

# Trustworthy User Modeling and Recommendation: Technical and Regulatory Perspectives

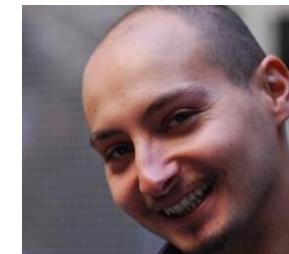
## Markus Schedl

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



## Vito Walter Anelli

Politecnico di Bari, Italy  
[vitowalter.anelli@poliba.it](mailto:vitowalter.anelli@poliba.it) | <https://sisinflab.poliba.it/people/vito-walter-anelli/> |  
@walteranelli



Politecnico  
di Bari

## Elisabeth Lex

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

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

# About Vito Walter Anelli



Politecnico  
di Bari



- Assistant Professor, SisinfLab Research Lab, Politecnico di Bari, Italy.  
<https://sisinflab.poliba.it/>
- Bsc/MSc in Computer Engineering, PhD in Electrical and Information Engineering.
- **Interests:** Knowledge-aware Recommender Systems, Evaluation, Privacy and Security for Recommender Systems, Fairness.

Contact: [vitowalter.anelli@poliba.it](mailto:vitowalter.anelli@poliba.it) |

<https://sisinflab.poliba.it/people/vito-walter-anelli/> | @walteranelli

# 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, natural language processing, computational social science

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

# **Overview (90 Mins.)**

## **1. Introduction - 20'**

Background, motivation, objectives, ethics guidelines for Trustworthy AI, Regulatory efforts

## **2. Fairness and Non-discrimination - 20'**

Categories of bias and fairness, relation to non-discrimination, definition and measurement of bias and fairness, algorithms to mitigate biases and improve fairness

## **3. Privacy and Security - 20'**

Privacy risks for recommender systems, privacy-preserving techniques, security risks for machine learning and RSs, attacks and defenses

## **4. Transparency and Explainability - 20'**

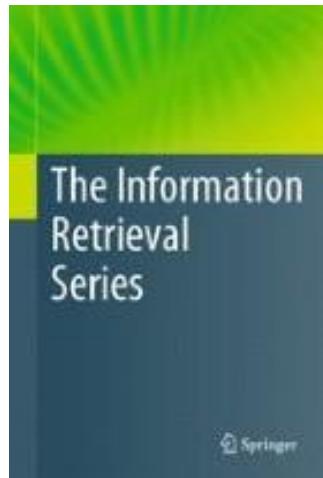
Explainability, justification and interpretability, explainability in UM and RecSys, algorithmic auditing

## **5. Open Challenges and Questions - 10'**

# Related Surveys

- Michael D. Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz: **Fairness in Information Access Systems**. Foundations and Trends in Information Retrieval 16(1-2): 1-177 (2022)
- Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, Shaoping Ma: **A Survey on the Fairness of Recommender Systems**. ACM Transactions on Information Systems 41(3): 52:1-52:43 (2023)
- Deldjoo, Y., Noia, T. D., & Merra, F. A. (2021). **A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial networks**. ACM Computing Surveys (CSUR), 54(2), 1-38.
- Darius Afchar, Alessandro B. Melchiorre, Markus Schedl, Romain Hennequin, Elena V. Epure, Manuel Moussallam: **Explainability in Music Recommender Systems**. AI Magazine 43(2): 190-208 (2022)
- Yongfeng Zhang, Xu Chen: **Explainable Recommendation: A Survey and New Perspectives**. Foundations and Trends in Information Retrieval 14(1): 1-101 (2020)
- Wenqi Fan et al.: **A Comprehensive Survey on Trustworthy Recommender Systems**. ACM Transactions on Recommender Systems (TORS), Just accepted.

# Related Book



Markus Schedl, Vito Walter Anelli, Elisabeth Lex: **Information Retrieval and Recommender Systems: Technical, Ethical, and Regulatory Perspectives**. Springer, to appear towards the end of 2024.

# **Part 1: Introduction**

# **Individual and Societal Impact of Recommender Systems**

- RecSys relying on advanced user models are ubiquitous nowadays, integrated in many online services and platforms
- Enable, control, and limit access to information, products, jobs, opportunities, etc.
- Influence individual human behavior
- Affect and are affected by various stakeholders (content consumers, content creators, platform/service providers, businesses, policy makers, etc.)
- Transformed from information access and decision-support systems into socio-technical systems  
→ **Raises many ethical questions**



dressipi

NETFLIX



Instagram



# A Paradigm Change

## System-centered



- *Focus on algorithmic performance (often accuracy)*
- *Single-stakeholder*
- *Technical perspective*
- *Data-driven: extraction and use of meaningful information from large amounts of data*

## Human-centered

- *Focus on benefit for the user and society, often beyond mere accuracy metrics*
- *Multi-stakeholder*
- *Multi-disciplinary (sociology, economics, ethics, psychology, law, policymaking, ...)*
- *Trustworthy*

# **Objectives of the Tutorial**

- Providing an introduction to **regulatory efforts to Trustworthy AI**, and their implications on RecSys
  - Raising awareness of **social and ethical implications** of UM and RecSys R&D
  - Providing **interdisciplinary perspectives** (technical, legal, ethical, regulatory) on some of the most important trustworthiness dimensions related to UM/RecSys: fairness/non-discrimination, privacy/security, transparency/explainability
- 
- Engage in interesting **discussions** (→ Slido, Zoom, and direct)

# **Ethical Guidelines and Regulatory Efforts**

# Why Regulations?

- Essential to address significant ethical, social & economic implications of RecSys
- How? Via e.g.:
  - establishing guidelines to prioritize well-being and promoting a healthier online ecosystem
  - mandating fairness assessments, transparency, and accountability to mitigate algorithmic biases
  - requiring (large) companies to disclose information about their RecSys and providing means for users to understand and control recommendations
  - implementing accuracy checks & content verification to help fighting the spread of disinformation
  - setting data protection and consent standards to safeguard user privacy

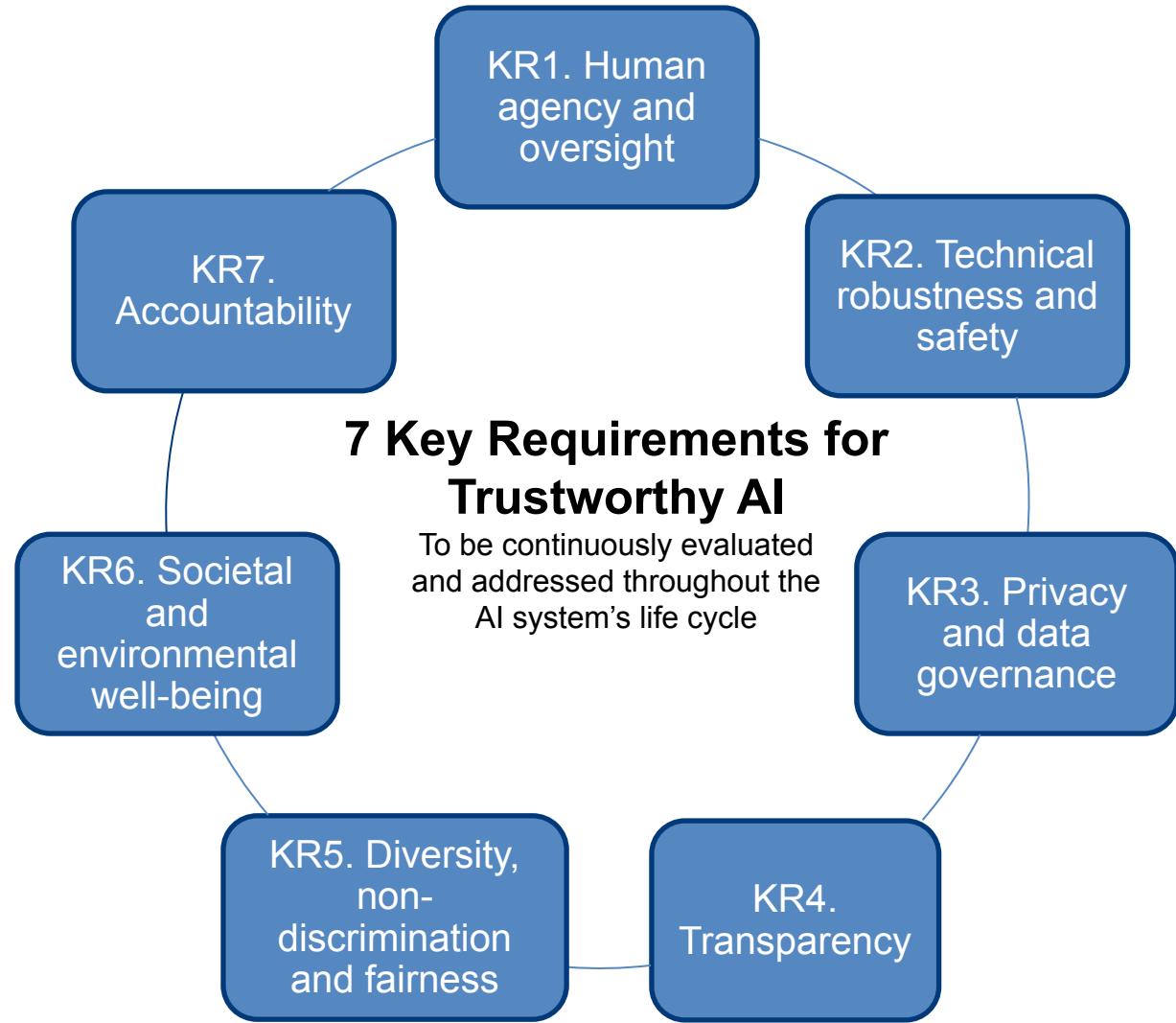
# European Regulatory Efforts

- EU Ethical Principles for Trustworthy AI  
(<https://op.europa.eu/s/pXjd>) - 2019
- EU Regulatory Framework Proposal on AI - 2021  
(<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>)

"The proposed rules will:

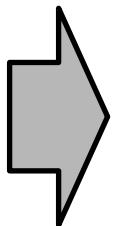
- address risks specifically created by AI applications;
- propose a list of high-risk applications;
- set clear requirements for AI systems for high risk applications;
- define specific obligations for AI users and providers of high risk applications;
- propose a conformity assessment before the AI system is put into service or placed on the market;
- propose enforcement after such an AI system is placed in the market;
- propose a governance structure at European and national level."



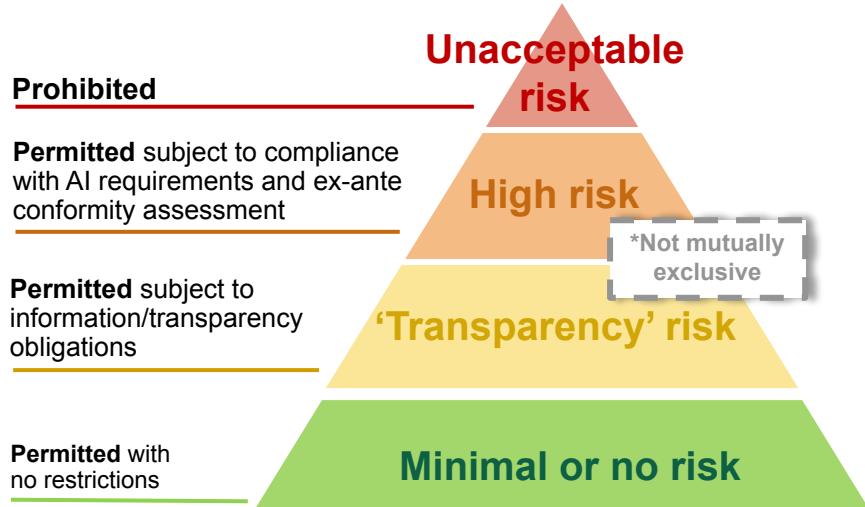


# Ethical Guidelines translated into Legal Req.

- KR1. Human agency and oversight
- KR2. Technical robustness and safety
- KR3. Privacy and data governance
- KR4. Transparency
- KR5. Diversity, non-discrimination and fairness
- KR6. Societal and environmental well-being
- KR7. Accountability



AI Act - Scope: AI systems (software products)



<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

Legal requirements:

- Robustness, accuracy and cybersecurity.
- Human oversight (measures built into the system and/or to be implemented by users).
- Ensure appropriate degree of transparency and provide users with information on system use, capabilities and limitations
- Technical documentation and logging capabilities (traceability and auditability).
- High-quality and representative training, validation and testing data.
- Risk management.

AI Act: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

# Digital Services Act

Scope: **digital services (e.g. recommender engines, online platforms)**: applies to very large online platforms (> 45 million monthly users in the EU)

- Transparency of recommender systems, online advertisement
- External & independent auditing, internal compliance function and public accountability
- Data sharing with authorities and researchers
- Crisis response cooperation

## Digital Services Act:

[https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment\\_en](https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en)

<https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>



# **Some Actions from Very Large Online Platforms**

- <https://blog.google/around-the-globe/google-europe/complying-with-the-digital-services-act/>
- <https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram/>
- [https://help.snapchat.com/hc/en-us/articles/17338132910484-Personalisation-on-Snapchat?  
ga=2.55027560.2100881955.1692971960-1974149196.1692971960](https://help.snapchat.com/hc/en-us/articles/17338132910484-Personalisation-on-Snapchat?ga=2.55027560.2100881955.1692971960-1974149196.1692971960)

Read more: <https://www.theverge.com/23845672/eu-digital-services-act-explained>

# China's AI Regulation Efforts

- Internet Information Service Algorithmic Recommendation Management Provisions  
(<https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/>)

**Article 4:** The provision of algorithmic recommendation services shall abide by laws and regulations, observe social morality and ethics, abide by commercial ethics and professional ethics, and respect the principles of fairness and justice, openness and transparency, science and reason, and sincerity and trustworthiness.

**Article 6:** Algorithmic recommendation service providers shall uphold mainstream value orientations, optimize algorithmic recommendation service mechanisms, vigorously disseminate positive energy, and advance the use of algorithms upwards and in the direction of good.

**Article 7:** Algorithmic recommendation service providers shall: fulfil their primary responsibility for algorithmic security, establish and complete management systems and technical measures for algorithmic mechanism examination and verification, technology ethics review, user registration, information dissemination examination and verification, security assessment and monitoring, security incident response and handling, data security and personal information protection, countering telecommunications and online fraud, etc.; formulate and disclose algorithmic recommendation service-related norms; and allocate specialized personnel and

**Article 8:** Algorithmic recommendation service providers shall regularly examine, verify, assess, and check algorithmic mechanisms, models, data, and application outcomes, etc., and may not set up algorithmic models that violate laws and regulations or ethics and morals, such as by leading users to addiction or excessive consumption.

**Article 21:** Where algorithmic recommendation service providers sell products or provide services to consumers, they shall protect consumers' fair trading rights, they may not use algorithms to commit acts of extending unreasonably differentiated treatment in trading conditions such as trading prices, etc., and other such unlawful activities, on the basis of consumers' tendencies, trading habits and other such characteristics.

<https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-ge-t-made-pub-90117>

# Differences between EU and China

Quick comparison: EU vs. China proposals on governing “recommender algorithms”



Common to both



- Respect for fundamental rights
- Obligation to provide access to vetted researchers for independent assessment of algorithms
- Specific obligations around the treatment of notices from users, including “trusted flaggers”, on potential illegal content
- Additional obligations for “very large online platforms”
- Specific obligations to deal with crisis situations affecting public health and security

- Transparency to individuals on the ads/recommendations they see
- Individual’s right to stop/modify the kind of recommendations they see
- Obligation to identify and dispose of illegal content
- Obligation to cooperate with relevant state authorities
- Obligation to conduct regular algorithm risk assessment and audit
- Obligation to designate points of contact/ legal reps; channels for complaint handling and redress

- **Obligation to not promote obsessive behaviors, excessive spending, behaviors that violate public order and morality**
- **Obligation to not endanger national security or social order, and to promote mainstream values/ spread positive energy**
- **Obligations for algorithms that can influence public opinion/ mobilise masses**
- **Obligation to not discriminate on pricing based on user preferences and habits**
- **Requirement to not set up fake accounts or falsely influence rankings/search results**
- **Obligation to safeguard worker interests where algorithm is used for managing gig economy workers**

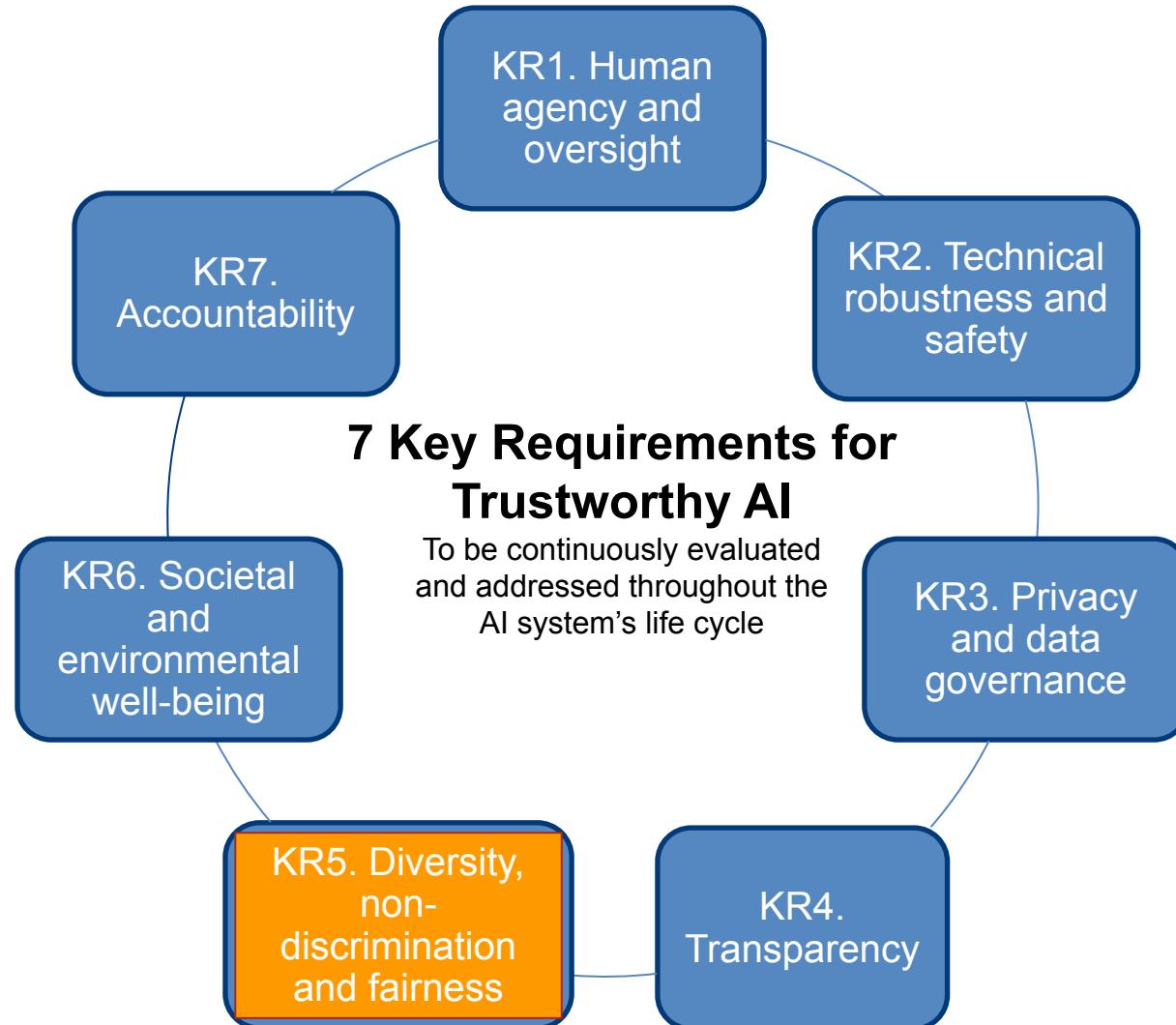
# **US Initiatives**

- 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>

**Part 2:**

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

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



# **Outline**

- EU guidelines and regulations
- Bias from various perspectives
- Relation to fairness and non-discrimination
- Measuring biases (ex.: demographic and popularity bias)
- Strategies to mitigate bias and improve fairness

# EU Regulations

- *EU AI Act (and Ethics Guidelines for Trustworthy AI)*

<https://artificialintelligenceact.eu>

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-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**

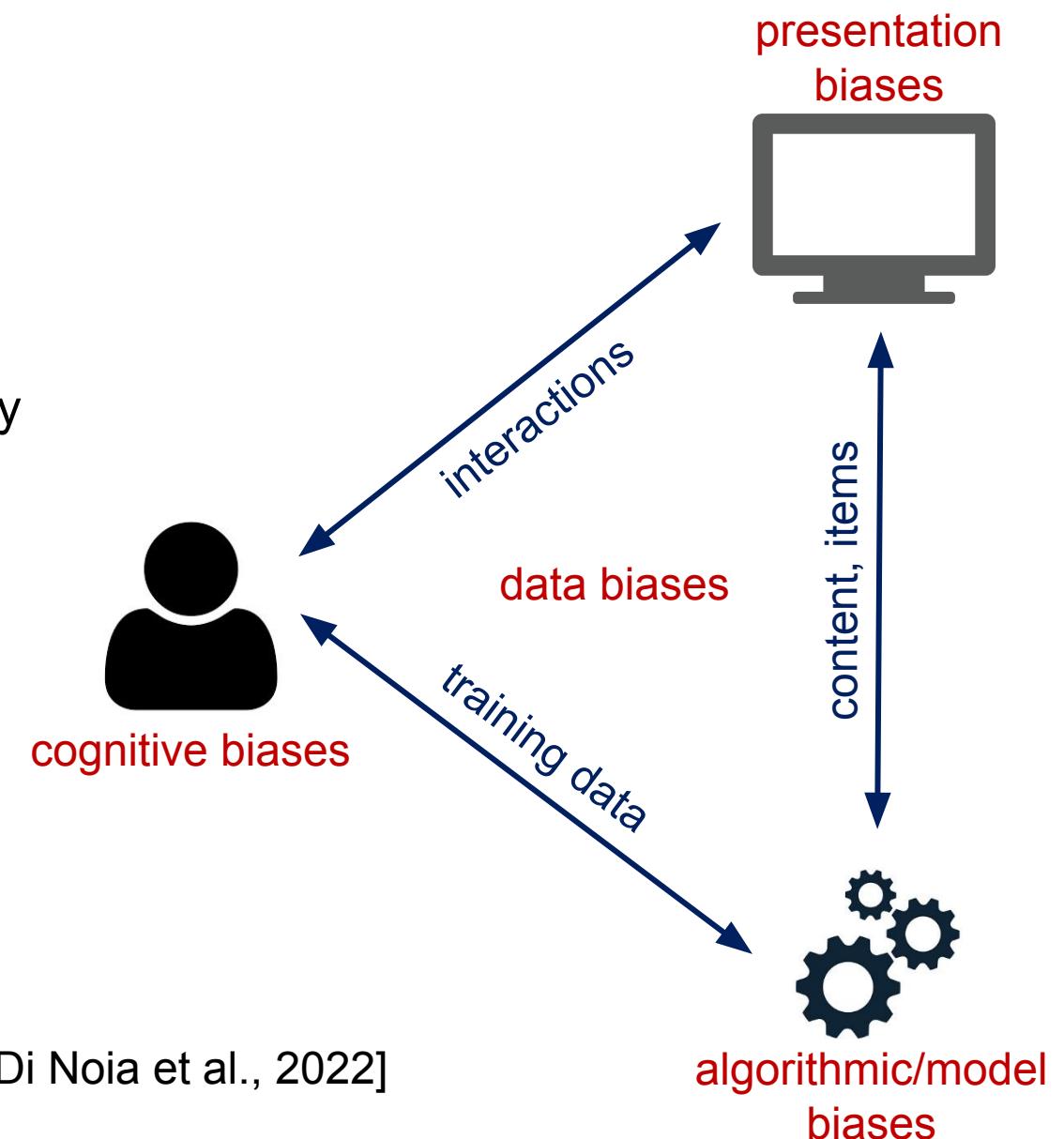
1. Equality between women and men must be ensured in all areas, including employment, work and pay.
2. The principle of equality shall not prevent the maintenance or adoption of measures providing for specific advantages in favour of the under-represented sex.



# Biases in Recommender Systems

Decisions made by 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., position, color, size of recommended items on screen
- *User cognition or perception*: e.g., serial position effect, confirmation bias



# 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]*

- **Popularity bias** → reinforcing already popular items/content, while limiting exposure of less popular ones (harmful for content creators and users)
- **Demographic bias** → disparate recommendation performance between users with different demographic characteristics

# When are Biases Problematic?

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 (e.g. case study popularity bias)

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

→ Finding an optimal level of popularity in recommendation results is tricky!  
(often, *popularity calibration* is aimed for) e.g. [Abdollahpouri et al., 2021; Lesota et al., 2021]

# **Bias Measurement**

# **Demographic and Popularity Bias**

# User Demographic Bias

[Melchiorre et al., 2021]

Assumption: A RS should perform *equally well* across all groups of users, according to a given evaluation metric.

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

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 ( $RecGap^\mu = 0$ ) iff it yields the same quality of recommendations, measured by  $\mu$ , for all groups of users.

# User Demographic Bias: Empirical Results

[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

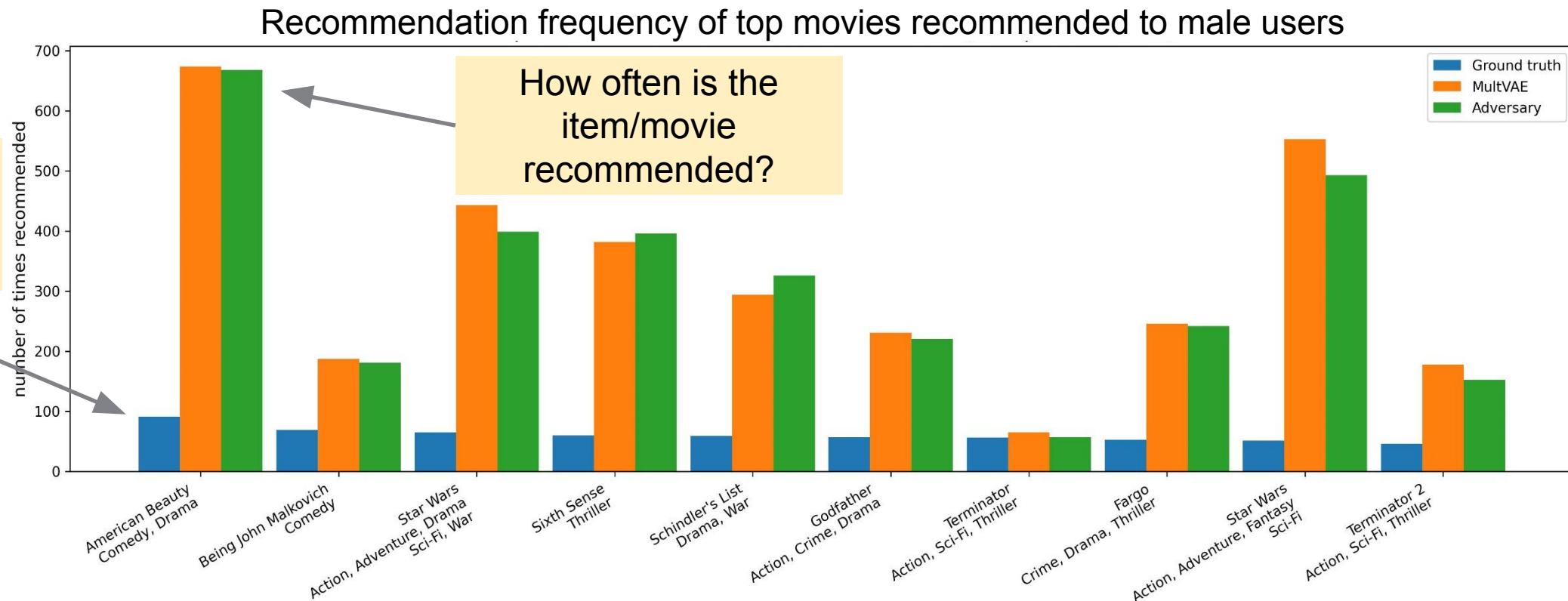
# Popularity Bias: Simple Example

[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: Delta Metrics

[Lesota et al., 2021]

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

- $pop$  some measure of popularity  
(e.g., total number of interactions with items, number of interacting users)
- $H_{u_i}$  historical interaction list of user  $u_i$ 's over items
- $R_{u_i}$  recommendation list created for user  $u_i$  (top recommendations at fixed cut-off)

# Popularity Bias: Delta Metrics

[Lesota et al., 2021]

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

$\% \Delta \xi$  relative popularity difference between items in  $H_{u_i}$  and  $R_{u_i}$  in terms of statistical measure  $\xi$  (e.g., mean, median, variance, skew)

Aggregate over all users (bias of the RS):  $\% \Delta \xi = Median(\% \Delta \xi(u_i))$

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

# 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

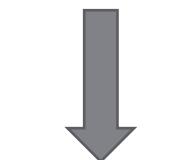
Popularity Bias can be combined with User Demographic Bias

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)

# **Bias Mitigation**

# Strategies to Mitigating Harmful Biases



## Pre-processing strategies

- Data rebalancing (e.g., upsample minority group, subsample majority group)  
e.g. [Melchiorre et al., 2021]

## In-processing strategies

- Regularization (e.g., include bias correction term/bias metric in loss function used to train a model)  
e.g. [Abdollahpouri et al., 2017]
- Adversarial learning (e.g., train a classifier that predicts the sensitive attribute and adapt model parameters to minimize performance of this classifier)  
e.g. [Ganhör et al., 2022]

## Post-processing strategies

- Filter items (e.g., remove items from overrepresented groups)
- Reweigh/Rerank recommendations in list  
e.g. [Ferraro et al., 2021]

# 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, but foremost due to male users' decrease in recommendation quality

# Mitigating Harmful Biases (In-processing Strategy)

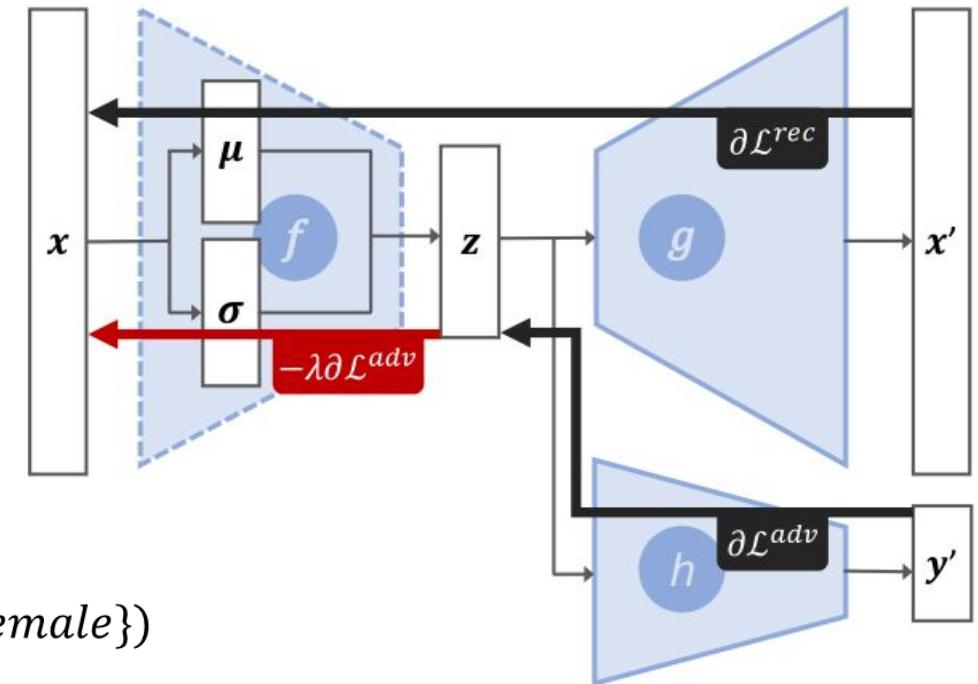
[Ganhör et al., 2022]

## 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 (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%

# 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%

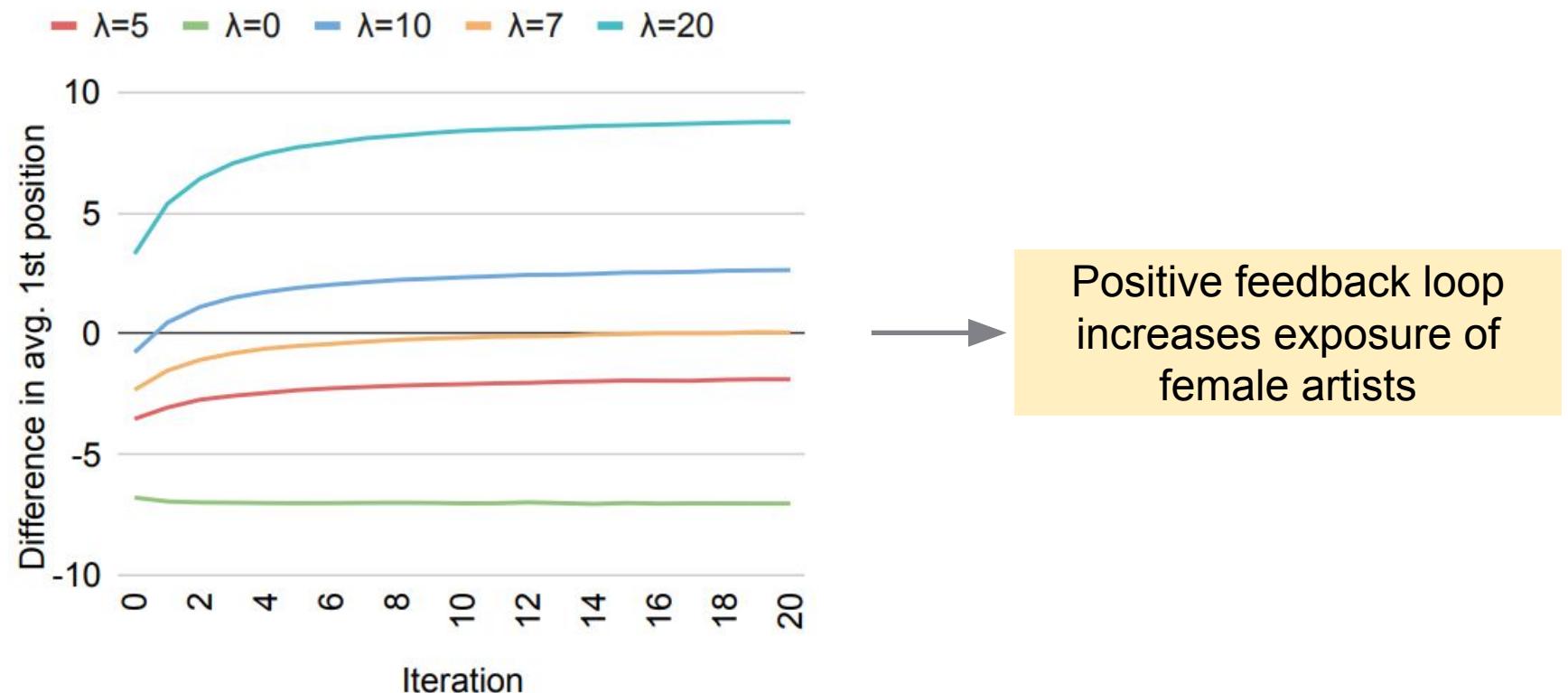
Female artists tend to occur further down in the recommendation lists  
→ position bias

# 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

- Recommender systems 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 or stakeholders differently)
- Most-researched biases include *popularity bias* and *demographic bias*
- Coping strategies include *pre-, in-, and post-processing* techniques
- Many open questions (e.g., perceived bias vs. offline metrics)

[Ferwerda et al., 2023; Lesota et al., 2023; Alves et al., 2024]

# References

- [Abdollahpouri et al., 2017]: *Controlling Popularity Bias in Learning-to-rank Recommendation*. In Proceedings of the 11th ACM Conference on Recommender Systems (RecSys), Como, Italy, 2017.
- [Abdollahpouri et al., 2021]: *User-centered Evaluation of Popularity Bias in Recommender Systems*, Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMA), Utrecht, The Netherlands, 2021.
- [Alves et al., 2024]: User Perception of Fairness-Calibrated Recommendations, Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, Cagliari, Sardinia, Italy, 2024
- [Chen et al., 2023]: *Bias and Debias in Recommender System: A Survey and Future Directions*, ACM Transactions on Information Systems 41(3), 67:1-39, 2023.
- [Di Noia et al. 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.
- [Ferwerda et al., 2023]: *I Don't Care How Popular You Are! Investigating Popularity Bias From a User's Perspective*, Proceedings of the 8th ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR), Austin, TX, USA, March 2023.
- [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), 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), 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), Amsterdam, the Netherlands, 2021.
- [Lesota et al., 2023]: *Computational Versus Perceived Popularity Miscalibration in Recommender Systems*, Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Taipei, Taiwan, 2023.
- [Melchiorre et al., 2020]: *Personality Bias of Music Recommendation Algorithms*, Proceedings of the 14th ACM Conference on Recommender Systems (RecSys), Virtual, 2020.
- [Melchiorre et al., 2021]: *Investigating Gender Fairness of Recommendation Algorithms in the Music Domain*, Information Processing & Management, 58(5), 2021.

# **Part 3: Privacy and Security**

# Motivation

Cost of a data breach 2023:  
Pharmaceutical industry impacts



Forbes

Nov 26, 2019, 11:40am EST | 34,050 views

## Data Privacy Will Be The Most Important Issue In The Next Decade

 **Mary Meehan** Contributor ⓘ  
Consumer Tech  
*I write about consumer insights and foresights to drive innovation.*

 Listen to article 7 minutes 



We now know that Big Data is always watching EV-GPJVRZYAVZC-UNSPASH

Airbus Suffers Data Breach: 3,000 Suppliers Leaked



# Regulations

- European Union
  - General Data Protection Regulation (GDPR) - empowers individuals with a robust set of rights, including the right to access, rectify, erase, and restrict the processing of their personal data. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- United States: complex patchwork of federal and state laws
  - Children's Online Privacy Protection Act (COPPA) - Empowers parents to act as gatekeepers of their children's online data. <https://uscode.house.gov/view.xhtml?path=/prelim@title15/chapter91&edition=prelim>
  - Gramm-Leach-Bliley Act (GLBA) - Financial institutions must provide customers with a clear explanation of what data they collect, how it is shared, and the right to opt out of information sharing with non-affiliated companies for marketing purpose. <https://www.ftc.gov/business-guidance/privacy-security/gramm-leach-bliley-act>
  - Health Insurance Portability and Accountability Act (HIPAA) - by setting privacy standards and by ensuring data security. <https://www.cdc.gov/phlp/publications/topic/hipaa.html>
  - California Consumer Privacy Act (CCPA) - Grants Californians rights similar to the GDPR, such as access, deletion, and opting out of data sales. <https://oag.ca.gov/privacy/ccpa>
  - California Privacy Rights Act (CPRA) - Further expands these rights and creates a dedicated enforcement agency. <https://cpa.ca.gov/regulations>
- China
  - Personal Information Protection Law (PIPL) - Grants individuals rights to access, rectify, erase, and restrict the processing of their personal data. It emphasizes informed consent and restricts the transfer of personal information outside China. [http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820\\_313088.html](http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html)

# Recommender Systems: User-centric Data

Medical records



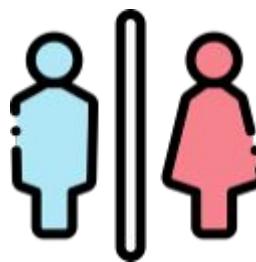
search logs



user preferences



user check-ins



Sensitive attribute (e.g., gender, disease)

# Privacy risks for Recommender Systems

- The **quality of the recommendations is correlated with the amount, richness, and freshness of the underlying user modeling data** (the same factors drive the severity of the privacy risk)
- Direct access to data: **unsolicited data collection, sharing data with third parties, unsolicited access by employees**
- Inference from User Preference Data
  - **Exposure of sensitive information**
  - **Targeted Advertising**
  - **Discrimination**
- Risks Imposed by other System Users
  - In collaborative approaches, users are compared with each other
  - **Create fake profiles to identify other users' preferences**
  - By observing **changes in item-to-item collaborative systems** an attacker may infer preferences of a target user

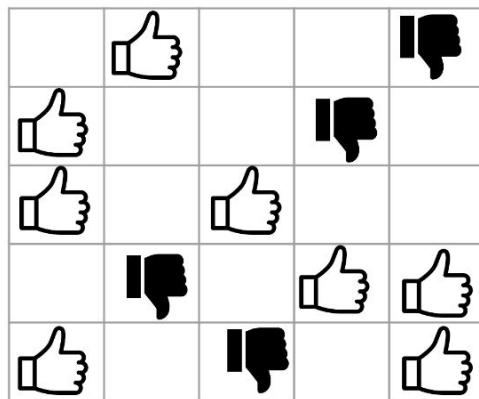
Friedman A., Knijnenburg B.P., Vanhecke K., Martens L., Berkovsky S. (2015) Privacy Aspects of Recommender Systems. In: Ricci F., Rokach L., Shapira B. (eds) Recommender Systems Handbook (2nd edition). Springer, Boston, MA.

# Netflix Prize

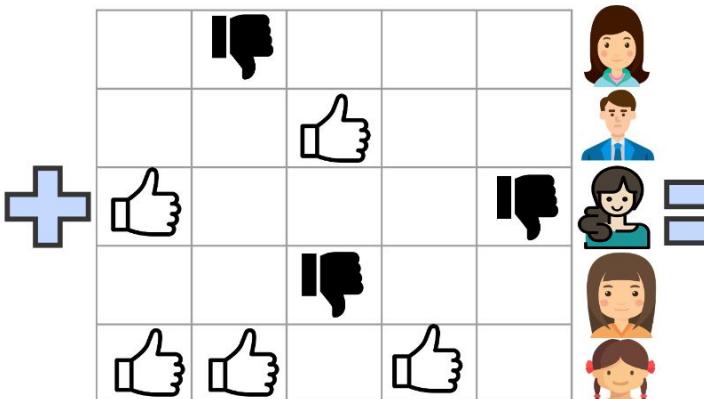
October 2006: Netflix announces Netflix Prize

10% of their users

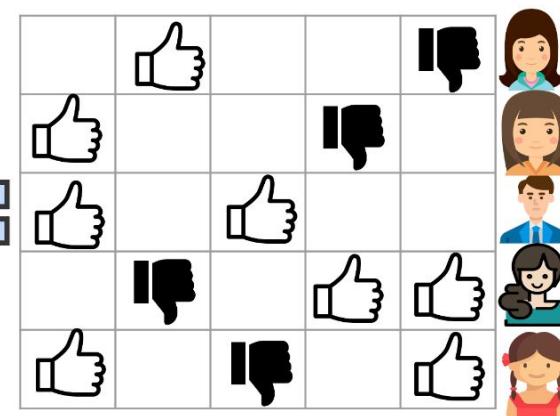
Average 200 ratings per user



 Anonymized  
Netflix data



 Public, incomplete  
IMDB

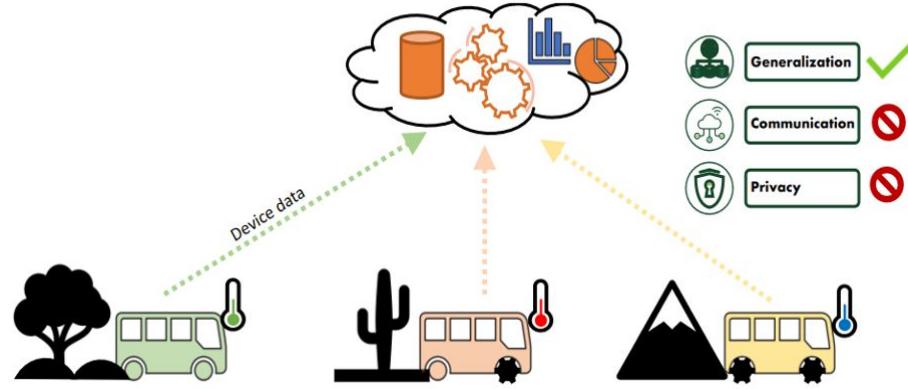


 Identified  
Netflix data

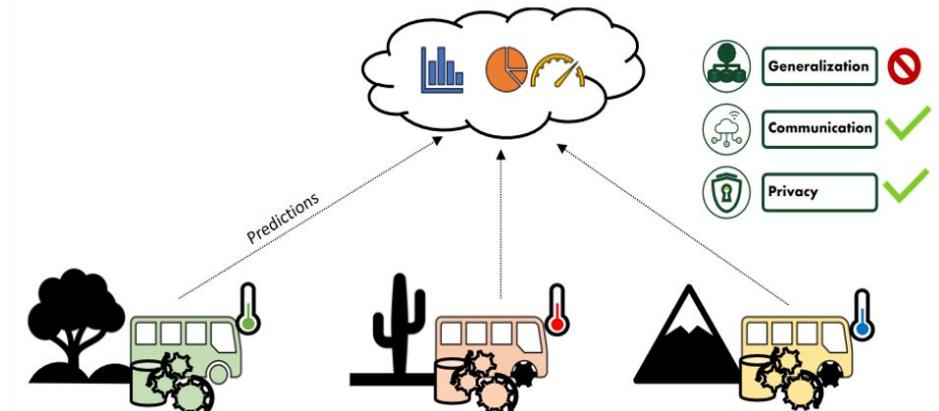
# Privacy-Preserving Machine Learning and Attacks

- What should we protect?
  - Input training **data**;
  - Output **predicted labels**;
  - Model information, including **parameters**, **architecture**, and **loss** function;
  - **Identifiable information**, such as which **site** a record comes from.
- Characteristics of an attack
  - What is the **target**? Data or Model?
  - What is the **knowledge** of the attacker? White- or Black-box?
  - What is the attacking **methodology**? Model extraction or encoding information?

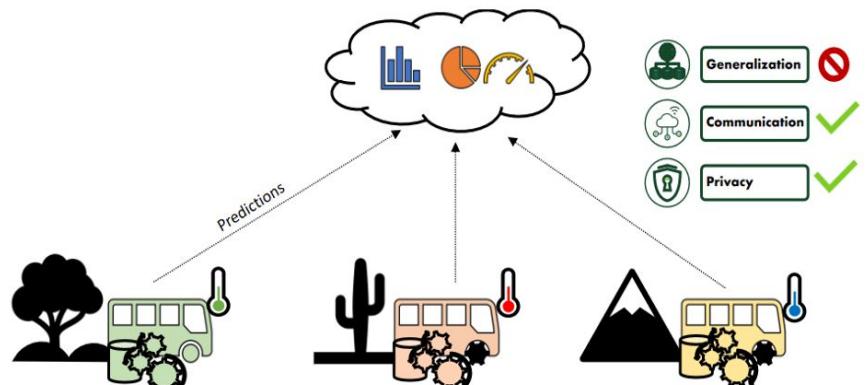
# Learning Paradigms



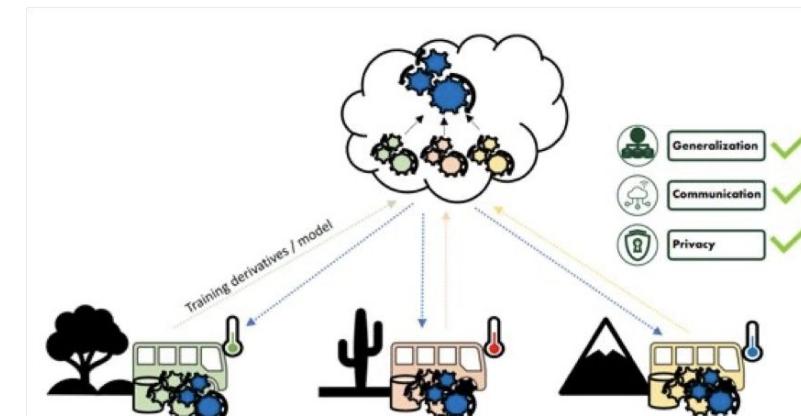
Centralized



Decentralized



Distributed



Federated

# Differential Privacy

- We want to learn nothing about individuals but still learn useful information about a population
- De-identified data is not so secure
  - e.g., Netflix ratings
- Releasing just statistics is still non-private
  - Which is the number of **faculty members** in the university who have **heart disease**?
  - Which is the number of **faculty members** in the university who have **heart disease and are not the President**?
  - **Now we know whether the President has heart disease or not**

# Differential Privacy in practice

Scenario: Suppose we have a small dataset containing information about **people's ages**, and we want to **calculate the average age of the individuals** in the dataset while preserving their privacy using differential privacy.

Person	Age
Alice	25
Bob	32
Carol	28
Dave	36
Eve	30

[Dwork, McSherry, Nissim and Smith, 2006]

# Define Privacy Parameters

DP formula:  $Pr[\mathcal{M}(\mathcal{X}) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(\mathcal{Y}) \in S]$  ( $\mathcal{X}$  and  $\mathcal{Y}$  being datasets differing by one element)

We need to define two key parameters for differential privacy:

- **Epsilon ( $\epsilon$ )**: A measure of **privacy budget** that controls how much noise will be added. **Smaller  $\epsilon$  values provide stronger privacy guarantees** but introduce more distortion.
- **Sensitivity ( $\Delta f$ )**: The **maximum change in the query result** when a single individual's data is added or removed from the dataset. In this case, the sensitivity of calculating the average age is 7 (36 - 25).

To calculate the differentially private average age, we add Laplace noise to the true average. The Laplace noise has a probability density function:

$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

where  $\mu$  is the true average age, and  $b$  is the scale parameter, which is determined by  $\epsilon$  and  $\Delta f$ :  $b = \frac{\Delta f}{\epsilon}$   
Let's say we choose  $\epsilon = 0.5$  (a relatively small privacy budget) for this example. Our goal is to calculate the differentially private average age.

# Calculate the Differentially Private Average Age

1. Calculate the true average age:

$$mu = \frac{25 + 32 + 28 + 36 + 30}{5} = 30.2$$

2. Calculate the scale parameter

$$b = \frac{7}{0.5} = 14$$

3. Generate Laplace noise  $\epsilon_1$  and  $\epsilon_2$  from the Laplace distribution with scale parameter b. These values will be different for each query:

- $\epsilon_1 = -7.92$  (sampled randomly)
- $\epsilon_2 = 12.45$  (sampled randomly)

4. Add the noise to the true average to obtain the differentially private average age:

$$DP\_average\_age = \mu + \epsilon_1 = 30.2 - 7.92 = 22.28$$

This is the differentially private average age that we can release while protecting individual privacy. Note that the released result is noisy and may not exactly match the true average, but it provides a privacy guarantee.

Differential privacy ensures that the presence or absence of any individual in the dataset has a limited impact on the query result, thereby protecting individual privacy while allowing for useful statistical analysis.

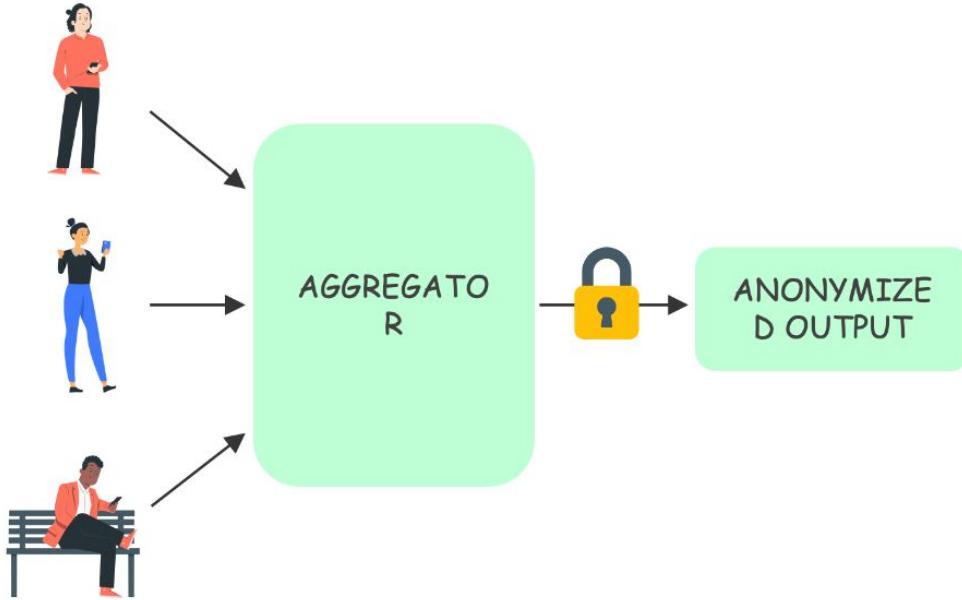
# Differential Privacy in ML and in short

- In ML, just adding noise to the output doesn't work. Non-convex functions are very sensitive (too much noise needed)
- Hard to characterize the dependence of final parameters on data
  - Take the standard algorithm and add appropriate noise during training

Overall, DP:

- Strong privacy guarantees
- No longer needed attack modeling
- Quantifiable privacy loss
- Composable mechanisms
- Useful for analyzing any algorithm

# Differential Privacy - Central vs Local

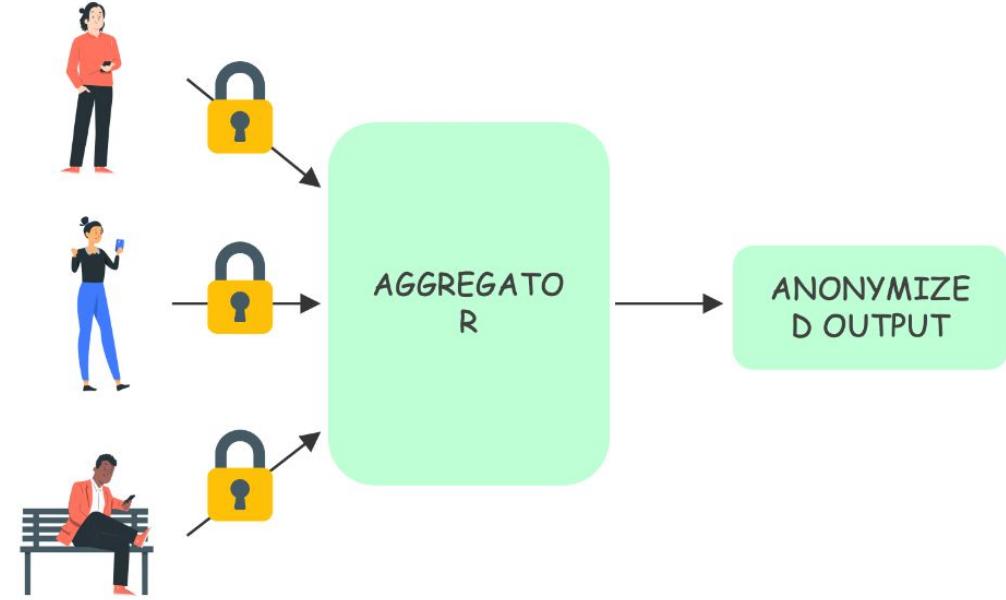


Central Differential Privacy

Higher accuracy

Trusted aggregator

(e.g., US Census)



Local Differential Privacy

Higher noise

No trust required

(e.g., Google RAPPOR, Apple Emojis)

# Secure Multi-Party Computation

Compute a function jointly while keeping the inputs secret

(Additive) Secret Sharing

Distribute random pieces of a secret (shares) among parties

A **secret share** is a piece of incomplete information about the initial secret value

	Antonio	Walter	Tommaso
Antonio (10 beers)	5 beers	3 beers	2 beers
Walter (20 beers)	-8 beers	10 beers	18 beers
Tommaso (30 beers)	0 beers	35 beers	-5 beers
	-3 beers	48 beers	15 beers

Average  
20 beers ;)

# **Homomorphic Encryption and overall privacy trade-off**

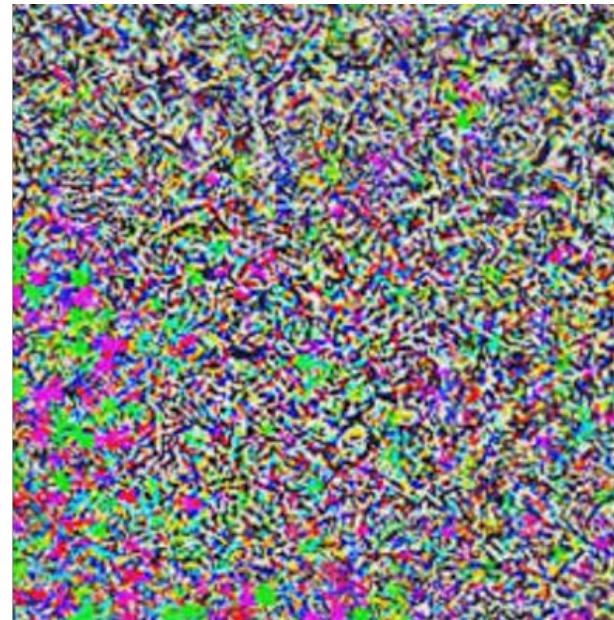
Homomorphic Encryption allows meaningful calculations on encrypted data:

- Only one party needed to encrypt and decrypt own data
- Can perform operations directly on encrypted data (without interactions)
- The result is equivalent to performing analogous operations without encryption!
- Computationally expensive
- Allows a little set of calculations (it varies from Partially HE to Fully HE)

The overall privacy trade-off (developing private ML needs a artful balance of efficiency-accuracy-privacy):

- HE and SMPC are often replaceable
  - HE: Little interaction and expensive computation
  - SMPC: Cheap computation and significant amount of interaction
- SMPC replaces computation with interaction, offering better practical performance
- DP replaces accuracy with efficiency
  - If the coordinator is trusted, send plain data to preserve more accuracy

# **Security risks for Machine Learning and RSs: Are our models actually secure?**



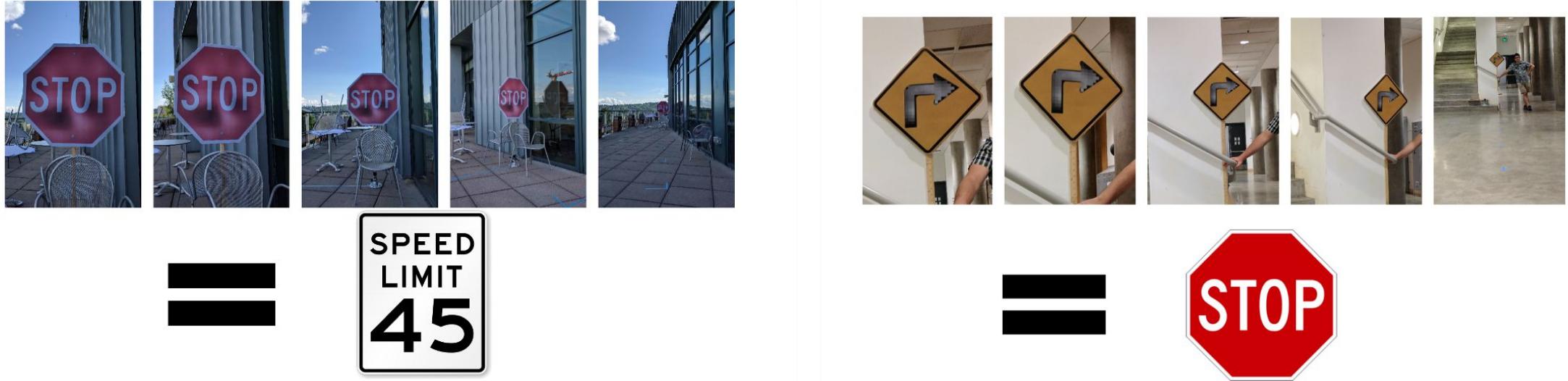
What the model sees is a Panda

.. but with an imperceptible noise

now the model sees a gibbon!

It is cute, isn't it?

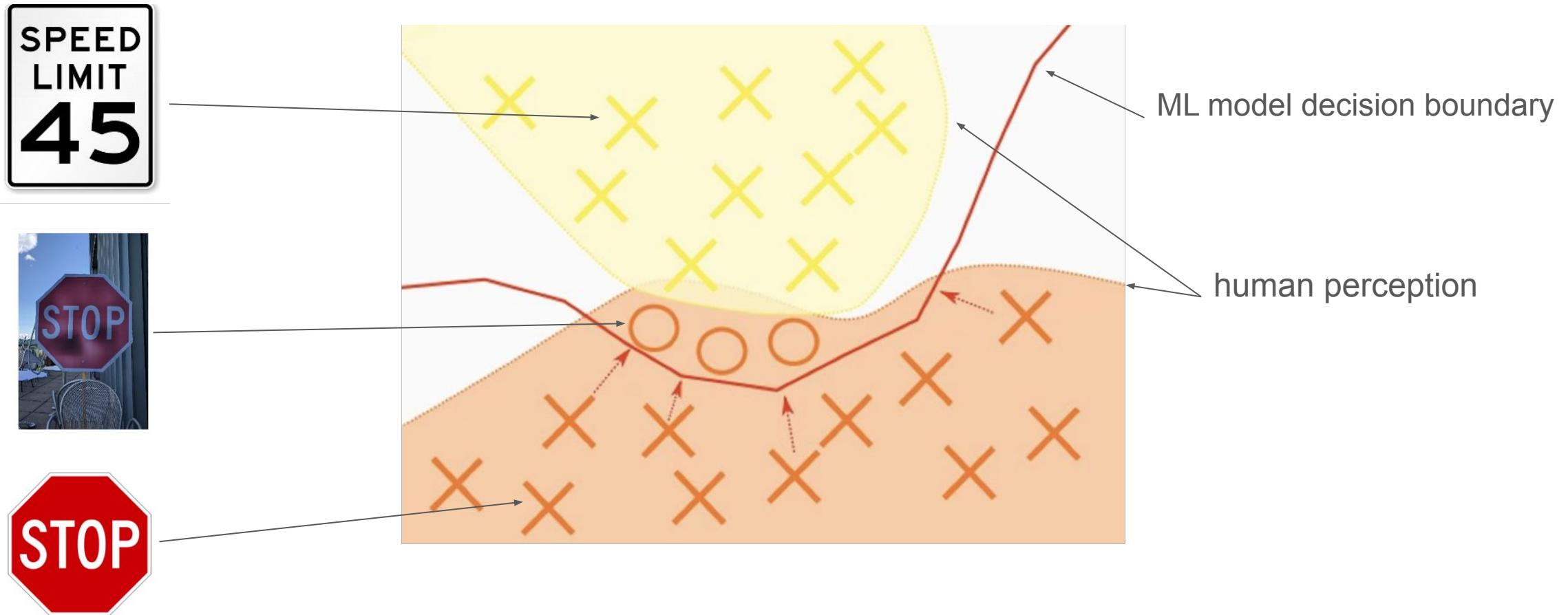
# The “unseen” security risks



Song, Dawn. "AI and Security: Lessons, Challenges & Future Directions", UC Berkeley, 2017

Evtimov, Ivan et. al. "Robust Physical-World Attacks on Machine Learning Models." arXiv preprint arXiv:1707.08945 (2017).

# How is that possible? Tell me Adversarial without telling me



# How can adversarial examples fool the model?

«... **adversarial examples can be directly attributed to the presence of non-robust features**: features (derived from patterns in the data distribution) that are highly predictive, yet brittle and (thus) incomprehensible to humans.»

«Adversarial vulnerability is a direct result of **our models' sensitivity to well-generalizing features** in the data.»

«...this perspective establishes adversarial vulnerability as a human-centric phenomenon, since, from the standard supervised learning point of view, **non-robust features can be as important as robust ones** »

[A. Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features, NIPS'19]

# Security for RSs: Taxonomy of attacks

## Attack's timing:

- Training time (data poisoning): occur **before the ML model is trained** or in the inference phase
- Inference time (evasive attack): occur after the ML model is trained or in the **inference phase and aim to evade detection** — or evade the decisions made by the learned model

## Model-specific attacks:

- **Attack on Binary Classification:** Label flipping attack, Kernel SVM [Biggio, B., Nelson, B., & Laskov, P. (ICML 2012). Poisoning attacks against support vector machines.]
- **Attack on unsupervised learning:** Clustering, Anomaly detection
- **Attack on matrix completion:** Hand Engineered Poisoning (Shilling attack in RecSys), Alternating minimization, Projected gadianet decent (PGA), Nuclear norm normalization, Mimicking user behavior, ML-optimized attacks [Vorobeychik, Y., & Kantarcioglu, M. (2018). Adversarial machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning]

## Attacker's knowledge:

White-box attack, Gray-box attack, Black-box attack

## Attacker's goal:

- Targeted attack: forces the classifier (ML model) to make predictions into a **target class label**.
- Untargeted attack (reliability attack): forces the classifier predictions into **any incorrect class label**.

# From classical ML to adversarial

Supervised learning (classification) problem  $\operatorname{argmin} J(\Theta, x, y)$

$$\operatorname{argmax}_{\Delta \text{adv}} J(\Theta, \overset{\Theta}{x} + \Delta \text{adv}, y) \quad \text{s.t.} \|\Delta \text{adv}\| \leq \epsilon$$

where delta adv is the adversarial perturbation and epsilon is the perturbation budget.

Algorithms that aim to find such adversarial perturbations are referred to as adversarial attacks.

Attacks in computer vision domain:

- L-BFGS [Szegedy et al., 2013] Uses Limited-memory BFGS (L-BFGS) algorithm to solve the problem **with linear memory requirement**
- FSGM - Fast Gradient Sign Method [Goodfellow et al., ICLR '15] Use the **gradient of the loss function** to work on the constraint problem.
- Carlini-Wagner [Carlini and Wagner, 2017a] **Refine the L-BFGS attack** to defeat Defensive distillation
- JSMA - Jacobian Saliency Map Attack [Papernot et al., 2015a] Construct an **input-output mapping** (forward derivatives) **to find the minimal perturbation**
- DeepFool [Moosavi-Dezfooli et al., 2015] Perform an **iterative attack** to find the closest decision boundary to a given input

# Countermeasures

## Proactive countermeasures

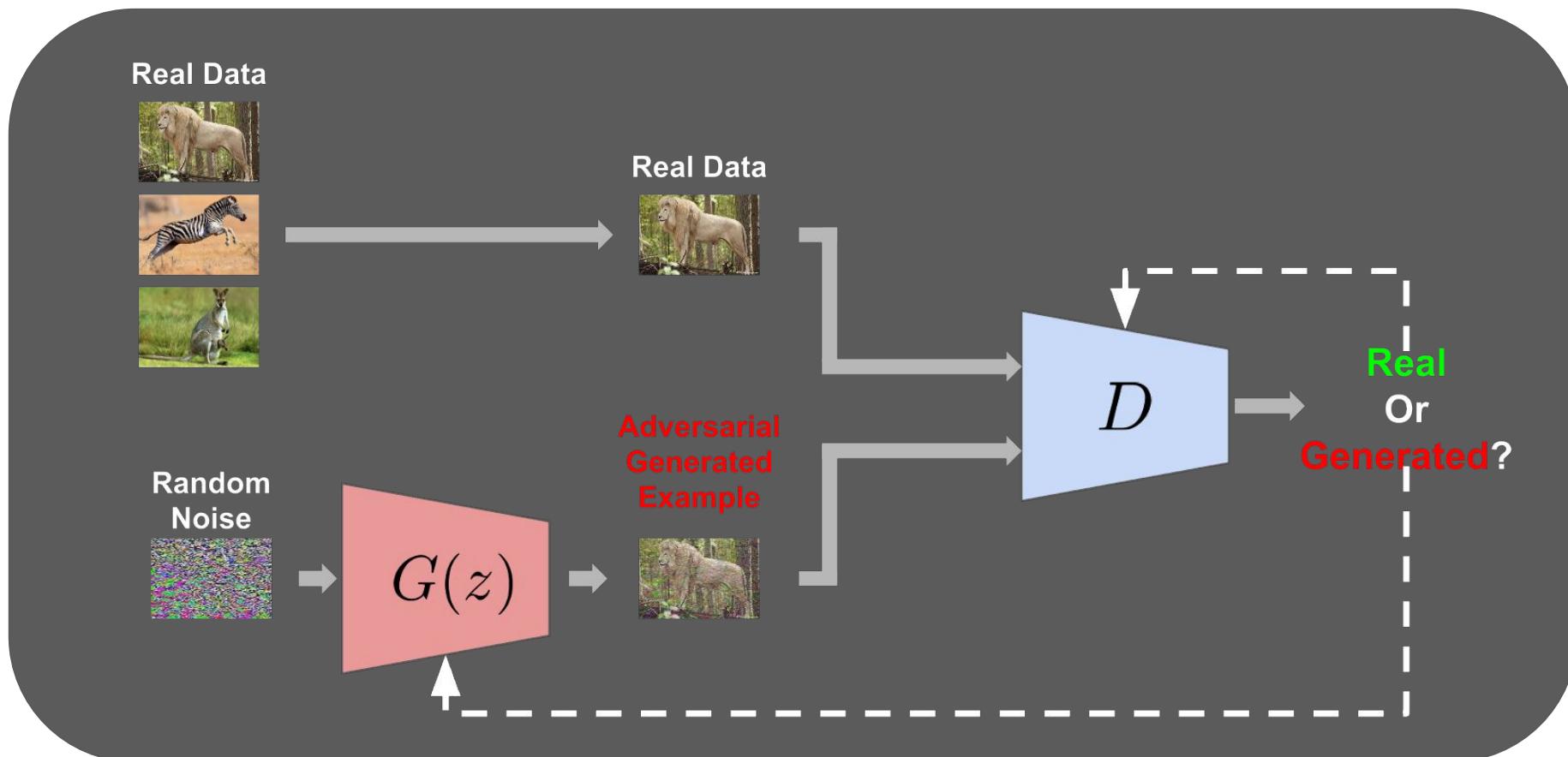
- **Adversarial Training** [Goodfellow et al., ICLR '15]
  - Additional training epochs with adversarial examples
- **Defensive Distillation** [Papernot et al., ISS'16]
  - Adapt distillation to increase the robustness of the network
- **Robust Optimization** [Madry et al., ICLR'18]
  - design robust DNN to prevent a specific class of adversarial examples

## Reactive countermeasures

- Adversarial Detecting
- Input Reconstruction
- Network Verification

# The MINMAX ATTACK-DEFENSE GAME

An effective example are Generative Adversarial Network (GAN) is trained with an ATTACKER that tries to alter a DEFENDER



# Security in Recommender Systems History

Attack Type	Years
<b>Hand-engineered shilling attacks</b> <ul style="list-style-type: none"><li>• Attack by leveraging interaction data</li><li>• Attack by exploiting semantic data</li><li>• Studying the impact of data characteristics on shilling attacks</li><li>• Detection and defense of shilling attack</li></ul>	<b>Early 2000-till now</b>
<b>Machine-learned Data Poisoning Optimization</b> <ul style="list-style-type: none"><li>• Factorization-based models</li><li>• Reinforcement Learning models</li><li>• Other recommendation models</li><li>• Defense</li></ul>	<b>Recently emerging</b>
<b>Adversarial machine-learned attacks</b> <ul style="list-style-type: none"><li>• Adversarial perturbations on model parameters</li><li>• Adversarial perturbation on content data</li><li>• Defense and robustification</li></ul>	<b>2016 till now (in RS field)</b>

# Hand-crafted shilling Attacks

Given a rating matrix with 'n' users and 'm' items, the goal is to add a limited number of fake (malicious) user profile with each profile having maximum 'C' ratings.

Attack such as random, popular, bandwagon, love-hate which are realized by building user profiles.

	3			5	
		4			
			2		4
	1		5	5	
			4	4	
	2				5

# Data Poisoning Optimization attacks

Main limitations of Hand-Engineered attacks:

- **No optimization procedure** to maximize the attacker's utility
- **Empirical techniques**
- **Heuristic**

In data-poisoning attacks against RS, a machine-learned optimization framework is built to learn an optimal user profile composition based on the attacker utility.

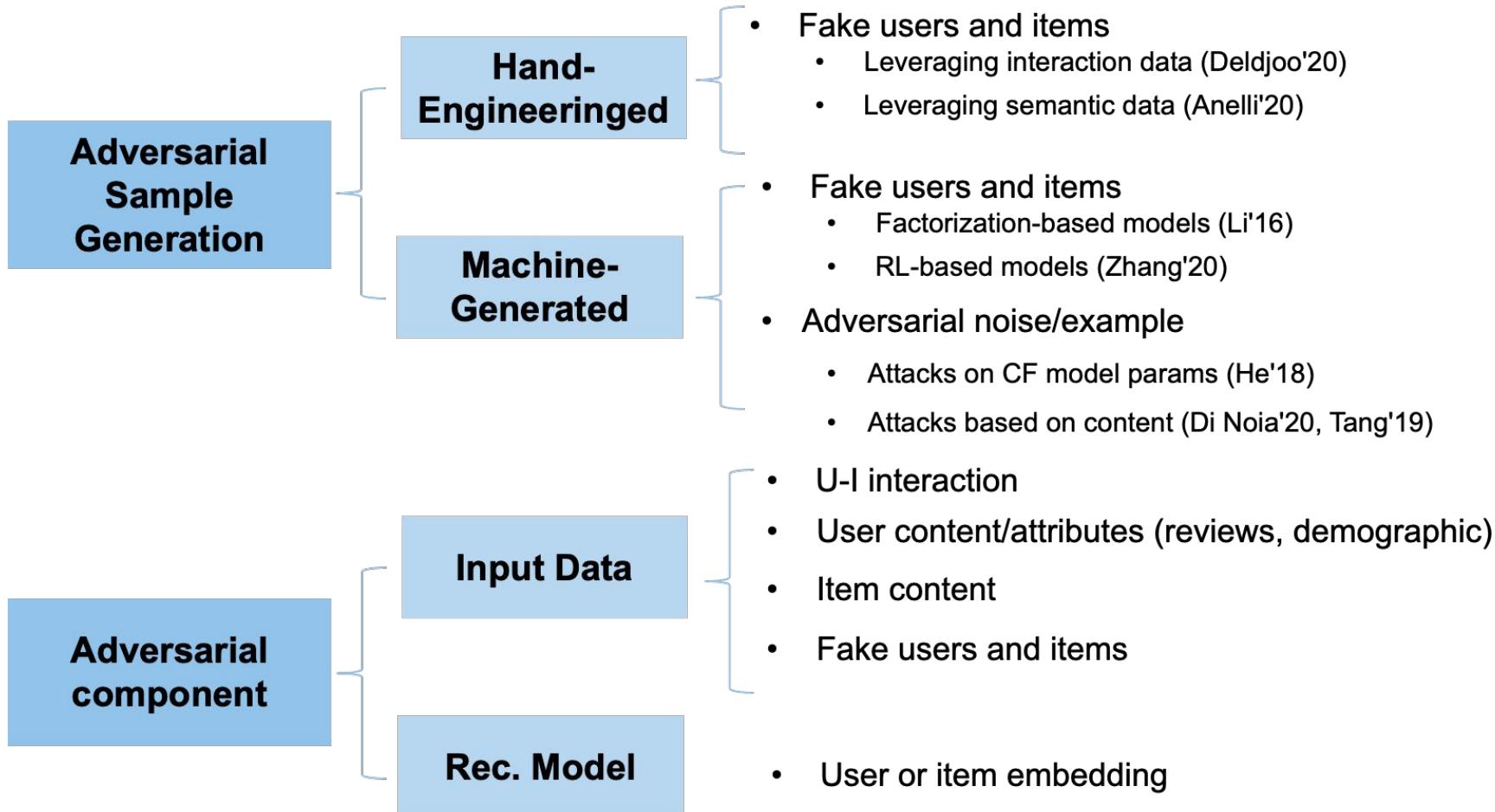
Every model could be optimized for data poisoning attacks. But, **Data Poisoning Optimization is DIFFERENT** from the classic optimization for a recommendation utility

What we need are:

- the **attacker utility**;
- the **target recommendation model**;
- an **optimization procedure** for that utility.

In the literature, we can find Factorization-based models, Reinforcement Learning-based models, Graph-based models, K-NN models, others

# Adversarial attacks to Recommender Systems



# Adversarial Personalized Ranking

Adversarial Perturbation on each embedding vector of user and item.

Adversarial Training used to robustify the model.

NDCG@100

Dataset	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$
BPR-MF	APR -4.7%	APR -12.5%	APR -31.0%
Yelp	-22.1%	-42.7%	-63.8%
Pinterest	-9.5%	-25.1%	-55.7%
Gowalla	-26.3% -2.9%	-53.0% -13.2%	-78.0% -29.2%

[XIANGNAN HE ET AL., SIGIR '18]

With VBPR we have Adversarial Multimedia Recommendation [XIANGNAN HE ET AL., TKDE'19]

# ATTACK Timing

## TRAINING TIME (Poisoning)

Image samples are perturbed and injected in the VRSs before the training.

- TAaMR Targeted Adversarial Attack against Multimedia Recommender Systems [Di Noia et al, 2020]
- VAR [Anelli et al, 2020, Anelli et al. 2021] 24 combinations of attack/defense strategies showing limited efficacy of defenses, and Study of Defensive Methods to Protect Visual Recommendation Against Adversarial Manipulation of Images

## TESTING TIME (Evasion)

Images are perturbed at inference time

- BlackBox-Model [Cohen et al, 2021] pixel-by-pixel perturbation to attack a Specific User (Grey-box settings) by segmenting Users with Similar Taste
- Adversarial Item Promotion [Zhouran et al, 2021] Directly optimize the perturbation with respect to a BPR-based loss function

# **Part 4:**

# **Transparency**

# Motivation

- Transparency of UM and RecSys allows insights into reasons for why certain items were recommended
- Critical for various stakeholders: users, developers, providers, policymakers
  - evaluate relevance and accuracy of results
  - system diagnostics and system performance
  - audit the system for potentially harmful biases or privacy violations
- Broader trend towards transparency - in particular in high-stakes domains
  - healthcare
  - employment
  - education

# 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



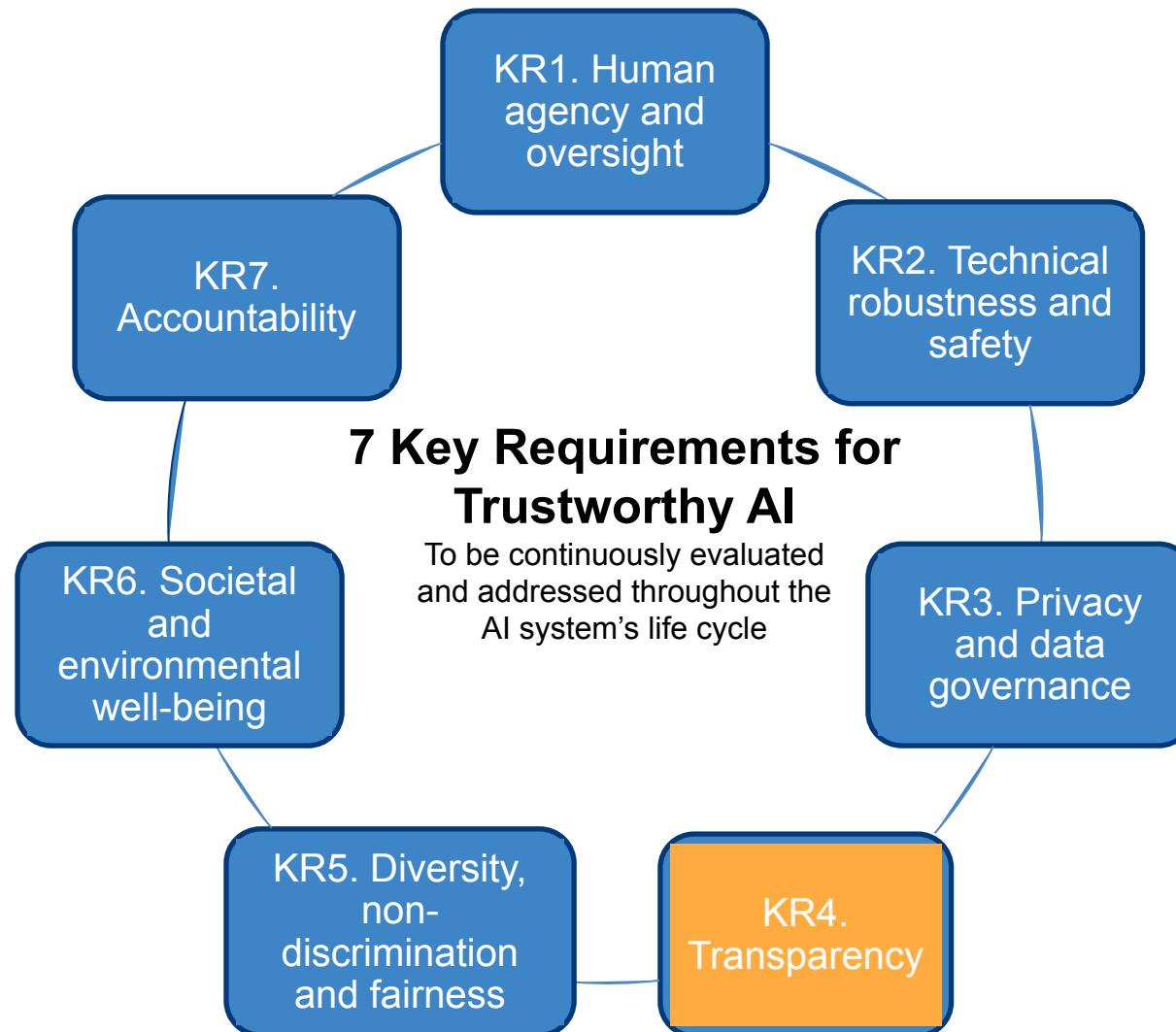
# EU Regulations

- 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
- AI Act
  - Transparency as a key requirement
  - Besides: technical documentation for high-risk use cases



<https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>  
<https://artificialintelligenceact.eu/the-act/>

# One of the requirements for 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 create fair outputs
- User perceptions of fairness
  - Explanations may lead to new behavior
  - Taking fair actions; at least, informed choices

# Explainability, Justification, Interpretability

- Explainability aims to answer *“the problem of why”*
- Justification refers to explaining to a user that a result is relevant and valuable
- Interpretability emphasizes understanding the internal workings of a model and the relationships between its inputs and outputs

## Frequently bought together



This item: LONELY  
PLANET Reiseführer  
Sardinien: Eigene We...  
€20<sup>50</sup> ✓prime

MARCO POLO  
Reiseführer Sardinien:  
Reisen mit Insider-...  
€16<sup>40</sup> ✓prime

Sardinien: Die  
schönsten Küsten- und  
Bergwanderungen. 7...  
€17<sup>40</sup> ✓prime

## Recommendations

We make recommendations based on your interests.

We examine the items you've purchased, items you've told us you own, and items you've rated. We compare your activity on our site with that of other customers, and using this comparison, recommend other items that may interest you on [Your Amazon](#) page.

Your recommendations change regularly, based on a number of factors, including when you purchase or rate a new item, and changes in the interests of other customers like you. Because your recommendations fluctuate, we suggest that you add items that interest you to your [Wish List](#) or [Shopping Cart](#).

# Explainability in UM and RecSys

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

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

# Explainability in UM and RecSys

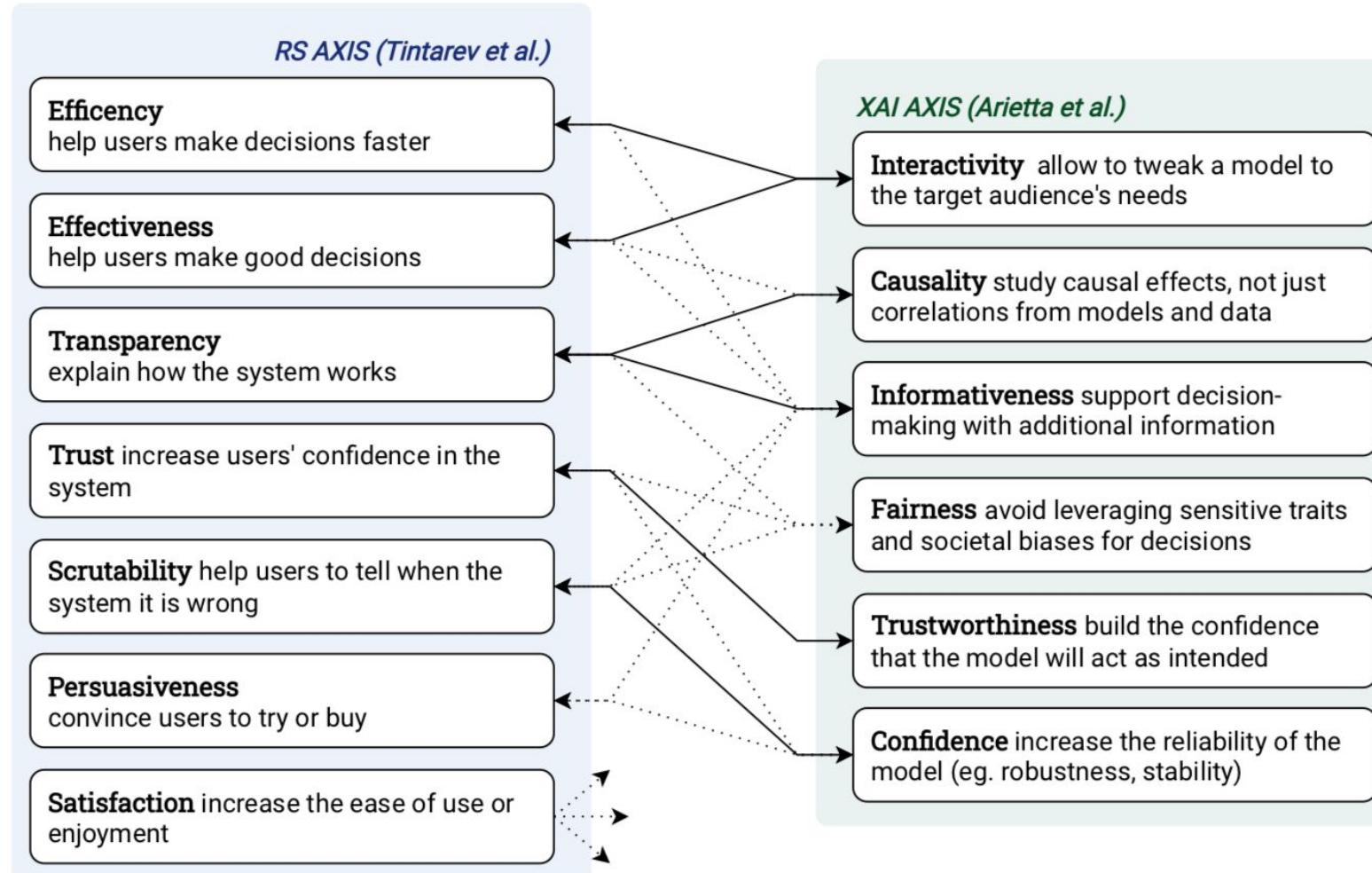
Complementary information

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

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

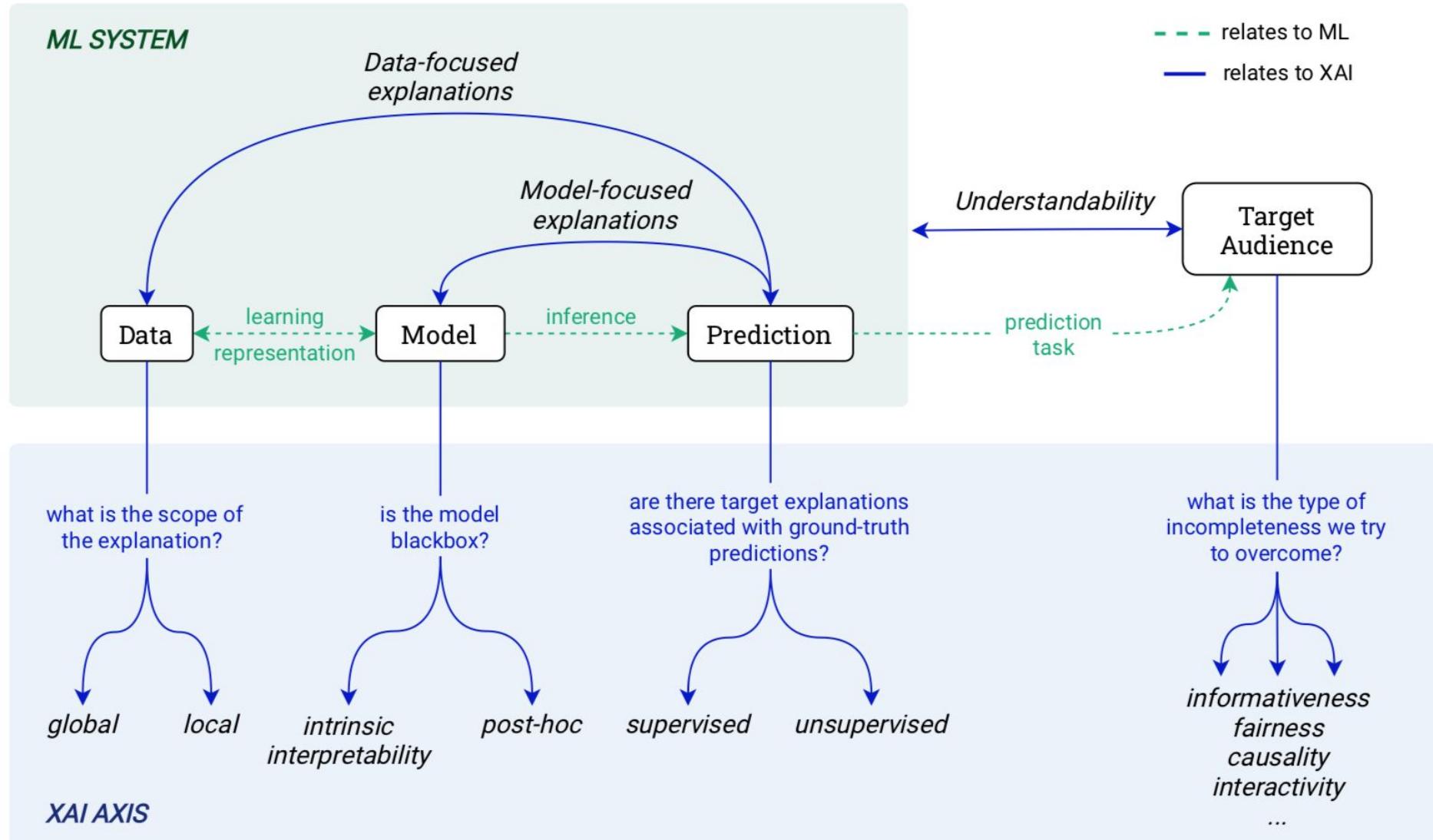
Helps ensure fairness regarding e.g. protected attributes. However: how to act upon them?

# Link to eXplainable AI (XAI)



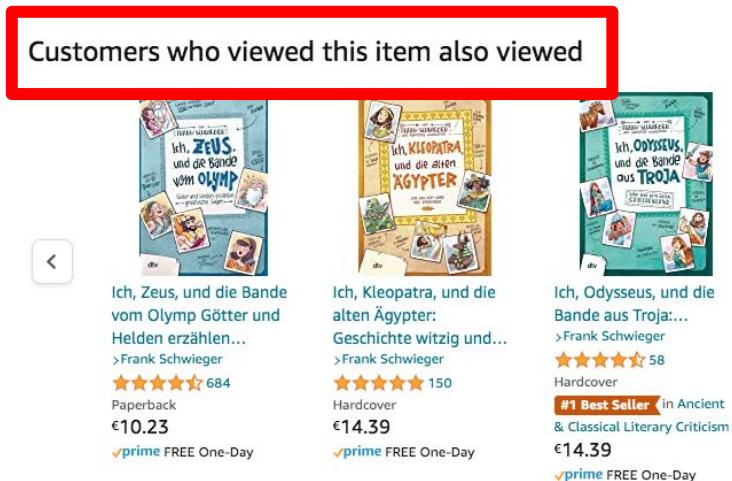
Afchar, D., Melchiorre, A. B., Schedl, M., Hennequin, R., Epure, E. V., & Moussallam, M. (2022). Explainability in Music Recommender Systems.  
<https://onlinelibrary.wiley.com/doi/full/10.1002/aaa.i.12056> & arXiv preprint arXiv:2201.10528.

# XAI Notions



# Local vs. Global

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



**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

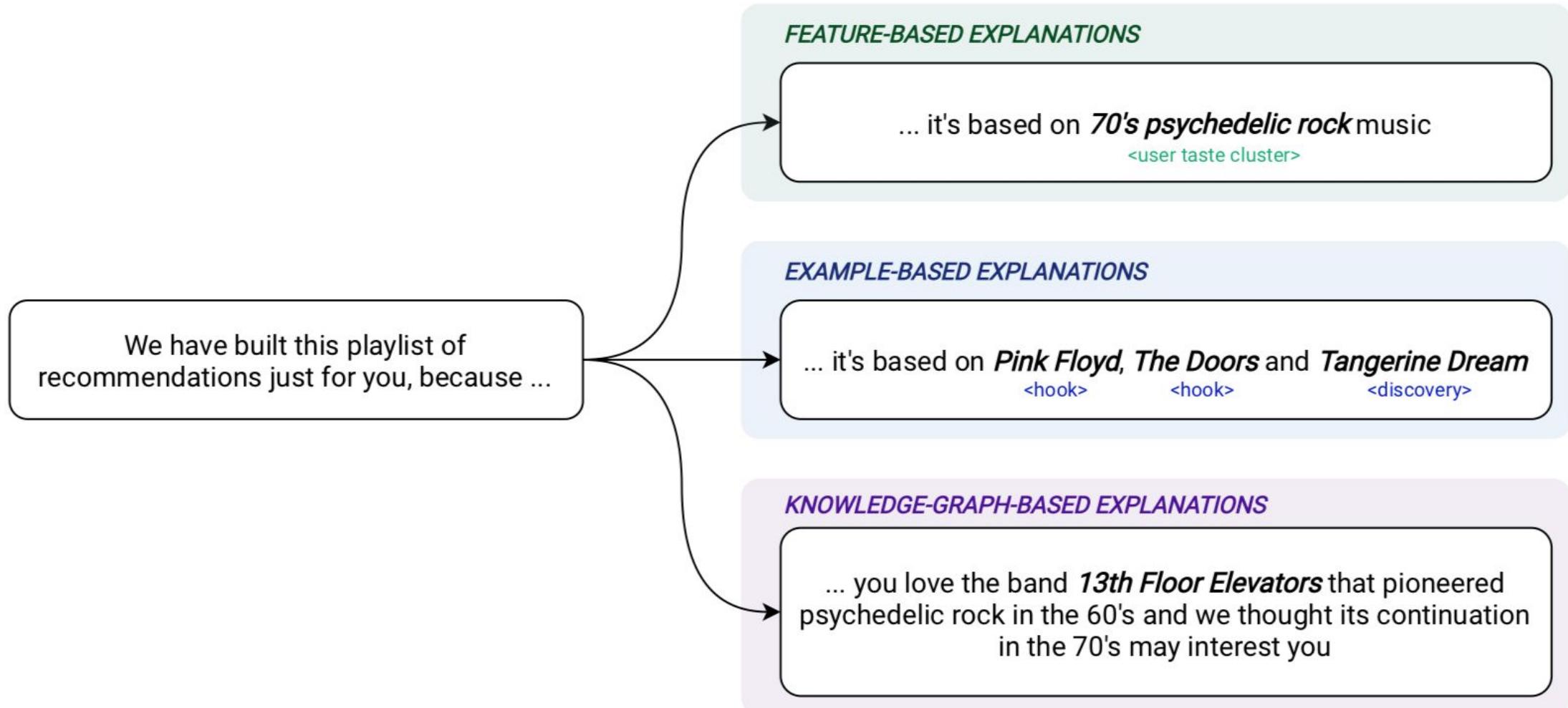
“The has recommended you the item because it maximizes the probability of being co-listened with your history, considering all other users listening history”

**Data:** explain data characteristics

Helps find irregularities in training data

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

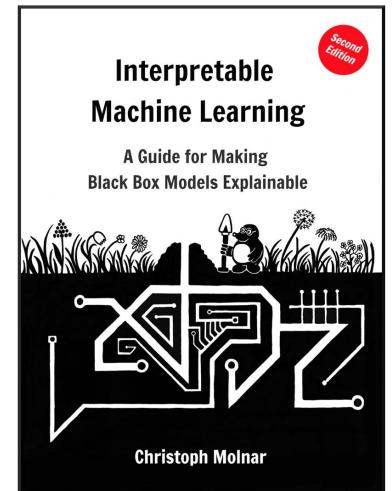
# Generating Explanations: Types



# Properties of Good Explanations

- Accuracy
- Fidelity
- Consistency
- Stability
- Comprehensibility

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



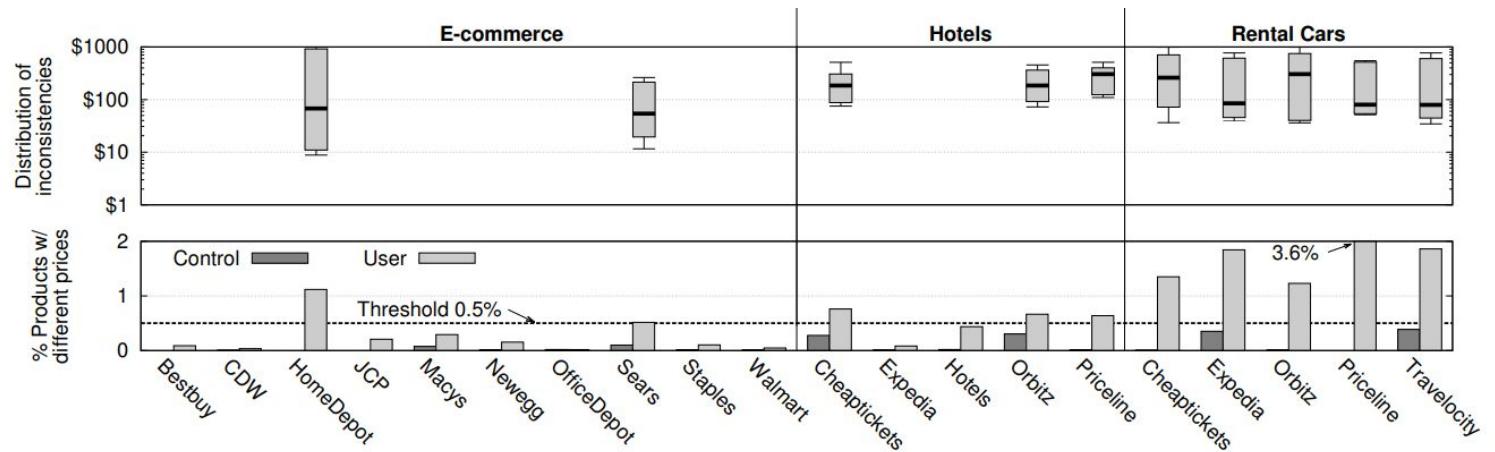
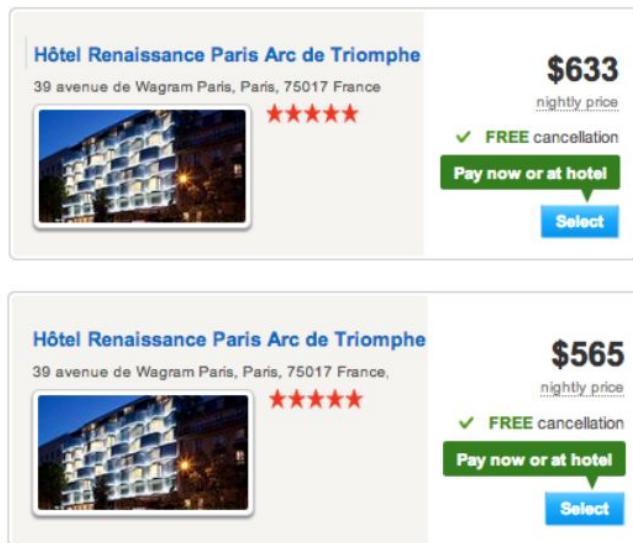
# Algorithm Auditing

- Aim: audit algorithms for biased, discriminatory, harmful behavior
  - alignment of systems with laws, regulations, ethics, ...
- Inspired by audits in finance, security, employment,...
- Involves third party external experts:
  - researchers
  - developers
  - policymakers
- Helped uncover bias in AI systems, e.g., 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



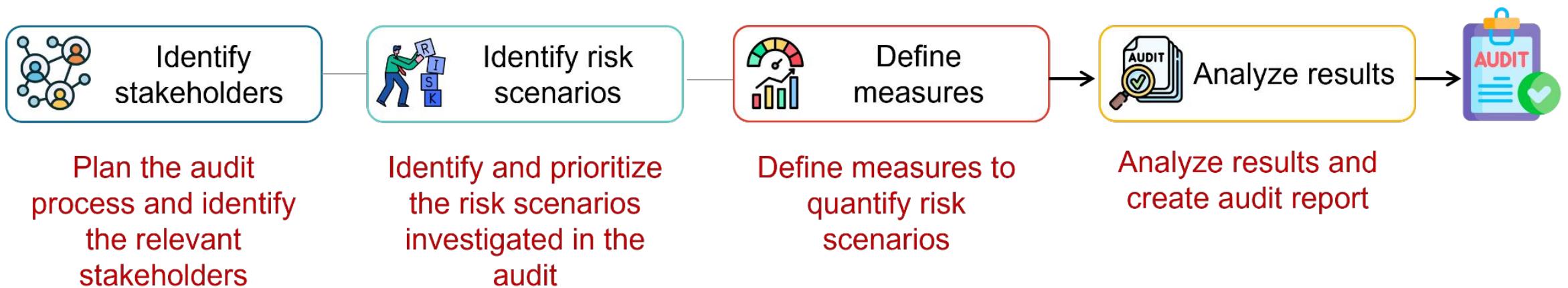
**Figure 3:** Percent of products with inconsistent prices (bottom), and the distribution of price differences for sites with  $\geq 0.5\%$  of products showing differences (top), across all users and searches for each web site. The top plot shows the mean (thick line), 25th and 75th percentile (box),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

# Auditing Process

- Meßmer and Degeling (2023): Risk-based approach
- Aim: guideline to enable audits according to the DSA



# 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
- 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!
- Solution: everyday algorithm auditing (DeVos et al., 2022)
  - everyday users detect problematic system behavior via day-to-day interactions with system

# 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.

# 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/>
- DeVos, A., Dhabalia, A., Shen, H., Holstein, K., & Eslami, M. (2022, April). Toward User-Driven Algorithm Auditing: Investigating users’ strategies for uncovering harmful algorithmic behavior. In CHI Conference on Human Factors in Computing Systems (pp. 1-19)

# **Part 5:**

# **Open Challenges**

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

- Which technological foundation do we need to debias data and algorithms in state-of-the-art recommender systems?
- How can we overcome the trade-off between accuracy and fairness?
- How should requirements and aims of various stakeholders (e.g., content creators and consumers, platform providers, policymakers) be accounted for?
- Do computational bias metrics really capture how users perceive fairness?
- What are economic and social consequences of biases resulting from RecSys technology adopted in high-risk areas (e.g., in recruitment, healthcare)?
- What are the legal implications of unfair or intransparent algorithms?

# Open Challenges (Privacy and Security)

- Privacy technical solutions present an inherent trade-off between privacy, accuracy, and efficiency. Randomization techniques increase privacy by lowering accuracy. Cryptographic and secure multi-party computation protocols increase privacy by lowering efficiency. How to choose the right solution?
- Privacy comparison is usually unfair, because of different datasets and different measures for accuracy. How do different privacy-protection techniques compare to each other when applied to the same dataset?
- A user profile is discrete, and modifying users' profiles changes the semantic of their behaviors. What is the best approach for attack designs?
- Most Security of RS focus on accuracy metrics. What is the impact of adversarial attacks and defenses in diversity, novelty and fairness of recommendations?
- Most of the modern approaches make use of computationally expensive models. What is the scalability and stability of learning of these models? Are there any other (better) learning frameworks to exploit?

# 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 **behavior**?
- What are effective explanation types for **different domains**?
- What do explanations **tell us about the user**? What ethical and privacy implications can arise?
- How to deal with increasingly complex models and their **inherent lack of transparency**?
- How to implement **transparency requirements** as mandated by different regulatory bodies?
- Who **audits algorithms** for problematic behavior?

# Related Surveys and Books

- Michael D. Ekstrand, Anubrata Das, Robin Burke, Fernando Diaz: **Fairness in Information Access Systems**. Foundations and Trends in Information Retrieval 16(1-2): 1-177 (2022)
- Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, Shaoping Ma: **A Survey on the Fairness of Recommender Systems**. ACM Transactions on Information Systems 41(3): 52:1-52:43 (2023)
- Deldjoo, Y., Noia, T. D., & Merra, F. A. (2021). **A Survey on Adversarial Recommender Systems: From Attack/Defense Strategies to Generative Adversarial networks**. ACM Computing Surveys (CSUR), 54(2), 1-38.
- Darius Afchar, Alessandro B. Melchiorre, Markus Schedl, Romain Hennequin, Elena V. Epure, Manuel Moussallam: **Explainability in Music Recommender Systems**. AI Magazine 43(2): 190-208 (2022)
- Yongfeng Zhang, Xu Chen: **Explainable Recommendation: A Survey and New Perspectives**. Foundations and Trends in Information Retrieval 14(1): 1-101 (2020)
- Wenqi Fan et al.: **A Comprehensive Survey on Trustworthy Recommender Systems**. (under review at ACM TORS SI TRS), pre-print: <https://arxiv.org/pdf/2209.10117.pdf>
- Markus Schedl, Vito Walter Anelli, Elisabeth Lex: **Information Retrieval and Recommender Systems: Technical, Ethical, and Regulatory Perspectives**. Springer, to appear in 2024.

# 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)

## Vito Walter Anelli

Politecnico di Bari, Italy  
[vitowalter.anelli@poliba.it](mailto:vitowalter.anelli@poliba.it) | <https://sisinflab.poliba.it/people/vito-walter-anelli/>

## Elisabeth Lex

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

## Acknowledgments

- Emilia Gómez (European Commission) for contributions to earlier variants of the tutorials
- Tommaso Di Noia, Antonio Ferrara (PolIBa); Felice Merra (Amazon)
- Austrian Science Fund; State of Styria, State of Upper Austria, Federal Ministry of Education, Science, and Research