Bias and Ethics of Using Machine Learning tools

AY 2025/2026

Taught Seminar: Feb 9–10 2026

Zee Talat

ztalat@ed.ac.uk

THE UNIVERSITY of EDINBURGH
Edinburgh Futures Institute

THE UNIVERSITY of EDINBURGH
informatics

Learning Goals

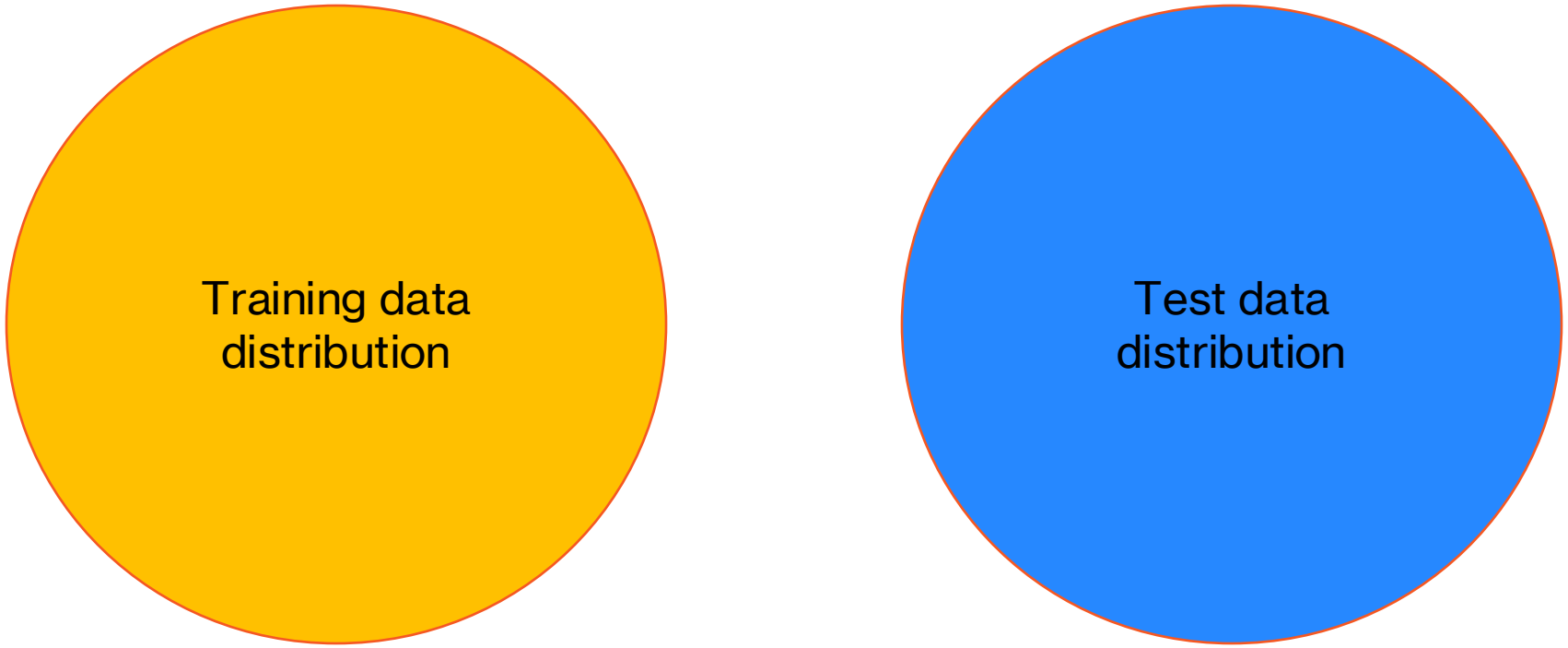
- Get an overview of how Machine Learning tools work
- Get an understanding of fairness concerns of machine learning
- Get tools for how to think about applying machine learning tools for computational social science

Part I
Machine Learning

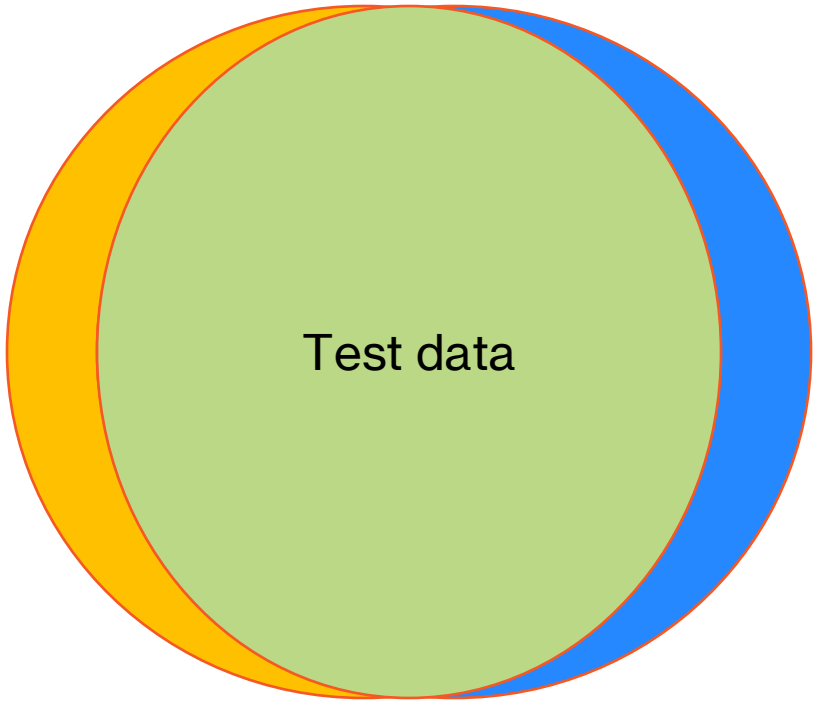
What is Machine Learning?

- Machine Learning (ML) are a series of optimisation techniques to develop a representation of data.
- Goal of optimisation is to minimise the expectation of error.

What is Machine Learning?



What is Machine Learning?



What is Machine Learning?

- Supervised machine learning
- Unsupervised machine learning

What is Supervised Learning?

- Combines data (e.g., text) with a label
  - The label can either be categorical for classification or continuous for regression.

Example:

“This is such an easy question!” Rating: ⭐ ⭐ ⭐ ⭐

What is Supervised Learning?

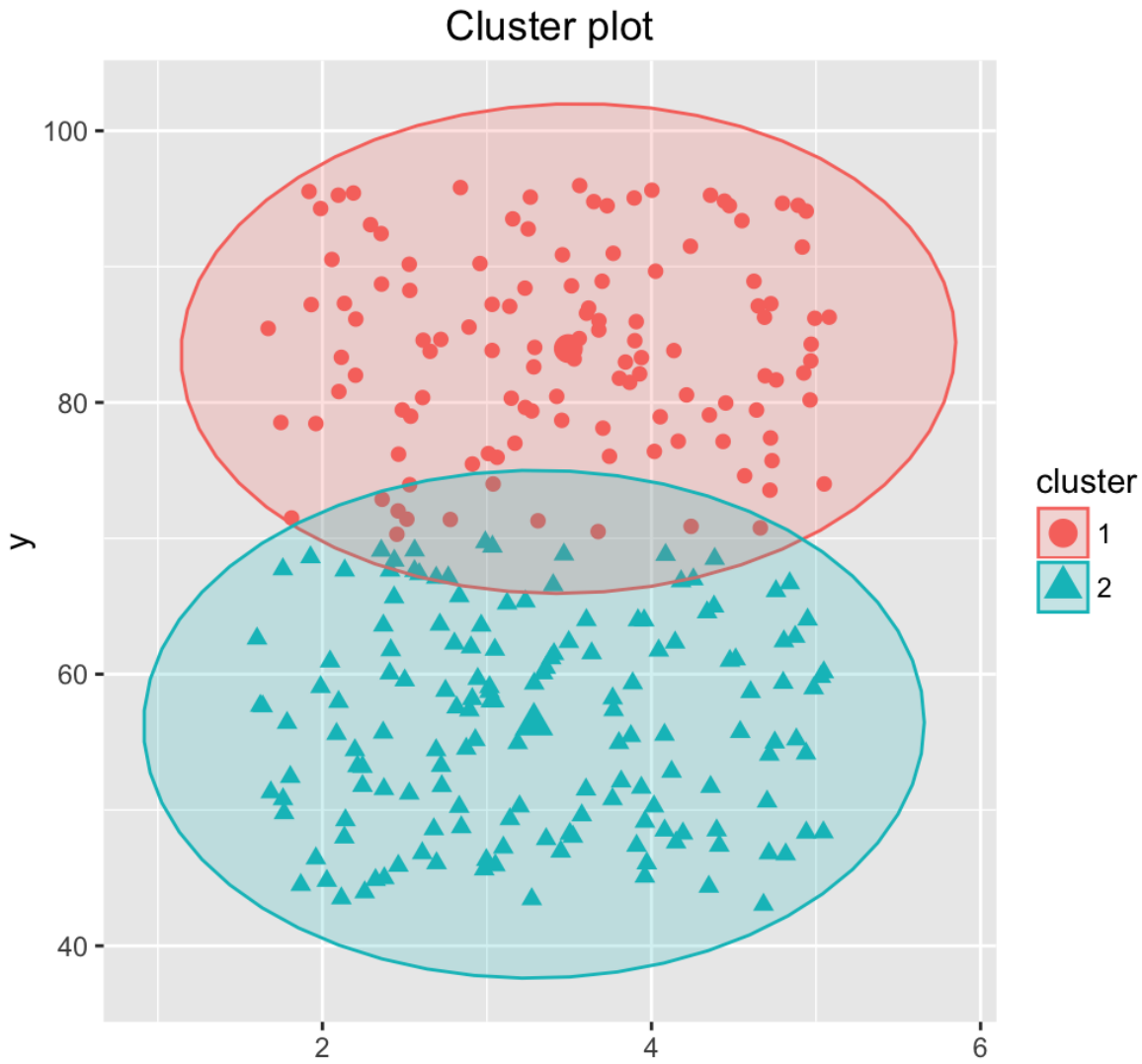
- Combines data (e.g., text) with a label
  - The label can either be categorical for classification or continuous for regression.

Example:

“This is such a dumb question!” Rating: 1.5/5

What is Unsupervised Learning?

- Label-free learning
  - So what does it learn?



Source: Assessing clustering tendency: A vital issue - Unsupervised Machine Learning. https://www.sthda.com/english/wiki/wiki.php?id_contents=7922

What is Generative AI?

- What kind of system are generative AI systems?

Returning to Error

- What kind of system are generative AI systems?

Part II
Fairness and Bias

Machine learning and patterns

- Do all patterns occur equally regularly in real life?
  - What is the ground truth?
- What is the interaction with data?

Fairness definitions

- Individual Fairness
  - About ensuring that
- Group Fairness
- Fairness through unawareness

Equalized Odds

- Equalized Odds
  - For all values $y \in Y, a \in A$

$$P(\hat{Y} = y \,|A = a, Y = y) = P(\hat{Y} = y|A = a', Y = y)$$

P: Probability

$\hat{Y}$: The predictions

Y: The ground truth (labelled data)

A: Protected characteristics

Loan Example

	Predicted Values	
Actual Values	True Positives	False Negatives
	False Positives	True Negatives

Loan Example

- We can give out 10 loans
- We have 100 applicants
  - 70 come from affluent backgrounds
  - 30 come from low-income backgrounds
- Y = {Granted, Rejected}
- 50% of candidates from both groups are bad candidates

Protected Attribute: A

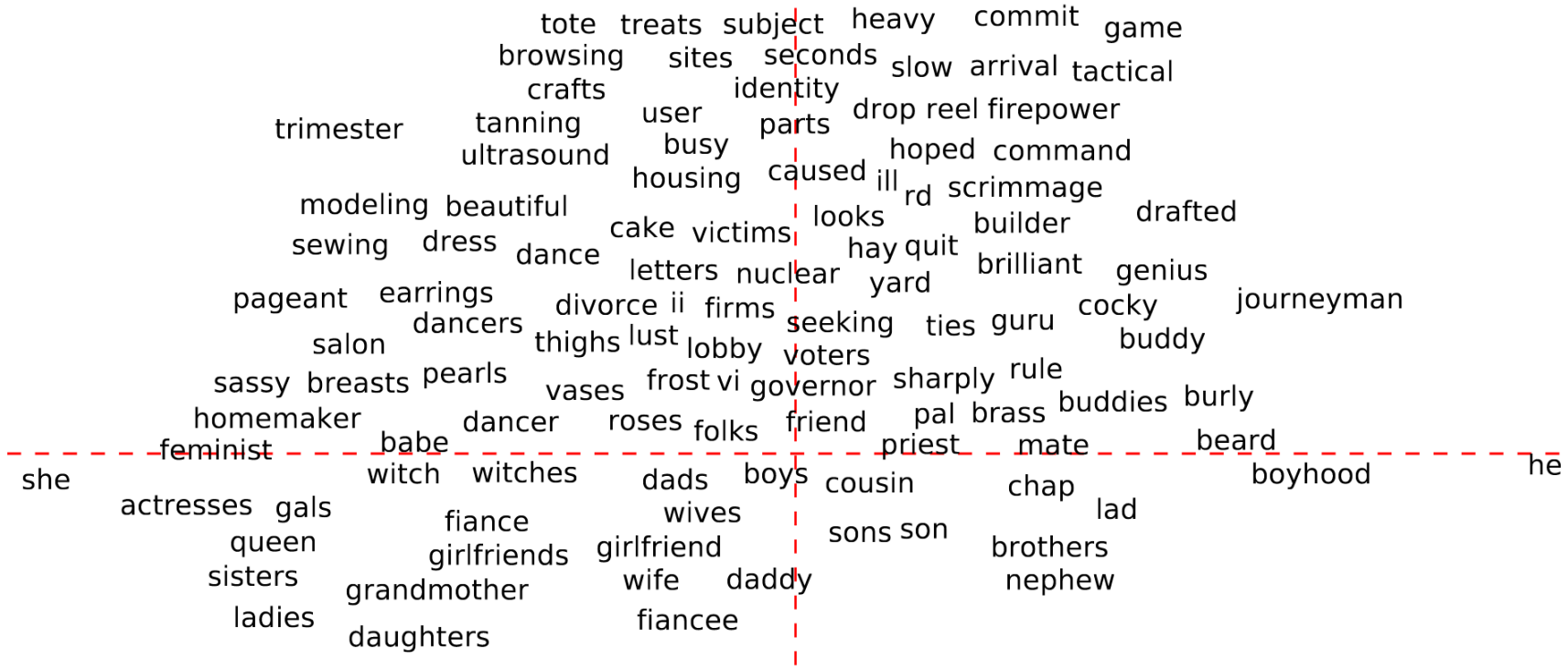
Loan Example

$$\text{Qualified applicants with affluent background} = 35$$
$$\text{Qualified applicants with low-income background} = 15$$

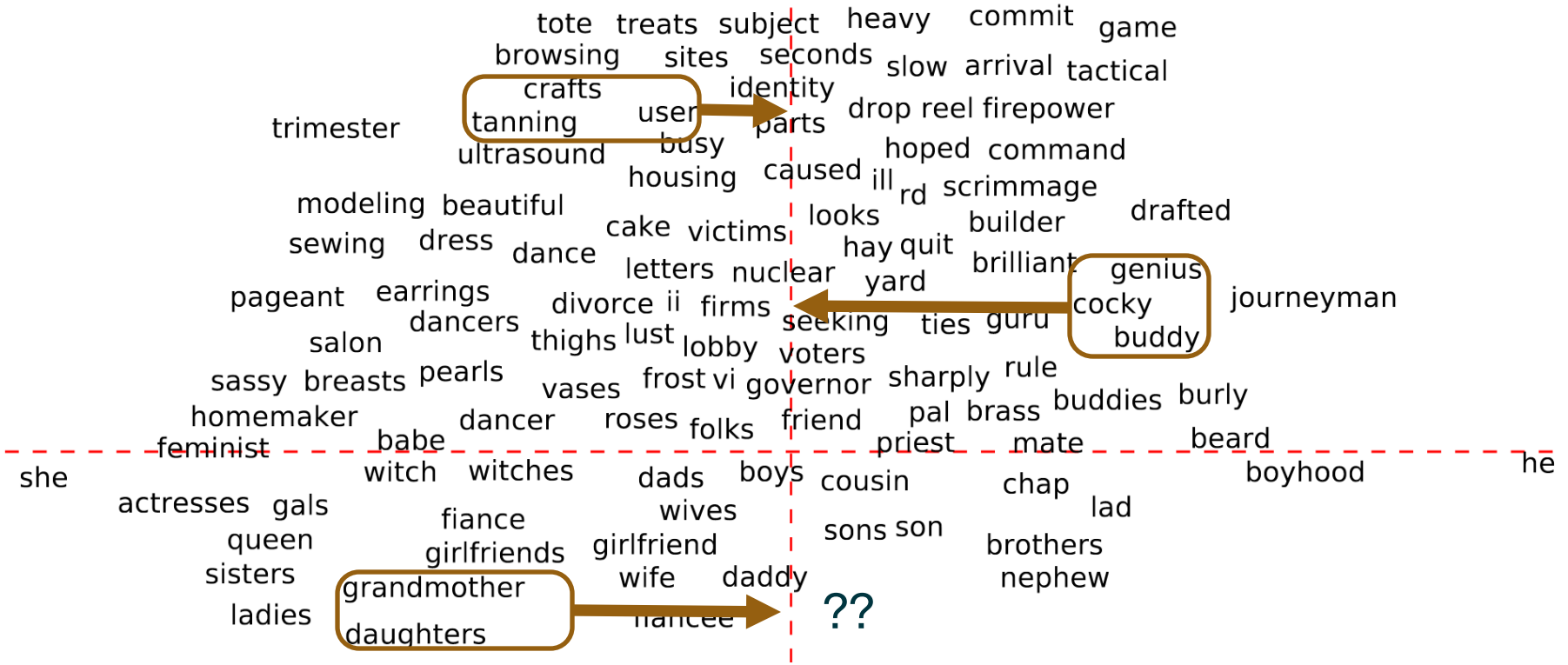
$$FPR_{affluent} = \frac{0}{35} = 0.0 \qquad TPR_{affluent} = \frac{7}{35} = 0.20$$

$$FPR_{low-income} = \frac{0}{15} = 0.0 \qquad TPR_{low-income} = \frac{3}{15} = 0.20$$

Methods in NLP



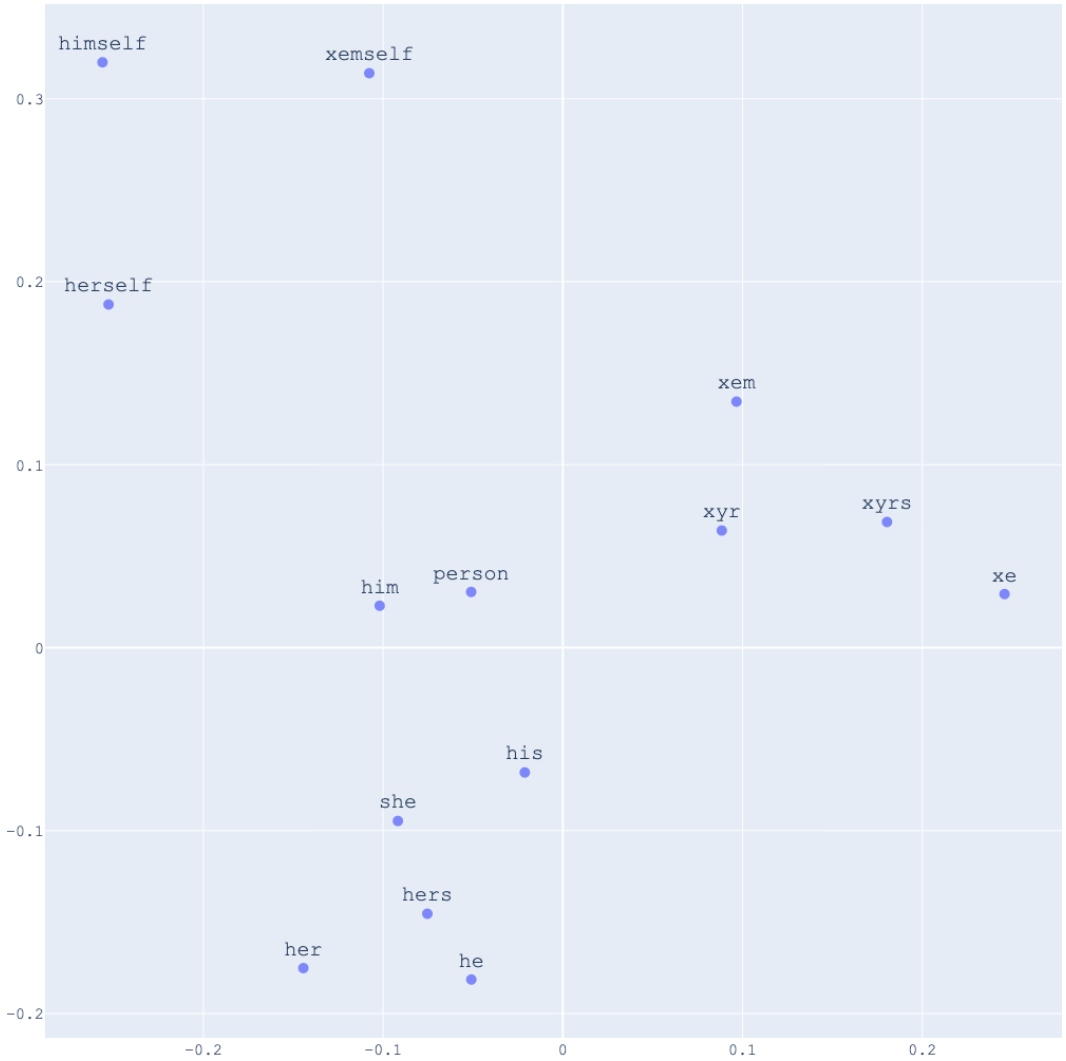
Methods in NLP



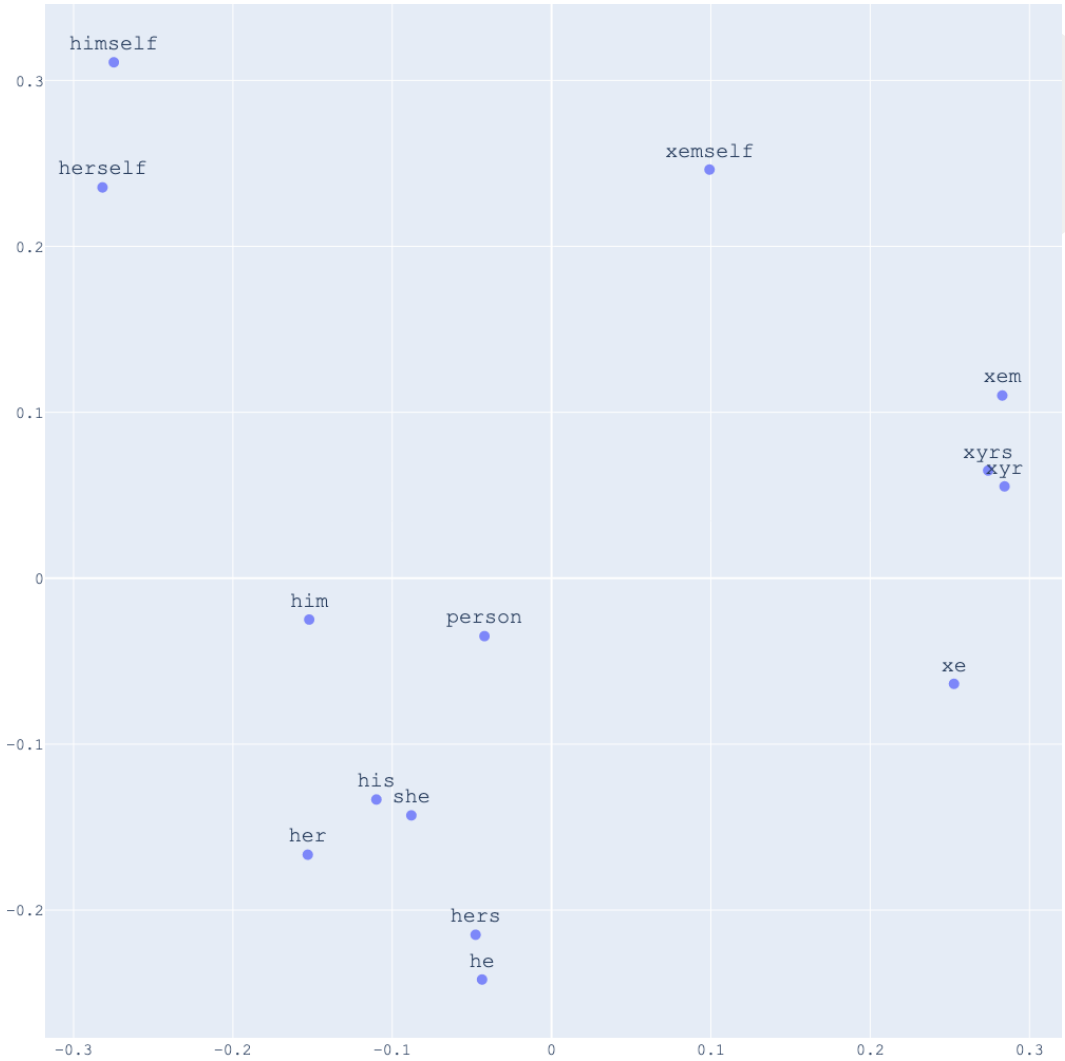
Extrinsic Bias Evaluations

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Examples of templates
Source: Dixon et al. (2018) Measuring and Mitigating Unintended Bias in Text Classification. AIES.



Adapter Tuned



original

Reinforcement Learning with Human Feedback (RLHF)

- Goal of RLHF is to use human rankings of model outputs to shift model weights

It’s Tricky: Evaluation and Mitigation in GenAI

- Evaluating generative AI is tricky

- Language is infinite
- Adversaries for fairness in AI
  - In non-generative AI it is some human
  - In generative AI it is the generative AI system

It’s Tricky: Hallucinations and Training GenAI systems

- Training language models optimises for two tasks:
  - Predicting the hidden word in a sequence of words
    - E.g., “Zee stood in front of the [MASK] today.”
  - Given one sentence, predicting if the next sentence follows the first.
    - E.g., “Zee stood in front of the group today.” and “Zee will have sore feet tonight.”

So what can we do?

- Narrow the scope of what AI systems can do
  - This affords greater control and specificity to them
- When developing evaluation methods
  - Ensure that our work is extensible
  - Resources and documentation is available
  - Guidelines for how to extend
- Explicate what the methods are and are not suited

Part III
Doing CSS with ML ethically

How do we do ethical work?

- Ensure methods match research questions/purposes
- Make sure that your questions are meaningful and valid
- Make sure the outcomes of your work are not harmful
- And that releasing your work/using it does not cause harm

What are ways CSS methods can harm?

- Two types of harms:
  - Representational harms
  - Allocative harms

What are ways CSS methods can harm?

- Language technologies are always socially biased
- Network methods can cause harms, in particular, through surveillance
- Data Science can cause harm by correlating actors with actions/attributes