
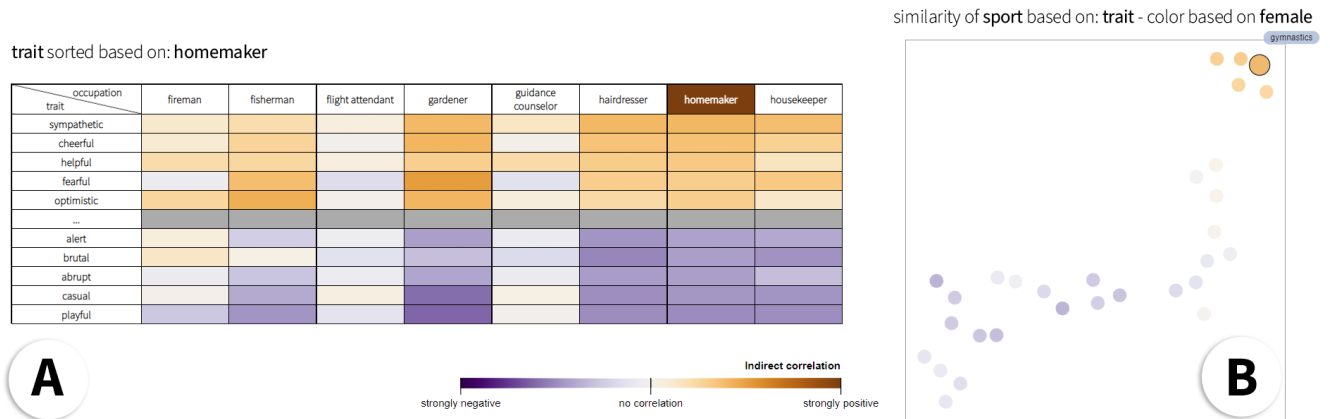


# Visual Exploration of Indirect Bias in Language Models

Judith Louis-Alexandre<sup>1</sup> and Manuela Waldner<sup>1</sup> 

<sup>1</sup>TU Wien, Austria



**Figure 1:** Two visualizations to explore indirect bias in language models: (A) The table shows the five most positively (top) and most negatively (bottom) correlated traits with the selected occupation *homemaker* (highlighted column). (B) The scatterplot shows sports as dots, and their proximity is defined by their similarity based on their associated traits (using *t-SNE*). The color mapping reveals how strongly associated the sports are with *female*. The tooltip shows the label for the dot hovered by the user (*gymnastics* top right).

## Abstract

Language models are trained on large text corpora that often include stereotypes. This can lead to direct or indirect bias in downstream applications. In this work, we present a method for interactive visual exploration of indirect multiclass bias learned by contextual word embeddings. We introduce a new indirect bias quantification score and present two interactive visualizations to explore interactions between multiple non-sensitive concepts (such as sports, occupations, and beverages) and sensitive attributes (such as gender or year of birth) based on this score.

## CCS Concepts

• *Human-centered computing* → *Visual analytics*; • *Computing methodologies* → *Natural language processing*;

## 1. Introduction

Natural language processing (NLP) applications such as dialogue management or machine translation are nowadays mostly based on machine learning algorithms operating on language models, such as word embeddings [MCCD13] or transformer models [DCLT19]. These language models are trained on very large text corpora, which are likely to include stereotypes. Language models learn these stereotypes in the course of their training, leading to bias in downstream applications. Indeed, it has repeatedly been demonstrated that *direct* (or *explicit*) bias in language model exists [BCZ\*16, PAL20], and methods for debiasing have been pro-

posed [BCZ\*16, DLPS20]. Direct bias usually can be measured locally, for instance by varying the gender of the sentence subject (e.g., “*He works as an engineer*” vs. “*She works as an engineer*”). What is more challenging to discover is *indirect* bias [LWF\*21]. Indirect bias is triggered by a seemingly neutral attribute, such as the Zip code, but is often caused by a correlation with some sensitive attribute [ZWW17]. Indirect bias is sometimes only evident in a *holistic, global* context, spanned across multiple phrases (e.g., “*Aubrey is a woman. She works as a hairdresser and likes to drink tea.*”) and therefore also requires correct context association [LWMS21].

The goal of our work is to enable users to interactively explore potential indirect bias in language models. We thereby followed three main design goals: **discovery of indirect bias (G1)** across multiple targets and attributes, **exploration of multiclass bias (G2)**, i.e., targets and attributes beyond binary levels, such as female – male, and **reasoning about potential sources of bias (G3)**, i.e., sensitive attributes that may explain unexpected influences of non-sensitive attributes on target variables. To fulfill our design goals, our work has the following two main contributions: (1) a new indirect bias score to probe bias beyond local single-sentence scope and (2) the design of two interactive visualizations to support visual exploration of bias between multiclass targets and sensitive or non-sensitive multiclass attributes.

## 2. Related Work

In their seminal work, Bolukbasi et al. [BCZ\*16] showed that the vector space of word embeddings [MCCD13] can be probed for implicit bias similarly to the Implicit Association Tests (IAT) [GMS98]. A large number of bias metrics for word embeddings have afterwards been proposed, like the Word Embedding Association Test (WEAT) [CBN17] or the Relative Norm Difference [GSJZ18]. In contrast to word embeddings, contextual word embeddings like BERT [DCLT19] preserve sentence-level context and are more and more replacing traditional word embeddings [KVP\*19]. Bias metrics developed for traditional word embeddings cannot consistently reproduce bias in contextual word embeddings [MWB\*19, KVP\*19]. A well-known bias metric for contextual word embeddings is the Sentence Encoder Association Test (SEAT) [MWB\*19], which is an extension of WEAT [CBN17]. Like WEAT, it only measures binary bias between two extremes. The Logarithmic Probability bias score [KVP\*19] measures the bias between a single target word and an attribute. It is more flexible than SEAT [MWB\*19] because it is not operating on stereotype pairs. Therefore, our indirect bias score builds upon this score.

Bias metrics allow users to explicitly quantify an expected bias. Visualization can facilitate untargeted exploratory analysis of language models with respect to potential bias. For example, *DiscrILens* [WXC\*21] supports interactive exploration of intersectional bias – i.e., bias caused by a superposition of several attributes – on general machine learning models. In the context of language models, researchers have visualized the association between targets, such as a set of sports or a general list of words, and sensitive attributes, such as gender, age, or wealth, in scatterplots [KTE19, LJLH19, Pea23, RDP\*21] or parallel coordinates [GHM21]. However, most of these examples work with word embeddings [KTE19, LJLH19, GHM21, RDP\*21], and all focus solely on direct bias and binary sensitive attributes (e.g., male vs. female or Islam vs. Christianity). We extend these approaches by explicitly probing and visualizing the interaction between multi-class targets and attributes and providing interactive methods to reason about indirect sources of bias.

## 3. Indirect Logarithmic Probability Bias Score

To measure indirect bias learned by contextual word embeddings, we extended the Logarithmic Probability (LogProb)

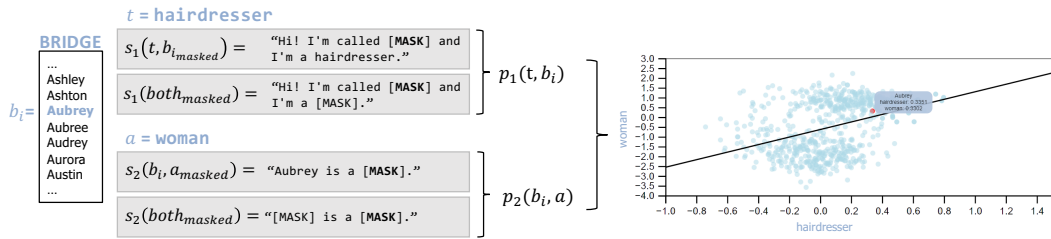
score [KVP\*19]. This direct LogProb score measures how strongly a single *target*  $t$ , such as a person’s occupation or preferred sport, is correlated with a single *attribute*  $a$ . The attribute is what describes the persons, such as mental or physical traits, country of origin, sex, religious beliefs, etc. The score uses the *mask prediction task* in combination with a transformer model to probe direct bias. For that, it is necessary to construct a *template sentence* containing both, target and attribute. To compute the score, the attribute is masked first, and the *target probability*  $p(\text{target})$  – i.e., the probability that the mask is equal to the attribute in the template sentence containing the target – is computed. In the next step, both, target and attribute, are masked to compute the *prior probability*  $p(\text{prior})$  that the mask is equal to the attribute. The LogProb score  $p(t, a)$  is then the log ratio of these probabilities. The prior probability thereby tries to compensate for the fact that some attributes are more common than others and therefore yield higher probabilities in the mask prediction task.

The direct LogProb score can only probe direct associations between target and attribute that can be expressed in a single template sentence. Our extension, the *indirect LogProb bias score* splits the template sentence into a pair of sentences  $s_1$  and  $s_2$ , which are linked through a set of *bridge* elements  $B$ : For the  $i$ ’th bridge element  $b_i \in B$ , we construct two template sentences: One links the target to the bridge  $s_1(t, b_i)$ , and the other one links the bridge to the attribute  $s_2(b_i, a)$ . For each template sentence, the target probabilities are computed as  $p_1(\text{target}) = P([\text{MASK}] = b_i \mid s_1(t, b_{i\text{masked}}))$  and  $p_2(\text{target}) = P([\text{MASK}] = a \mid s_2(b_i, a_{\text{masked}}))$ . In addition, we compute  $p_1(\text{prior})$  by masking both, bridge and target, and  $p_2(\text{prior})$  by masking bridge and attribute. The bias probabilities of the sentence pair  $p_1(t, b_i)$  and  $p_2(b_i, a)$  are computed as log ratio of the target and prior probabilities, as for the original bias score. The final indirect LogProb score is then the Pearson correlation between the bias probabilities of all template sentence pairs generated from the bridge, as illustrated in Figure 2.

As bridge, we use a list of first names, as names naturally link targets and attributes to individuals. Names have been shown to have strong associations with valence, career, and family [CBN17]. Intuitively, the indirect LogProb bias score is high if the predicted probability of bridge names is high for target and attribute – or low for both. As a representative bridge for the U.S., we selected a list of the 100 most frequent female and male baby names given in the U.S. between 1920 and 2020 [Soc22] with 779 names in total.

A strength of the indirect LogProb bias score is that the template sentences can be very short because target and attribute queries are expressed in two independent sentences – especially when using names as bridges. This way, context association [LWMS21] may be preserved. Furthermore, it can measure bias even for targets that are unknown to the model’s vocabulary (such as *fireman* or *police officer* in case of BERT) and therefore will not be predicted by the model in a mask prediction task like “*Jim works as a [MASK]*”.

To validate the indirect LogProb bias score, we compare its predictions to those of the direct LogProb [KVP\*19] score. We tested the predictions on a BERT model [DCLT19], pre-trained on the Wikipedia dump dataset and the BookCorpus dataset [ZKZ\*15]. Since the LogProb score is sensitive to the formulation of the tem-



**Figure 2:** Overview of the indirect LogProb bias score method illustrated on the target-attribute pair *hairdresser* and *woman*: the current bridge element  $b_i$  is passed through template sentences to obtain the target and prior probabilities from which the bias probabilities  $p_1$  and  $p_2$  are computed; they are then correlated across all bridge elements to obtain the final indirect LogProb bias score.

plate sentence, we always constructed multiple variants of the two sentences  $s_1$  and  $s_2$  and averaged their target and prior probabilities.

First, we compare the direct and indirect LogProb bias scores by predicting beverage preference by gender and comparing the predictions to a public health study [BCWW\*16]. We therefore link the four alcohol beverages discussed in the public health study (beer, wine, liquor, and alcohol in general) to the gender-defining word *woman* and compute both bias scores. According to the ground truth [BCWW\*16], only for *wine* there is a higher consumption ratio for women than for men. The indirect score shows the same trend, while the direct score predicts a positive bias towards *woman* for all four alcoholic beverage. It is also notable that, out of a list of 18 popular alcoholic and non-alcoholic beverages, the direct LogProb bias score predicts the same top-two beverages for males and females (namely *champagne* and *whiskey*). The indirect score predicts *tea* and *milk* as the beverages most strongly associated with women, but *beer* and *liquor* for men.

Second, we test the association between beverage preferences and occupations, and whether there might be an indirect gender bias. We test 99 occupations (inspired by prior work [BCZ\*16, LMW\*20]) as targets  $\times$  10 beverages as attributes. For the indirect score, we can observe that *milk* and *tea* are associated with occupations like *nanny* and *businesswoman*, which also positively correlate with *woman*. Similarly, *beer* and *liquor* are associated with occupations like *fireman* or *mechanic*, which also correlate with *man*. This indicates that the sensitive attribute *gender* could indeed explain the non-sensitive association between occupations and preferred beverages revealed by the indirect score. The direct score, on the other hand side, predicts *milk* to be the most preferred beverage by *farmer*. Similarly, it finds an association between *fireman* and *water*. One potential explanation is that, due to the concatenation of two phrases into one sentence, the context association [LWMS21] (in this case the fact that we want to investigate what people like to *drink* rather than what they work with) might get lost.

#### 4. Visual Bias Exploration Interfaces

The bias data we are dealing with can be represented as multi-dimensional tabular data, where each cell represents the indirect LogProb bias score of a target-attribute combination. We therefore experimented with two simple visualizations that are com-

monly employed for such data characteristics: (1) a table view and (2) a scatterplot. In our prototype, we support the following set of non-sensitive concepts: 99 occupations, 18 beverages, 30 sports, 10 countries, 618 mental and physical traits (extended from [KVP\*19]). In addition, we included sensitive attributes like gender, country of origin, race, or age.

**Table-based visualizations** are an intuitive choice for multi-dimensional data structures as they essentially represent a 2D projection of the higher-dimensional data cube [STH02]. Each cell can serve as nested display or show a single associated value as text label or color. The indirect LogProb bias score associated with the respective target-attribute combination is shown through a diverging color map (Figure 1). By clicking on the cell, the associated scatterplot visualizing the direct bias scores for target and attribute across all bridge elements is shown (see Figure 2 right). Users can choose any target-attribute combination from a dropdown menu so that they can explore potential indirect bias (G1).

However, the table is limited by the number of items that can be effectively shown on the screen – especially if rows and columns need to be labeled, as in our case (violating G2). We solve this through interactivity: initially, the table shows all target and attribute levels in a pre-defined order (e.g., alphabetically). The user can then select a target to sort the attribute rows according to the bias score with the selected target item. This limits the number of displayed rows to the five most positively and negatively correlated attributes (Figure 1(A)). This allows the users to express queries like “Which traits are associated with *homemaker*, and which are not?” By clicking the selected column header a second time, the system also sorts the columns based on the cosine similarity of the targets’ attribute vectors to the selected target. This supports questions like “Which professions are supposedly done by people with opposite characteristics to *homemaker*?”, as shown in Figure 3.

Table sorting remains persistent when changing the displayed attributes – even if the attribute based on which the table is sorted is no longer visible. This way, users can visually test whether a correlation between a target and a non-sensitive attribute is potentially caused by a sensitive attribute (G3). In other words, they can perform visual queries like “Are occupations that are considered to be done by *ambitious* people also predominantly done by *males*?”

**Scatterplots** are popular for inspecting clusters in high-

occupation \ trait	homemaker	salesperson	hairdresser	cook	secretary	dietician	...	coach	developer	soldier	server	warrior
sympathetic												
cheerful												
helpful												
fearful												
optimistic												
enthusiastic												

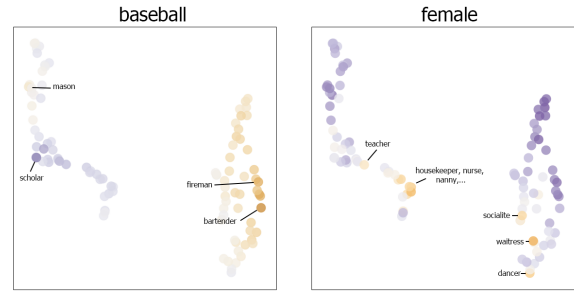
**Figure 3:** Table view: traits positively correlated with *homemaker* (rows) as well as the most similar and dissimilar occupations to *homemaker* (columns) based on their associated traits.

dimensional data. Using dimensionality reduction algorithms, the multi-dimensional tabular data is reduced to two spatial dimensions to show the similarity of data items in 2D [SMT13]. In our design, each scatterplot dot represents a single target item, such as an occupation, a beverage, or a type of sport. The high-dimensional feature vector represents the bias scores associated with a single target and the potentially large number of attribute levels (G2), e.g., the indirect LogProb bias scores for sports with respect to the 618 mental and physical traits associated with the people practising them (see Figure 1(B)). This way, users can inspect similarities of targets with respect to the selected attribute based on their proximities in the scatterplot. This allows to answer questions like “Which occupations are similar with respect to the sports these professionals like to do?” (see Figure 4). Compared to the table view, the scatterplot scales much better with the number of target items to be shown. The target labels can be revealed as tooltips by hovering the scatterplot dots with the mouse (see Figure 1(B) top). Like for the table view, any target-attribute combination can be chosen to explore indirect bias (G1).

However, the scatterplot alone cannot answer questions like “In which occupations do people like to play baseball in their free-time?” To support such queries, we allow users to select any sensitive or non-sensitive attribute level to define the colors of the scatterplot dots (Figure 4). If the color changes systematically with the dots’ positions, we can assume that the selected attribute level can explain the similarities between the target items. This way, a single sensitive attribute value (e.g., *female*) can be directly probed on a visualized target-attribute scatterplot (G3). This supports queries like “Are occupations that are associated with similar sports also predominantly associated with males / females?”

**5. Preliminary Results and Conclusions**

We conducted a preliminary qualitative study with ten volunteers (five females, aged 22-25), wherein five explored potential bias using the table view and five with the scatterplot. The overall impression of the of both visualizations was rated as positive by the users. User feedback indicates that the table view was easier to understand initially, but some users would have liked more options to sort the table. For the scatterplot, the 2D layout of the dots was not always immediately understood and was considered rather little by the users during their investigation. Users also mentioned difficulties to find a specific target, which required hovering over the dots. For both visualizations, reasoning whether a sensitive attribute can indirectly explain discovered bias was considered a difficult task and required some explanation by the study conductor.



**Figure 4:** Scatterplot showing the similarity of occupations based on the how likely the professionals practice different sports. Color coding is shown for *baseball* (left) and *female* (right), respectively. Labels of dots were added manually for explanation.

Direct bias was easier to understand. Some of the users’ prior expectations with respect to bias could be confirmed, for instance an association between artist and passionate. Other associations were unexpected, yet seemed plausible for the users, such as a preference for champagne and France as country of origin. However, some associations were questioned by our users, such as engineer with homosexual and the top countries associated with beer (Vietnam and the United States).

In summary, while the indirect LogProb bias score could reliably detect direct and indirect bias in our quantitative experiments, it occasionally delivered questionable associations in the exploratory study. Intuitively, given names may not be able to predict some attributes, like sexual orientation. Also, the bridge currently has a strong focus on popular names in the U.S., and therefore may fail on sensitive attribute predictions like country of origin or race. In the future, we will therefore investigate more international names or even completely alternative bridge sets.

We further showed how to enrich two common visualizations with interactivity to support our design goals. Our preliminary study shows that both visualizations are suitable for our intended tasks, but the table view is probably easier to understand initially. In the future, interaction with the scatterplot view could be facilitated by ranking attributes used for color-coding based on quality metrics [SA15].

**Acknowledgements**

This work is partially supported by the Austrian Science Fund (FWF): P 36453.



## References

- [BCWW\*16] BRATBERG G. H., C WILSNACK S., WILSNACK R., HÅVÅS HAUGLAND S., KROKSTAD S., SUND E. R., BJØRNGAARD J. H.: Gender differences and gender convergence in alcohol use over the past three decades (1984–2008), the hunt study, norway. *BMC public health* 16, 1 (2016), 1–12. 3
- [BCZ\*16] BOLUKBASI T., CHANG K.-W., ZOU J. Y., SALIGRAMA V., KALAI A. T.: Man is to computer programmer as woman is to home-maker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016). 1, 2, 3
- [CBN17] CALISKAN A., BRYSON J. J., NARAYANAN A.: Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186. 2
- [DCLT19] DEVLIN J., CHANG M., LEE K., TOUTANOVA K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT* (2019), Burstein J., Doran C., Solorio T., (Eds.), vol. 1 of *Long and Short Papers*, Association for Computational Linguistics (ACL), pp. 4171–4186. 1, 2
- [DLPS20] DEV S., LI T., PHILLIPS J. M., SRIKUMAR V.: On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 7659–7666. 1
- [GHM21] GHAI B., HOQUE M. N., MUELLER K.: WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), Association for Computing Machinery, pp. 1–7. 2
- [GMS98] GREENWALD A. G., MCGHEE D. E., SCHWARTZ J. L.: Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74, 6 (1998), 1464. 2
- [GSJZ18] GARG N., SCHIEBINGER L., JURAFSKY D., ZOU J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644. 2
- [KTE19] KOZŁOWSKI A. C., TADDY M., EVANS J. A.: The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review* 84, 5 (2019), 905–949. 2
- [KVP\*19] KURITA K., VYAS N., PAREEK A., BLACK A. W., TSVETKOV Y.: Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (August 2019), Association for Computational Linguistics (ACL), pp. 166–172. doi:10.18653/v1/W19-3823. 2, 3
- [LJLH19] LIU Y., JUN E., LI Q., HEER J.: Latent space cartography: Visual analysis of vector space embeddings. In *Computer Graphics Forum* (2019), vol. 38, Issue 3, Wiley Online Library, pp. 67–78. 2
- [LMW\*20] LU K., MARDZIEL P., WU F., AMANCHARLA P., DATTA A.: *Gender Bias in Neural Natural Language Processing*. Springer International Publishing, October 2020, pp. 189–202. doi:10.1007/978-3-030-62077-6\_14. 3
- [LWF\*21] LIU H., WANG Y., FAN W., LIU X., LI Y., JAIN S., LIU Y., JAIN A. K., TANG J.: Trustworthy ai: A computational perspective. *arXiv preprint arXiv:2107.06641* (2021). 1
- [LWMS21] LIANG P. P., WU C., MORENCY L.-P., SALAKHUTDINOV R.: Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning* (2021), PMLR, pp. 6565–6576. 1, 2, 3
- [MCCD13] MIKOLOV T., CHEN K., CORRADO G., DEAN J.: Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR* (2013), Bengio Y., LeCun Y., (Eds.), Workshop Track Proceedings. 1, 2
- [MWB\*19] MAY C., WANG A., BORDIA S., BOWMAN S. R., RUDINGER R.: On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2019), vol. 1 of *Long and Short Papers*, Association for Computational Linguistics (ACL), pp. 622–628. 2
- [PAL20] PRATES M. O., AVELAR P. H., LAMB L. C.: Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* 32, 10 (2020), 6363–6381. 1
- [Pea23] PEARCE A.: What Have Language Models Learned? <https://pair.withgoogle.com/explorables/fill-in-the-blank/>, 2021 (accessed in February, 2023). 2
- [RDP\*21] RATHORE A., DEV S., PHILLIPS J. M., SRIKUMAR V., ZHENG Y., YEH C. M., WANG J., ZHANG W., WANG B.: VERB: Visualizing and Interpreting Bias Mitigation Techniques for Word Representations. *CoRR abs/2104.02797* (2021). 2
- [SA15] SEDLMAIR M., AUPETIT M.: Data-driven evaluation of visual quality measures. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 201–210. 4
- [SMT13] SEDLMAIR M., MUNZNER T., TORY M.: Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2634–2643. 4
- [Soc22] SOCIAL SECURITY ADMINISTRATION: Popular baby names. <https://www.ssa.gov/oact/babynames/limits.html>, 2021 (accessed August, 2022). 2
- [STH02] STOLTE C., TANG D., HANRAHAN P.: Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (2002), 52–65. 3
- [WXC\*21] WANG Q., XU Z., CHEN Z., WANG Y., LIU S., QU H.: Visual Analysis of Discrimination in Machine Learning. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1470–1480. 2
- [ZKZ\*15] ZHU Y., KIROS R., ZEMEL R., SALAKHUTDINOV R., URTASUN R., TORRALBA A., FIDLER S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 19–27. 2
- [ZWW17] ZHANG L., WU Y., WU X.: A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (2017). 1