x

# Learning to Diversify from Human Judgments: Research Directions and Open Challenges

**Emily Denton***
Google Research
dentone@google.com

**Alex Beutel**
Google Research
alexbeutel@google.com

**Hansa Srinivasan***
Google Research
hansas@google.com

**Tulsee Doshi**
Google Research
tulsee@google.com

**Dylan Baker**
Google Research
dylanbaker@google.com

**Ed H. Chi**
Google Research
edchi@google.com

**Jilin Chen**
Google Research
jilinc@google.com

*Both authors contributed equally.

*CHI'20,*, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

## Abstract

Algorithmic ranking and retrieval systems have enormous influence over online media consumption, but run the risk of reflecting and reinforcing social biases. In this work, we outline a proposed research direction aimed at developing algorithmic techniques to increase diversity in such systems and pose open questions and challenges that arise from considering this problem in the realm of image sets.

## Author Keywords

Diversity; Fairness; Ranking.

## CCS Concepts

**•Information systems → Information retrieval diversity;**
**•Human-centered computing → Empirical studies in HCI;**

## Introduction

Data driven algorithmic systems increasingly serve much of the media individuals come into contact with online, with the goal of improving user satisfaction. However, ranking and retrieval systems have received criticism for reflecting and reinforcing underlying social biases, with harmful consequences, specifically for marginalized or underprivileged groups [15, 4, 10]. This motivates the development of algorithmic techniques for improving socially salient diversity in ranking and retrieval systems.

Existing frameworks for improving demographic diversity in ranking and retrieval systems tend to focus on distributional properties of the result set with respect to a given set of social categories (e.g. race or gender) associated with each data instance. The categorical structure, and the process of associating data instances with category values is typically de-centered in these frameworks. In this work, we motivate and propose a framework that moves beyond rigid, discrete, and ascribed categories and instead relies on subjective judgements from a large pool of diverse individuals. We consider a specific instantiation of this approach applied to image sets, without attribute annotations for the images and detail key challenges and directions for future work.

## Background and Motivation

Film, television, advertising, and other media play a significant role in shaping individuals understanding of their own social identities [12, 6] and how individuals and social groups are perceived by others [1, 7]. Underrepresentation, stereotyped portrayals, or otherwise biased representations of social groups can in turn deepen social inequalities by reinforcing sexist, racist, heteronormative, classist, ablist, or other oppressive worldviews [11].

Automated algorithmic retrieval, ranking, and curation tools increasingly structure the content individuals see and interact with online. Thus, these algorithmic systems also play a significant role in structuring cultural representations of individuals and groups. Just as traditional forms of media can reproduce or reify existing social stereotypes, so too can systems serving online content. Several recent studies have examined how such systems can perpetuate existing racial and gender stereotypes [15, 4, 10]. In recommender systems, the problem of filter bubbles describes the information isolation that can occur with personalization [16], harming the diversity of media presented to users, as occurred in an examination of book recommendation [5].

There is a significant literature studying ‘diversity’ in ranking and retrieval systems [20, 17, 13, 8]. However, as [14] points out, this literature tends to strip the term ‘diversity’ of its social meaning and instead uses it to reference arbitrary heterogeneity. In this work, we follow [14] and use diversity to reference variety along social axes that structure differential access to power and privilege in the world today.

Several recent works have developed ranking and retrieval frameworks that seek certain distributional properties across discrete demographic groups associated with each data instance in a set. For example, statistical fairness definitions such as demographic parity and equality of odds have been adapted for ranking and retrieval [19, 3, 21, 22, 2] and [14] introduces metrics from social choice to score the diversity of image sets.

Previous methods for improving diversity in ranking or retrieval tend to take the categories they seek diversity with respect to as given. Following algorithmic fairness work, the categories are framed in terms of sensitive attributes that can take one of k values and the annotations are frequently assumed to be given (although some works allow for unknown or hidden attributes, e.g. [2]). While this simplistic framing has notable utility in certain circumstances, in others it can be ineffective or even actively harmful. For example, increasing gender diversity in a ranking or retrieval system serving visual media hinges on challenging questions of representation. In this case, reducing race, gender, and other complex and situated social constructs to one-dimensional, third-party or algorithmically ascribed, categorical assignments should be avoided. Inaccuracies always abound when personally held identities are inferred and the nuance and contextuality of identity may be lost.

Furthermore, for some axes relevant to diversity, e.g. gender, labelling data instances via crowd-sourced or algorithmic inferences can be an act of violence in and of itself [18, 9].

## Proposed research direction

In our work, we aim to move beyond ascribing rigid, fixed, and one-dimensional categories of race, gender, etc. to data instances. Given that we seek to measure diversity of representation along social axes salient to power and privilege, we cannot sidestep algorithmically encoding markers of these social categories. However, we propose to derive this encoding from human input in a manner that relies on self-reported identity and provides individuals with the flexibility to determine what makes them feel represented. Critically, our approach does not require us to define and ascribe categories to people – either algorithmically or through third-party human raters.

More specifically, our proposed measure is designed to capture the degree to which a *diverse group of individuals see themselves represented in the result set*. To reiterate: this approach can be contrasted with previous methods that capture the degree to which *different sensitive attributes appear in the result set*.

At an algorithmic level, our method leverages a determinental point process (DPP), which is a probabilistic model that gives higher probability to sets that have greater spread, as specified by a predefined distance metric. Previous works on increasing variety in ranking and retrieval have leveraged DPPs with similarity kernels learned from item data and annotations that are treated as a given [13, 8]. In our work, we propose to learn an embedding space that reflects human judgements about feelings of representations in various pieces of visual media. The embedding space should place items that an individual (or group of individuals) feel well represented by close together. In contrast, if an individual judges an item to poorly represent their identities, it should be far in the embedding space from items the same individual judges as well representative of their identities. Specifically, we ask raters to select subsets of items they feel well represent some aspect(s) of their identity. We subsequently ask them to select items that they feel poorly represent those aspects of their identity. Using a triplet loss, we train the embedding model to place items in the former set close together while pushing items in the latter set away from this cluster.

We now detail a set of challenges that have emerged from ongoing research efforts to apply this framework to image sets. First, we recognize the fundamental limitations of an algorithmic approach aimed at quantifying representation in visual media. While we strive to avoid algorithmic inferences of identity categories, our framework still encodes image characteristics that act as signifiers of different social categories. As we endeavor to be respectful and nuanced in this encoding, we recognize the potential for this work to be overly reductive and even harmful. This fundamental tension indicates the importance of this work progressing in a socially engaged manner.

Second, we emphasize that the notion of diversity of representation is a challenging thing to formalize in computational terms. It inherently deals with messy, subjective, and personal perceptions and identities. It means different things to different people in different contexts; critically, there is no one right perspective (although some are arguably preferable to others). Our framework revolves around the notion of a diverse group of individuals feeling well represented by the visual media within the set. This framing ties diversity to human judgements of feeling represented, but leaves open questions of what constitutes a diverse set of individuals, and to what extent different individuals need to feel represented. It also leaves an open question as to the best way to solicit feelings of representation from human raters in a form that is amenable to learning the embedding space.

Finally, in order to build our embedding space we must solicit human judgements at a large scale. In machine learning, crowdsource platforms are commonly used for human labelling or rating, but are typically designed for studies where the relationship between the rater's identities and perception are outside the scope, so there are limited options for specification of the rater pool. As a result, traditional crowdsourcing platforms that are cost-effective at large scales have no guarantee of diversity along a variety of socially salient axes. This raises questions of whether we can scale insights from a small pool of diverse, high quality raters and whether crowdsourcing platforms could aid in this scaling.

## REFERENCES

[1] Mari Castañeda. 2018. The Power of (Mis)Representation: Why Racial and Ethnic Stereotypes in the Media Matter.

[2] L. Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. 2016. How to be Fair and Diverse? (10 2016).

[3] L. Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. 2018. Fair and Diverse DPP-Based Data Summarization. In *ICML*.

[4] P.D. Clough, Jo Bates, and Jahna Otterbacher. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results.

[5] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*.

[6] Stephanie A. Fryberg and Sarah S. M. Townsend. 2008. The psychology of invisibility.

[7] Elfriede Fürsich. 2010. Media and the representation of Others. *International Social Science Journal* 61 (03 2010), 113 – 130.

[8] Mike Gartrell, Elvis Dohmatob, and Jon Alberdi. 2018. Deep Determinantal Point Processes. *arXiv preprint arXiv:1811.07245* (2018).

[9] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

[10] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*.

[11] Kassia E. Kulaszewicz. 2015. Racism and the Media: A Textual Analysis.

[12] Peter Leavitt, Rebecca Covarrubias, Yvonne Perez, and Stephanie Fryberg. 2015. "Frozen in Time": The Impact of Native American Media Representations on Identity and Self-Understanding. *Journal of Social Issues* 71 (03 2015).

[13] Zelda Mariet, Mike Gartrell, and Suvrit Sra. 2019. Learning determinantal point processes by corrective negative sampling. In *The 22nd International Conference on Artificial Intelligence and Statistics*. 2251–2260.

[14] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and Inclusion Metrics in Subset Selection. In *AIES*.

[15] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[16] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

[17] Davood Rafiei, Krishna Bharat, and Anand Shukla. 2010. Diversifying web search results. In *Proceedings of the 19th international conference on World wide web*.

[18] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019).

[19] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery  Data Mining (KDD '18)*. 2219–2228.

[20] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H Chi, and Jennifer Gillenwater. 2018. Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*.

[21] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*.

[22] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm.