Ethics, Fairness, and Bias in Machine Learning

17 October 2023

Zeerak Talat
MBZUAI
z@zeerak.org | @zeeraktalat

Outline

Discussion-based lecture

Today:

- Natural Language Processing
- Computer Vision
- Machine Learning
- Cross-polination between NLP/CV
- LLMs
- (Y)our Individual responsibility
- Discussing case studies

Take-Aways

At the end of this lecture you should have

- An idea of how to think about ethical issues
- An understanding of your role in research and development of AI tools
- A way to question what you think is acceptable work
- An understanding of how (AI) technologies influence the world around us

Crucially: How ***your perspective*** *can change AI to make it less harmful.*

Why should you care?

Created a dataset for hate speech detection and got this table of features.

Q: *What problem arises when you look at these features?*

Feature (sexism)	Feature (racism)
'xist'	'sl'
'sexi'	'sla'
'ka'	'slam'
'sex'	'isla'
'kat'	'l'
'exis'	'a'
'xis'	'isl'
'exi'	'lam'
'xi'	'i'
'bitc'	'e'
'ist'	'mu'
'bit'	's'
'itch'	'am'
'itc'	'm'
'fem'	'la'
'ex'	'is'
'bi'	'slim'
'irl'	'musl'
'wom'	'usli'
'girl'	'lim'

Table 5: Most indicative character n-gram features for hate-speech detection

NLP: Content Moderation

Content moderation is the process of determining what is acceptable and what is not.

Models reproduce what is available to them in their datasets.

Q: *What issues arise with this approach?*

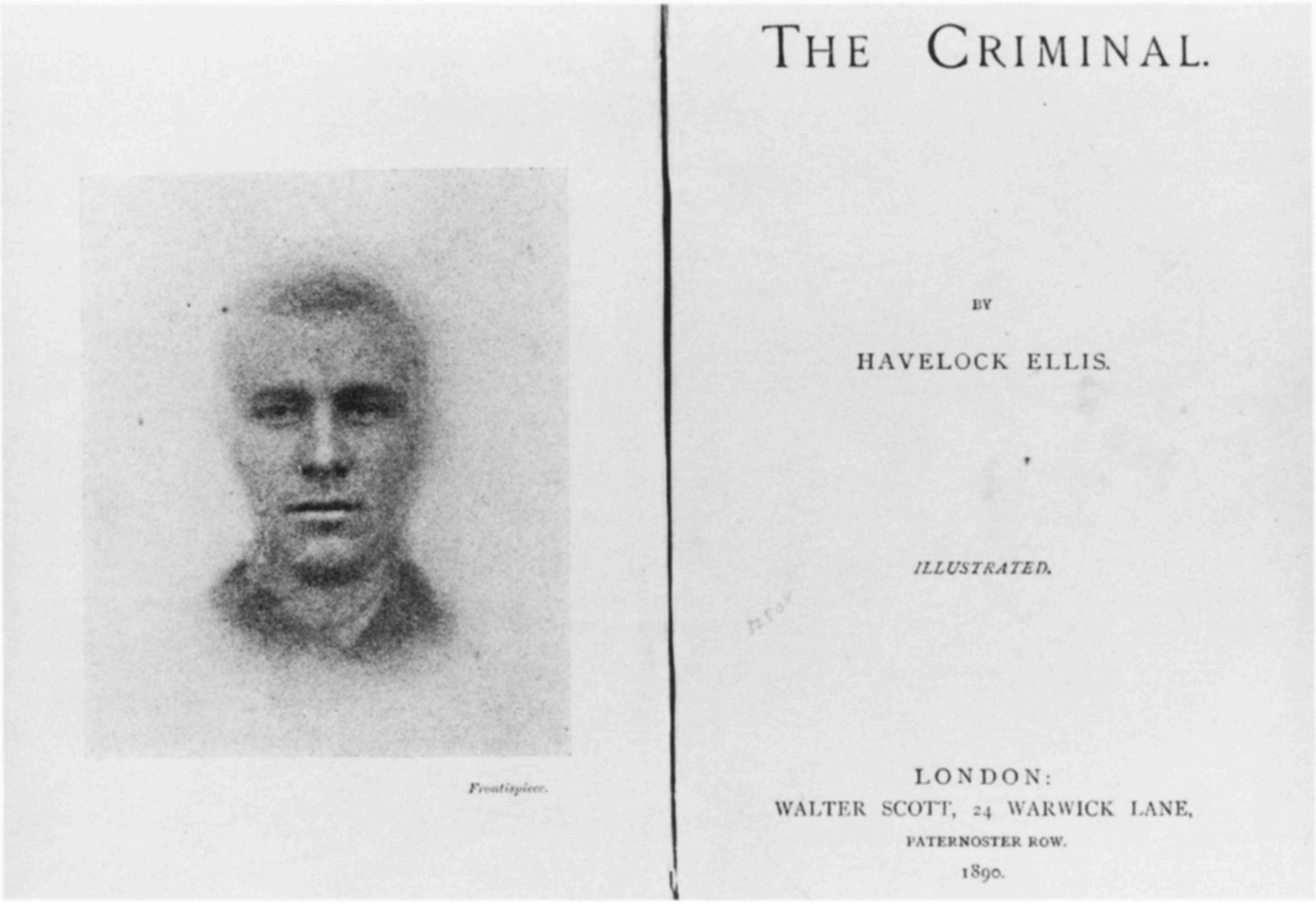
Q: *How can we address such issues, without changing data or model?*

Q: *What are the limits of those approaches?*

ML & Statistics

A little history: Francis Galton and Eugenics

- *Goal:* Classify good/bad human traits
- *Method*
    - Average intra-group diffs
    - Highly inter-group diffs



A Galtonoian Composite as shown by Alan Sekula: The Body and the Archive (1986). October. MIT Press

The Distributional Hypothesis

The Distributional Hypothesis describes a frequentist approach to salience: What frequently co-occurs should be treated related

For NLP Tokens frequently co-occurring with the same tokens ➡ Similar semantically

For ML Frequently co-occurring patterns ➡ Highly salient patterns

Used to draw decision boundaries between classes & cluster information within classes.

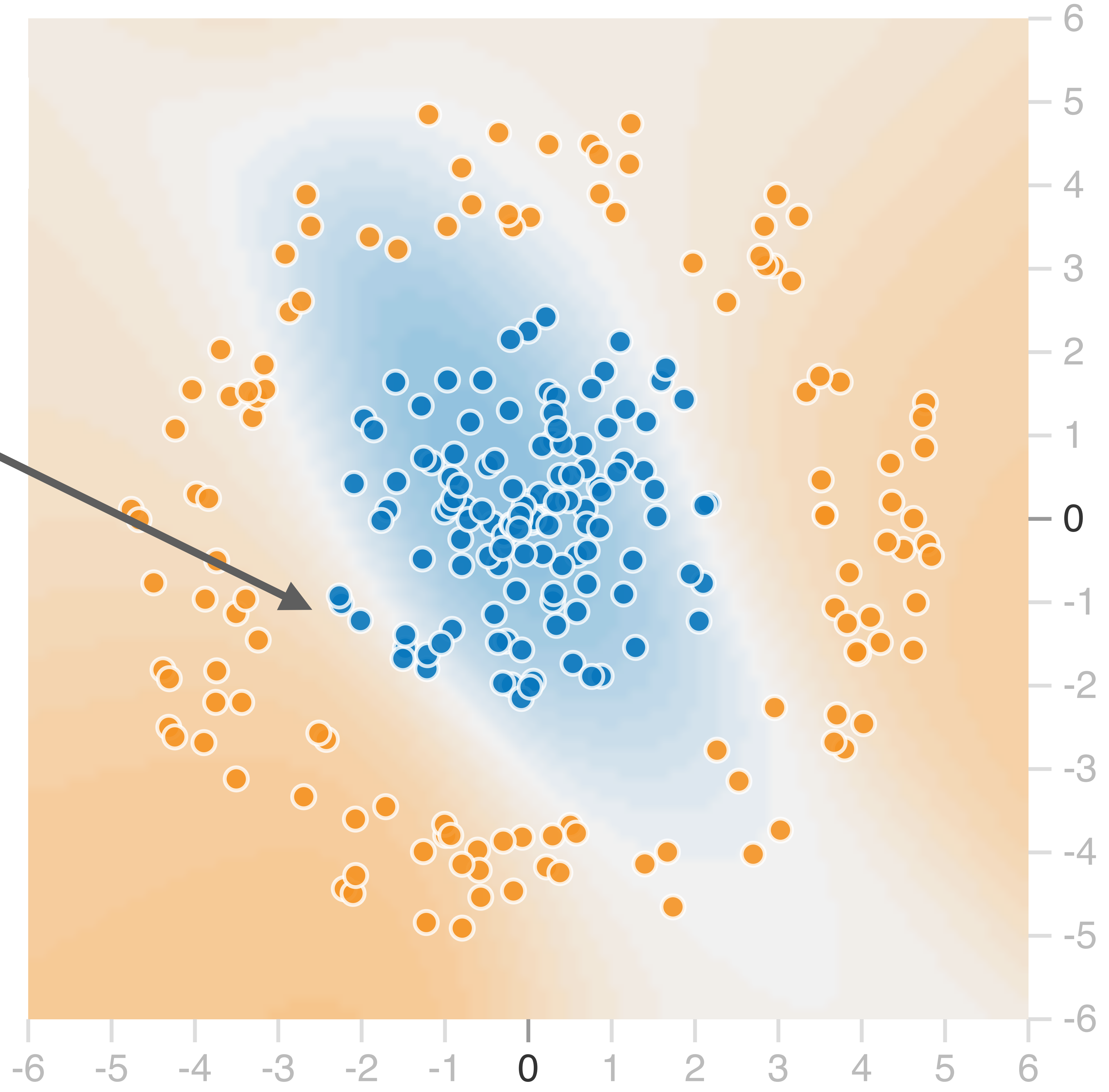
The Distributional Hypothesis

Infrequent information at the edge of the vector space ➡ incorrect classification / Infrequent generation

Full breadth of data impossible to collect

Q: *Would this change with a full sample?*

Q: *When does it matter that we have a narrow sample?*



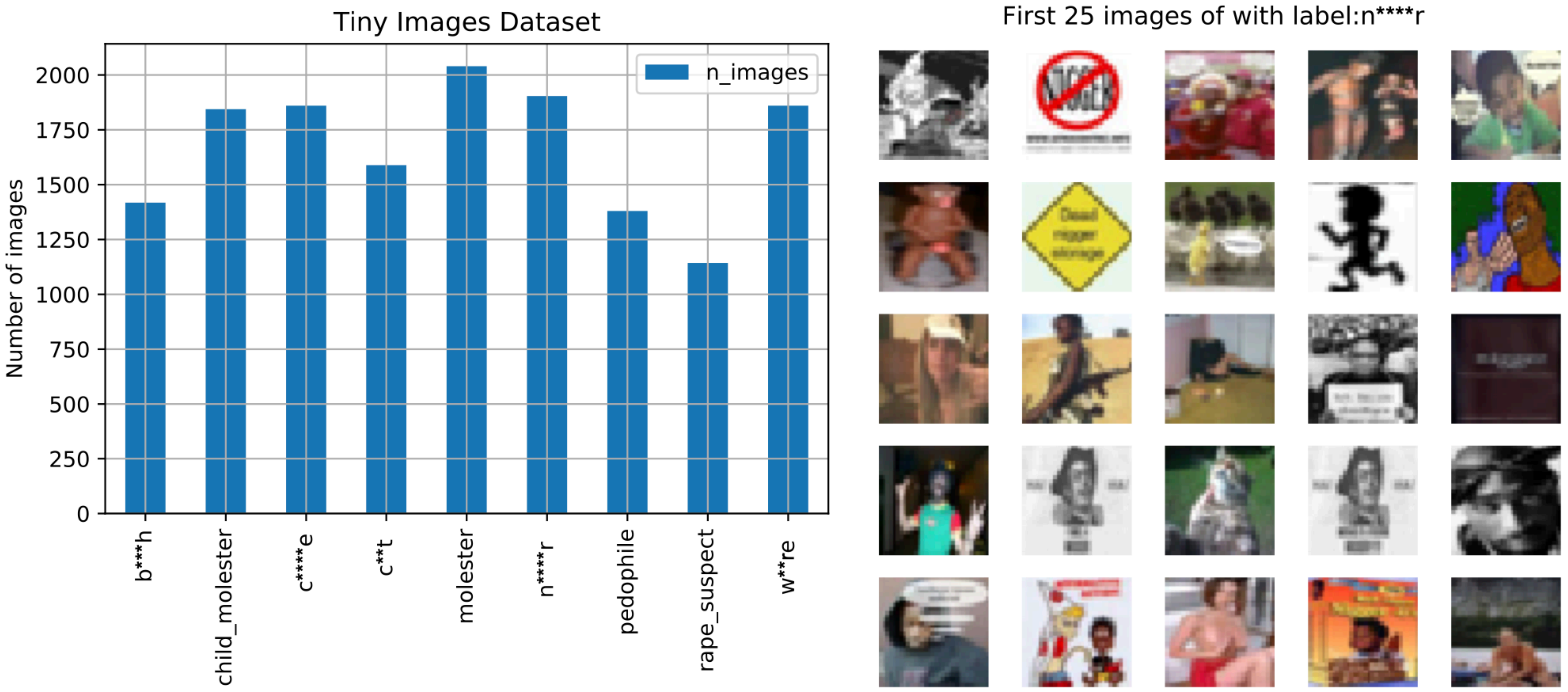
Classification boundary for trained on toy dataset from TensorFlow Playground.

Computer Vision: Face Recognition

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Computer Vision & NLP: ImageNet



(a) Class-wise counts of the offensive classes

(b) Samples from the class labelled n****r

Figure 1: Results from the *80 Million Tiny Images* dataset exemplifying the toxicities of it's label space

Prahbu and Birhane

Attempted (technical) Solutions: ML

Protected Attribute: A (e.g., Male/Female person)

Predicted Class: O (An outcome e.g., gets admitted to MBZ)

Predictive attribute: Y (e.g., Variable that indicates degree attainment)

Demographic Parity:

$$\mathbb{P}(O = 1 | A = 0) = \mathbb{P}(O = 1 | A = 1)$$

Equalized Odds:

$$FNR = \mathbb{P}(O = 0, A = 1, Y = 1) = \mathbb{P}(O = 0 | A = 1, Y = 0)$$

$$FPR = \mathbb{P}(O = 1, A = 1, Y = 1) = \mathbb{P}(O = 1 | A = 1, Y = 0)$$

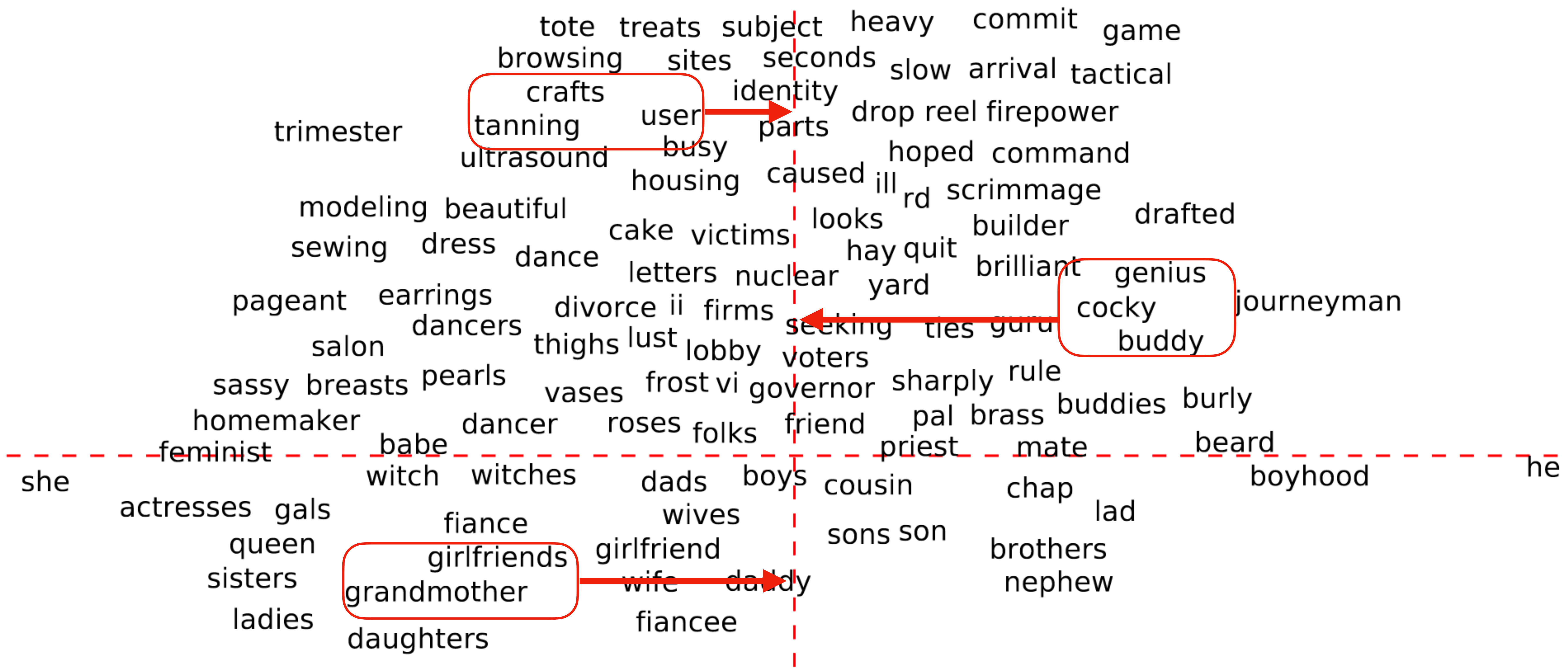
Attempted (technical) Solutions: CV

...

Attempted (technical) Solutions: NLP

Measurement (e.g., Bolukbasi et al., 2016; Nangia et al., 2020)
Debiasing vector representations (e.g., Bolukbasi et al., 2016; Kaneko and Bollegala, 2019)
Counter-factuals / Invariance (e.g., Liu and Avci, 2019)
Value alignment (e.g., Solaiman and Dennison, 2021)

Attempted (technical) Solutions: NLP



Bolukbasi et al. 2016

Attempted (technical) Solutions: NLP

Method	**Sentence**	**Probability**
Baseline	I am *gay*	0.915
	I am straight	0.085
Our Method	*I am* gay	0.141
	I am straight	0.144

Table 1: Toxicity probabilities for samples of a baseline CNN model and our proposed method. Words are shaded based on their attribution and italicized if attribution is > 0.

Liu and Avci, 2019.

Attempted (technical) Solutions: NLP

Value Alignment

- Just prompt engineering and penalizing models for bad completions
- Also what is done using RLHF

Evaluation Paradigms

Intrinsic - Fixing model representations (i.e., gender bias in model representations)
Extrinsic - Evaluating on a downstream task (i.e., discriminatory classifications)

Q: *Which evaluation paradigm would you prefer? Why?*

Generative AI

- Alignment with human values
  - Done through fine-tuning on datasets
  - Through RLHF
- Blocklists

Partial Views and Subjective Knowledge

A particularly starry night in August

Partial Views and Subjective Knowledge

A particularly starry night in August

Partial Views and Subjective Knowledge

Our knowledges and experiences provide the background for how we view the world

- E.g., Face Recognition example

Partial views are okay — important thing is to critically examine what we might be missing

Summary

Discussed different ethical issues
- Content moderation
- Face Detection
- The distributional hypothesis / Frequency

Generative AI and its issues

Different approaches to addressing harms

How we as researchers impact technology

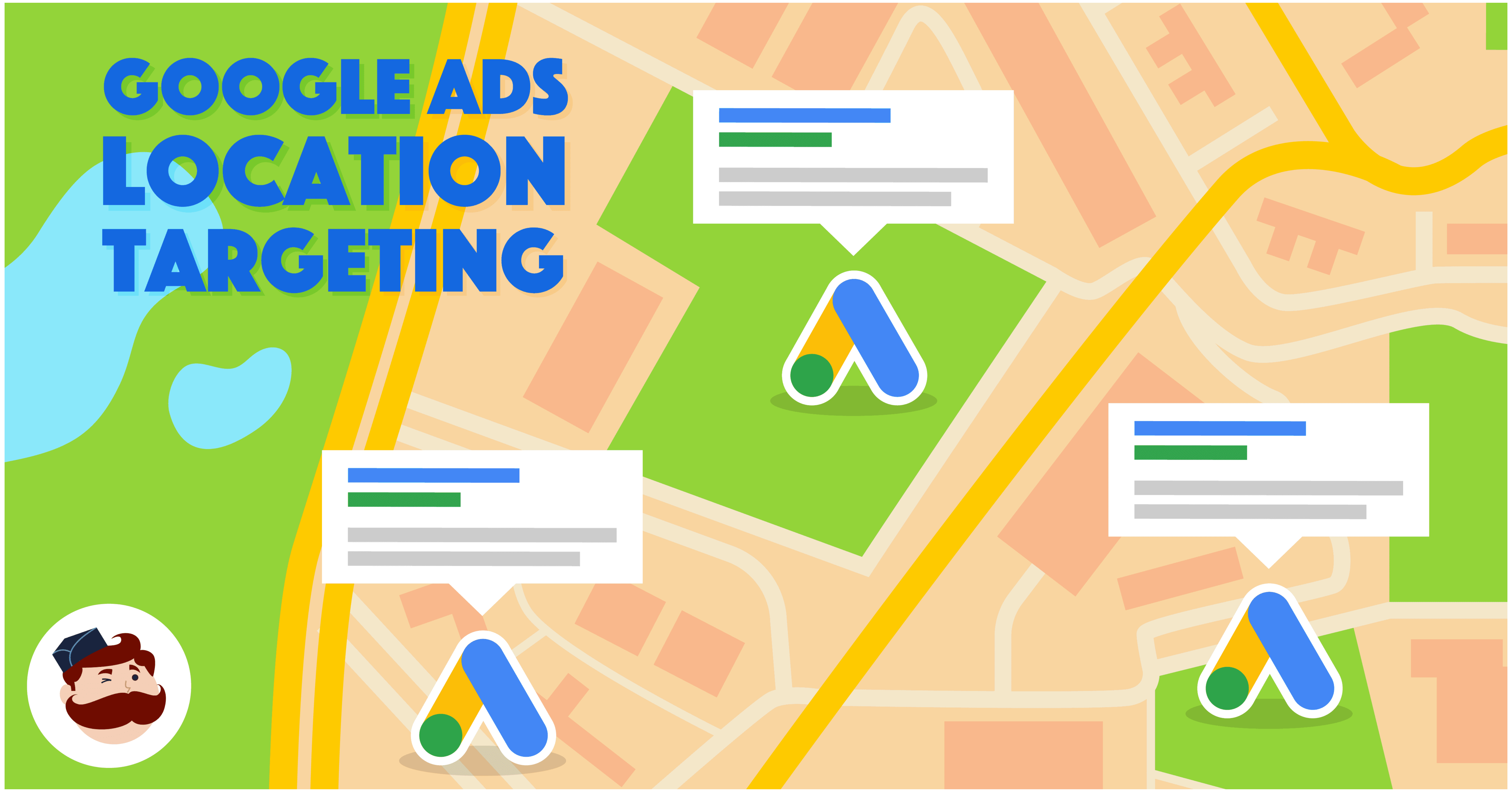
References

1. Birhane, Abeba, and Vinay Prabhu. “Large Image Datasets: A Pyrrhic Win for Computer Vision?” In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1537–47. Computer Vision Foundation, 2021. https://openaccess.thecvf.com/content/WACV2021/html/Birhane_Large_Image_Datasets_A_Pyrrhic_Win_for_Computer_Vision_WACV_2021_paper.html.
2. Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. “Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings.” In *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Vol. 29. Curran Associates, Inc., 2016. https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.
3. Buolamwini, Joy, and Timnit Gebru. “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.” In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A. Friedler and Christo Wilson, 81:77–91. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, 2018. http://proceedings.mlr.press/v81/buolamwini18a.html.
4. Costanza-Chock, Sasha. “Design Justice, A.I., and Escape from the Matrix of Domination.” *Journal of Design and Science*, July 16, 2018. https://doi.org/10.21428/96c8d426.
5. Dias Oliva, Thiago, Dennys Marcelo Antonialli, and Alessandra Gomes. “Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online.” *Sexuality & Culture* 25, no. 2 (April 2021): 700–732. https://doi.org/10.1007/s12119-020-09790-w.
6. Douglas, Mary. Purity and Danger: An Analysis of the Concepts of Pollution and Taboo. Repr. London: Routledge, 1978.
7. Dunn, Jonathan. “Mapping Languages: The Corpus of Global Language Use.” *Language Resources and Evaluation* 54, no. 4 (December 2020): 999–1018. https://doi.org/10.1007/s10579-020-09489-2.
8. Hall, Stuart. “Race, the Floating Signifier.” Lecture, Media Education Foundation, 1997. https://shop.mediaed.org/race-the-floating-signifier-p173.aspx.
9. ———. “The Spectacle of the Other.” In Representation: Cultural Representations and Signifying Practices, Vol. 7. Sage London, 1997.
10. Haraway, Donna. “Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective.” *Feminist Studies* 14, no. 3 (1988): 575–99. https://doi.org/10.2307/3178066.
11. Kalluri, Pratyusha. “Don’t Ask If Artificial Intelligence Is Good or Fair, Ask How It Shifts Power.” *Nature* 583, no. 7815 (July 9, 2020): 169–169. https://doi.org/10.1038/d41586-020-02003-2.
12. Liu, Frederick, and Besim Avci. “Incorporating Priors with Feature Attribution on Text Classification.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6274–83. Florence, Italy: Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/P19-1631.
13. Sekula, Allan. “The Body and the Archive.” *October* 39 (1986): 3. https://doi.org/10.2307/778312.
14. Solaiman, Irene, and Christy Dennison. “Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets.” *arXiv:2106.10328 [Cs]*, November 23, 2021. http://arxiv.org/abs/2106.10328.

Case I: Autonomous Weaponry

NATO Foundation Dossier on Autonomous Weaponry. 2020.

Case II: Advertisement



Algorithmic Global. Understanding Location Targeting in Google Ads. 2020.

Case III: Combining Disparate Sources

Ragha Vasudevan. Combining Data Sources: Approaches & Considerations. 2017.

Case IV: Automatic Speech Recognition

TECH

Prisons are using Amazon Transcribe and AI to monitor inmates' phone calls

A new report sheds light on companies like LEO Technologies, whose AI-scanning audio software employs Amazon speech-to-text recognition.

Andrew Paul. Prisons are using Amazon Transcribe and AI to monitor inmates' phone calls. 2021