

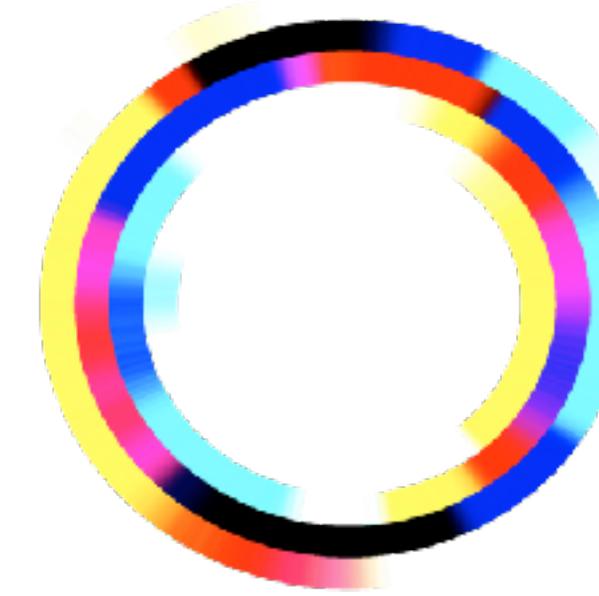
Responsible AI & GLAM: challenges and opportunities

Laura Hollink

Human-Centered Data Analytics Group
Centrum Wiskunde & Informatica



Cultural AI
a lab for
culturally
valued AI



**AI,
media and
democracy**



AI in the GLAM sector

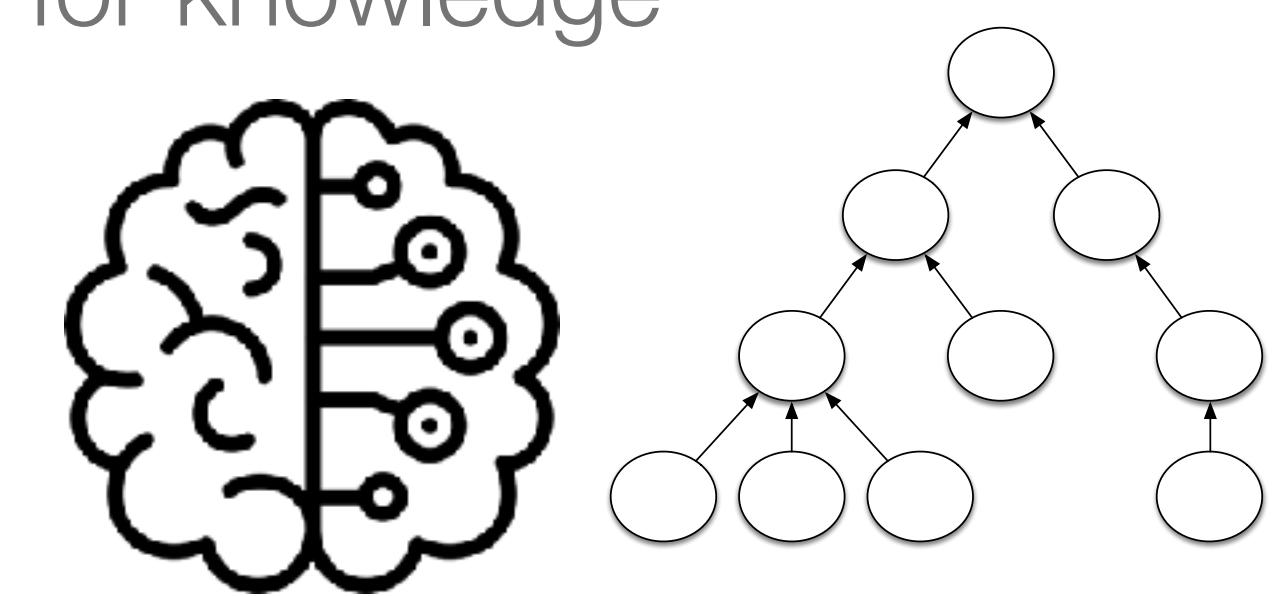
- Recommender systems
- Automatic classification, tagging
- Metadata creation and enrichment
- Handwriting recognition, OCR, etc.

A woman with long dark hair and glasses is sitting at a desk, looking intently at a computer monitor. She has her hand to her chin in a thoughtful pose. The background shows a cluttered office environment with books and papers.

Transparency?
Privacy?
Inclusivity?
Diversity?

Responsible AI

- A broad research field related to developing, assessing and deploying AI in an ethical way.
 - Fairness, bias, non-discrimination, diversity, privacy, security, transparency, accountability, etc.
 - Relevant for machine learning (incl. deep learning/generative AI) but also for knowledge representation and reasoning (e.g. knowledge graphs, thesauri)

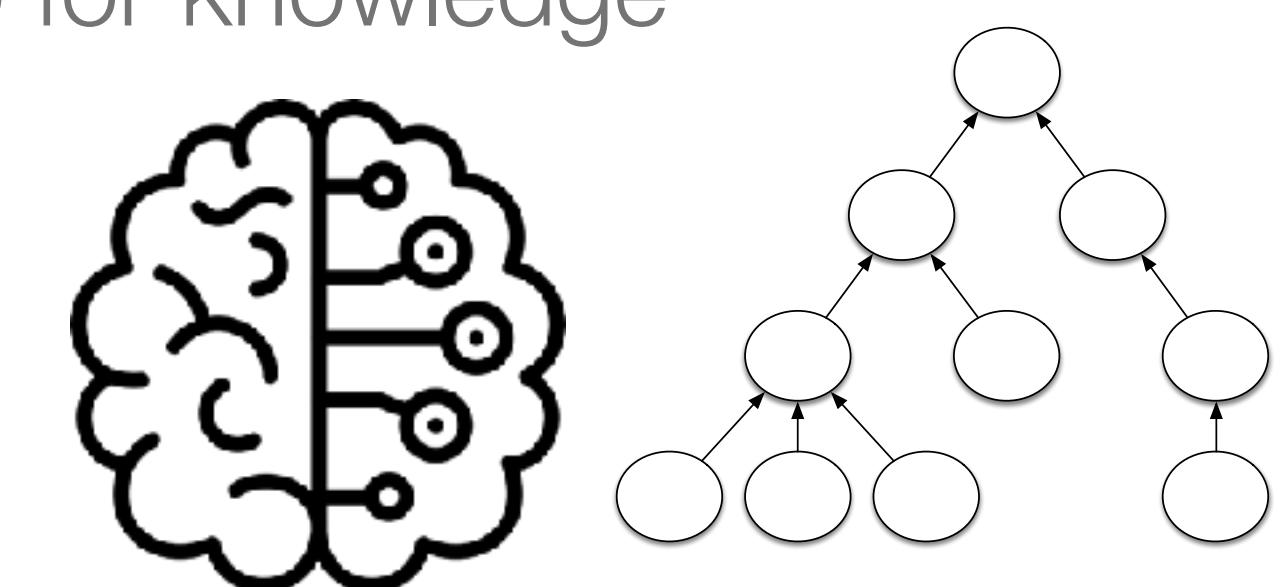
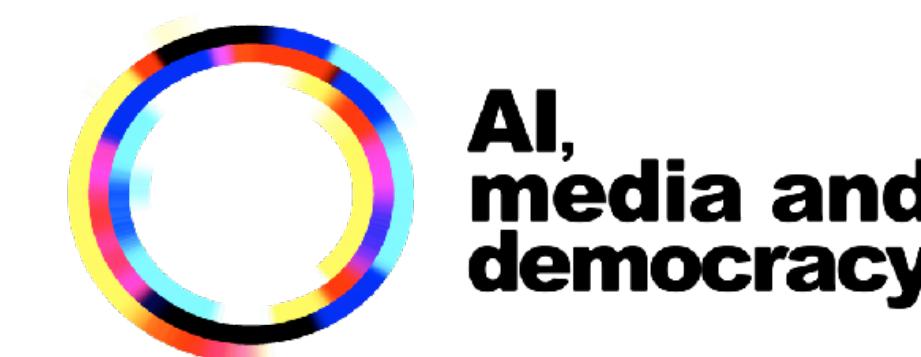


Responsible AI

- A broad research field related to developing, assessing and deploying AI in an ethical way.
 - Fairness, bias, non-discrimination, diversity, privacy, security, transparency, accountability, etc.
 - Relevant for machine learning (incl. deep learning/generative AI) but also for knowledge representation and reasoning (e.g. knowledge graphs, thesauri)

This talk

- **What happens in the AI research community that is relevant for GLAM?**
 - e.g. inclusivity and, in the Netherlands: decolonisation of heritage data
- **What is (or can be) the role of GLAM in creating responsible AI?**
- Examples from



Acknowledgements

- This talk is based on the results of - and discussions with - past and present PhD students and postdocs at CWI, in the Cultural AI Lab and in the AI, Media and Democracy Lab.



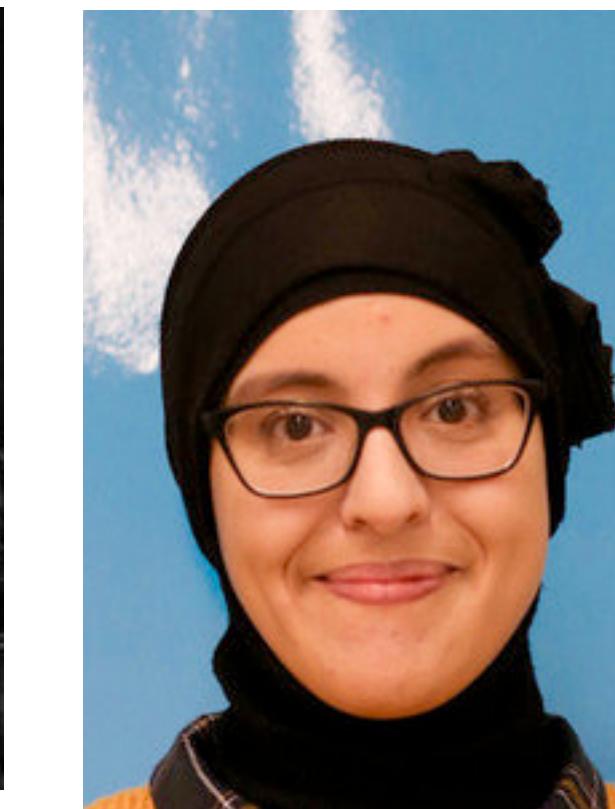
Tessel
Bogaard



Andrei
Nesterov



Savvina
Daniil



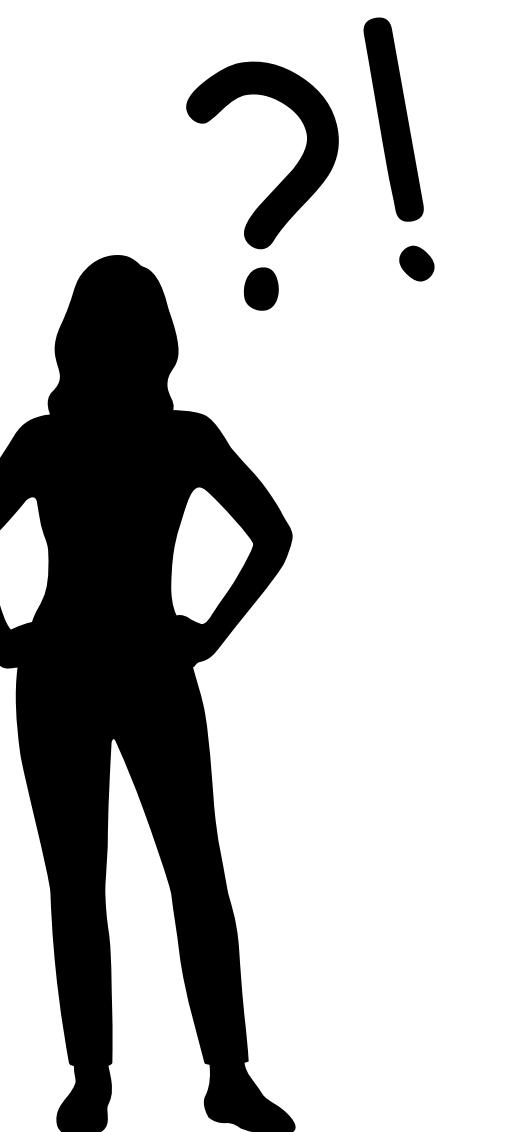
Manel
Slokom



Sanne
Vrijenhoek

Different goals for responsible AI

What do we mean when we say we want
'fairness' or 'diversity'?

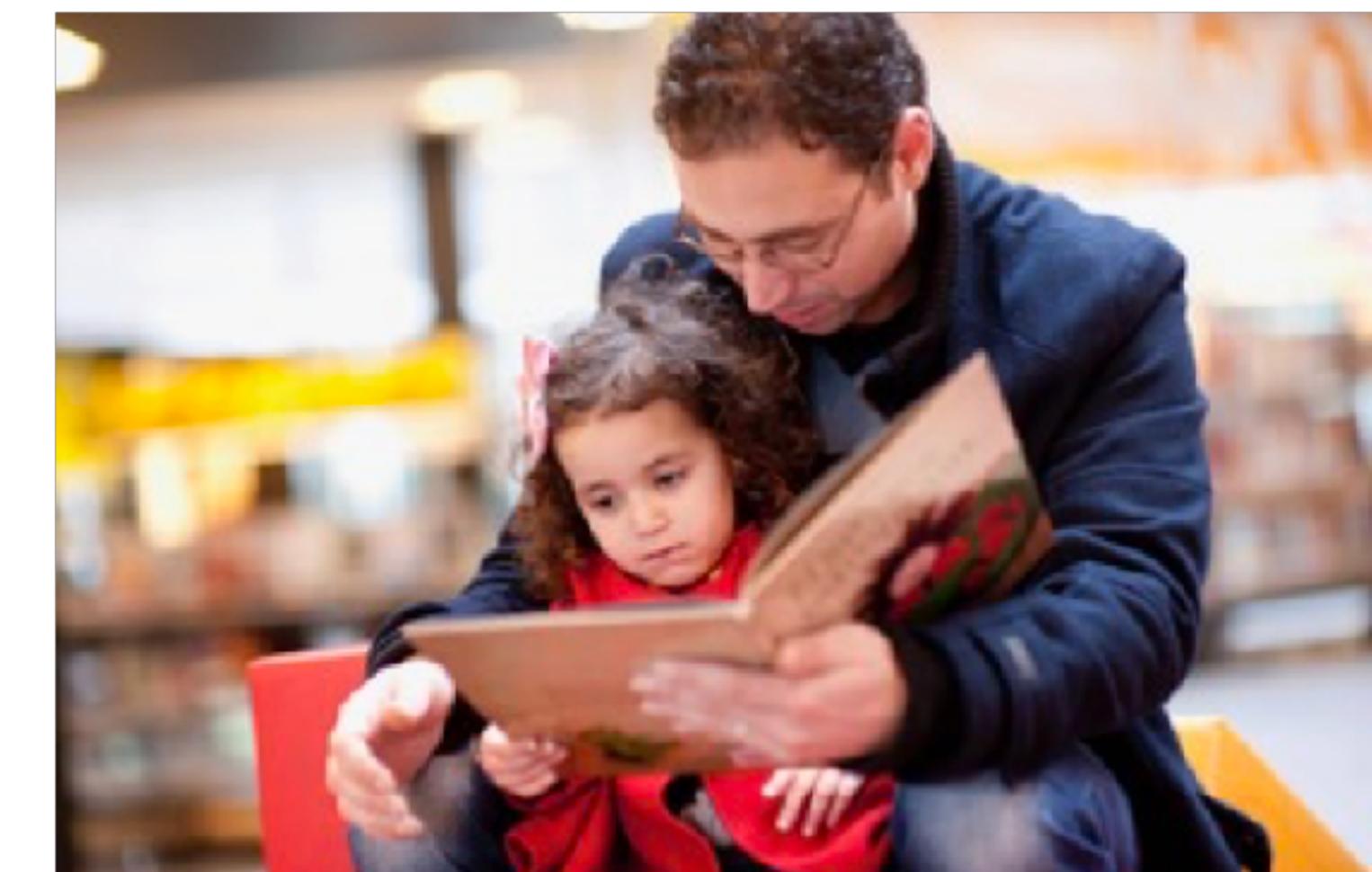


Different goals for responsible AI systems - examples from the media and culture sector.

- National Library of the Netherlands aims to be **Neutral**: “*We do not develop or use AI applications that actively aim to manipulate people's behavior or thinking.*”

AI in Libraries: Seven Principles
Jan Willem Van Wessel <https://doi.org/10.5281/zenodo.3865343>

- Recommendations should be as close as possible to the items someone would consume on their own?
- Recommendations should offer a wide variety of items?



Different goals for responsible AI systems - examples from the media and culture sector.

- **Diversity** is often mentioned as a goal of news recommender systems [1, 2]. [3] define diversity metrics depending on the role of media in democracy:
 - **Participatory model:** media should give citizens what they need to be (politically) engaged -> recommendations should be a reflection of the real political world, with a larger share for more prevalent opinions.
 - **Critical model:** media should critically reflect on the status quo -> recommendations should highlight ‘alternative voices’, i.e. content from people from minority or marginalised groups.



[1] Balazs Bodo. 2019. Selling News to Audiences – A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media. *Digital Journalism* 0, 0 (2019), 1–22.

[2] Helberger, N., K. Karppinen, L. D'Acunto. 2018. Exposure Diversity as a Design Principle for Recommender Systems. *Information, Communication & Society* 21(2):191–207.

[3] S. Vrijenhoek, M. Kaya, N. Metoui, J. Möller, D. Odijk, and N. Helberger. Recommenders with a Mission: Assessing Diversity in News Recommendations. In Proc of CHIIR '21

Ongoing work: we study varying notions of ‘diversity’

- Interviews with 3 public organisations - a library, a news organisation and a TV broadcaster - about how they see “diversity” in the context of a recommender system.



Savvina
Daniil



Sanne
Vrijenhoek

Ongoing work: we study varying notions of ‘diversity’

- Interviews with 3 public organisations - a library, a news organisation and a TV broadcaster - about how they see “diversity” in the context of a recommender system.

“ensuring that [...] everyone feels that there is something for [them]” and “[that people] recognise themselves in the author or in the main characters or the topics []”



Savvina
Daniil



Sanne
Vrijenhoek

Ongoing work: we study varying notions of ‘diversity’

- Interviews with 3 public organisations - a library, a news organisation and a TV broadcaster - about how they see “diversity” in the context of a recommender system.

“ensuring that [...] everyone feels that there is something for [them]” and “[that people] recognise themselves in the author or in the main characters or the topics []”

*“If you’re diverse, you don’t take a stance.
Because you show everything”*



Savvina
Daniil



Sanne
Vrijenhoek

Ongoing work: we study varying notions of ‘diversity’

- Interviews with 3 public organisations - a library, a news organisation and a TV broadcaster - about how they see “diversity” in the context of a recommender system.



Savvina
Daniil



Sanne
Vrijenhoek

“ensuring that [...] everyone feels that there is something for [them]” and “[that people] recognise themselves in the author or in the main characters or the topics []”

*“If you’re diverse, you don’t take a stance.
Because you show everything”*

We found differences in diversity

- **goals (e.g. diverse content vs. a diverse user base)**
- **granularity (e.g. diverse lists vs. diverse items)**
- **characteristics to consider (gender, ability, genre, etc.).**

Ongoing work: we study varying notions of ‘diversity’

- Interviews with 3 public organisations - a library, a news organisation and a TV broadcaster - about how they see “diversity” in the context of a recommender system.



Savvina
Daniil



Sanne
Vrijenhoek

“ensuring that [...] everyone feels that there is something for [them]” and “[that people] recognise themselves in the author or in the main characters or the topics []”

*“If you’re diverse, you don’t take a stance.
Because you show everything”*

In AI, ‘diversity’ usually means: items in a list are sufficiently different from each other, often in terms of genre, topic, producer, etc.

We found differences in diversity

- **goals (e.g. diverse content vs. a diverse user base)**
- **granularity (e.g. diverse lists vs. diverse items)**
- **characteristics to consider (gender, ability, genre, etc.).**

With the widespread use of artificial intelligence (AI) systems and applications in our everyday lives, accounting for fairness has gained significant importance in designing and engineering of such systems. AI systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that these do not reflect discriminatory behavior toward certain groups or populations. More recently, some work has been developed in traditional machine learning and deep learning that address such challenges in different domains. With the commercialization of these systems, researchers are becoming more aware of the biases that these applications can contain and are attempting to address them. In this survey, we investigated different real-world applications that have shown biases in various ways, and we listed different sources of biases that can affect AI applications. We then created a taxonomy for fairness definitions that machine learning researchers have defined, avoid the existing bias in AI systems. In addition to that, we examined different domains and industries in AI showing what researchers have observed with regard to unfair outcomes in the state-of-the-art methods and ways they have tried to address them. There are still many future directions and solutions that can be taken to mitigate the problem of bias in AI systems. We are hoping that this survey will motivate researchers to tackle these issues in the near future by observing existing work in their respective fields.

CCS Concepts: • Computing methodologies → Artificial intelligence

Additional Key Words and Phrases: Fairness and bias in artificial intelligence, machine learning, deep learning, natural language processing, representation learning

ACM Reference format:
Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 93 pages.
<https://doi.org/10.1145/3487607>

Varying notions of ‘fairness’ in Mehrabi et al.

- ...
- **Definition 2 (Treatment equality)** “*Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories*”
- **Definition 8 (Counterfactual Fairness)** “*a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group.*”
- ...

Ninareh Mehrabi,
Fred Morstatter,
Nripsuta Saxena,
Kristina Lerman, and
Aram Galstyan. 2021.
A Survey on Bias and
Fairness in Machine
Learning. ACM
Comput. Surv. 54, 6,
Article 115 (July
2022)

With the widespread use of artificial intelligence (AI) systems and applications in our everyday lives, accounting for fairness has gained significant importance in designing and engineering of such systems. AI systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that these decisions do not reflect discriminatory behavior toward certain groups or populations. More recently, some work has been developed in traditional machine learning and deep learning that address such challenges in different domains. With the commercialization of these systems, researchers are becoming more aware of the biases that these applications can contain and are attempting to address them. In this survey, we investigated different real-world applications that have shown biases in various ways, and we listed different sources of biases that can affect AI applications. We then created a taxonomy for fairness definitions that machine learning researchers have defined, avoid the existing bias in AI systems. In addition to that, we examined different domains and industries in AI showing what researchers have observed with regard to unfair outcomes in the state-of-the-art methods and ways they have tried to address them. There are still many future directions and solutions that can be taken to mitigate the problem of bias in AI systems. We are hoping that this survey will motivate researchers to tackle these issues in the near future by observing existing work in their respective fields.

CCS Concepts: • Computing methodologies → Artificial intelligence

Additional Key Words and Phrases: Fairness and bias in artificial intelligence, machine learning, deep learning, natural language processing, representation learning

ACM Reference format:
Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 93 pages.
<https://doi.org/10.1145/3487607>

Varying notions of ‘fairness’ in Mehrabi et al.

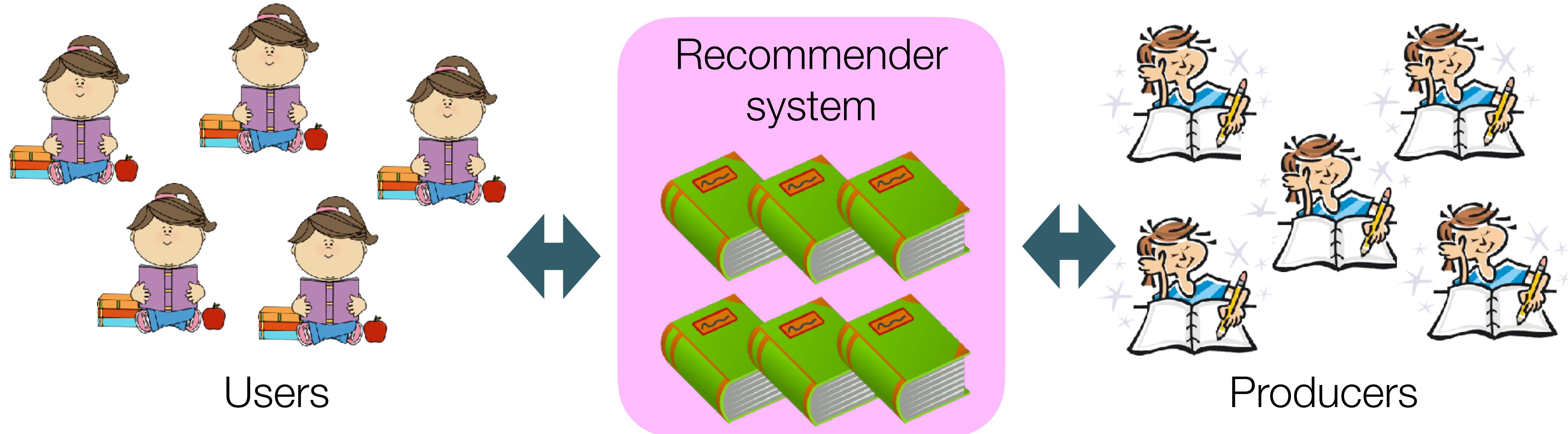
- ...
- **Definition 2 (Treatment equality)** “*Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories*”
- **Definition 8 (Counterfactual Fairness)** “*a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group.*”
- ...

Group-fairness

Individual-fairness

Ninareh Mehrabi,
Fred Morstatter,
Nripsuta Saxena,
Kristina Lerman, and
Aram Galstyan. 2021.
A Survey on Bias and
Fairness in Machine
Learning. ACM
Comput. Surv. 54, 6,
Article 115 (July
2022)

User-fairness and producer-fairness



Both are relevant for GLAM.

- E.g. does art by female artists get the same visibility as male artists?
- E.g. do readers from minority groups get the same quality recommendations?

(Un)availability of data

What data do we need/have
to measure whether AI is fair, diverse, etc?



Sensitive data

Many of these tools/approaches/metric require to know who is the ‘protected class’



[Photo by Marcin Bajer on flickr.com, CC BY-NC 2.0](#)

Sensitive data

Many of these tools/approaches/metric require to know who is the ‘protected class’

- Legally protected characteristics: race, color, national origin, religion, sex, age, or disability. (or see <https://mensenrechten.nl/en/node/3>)



[Photo by Marcin Bajer on flickr.com, CC BY-NC 2.0](#)

Sensitive data

- To study user-fairness, we need sensitive data about users
- To study producer-fairness, we need sensitive data about producers



Low availability
In GLAM



High availability
In GLAM

Sensitive data

- To study user-fairness, we need sensitive data about users
- To study producer-fairness, we need sensitive data about producers



Low availability
In GLAM



High availability
In GLAM

		Items					
		1	2	...	i	...	m
Users	1	5	3		1	2	
	2		2				4
	:			5			
	u	3	4		2	1	
	:						4
	n			3	2		

- In the AI research community, we generally have neither.

Image from

https://www.researchgate.net/figure/Sample-of-user-item-matrix_fig1_284737564

Popularity bias

- ◆ A known phenomenon in recommender systems “where popular items tend to be suggested over long-tail ones, even if the latter would be of reasonable interest for individuals”
- ◆ Can be studied with just user-item matrices

Emre Yalcin, Alper Bilge.
Investigating and counteracting
popularity bias in group
recommendations, Information
Processing & Management, 58(5)
2021.

		<i>Items</i>					
		<i>I</i>	<i>2</i>	...	<i>i</i>	...	<i>m</i>
<i>I</i>	<i>1</i>	5	3		1	2	
	2		2				4
:				5			
<i>u</i>	3	4		2	1		
:						4	
<i>n</i>			3	2			

Image from
https://www.researchgate.net/figure/Sample-of-user-item-matrix_fig1_284737564

Example study on popularity bias

- Abdollahpouri et al. study this from a user perspective:
 - **User groups:** Niche users, Diverse users, Blockbuster users.
 - **RQ:** How does popularity bias affect each group?
 - **Results:** All algorithms were extremely unfair to users with lesser interest in popular items.
 - Similar studies have been done on e.g. books and music.

Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.:
The unfairness of popularity bias in recommendation.
arXiv preprint
arXiv:1907.13286 (2019)

The Unfairness of Popularity Bias in Recommendation*

Hosseini Abdollahpouri
University of Colorado Boulder
Boulder, USA
hosseini.abdollahpouri@colorado.edu

Mansoury Mansoury
Eindhoven University of Technology
Eindhoven, Netherlands
mansoury@tue.nl

Robin Burke
University of Colorado Boulder
Boulder, USA
robin.burke@colorado.edu

Baneshad Mobasher
DePaul University
Chicago, USA
mobasher@cs.depaul.edu

ABSTRACT

Recommender systems are known to suffer from the popularity bias problem: the frequently rated items get a lot of exposure while less popular ones are under-represented in the recommendations. Research in this area has been mainly focusing on finding ways to tackle this issue by increasing the number of recommended long-tail items or otherwise. In this paper, however, we look at this problem from the user's perspective: we want to see how popularity bias causes the recommendations to deviate from what the user expects to get from the recommender system. We define three different groups of users according to their interest in popular items (Niche, Diverse and Blockbuster users) and show the impact of popularity bias on the users in each group. Our experimental results on a movie dataset show that many recommendation algorithms the recommender systems the users get are extremely concentrated on popular items even if a user is interested in long-tail and non-popular items showing an extreme bias disparity.

KEYWORDS

Recommender system, Popularity bias, Fairness, Long-tail recommendation

1 INTRODUCTION

Recommender systems have been widely used in a variety of different domains such as movies, music, video gaming etc. Their goal is to help users find relevant items which are difficult or otherwise time-consuming to find in the absence of such systems.

Different types of algorithms are being used for recommendation depending on the domain or other constraints such as the availability of the data about users or items. One of the most widely-used classes of algorithms for recommendation is collaborative filtering. In these algorithms the recommendations for each user are being generated based on the rating information from other users and items. Unlike some types of algorithms such as content-based recommendation, collaborative algorithms do not use the content information.

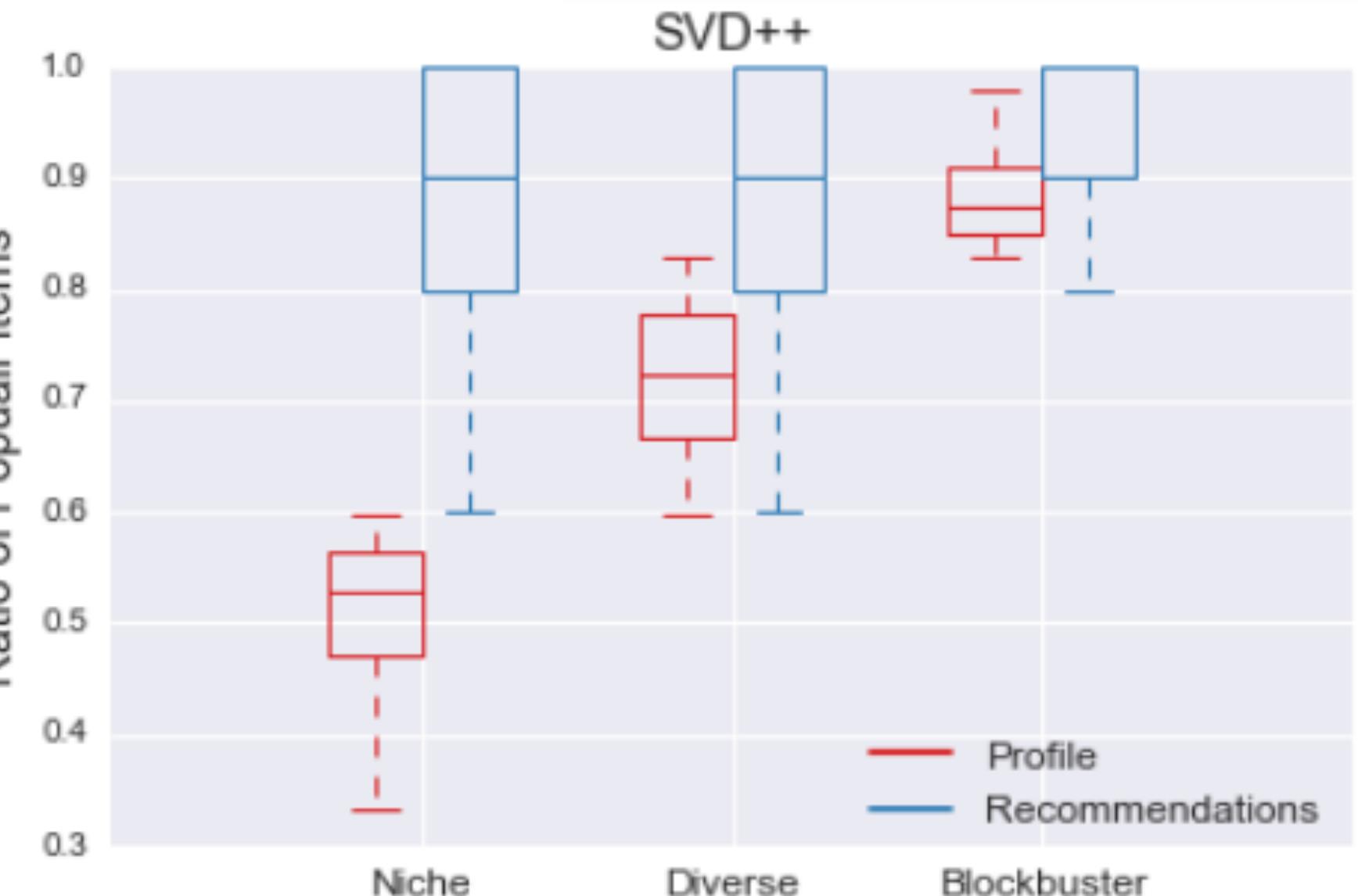
One of limitations of collaborative recommenders is the problem of popularity bias [1, 7]: popular items are being recommended too frequently while the majority of other items do not get the deserved attention.

Popularity bias could be problematic for a variety of different reasons: Long-tail/non-popular items are important for generating a fairer understanding of user's preferences. Systems that are using learning to explore each user's profile will typically need to present more long tail items because these are the ones that the user is less likely to have rated and where user's preferences are more likely to be diverse [15, 2].

In addition long-tail recommendations can also be understood as a social goal. A market that suffers from popularity bias will lack opportunities to discover more obscure products and will be, by definition, dominated by a few large brands or well-known artists [10]. Such a market will be more homogeneous and offer fewer opportunities for innovation and creativity.

In this paper, however, we look at the popularity bias from a different perspective: the user's. Take for example the user described in figure 1: for example, the user has rated 18 long-tail (non-popular) items and 70 popular items. So it is reasonable to expect the recommendations to keep the same ratio of popular and non-popular items. However, most of the recommendation algorithms produce a list of recommendations that is over-concentrated on popular items. When it is even close to 100% popular items. In this paper, we show how different recommendation algorithms are propagating this bias differently for different users. In particular, we define three groups of users according to their degree of interest towards popular items and show how they propagate popularity bias across groups higher than the others. For instance, the niche users (users with the lowest degree of interest towards popular items) are affected the most by this bias. We also show these niche users tend, in general, users with lower interest in popular items, are more active (they have rated more items in the system and therefore they should be considered as important stakeholders).

*Copyright: 2019 for this paper by its author. Licensed under Creative Commons License Attribution International 4.0 (CC BY 4.0). Presented at the 19th ACM workshop held in conjunction with the 19th ACM Conference on Recommender Systems (RecSys), 2019, Copenhagen, Denmark.



(b) SVD++

Example study on popularity bias

- Abdollahpouri et al. study this from a user perspective:
- **User groups:** Niche users, Diverse users, Blockbuster users.
- **RQ:** How does popularity bias affect each group?
- **Results:** All algorithms were extremely unfair to users with lesser interest in popular items.
- Similar studies have been done on e.g. books and movies

Has the focus of the field on popularity bias been mostly data-availability-driven, rather than interest-driven?

Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.:
The unfairness of popularity bias in recommendation.
arXiv preprint
arXiv:1907.13286 (2019)

The Unfairness of Popularity Bias in Recommendation*

Hosseini Abdollahpouri
University of Colorado Boulder
Boulder, USA
hosseini.abdollahpouri@colorado.edu

Robin Burke
University of Colorado Boulder
Boulder, USA
robin.burke@colorado.edu

Baneshad Mobasher

DéPaul University

Chicago, USA

mobasher@cs.depaul.edu

ABSTRACT

Recommendation systems are known to suffer from the popularity bias problem: the frequently rated items get a lot of exposure while less popular ones are under-represented in the recommendations. Research in this area has been mainly focusing on finding ways to tackle this issue by increasing the number of recommended long-tail items or otherwise. The overall catalog coverage. In this paper, however, we look at this problem from the user's perspective: we want to see how popularity bias causes the recommended items to deviate from what the user expects to get from the recommendation system. We define three different groups of users according to their interest in popular items (Niche, Diverse and Blockbuster users) and show the impact of popularity bias on the users in each group. Our experimental results on a movie dataset show that many recommendation algorithms the recommendation engines the users get are extremely concentrated on popular items even if a user is interested in long-tail and non-popular items showing an extreme bias disparity.

KEYWORDS

Recommendation system, Popularity bias, Fairness, Long-tail recommendation

1 INTRODUCTION

Recommendation systems have been widely used in a variety of different domains such as movies, music, video gaming etc. Their goal is to help users find relevant items which are difficult or otherwise time-consuming to find in the absence of such systems.

Different types of algorithms are being used for recommendation depending on the domain or other constraints such as the availability of the data about users or items. One of the most widely-used classes of algorithms for recommendation is collaborative filtering. In these algorithms the recommendations for each user are being generated based on the rating information from other users and items. Unlike other types of algorithms such as content-based recommendation, collaborative algorithms do not use the content information.

One of limitations of collaborative recommenders is the problem of popularity bias [1, 7]: popular items are being recommended too

*Copyright © 2019 for this paper by its author. The paper is copyrighted under Creative Commons License. Available online at <http://dx.doi.org/10.1145/3331184.3331209>. Presented at the 19th ACM SIGKDD Conference on Recommender Systems (RecSys), 2019, in Copenhagen, Denmark.

MOSCOW UNIVERSITY
Eindhoven University of Technology
Eindhoven, Netherlands
mansoury@mathworks.nl

Baneshad Mobasher
DePaul University
Chicago, USA
mobasher@cs.depaul.edu



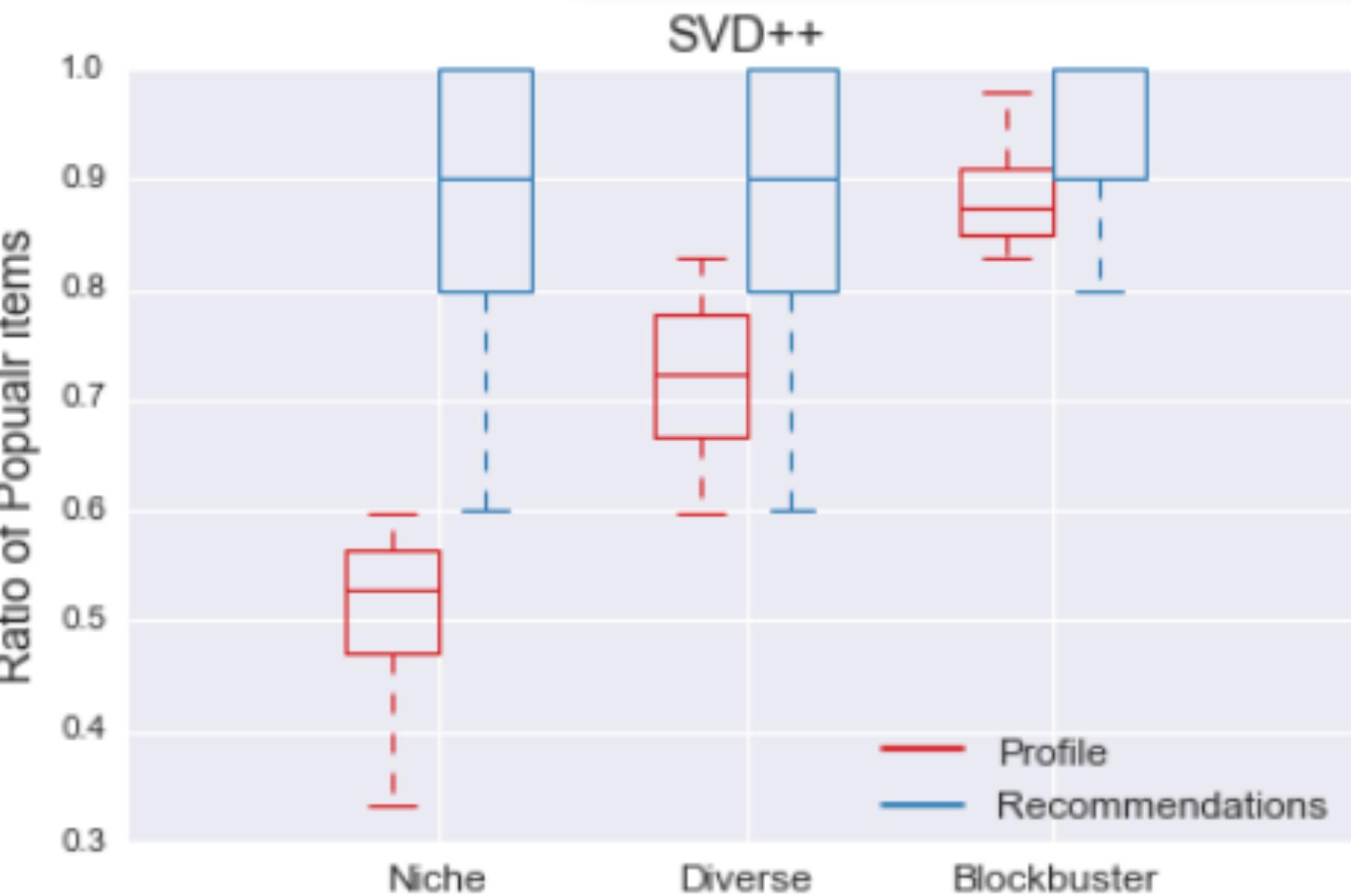
Figure 1: The inconsistency between user's expectation and the given recommendations

frequently while the majority of other items do not get the deserved attention.

Popularity bias could be problematic for a variety of different reasons: Long-tail/non-popular items are important for generating a fairer understanding of user's preferences. Systems that are using learning to explore each user's profile will typically need to prevent users from rating tail items because these are the ones that the user is less likely to have rated and where user's preferences are more likely to be diverse [15, 2].

In addition, long-tail recommendations can also be understood as a social goal. A market that suffers from popularity bias will lack opportunities to discover more obscure products and will be, by definition, dominated by a few large brands or well-known artists [10]. Such a market will be more homogeneous and offer fewer opportunities for innovation and creativity.

In this paper, however, we look at the popularity bias from a different perspective: the user's. Take at figure 1: for example, the user has rated 18 long-tail (non-popular) items and 70 popular items. So it is reasonable to expect the recommendation to keep the same ratio of popular and non-popular items. However, most of the recommendation algorithms produce a list of recommendations that is over-concentrated on popular items. Often it is even close to 100% popular items. In this paper, we show how different recommendation algorithms are propagating this bias differently for different users. In particular, we define three groups of users according to their degree of interest towards popular items and show how they propagate to some groups higher than the others. For instance, the niche users (users with the lowest degree of interest towards popular items) are affected the most by this bias. We also show these niche users tend, in general, users with lower interest in popular items, are more active (they have rated more items in the system and therefore they should be considered as important stakeholders).



(b) SVD++

Example study on popularity bias

- Abdollahpouri et al. study this from a user perspective:
 - **User groups:** Niche users, Diverse users, Blockbuster users.
 - **RQ:** How does popularity bias affect each group?
 - **Results:** All algorithms were extremely unfair to users with lesser interest in popular items.
- Similar studies have been done on e.g. books and movies

Has the focus of the field on popularity bias been mostly data-availability-driven, rather than interest-driven?

Abdollahpouri, H.,
Mansoury, M., Burke,
R., Mobasher, B.:
The unfairness of
popularity bias in
recommendation.
arXiv preprint
arXiv:1907.13286
(2019)

The Unfairness of Popularity Bias in Recommendation*

Hossein Abdollahpouri
University of Colorado Boulder
Boulder, USA
hossein.abdollahpouri@colorado.edu

Robin Burke
University of Colorado Boulder
Boulder, USA
robin.burke@colorado.edu

ABSTRACT

Recommender systems are known to suffer from the popularity bias problem: the frequently rated items get a lot of exposure while less popular ones are under-represented in the recommendations. Research in this area has been mainly focusing on finding ways to tackle this issue by increasing the number of recommended long-tail items or otherwise. The overall catalog coverage. In this paper, however, we look at this problem from the user's perspective: we want to see how popularity bias causes the recommendation to deviate from what the user expects to get from the recommender system. We define three different groups of users according to their interest in popular items (Niche, Diverse and Blockbuster users) and show the impact of popularity bias on the users in each group. Our experimental results on a movie dataset show that many recommendation algorithms the recommender systems the users get are extremely concentrated on popular items even if a user is interested in long-tail and non-popular items showing an extreme bias disparity.

KEYWORDS
Recommender system, Popularity bias, Fairness, Long-tail recommendation

1 INTRODUCTION

Recommender systems have been widely used in a variety of different domains such as movies, music, video gaming etc. Their goal is to help users find relevant items which are difficult or otherwise time-consuming to find in the absence of such systems.

Different types of algorithms are being used for recommendation depending on the domain or other constraints such as the availability of the data about users or items. One of the most widely-used classes of algorithms for recommendation is collaborative filtering. In these algorithms the recommendations for each user are being generated based on the rating information from other users and items. Unlike other types of algorithms such as content-based recommendation, collaborative algorithms do not use the content information.

One of limitations of collaborative recommenders is the problem of popularity bias [1, 7]: popular items are being recommended

*Copyright © 2019 for this paper by its author. The paper is copyrighted under Creative Commons License. Available online at <http://dx.doi.org/10.1145/3331003.3331040>. Presented at the 2019 ACM workshop held in conjunction with the 19th ACM Conference on Recommender Systems (RecSys), 2019, in Copenhagen, Denmark.

MOSCOW UNIVERSITY
Eindhoven University of Technology
Eindhoven, Netherlands
mansoury@mathworks.nl

Baneshad Mobasher
DePaul University
Chicago, USA
mobasher@cs.depaul.edu



Figure 1: The inconsistency between user's expectation and the given recommendations

frequently while the majority of other items do not get the deserved attention.

Popularity bias could be problematic for a variety of different reasons: Long-tail/non-popular items are important for generating a fairer understanding of user's preferences. Systems that are using learning to explore each user's profile will typically need to present more long tail items because these are the ones that the user is less likely to have rated and where user's preferences are more likely to be diverse [15, 2].

In addition, long-tail recommendations can also be understood as a social goal. A market that suffers from popularity bias will lack opportunities to discover more obscure products and will be, by definition, dominated by a few large brands or well-known artists [10]. Such a market will be more homogeneous and offer fewer opportunities for innovation and creativity.

In this paper, however, we look at the popularity bias from a different perspective: the user's. Take at figure 1: for example, the user has rated 10 long-tail (non-popular) items and 70 popular items. So it is reasonable to expect the recommendations to keep the same ratio of popular and non-popular items. However, most of the recommendation algorithms produce a list of recommendations that is over-concentrated on popular items. Often it is even close to 100% popular items. In this paper, we show how different recommendation algorithms are propagating this bias differently for different users. In particular, we define three groups of users according to their degree of interest towards popular items and show how the bias propagates for some groups higher than others. For instance, the niche users (users with the lowest degree of interest towards popular items) are affected the most by this bias. We also show these niche users tend, in general, users with lower interest in popular items, are more active (they have rated more items in the system) and therefore they should be considered as important stakeholders.



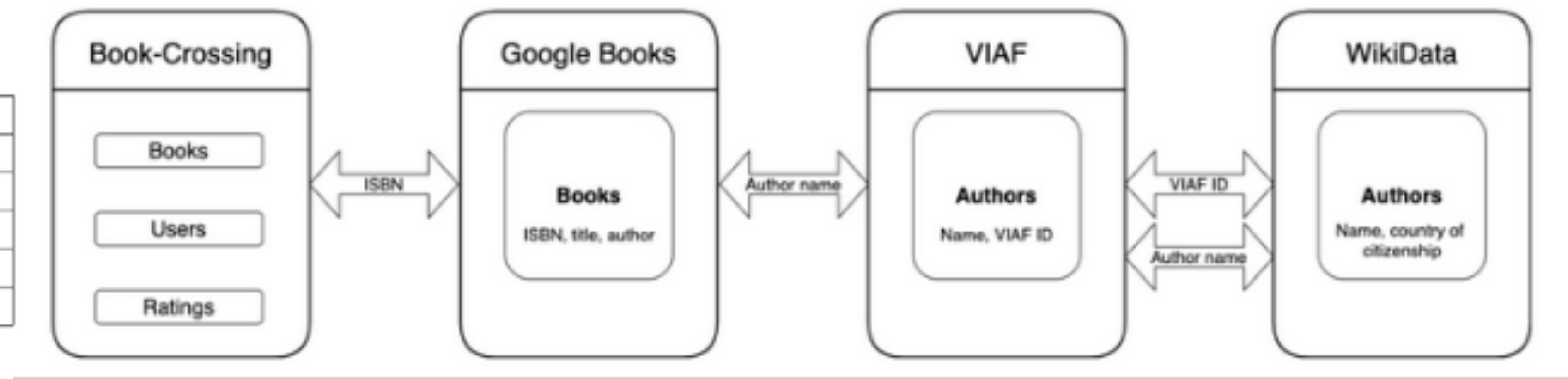
High availability of ‘producer data’ in GLAM means there is a potential role of GLAM to help shape new research directions.

(b) SVD++

Producer-fairness: using the LOD cloud to get (sensitive) data about book authors

We developed a pipeline to add sensitive characteristics to the well-known Book-Crossing dataset.

		Items				
		1	2	... i	...	m
Users	1	5	3		1	2
	2		2			4
u			5			
	3	3	4	2	1	
:						4
	n			3	2	



Hidden Author Bias in Book Recommendation*

Savvina Daniil

s.daniil@cwi.nl

Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

Mirjam Cuper

mirjam.cuper@kb.nl

National Library of the Netherlands
The Hague, The Netherlands

Jacco van Ossenbruggen

jacco.van.ossenbruggen@cwi.nl

Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Laura Hollink

l.hollink@cwi.nl

Centrum Wiskunde & Info
Amsterdam, The Netherl

personal information of the user. That by omitting sensitive information can lead to manifest in such a system. Collaborative filtering approaches are still known to suffer from popularity bias. In short, popularity bias is an algorithmic bias that occurs when originally popular in the training set items are recommended more often and thus have their popularity reinforced.

ABSTRACT

Collaborative filtering algorithms have the advantage of not requiring sensitive user or item information to provide recommendations. However, they still suffer from fairness related issues, like popularity bias. In this work, we argue that popularity bias often leads to other biases that are not obvious when additional user or item information is not provided to the researcher. We examine our hypothesis that by omitting sensitive information can lead to manifest in such a system.

Savvina
Daniil



Author data in Wikidata

Zadie Smith (Q140052)

British novelist, essayist, and short-story writer

Zadie Adeline Smith

▼ In more languages

Configure

Language	Label	Description	Also known as
English	Zadie Smith	British novelist, essayist, and short-story writer	Zadie Adeline Smith
Dutch	Zadie Smith	Brits schrijfster	
German	Zadie Smith	britische Schriftstellerin	
French	Zadie Smith	écrivaine britannique	

All entered languages

Statements

instance of	 human	edit
» 3 references		

image		 sex or gender	 female
		» 2 references	

country of citizenship	 United Kingdom	» 2 references
------------------------	--	--------------------------------

Hidden Author Bias in Book Recommendation*

Savvina Daniil

s.daniil@cwi.nl

Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

Mirjam Cuper

mirjam.cuper@kb.nl

National Library of the Netherlands
The Hague, The Netherlands

Jacco van Ossenbruggen

jacco.van.ossenbruggen@cwi.nl

Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Delft University of Technology

Delft, The Netherlands

Laura Hollink

l.hollink@cwi.nl

Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

personal information of the user. However, it is well-known that by omitting sensitive information from the user profile, it can manifest in such a system. This is because collaborative filtering approaches are still known to suffer from popularity bias. Short, popularity bias is an algorithmic bias that occurs when items originally popular in the training set are recommended more often and thus have their popularity reinforced.



Savvina
Daniil

Author data in Wikidata

Zadie Smith (Q140052)

British novelist, essayist, and short-story writer

Zadie Adeline Smith

▼ In more languages

Configure

Language	Label	Description	Also known as
English	Zadie Smith	British novelist, essayist, and short-story writer	Zadie Adeline Smith
Dutch	Zadie Smith	Brits schrijfster	
German	Zadie Smith	britische Schriftstellerin	
French	Zadie Smith	écrivaine britannique	

All entered languages

Statements

instance of human [edit](#)

▶ 3 references [+ add value](#)

image

sex or gender female [2 references](#)

country of citizenship United Kingdom [2 references](#)

Hidden Author Bias in Book Recommendation*

Savvina Daniil

s.daniil@cwi.nl

Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

Mirjam Cuper

mirjam.cuper@kb.nl

National Library of the Netherlands
The Hague, The Netherlands

Jacco van Ossenbruggen

jacco.van.ossenbruggen@cwi.nl

Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Delft University of Technology

The Hague, The Netherlands

Laura Hollink

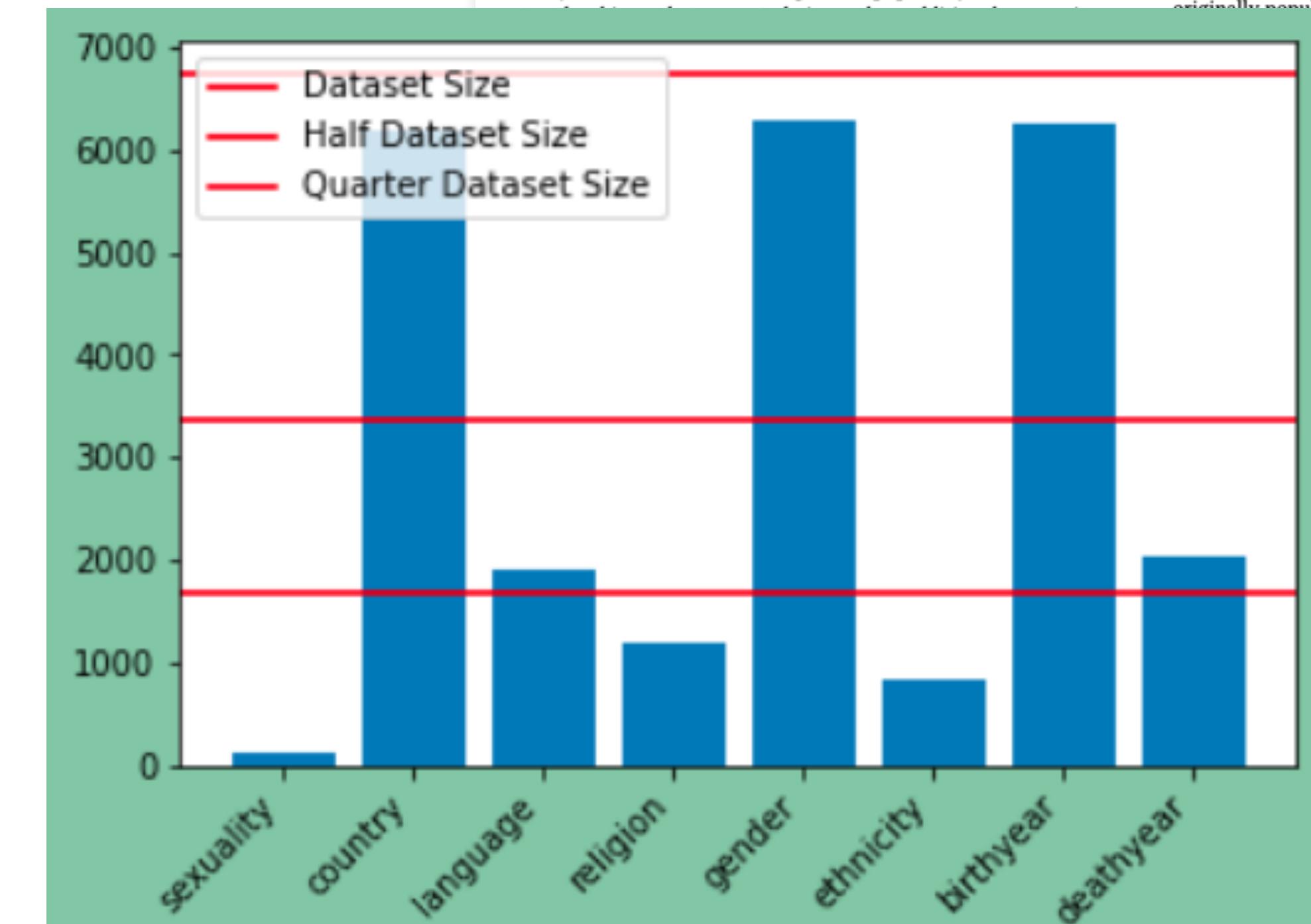
l.hollink@cwi.nl

Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

personal information of the user. However, by omitting sensitive information, it is possible for bias to manifest in such a system. Collaborative filtering approaches are still known to suffer from fairness related issues, like popularity bias. In this work, we argue that popularity bias often leads to unfair recommendations.

ABSTRACT

Collaborative filtering algorithms have the advantage of not requiring sensitive user or item information to provide recommendations. However, they still suffer from fairness related issues, like popularity bias. In this work, we argue that popularity bias often leads to unfair recommendations.



Savvina
Daniil

Author data in Wikidata

Zadie Smith (Q140052)

British novelist, essayist, and short-story writer

Zadie Adeline Smith

▼ In more languages

Configure

Language	Label	Description	Also known as
English	Zadie Smith	British novelist, essayist, and short-story writer	Zadie Adeline Smith
Dutch	Zadie Smith	Brits schrijfster	
German	Zadie Smith	britische Schriftstellerin	
French	Zadie Smith	écrivaine britannique	

All entered languages

Statements

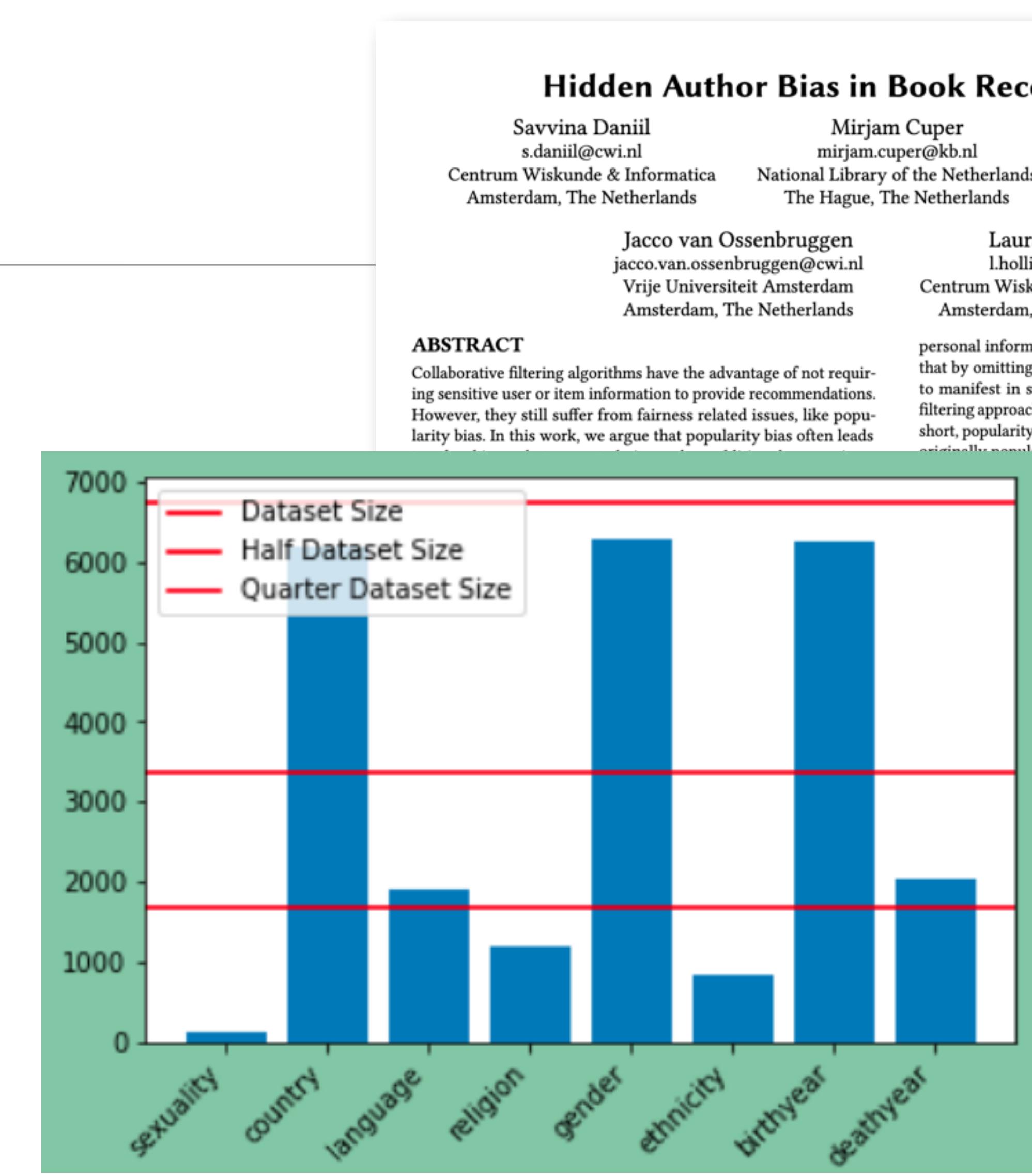
instance of human [edit](#)

▶ 3 references [+ add value](#)

image

sex or gender female [2 references](#)

country of citizenship United Kingdom [2 references](#)

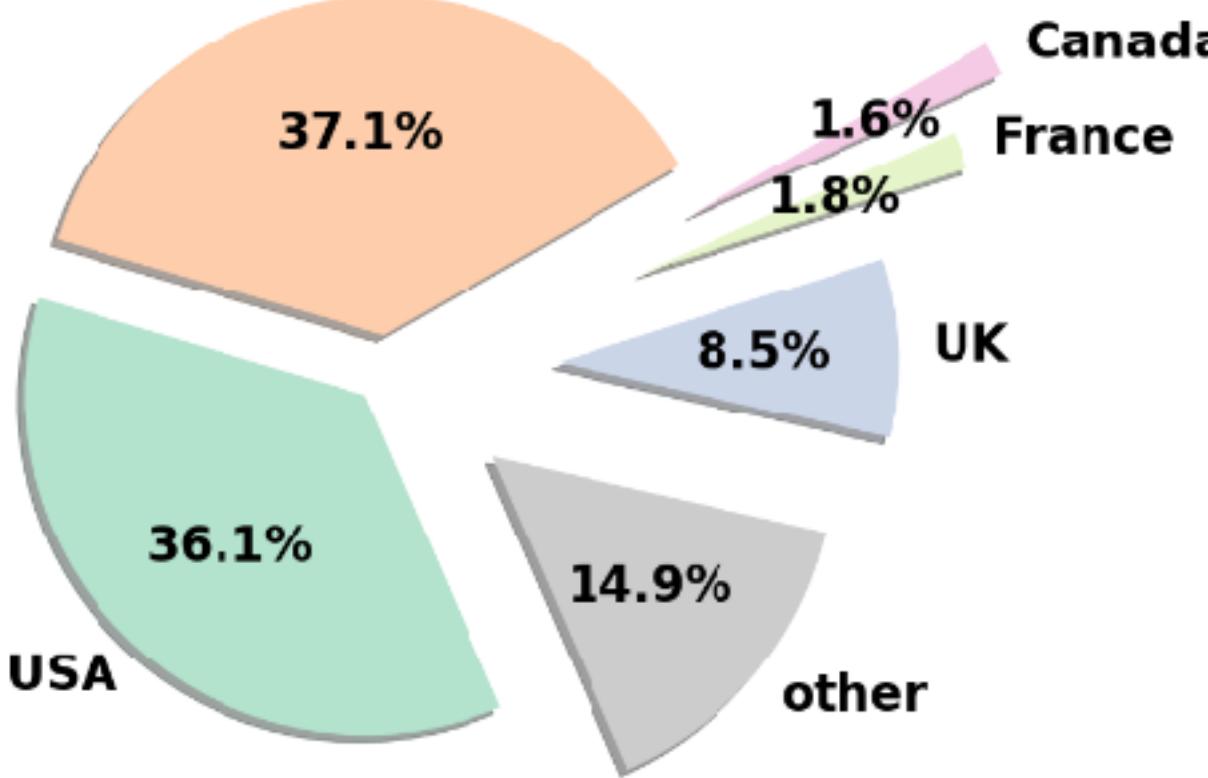


Savvina
Daniil

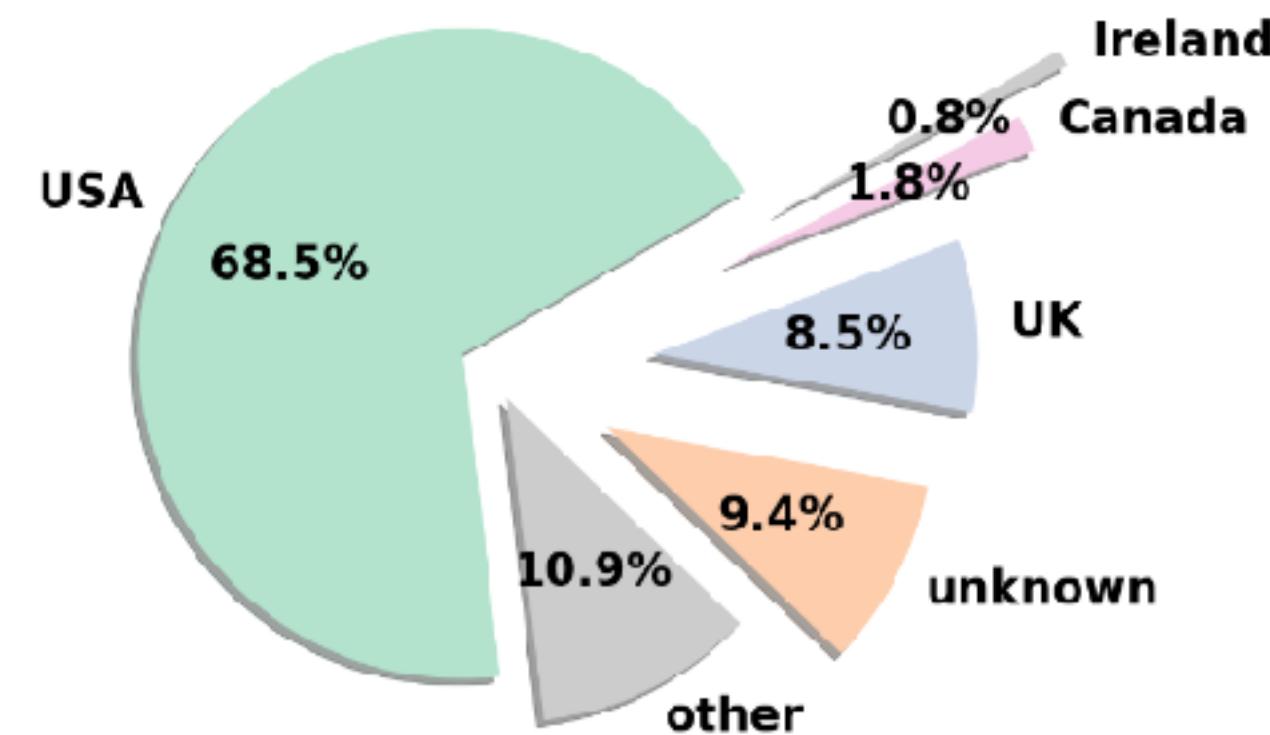
With that we were able to study not only popularity bias but also author-bias in book recommendation.

- First study is on country of citizenship-bias. We also have data on age and gender.

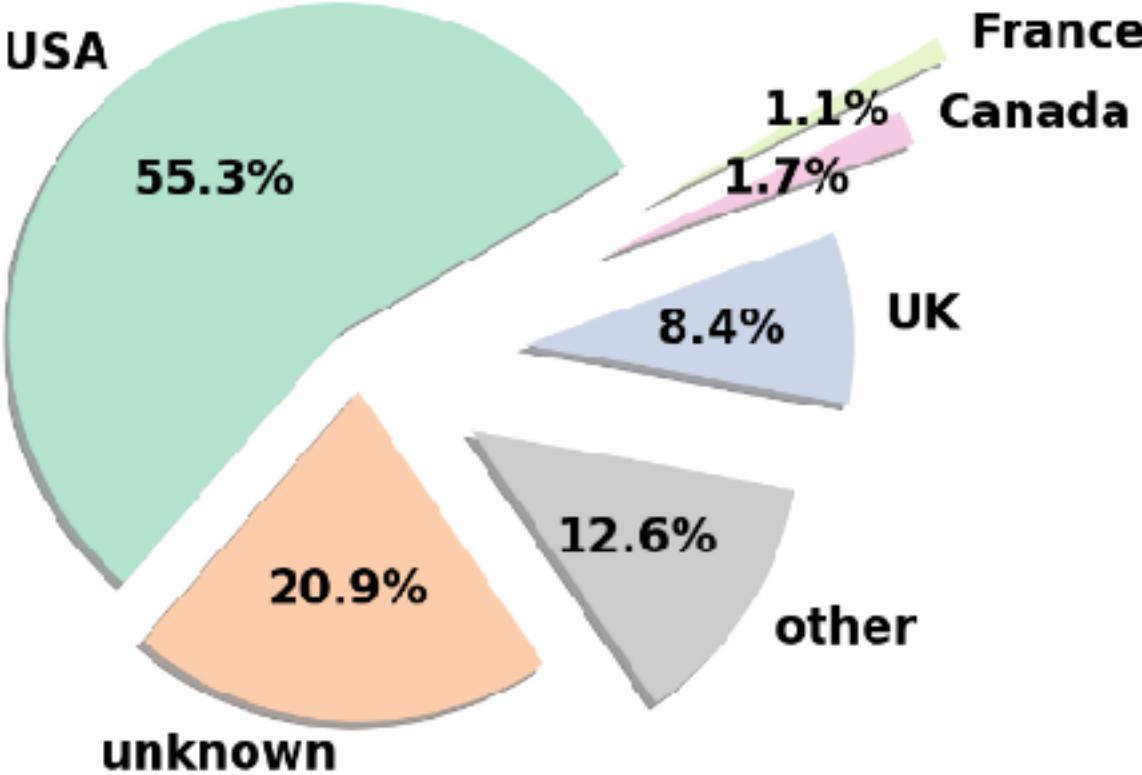
Measuring bias in datasets: the effect of dataset selection



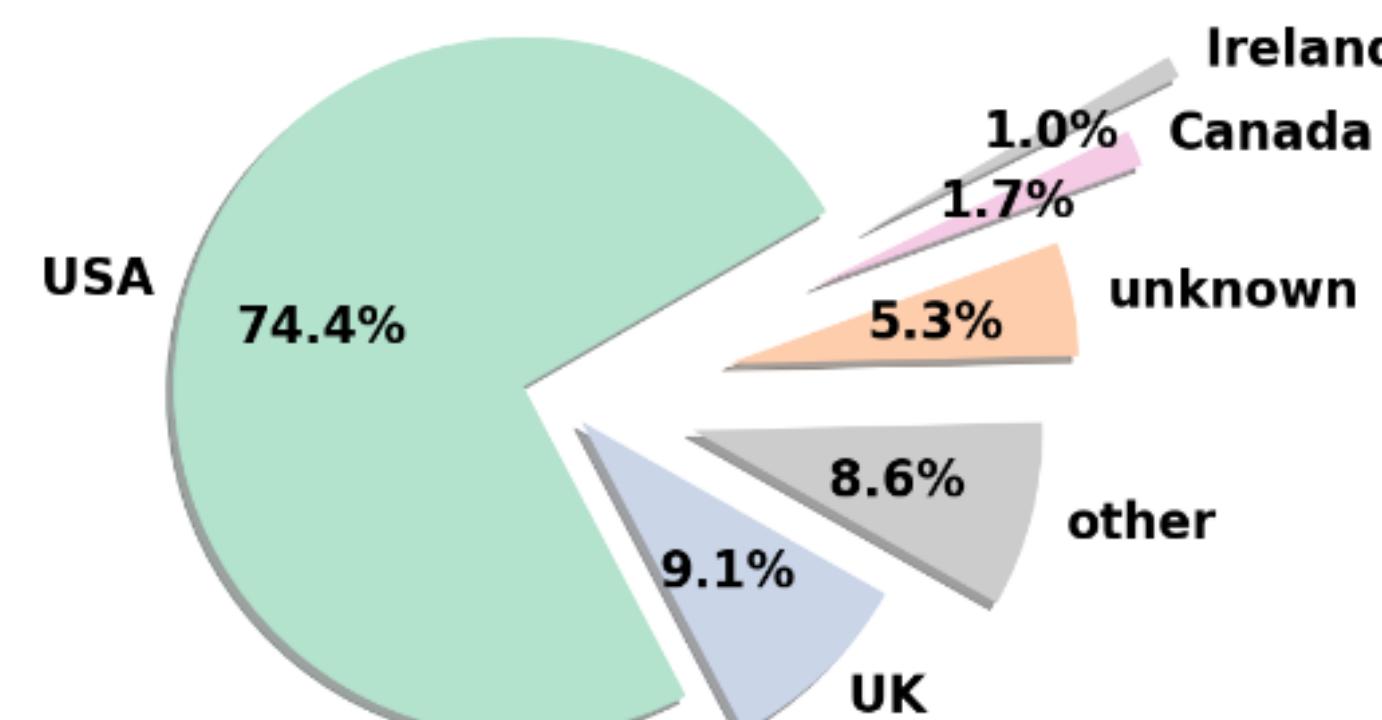
(a) Country distribution in the entire book dataset.



(b) Country distribution in the book dataset with Fairbook cut offs.



(a) Country distribution in the entire ratings dataset.



(b) Country distribution in the ratings dataset with Fairbook cut offs.

Hidden Author Bias in Book Recommendation*

Savvina Daniil
s.daniil@cwi.nl
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands

Mirjam Cuper
mirjam.cuper@kb.nl
National Library of the Netherlands
The Hague, The Netherlands

Jacco van Ossenbruggen
jacco.van.ossenbruggen@cwi.nl
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Laura Hollink
l.hollink@cwi.nl
Centrum Wiskunde & Info
Amsterdam, The Netherl



Savvina
Daniil

ABSTRACT

Collaborative filtering algorithms have the advantage of not requiring sensitive user or item information to provide recommendations. However, they still suffer from fairness related issues, like popularity bias. In this work, we argue that popularity bias often leads to other biases that are not obvious when additional user or item information is not provided to the researcher. We examine our hy-

personal information of the user that by omitting sensitive information to manifest in such a system. Collaborative filtering approaches are still known to be biased. In short, popularity bias is an algorithmic bias that occurs when originally popular in the training set items are recommended more often and thus have their popularity increased.

Measuring bias in the output of recommender algorithms

- Most algorithms (in fact, all but matrix factorization algorithms) over-represent U.S.-authored books in their recommendations.
- Algorithms that display a bias in favor of U.S. authors are also the ones that display a popularity bias.

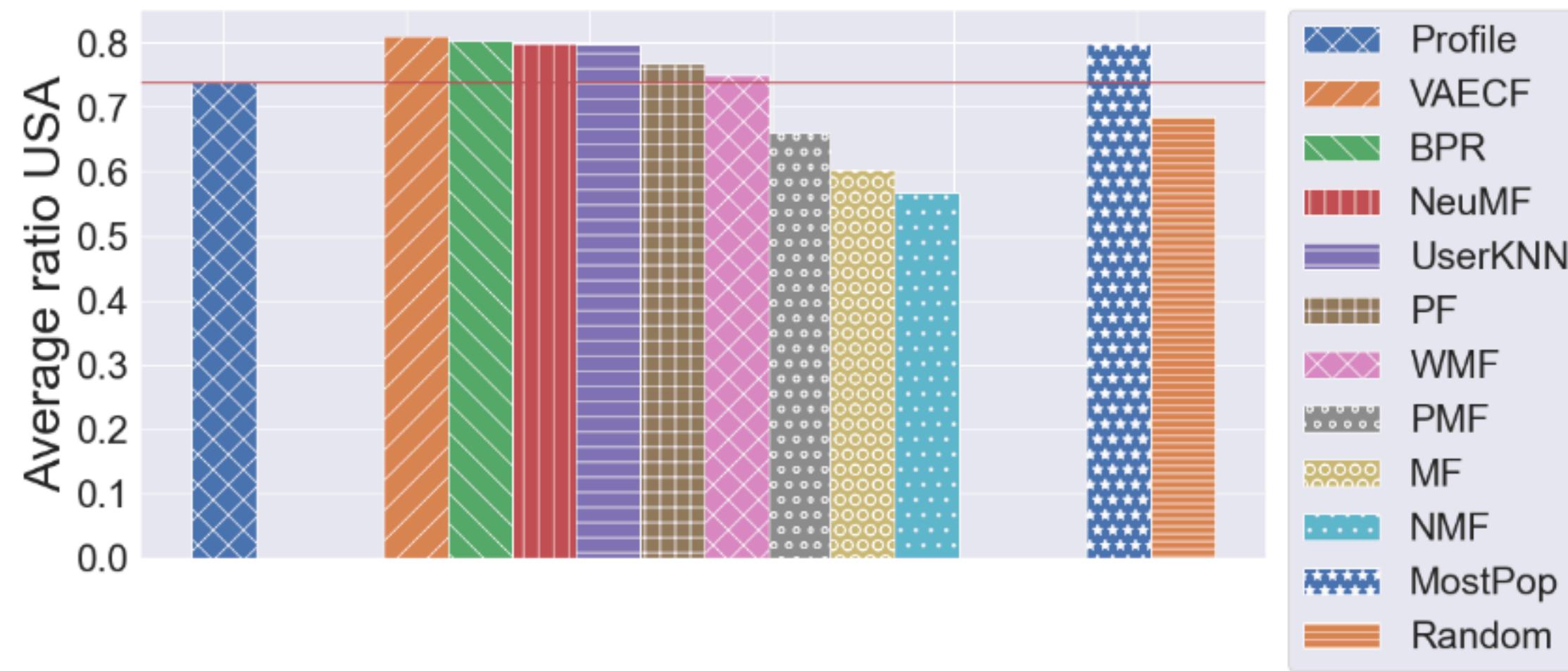


Figure 4: Average ratio of recommended books by every algorithm that were written by US citizens. Comparison with the average ratio of American-authored books in the users' profiles.

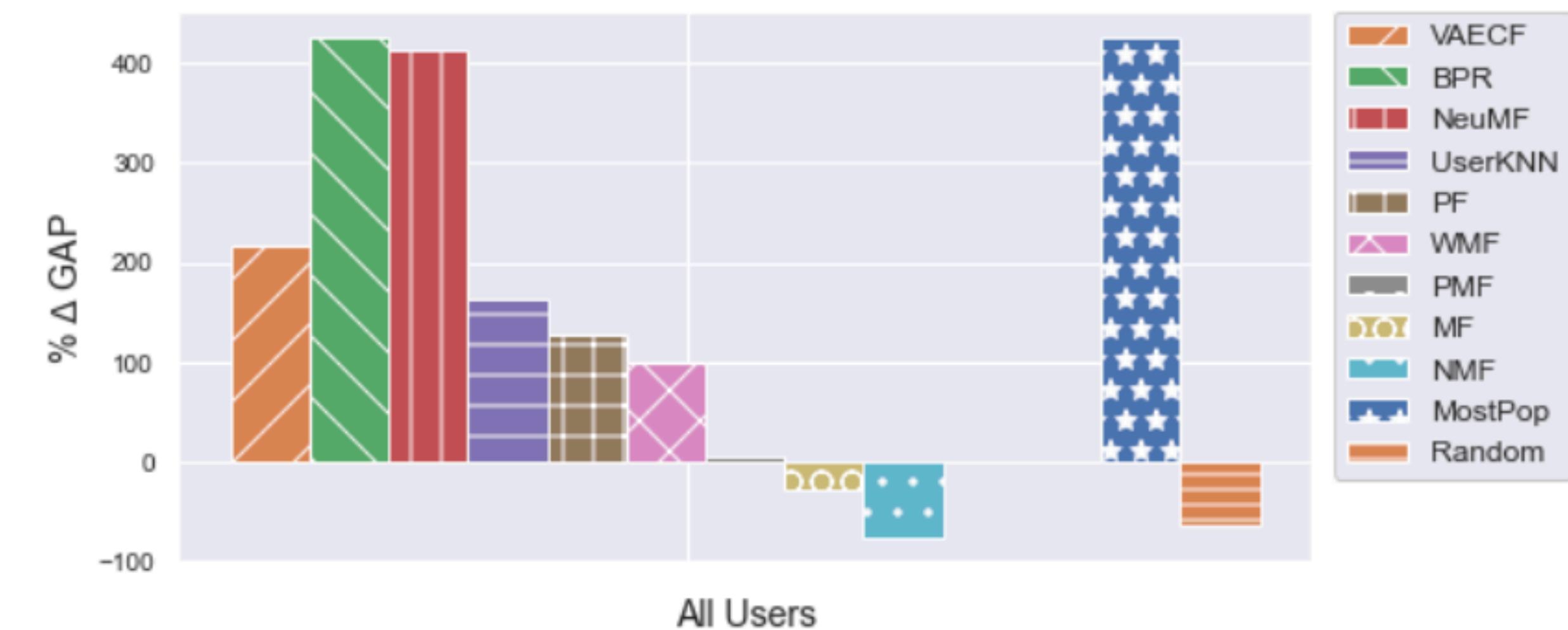
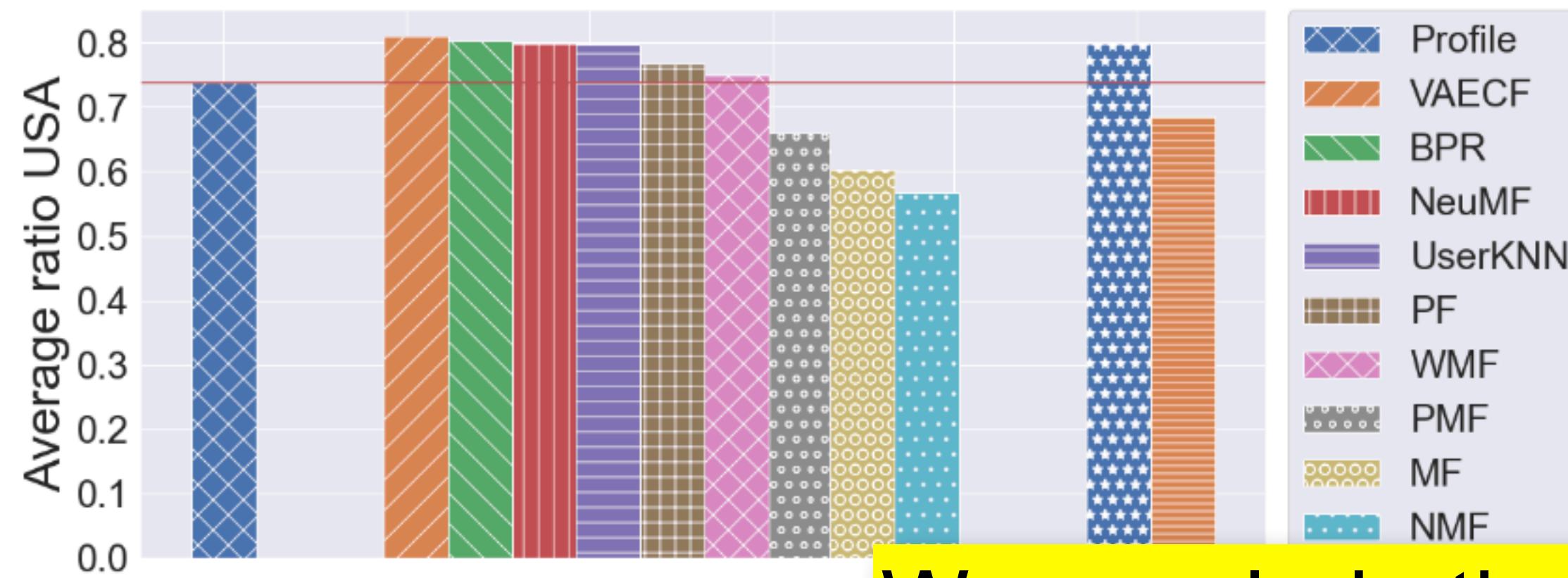


Figure 5: Relative increase in average popularity between profile and recommendation by every algorithm, averaged over all users.

Measuring bias in the output of recommender algorithms

- Most algorithms (in fact, all but matrix factorization algorithms) over-represent U.S.-authored books in their recommendations.
- Algorithms that display a bias in favor of U.S. authors are also the ones that display a popularity bias.



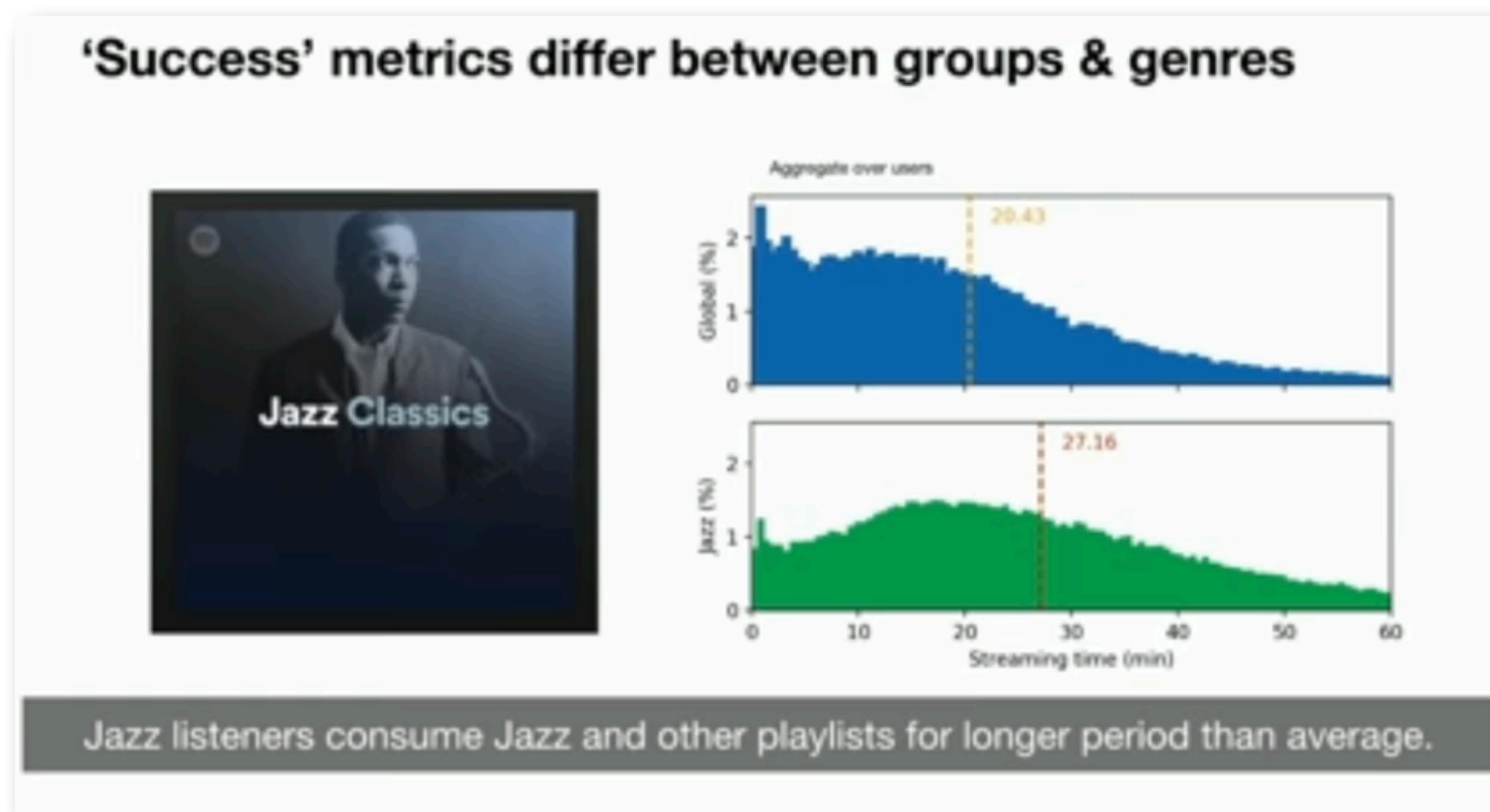
We conclude that the relatively accepted and harmless phenomenon of **popularity bias** leads to undesired forms of bias. We call this ‘hidden bias.’

Figure 4: Average ratio of recommended books by U.S. authors for each algorithm that were written by U.S. citizens, relative to the average ratio of American-authored books in the user profiles.

average popularity between every algorithm, averaged over all users.

Infer user information from interaction with the system

- Interaction signals user interest
- You can study e.g. if the system performs equally well for each user group.
- But: different groups might require different success metrics.



Slide from Henriette Cramer
on FAT* 2019 Translation Tutorial:
Challenges of incorporating algorithmic fairness
<https://www.youtube.com/watch?v=UicKZv93SOY>

Example study on defining user groups based on interaction with the historic newspaper archive of the National Library of the Netherlands

- We assume (faceted) queries and clicks on documents represent users' interests.
- We take subsets of the usage logs that show a particular user interest - and analyse behaviour within these subsets.

Bogaard, Tessel, Laura Hollink, Jan Wielemaker, Jacco van Ossenbruggen, and Lynda Hardman. "Metadata categorization for identifying search patterns in a digital library." *Journal of Documentation* (2018).

The screenshot shows the Delpher newspaper archive interface. At the top, there is a search bar with the word 'batavia' and a dropdown menu set to 'Kranten'. Below the search bar, the results are displayed. A sidebar on the left contains facets for 'Periode' (18e eeuw, 19e eeuw, 20e eeuw), 'Verspreidingsgebied' (Landelijk, Nederlands-Indië / Indonesië, Nederlandse Antillen, Regionaal/lokaal, Suriname, onbekend), and 'Soort bericht' (Advertentie). The main area displays three search results, each with a thumbnail, title, and details. The first result is an advertisement from 'De locomotief : Samarangsche handels- en advertentie-blad' dated 04-04-1901. The second result is a 'Familiebericht' from 'Het nieuws van den dag voor Nederlandsch-Indië' dated 30-09-1939. The third result is a 'BATAVIA' entry from 'Java government gazette' dated 26-06-1813. Each result has a star icon on the right.

Delpher

Kranten batavia

2.023.057 krantenartikelen gevonden voor: batavia × Nederlands-Indië / Indonesië ×

Nederlandse Antillen × Wissen

Sorteer op relevantie

Weergave

Periode

- 18e eeuw (7)
- 19e eeuw (751416)
- 20e eeuw (1271634)

Verspreidingsgebied

- Landelijk (690904)
- Nederlands-Indië / Indonesië (2019087)
- Nederlandse Antillen (3970)
- Regionaal/lokaal (696497)
- Suriname (8557)
- onbekend (3548)

Soort bericht

- Advertentie (980056)

facets

2.023.057 krantenartikelen gevonden voor: batavia × Nederlands-Indië / Indonesië ×

Nederlandse Antillen × Wissen

Sorteer op relevantie

Weergave

Advertentie
BATAVIA- ...

Krantentitel
De locomotief : Samarangsche handels- en advertentie-blad

Datum
04-04-1901

Familiebericht
BATAVIA ...

Krantentitel
Het nieuws van den dag voor Nederlandsch-Indië

Datum
30-09-1939

BATAVIA,
BATAVIA, ...

Krantentitel
Java government gazette

Datum
26-06-1813

search results

Example study on defining user groups based on interaction with the historic newspaper archive of the National Library of the Netherlands

- Different user interests connected to differ user behaviours:
 - **Users interested in WOII:** long sessions, many (complex) queries, clicks and downloads (indications of success?)
 - **Users interested in family announcements:** short sessions, few clicks and downloads, many unique queries, usage of quotes



- Recommendations to the Library (selection)
 - include a facet to easily select the WOII period
 - prioritise post-correction of OCR tools for articles from Surinam

Example study on defining user groups based on interaction with the historic newspaper archive of the National Library of the Netherlands

- Different user interests connected to differ user behaviours:
 - **Users interested in WOII:** long sessions, many (complex) queries, clicks and downloads (indications of success?)
 - **Users interested in family announcements:** short sessions, few clicks and downloads, many unique queries, usage of quotes



- Recommendations to the Library (selection)
 - include a facet to easily select the WOII period
 - prioritise post-correction of OCR tools for articles from Surinam

Alternative: clustering of user interests

- Results show 5 large clusters that are stable over time, plus several smaller, less stable clusters.
- Stable clusters show different user behaviour, as above.

T. Bogaard, L. Hollink, J. Wielemaker, L. Hardman, and J. van Ossenbruggen. Searching for Old News: User Interests and Behavior within a National Collection. In Proc of CHIIR '19.

Example study on defining user groups based on interaction with the historic newspaper archive of the National Library of the Netherlands

- Different user interests connected to differ user behaviours:
 - **Users interested in WOII:** long sessions, many (complex) queries, clicks and downloads (indications of success?)
 - **Users interested in family announcements:** short sessions, few clicks and downloads, many unique queries, usage of quotes



Note: present studies not about responsible AI.

Main point: user behaviour data can be used as proxy for personal data

- Recommendations to the Library (selection)
 - include a facet to easily select the WOII period
 - prioritise post-correction of OCR tools for articles from Surinam

Alternative: clustering of user interests

- Results show 5 large clusters that are stable over time, plus several smaller, less stable clusters.
- Stable clusters show different user behaviour, as above.

T. Bogaard, L. Hollink, J. Wielemaker, L. Hardman, and J. van Ossenbruggen. Searching for Old News: User Interests and Behavior within a National Collection. In Proc of CHIIR '19.

Biassed perspectives in data and metadata

For example, data that is created, collected, described from an outdated, colonial perspective.



Heritage collections have been compiled over long periods of time

 [Collections](#) [Explore](#) [Exhibitions](#) [Blog](#)

 [See this page on our new Europeana experience](#)

 [Return to Home](#) / [Results](#) / [Item](#)

We want your feedback on our new item page, use our feedback button to leave your comments.



Exotic visitors for London_x000D_H H

Exotic visitors for London_x000D_
H H the Mangku Negoro , reigning Prince of Surakarta (Java) with his wife
and child , who are now on their way to Holland . They will spend a few
weeks in London ._x000D_
25 June 1926

Created by
[TopFoto](#)

Screenshot. Europeana catalogue: <https://classic.europeana.eu/portal/en/record/2024904/>
https://www_topfoto_co_uk_asset_3022471

Detection is not straightforward -> Context is key

 [Collections](#) [Explore](#) [Exhibitions](#) [Blog](#)

[!\[\]\(86e76716e719e1794580a2abb382bc2a_img.jpg\) See this page on our new Europeana experience](#)

[!\[\]\(44298d4d999c16adc91d7accd5867a90_img.jpg\) Return to Home](#) / [Item](#)

We want your feedback on our new item page, use our feedback button to leave your comments.



Exotic cultivated mushrooms - Cinnamon

Exotic cultivated mushrooms - Cinnamon Cap _x000D_
credit: Marie-Louise Avery / thePictureKitchen / TopFoto

Created by
thePictureKitchen / EUFD

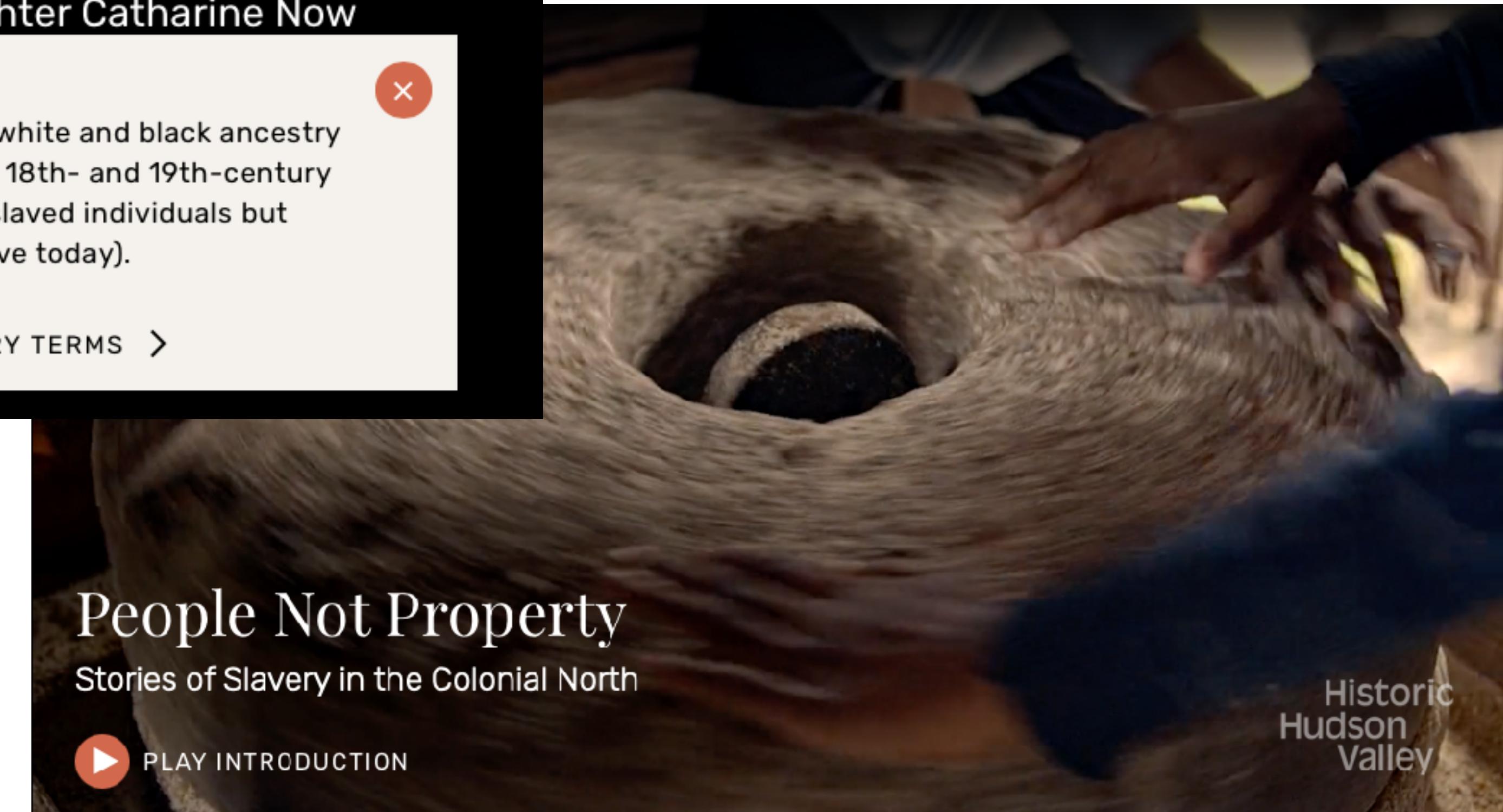
Screenshot. Europeana catalogue: <https://www.europeana.eu/en/item/2024904/>
https://www_topfoto_co_uk_asset_1827839

Different strategies to handle contentious terms

I Certify that Ruth a **Mullato** Girl a Slave born in my House was given by me to my Daughter Catharine Now the Wife of Pierre Val said Ruth was a Child of the said Pierre Val Hand this 28th Day of Geo Clinton

Mullato
A person of mixed white and black ancestry (commonly used in 18th- and 19th-century descriptions of enslaved individuals but considered offensive today).

[VIEW ALL GLOSSARY TERMS >](#)



Different strategies to handle contentious terms

The screenshot shows the Amsterdam Museum website. At the top left is a 'HOME - NIEUWS' link. The main title is 'AMSTERDAM MUSEUM GEBRUIKT TERM 'GOUDEN EEUW' NIET MEER' in large red letters, dated '12 SEPTEMBER 2019'. Below this is a paragraph explaining the museum's decision to stop using the term 'Gouden Eeuw'. The text states: 'Het Amsterdam Museum zal vanaf heden de term 'Gouden Eeuw' niet meer gebruiken om de periode van de 17e eeuw aan te duiden. Volgens het museum dekt de term de lading van de 17e eeuw niet. Het Amsterdam Museum is al geruime tijd actief om voor steeds meer mensen relevant te zijn en ziet het afstand doen van de term 'Gouden Eeuw' als stap om andere perspectieven op die tijd mogelijk te maken.' To the right of the text is a large red circular logo with 'AMSTERDAM MUSEUM' written around it. Below the logo is a red button with 'BOEK ONLINE TICKETS' and a search bar with the placeholder 'Zoeken...'. At the bottom right is a grey box containing social media icons (Twitter, Facebook, YouTube, Instagram, iZI) and the address 'KALVERSTRAAT 92 AMSTERDAM' along with opening hours.

HOME - NIEUWS

AMSTERDAM MUSEUM GEBRUIKT TERM 'GOUDEN EEUW' NIET MEER

12 SEPTEMBER 2019

Het Amsterdam Museum zal vanaf heden de term 'Gouden Eeuw' niet meer gebruiken om de periode van de 17e eeuw aan te duiden. Volgens het museum dekt de term de lading van de 17e eeuw niet. Het Amsterdam Museum is al geruime tijd actief om voor steeds meer mensen relevant te zijn en ziet het afstand doen van de term 'Gouden Eeuw' als stap om andere perspectieven op die tijd mogelijk te maken.

AMSTERDAM MUSEUM

BOEK ONLINE
TICKETS

Zoeken...

KALVERSTRAAT 92
AMSTERDAM

Dagelijks geopend
van 10:00 tot 17:00
uur

Amsterdam museum does not use the term 'golden age' any more

Screenshot. Amsterdam Museum gebruikt term 'Gouden Eeuw' niet meer. https://www.amsterdammuseum.nl/nieuws/gouden_eeuw

Different strategies to handle contentious terms



The screenshot shows a green header with the National Archives logo (a stylized 'N' inside a circle) and a 'Menu' button. Below the header, a green bar contains the word 'Home' in white. The main content area has a white background. A section titled 'Taalgebruik in onze archieven' (Language use in our archives) is visible. The text discusses the presence of potentially hurtful or discriminatory language in historical descriptions and the archive's choice to retain them. At the bottom, there is a footer with the text 'Screenshot 2020? Het Nationaal Archief. Taalgebruik in onze archieven.' and a link 'https://www.nationaalarchief.nl/taalgebruik-in-onze-archieven'.

Taalgebruik in onze archieven

Op onze website kunt u archieven doorzoeken met behulp van beschrijvingen en toegangen die vaak net zo oud zijn als de archieven zelf. De mogelijkheid bestaat dat u woorden tegenkomt die toen acceptabel waren, maar nu als kwetsend, racistisch of discriminerend ervaren kunnen worden.

Het Nationaal Archief kiest ervoor deze oorspronkelijke beschrijvingen te behouden, omdat deze een beeld geven van de tijd waarin ze zijn gemaakt of in de collectie zijn opgenomen. We onderzoeken de mogelijkheid om taal die in het verleden acceptabel en gangbaar waren, te verklaren en te voorzien van hedendaagse alternatieven.

Screenshot 2020? Het Nationaal Archief. Taalgebruik in onze archieven.
<https://www.nationaalarchief.nl/taalgebruik-in-onze-archieven>

Language in our archives

You may encounter words that were acceptable then, but can be experienced as hurtful, racist or discriminating now.

The National Archive chooses to keep the original descriptions, because...

Ongoing process

The screenshot shows the header of the National Archive website. The header features the 'national archief' logo with a stylized 'N' and 'A'. It includes a navigation menu with links to 'Home', 'Onderzoeken' (Research), 'Beleven' (Experience), 'Archiveren' (Archive), and 'Menu' with a three-line icon. Below the header, a green sidebar on the left has 'Home' and 'Taalgebruik' (Language use) underlined. The main content area has a green header bar with 'Home' underlined. The main title 'Taalgebruik in onze archieven' is displayed in large bold letters. A paragraph discusses the use of language in inventories, mentioning that some words were acceptable then but may be hurtful or discriminatory now. The text is in Dutch.

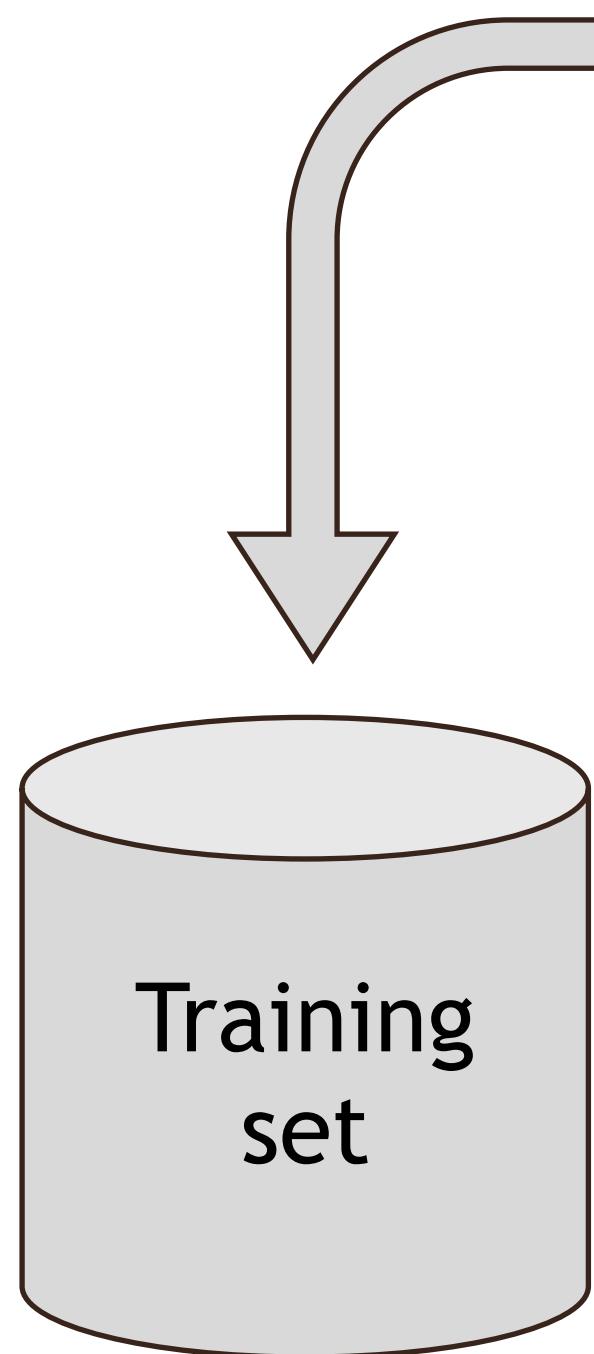
Language in our archives

You may encounter words that were acceptable then, but can be experienced as hurtful, racist or discriminating now.

The National Archive currently investigates the possibilities to adapt, explain or replace this language in the inventories.

Screenshots: Screenshot 2023. Het Nationaal Archief. Taalgebruik in onze archieven.
<https://www.nationaalarchief.nl/taalgebruik-in-onze-archieven>

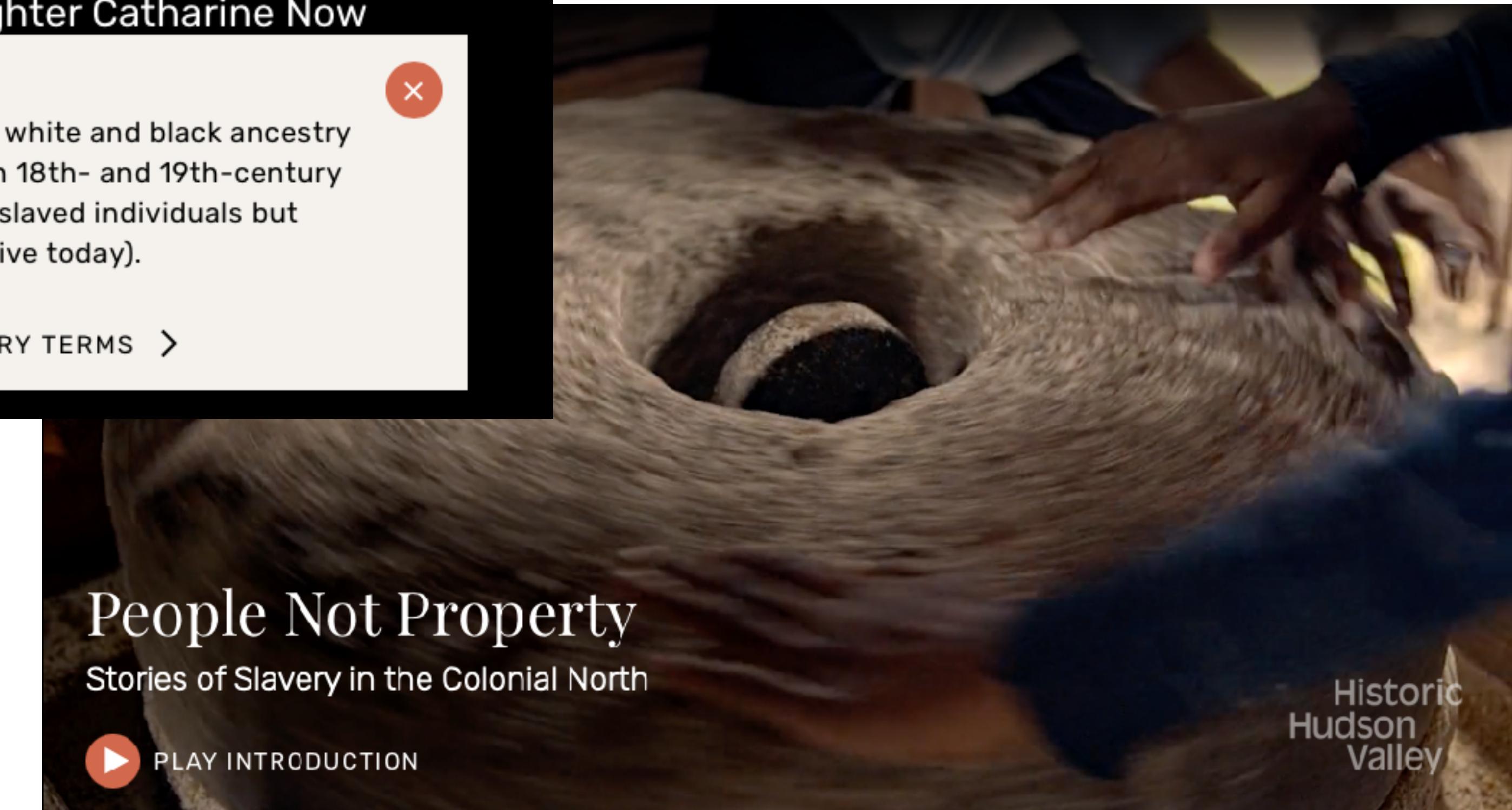
Biassed terminology might have consequences outside the archive



I Certify that Ruth a **Mullato** Girl a Slave born in my House was given by me to my Daughter Catharine Now the Wife of Pierre Val... said Ruth was a Child of the said Pierre Val... Hand this 28th Day of Geo Clinton

Mullato
A person of mixed white and black ancestry (commonly used in 18th- and 19th-century descriptions of enslaved individuals but considered offensive today).

[VIEW ALL GLOSSARY TERMS >](#)



We developed a knowledge graph of contentious terminology

Based on domain expert knowledge about contentious words in the cultural sector



An Unfinished Guide
to Word Choices
in the Cultural Sector

GLOSSARY OF TERMS

Exotic

HISTORY, USE & POSSIBLE SENSITIVITIES

This term is derived from the Ancient Greek word "exōtikós," literally meaning "from the outside." It entered the Dutch language with the meaning of foreign/alien, which it still has today. The term has become intertwined with ideas about the (racialized and sexualized) Other.

The term "exotic" is commonly used to describe plants and animals, but is also used for people (usually people of color), where it has a connotation of being different from the norm, especially in reference to appearance and name (for example "what an exotic name!"). Sometimes it has a sensual connotation.

SUGGESTIONS

- Applicable when referring to plant and animal species. It is, however, contested to use the term to describe people.

ESWC 2023

A Knowledge Graph of Contentious Terminology for Inclusive Representation of Cultural Heritage

Andrei Nesterov¹ [0000-0001-9697-101X], Laura Hollink¹ [0000-0002-6865-0021],
Marieke van Erp² [0000-0001-9195-8203], and Jacco van Ossenbruggen³ [0000-0002-7748-4715]

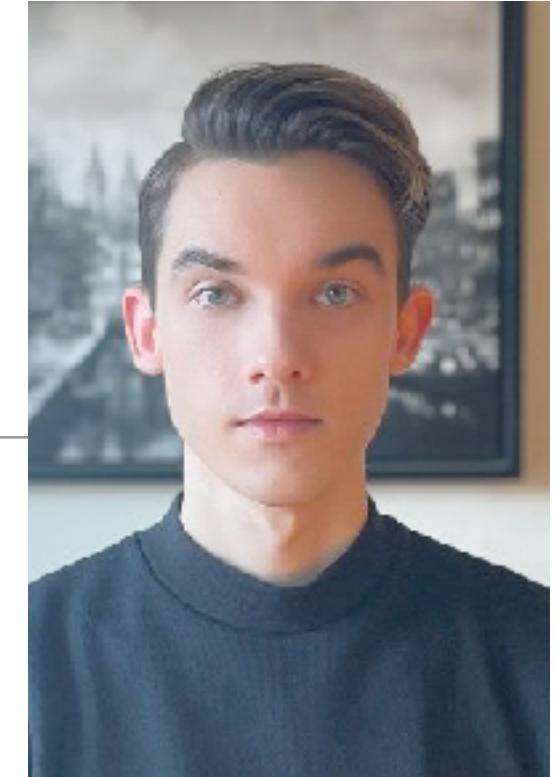
¹ Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
{nesterov,hollink}@cwi.nl

² KNAW Humanities Cluster, Amsterdam, The Netherlands
marieke.van.erp@dh.huc.knaw.nl

³ VU University Amsterdam, Amsterdam, The Netherlands
jacco.van.ossenbruggen@vu.nl

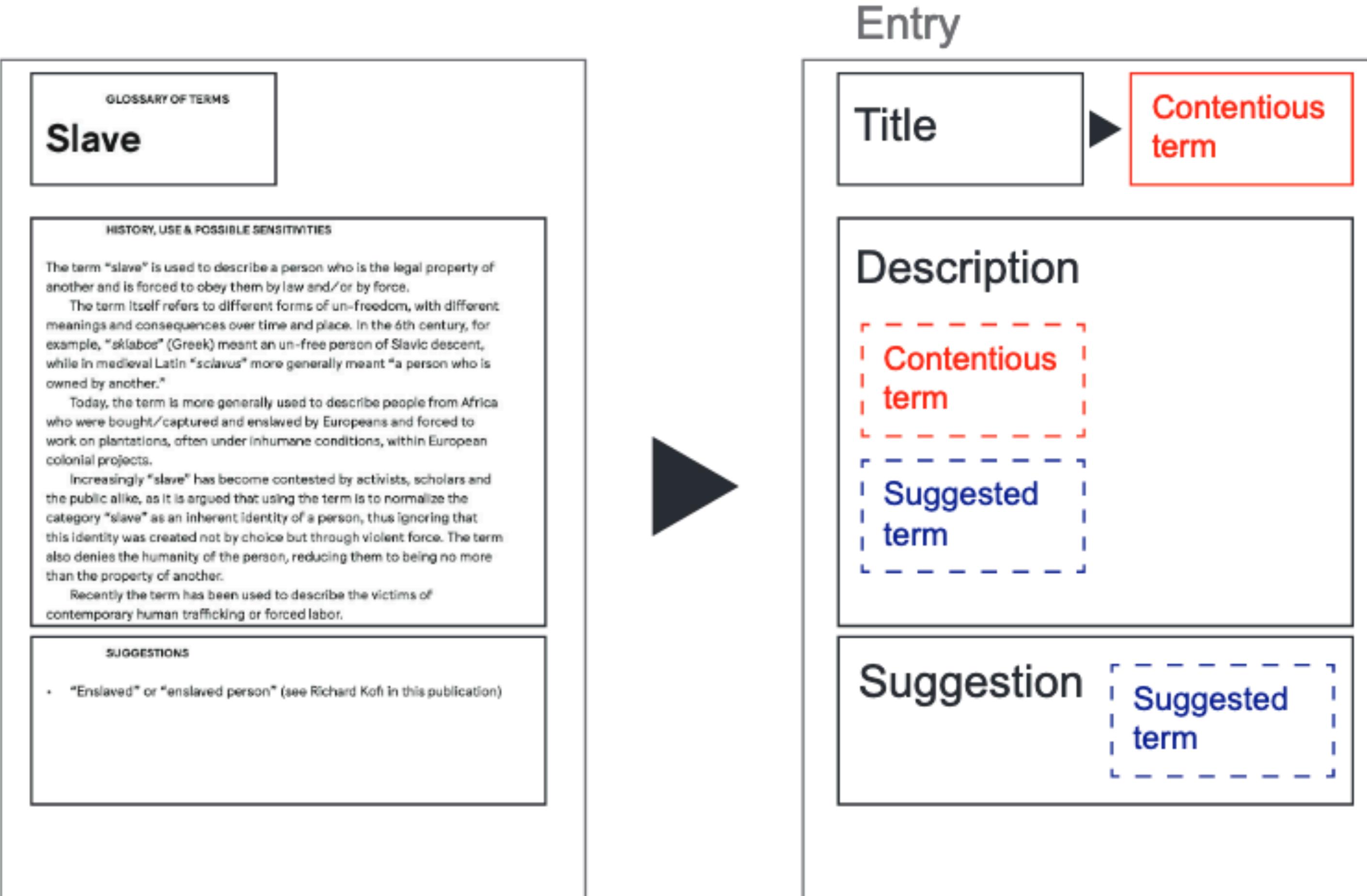
Abstract. Cultural heritage collections available as linked open data (LOD) may contain harmful stereotypes about people and cultures, for example, in outdated textual descriptions of objects. Galleries, libraries, archives, and museums (GLAM) have suggested various approaches to tackle potentially problematic content in digital collections. However, the domain expertise and discussions about words and phrases used in

Modest, Wayne & Lelijveld, Robin (editors) 2018.
Words Matter, Work in Progress I. National Museum of
World Cultures. <https://www.materiaculture.nl/en/publications/words-matter>



Andrei
Nesterov

We developed a knowledge graph of contentious terminology



We developed a knowledge graph of contentious terminology

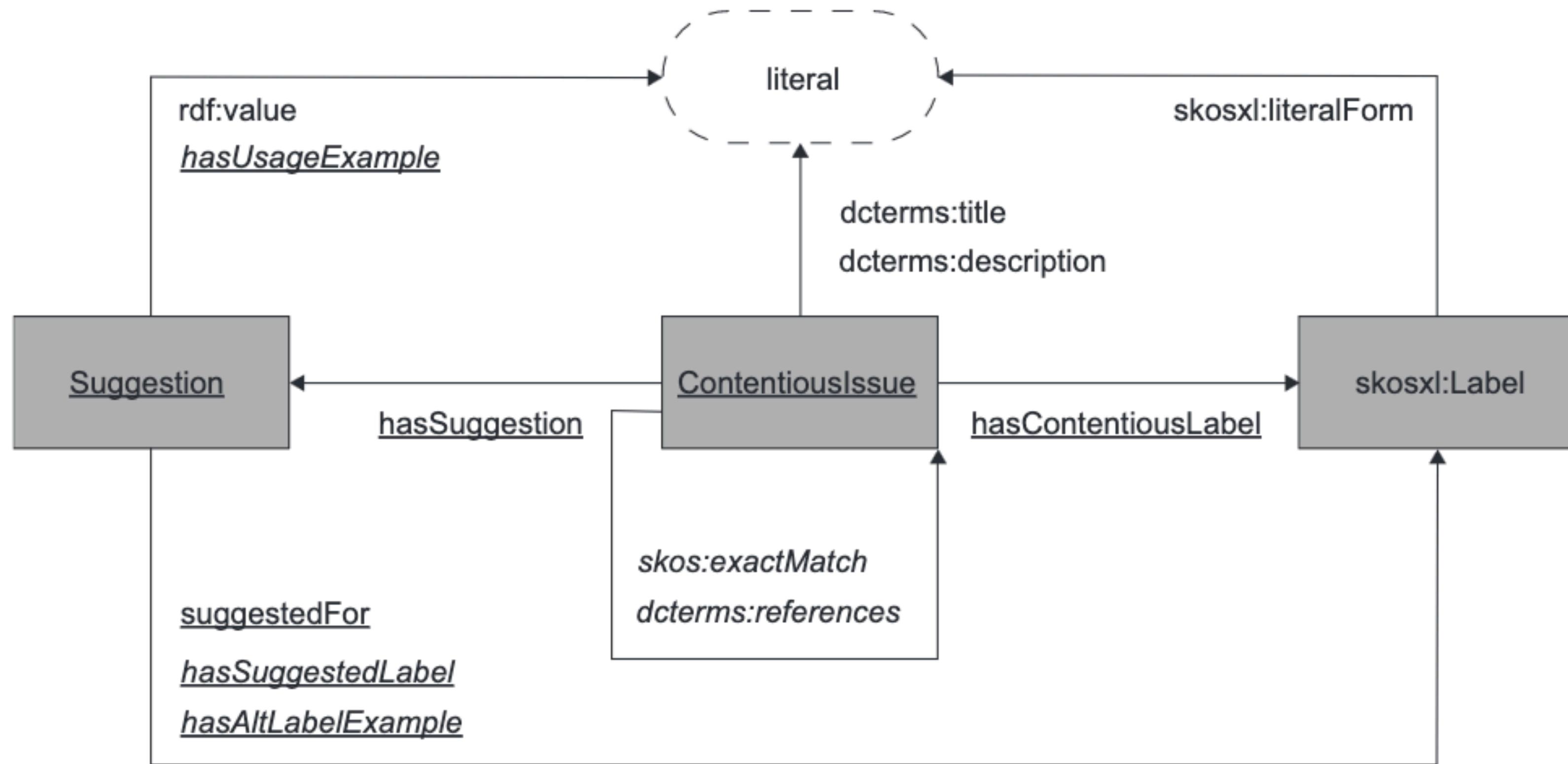


Fig. 2. The knowledge graph schema with custom classes and properties underlined. The italicized properties are optional.

- 75 English and 83 Dutch potentially contentious terms
- Linked to suggestions, explanations, examples
- Linked to other LOD resources:
 - WordNet
 - Wikidata
 - Getty AAT
 - NMVM Thesaurus

The resulting resource has been made openly available with a CC BY-SA 4.0 license following FAIR practices.

<https://github.com/cultural-ai/wordsmatter/>

We developed an annotated text corpus of contentious terminology

"De vrouw tegenover hem was nog maar een meisje, twintig naar schatting.
Een nauwsluitend zwart manteltje en rok, witte satijnen blouse, een kleine,
chique, zwarte toque, modieus gedragen op één oor.
Ze had een mooi, **exotisch** gezichtje, mat-witte huid, grote bruine oogen,
git-zwart haar.
Ze rookte een sigaret in een langen houder.
Haar gemanicuurde handen hadden donkerroode nagels."

8 annotators per sample

4: contentious, 3: not contentious, 1: I don't know



Andrei
Nesterov



Ryan
Brate

K-CAP 2021

Capturing Contentiousness: Constructing the Contentious Terms in Context Corpus

Ryan Brate^{*}
KNAW Humanities Cluster
Amsterdam, The Netherlands
ryan.brate@dh.huc.knaw.nl

Andrei Nesterov^{*}
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
nesterov@cwi.nl

Valentin Vogelmann
KNAW Humanities Cluster
Amsterdam, The Netherlands
valentin.vogelmann@dh.huc.knaw.nl

Jacco van Ossenbruggen
VU University Amsterdam
Amsterdam, The Netherlands
jacco.van.ossenbruggen@vu.nl

Laura Hollink
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
l.hollink@cwi.nl

Marieke van Erp
KNAW Humanities Cluster
Amsterdam, The Netherlands
marieke.van.erp@dh.huc.knaw.nl

ABSTRACT
Recent initiatives by cultural heritage institutions in addressing outdated and offensive language used in their collections demonstrate the need for further understanding into when terms are problematic or contentious. This paper presents an annotated dataset of 2,715 unique samples of terms in context, drawn from a historical newspaper archive, collating 21,830 annotations of contentiousness from expert and crowd workers.

We describe the contents of the corpus by analysing inter-rater agreement and differences between experts and crowd workers. In addition, we demonstrate the potential of the corpus for automated detection of contentiousness. We show that a simple classifier applied to the embedding representation of a target word provides a better than baseline performance in predicting contentiousness. We find that the term itself and the context play a role in whether a term is considered contentious.

CCS CONCEPTS

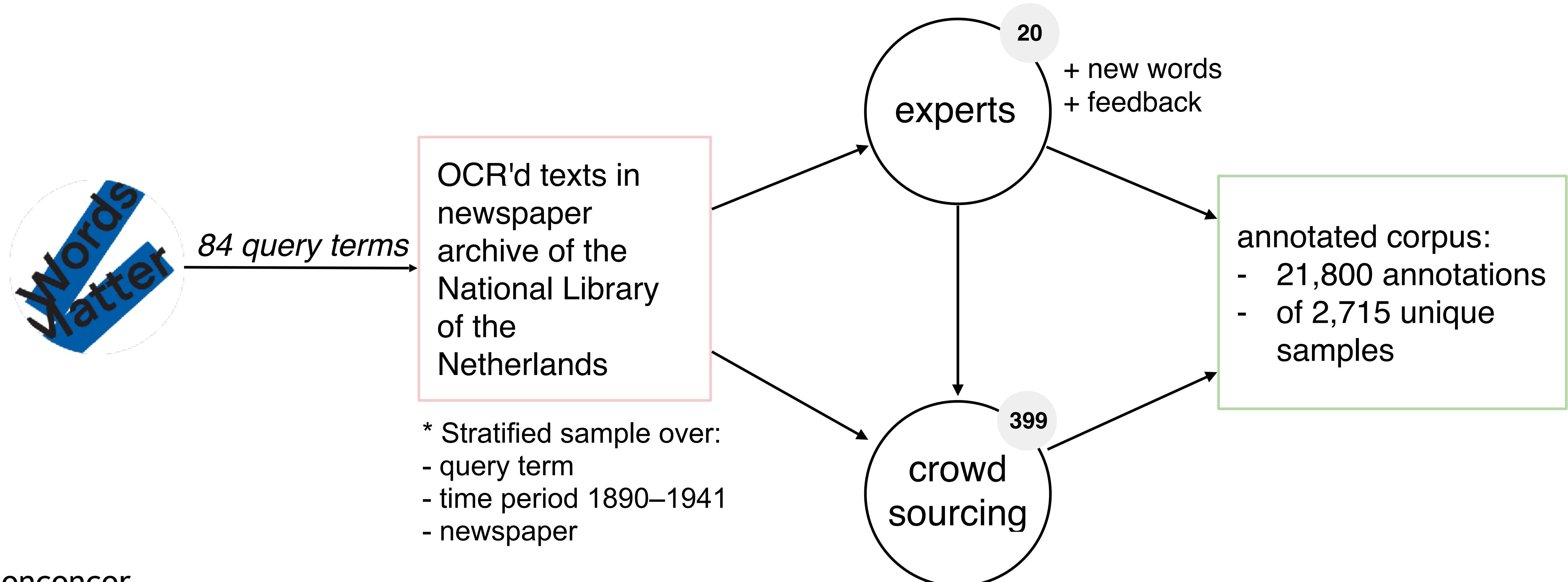
NOTE: Some examples in this paper may be shocking or offensive. They are provided as illustration or explanation of the work and do not reflect the opinion of the authors or their organisations.

1 INTRODUCTION AND MOTIVATION

Cultural heritage institutions harbour vast collections that have often been compiled over long periods of time. Collection and documentation practices therefore reflect the cultural and social norms of the various time periods during which they were compiled. As a result, they may contain terms that are inappropriate in modern society. An example of a contentious term that we find in historical documents is 'half-blood' to denote people of mixed descent. Nowadays, this term is considered offensive when discussing people, although it is still acceptable when discussing for example animals or plants.

Many institutions recognise the problem of outdated language in their collections. For example, the Amsterdam Museum published a statement in 2019 that they would not use the term 'Golden Age'

We developed an annotated text corpus of contentious terminology



Conconcor
(potentially contentious words, text snippets in which they occur,
annotators' responses, and metadata of the newspaper articles)
is available from <https://github.com/cultural-ai/ConConCor>

Large scale manual annotation of contentiousness: lessons learned

Inter-rater agreement is low:

- $\alpha = 0.54$ among experts
- $\alpha = 0.31$ for crowd annotators

but can be improved (to $\alpha = 0.50$)

by filtering out underperforming annotators:

- using control questions?
- using pairwise agreement between annotators?

Large scale manual annotation of contentiousness: lessons learned

Inter-rater agreement is low:

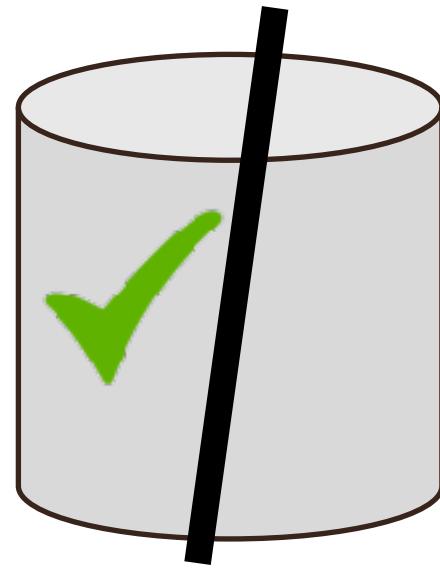
- $\alpha = 0.54$ among experts
- $\alpha = 0.31$ for crowd annotators

but can be improved (to $\alpha = 0.50$)

by filtering out underperforming annotators:

- using control questions?
- using pairwise agreement between annotators?

Multiple annotators helps to get reliable data:
on half of the samples,
over 80% of
annotators agreed
with each other.



Large scale manual annotation of contentiousness: lessons learned

Inter-rater agreement is low:

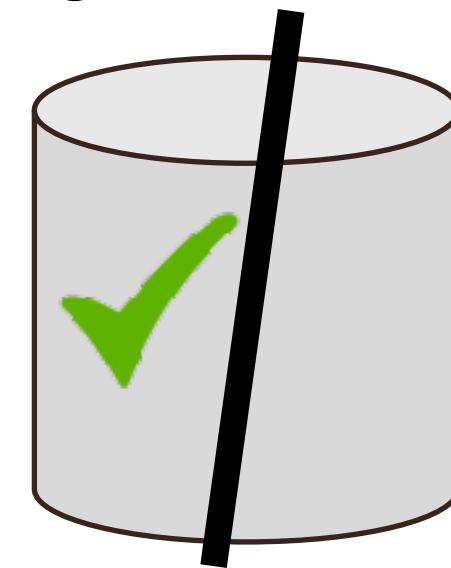
- $\alpha = 0.54$ among experts
- $\alpha = 0.31$ for crowd annotators

but can be improved (to $\alpha = 0.50$)

by filtering out underperforming annotators:

- using control questions?
- using pairwise agreement between annotators?

Multiple annotators helps to get reliable data:
on half of the samples,
over 80% of
annotators agreed
with each other.



Context is necessary to judge contentiousness:
most words are sometimes contentious and
sometimes noncontentious, depending on the
context.

Large scale manual annotation of contentiousness: lessons learned

Inter-rater agreement is low:

- $\alpha = 0.54$ among experts
- $\alpha = 0.31$ for crowd annotators

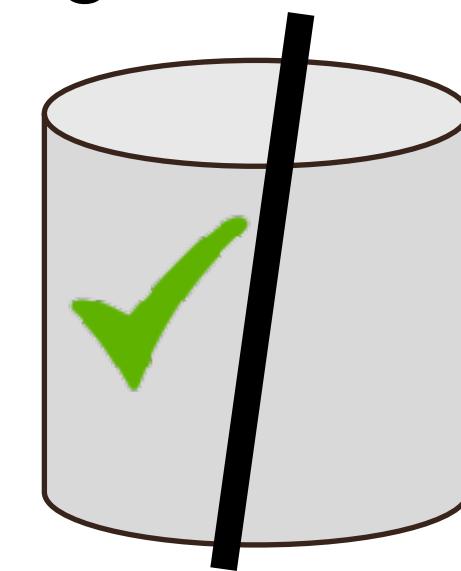
but can be improved (to $\alpha = 0.50$)

by filtering out underperforming annotators:

- using control questions?
- using pairwise agreement between annotators?

First experiments demonstrate that the corpus can be used to train a model to predict contentiousness
baseline: balanced accuracy = [0.54-0.55]
model: balanced accuracy = [0.76-0.78]

Multiple annotators helps to get reliable data:
on half of the samples,
over 80% of
annotators agreed
with each other.

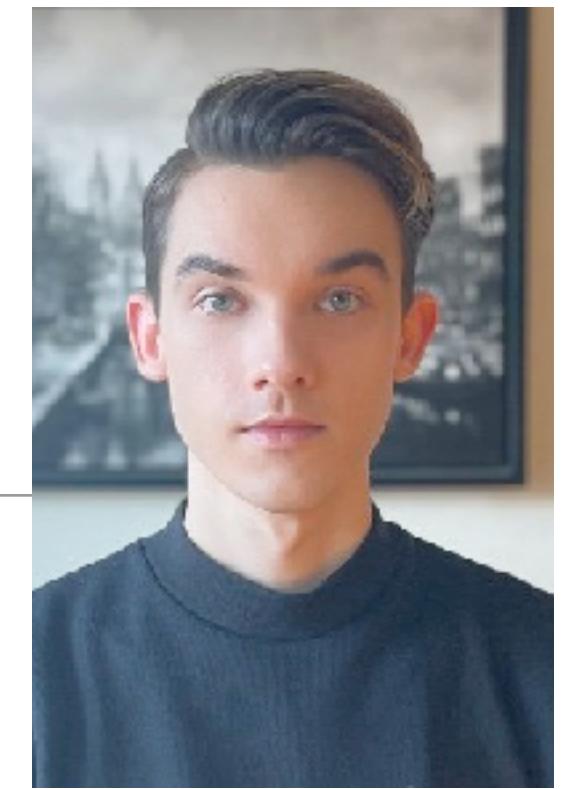


Context is necessary to judge contentiousness:
most words are sometimes contentious and
sometimes noncontentious, depending on the
context.

Ongoing work: We study how contentious terms are used in Linked Open Data



→ LOD: Wikidata, The Getty Art & Architecture Thesaurus, WordNet (English and Dutch)



Andrei
Nesterov

Visit the main page

Half-breed (Q17144151)

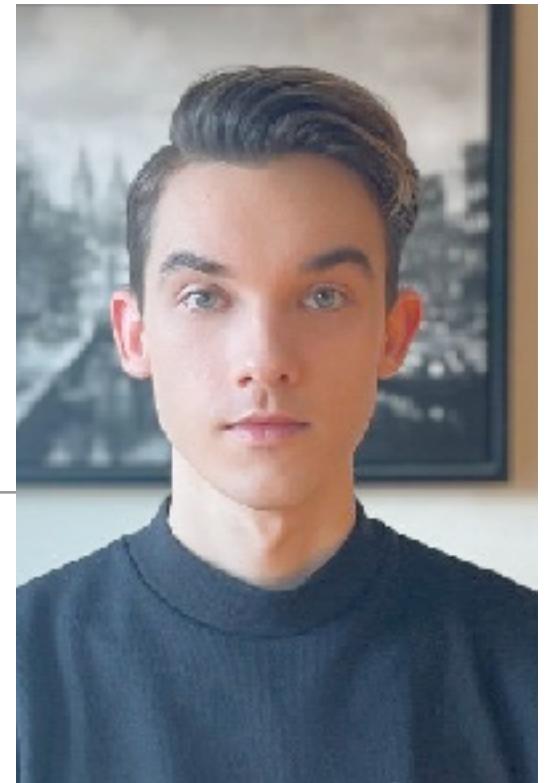
obsolete term for mixed Native American and European ancestry

▼ In more languages

Configure

Language	Label	Description
English	Half-breed	obsolete term for mixed Native American and European ancestry
Dutch	No label defined	No description defined
Croatian	No label defined	No description defined
Italian	No label defined	No description defined

Ongoing work: We study how contentious terms are used in Linked Open Data



Results:

- Contentious terms are used on a large scale **in preferred labels, alternative labels and descriptions.**
- The LOD community is trying to address the issue in various ways:
 - Some LOD datasets mention it in their **guidelines** for editors
 - All LOD datasets contain **properties** that can be used to mark labels as offensive/slur/outdated, etc.
 - In all LOD datasets, we found cases where editors choose **words** to flag something as offensive/slur/outdated, etc.
 - All of the above methods are used sparsely and inconsistently.

Potentially large effects outside single LOD resources:

[https://babelnet.org/synset?
id=bn%3A00037547n&orig=h
oseksuele&lang=NL](https://babelnet.org/synset?id=bn%3A00037547n&orig=hoseksuele&lang=NL)

Thank you!

[https://www.cwi.nl/en/groups/human-centered-data-analytics/
cultural-ai.nl/
aim4dem.nl/](https://www.cwi.nl/en/groups/human-centered-data-analytics/cultural-ai.nl/)