribunales. La pueden presentar los directamente afectados por una acción u omisión que importe discriminación arbitraria, o su representante legal o quien tenga de hecho el cuidado personal o la educación del afectado, circunstancia esta última que deberá señalarse en la presentación. Esta acción debe ser presentada ante el juez de letras de su domicilio o ante el del domicilio del responsable de dicha acción u omisión.

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)

1. EXPLAINABILITY & TRANSPARENCY

FAT in Recommender Systems

- Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250). ACM.
- Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems* (pp. 830-831). ACM.

FAT in Recommender Systems (movies)

- Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250). ACM.

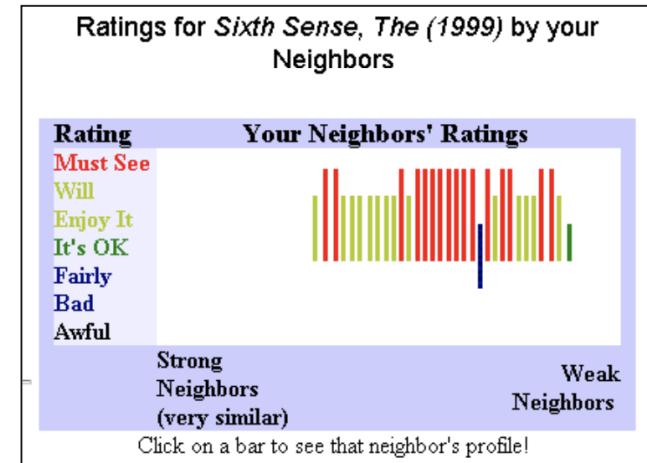
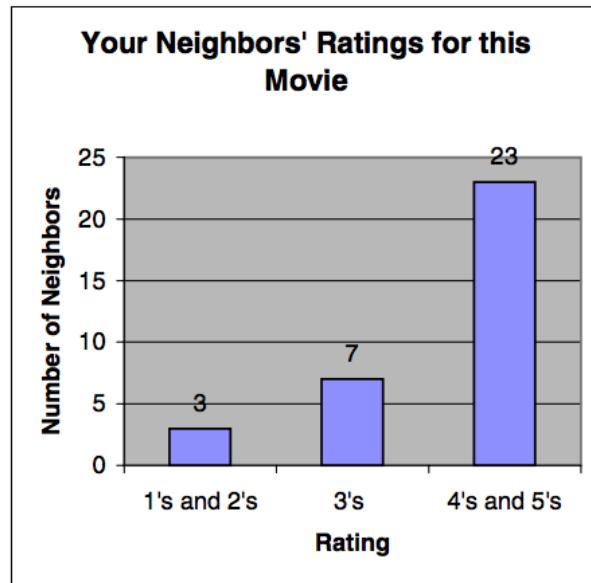
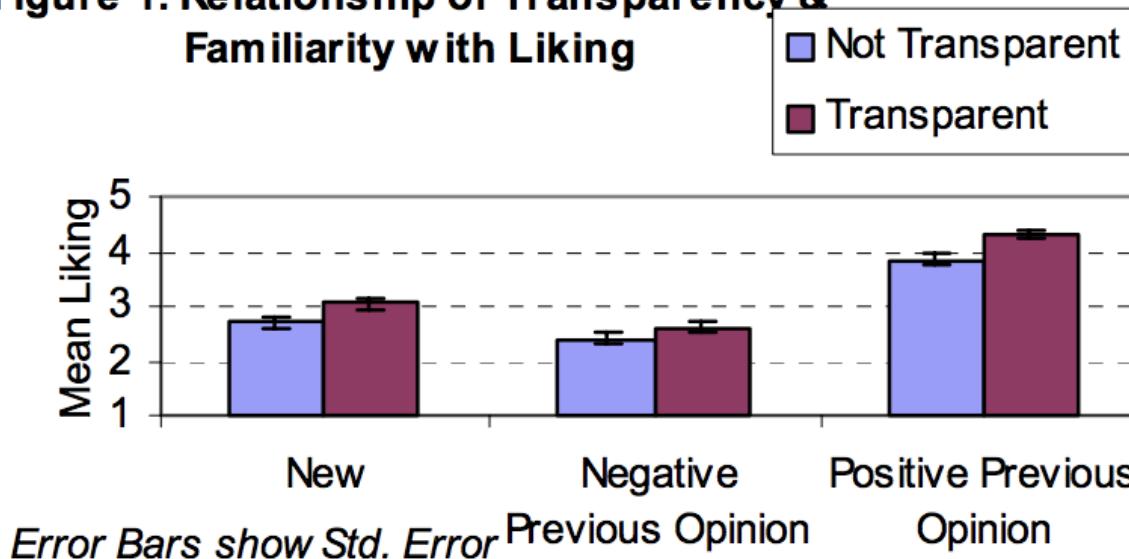


Figure 4. A screen explaining the recommendation for the movie “The Sixth Sense.” Each bar represents a rating of a neighbor. Upwardly trending bars are positive ratings, while downward trending ones are negative. The x-axis represents similarity to the user.

FAT in Recommender Systems (music)

- Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems* (pp. 830-831). ACM.

Figure 1: Relationship of Transparency & Familiarity with Liking



FAT in Recommender Systems II

- Tintarev, N., & Masthoff, J. (2007). A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop* (pp. 801-810). IEEE.
- Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5), 399-439.
- Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook* (pp. 353-382). Springer, Boston, MA.

FAT in Recommender Systems II

- Tintarev, N., & Masthoff, J. (2007). A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop* (pp. 801-810). IEEE.
- Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5), 399-439.
- Tintarev, N., & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In *Recommender systems handbook* (pp. 353-382). Springer, Boston, MA.

RecSys: Explanatory Goals and Definitions

Aim	Definition
Transparency (Tra.)	Explain how the system works
Scrutability (Scr.)	Allow users to tell the system it is wrong
Trust	Increase users' confidence in the system
Effectiveness (Efk.)	Help users make good decisions
Persuasiveness (Pers.)	Convince users to try or buy
Efficiency (Efc.)	Help users make decisions faster
Satisfaction (Sat.)	Increase the ease of usability or enjoyment

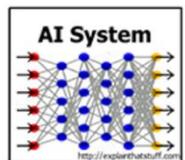
Tintarev, N., & Masthoff, J. (2007). A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop* (pp. 801-810). IEEE.

XAI (2017)

- XAI is a term coined by David Gunning, program manager at DARPA

Explainable Artificial
Intelligence (XAI)

Mr. David Gunning



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?



Mr. David Gunning

Information Innovation Office (I2O)

Program Manager

Figure 1. The Need for Explainable AI

XAI for Recommender Systems

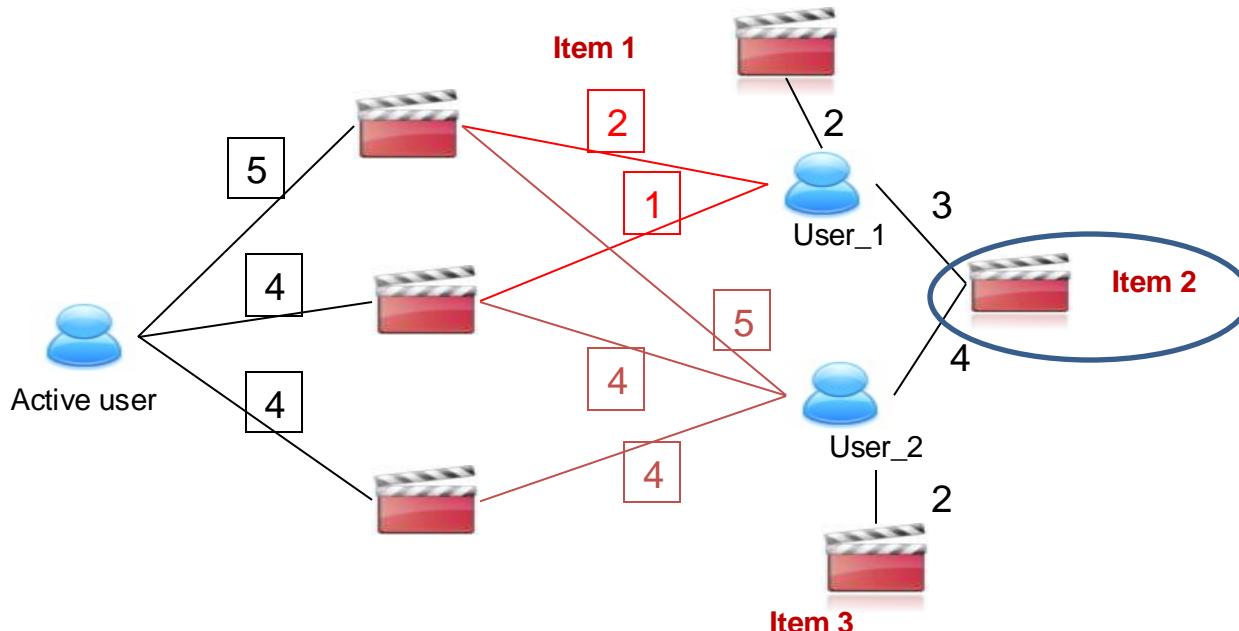
- First generation of approaches for Recommender Systems were easily to explain: User and Item based CF, Content-based, Rule-based
- The Second generation of RecSys, based on Matrix Factorization made the process more difficult: latent user and item representation
- The Third generation based on Deep Learning makes accountability and transparency even more difficult !

1st generation of RecSys

- Algorithms were simple and intuitive (User-based KNN, Item-Based KNN, Content-based, Case-based)
- Provide explanations for items recommended would not require a big engineering effort

1st generation of RecSys

- User Based KNN



Explanation: Users who have similar ratings with you highly rated this item

1st generation of RecSys

- Content Based



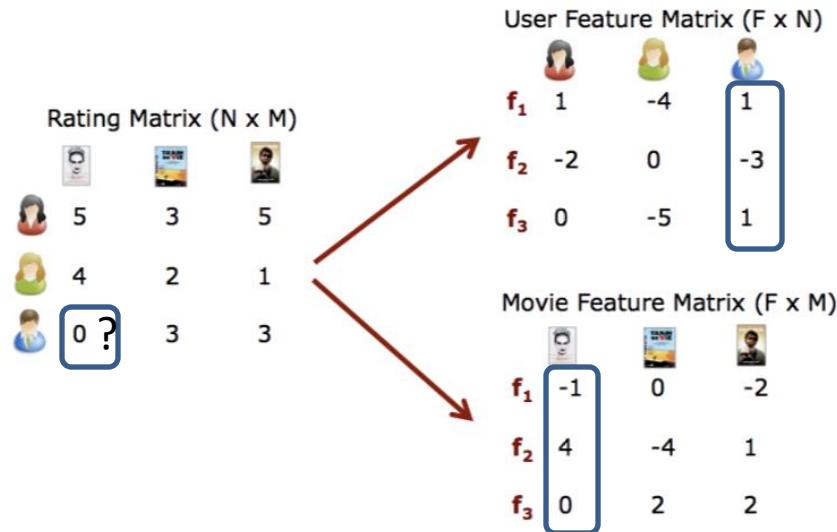
Explanation: This items has similar content (features: description, actors, director, genre) to what you have liked in the past

XAI for Recommender Systems

- First generation of approaches for Recommender Systems were easily to explain: User and Item based CF, Content-based, Rule-based
- The Second generation of RecSys, based on Matrix Factorization made the process more difficult: latent user and item representation
- The Third generation based on Deep Learning makes accountability and transparency even more difficult !

2nd generation of RecSys

- Matrix Factorization - latent factor models: difficult to explain



$$\min_{q^*, p^*} \sum_{(u,i) \in K} (r_{ui} - q_i^T \cdot p_u)^2 + \lambda(\|q_i\|^2 + \|p_u\|^2)$$

2nd generation of RecSys

- Alternatives: try to assign explicit meaning to latent factor models:
 - Zhang, Y., Lai, G., Zhang, M., Zhang, Y., Liu, Y., & Ma, S. (2014). **Explicit factor models for explainable recommendation based on phrase-level sentiment analysis.** In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 83-92). ACM.
 - Chen, X., Qin, Z., Zhang, Y., & Xu, T. (2016). **Learning to rank features for recommendation over multiple categories.** In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 305-314). ACM.
 - Wang, N., Wang, H., Jia, Y., & Yin, Y. (2018). **Explainable recommendation via multi-task learning in opinionated text data.** In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 165-174). ACM.

2nd generation of RecSys

- Zhang et al (2014) “Explicit factor models for explainable recommendation based on phrase-level sentiment analysis”

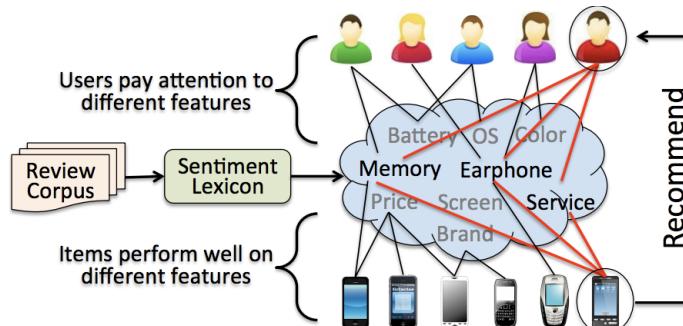


Figure 1: The product feature word and user opinion word pairs are extracted from user review corpus to construct the sentiment lexicon, and the feature word set further serves as the explicit feature space. An item would be recommended if it performs well on the features that a user cares.

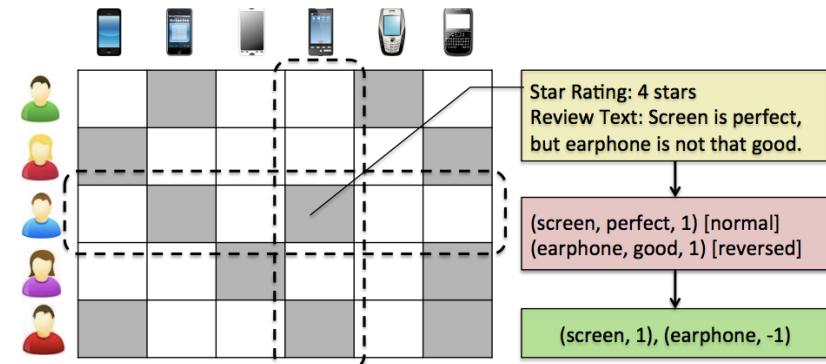


Figure 2: An example of user-item review matrix, where each shaded block is a review made by a user towards an item; the entries included in the review are extracted, and further transformed to feature scores while considering the negation words.

XAI for Recommender Systems

- First generation of approaches for Recommender Systems were easily to explain: User and Item based CF, Content-based, Rule-based
- The Second generation of RecSys, based on Matrix Factorization made the process more difficult: latent user and item representation
- The Third generation based on Deep Learning makes accountability and transparency even more difficult !

3rd Generation of RecSys

- In Matrix factorization we had one level of interactions, with deep learning we can have many! Making explanations more complex

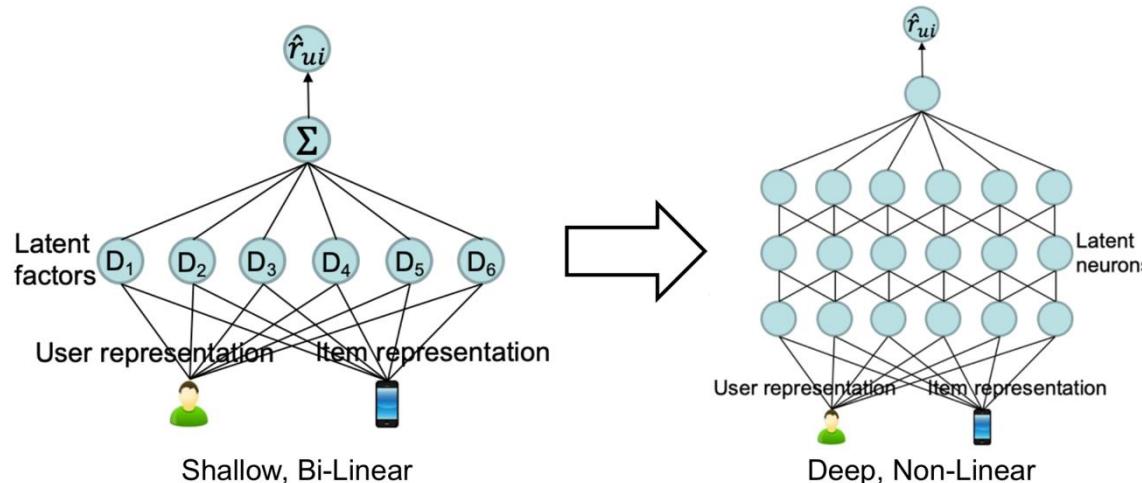


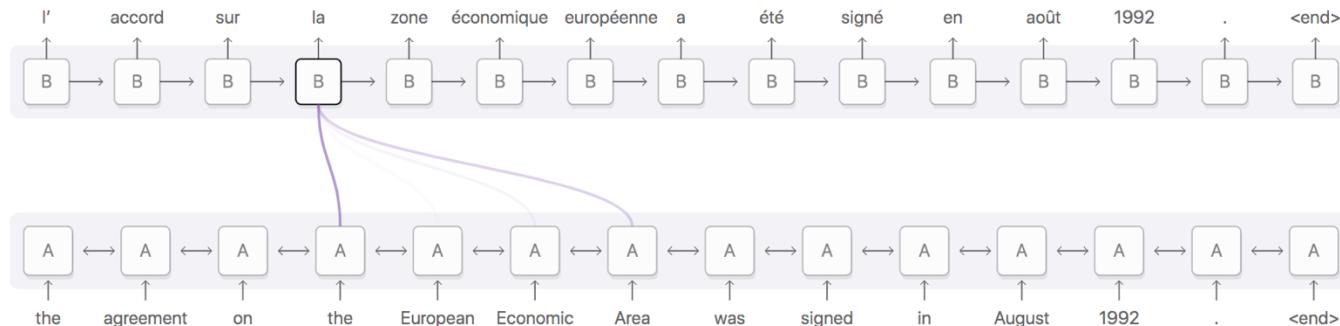
Image from Zhang et al. (2019) Tutorial on Explainable Recommendation and Search

3rd Generation of RecSys

- Alternatives: use attention mechanism within the neural architecture (over text or images)
- Generate explanations directly (Natural Language Generation)
- Use a model agnostic approach: generate explanations after recommendation (LIME, SHAP, etc.)

Neural Attention

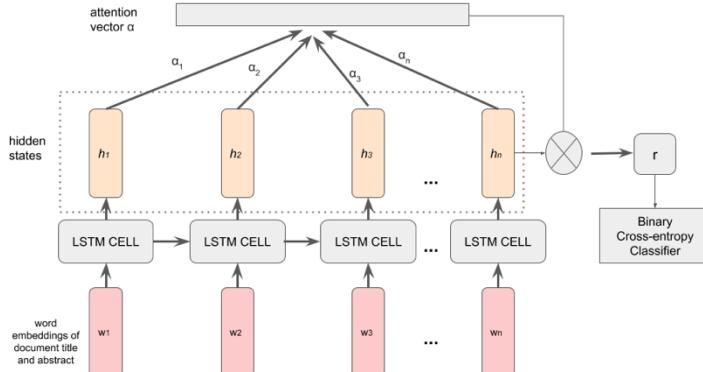
- Attention in neural networks is a mechanism which allows the model to focus selectively during the learning process.
- Eventually, we can observe where the network was attending to in order to make a prediction.



Olah, C., & Carter, S. (2016). Attention and augmented recurrent neural networks. *Distill*, 1(9), e1. <http://doi.org/10.23915/distill.00001>

Neural Attention

- Example of document classification: Does the model attend to reasonable words ?



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

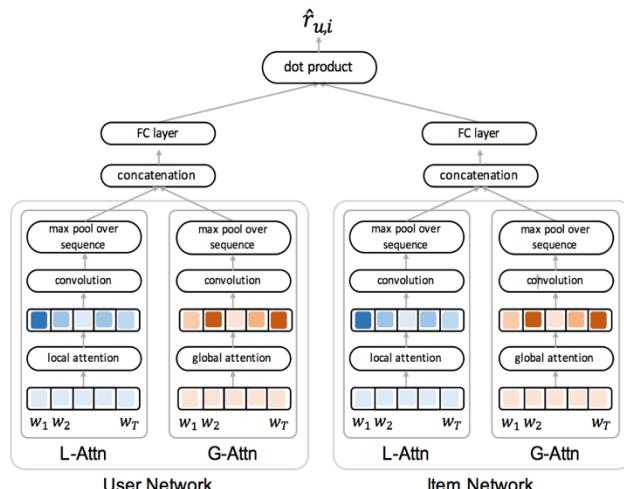
Prediction: Not Relevant (NR)

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

3rd Generation of RecSys

- Seo, S., Huang, J., Yang, H., & Liu, Y. (2017). **Interpretable convolutional neural networks with dual local and global attention for review rating prediction.** RecSys 2017.



Yelp (user), L-Attn-only model: local attention

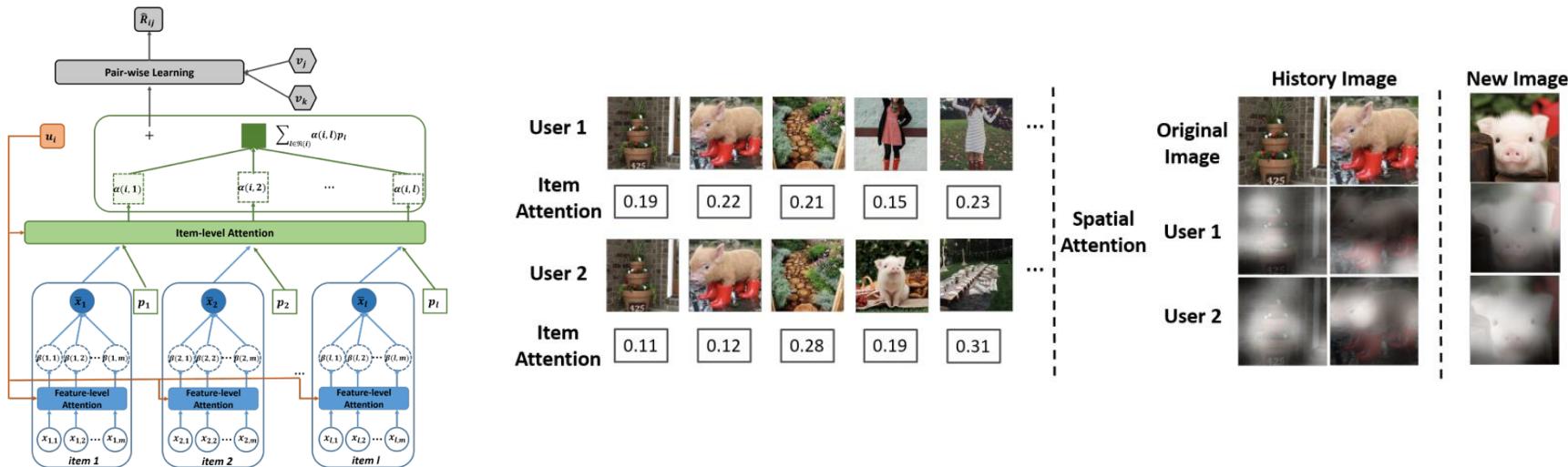
They carry some rare things that you can't find anywhere else. The staff is pretty damn cool too best in Arizona. I prefer ma-and-pa. They treat you the best and they value your business extreme. They are good people great atmosphere and music. I definitely believe that Lux has the best coffee I've ever had at this point. Screw all my previous reviews. This place has coffee down, they make damn good toast too.

Yelp (user), G-Attn-only model: global attention

They carry some rare things that you can't find anywhere else. The staff is pretty damn cool too best in Arizona. I prefer ma-and-pa. They treat you the best and they value your business extreme. They are good people great atmosphere and music. I definitely believe that Lux has the best coffee I've ever had at this point. Screw all my previous reviews. This place has coffee down, they make damn good toast too.

3rd Generation of RecSys

- Chen, J., Zhang, H., He, X., Nie, L., Liu, W., & Chua, T. S. (2017). Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. SIGIR 2017.



3rd Generation of RecSys

- Li, P., Wang, Z., Ren, Z., Bing, L., & Lam, W. (2017,). Neural rating regression with abstractive tips generation for recommendation.
SIGIR 2017

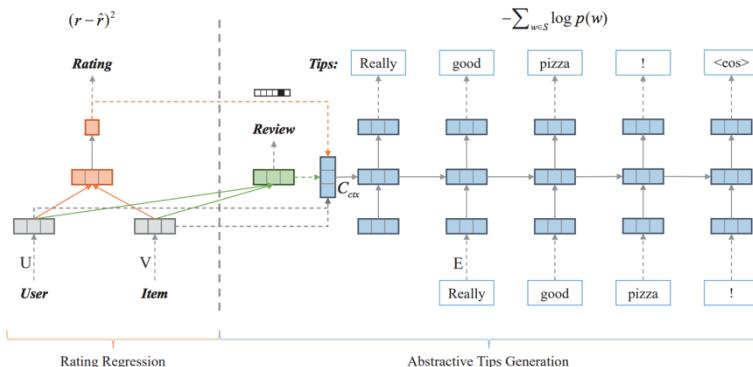


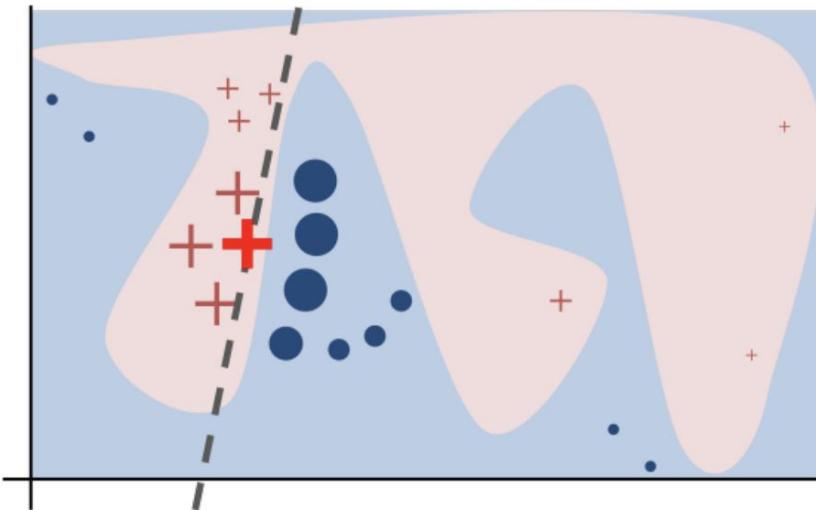
Figure 2: Our proposed framework NRT for rating regression and abstractive tips generation.

Table 10: Examples of the predicted ratings and the generated tips. The first line of each group shows the generated rating and tips. The second line shows the ground truth.

Rating	Tips
4.64 5	<i>This is a great product for a great price.</i> Great product at a great price.
4.87 5	<i>I purchased this as a replacement and it is a perfect fit and the sound is excellent.</i> Amazing sound.
4.69 4	<i>I have been using these for a couple of months.</i> Plenty of wire gets signals and power to my amp just fine quality wise.
4.87 5	<i>One of my favorite movies.</i> This is a movie that is not to be missed.
4.07 4	<i>Why do people hate this film.</i> Universal why didnt your company release this edition in 1999.

Adapting current XAI approaches to RecSys

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



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

Prediction probabilities

atheism	0.58
christian	0.42

atheism

Posting	0.15
Host	0.14
NNTP	0.11
edu	0.04
have	0.01
There	0.01

christian

Text with highlighted words

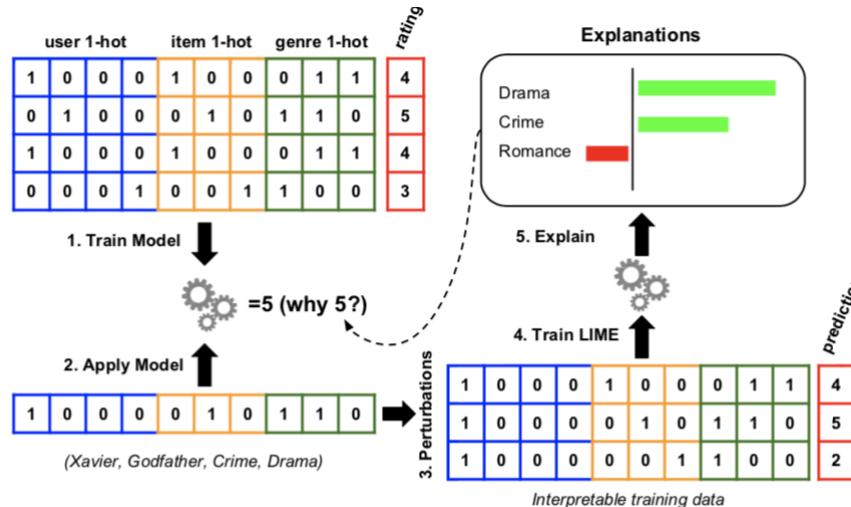
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I **have** and I **have** not seen an answer on the net. If anyone has a contact please post on the net or email me.

Adapting current XAI approaches to RecSys

- Adapting LIME to recommendation: Nóbrega, C., & Marinho, L. (2019). Towards explaining recommendations through local surrogate models. ACM/SIGAPP Symposium on Applied Computing.



This survey is not exhaustive

- I strongly recommend visiting
- <https://sites.google.com/view/ears-tutorial/>



Yongfeng Zhang

Assistant Professor

Rutgers University, New
Brunswick, NJ, USA



Jiaxin Mao

Postdoc

Tsinghua University, Beijing,
China



Qingyao Ai

Assistant Professor

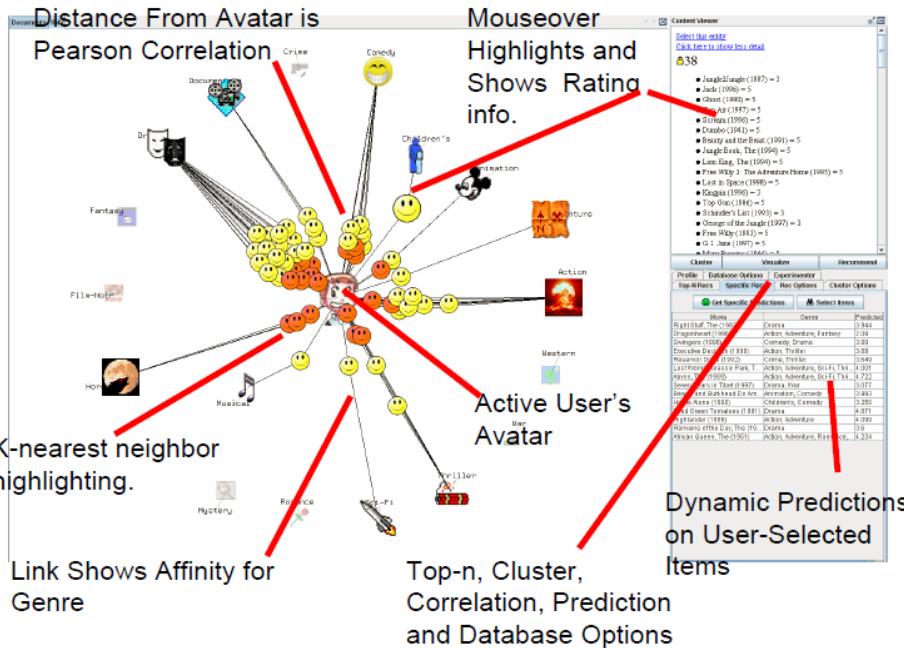
University of Utah, Salt Lake
City, UT, USA

Accountability: the Role of Interactive Visualization

- PeerChooser (O'Donovan et al, 2008)
- SmallWorlds (Gretarsson et al, 2010)
- TasteWeights (Bostandjev et al. 2012, Knijnenburg et al. 2012)
- TalkExplorer/Aduna (Verbert et al. 2013)
- SetFusion (Parra et al., 2014)
- Moodplay (Andjelkovic et al., 2016)
- 3D Inspector (Loepp et al, 2017)

Peerchooser - 1

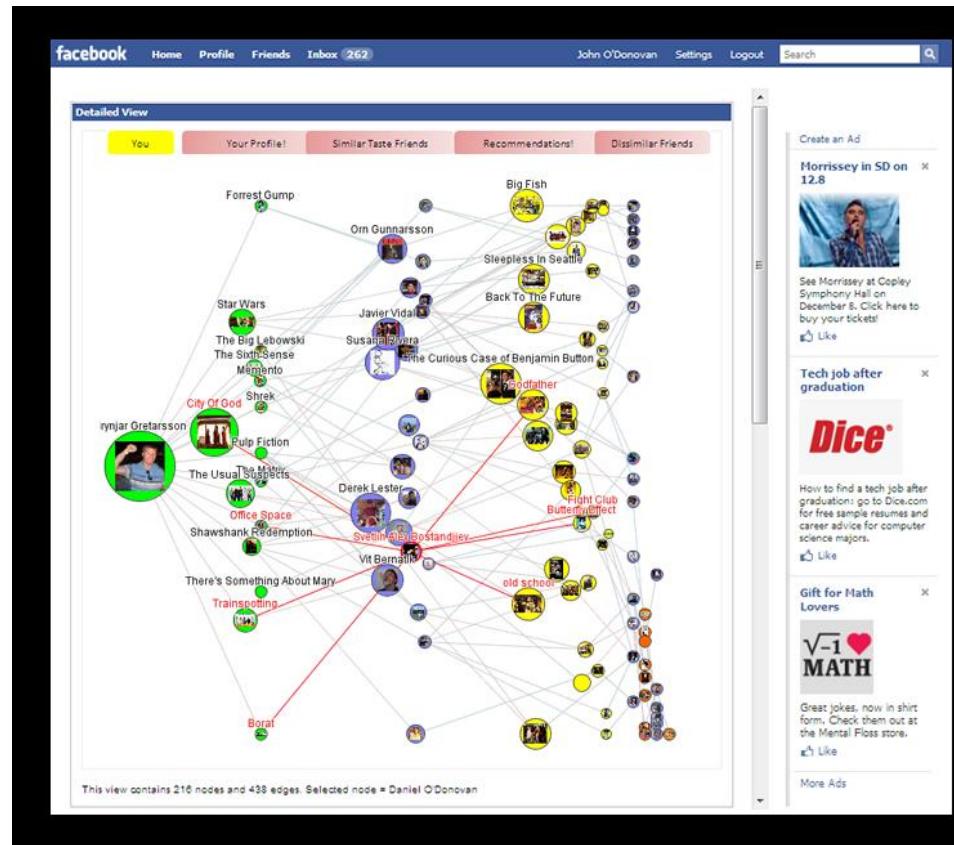
- PeerChooser (CHI 2008)
- John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, Tobias Hollerer



2: Annotated Screenshot of PeerChooser's Interactive Interface.

Smallworlds – 2

- SmallWorlds:
Visualizing Social
Recommendations
(IEEE-TVCG 2010)
- Brynjar Gretarsson,
John O'Donovan ,
Svetlin Bostandjiev,
Christopher Hall,
Tobias Höllerer
- User study with 17
users



TasteWeights - 3

- TasteWeights: a visual interactive hybrid recommender system (RecSys 2012)
- Bostandjiev, S., O'Donovan, J., & Höllerer, T.



Controllability:
Sliders that let users control the importance of preferences and contexts

Inspectability: lines that connect recommended items with contexts and user preferences

Also : Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., & Kobsa, A. (2012). Inspectability and control in social recommenders. In Proceedings of ACM RecSys.

TalkExplorer - 4

- Visualizing recommendations to support exploration, transparency and controllability (IUI 2014)
- Verbert, K., Parra, D., Brusilovsky, P., & Duval, E.

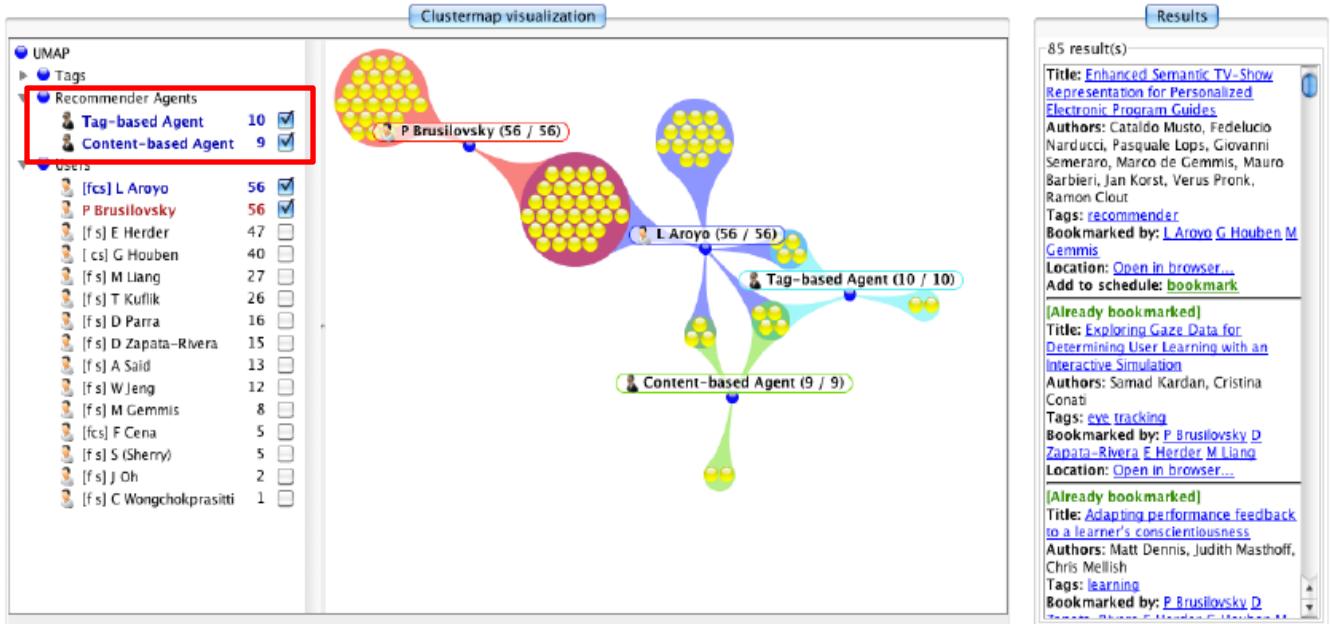


Figure 2: TalkExplorer

SetFusion - 5

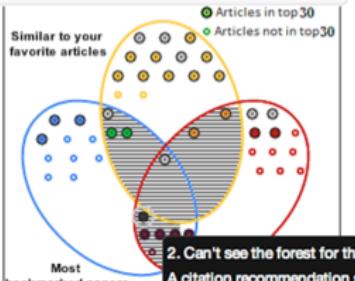
- See what you want to see: visual user-driven approach for hybrid recommendation (IUI 2014)
- Denis Parra, Peter Brusilovsky, and Christoph Trattner.

Tune weights of the recommender methods: (b)

Most bookmarked papers
Similar to your favorite articles
Frequently cited authors in ACM DL

Update Recommendation List →

* Hover over circles to explore articles
* Click on the diagram to highlight subsets



(a)

2. Can't see the forest for the trees? A citation recommendation system [\[see abstract\]](#)
by C. Lee Giles, Cornelia Caragea, Adrian Silvescu, Prasenjit Mitra

3. When thumbnails are and are not enough: Factors behind users [\[see abstract\]](#)
by Dan Albertson

7. Gendered Artifacts and User Agency [\[see abstract\]](#)
by Andrea R. Marshall, Jennifer A. Rode

8. Two Paths to Motivation through Game Design Elements: Reward-Based Gamification and Meaningful Gamification [\[see abstract\]](#)
by Scott Nicholson

9. Automatic Identifying Search Tactic in Individual Information Seeking: A Hidden Markov Model Approach [\[see abstract\]](#)
by Zhen Yue, Shuguang Han, Daqing He

11. Old Maps and Open Data Networks [\[see abstract\]](#)
by Werner Robitz, Carl Lagoze, Bernhard Haslhofer, Keith Newman, Amanda Stefanik

14. Effects of User Identity Information On Key Answer Outcomes in Social Q&A [\[see abstract\]](#)
by Erik Choi, Craig Scott, Chirag Shah

15. Ebooks and cross generational perceived privacy issues [\[see abstract\]](#)
by Jennifer Sue Thiele, Renee Kapusniak

16. Toward a mesoscopic analysis of the temporal evolution of scientific collaboration networks [\[see abstract\]](#)

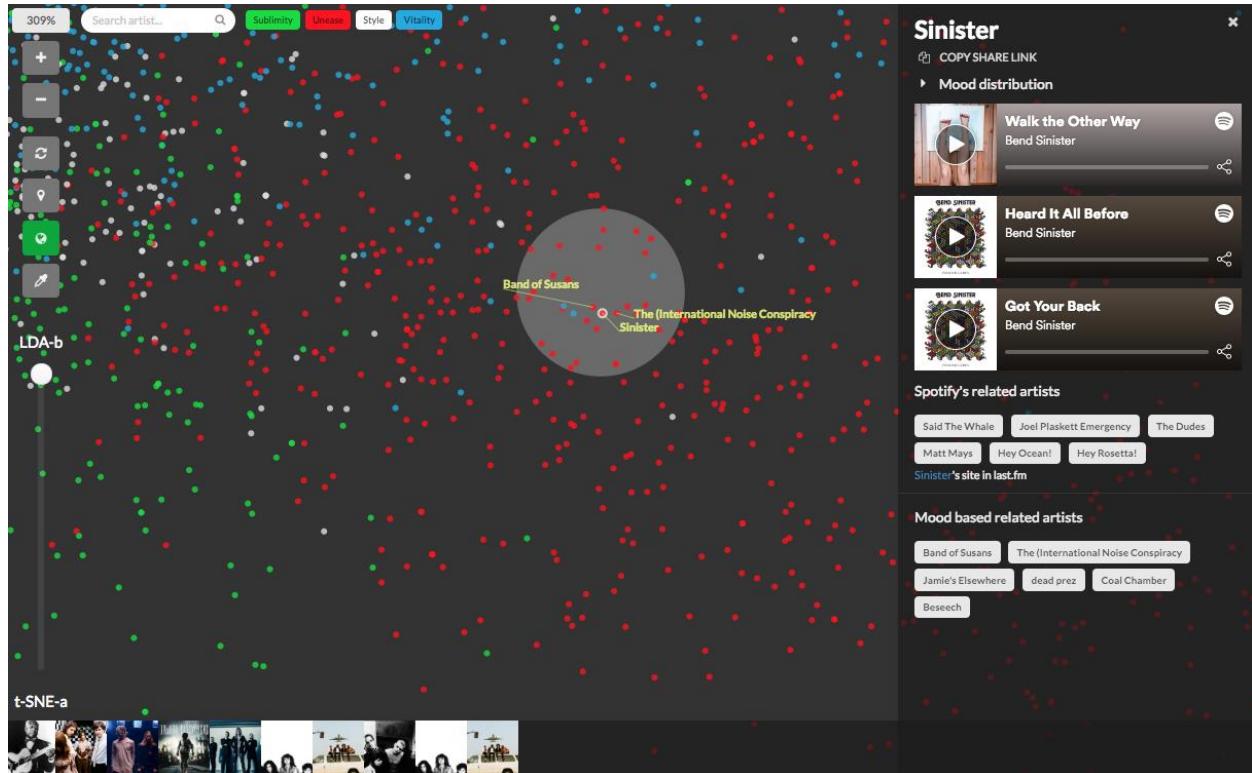
SetFusion: A Controllable Hybrid Recommender

Parra, D., Brusilovsky, P., Trattner, C.

IUI 2014, Haifa, Israel

Moodplay - 6

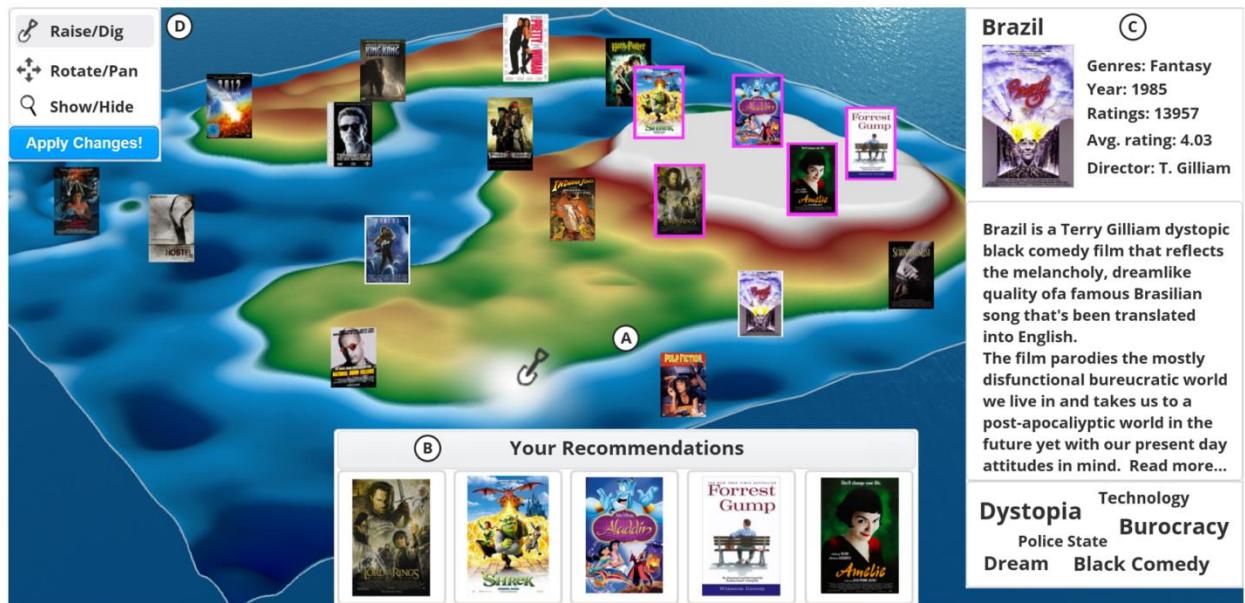
- Moodplay: Interactive Mood-based Music Discovery and Recommendation.
(UMAP 2016)
- Ivana Andjelkovic,
Denis Parra, and John
O'Donovan.



<http://moodplay.pythonanywhere.com/>

3D Landscape - 7

- A 3D item space visualization for presenting and manipulating user preferences in collaborative filtering (IUI 2017)
- Kunkel, J., Loepp, B., & Ziegler, J.

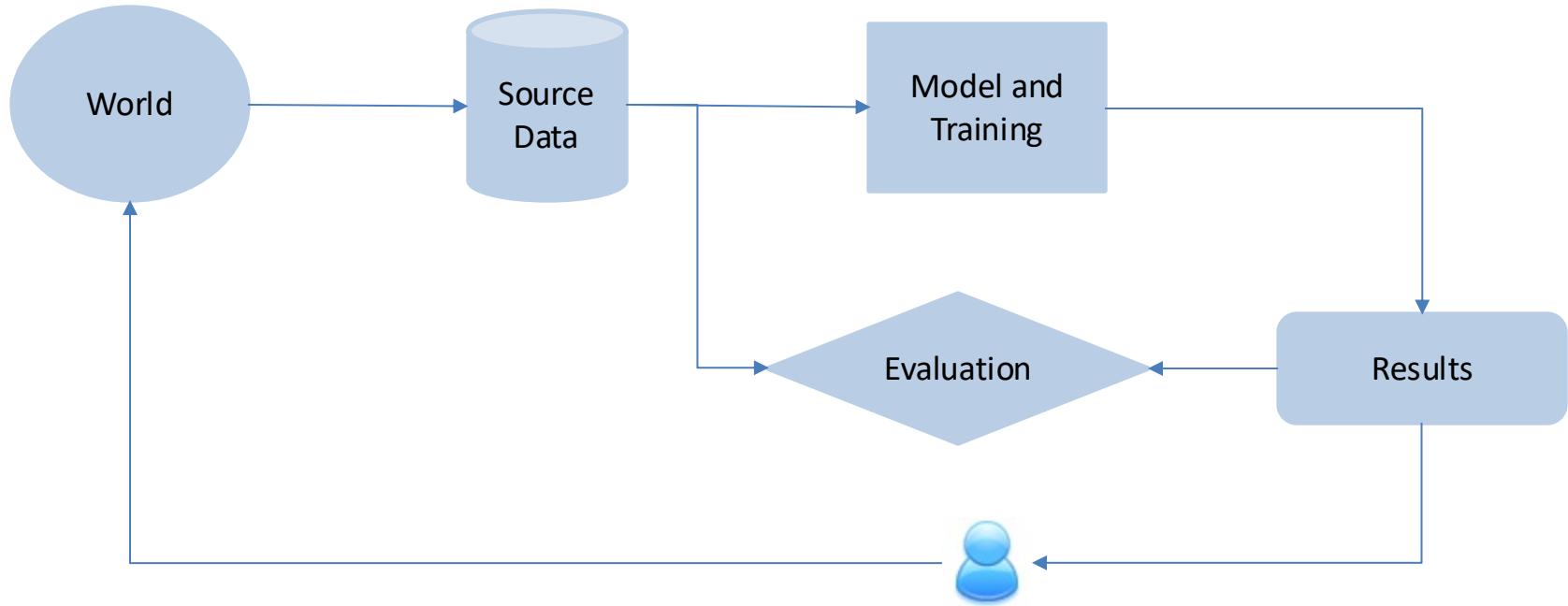


Open Challenges

- Recent advance in NLP models, neural attention architectures (the transformer), and generative models provide a big potential for this area. Notice that interpretable != transparent.
- Visualization has not been deeply explored for supporting transparency and explainability in recommender systems, and it is an open field for further research.
- For a glimpse of what can be done combining the aforementioned points, check <https://distill.pub> as well as <https://visxai.io>

2. FAIRNESS IN RECSYS

Where does Unfairness come from ?



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

World & Data Bias

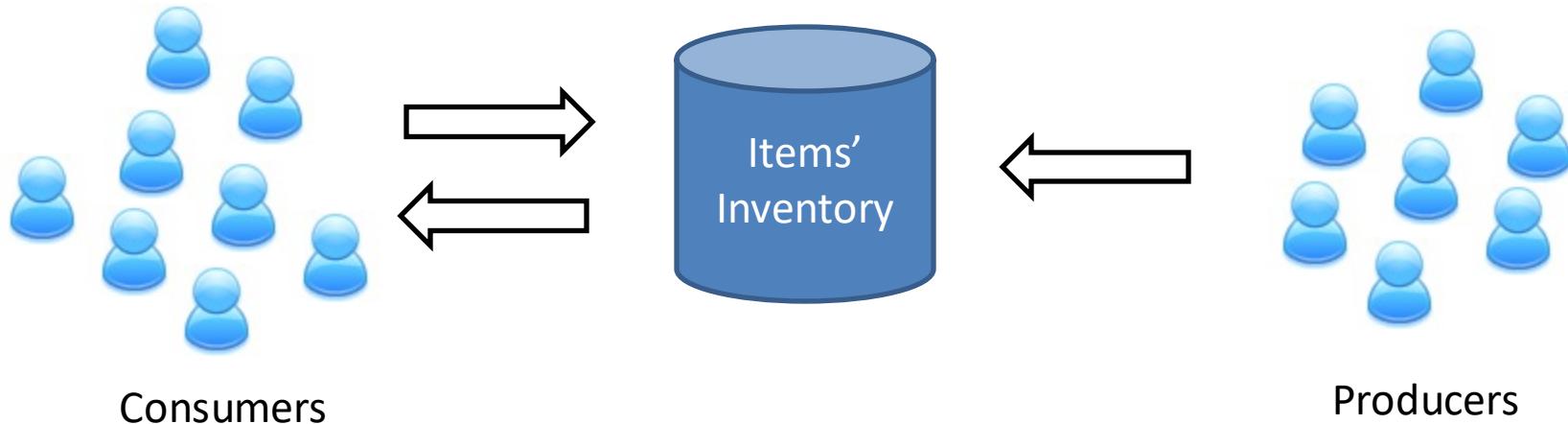


Figure 5: Words most associated with women (left) and men (right), estimated with *Pointwise Mutual Information*. Font size is inversely proportional to PMI rank. Color encodes frequency (the darker, the more frequent).

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.

Consumer vs. Producer Bias

- The figure, from Ekstrand, Diaz, Burke (2019) show different stakeholders on Information Access Systems



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

Consumer Bias in RecSys

All The Cool Kids, How Do They Fit In? Popularity and Demographic Biases in Recommender Evaluation and Effectiveness^{*†}

Michael D. Ekstrand
Mucun Tian

Ion Madrazo Azpiazu
Jennifer D. Ekstrand
Oghenemaro Anuyah
David McNeill
Maria Soledad Pera

People and Information Research Team, Dept. of Computer Science, Boise State University

MICHAELEKSTRAND@BOISESTATE.EDU

MUCUNTIAN@U.BOISESTATE.EDU

IONMADRAZO@U.BOISESTATE.EDU

JENNIFEREKSTRAND@U.BOISESTATE.EDU

OGHENEMAROANUYAH@U.BOISESTATE.EDU

DAVIDMCNEILL@U.BOISESTATE.EDU

SOLEPERA@BOISESTATE.EDU

Biases in RecSys

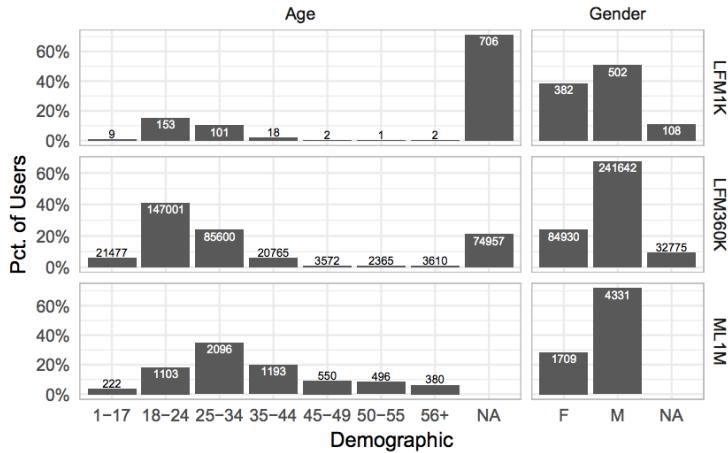


Figure 1: User distribution by demographic group. Numbers in bars are the number of users in that bin.

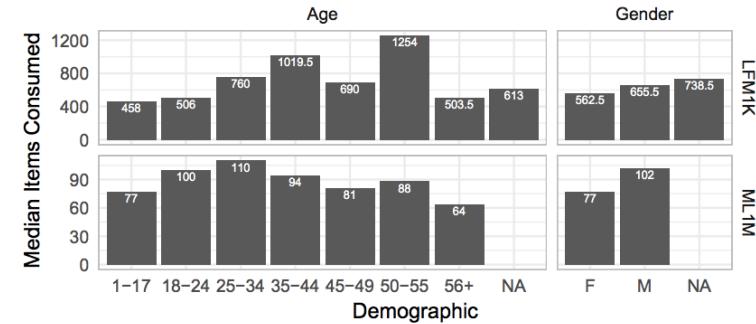


Figure 2: Median items consumed by users in each demographic group. We omit LFM360K since it only contains each user's top 50 artists.

Biases in RecSys

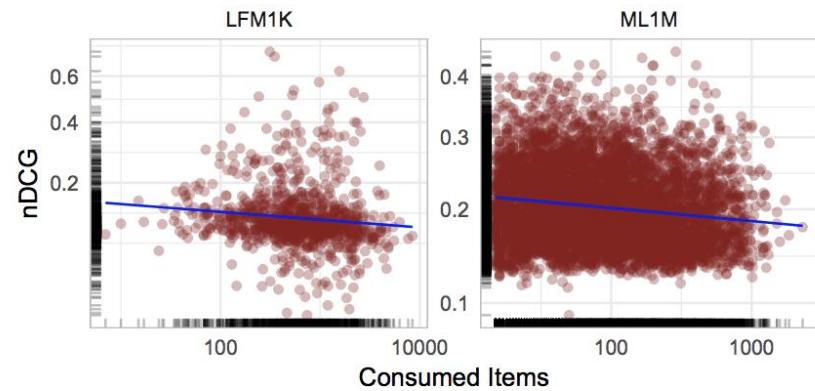
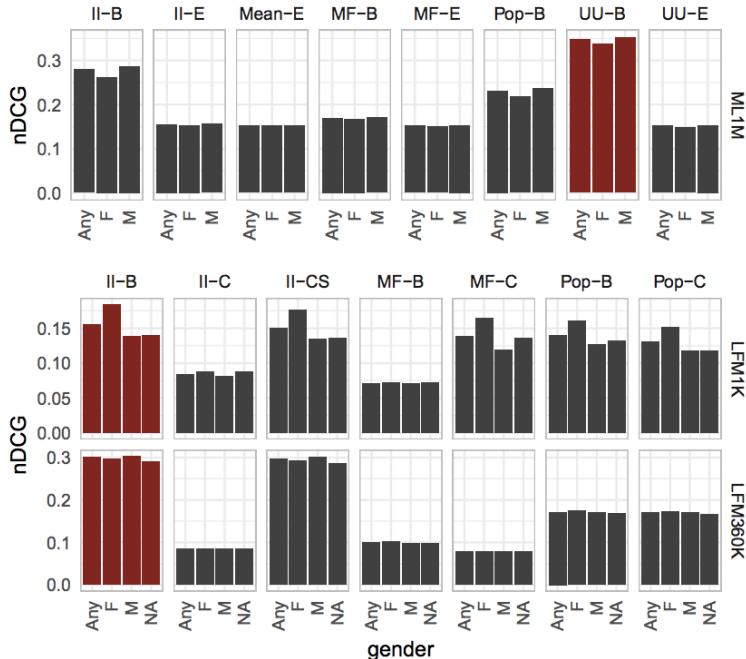
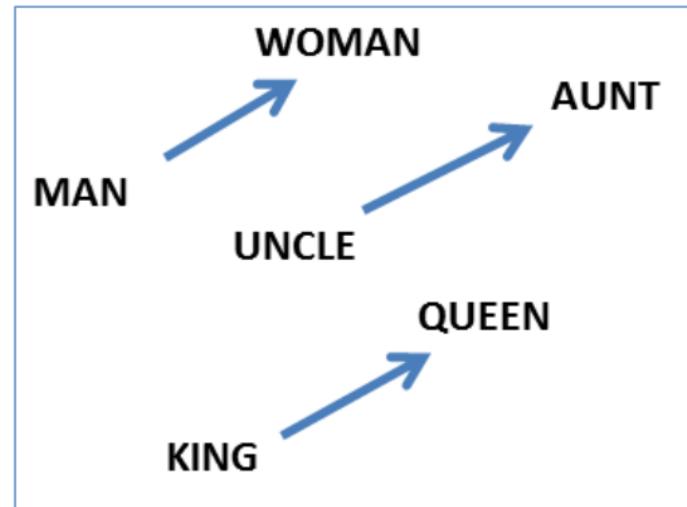
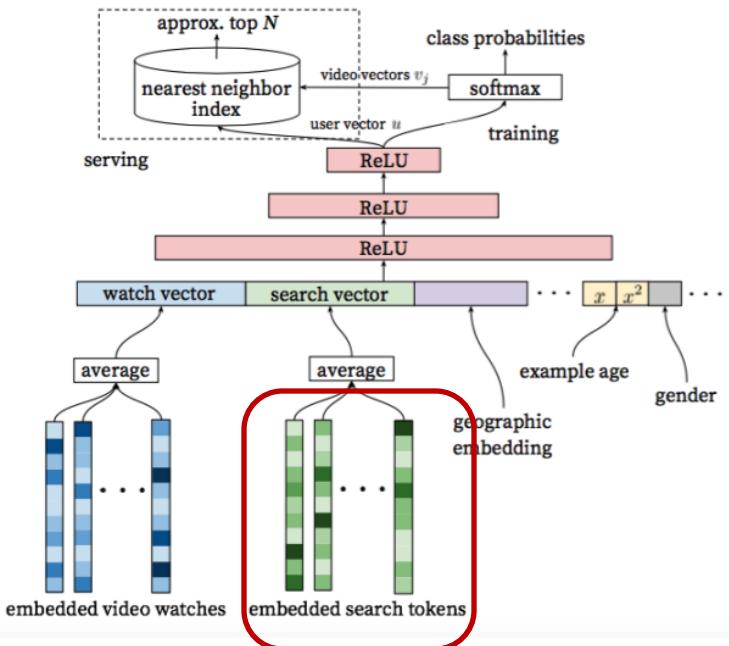


Figure 5: Models predicting nDCG with profile size.

Language models (word2vec, Glove, etc.)



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

Bias in Language Models

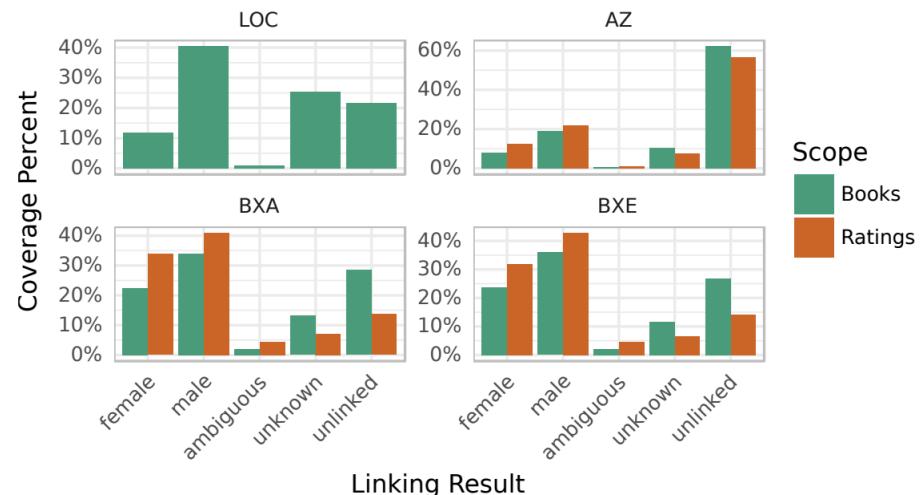
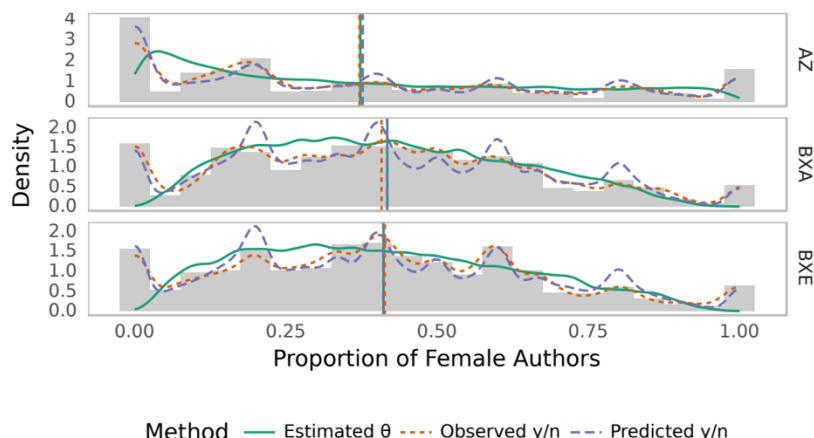
- In addition to their ability to learn word meaning from text, embeddings, alas, also **reproduce the implicit biases and stereotypes** that were latent in the text.
- Bolukbasi et al. (2016) found that the closest occupation to ‘man’ - ‘computer programmer’ + ‘woman’ in word2vec embeddings trained on news text is ‘homemaker’
- <https://web.stanford.edu/~jurafsky/slp3/6.pdf>

Debiasing in Language Models

- Recent research focuses on ways to try to remove these kinds of biases
- By developing a transformation of the embedding space that removes gender stereotypes but preserves definitional gender (Bolukbasi et al. 2016, Zhao et al. 2017) or changing the training procedure (Zhao et al., 2018).
- However, although these sorts of debiasing may reduce bias in embeddings, they do not eliminate it (Gonen and Goldberg, 2019), and this remains an open problem

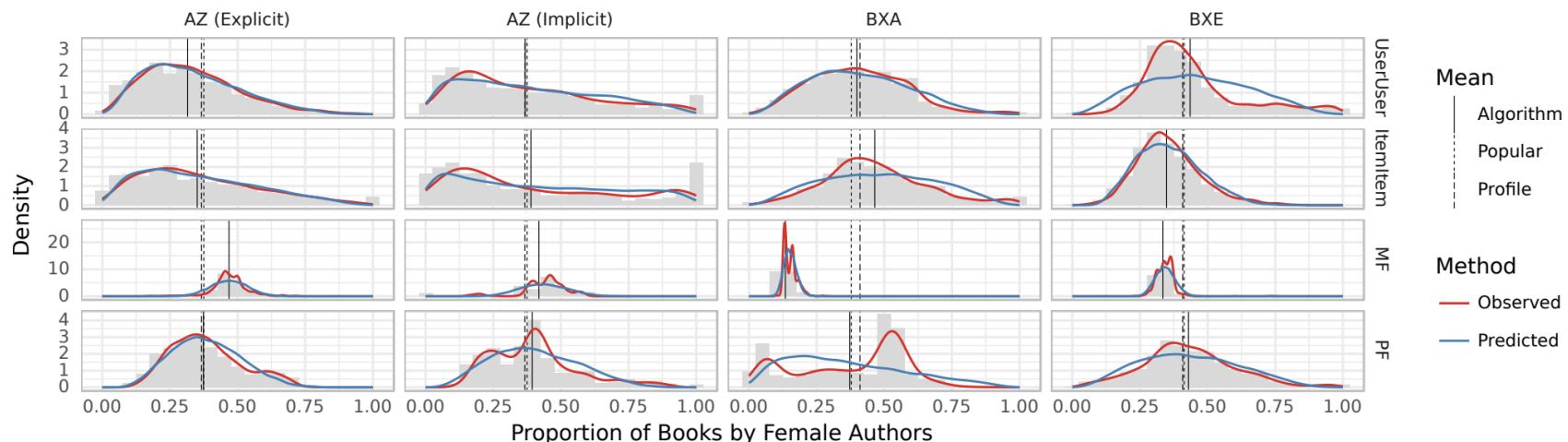
Producer Bias

- Ekstrand, M. D., Tian, M., Kazi, M. R. I., Mehrpouyan, H., & Kluver, D. (2018). Exploring author gender in book rating and recommendation. ACM RecSys 2018.



Producer Bias

- Ekstrand, M. D., Tian, M., Kazi, M. R. I., Mehrpouyan, H., & Kluver, D. (2018). Exploring author gender in book rating and recommendation. ACM RecSys 2018.



Filter Bubble

- The term filter bubble was popularized by Eli Pariser in his book “The filter bubble: What the Internet is hiding from you”.
- It refers to echo chambers and feedback loops: people gets stacked into a bubble without much option to escape and consume more diverse content.
- Chaney, A. J., Stewart, B. M., & Engelhardt, B. E. (2018). How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. ACM RecSys.

Fairness methods in 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.

Recent articles (RecSys 2020)

- Tobias Schnabel and Paul N. Bennett. 2020. *Debiasing Item-to-Item Recommendations With Small Annotated Datasets*.
- Mesut Kaya, Derek Bridge, and Nava Tintarev. 2020. *Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance*.
- Jin Huang, Harrie Oosterhuis, Maarten de Rijke, and Herke van Hoof. 2020. *Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning based Recommender Systems*.
- FAccTRec workshop:

Open Challenges for Fairness in RecSys

- DATA: Most datasets do not have information to investigate these issues, identifying biases is an open area of research.
- There is no one-size-fits-all solution for fairness: What accurately represents the world? What accurately represents the world as it could or should be?
- We should consider both consumer and producer forms of bias: recommending most popular might be easy to implement and effective, but we are not promoting new producers.

Summary

- In this talk, I have presented, motivated, and defined several aspects of FAT, with a focus on the context of RecSys.
- I have also surveyed several works and areas of research related to FAT and XAI. There are many open research questions to address, and decisions to make in order to progress making RecSys useful but also fair.

dparra@ing.puc.cl

THANKS!

