

# Retrieval and Recommendation Systems at the Crossroads of Artificial Intelligence, Ethics, and Regulation

## Markus Schedl

Johannes Kepler University Linz, Austria  
Linz Institute of Technology, Austria  
[markus.schedl@jku.at](mailto:markus.schedl@jku.at) | [www.mschedl.eu](http://www.mschedl.eu) | @m\_schedl



## Emilia Gómez

Joint Research Centre, European Commission  
Universitat Pompeu Fabra, Spain  
[emilia.gomez-gutierrez@ec.europa.eu](mailto:emilia.gomez-gutierrez@ec.europa.eu) | [https://emiliagomez.com](http://emiliagomez.com) | @emiliagogu



## Elisabeth Lex

Graz University of Technology, Austria  
[elisabeth.lex@tugraz.at](mailto:elisabeth.lex@tugraz.at) | [https://elisabethlex.info](http://elisabethlex.info)



# About Markus Schedl

- Full Professor at Johannes Kepler University (JKU) Linz, Austria
- Head of *Multimedia Mining and Search* (MMS) group at Institute of Computational Perception
- Head of *Human-centered Artificial Intelligence* (HCAI) group at Linz Institute of Technology (LIT), AI Lab
- Lab: <https://hcai.at> | <https://www.jku.at/en/institute-of-computational-perception>
- Interests: recommender systems, user modeling, information retrieval, machine learning, natural language processing, multimedia, data analysis, and web mining

Contact: [markus.schedl@jku.at](mailto:markus.schedl@jku.at) | [www.mschedl.eu](http://www.mschedl.eu) | @m\_schedl

# About Emilia Gómez



HUMAINT  
HUMAN behaviour and  
MACHINE INTEllIGENCE



- PI HUMAINT team, European Commission, Joint Research Centre, European Commission, Seville.  
Lab page: <https://ec.europa.eu/jrc/communities/en/community/humaint>
- Guest Professor, Music Information Research Lab, Universitat Pompeu Fabra, Barcelona. <https://www.upf.edu/web/mtg/>
- **Interests:** music IR, content-based description, recommendation, social impact of data-driven algorithms (bias/fairness, transparency, children, facial processing)..

Contact: [emilia.gomez@upf.edu](mailto:emilia.gomez@upf.edu) | [emiliagomez.com](http://emiliagomez.com) | [@emiliagogu](https://twitter.com/emiliagogu)

# About Elisabeth Lex

- Assoc. Prof at Graz University of Technology, Austria
- PI Recommender Systems & Social Computing Lab at Institute of Interactive Systems and Data Science (ISDS)
- Lab page: <https://socialcomplab.github.io>
- Interests: user modeling, recommender systems, information retrieval, computational social science

Contact: [elisabeth.lex@tugraz.at](mailto:elisabeth.lex@tugraz.at) | <https://elisabethlex.info> | @elisab79

# **What about you?**

- Share your views on our tutorial #sigir22ethics
- Interact, ask/vote questions, on site & online

Join at  
**slido.com**  
**#sigir22ethics**



# Overview

- **Introduction**  
Background, motivation, objectives, relevance to community, recent political and legal regulations
- **Fairness and non-discrimination**  
Categories of bias and fairness, relation to non-discrimination, definition and measurement of bias and fairness, algorithms to mitigate biases and improve fairness
- **Diversity**  
Categories of diversity, diversity in the research community, diversity by design
- **Transparency**  
Categories of transparency, explainability and justification, traceability and auditability, documentation
- **Open Challenges**

**Tutorial Slides:**

<https://socialcomplab.github.io/Retrieval-RecSys-AI-Ethics-Regulation-Tutorial-SIGIR22>

# **Introduction**

# Information Retrieval (IR) and Recommender Systems (RS) are Ubiquitous

amazon

eBay

Etsy

purchases

last.fm

Spotify

music

XING 

LinkedIn

jobs



social networking / information



movies / tv

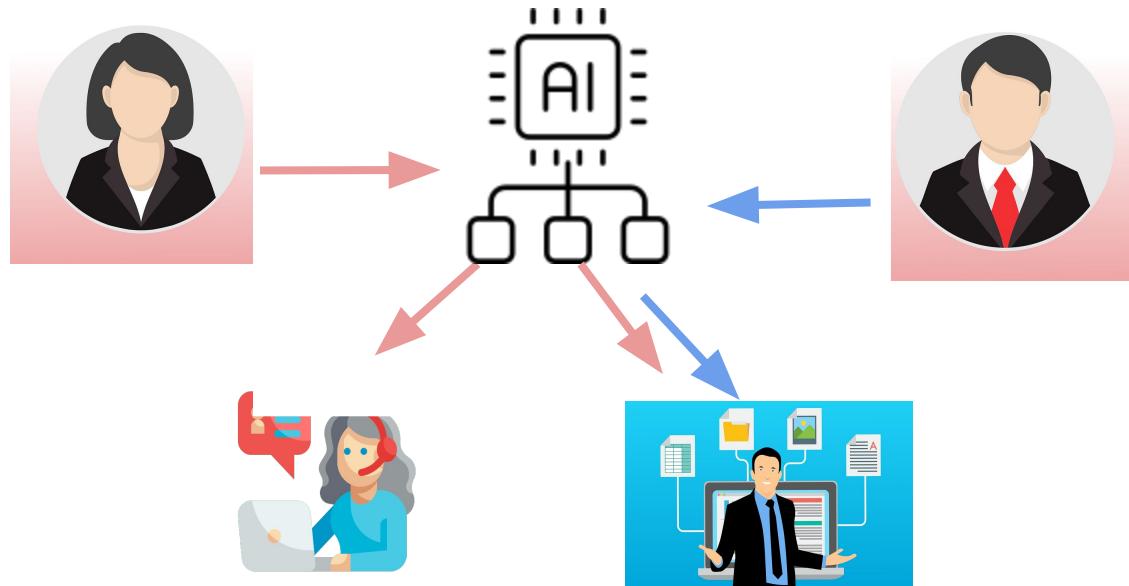


airbnb

travel

# Societal Impacts of IR & RS

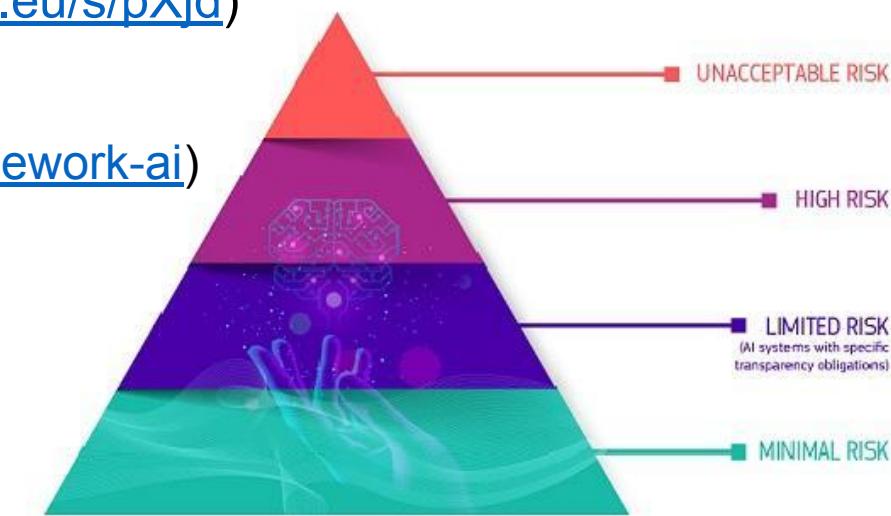
- From decision support / information seeking tools → socio-technical systems
- Create, control, limit exposure & access, shape opinion, influence behaviour:
  - e.g., jobs, products, information, opportunities



**Raises Ethical Questions**

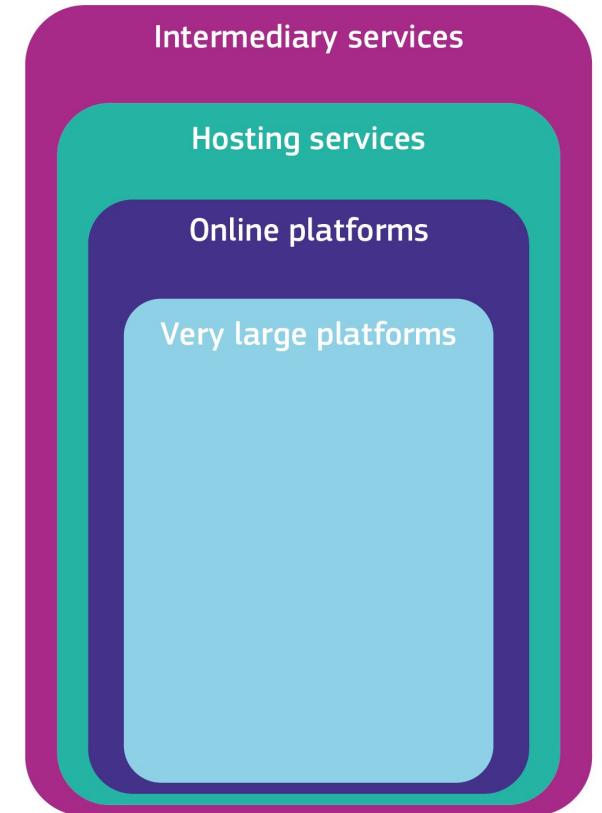
# Not Only a Technological or Algorithmic Problem

- **Multidisciplinary perspective:** law, ethics, sociology, economics, psychology, etc.
- EU Charter of Fundamental Rights  
([https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights\\_en](https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en))
- RS & SE as part of Artificial Intelligence:
  - EU Ethical Principles for Trustworthy AI (<https://op.europa.eu/s/pXjd>)
  - EU Regulatory Framework proposal on AI - 2021  
(<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>)
    - Prohibited: e.g. over-manipulation, social scoring.
    - High-risk: e.g. access to education, recruiting.



# EU Digital Services Act

- Some new obligations (Online platforms and Search Engines), e.g.:
  - Transparency of recommender systems
  - User-facing transparency of online advertising
  - Risk assessment and mitigation measures
  - External & independent auditing, internal compliance function and public accountability
  - Data sharing with authorities and researchers
  - Codes of conduct
  - Crisis response cooperation



# Chinese AI Governance Approaches

## THREE APPROACHES TO CHINESE AI GOVERNANCE

Organization	Focus of Approach	Relevant Documents
Cyberspace Administration of China	Rules for online algorithms, with a focus on public opinion	<ul style="list-style-type: none"><li>- Internet Information Service Algorithmic Recommendation Management Provisions</li><li>- Guiding Opinions on Strengthening Overall Governance of Internet Information Service Algorithms</li></ul>
China Academy of Information and Communications Technology	Tools for testing and certification of "trustworthy AI" systems	<ul style="list-style-type: none"><li>- Trustworthy AI white paper</li><li>- Trustworthy Facial Recognition Applications and Protections Plan</li></ul>
Ministry of Science and Technology	Establishing AI ethics principles and creating tech ethics review boards within companies and research institutions	<ul style="list-style-type: none"><li>- Guiding Opinions on Strengthening Ethical Governance of Science and Technology</li><li>- Ethical Norms for New Generation Artificial Intelligence</li></ul>

# US Initiatives

- The Artificial Intelligence Initiative Act (116th Congress 2019-2020, S.1558):  
<https://www.congress.gov/bill/116th-congress/senate-bill/1558/text>
- White House's Office of Science and Technology Policy released a draft *Guidance for Regulation of Artificial Intelligence Applications*:  
<https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf>
- Regulations in different states, e.g. California on Automated Decision Systems for Employment and Housing.  
<https://www.dfeh.ca.gov/wp-content/uploads/sites/32/2022/03/AttachB-ModtoEmployReqAutomated-DecisionSystems.pdf>

# **Ethical Issues of IR and RS**

- Fairness and non-discrimination (Part 1)
- Diversity (Part 2)
- Transparency (Part 3)

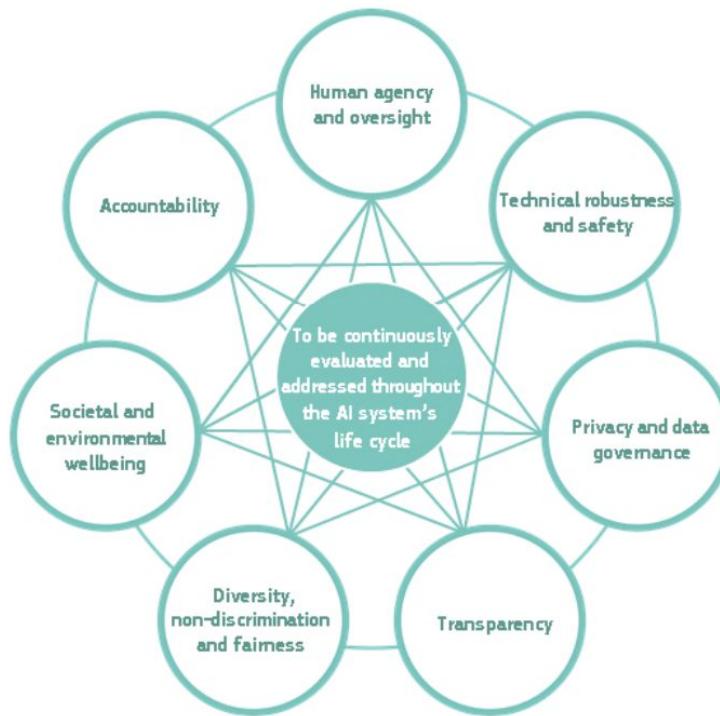
**Part 1:**

**Bias, Fairness, and Non-discrimination**

# **Outline**

- EU Regulation
- Bias from various perspectives
- Relation to fairness and non-discrimination
- Measuring biases
- Strategies to mitigate bias and improve fairness

# Non-discrimination and Fairness are Key Requirements for Trustworthy AI



High-level Expert Group on AI, European Commission, Ethics Guidelines for Trustworthy AI,  
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

# EU Regulations

- EU Regulatory Framework for AI  
(<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>)
- EU Charter of Fundamental Rights  
([https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights\\_en](https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_en))



## **Article 21: Non-discrimination**

1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.
2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited.

## **Article 23: Equality between women and men**

1. Equality between women and men must be ensured in all areas, including employment, work and pay.
2. The principle of equality shall not prevent the maintenance or adoption of measures providing for specific advantages in favour of the under-represented sex.

# Biases from a High-level Perspective



- **Societal Bias:** Discrepancy between how the world should be and how it actually is (e.g., equal representation of genders in jobs/positions vs. actual over/underrepresentation of genders)



# Biases from a High-level Perspective

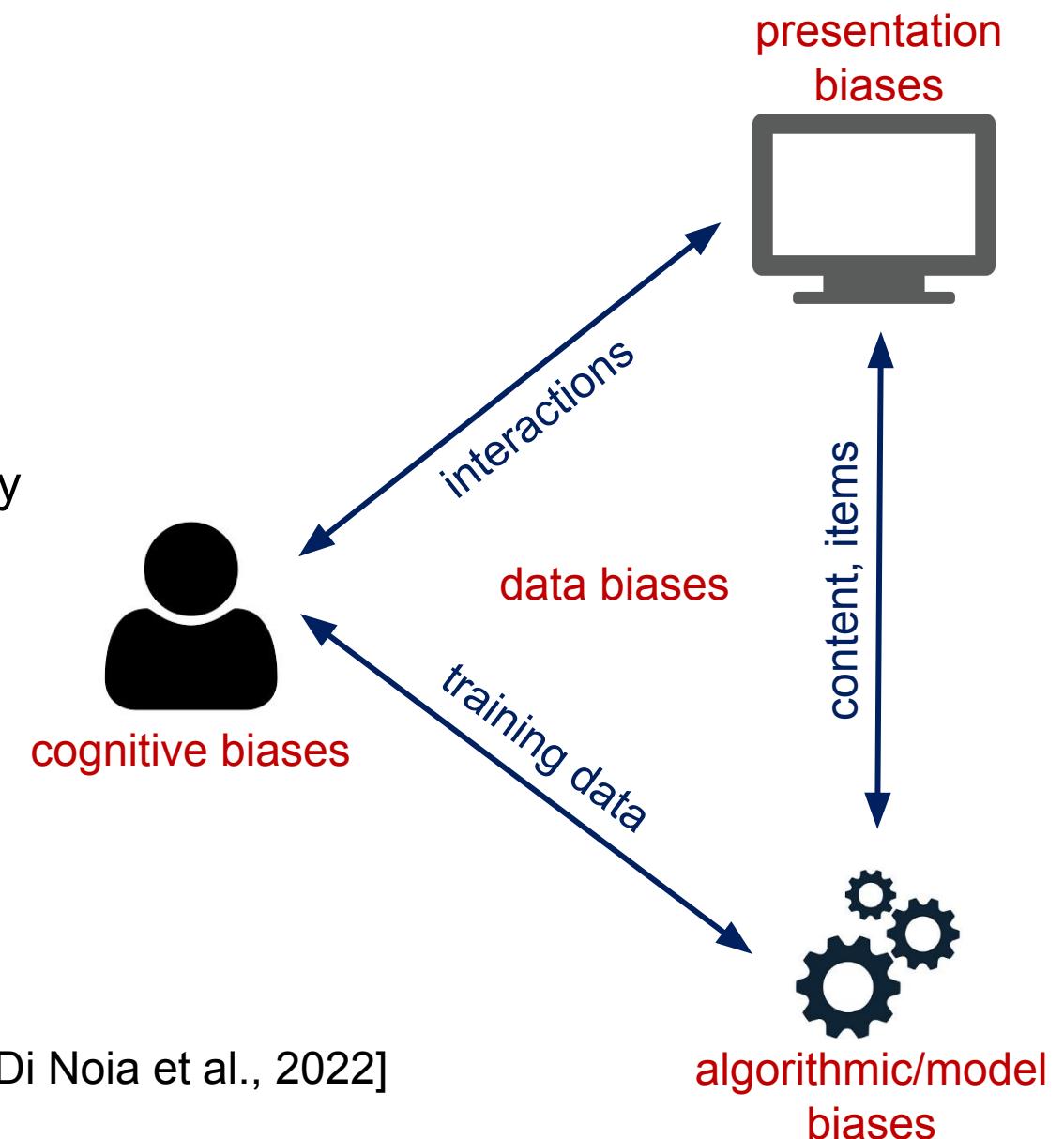


- **Societal Bias:** Discrepancy between how the world should be and how it actually is (e.g., equal representation of genders in jobs/positions vs. actual over/underrepresentation of genders)
- **Statistical Bias:** Discrepancy between how the world is and how it is encoded in the system or created machine learning model (e.g., data does not reflect population at large; in RSSs often a community bias)

# Biases in Retrieval and Recommender Systems

Decisions made by IR and RSs are affected by various biases (influencing each other), originating from:

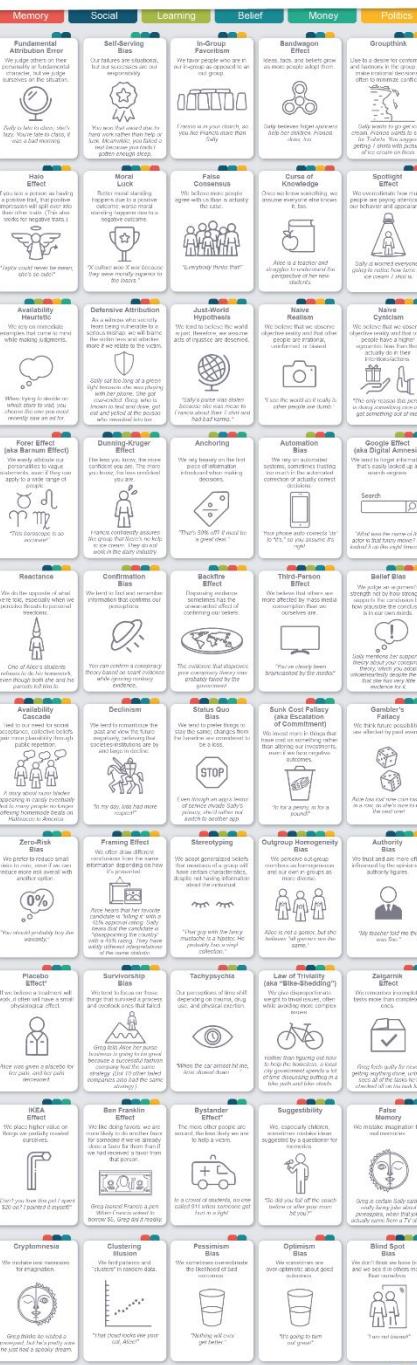
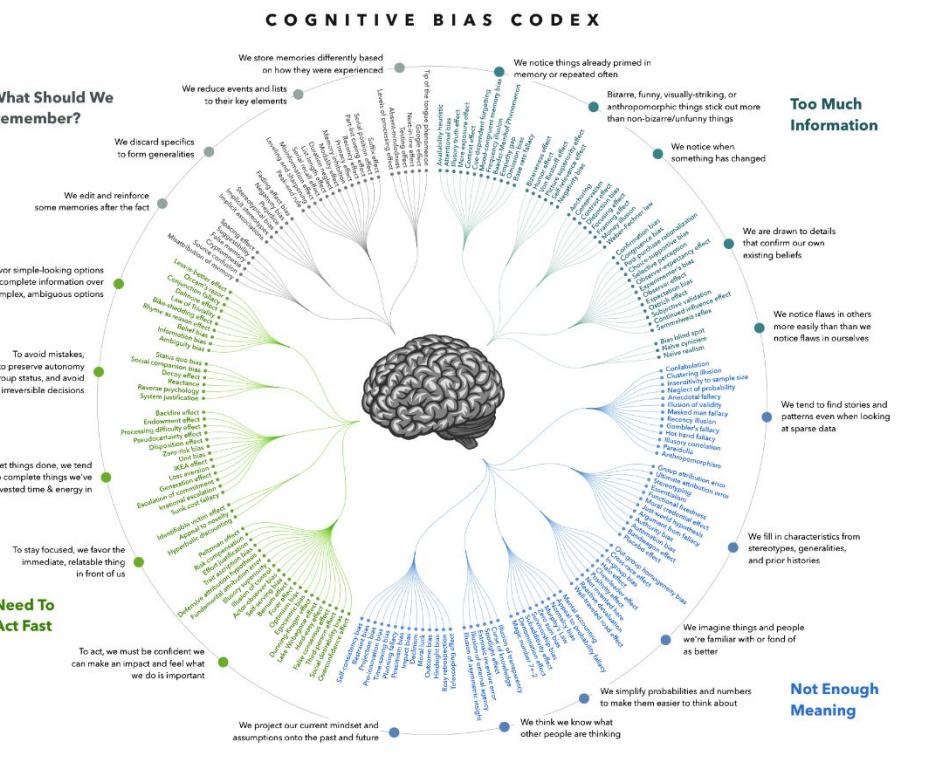
- *Data*: e.g., unbalanced dataset w.r.t. group of users → demographic bias, community bias
- *Algorithms*: e.g., reinforcing stereotypes or amplify already popular content (“rich get richer” effect) → popularity bias
- *Presentation*: e.g., positions of recommended items on screen
- *User cognition or perception*: e.g., serial position effect, confirmation bias



# Biases in IR and RSs

Additional cognitive biases:

- [https://en.wikipedia.org/wiki/List\\_of\\_cognitive\\_biases](https://en.wikipedia.org/wiki/List_of_cognitive_biases)
- [https://commons.wikimedia.org/wiki/File:Cognitive\\_bias\\_codex\\_en.svg](https://commons.wikimedia.org/wiki/File:Cognitive_bias_codex_en.svg)
- <https://www.visualcapitalist.com/50-cognitive-biases-in-the-modern-world>



# When are Biases Problematic?

Biases can result in different treatment of users or groups of users

*“The system systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others.”* [Friedman and Nissenbaum, 1996]

However, not all biases are bad

- Trade-off between personalization and fairness, i.e., the RS has to favor items that the user is likely to consume
- Case study: Popularity bias
  - Should a system recommend *all* content items with the same likelihood?
  - Should the popularity of items in the recommendation list match the popularity of items in the user’s consumption history (“calibration”)?
  - Should it match with the item popularity in the consumption history of all users of the RS?

Making things even more complicated: multiple stakeholders are involved  
(e.g., content producers, content consumers, platform providers, policymakers)

**Have you ever experienced popularity bias  
when using retrieval or recommender  
systems? Using which one(s)?**

sli.do

#sigir22ethics

Join at

**slido.com**

**#sigir22ethics**



# Fair for Whom?

- **Individual fairness:**  
Similar users are treated in a similar fashion (e.g., users with similar skills receive job recommendations within the same pay grade)
- **Group fairness:**  
Different groups of users defined by some sensitive or protected attribute (e.g., gender, age, or ethnicity) are treated in the same way. Accordingly, unfairness is defined as “systematically and unfairly discriminat[ing] against certain individuals or groups of individuals in favor of others.”

# **Bias Measurement**

# User Demographic Bias

[Melchiorre et al., 2021]

Metric:  $RecGap$  measures performance difference of the RS for different user groups

$$RecGap^\mu = \frac{\sum_{\langle g, g' \rangle \in G^{pair}} \left| \frac{\sum_{u \in U_g} \mu(u)}{|U_g|} - \frac{\sum_{u' \in U_{g'}} \mu(u')}{|U_{g'}|} \right|}{|G^{pair}|}$$

Average difference in performance metric  $\mu$  between all pairs of user groups  $G^{pair}$

$\mu$  precision, recall, NDCG, or beyond-accuracy metrics (e.g., coverage or diversity)

$U_g$  set of users in group  $g$ , e.g. defined by gender, ethnicity, age, country

→  $RecGap$  considers a RS to be fair if it performs *equally good* across the groups

# User Demographic Bias

[Melchiorre et al., 2021]

Metric: *RecGap* measures performance difference of system for different user groups



Model	Scenario	All	M/F	<i>RecGap</i>
POP	STANDARD	.046	.045/.049	.004 (f)
	RESAMPLED	.045	.044/.051	.007 (f) †
ItemKNN	STANDARD	.301	.313/.259	.054 (m) †
	RESAMPLED	.292	.304/.250	.054 (m) †
BPR	STANDARD	.127	.129/.117	.012 (m) †
	RESAMPLED	.123	.124/.116	.008 (m)
ALS	STANDARD	.241	.251/.205	.046 (m) †
	RESAMPLED	.238	.248/.204	.044 (m) †
SLIM	STANDARD	.364	.378/.315	.063 (m) †
	RESAMPLED	.359	.372/.312	.060 (m) †
MultiVAE	STANDARD	.192	.197/.173	.024 (m) †
	RESAMPLED	.183	.188/.166	.023 (m) †



- Majority of CF-based algorithm provide worse recommendations to female than to male users (w.r.t. NDCG and Recall)
- Mostly inverse relationship between accuracy (NDCG, Recall) and fairness

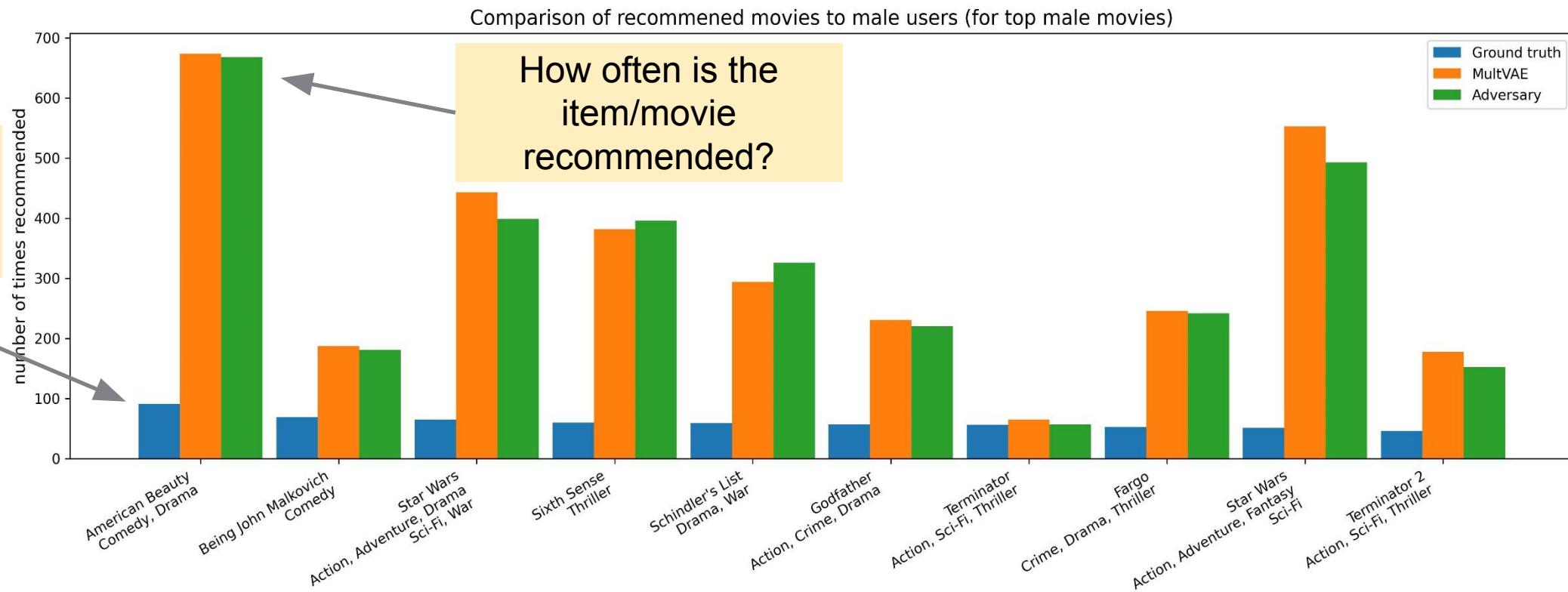
# Popularity Bias: Simple Example

[Lesota et al., 2021]

Metric: Difference between an item's recommendation frequency and consumption frequency in user profiles



How often is the item/movie consumed?



# Popularity Bias: More Formal / Delta Metrics

[Lesota et al., 2021]

Metrics: “*Delta*” metrics and *distribution-based metrics*

Assumption: Users prefer “calibrated” recommendations, i.e., the RS should mimic the input distribution w.r.t. an attribute (popularity in our case):  $\text{pop}(H_{u_i}(p_j)) \sim \text{pop}(R_{u_i}(p_j))$

$\text{pop}$  some measure of popularity

(e.g., number of interactions with item  $p_j$ , over all users, or number of users)

$H_{u_i}$  list of user  $u_i$ ’s interaction history (over items  $p_j$ )

$R_{u_i}$  recommendation list created for user  $u_i$  (top recommendations at fixed cut-off)

Delta metrics: *statistical moments* of popularity differences between items in  $H_{u_i}$  and  $R_{u_i}$

Distribution-based metrics: difference between popularity distributions (e.g., Kullback-Leibler divergence or Kendall’s  $\tau$ )

# Popularity Bias: Delta Metrics

[Lesota et al., 2021]

Metrics: “*Delta*” metrics

$\% \Delta \xi$  “percent Delta Xi” ~ relative popularity difference in terms of statistical measure  $\xi$

$$\% \Delta \xi(u_i) = \frac{\xi(R_{u_i}(p_j)) - \xi(H_{u_i}(p_j))}{\xi(H_{u_i}(p_j))} * 100$$

$\xi$  statistical measure or moment of interest (mean, median, variance, skew, etc.)

- Positive  $\% \Delta \text{Mean}$  and  $\% \Delta \text{Median}$  indicate that more popular tracks are recommended to user  $u_i$  than warranted given his or her consumption profile (“miscalibration”)
- Positive  $\% \Delta \text{Variance}$  indicate that recommendation list is more diverse w.r.t. covering differently popular items than user  $u_i$ ’s consumption profile

Aggregate over all users (a RS’s bias):  $\% \Delta \xi = \text{Median}(\% \Delta \xi(u_i))$

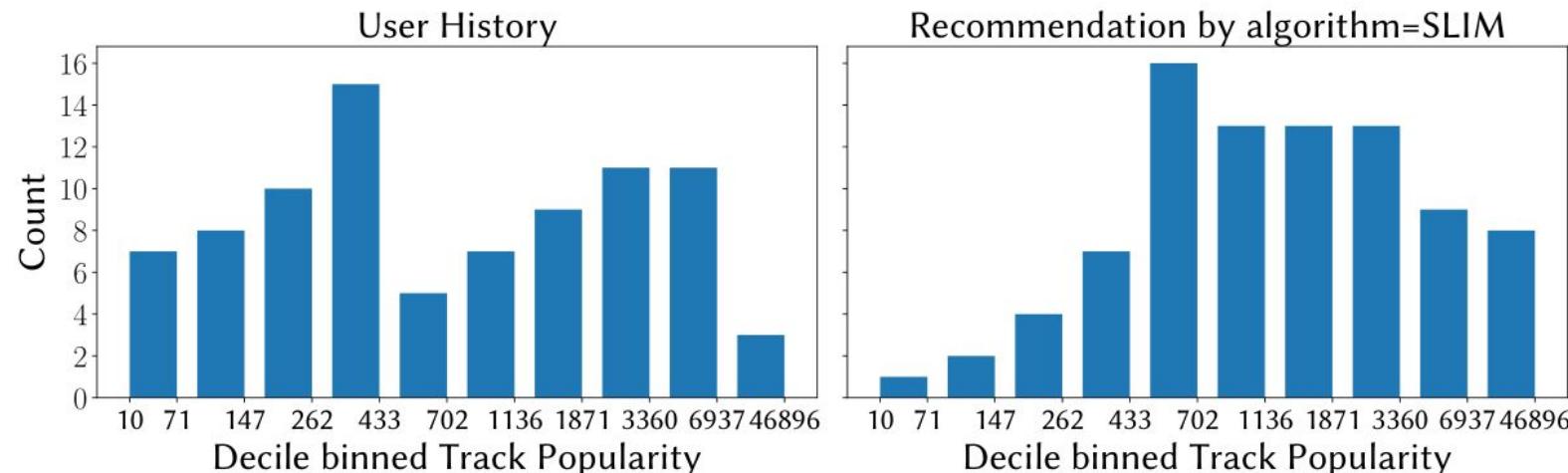
# Popularity Bias: Distribution-based Metrics

[Lesota et al., 2021]

Metrics: *Distribution-based metrics*

Considers the binned and normalized item popularities as (probability) distribution and computes:

- Kullback-Leibler (KL) divergence: ~dissimilarity between the two distributions
- Kendall's  $\tau$ : ~degree to which the order of bins is the same for the two distributions when ranked according to the respective counts



# Popularity Bias: Empirical Results

Alg.	Users	%ΔMean	%ΔMedian	%ΔVar.	%ΔSkew	%ΔKurtosis	KL	Kendall's $\tau$	NDCG@10
RAND	All	-91.8	-87.2	-99.5	11.5	15.3	3.904	0.165	0.000
	ΔFemale	-1.8	-3.5	-0.2	+0.0	-3.5	+0.976	-0.189	-0.000
	ΔMale	+0.5	+1.1	+0.1	-0.0	+1.3	-0.281	+0.053	+0.000
POP	All	432.5	975.2	487.0	-58.0	-87.0	6.023	0.057	0.045
	ΔFemale	+11.0	+282.1	-172.2	-2.1	-1.9	+1.626	-0.033	+0.003
	ΔMale	-2.8	-115.8	+55.9	+0.5	+0.5	-0.380	+0.016	-0.001
ALS	All	121.8	316.6	72.6	-25.2	-43.9	4.368	0.046	0.184
	ΔFemale	+9.9	+27.4	-7.1	-3.2	-5.4	+0.467	+0.110	-0.017
	ΔMale	-2.7	-6.6	+1.6	+0.8	+1.5	-0.121	-0.023	+0.005
BPR	All	-49.0	-3.7	-87.4	-14.8	-29.4	1.202	0.268	0.129
	ΔFemale	+5.2	+7.7	+2.1	-1.4	-3.9	+0.476	-0.043	-0.011
	ΔMale	-1.1	-1.9	-0.6	+0.4	+1.1	-0.110	+0.010	+0.003
ItemKNN	All	9.6	4.6	5.7	-14.3	-29.0	0.175	0.423	0.301
	ΔFemale	+2.0	+5.8	-2.6	-2.1	-3.2	+0.128	-0.037	-0.042
	ΔMale	-0.5	-1.3	+0.9	+0.8	+0.9	-0.020	+0.008	+0.012
SLIM	All	49.8	99.8	56.0	-12.5	-26.0	0.424	0.189	0.365
	ΔFemale	-6.4	-13.1	-17.4	-1.7	-4.6	+0.217	+0.052	-0.048
	ΔMale	+1.9	+3.9	+5.6	+0.6	+1.1	-0.029	-0.012	+0.014
VAE	All	303.9	736.3	351.0	-45.2	-70.1	4.823	-0.028	0.191
	ΔFemale	+10.1	+56.4	-69.3	-6.2	-6.6	+0.633	+0.146	-0.020
	ΔMale	-2.3	-20.4	+17.3	+1.8	+2.1	-0.161	-0.042	+0.006

- Most RS algorithms are prone to popularity bias (%ΔMean)
- ALS and VAE particularly
- ItemKNN least
- ALS and VAE increase also diversity (%ΔVar.)

# Popularity Bias: Empirical Results

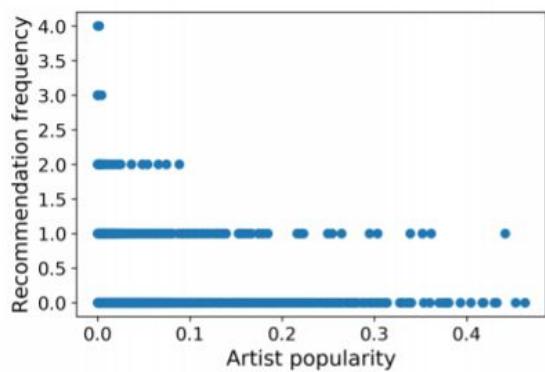
Popularity Bias can be combined with User Demographic Bias

Alg.	Users	%ΔMean	%ΔMedian	%ΔVar.	%ΔSkew	%ΔKurtosis	KL	Kendall's $\tau$	NDCG@10
RAND	All	-91.8	-87.2	-99.5	11.5	15.3	3.904	0.165	0.000
	ΔFemale	-1.8	-3.5	-0.2	+0.0	-3.5	+0.976	-0.189	-0.000
	ΔMale	+0.5	+1.1	+0.1	-0.0	+1.3	-0.281	+0.053	+0.000
POP	All	432.5	975.2	487.0	-58.0	-87.0	6.023	0.057	0.045
	ΔFemale	+11.0	+282.1	-172.2	-2.1	-1.9	+1.626	-0.033	+0.003
	ΔMale	-2.8	-115.8	+55.9	+0.5	+0.5	-0.380	+0.016	-0.001
ALS	All	121.8	316.6	72.6	-25.2	-43.9	4.368	0.046	0.184
	ΔFemale	+9.9	+27.4	-7.1	-3.2	-5.4	+0.467	+0.110	-0.017
	ΔMale	-2.7	-6.6	+1.6	+0.8	+1.5	-0.121	-0.023	+0.005
BPR	All	-49.0	-3.7	-87.4	-14.8	-29.4	1.202	0.268	0.129
	ΔFemale	+5.2	+7.7	+2.1	-1.4	-3.9	+0.476	-0.043	-0.011
	ΔMale	-1.1	-1.9	-0.6	+0.4	+1.1	-0.110	+0.010	+0.003
ItemKNN	All	9.6	4.6	5.7	-14.3	-29.0	0.175	0.423	0.301
	ΔFemale	+2.0	+5.8	-2.6	-2.1	-3.2	+0.128	-0.037	-0.042
	ΔMale	-0.5	-1.3	+0.9	+0.8	+0.9	-0.020	+0.008	+0.012
SLIM	All	49.8	99.8	56.0	-12.5	-26.0	0.424	0.189	0.365
	ΔFemale	-6.4	-13.1	-17.4	-1.7	-4.6	+0.217	+0.052	-0.048
	ΔMale	+1.9	+3.9	+5.6	+0.6	+1.1	-0.029	-0.012	+0.014
VAE	All	303.9	736.3	351.0	-45.2	-70.1	4.823	-0.028	0.191
	ΔFemale	+10.1	+56.4	-69.3	-6.2	-6.6	+0.633	+0.146	-0.020
	ΔMale	-2.3	-20.4	+17.3	+1.8	+2.1	-0.161	-0.042	+0.006

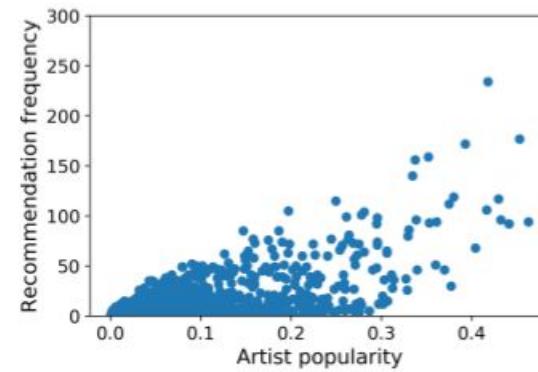
Most RS create an even higher popularity bias for female users than for male users (+/- values are relative to values in row All)

# Popularity Bias: Another Variant

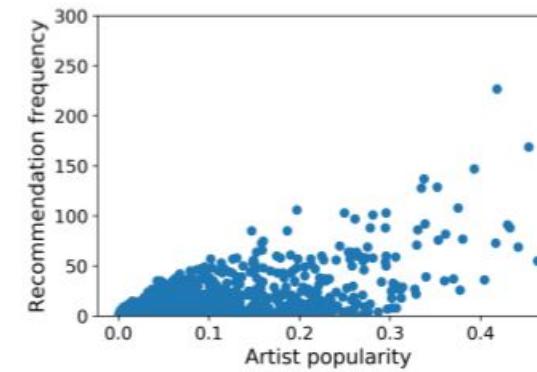
- RQ: Is the popularity level/*mainstreaminess* of users' listening preferences accurately reflected in recommendations made by algorithms?
- ~3K Last.fm users of different mainstreaminess (low, medium, high), selected from LFM-1b (dataset of 1B music listening records from Last.fm)
- Algorithms: User-based CF (KNN), NMF, UserItemAvg, Random, Most Popular
- Correlation between (artist) popularity and frequency of recommendation:



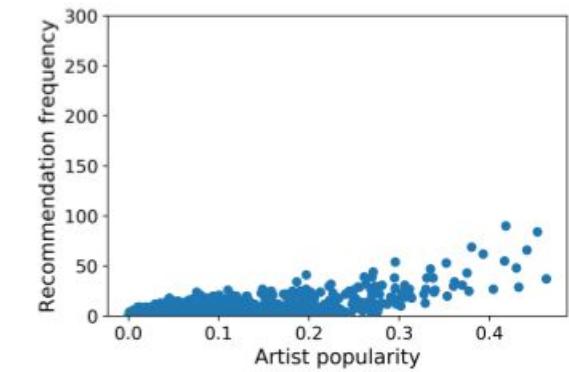
(a) Random.



(d) UserKNN.



(e) UserKNNAvg.



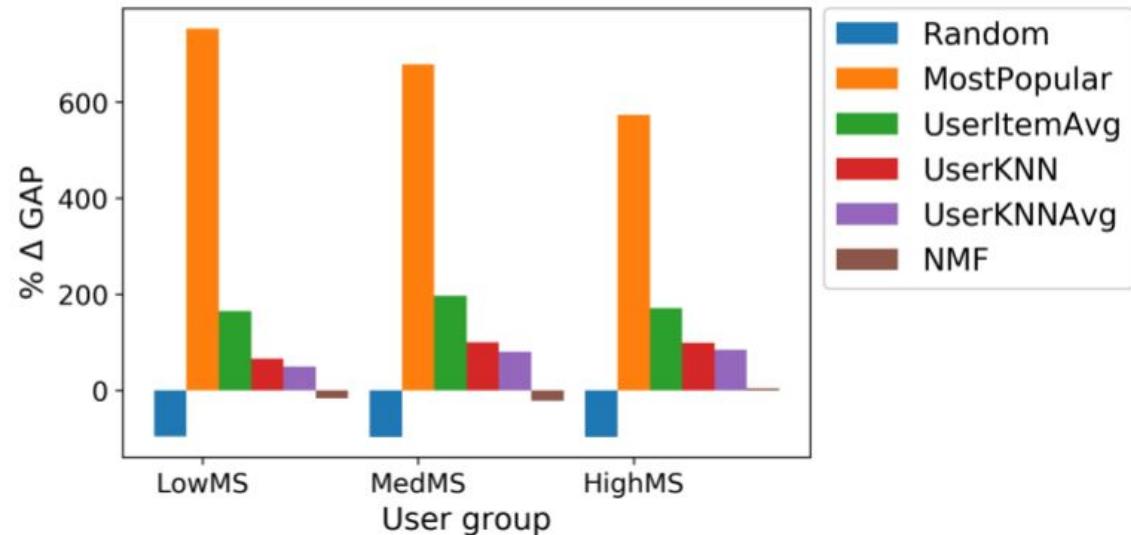
(f) NMF.

All RS algorithms favor popular artists (except for Random),  
irrespective of user preferences

[Kowald et al., 2020]

# Popularity Bias: Another Variant

- RQ: Is the popularity level/*mainstreaminess* of users' listening preferences accurately reflected in recommendations made by algorithms?
- Taking user preferences towards popular artists (mainstreaminess) into account:
- Metric: Difference in GAP (Group Average Popularity)
- $\Delta GAP = \frac{GAP(group, rec) - GAP(group, pref)}{GAP(group, pref)}$
- Measures extent to which popularity of recommendations exceed popularity of items in user profile



Most RS algorithms favor popular artists, irrespective of user preferences

# Personality Bias

- RQ: Do (music) recommender algorithms treat users with different personality traits equally?
- ~18K Twitter users (extracted music listening events; inferred personality traits from posts)
- Traits (high/low groups): openness, conscientiousness, extraversion, agreeableness, neuroticism
- Algorithms: SLIM, EASE (shallow AE), Mult-VAE

Neurotic people seem to have a narrow music taste



Group		Agr.	Con.	Ext.	Neu.	Ope.
High	No. unique tracks/user (mean and std.)	$19.1 \pm 24.4$	$19.2 \pm 25.5$	$20.0 \pm 26.3$	$16.2 \pm 18.4$	$19.5 \pm 24.9$
	No. unique tracks	15,694	15,674	15,655	15,429	15,652
	No. listening events	208,054	206,179	217,895	177,892	209,741
Low	No. unique tracks/user (mean and std.)	$17.3 \pm 21.7$	$17.2 \pm 20.4$	$16.4 \pm 19.2$	$20.3 \pm 26.9$	$16.9 \pm 21.1$
	No. unique tracks	15,664	15,695	15,672	15,607	15,619
	No. listening events	187,002	188,877	177,161	217,164	185,315

[Melchiorre et al., 2020]

# Personality Bias: Empirical Results

- RQ: Do music recommender algorithms treat users with different personality traits equally?
- Summary of results:
  - *Open* users receive worse recommendations (than narrow-minded users)
  - *Neurotics* receive better recommendations
  - *Extraverts* receive worse recommendations
  - *Conscientious* users get worse recommendations
  - Differences for *agreeableness* not very pronounced

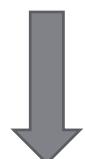
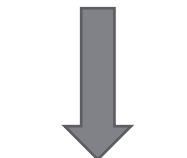
Performance of RS algorithms differs substantially between users of different personality

[Melchiorre et al., 2020]

Trait	Algorithm	@5		
		All	High	Low
Agr.	EASE	0.0311	0.0295	0.0327
	SLIM	0.0279	0.0263	0.0295
	Mult-VAE	0.0380	<b>0.0385*</b>	<b>0.0374*</b>
Con.	EASE	0.0311	<b>0.0274*</b>	<b>0.0349*</b>
	SLIM	0.0279	<b>0.0241***</b>	<b>0.0319***</b>
	Mult-VAE	0.0380	0.0353	0.0407
Ext.	EASE	0.0311	<b>0.0266**</b>	<b>0.0355**</b>
	SLIM	0.0279	<b>0.0242**</b>	<b>0.0317**</b>
	Mult-VAE	0.0380	<b>0.0340**</b>	<b>0.0417**</b>
Neu.	EASE	0.0311	<b>0.0366***</b>	<b>0.0257***</b>
	SLIM	0.0279	<b>0.0335***</b>	<b>0.0224***</b>
	Mult-VAE	0.0380	<b>0.0436***</b>	<b>0.0324***</b>
Ope.	EASE	0.0311	<b>0.0221***</b>	<b>0.0400***</b>
	SLIM	0.0279	<b>0.0196***</b>	<b>0.0363***</b>
	Mult-VAE	0.0380	<b>0.0285***</b>	<b>0.0473***</b>

# **Bias Mitigation**

# Strategies to Mitigating Harmful Biases



## Pre-processing strategies

- Data rebalancing (e.g., upsample minority group, subsample majority group)

## In-processing strategies

- Regularization (e.g., include bias correction term/bias metric in loss function used to train a model)
- Adversarial learning (e.g., train a classifier that predicts the sensitive attribute and adapt model parameters to minimize performance of this classifier)

## Post-processing strategies

- Reweigh/Rerank items in recommendation list
- Filter items (e.g., remove items from overrepresented groups)

# Mitigating Harmful Biases (Pre-processing Strategy)

## Ex.: Data Rebalancing

[Melchiorre et al., 2021]

Upsample data points by female user (to same amount created by male users)



last.fm

Model	Scenario	All	M/F	<i>RecGap</i>
POP	STANDARD	.046	.045/.049	.004 (f)
	RESAMPLED	.045	.044/.051	.007 (f) †
ItemKNN	STANDARD	.301	.313/.259	.054 (m) †
	RESAMPLED	.292	.304/.250	.054 (m) †
BPR	STANDARD	.127	.129/.117	.012 (m) †
	RESAMPLED	.123	.124/.116	.008 (m)
ALS	STANDARD	.241	.251/.205	.046 (m) †
	RESAMPLED	.238	.248/.204	.044 (m) †
SLIM	STANDARD	.364	.378/.315	<b>.063</b> (m) †
	RESAMPLED	.359	.372/.312	<b>.060</b> (m) †
MultiVAE	STANDARD	.192	.197/.173	.024 (m) †
	RESAMPLED	.183	.188/.166	.023 (m) †



NDCG gap between male  
and female users narrows

# Mitigating Harmful Biases (In-processing Strategy)

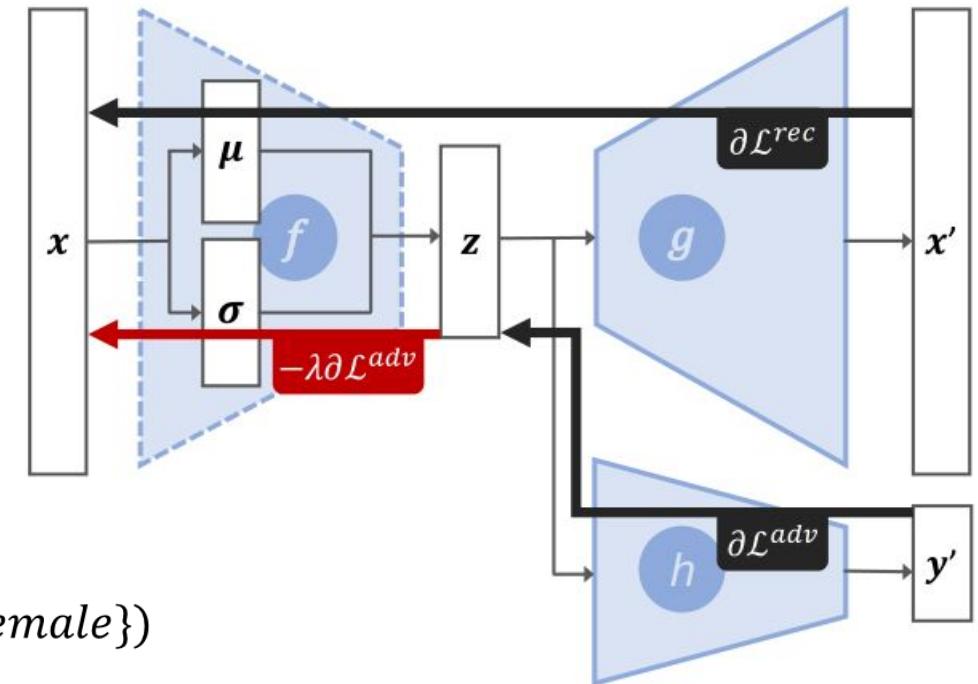
[Ganhör et al., 2022]

## Ex.: Adversarial Learning

Unlearn implicit information of protected attributes while preserving accuracy

Adversarial Mult-VAE architecture:

- $f(\cdot)$  encoder network
- $g(\cdot)$  decoder network
- $h(\cdot)$  adversarial network
- $x$  multi-hot encoded vector of item interactions
- $x'$  reconstruction of  $x$
- $z$  latent representation
- $y'$  prediction of protected attribute (e.g., gender  $\in \{male, female\}$ )



$$\arg \min_{f,g} \arg \max_h \mathcal{L}^{rec}(x) - \mathcal{L}^{adv}(x, y)$$

# Mitigating Harmful Biases (In-processing Strategy)

## Ex.: Adversarial Learning

[Ganhör et al., 2022]

Unlearn implicit information of protected attributes while preserving accuracy



Dataset	Model	Bias↓		Performance↑	
		Acc	BAcc	NDCG	Recall
ML-1M	MULTVAE <sub>BEST</sub>	0.692	0.707	<b>0.621</b>	<b>0.596</b>
	MULTVAE <sub>LAST</sub>	0.699	0.693	0.591†	0.566†
	ADV-MULTVAE	<b>0.565</b>	<b>0.572</b>	0.593†	0.569†
LFM2B-DB	MULTVAE <sub>BEST</sub>	0.703	0.717	<b>0.211</b>	<b>0.192</b>
	MULTVAE <sub>LAST</sub>	0.709	0.717	0.206†	0.189†
	ADV-MULTVAE	<b>0.631</b>	<b>0.609</b>	0.206†	0.189†



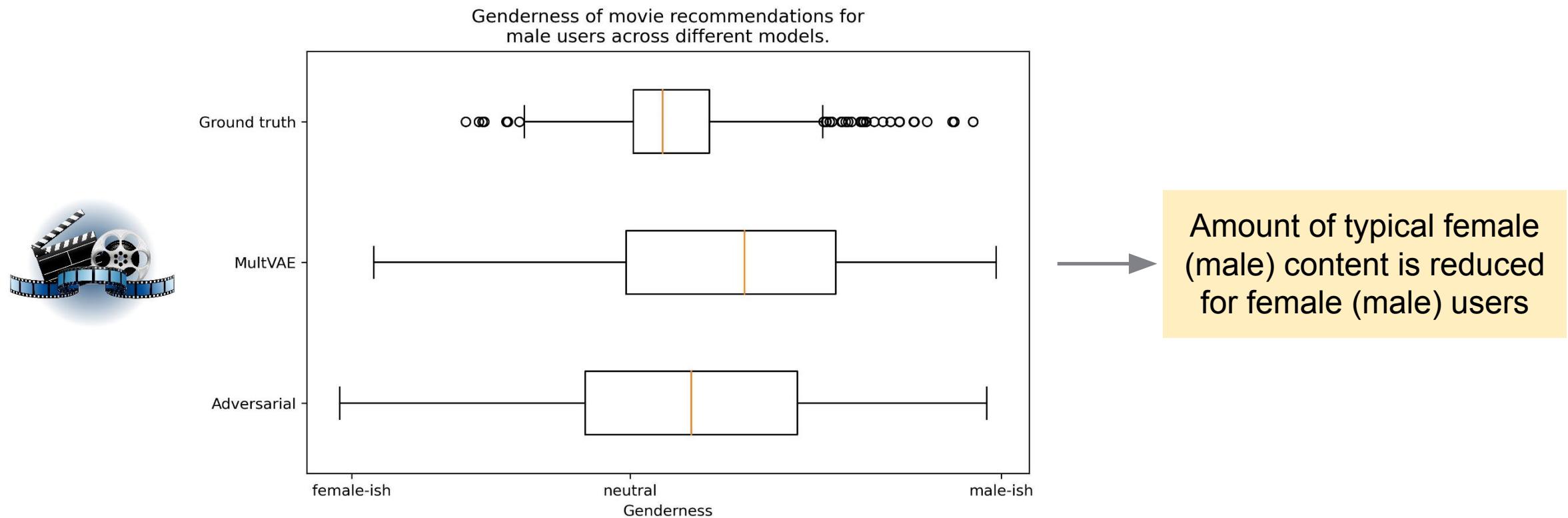
Substantial reduction of encoded protected information at expense of a marginal performance decrease

# Mitigating Harmful Biases (In-processing Strategy)

Ex.: Adversarial Learning

[Ganhör et al., 2022]

Unlearn implicit information of protected attributes while preserving accuracy



# Mitigating Harmful Biases (Post-processing Strategy)

Ex.: Reranking

[Ferraro et al., 2021]

Penalize/downrank content by the majority group (male artists) by  $\lambda$  positions in the recommendation list, created with ALS CF approach



	Algo	Avg position		% females rec.
		1st female	1st male	
LFM-1b	ALS	6.7717	0.6142	25.44
	POP	0.1325	1.7299	32.44
	RND	3.3015	0.3046	23.30
LFM-360k	ALS	8.3165	0.7136	26.27
	POP	0.9191	0.2713	29.31
	RND	3.3973	0.2951	22.77

cf. female artists in dataset: 23.25%

cf. female artists in dataset: 22.67%

**Do you think such a system is fair?  
Discriminates against women? Against men?**

sli.do

#sigir22ethics

Join at  
[slido.com](https://slido.com)  
**#sigir22ethics**



# Mitigating Harmful Biases (Post-processing Strategy)

Ex.: Reranking

[Ferraro et al., 2021]

Penalize/downrank content by the majority group (male artists) by  $\lambda$  positions in the recommendation list, created with ALS CF approach



	Algo	Avg position		% females rec.
		1st female	1st male	
LFM-1b	ALS	6.7717	0.6142	25.44
	POP	0.1325	1.7299	32.44
	RND	3.3015	0.3046	23.30
LFM-360k	ALS	8.3165	0.7136	26.27
	POP	0.9191	0.2713	29.31
	RND	3.3973	0.2951	22.77

cf. female artists in dataset: 23.25%

cf. female artists in dataset: 22.67%



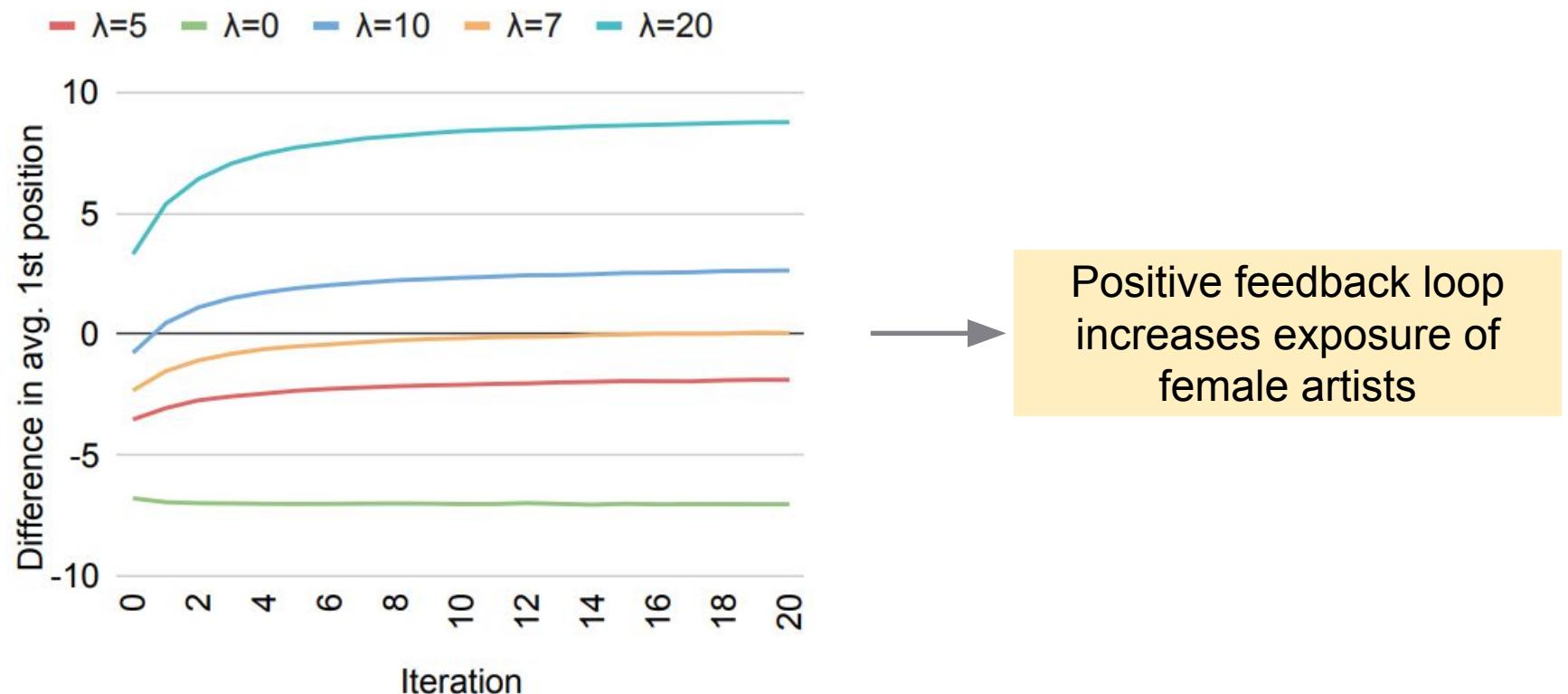
Female artists tend to occur further down in the recommendation lists

# Mitigating Harmful Biases (Post-processing Strategy)

## Ex.: Reranking

[Ferraro et al., 2021]

- Penalize/downrank content by the majority group (male artists) by  $\lambda$  positions
- Simulation study: In each iteration it is assumed that the top-10 recommendations are interacted with by the user, and the RS (ALS) is retrained accordingly



# **Summary**

- Biases are everywhere, not only in computer systems
- All IR and RSs have to cope with a variety of biases
- Some of them are desired, because they enable personalized results
- Some of them cause unfair behavior (i.e., treat different users/stakeholders differently)
- Most researched biases include popularity bias and demographic biases
- Coping strategies include pre-, in-, and post-processing techniques

# References

- [Chen et al., 2020]: *Bias and Debias in Recommender System: A Survey and Future Directions*, CoRR abs/2010.03240, <https://arxiv.org/abs/2010.03240>, 2020.
- [Di Noia et al.m 2022]: *Recommender systems under European AI regulations*. Communications of the ACM 65(4): 69-73, 2022.
- [Ekstrand et al., 2021]: *Fairness and Discrimination in Information Access Systems*, CoRR abs/2105.05779, 2021.
- [Ferraro et al., 2021]: *Break the Loop: Gender Imbalance in Music Recommenders*, Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR), Canberra, Australia, 2021.
- [Friedman and Nissenbaum, 1996]: *Bias in Computer Systems*, ACM Transactions on Information Systems 14(3):330-347, 1996.
- [Ganhör et al., 2022]: *Mitigating Consumer Biases in Recommendations with Adversarial Training*, Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2022), Madrid, Spain, 2022.
- [Kowald et al., 2020]: *The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study*, Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020), Lisbon, Portugal, 2020.
- [Lesota et al., 2021]: *Analyzing Item Popularity Bias of Music Recommender Systems: Are Different Genders Equally Affected?*, Proceedings of the 15th ACM Conference on Recommender Systems (RecSys 2021), Amsterdam, the Netherlands, 2021.
- [Melchiorre et al., 2020]: *Personality Bias of Music Recommendation Algorithms*, Proceedings of the 14th ACM Conference on Recommender Systems (RecSys 2020), Virtual, 2020.
- [Melchiorre et al., 2021]: *Investigating Gender Fairness of Recommendation Algorithms in the Music Domain*, Information Processing & Management, 58(5), 2021.

# **Part 2: Diversity**

# **Outline**

- **Motivation**
- Demographic diversity
- Diversity by design

# Why diversity?

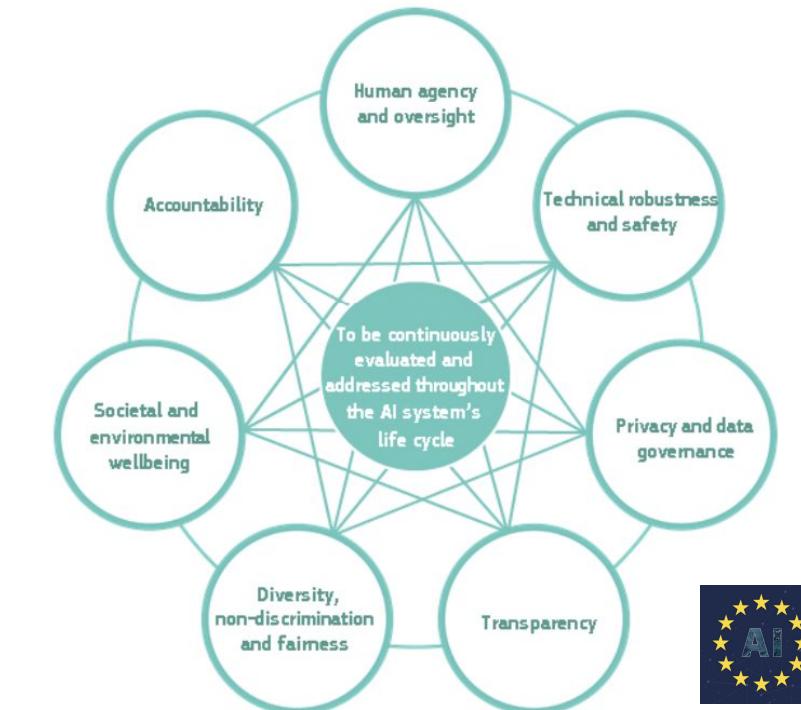
== Innovation, creativity; Lack of diversity == bias

Ensuring diversity and inclusion (UNESCO)

- Respect, protection and promotion of diversity.
- Consider personal choices, including the optional use of AI systems and its co-design.
- Overcome lack of necessary technological infrastructure, education and skills, as well as legal frameworks.

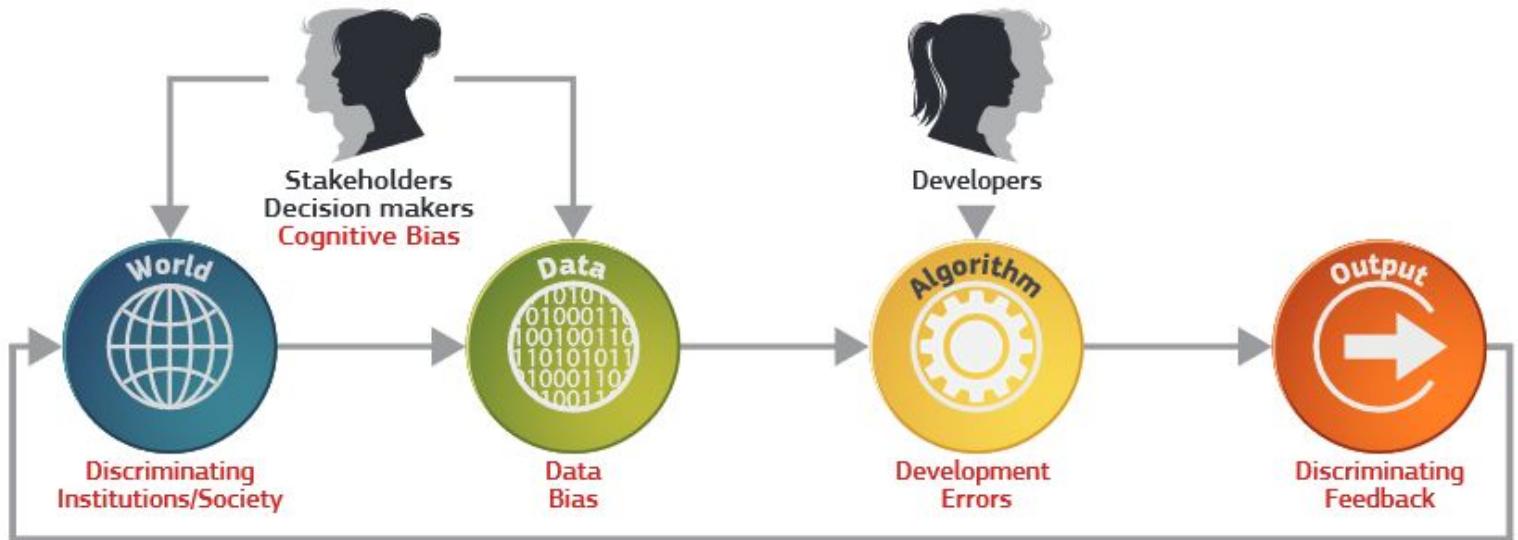
Diversity in one of the 7 requirements for trustworthy AI (EC-HLEG)

- Accessibility and universal design.
- Consideration and involvement of all affected stakeholders.



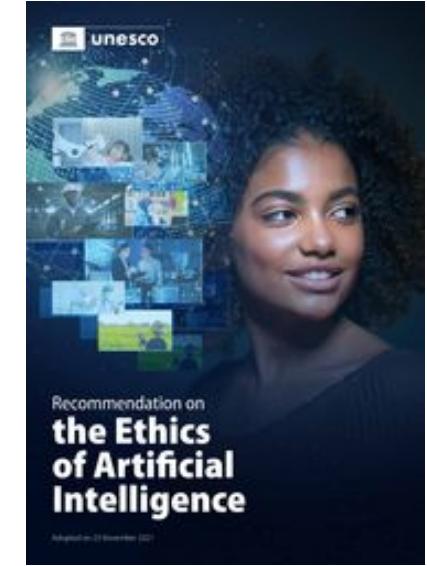
# Diversity as a transversal value in the design process

- Application
- Data sample
- Annotation
- Algorithmic model
- Evaluation strategy
- User interface



# Diversity dimensions

- Race
- Colour
- Descent
- Gender
- Age
- Language
- Religion
- Political opinion
- National origin
- Ethnic origin
- Social origin
- Economic or social condition of birth
- Disability
- + lifestyle choices, beliefs, opinions, expressions or personal experiences, including the optional use of AI systems and its co-design.
- Culture
- **Scientific research: methodologies, disciplines, topics**



- Recommendation on the Ethics of Artificial Intelligence, UNESCO <https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>
- Porcaro, L., Castillo, C. and Gómez, E., 2021. Diversity by Design in Music Recommender Systems. Transactions of the International Society for Music Information Retrieval, 4(1), pp.114–126. DOI: <http://doi.org/10.5334/tismir.106>

# Outline

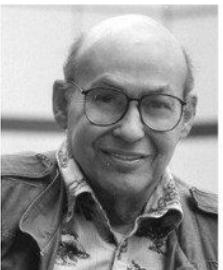
- Motivation
- **Demographic diversity**, or diversity in research communities
- Diversity by design

# Diversity in AI

## 1956 Dartmouth Conference: The Founding Fathers of AI



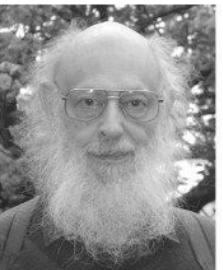
John McCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



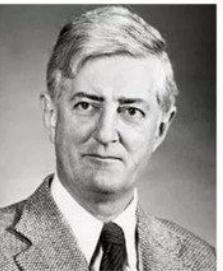
Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge



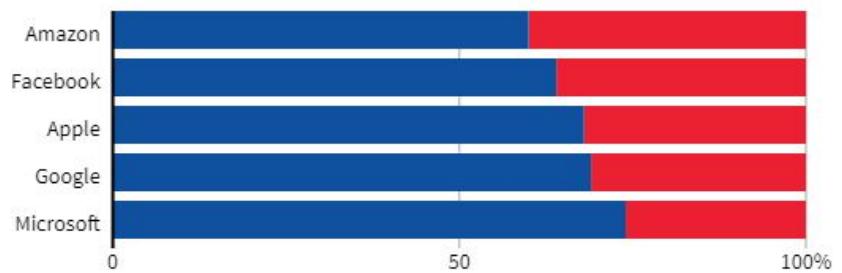
Nathaniel Rochester



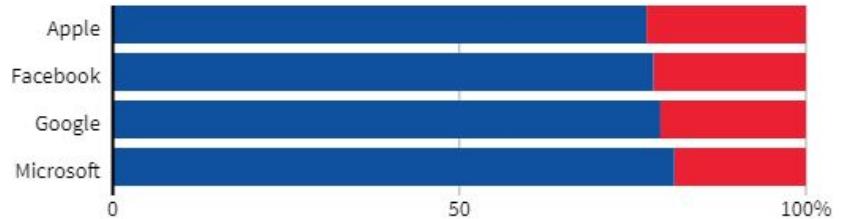
Trenchard More

### GLOBAL HEADCOUNT

■ Male ■ Female



### EMPLOYEES IN TECHNICAL ROLES



<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scaps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

4

**AI has a diversity challenge:** In 2019, 45% new U.S. resident AI PhD graduates were white—by comparison, 2.4% were African American and 3.2% were Hispanic.

Zhang, D., et al. The AI Index 2021 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, March 2021.

# How diverse is our group?

sli.do

#sigir22ethics

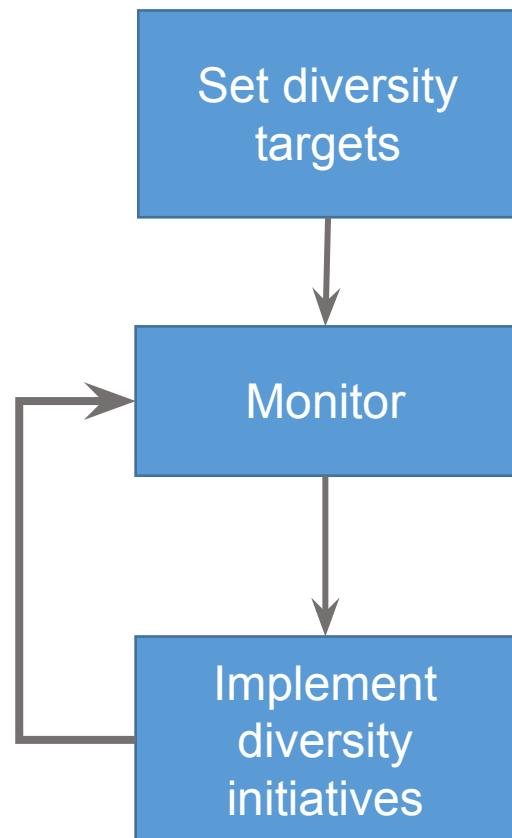
Join at

[slido.com](https://slido.com)

**#sigir22ethics**



# How to enhance diversity?

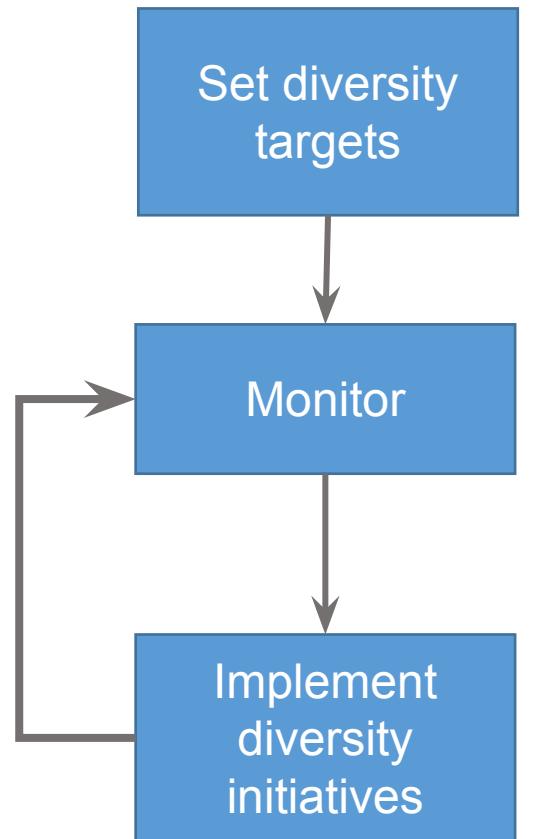


1

# Setting diversity targets

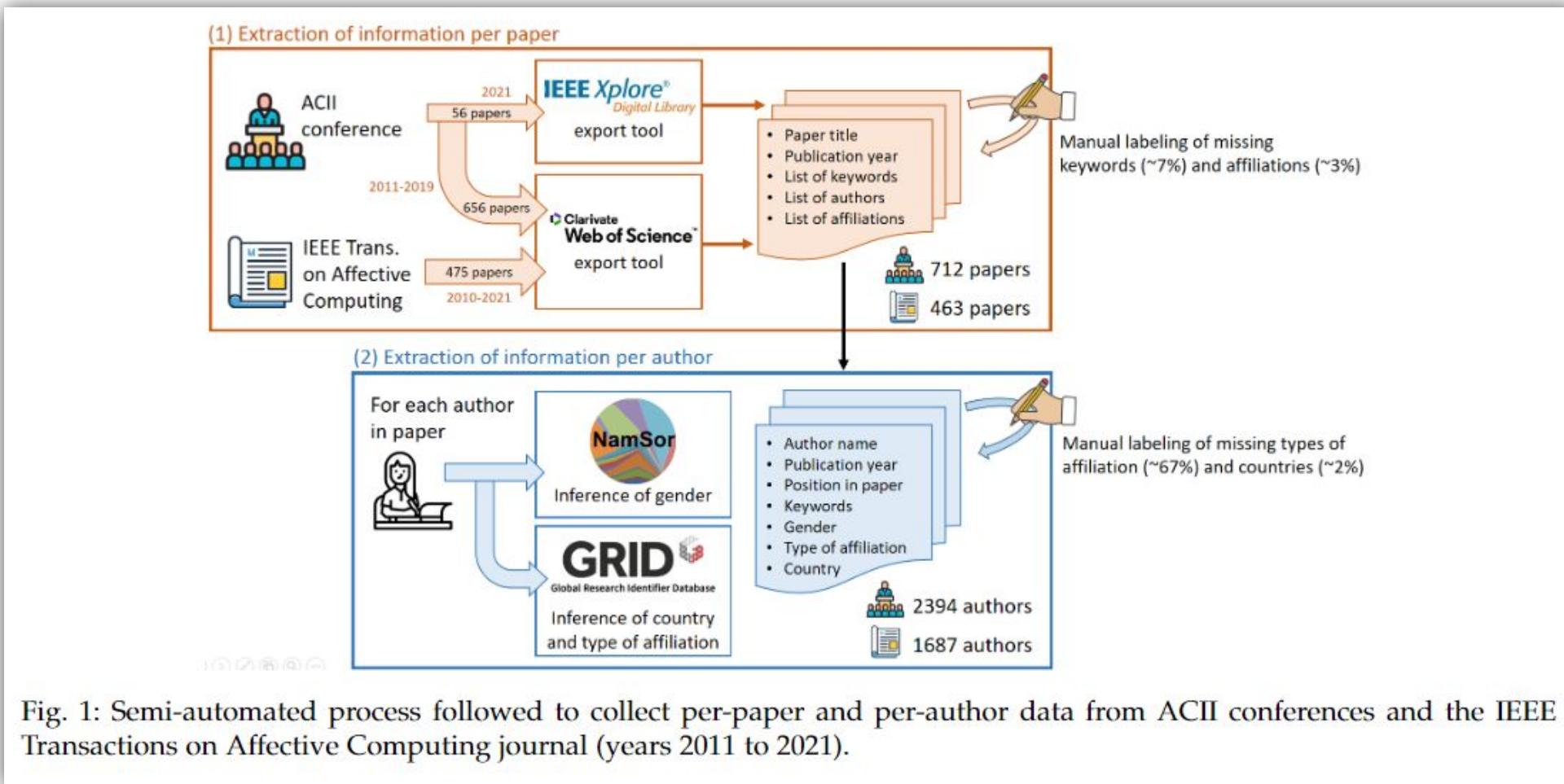
- Dimensions:
  - Gender, sexual orientation
  - Age, seniority
  - Racial, ethnicity / geographical origin or location
  - Institution type: academia, industry, government,...
  - Disabilities
  - Topics: disciplines, methodologies, aspects
- Targets: increase diversity, a collective decision?

# How to enhance diversity?



2

# Monitoring diversity: gather data



# Monitoring diversity: indicators

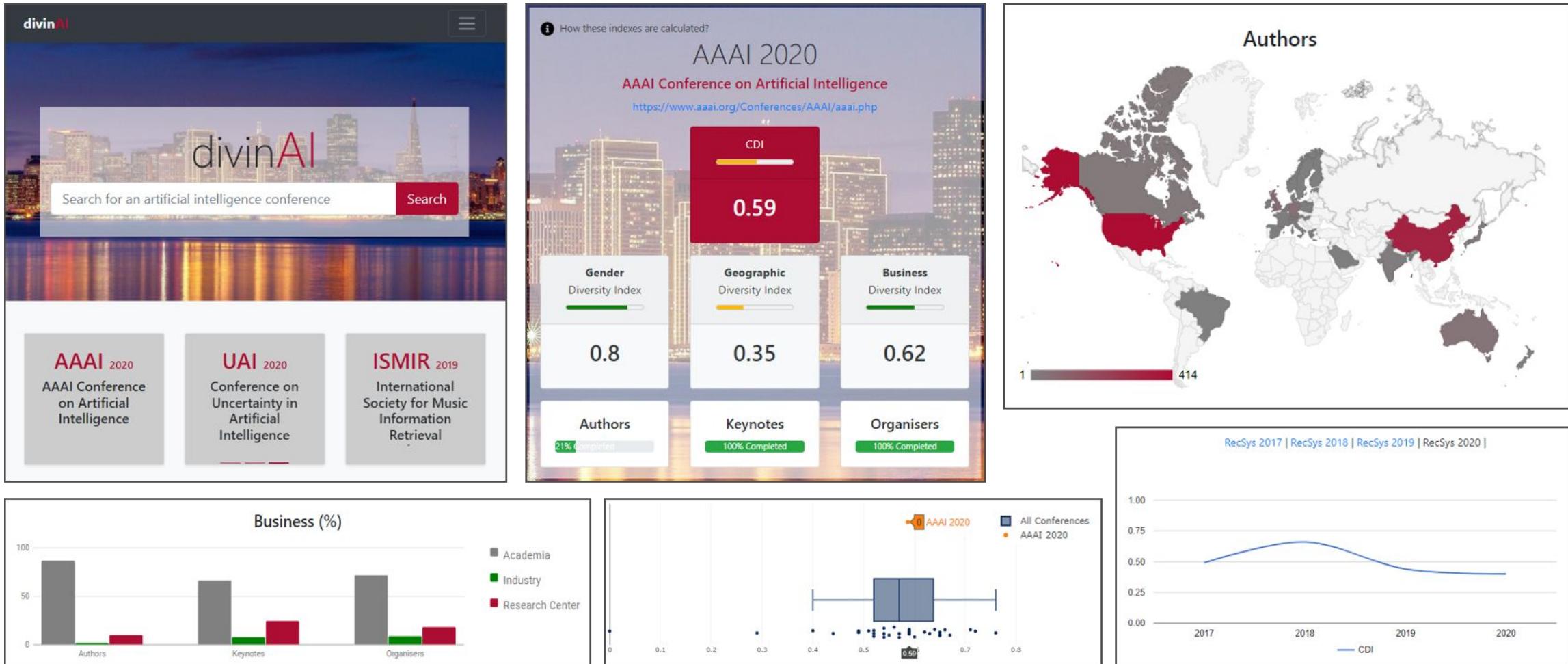
MAIN FOCUS	YEARS	METRICS	DIVERSITY DIMENSIONS							
			Gender	Sex. orient.	Ethnicity	Age	Countries	Institutions	Topics	Cross
BiasWatchNeuro [13] Neuroscience: keynote speakers in >50 conferences and 4 journals	2015-2021	Women rate with respect to "base" rate	x							
Neuroscience: speakers in 18 conferences [14]	2019-2020	Percentages		x						
Geoscience: 9 societies, 25 journals and 10 conferences (organisation committee members) [15]	2016	Percentages	x		x			x		
Geoscience: speakers at 1 conference [16]	2017	Percentages	x		x			x	x	
STEM: 1 society and 1 conference (speakers, attendees and poster presenters) [18]	2011-2015	Percentages	x					x	x	
Medicine: speakers at 1 conference [20]	2016-2018	Percentages, speaking time	x							
AI Index Report [22] AI: survey data obtained from under-represented group members (women, queer, black) and participants in 1 workshop	2015-2020	Percentages	x	x	x		x	x		
AI Watch Index [23] AI: authors, keynote speakers and PC members in 5 top-tier conferences	2016-2020	4 diversity indexes	x			x	x	x	x	
Affective Computing: authors, keynote speakers and PC members in ACII conference [11]	2005-2019	4 diversity indexes, percentages	x			x	x	x	x	
This work Affective Computing: ACII conference (authors, keynote speakers and PC members), TAFFC journal and AAAC association	2011-2021	8 diversity indexes, percentages, clustering	x			x	x	x	x	x

TABLE 1: Main focus, years, metrics and dimensions analysed in state-of-the-art diversity studies. The last row corresponds to the current work.

# Diversity indicators

- Based on dual-concept diversity (McDonald and Dimmick, 2003)
  - Variety: number of categories in a population.
  - Balance: evenness of distribution across categories.
- Examples: Shannon, Pielou, Simpson, Herfindahl-Hirschman.
- Add a 3rd dimension to account for similarity among categories -  
disparity: Rao-Stirling index (stirling, 2007)
- Weighting of dimensions.

# divinAI

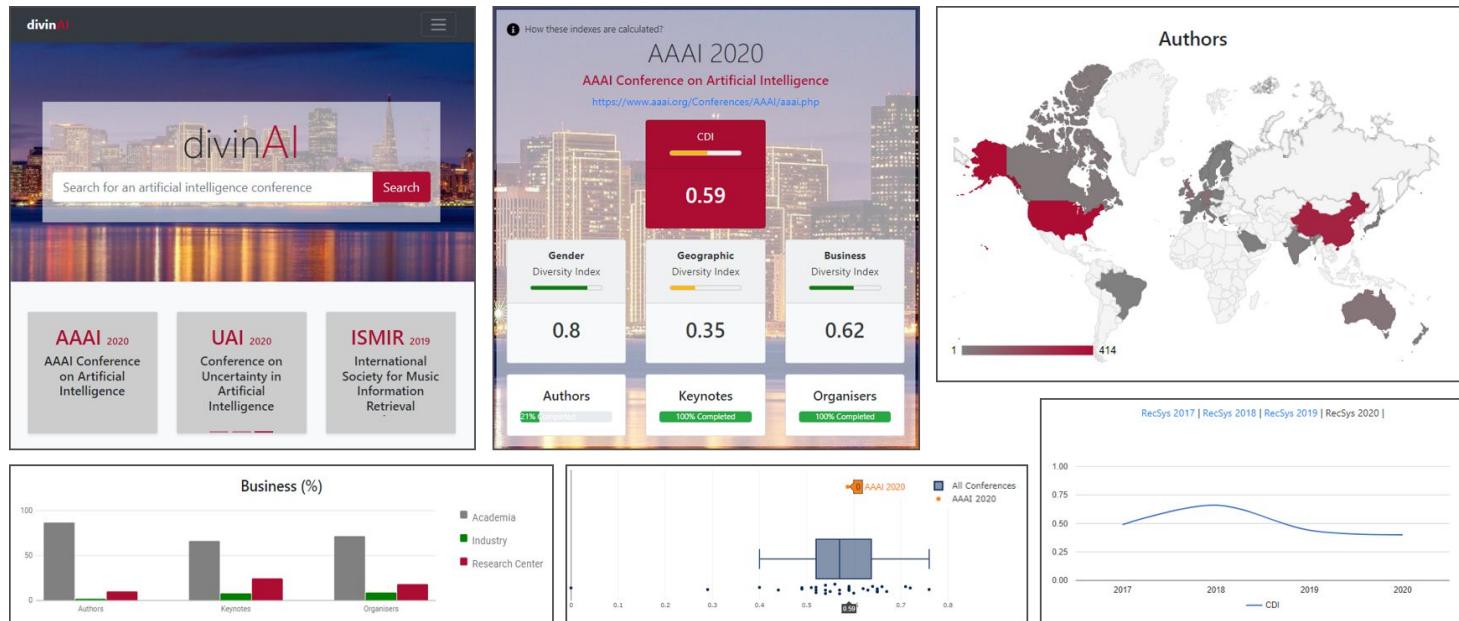


<https://divinai.org/>

Freire, A.; Porcaro, L.; and Gomez, E. 2021. Measuring diversity of artificial intelligence conferences. *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, 39–50. PMLR.

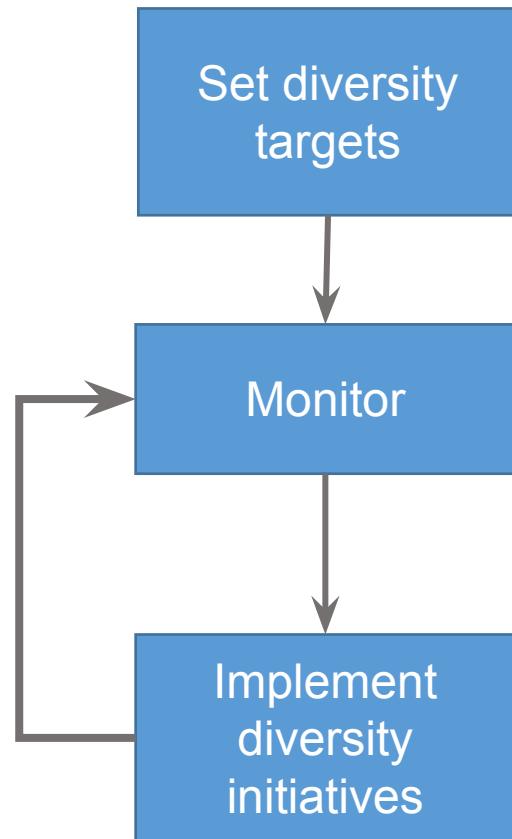
# Monitoring diversity: challenges

- Lack of curated data (country, gender, institution type, topics)
- Ethical concerns:
  - Privacy (personal data) – anonymization, secure storage, consent.
  - Labelling (over-simplification, mislabeling) – self-assignment, manual corrections.

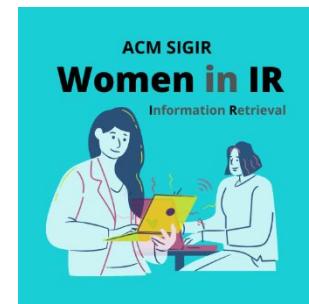


<https://divinai.org/>

# How to enhance diversity?



3



# Diversity groups

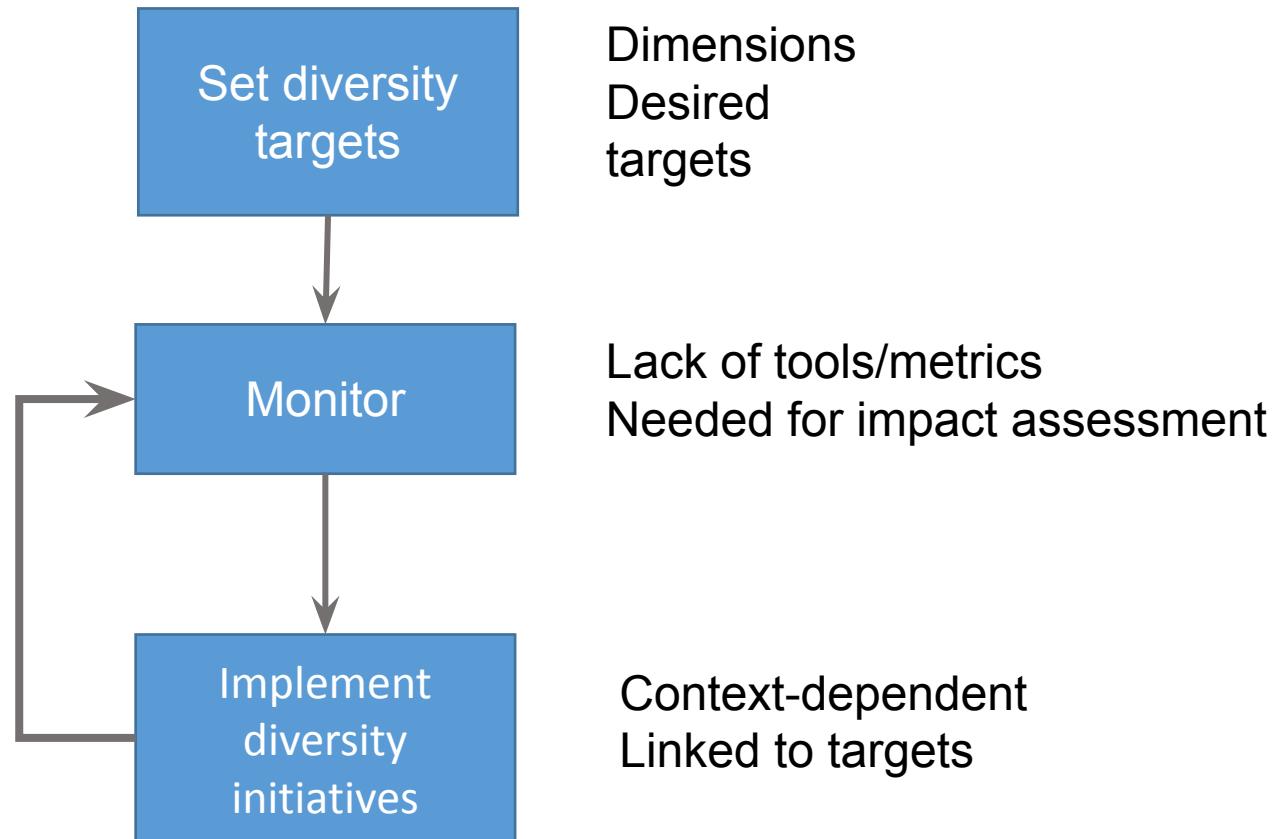
AFFINITY GROUP	SINCE	FOCUS
Women in ML (WiML) <sup>19</sup>	2007	Enhance the experience of women in ML, in order to help them succeed professionally and increase their impact in the community.
Women in MIR (WiMIR) <sup>20</sup>	2012	Promote the role of, and increase opportunities for, women, trans or non-binary at any career stage in the field of music information retrieval.
Women in RecSys <sup>21</sup>	2014	Foster diversity and celebrate female role models in the recommender systems research community.
Women in CV (WiCV) <sup>22</sup>	2015	Foster the career and mitigate the isolation of female researchers working on computer vision.
Black in AI <sup>23</sup>	2017	Increasing the presence and inclusion of Black people in the field of AI.
Widening NLP (WiNLP) <sup>24</sup>	2017	Help to promote and support ideas and voices of under-represented groups in the Natural Language Processing community.
LatinX in AI <sup>25</sup>	2018	Latin professionals working on AI, ML and Data Science.
Queer in AI <sup>26</sup>	2019	People with diverse non-normative sexual orientations, romantic orientations and/or genders, corresponding to acronyms like LGBTQ+.
{Dis}Ability in AI <sup>27</sup>	2019	All those who experience barriers in accessing education due to having or being considered to have an impairment (e.g. physical or sensory impairments, people with learning difficulties, people with mental health or autism spectrum conditions).
Indigenous AI <sup>28</sup>	2019	Design and create AI from an ethical position that centers Indigenous concerns. The Indigenous term covers diverse communities in Aotearoa, Australia, North America and the Pacific.
African Women in AI (AWAI) <sup>29</sup>	2022	Promote knowledge sharing within the African women AI and ML community.

# Diversity initiatives

- Rotation of conference location
- Specific workshops
- Dedicated panel at main conference
- Social events (**Diversity, Equity and Inclusion Lunch July 13<sup>th</sup> 12:30**)
- Call for activities
- Diversity and inclusion chairs – **DEI**
- Directory/profiles
- Mailing list
- Financial support
- Mentoring program
- Journal/call



# How to enhance diversity?



# **Outline**

- Motivation
- Diversity in research communities
- **Diversity by design**

# Outline

- Motivation
- Diversity in research communities
- **Diversity by design:** the case of Music Recommender Systems

(thanks to Lorenzo Porcaro, Carlos Castillo)

# Outline

- Motivation
- Diversity in research communities
- **Diversity by design:** the case of Music Recommender Systems
  1. Music, diversity & recommender systems

# When listening to music...

<https://vimeo.com/679768136>

Diversity

Differences

# The MIR perspective

1. [Demographic Diversity] What is the demographic makeup of MIR as a profession?
2. [Cultural Diversity] Whose music and which music gets to be the focus of MIR's influential scientific practices?
3. [Methodological Diversity] How can MIR equip itself with epistemologies and ontologies of music responsive to a greater diversity of musical cultures?
4. [Goal Diversity] Could MIR cultivate a more plural set of orientations and institutional partners so as to include non-commercial, publicly-oriented initiatives aimed at enhancing human musical flourishing?

# The MIR perspective

1. [Demographic Diversity] What is the demographic makeup of MIR as a profession?
2. **[Cultural Diversity]** Whose music and which music gets to be the focus of MIR's influential scientific practices?
3. [Methodological Diversity] How can MIR equip itself with epistemologies and ontologies of music responsive to a greater diversity of musical cultures?
4. [Goal Diversity] Could MIR cultivate a more plural set of orientations and institutional partners so as to include non-commercial, publicly-oriented initiatives aimed at enhancing human musical flourishing?

# The MIR perspective

1. [Demographic Diversity] What is the demographic makeup of MIR as a profession?
2. [Cultural Diversity] Whose music and which music gets to be the focus of MIR's influential scientific practices?
3. **[Methodological Diversity]** How can MIR equip itself with epistemologies and ontologies of music responsive to a greater diversity of musical cultures?
4. [Goal Diversity] Could MIR cultivate a more plural set of orientations and institutional partners so as to include non-commercial, publicly-oriented initiatives aimed at enhancing human musical flourishing?

# The MIR perspective

1. [Demographic Diversity] What is the demographic makeup of MIR as a profession?
2. [Cultural Diversity] Whose music and which music gets to be the focus of MIR's influential scientific practices?
3. [Methodological Diversity] How can MIR equip itself with epistemologies and ontologies of music responsive to a greater diversity of musical cultures?
4. **[Goal Diversity]** Could MIR cultivate a more plural set of orientations and institutional partners so as to include non-commercial, publicly-oriented initiatives aimed at enhancing human musical flourishing?

# The MIR perspective

1. [Demographic Diversity] What is the demographic makeup of MIR as a profession?
2. [Cultural Diversity] Whose music and which music gets to be the focus of MIR's influential scientific practices?
3. [Methodological Diversity] How can MIR equip itself with epistemologies and ontologies of music responsive to a greater diversity of musical cultures?
4. [Goal Diversity] Could MIR cultivate a more plural set of orientations and institutional partners so as to include non-commercial, publicly-oriented initiatives aimed at enhancing human musical flourishing?

# The Media Perspective (1/2)

Deconstructing the diversity principle:

- ❖ [Source diversity] The range of information providers  
*e.g. artists and record labels.*
- ❖ [Content diversity] The range of information provided  
*e.g. tracks, albums.*
- ❖ [Exposure diversity] The range of information accessed by people  
*e.g. what listeners choose to listen.*

# The Media Perspective (1/2)

Deconstructing the diversity principle:

- ❖ [Source diversity] The range of information providers  
*e.g. artists and record labels.*
- ❖ [Content diversity] The range of information provided  
*e.g. tracks, albums.*
- ❖ [Exposure diversity] The range of information accessed by people  
*e.g. what listeners choose to listen.*

# The Media Perspective (1/2)

Deconstructing the diversity principle:

- ❖ [Source diversity] The range of information providers  
*e.g. artists and record labels.*
- ❖ [Content diversity] The range of information provided  
*e.g. tracks, albums.*
- ❖ [Exposure diversity] The range of information accessed by people  
*e.g. what listeners choose to listen.*

# The Media Perspective (1/2)

Deconstructing the diversity principle:

- ❖ [Source diversity] The range of information providers  
*e.g. artists and record labels.*
- ❖ [Content diversity] The range of information provided  
*e.g. tracks, albums.*
- ❖ [Exposure diversity] The range of information accessed by people  
*e.g. what listeners choose to listen.*

# **Which aspect of diversity have you considered/is more relevant for your research?**

sli.do

#sigir22ethics

Join at

**slido.com**

**#sigir22ethics**



# The Media Perspective (2/2)

Diversity by design: the creation of an architecture or service that helps people to make diverse choices.

- ❖ [Individual autonomy perspective] Provide people with a tool for exploiting their interests  
*e.g. calibrated recommendations.*
- ❖ [Deliberative perspective] promote public awareness by showing divergent opinions  
*e.g. make listeners explore music far from their preferences.*
- ❖ [Adversarial perspective] enhance the visibility of underrepresented opinions  
*e.g. promote underrepresented groups e.g. subcultures or non-mainstream musical styles.*

Helberger, N. (2011). Diversity by design. *Journal of Information Policy*, 1(2011), 441–469. <https://doi.org/10.5325/jinfopol.1.2011.0441>

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information Communication and Society*, 21(2), 191–207. <https://doi.org/10.1080/1369118X.2016.1271900>

# The Media Perspective (2/2)

Diversity by design: the creation of an architecture or service that helps people to make diverse choices.

- ❖ [Individual autonomy perspective] Provide people with a tool for exploiting their interests  
*e.g. calibrated recommendations.*
- ❖ [Deliberative perspective] promote public awareness by showing divergent opinions  
*e.g. make listeners explore music far from their preferences.*
- ❖ [Adversarial perspective] enhance the visibility of underrepresented opinions  
*e.g. promote underrepresented groups e.g. subcultures or non-mainstream musical styles.*

Helberger, N. (2011). Diversity by design. *Journal of Information Policy*, 1(2011), 441–469. <https://doi.org/10.5325/jinfopol.1.2011.0441>

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information Communication and Society*, 21(2), 191–207. <https://doi.org/10.1080/1369118X.2016.1271900>

# The Media Perspective (2/2)

Diversity by design: the creation of an architecture or service that helps people to make diverse choices.

- ❖ [Individual autonomy perspective] Provide people with a tool for exploiting their interests  
*e.g. calibrated recommendations.*
- ❖ [Deliberative perspective] promote public awareness by showing divergent opinions  
*e.g. make listeners explore music far from their preferences.*
- ❖ [Adversarial perspective] enhance the visibility of underrepresented opinions  
*e.g. promote underrepresented groups e.g. subcultures or non-mainstream musical styles.*

Helberger, N. (2011). Diversity by design. *Journal of Information Policy*, 1(2011), 441–469. <https://doi.org/10.5325/jinfopol.1.2011.0441>

Helberger, N., Karppinen, K., & D'Acunto, L. (2018). Exposure diversity as a design principle for recommender systems. *Information Communication and Society*, 21(2), 191–207. <https://doi.org/10.1080/1369118X.2016.1271900>

# The Media Perspective (2/2)

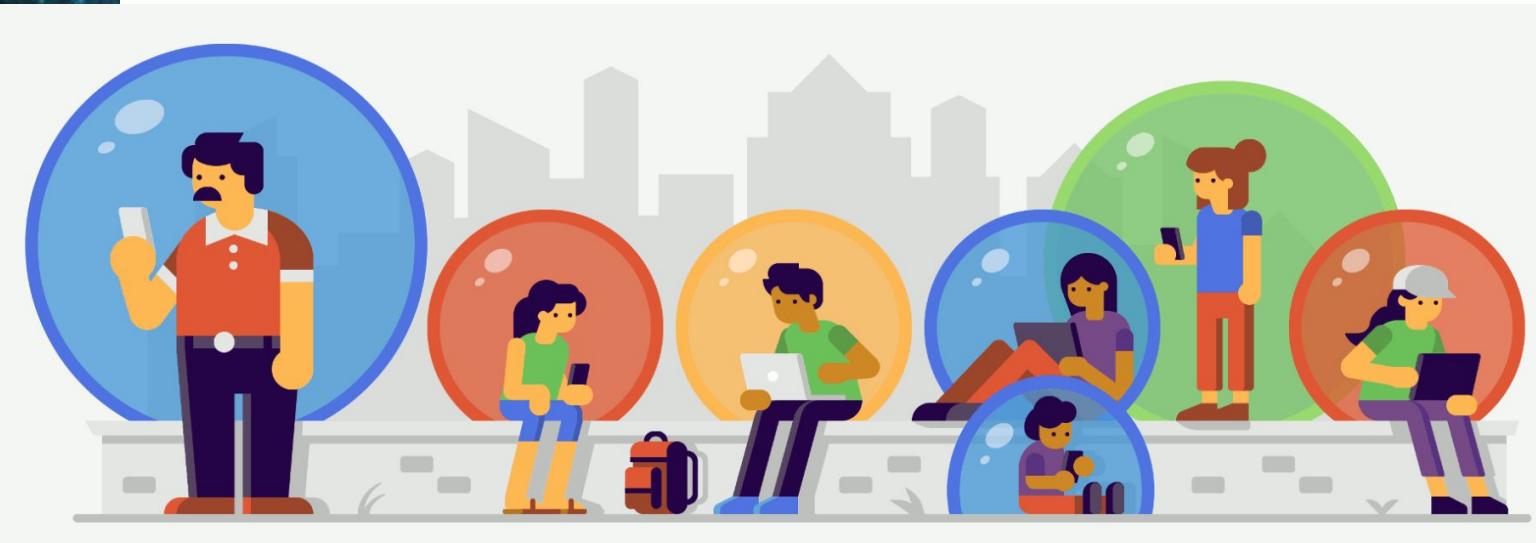
Diversity by design: the creation of an architecture or service that helps people to make diverse choices.

- ❖ [Individual autonomy perspective] Provide people with a tool for exploiting their interests  
*e.g. calibrated recommendations.*
- ❖ [Deliberative perspective] promote public awareness by showing divergent opinions  
*e.g. make listeners explore music far from their preferences.*
- ❖ [Adversarial perspective] enhance the visibility of underrepresented opinions  
*e.g. promote underrepresented groups e.g. subcultures or non-mainstream musical styles.*

# Echo Chambers



# Filter Bubbles



# Cyber Fragmentation

# Music recommendation algorithms are unfair to female artists, but we can change that

Representation of women & gender minorities in the music industry is low, and streaming services mimic this bias

COUNTRY

Martina McBride 'Felt Like We'd Been Erased' When Spotify Didn't Recommend a Single Female Country Artist



By Annie Reuter  
9/16/2019

## Why Spotify's music recommendations always seem so spot on

Spotify knows what you like to hear, and isn't afraid to tell you.



The YouTube algorithm is becoming scarily good at recommending your next listen. I, for one, welcome our new overlord.



33



Comments 1K



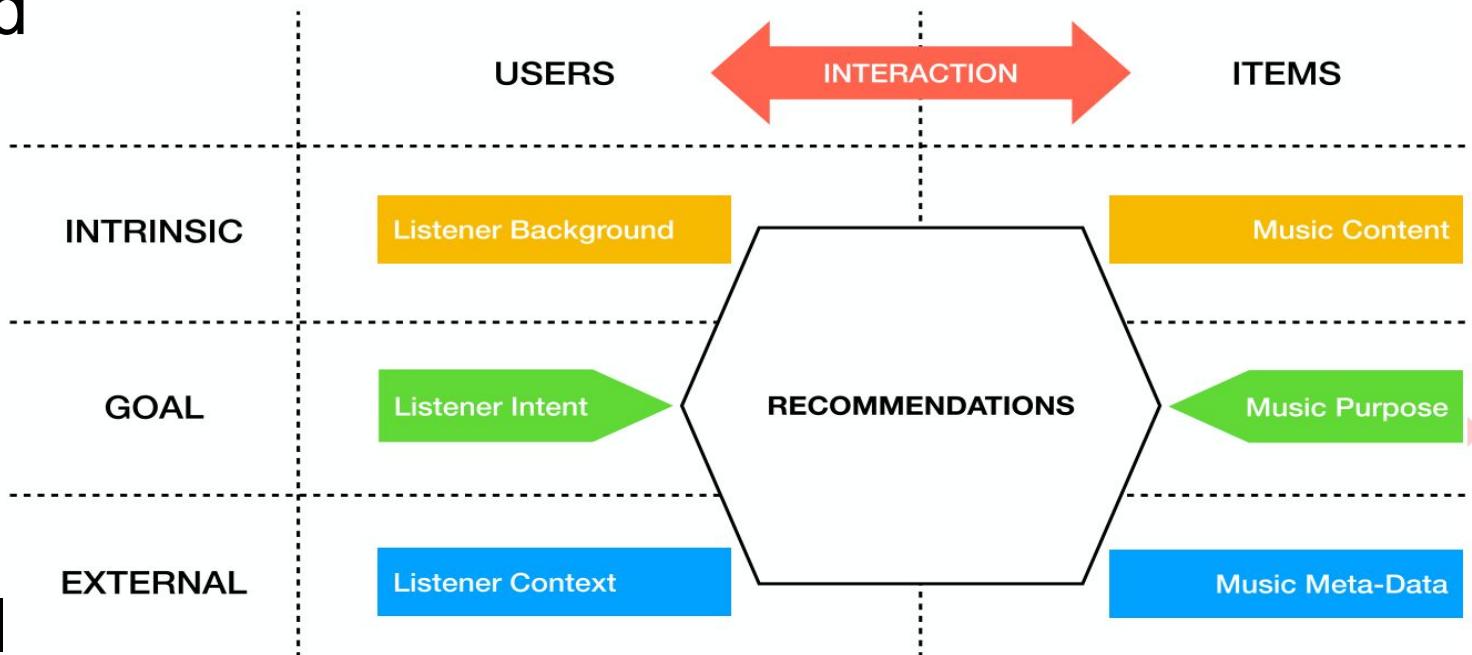
Shout out to the YouTube algorithm for bringing us here. We have similar tastes? I think.

# Outline

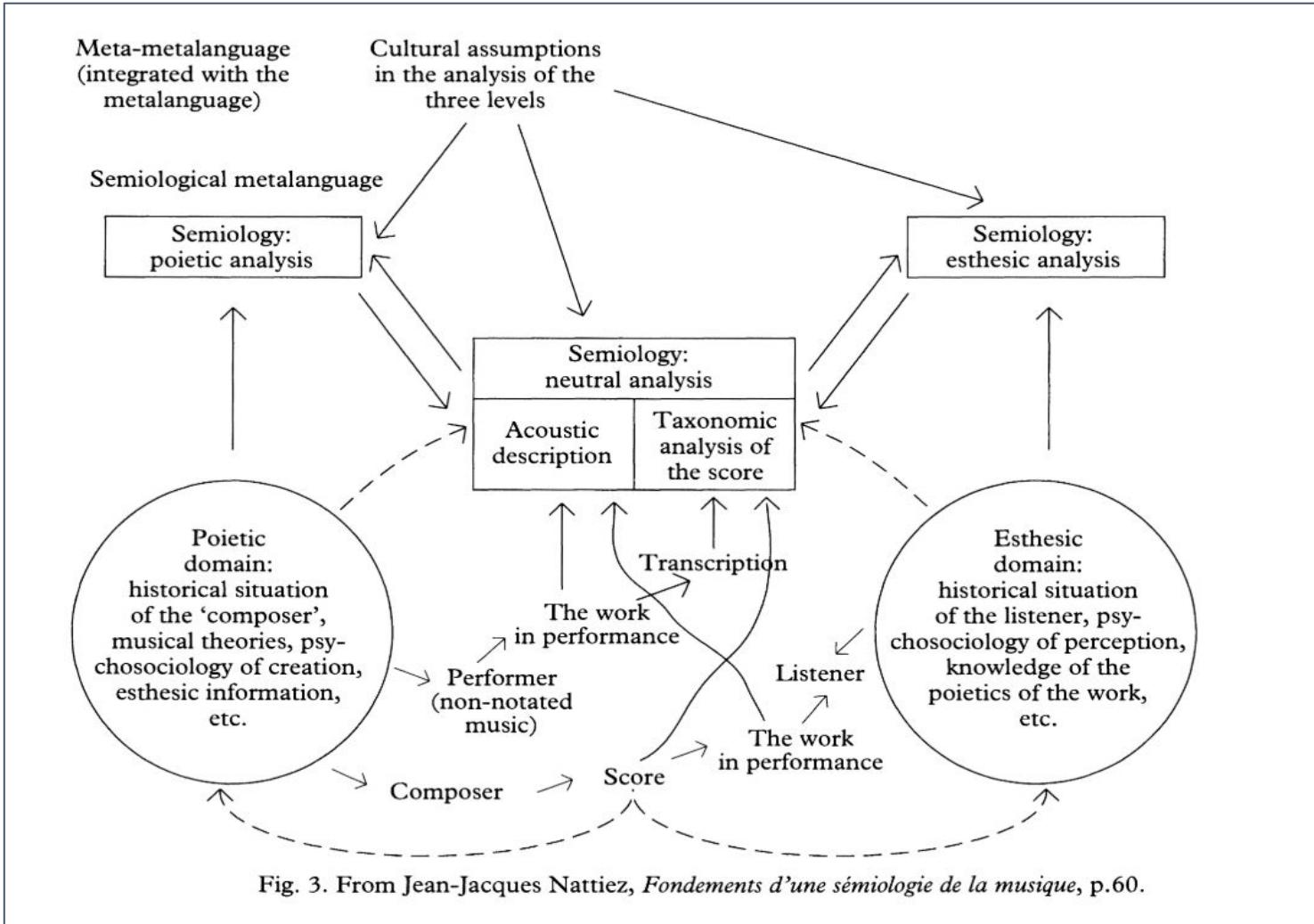
- Motivation
- Diversity in research communities
- **Diversity by design:** the case of Music Recommender Systems
  1. Music, diversity & recommender systems
  2. Examples from the Music IR literature

# The (Music) Recommender Systems Framework

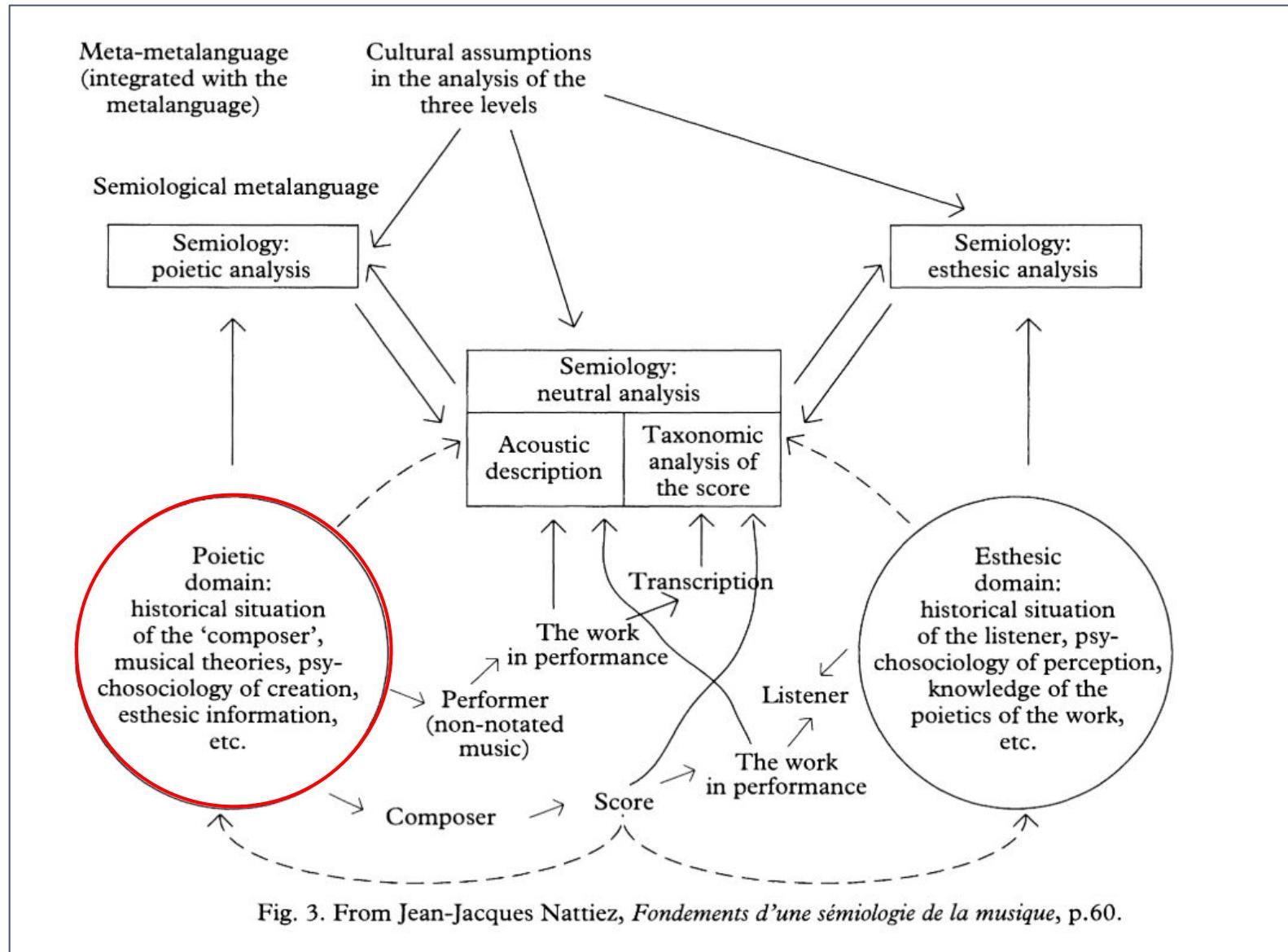
- Music is often consumed passively, sequentially, repetitively.
- Small duration of the items is quite small, big size of the catalogue.
- Listening intent, and context are fundamental aspects.



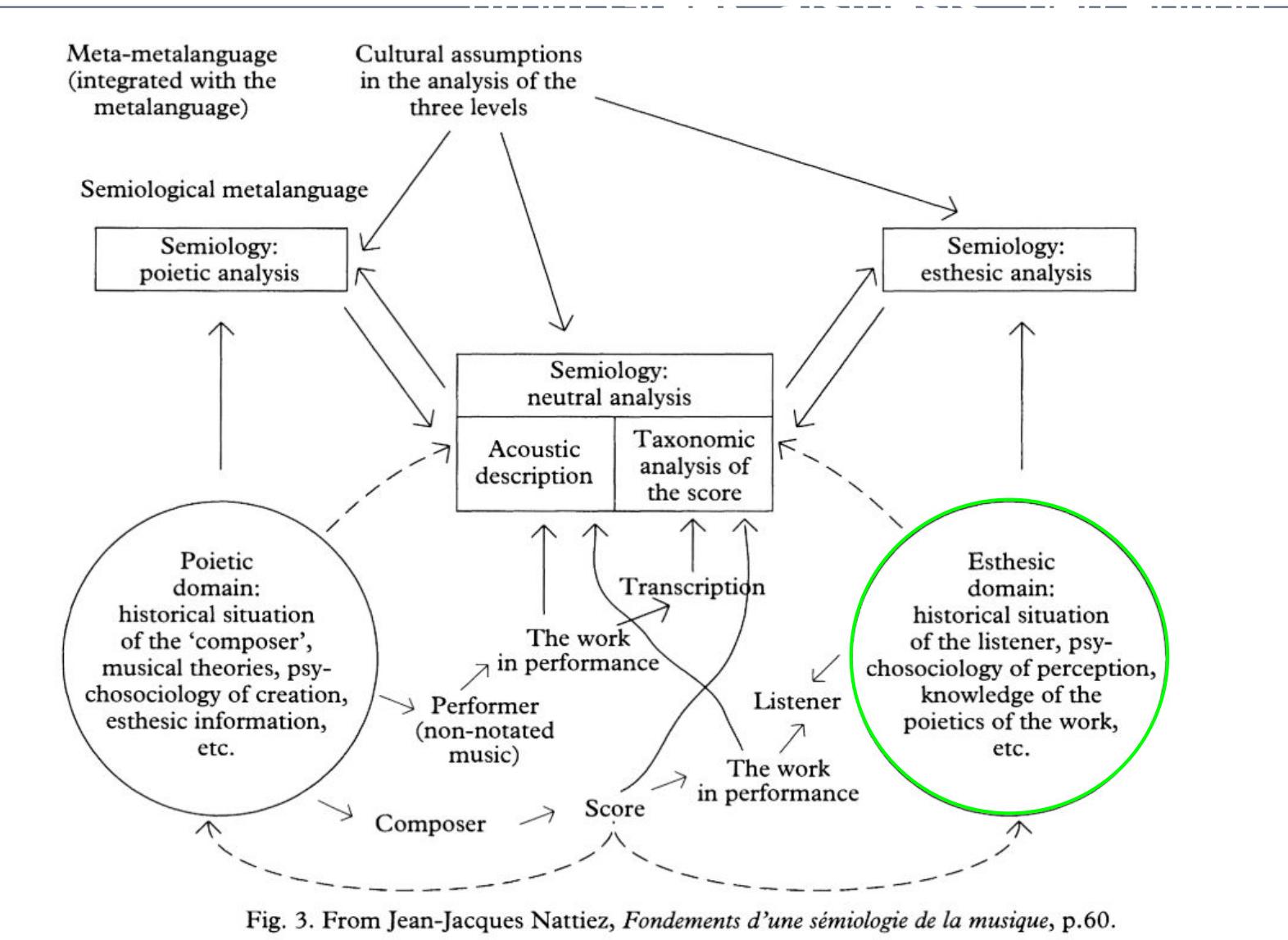
**Semiology (“study of signs”)** → Discipline that studies the phenomena of signification and communication.



# Poietic Domain (from Greek: poiētikós, 'creative') - The Item Side



# Esthetic domain - The User Side



# Poietic Domain - the Item side

- How often a user listen to each track in her collection on average (count of different items with which users interact).

$$diversity_u = \frac{\text{total number of playcounts of } u}{|\text{unique items } u \text{ listened to}|}$$

- Distinct genre tags in a user listening profile.

$$diversity_u = |\text{unique genre tags that describes } u \text{ music taste}|$$

# Poietic Domain - the Item side

- How often a user listen to each track in her collection on average (count of different items with which users interact).

$$diversity_u = \frac{\text{total number of playcounts of } u}{|\text{unique items } u \text{ listened to}|}$$

- Distinct genre tags in a user listening profile.

$$diversity_u = |\text{unique genre tags that describes } u \text{ music taste}|$$

**Pro:** Not complex formulation and relatively simple implementation.

**Cons:** No use of any additional features to differentiate between items.

# Poietic Domain - the Item side

Diversity as distribution of the user-item interactions + distance spaces containing additional information.

*Rao-Stirling Index:*

- $p_i$  and  $p_j$  := fraction of streams from genres  $i$  and  $j$
- $d(i, j)$  := dissimilarity of the two genres
- $K$  := genres listened to by a user

$$d_{RS}(p) = \sum_{i,j \in K} p_i \times p_j \times d(i, j)$$

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15), 707–719. <https://doi.org/10.1098/rsif.2007.0213>

Way, S. F., Gil, S., Anderson, I., & Clauset, A. (2019). Environmental Changes and the Dynamics of Musical Identity. *Proceedings of the International AAAI Conference on Web and Social Media*, 1–10. <http://arxiv.org/abs/1904.04948>

# Poietic Domain - the Item side

Diversity as distribution of the user-item interactions + distance spaces containing additional information.

*Rao-Stirling Index:*

- $p_i$  and  $p_j$  := fraction of streams from genres  $i$  and  $j$
- $d(i, j)$  := dissimilarity of the two genres
- $K$  := genres listened to by a user

$$d_{RS}(p) = \sum_{i,j \in K} p_i \times p_j \times d(i, j)$$

**Pro:** Items' fine-grained features for estimating diversity.

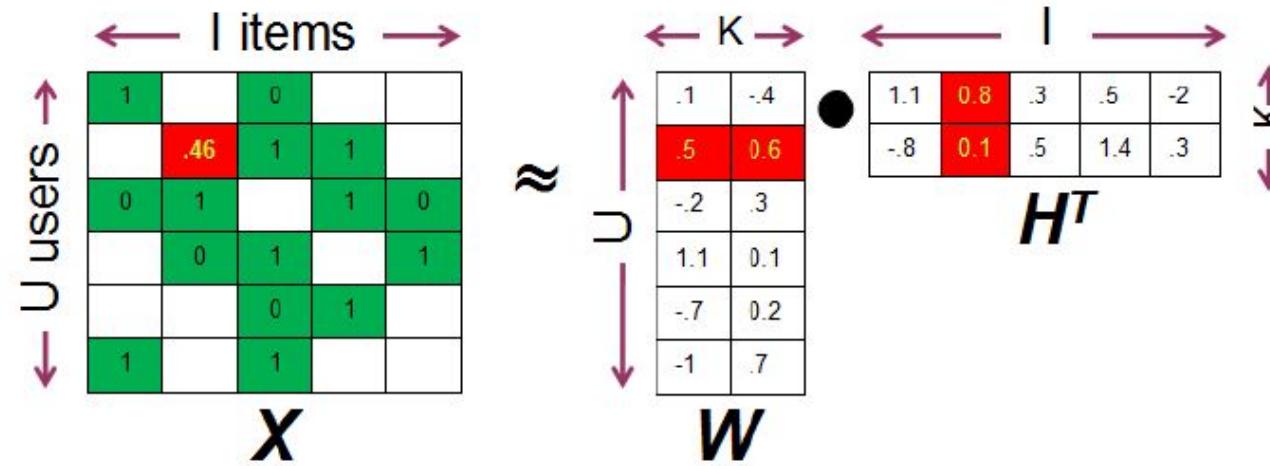
**Cons:** Expensive in terms of data and computational resources. “Dissimilarity as a research problem”

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of The Royal Society Interface*, 4(15), 707–719. <https://doi.org/10.1098/rsif.2007.0213>

Way, S. F., Gil, S., Anderson, I., & Clauset, A. (2019). Environmental Changes and the Dynamics of Musical Identity. *Proceedings of the International AAAI Conference on Web and Social Media*, 1–10. <http://arxiv.org/abs/1904.04948>

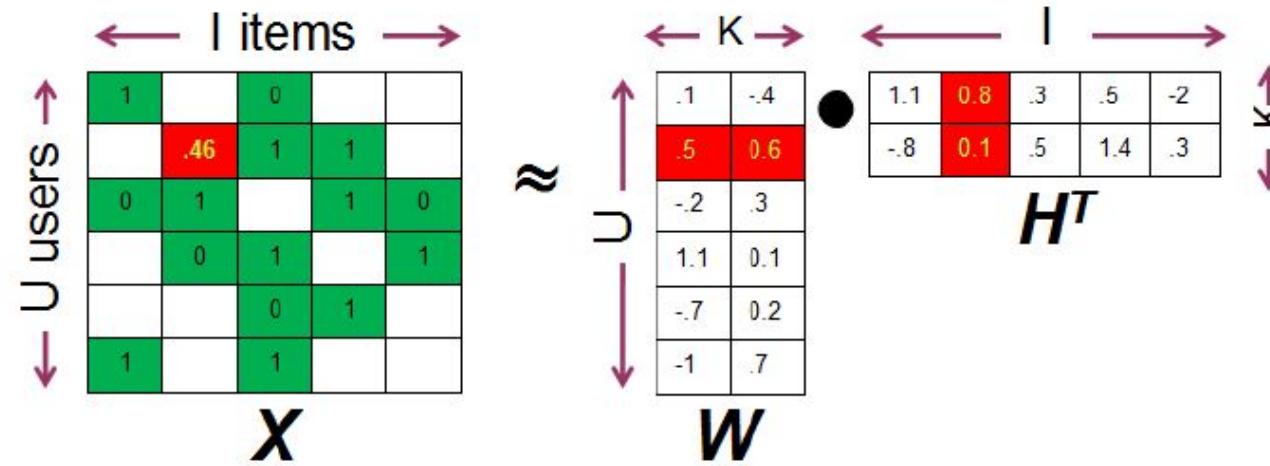
# Poietic Domain - the Item side

Diversity as the distance between item vectors in the Matrix Factorization space.



# Poietic Domain - the Item side

Diversity as the distance between item vectors in the Matrix Factorization space.



**Pro:** Required only the user-item interaction matrix.

**Cons:** Little interpretability of the latent space.

# Poietic Domain - the Item side

- ❖ Measuring item diversity connected with the users' behaviours (*exposure diversity*).
- ❖ *Content and source diversity* considered in works centered on music lists (e.g. playlists).
- ❖ *The user is left aside!*  
Grouping users by their diversity = grouping them by the diversity of the items they consumed.

# **Esthetic domain – the User side** (*Individual aspects*)

**Personality traits** → Big Five personality traits (OCEAN):

- **Openness to Experience**
- **Conscientiousness**
- **Extraversion**
- **Agreeableness**
- **Neuroticism**

# **Esthetic domain – the User side** (*Individual aspects*)

**Personality traits** → Big Five personality traits (OCEAN):

- Openness to Experience
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism



“conscientious participants are increasingly satisfied when provided a higher degree of diversity”

Ferwerda, B., Graus, M., Vall, A., Tkalcic, M., & Schedl, M. (2016). The influence of users' personality traits on satisfaction and attractiveness of diversified recommendation lists. Proceedings of the 4th Workshop on Emotions and Personality in Personalized Systems (EMPIRE), at the 10th Conference on Recommender Systems (RecSys), 1680, 43–47.

# Esthetic domain – the User side (*Individual aspects*)

## Personal values

Conservation (*caring about one's safety in every aspects of one's life*)

Openness to Change (*caring about independence and discovery*)

Self-Transcendence (*caring for the world*)

Self-Enhancement (*caring for oneself*)

Hedonism

Manolios, S., Hanjalic, A., & Liem, C. C. S. (2019). *The influence of personal values on music taste: Towards value-based music recommendations*. Proceedings of the 13th ACM Conference on Recommender Systems (RecSys), September 2019, 501–505. <https://doi.org/10.1145/3298689.3347021>

## Musical Sophistication

Active Musical Engagement (*how much time and money resources spent on music*)

Self-reported Perceptual Abilities (*accuracy of musical listening skills*)

Musical Training (*amount of formal musical training received*)

Self-reported Singing Abilities (*accuracy of one's own singing*)

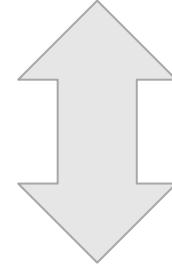
Sophisticated Emotional Engagement with Music (*ability to talk about emotions that music expresses*)

Ferwerda, B., & Tkalcic, M. (2019). Exploring online music listening behaviors of musically sophisticated users. ACM UMAP 2019 Adjunct - Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, 33–37. <https://doi.org/10.1145/3314183.3324974>

# **Esthetic domain – the User side** *(Individual aspects)*

**metric-based diversity**

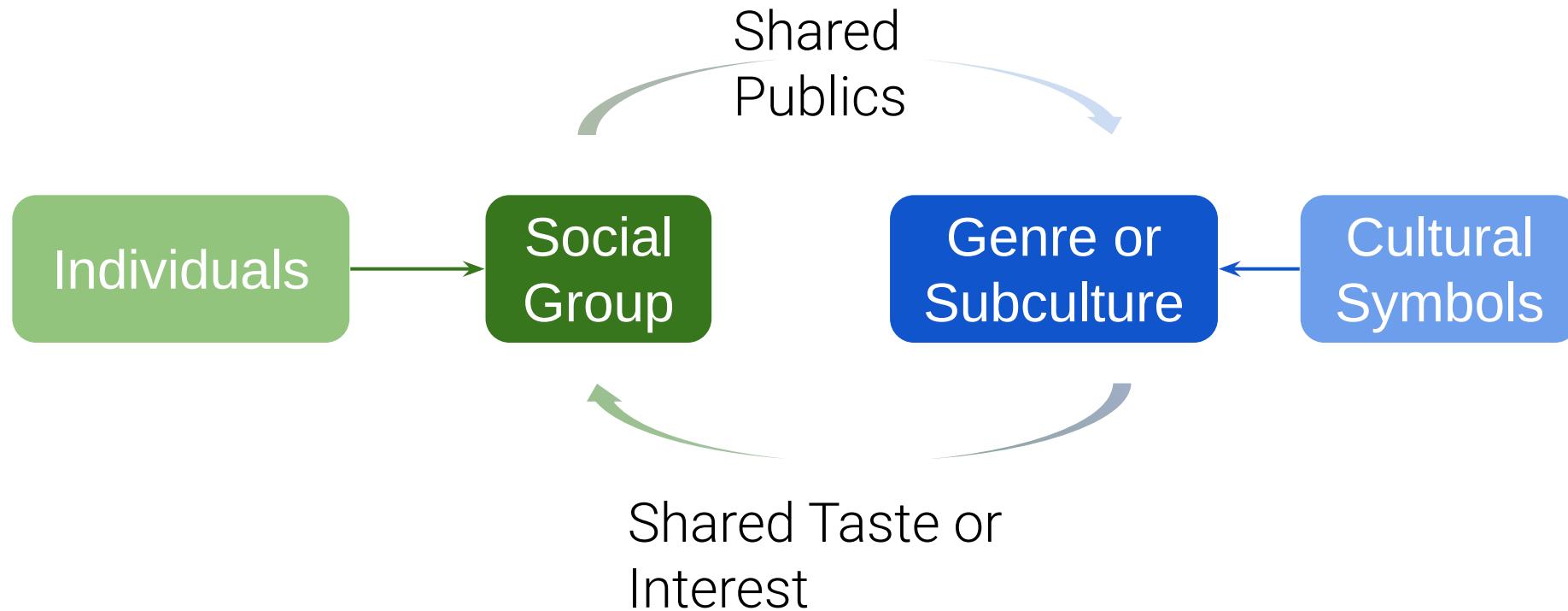
*extractable by algorithmic processes*



**perceived diversity**

*how people evaluate a degree of diversity*

# **Esthetic domain – the User side** *(Collective aspects)*



# **Esthetic domain – the User side** (*Collective aspects*)

- ❖ Lack of data publicly available:

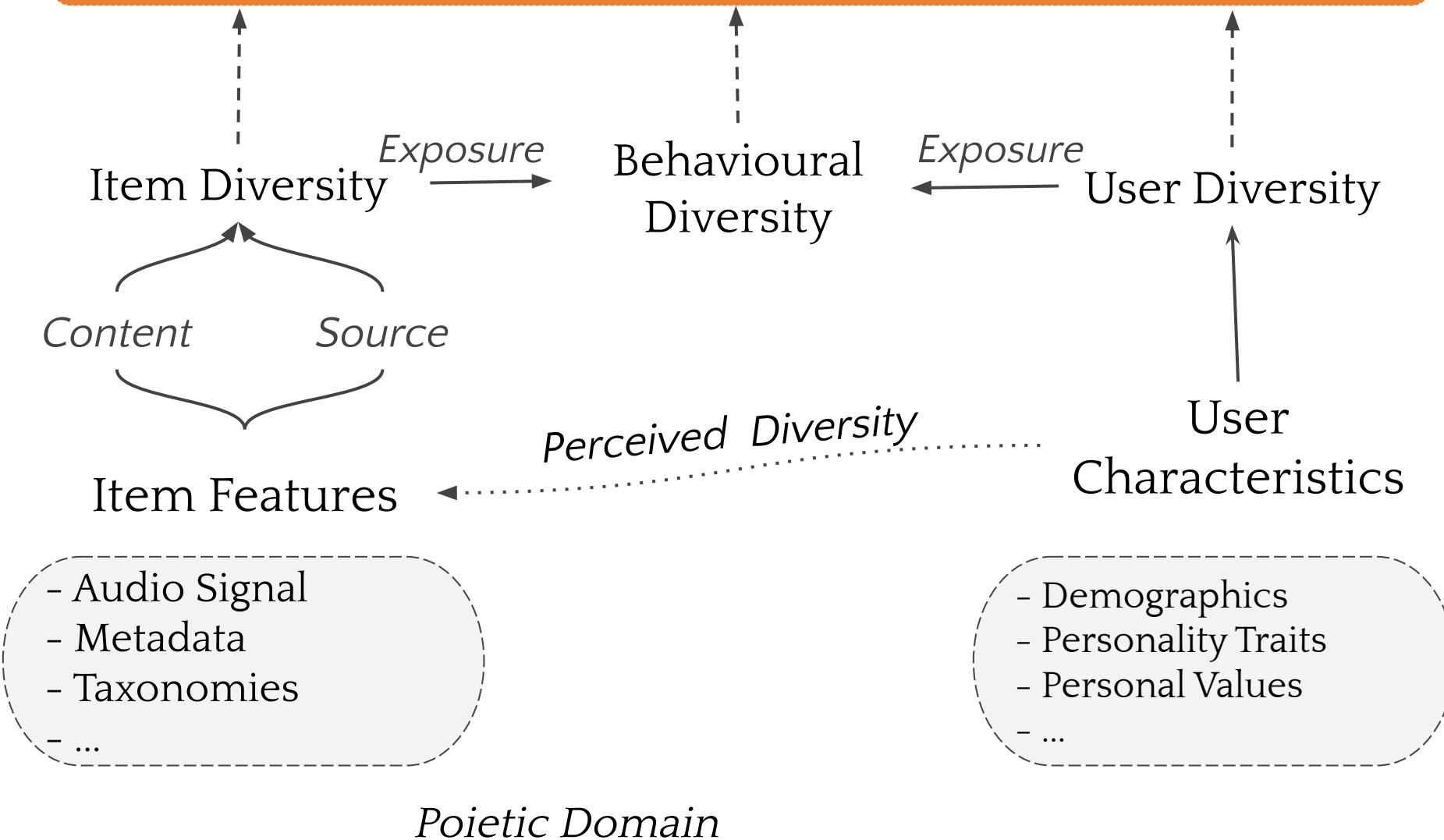
- Demographic information (Cross-country analysis)

e.g. Ferwerda, B., Vall, A., Tkalcic, M., & Schedl, M. (2016). *Exploring Music Diversity Needs Across Countries*. Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16), 287–288. <https://doi.org/10.1145/2930238.2930262>

- Socio-economic factors (Sociological-informed analysis)

e.g. Park, M., Weber, I., Naaman, M., & Vieweg, S. (2016). *Understanding Musical Diversity via Online Social Media*. Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM'16). <http://arxiv.org/abs/1604.02522>

# Diversity by design in Music RS



Those who study music should be concerned about the loss of **cultural diversity** for the same reason that biologists worry about the loss of biodiversity: we don't yet know what the loss will mean, but we do know that the loss will be **irreversible.**

# Additional references

- Evaluation procedures, algorithmic solutions and empirical results in recommender systems research (Castells et al. 2015)
- Diversity-related metrics (Kaminskas and Bridge, 2016)
- Overview of recommender systems diversification techniques: algorithmic solutions and evaluation practices (Junaver and Požrl, 2017)
- Role of diversity in Big Data applications: selection task (Drosou et al., 2017)
- Bias (data, algorithm, user interaction) on web systems (Baeza-Yates, 2018)

Castells, P., Hurley, N. J., and Vargas, S. (2015). Novelty and diversity in recommender systems. In Ricci, F., Rokach, L., and Shapira, B., editors, Recommender Systems Handbook, pages 881–918. Springer, Boston, MA. DOI: [https://doi.org/10.1007/978-1-4899-7637-6\\_26](https://doi.org/10.1007/978-1-4899-7637-6_26)

Kaminskas, M., and Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Transactions on Interactive Intelligent Systems, 7(1): 1–42. DOI: <https://doi.org/10.1145/2926720>

Kunaver, M., and Požrl, T. (2017). Diversity in recommender systems: A survey. Knowledge-Based Systems, 123: 154–162. DOI: <https://doi.org/10.1016/j.knosys.2017.02.009>

Drosou, M., Jagadish, H., Pitoura, E., and Stoyanovich, J. (2017). Diversity in big data: A review. Big Data, 5(2): 73–84. DOI: <https://doi.org/10.1089/big.2016.0054>

Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6): 54–61. DOI: <https://doi.org/10.1145/3209581>

# **Part 3: Transparency**

# Outline

- Motivation & EU regulations
- Categories of Transparency
- Explainability
- Traceability and Auditability
- Documentation

# Motivation

- IR and RS systems should be able to *explain their decisions*
  - why are results shown to a user
  - how were results retrieved
  - help user assess whether to trust the system
- Particularly when decision making involves sensitive aspects
- More reasons:
  - Reproducibility
  - Accountability
  - System diagnostics & performance

# EU Regulations

- Transparency key feature of EU law
- Also: expression of fairness principle related to processing personal data as described in Article 8 of the Charter of Fundamental Rights of the EU
- EU General Data Protection Regulation (GDPR)
  - Transparency overarching obligation
- 3 central areas:
  - Provision of information to data subjects related to fair processing
  - How data controllers communicate with data subjects in relation to their rights under GDPR
  - How data controllers facilitate the exercise by data subjects of their rights
- Compliance with transparency required related to data processing under Directive 2016/680



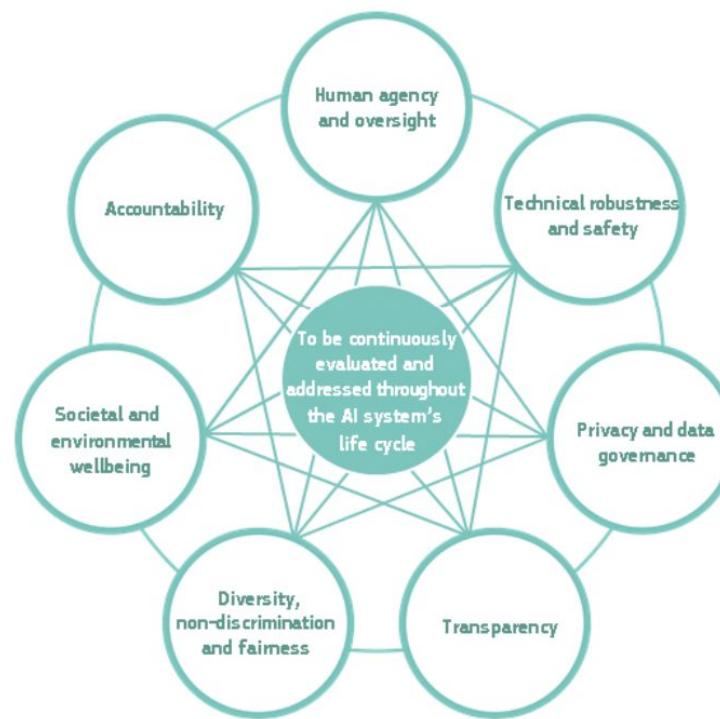
# EU Regulations

- Digital Services Act
  - Online platforms & search engines need to be transparent in terms of recommender systems
  - Plus, advertisements
  - Requirements depend on size of platform measured by number of users
- Artificial Intelligence Act
  - Transparency as a key requirement
  - Besides: technical documentation for high-risk use cases



<https://www.europarl.europa.eu/news/en/press-room/20220412IPR27111/digital-services-act-agreement-for-a-transparent-and-safe-online-environment>  
<https://eur-lex.europa.eu/legal-content/NL/TXT/HTML/?uri=CELEX:52021PC0206&from=EN>

# One of the requirements for trustworthy AI



High-level Expert Group on AI, European Commission, Ethics Guidelines for Trustworthy AI,  
<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

# **Transparency and Fairness**

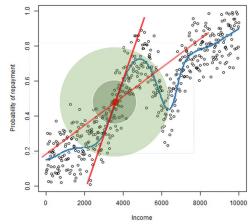
- Fair systems not possible if systems are opaque
  - How do algorithms work: what is in the data
  - How are end users affected
- Transparency enables audits
  - How does the system work
  - And: does system creates fair outputs
- User perceptions of fairness
  - IR /RS explanations may lead to new behavior
  - Taking fair actions; at least, informed choices

# **Outline**

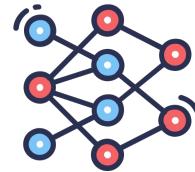
- Motivation & EU regulations
- **Categories of Transparency**
- Explainability
- Traceability and Auditability
- Documentation

# Major Aspects of Transparency

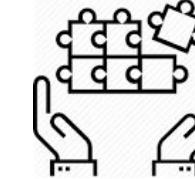
Algorithmic  
Transparency



Simulability



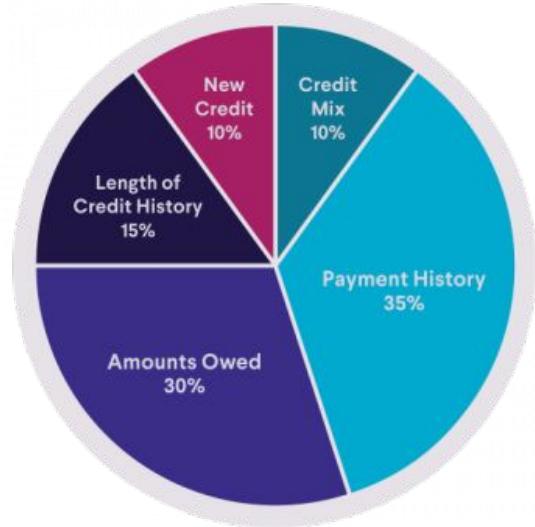
Decomposability



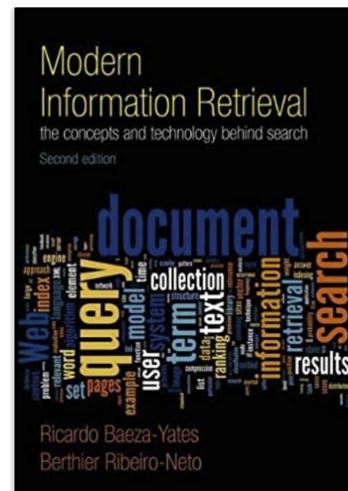
Related concepts: Explainability, Interpretability, Understandability, Black boxes

# Transparency - Understandability

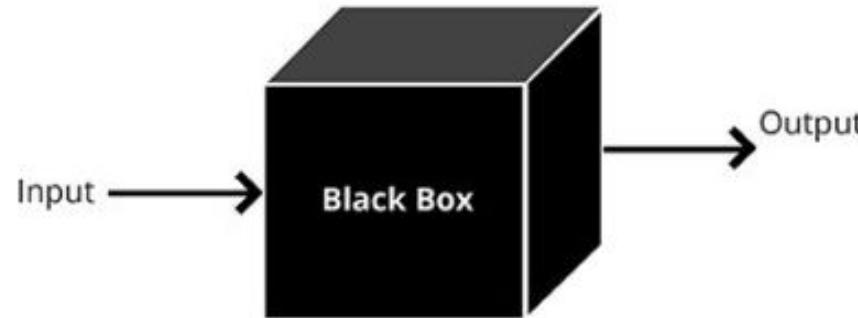
- Decision made by an algorithm should be understandable by those affected by the decision
  - Why was a decision reached based on a given input?



<https://www.sofi.com/learn/content/do-personal-loans-hurt-credit/>



# The Problem of Black Boxes



- Contemporary IR & RS based on complex models: deep learning, ML
- We do not understand what is going on in the box
- Hard for users to understand why output is relevant - trust the prediction?

**Do you think it is sufficient to disclose how  
algorithms came to their decision and tell how  
human could reverse the decision? Why yes?  
Why now?**

sli.do

#sigir22ethics

Join at

[slido.com](https://slido.com)

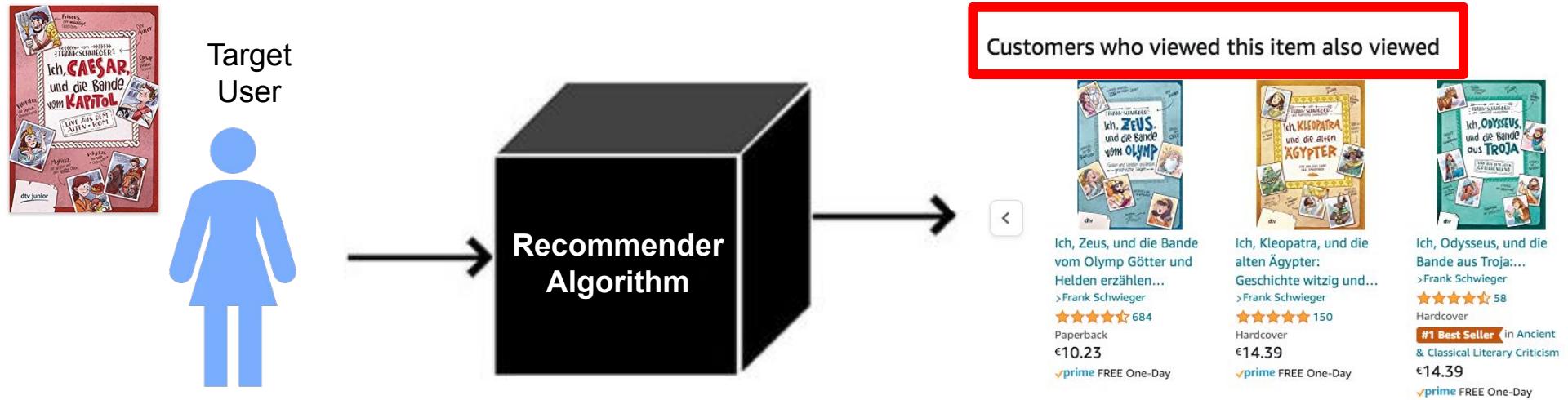
**#sigir22ethics**



# Outline

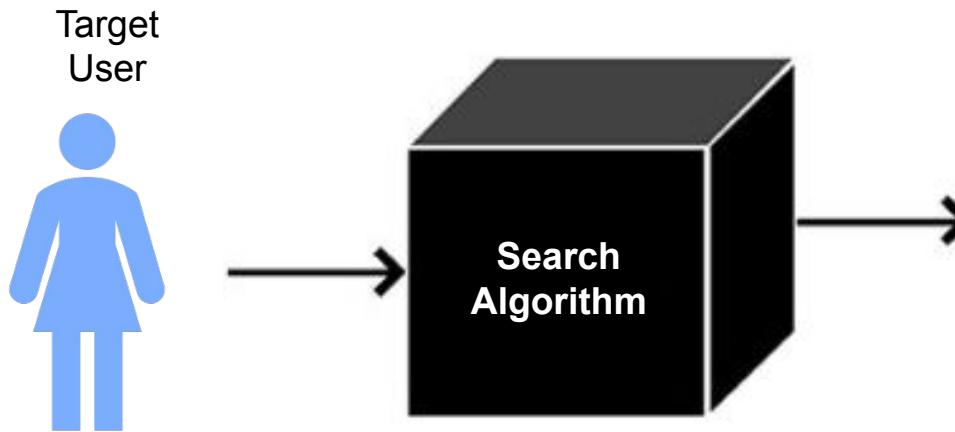
- Motivation & EU regulations
- Categories of Transparency
- **Explainability**
- Traceability and Auditability
- Documentation

# Explanations in Recommender Systems



Task: Given user-item pair, provide **explanation** to justify why item is recommended to the user

# Explanations in IR



About 191.000 results (0,38 seconds)

<https://sigir.org> › sigir2022

**SIGIR 2022 - The 45th International ACM SIGIR Conference ...**

ACM **SIGIR** is the Annual Conference of the Association for Computing Machinery Special Interest Group in Information Retrieval. In 2022, it comes to [pain](#).

#### Call for Full Papers

The 45th ACM SIGIR conference, will be run as a hybrid ...

#### Accepted papers

Hybrid Transformer with Multi-level Fusion for Multimodal ...

#### Call for Short Papers

The 45th ACM SIGIR conference, will be run as a hybrid ...

#### Workshops

The SIGIR 2022 workshop program will host 8 compelling ...

About this result BETA

#### Source

SIGIR is the Association for Computing Machinery's Special Interest Group on Information Retrieval. The scope of the group's specialty is the theory and application of computers to the acquisition, ... [Wikipedia](#)

- <https://sigir.org/sigir2022/>
- Your connection to this site is **secure**

[More about this page](#)

This is a search result, not an ad. Only ads are paid, and they'll always be labeled with "Sponsored" or "Ad."

Explanations in the form of search snippets, query terms highlighted

Additional information to the search result

# Why Explainability?

- Increasingly important role in user interactions with systems
  - Trust in the system
  - Accountability
- Model validation
- Biases, unfairness, problems with training data, legal requirements
- Improvements of model
  - Reliability, robustness,..

# **What makes a good explanation?**

sli.do

#sigir22ethics

Join at

**slido.com**

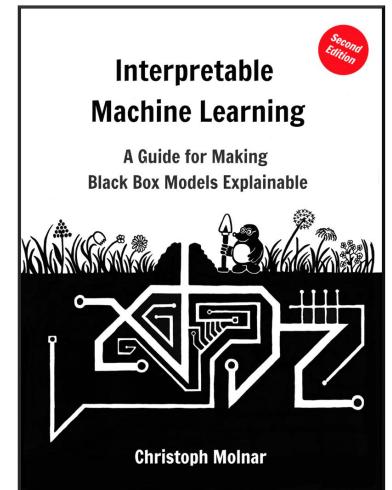
**#sigir22ethics**



# Properties of Good Explanations

- Accuracy
- Fidelity
- Consistency
- Stability
- Comprehensibility

→ see: <https://christophm.github.io/interpretable-ml-book/>



# Explainability in Recommender Systems

“To make clear by giving a detailed description” (Tintarev et al.)

“Explainable recommendation to answer the question of why” (Zhang et al.)

# Explainability in Recommender Systems

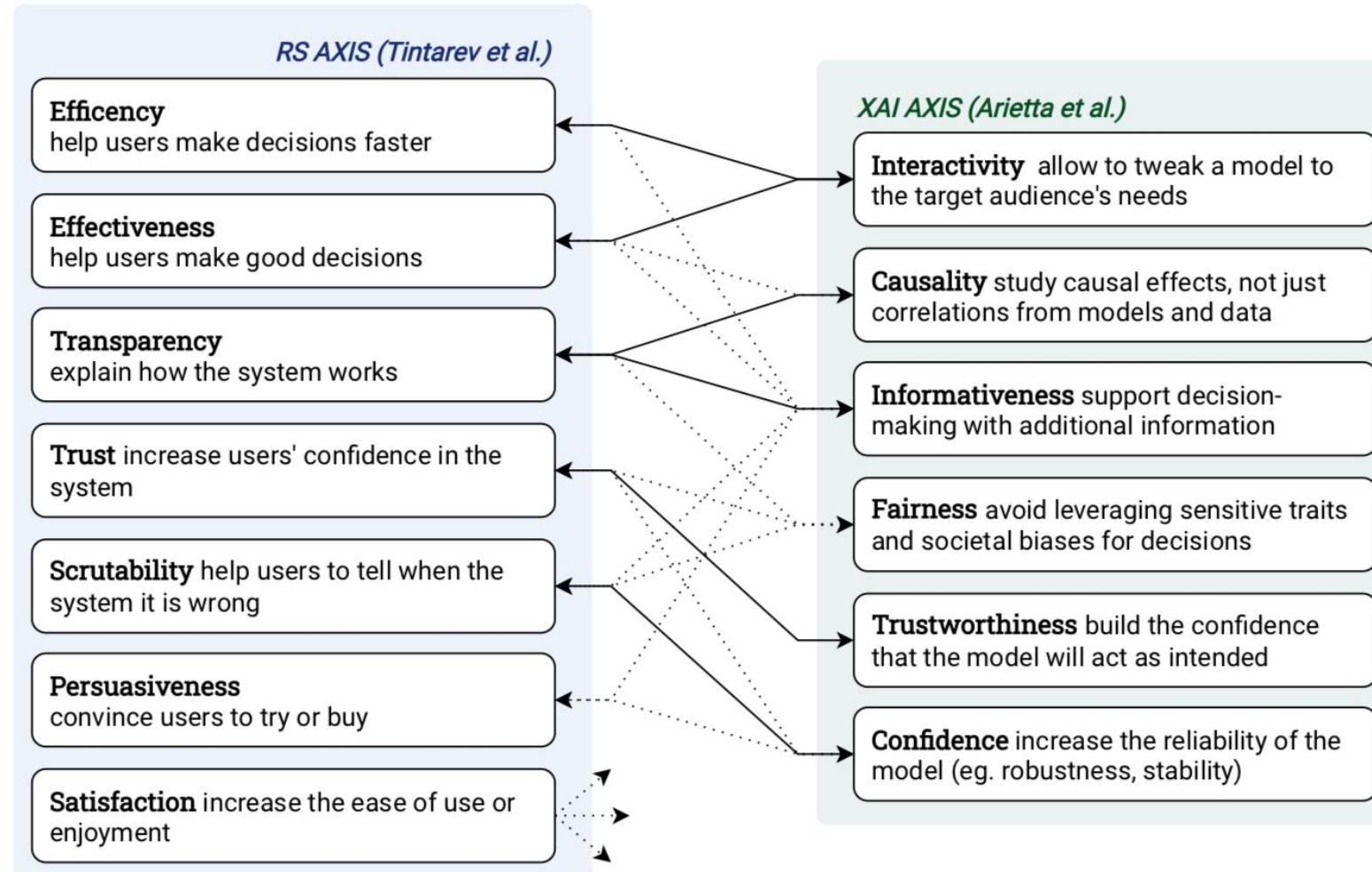
Complementary information

“To make clear by giving a **detailed description**” (Tintarev et al.)

“Explainable recommendation to answer the question of **why**” (Zhang et al.)

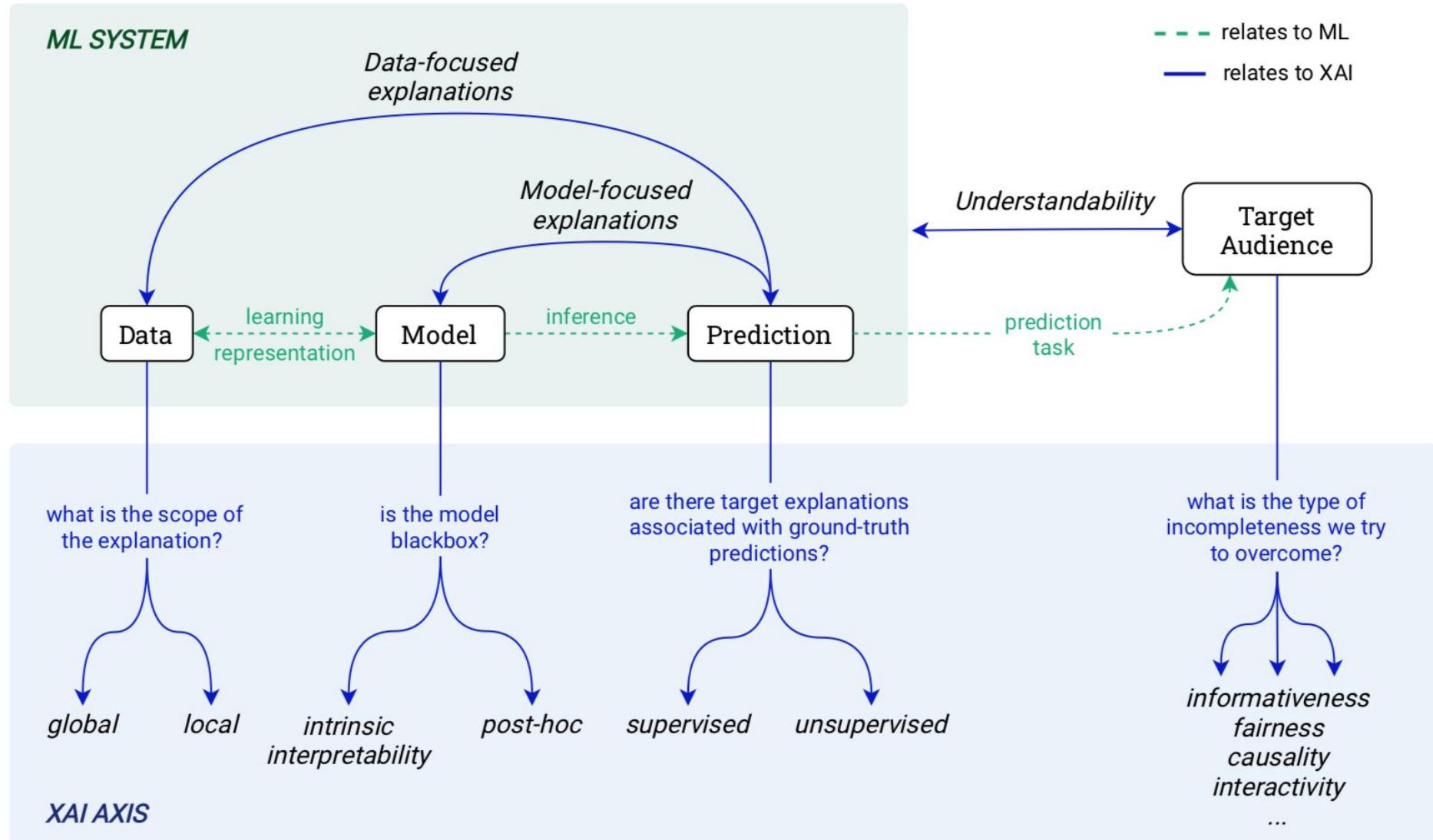
Helps ensure fairness regarding e.g. protected attributes. However: how to act upon them?

# Explainability: Link to eXplainable AI (XAI)



Afchar, D., Melchiorre, A. B., Schedl, M., Hennequin, R., Epure, E. V., & Moussallam, M. (2022). Explainability in Music Recommender Systems.  
<https://onlinelibrary.wiley.com/doi/full/10.1002/aaa.i.12056> & arXiv preprint arXiv:2201.10528.

# XAI Notions



# Local vs. Global

**Local:** explain model decision for particular user-item pair

Explain single predictions

Customers who viewed this item also viewed



**Global:** explain model logic

Tells us about the average behavior of the model

Helps detect systematic biases of the model

# Intrinsic vs. Post-hoc

**Intrinsic:** interpretability inherent in the model

“White-box models”

Ex.: item kNN model

“We recommend you <artist> because it is similar to <artist(s)>”

**Post-hoc:** apply external technique to create interpretability

Applied for black box models

“We recommend you <artist> because it has <features> that you might like”

# Model vs. Data

**Model:** explaining learned model and parameters

Can lead to adjustments and regularization, e.g. to balance fairness and accuracy

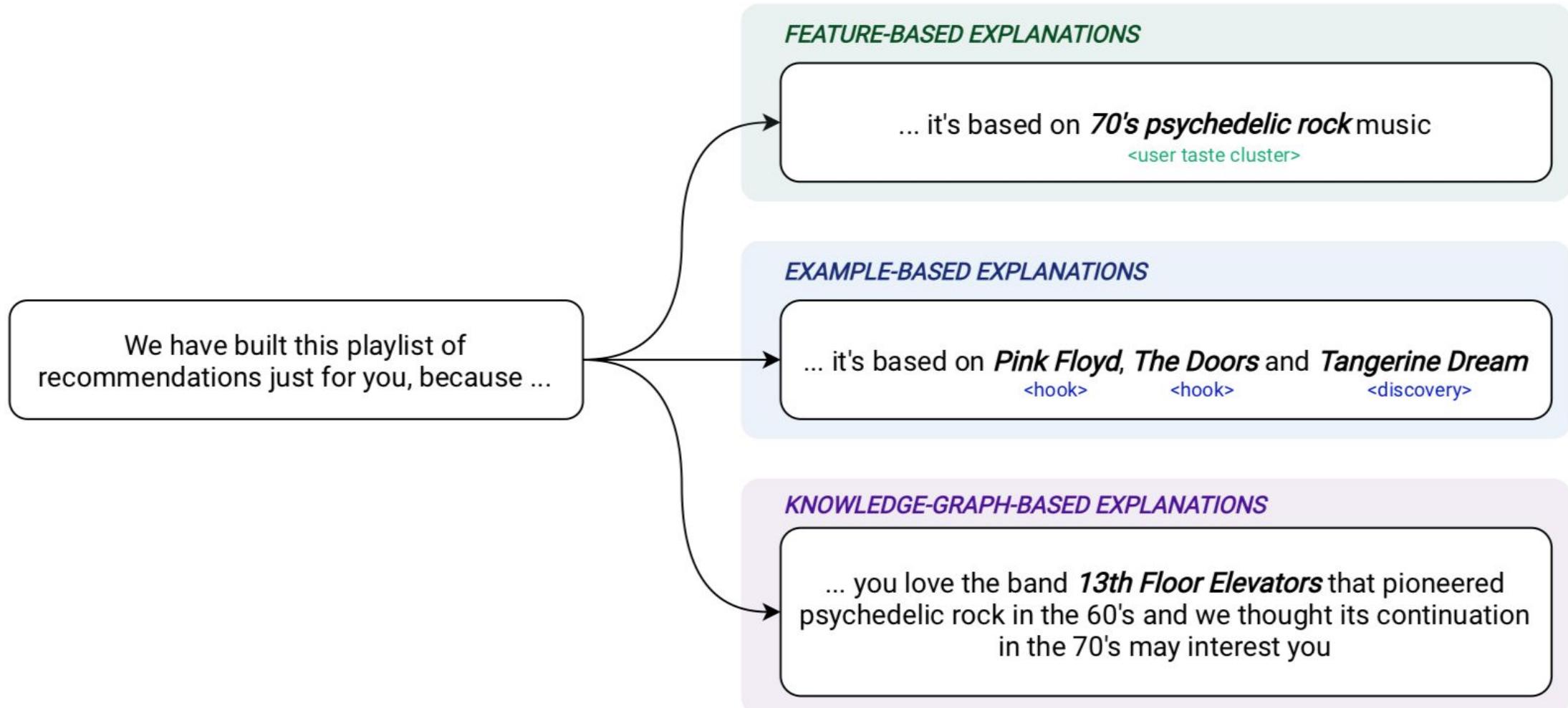
“The has recommended you the item because it maximizes the probability of being co-listened with your history, considering all other users listening history”

**Data:** explain data characteristics

Helps find irregularities in training data

“why are those items co-listened in the first place?”

# Generating Explanations: Types



# Selected Further Resources

- Afchar, D., Melchiorre, A. B., Schedl, M., Hennequin, R., Epure, E. V., & Moussallam, M. (2022). Explainability in Music Recommender Systems.  
<https://onlinelibrary.wiley.com/doi/full/10.1002/aaai.12056> & arXiv preprint arXiv:2201.10528.
- Yongfeng Zhang and Xu Chen (2020), “Explainable Recommendation: A Survey and New Perspectives”, Foundations and Trends® in Information Retrieval: Vol. 14, No. 1, pp 1–101. DOI: 10.1561/1500000066.
- Tintarev, N., & Masthoff, J. (2022). Beyond explaining single item recommendations. In Recommender Systems Handbook(pp. 711-756). Springer, New York, NY.
- Zhang, Y., Zhang, Y., Zhang, M., & Shah, C. (2019, July). EARS 2019: The 2nd international workshop on explainable recommendation and search. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1438-1440).
- EARS tutorial: <https://sites.google.com/view/ears-tutorial/>

# Outline

- Motivation & EU regulations
- Categories of Transparency
- Explainability
- **Traceability and Auditability**
- Documentation

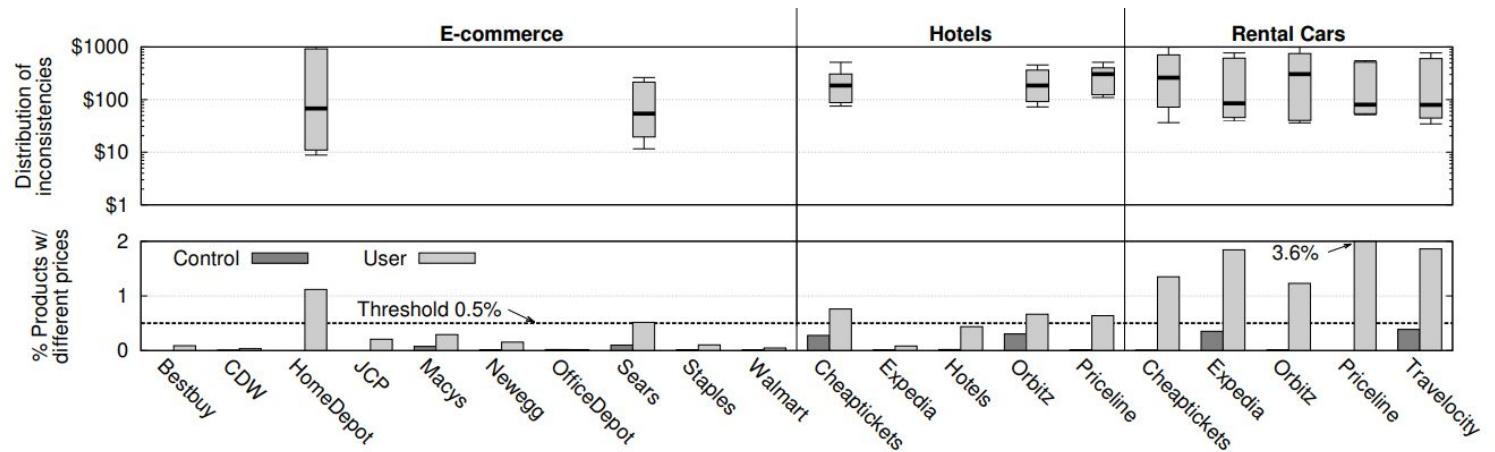
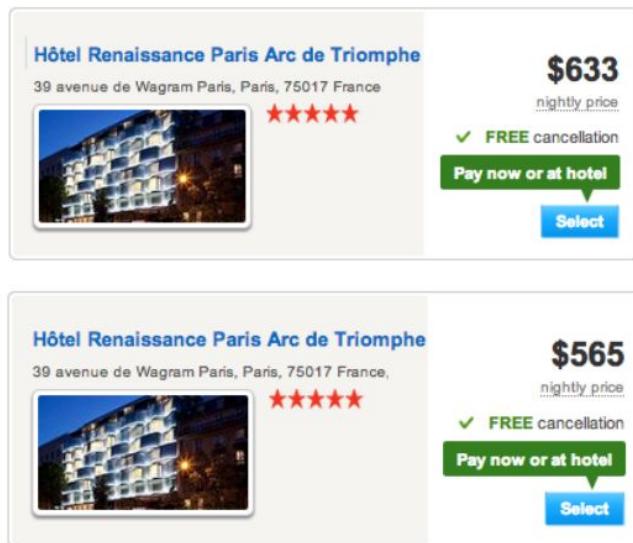
# Algorithm Auditing

- Area receives increased attention in various communities: CSCW, HCI, ML
- Aim: audit algorithms for biased, discriminatory, harmful behavior
  - alignment of systems with laws, regulations, ethics, ...
- Inspired by audits in finance, security, employment,...
- Involves third part external experts:
  - researchers
  - developers
  - policymakers
- Helped uncover bias in search engines, housing, hiring, e-commerce → see  
<https://arxiv.org/pdf/2105.02980.pdf> for cases

# Algorithm Auditing

Audit e-commerce sites for discrimination & price steering (Hannak et al., 2014)

- Web scraping + Amazon MTurk users as testers to audit e-commerce sites



**Figure 3:** Percent of products with inconsistent prices (bottom), and the distribution of price differences for sites with  $\geq 0.5\%$  of products showing differences (top), across all users and searches for each web site. The top plot shows the mean (thick line), 25th and 75th percentile (box), and 5th and 95th percentile (whisker).

**Figure 4:** Example of price discrimination. The top result was served to the AMT user, while the bottom result was served to the comparison and control.

<https://personalization.ccs.neu.edu>

# Types of Algorithm Auditing Methods

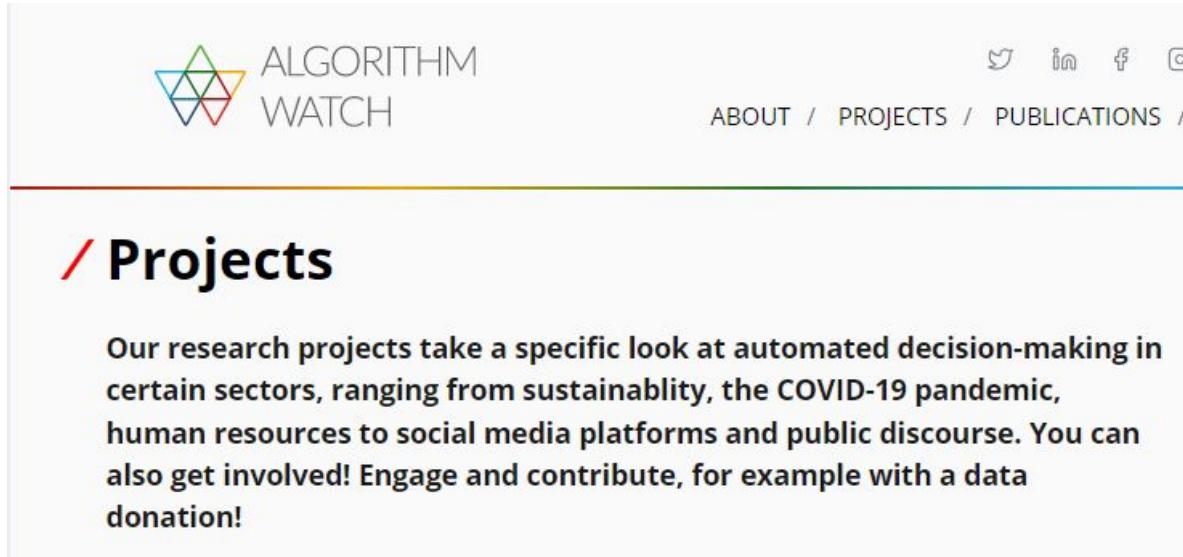
Taxonomy by Sandvig et al.:

- Code audits
  - access to code and system design
- Noninvasive user audits
  - surveys
- Scraping audits
  - send repeated queries to test behavior of system under variety of conditions
- Sock puppet audits
  - researchers generate fake accounts to study system behavior for different user characteristics or patterns of behavior
- Crowdsourced/collaborative audits
  - researchers hire crowdworkers as testers

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (2014).

# Limits of Algorithm Auditing Methods

- Auditing requires technical expertise that might not always be available
  - Frequently: NGOs like AlgorithmWatch doing audits



The screenshot shows the AlgorithmWatch website's homepage. At the top left is the logo, which consists of three overlapping triangles (green, blue, and red) forming a larger triangle, with the text "ALGORITHM WATCH" next to it. At the top right are social media icons for Twitter, LinkedIn, Facebook, and YouTube. Below the header is a navigation bar with links for "ABOUT / PROJECTS / PUBLICATIONS /". A horizontal line separates the header from the main content area. The main content area has a light gray background and features a large, bold, black heading "/ Projects". Below this heading is a paragraph of text: "Our research projects take a specific look at automated decision-making in certain sectors, ranging from sustainability, the COVID-19 pandemic, human resources to social media platforms and public discourse. You can also get involved! Engage and contribute, for example with a data donation!"

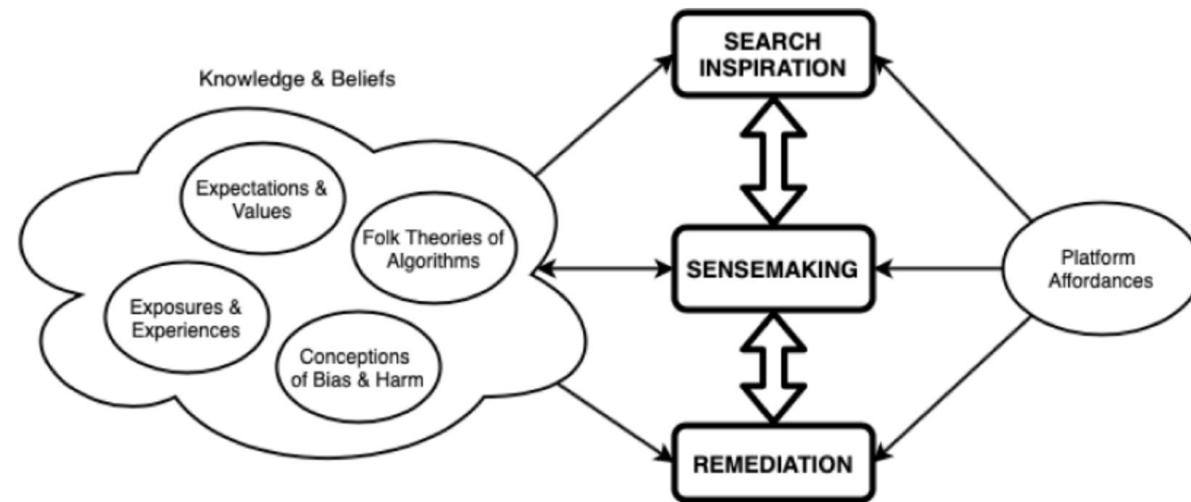
<https://algorithmwatch.org>

# Limits of Algorithm Auditing Methods

- Many harmful algorithmic behaviors are hard to detect outside situated contexts
  - bias happens in specific social / cultural dynamics
  - challenging to anticipate real-world contexts
- Crowdworkers may not represent demographics of investigated system
  - biases might still be undetected
- Expert-driven audits might miss harmful behavior!

# Everyday Algorithm Auditing

- Idea: everyday users detect problematic system behavior via day-to-day interactions with system
- Recent work looked at what strategies users apply in such user-driven audits



# Examples: Everyday Algorithm Auditing

<https://arxiv.org/pdf/2105.02980.pdf>

Domains	Cases	Descriptions
Search	Google Image Search [65]	Researcher Noble searched “black girls” on Google and found out the results were primarily associated with pornography.
Rating/review	<b>Yelp advertising bias [29]</b>	Many small business owners on Yelp came together to investigate Yelp’s potential bias against businesses that do not advertise with Yelp.
	<b>Booking.com quality bias [28]</b>	A group of users on Booking.com scrutinized its rating algorithm after realizing the ratings appeared mismatched with their expectations.
Recommendation systems	YouTube LGBTQ+ demonetization [73]	A group of YouTubers found that the YouTube recommendation algorithm demonetizes LGBTQ+ content, resulting in a huge loss of advertising revenue for LGBTQ+ content creators.
	Google Maps [34]	A group of users reported that when they searched for the N-word on Google Maps, it directed them to the Capitol building, the White House, and Howard University, a historically Black institution. Other users joined the effort and uncovered other errors.
TikTok	recommendation algorithm [54, 82]	A group of users found that TikTok’s “For You Page” algorithm suppresses content created by people of certain social identities, including LGBTQ+ users and people of color. As a result, they worked together to amplify the suppressed content.

# **Outline**

- Motivation & EU regulations
- Categories of Transparency
- Explainability
- Traceability and Auditability
- **Documentation**

# Datasheets for Datasets

- Aim: transparency on datasets used to train and evaluate ML models
  - dataset creation process, possible sources of bias
- Questions: motivation, composition, collection, pre-processing, labeling, intended uses, distribution, and maintenance.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92.

## Datasheets for Datasets

This template contains a set of questions covering the information that a datasheet for a dataset might contain, as well as a workflow for dataset creators to use when answering these questions.

The questions are grouped into seven sections that roughly match the key stages of the dataset creation, maintenance, and distribution process. By grouping the questions in this way, we encourage dataset creators to reflect on the process of creating, distributing, and maintaining datasets, and even to modify this process in response to that reflection. We recommend that dataset creators read through the questions in all sections prior to any data collection so as to flag potential issues early on, and then provide answers to the questions in each section during the relevant stage of the process.

We emphasize that the questions are intended to be used as a starting point for dataset creators to customize. Not all questions will be applicable to all datasets, and dataset creators will likely need to add, revise, or remove questions to better fit their specific circumstances and needs.

To prompt dataset creators to provide sufficient information, all questions are worded so as to discourage yes/no answers. The questions are not intended to serve as a checklist, and dataset creators must be as transparent and forthcoming as possible for datasheets to be useful to dataset consumers.

## Questions

### Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
- **Any other comments?**

### Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
- **How many instances are there in total (of each type, if appropriate)?**

# Model Cards

- Aim: transparent model reporting
  - performance characteristics of trained ML model
- Idea: release model cards in addition to datasets
- Contains:
  - model details, intended use, metrics, training data, evaluation data, ethical considerations

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT\* '19). Association for Computing Machinery, New York, NY, USA, 220–229.  
<https://doi.org/10.1145/3287560.3287596>

## Model Card

- **Model Details.** Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use.** Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors.** Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics.** Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data.** Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data.** May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

**Have you used datasheets / model cards in  
your work or have you created such  
documentation?**

sli.do

#sigir22ethics

Join at

[slido.com](https://slido.com)

**#sigir22ethics**



# **Open Challenges**

# **Open Challenges (Bias and Fairness)**

- Which technological foundation do we need to debias state-of-the-art IR and RS algorithms?
- How should requirements and aims of various stakeholders be accounted for?
- Do computational bias metrics really capture how users perceive fairness?
- What are economic and social consequences of biases resulting from IR and RS technology adopted in high-risk areas (e.g., in recruitment, healthcare)?
- What are the legal implications of unfair or intransparent algorithms?

# **Open Challenges (Diversity & Social Impact)**

- How to collectively set targets and indicators for diversity?
- Which methodologies should be put in place to assess the short-term and long-term social impact of algorithms, to be able to maximize opportunities while avoiding risks?
- Which data may researchers need from real-world scenarios where IR and RS algorithms are developed to carry out a multi-perspective evaluation, e.g. including fairness, diversity, transparency or impact?

# **Open Challenges (Transparency)**

- What **level of transparency** is useful for the needs of different stakeholders and how can transparency be **adjusted** depending on varying needs?
- What is the relation between **explanations** and **perceived fairness**?
- What are effective explanation types for **different** retrieval and recommendation **domains**?
- What do explanations **tell us about the user**? What ethical and privacy implications can arise?

# **Thank You!**

## **Markus Schedl**

Johannes Kepler University Linz, Austria  
Linz Institute of Technology, Austria  
[markus.schedl@jku.at](mailto:markus.schedl@jku.at) | [www.mschedl.eu](http://www.mschedl.eu)

## **Emilia Gómez**

Joint Research Centre, European Commission, Spain  
Universitat Pompeu Fabra, Spain  
[emilia.gomez-gutierrez@ec.europa.eu](mailto:emilia.gomez-gutierrez@ec.europa.eu) | <https://emiliagomez.com>

## **Elisabeth Lex**

Graz University of Technology, Austria  
[elisabeth.lex@tugraz.at](mailto:elisabeth.lex@tugraz.at) | <https://elisabethlex.info>