The Simpson and Bias Amplification Paradoxes

Or, What Can Go Wrong With My Evaluation?

Yanai Elazar - 20.3.2024 - Dagstuhl Seminar

Hi There

Yanai Elazar

Postdoc @ Allen Institute for AI & University of Washington

Hi There

Yanai Elazar

Postdoc @ Allen Institute for AI & University of Washington

I work on

The Science of “Language Models”

- How, when, and what make them work, and not work
- Connecting training data to model behavior

The Simpson's Paradox

A brief introduction to the drunk

The beer paradox

Biases in University Admissions?

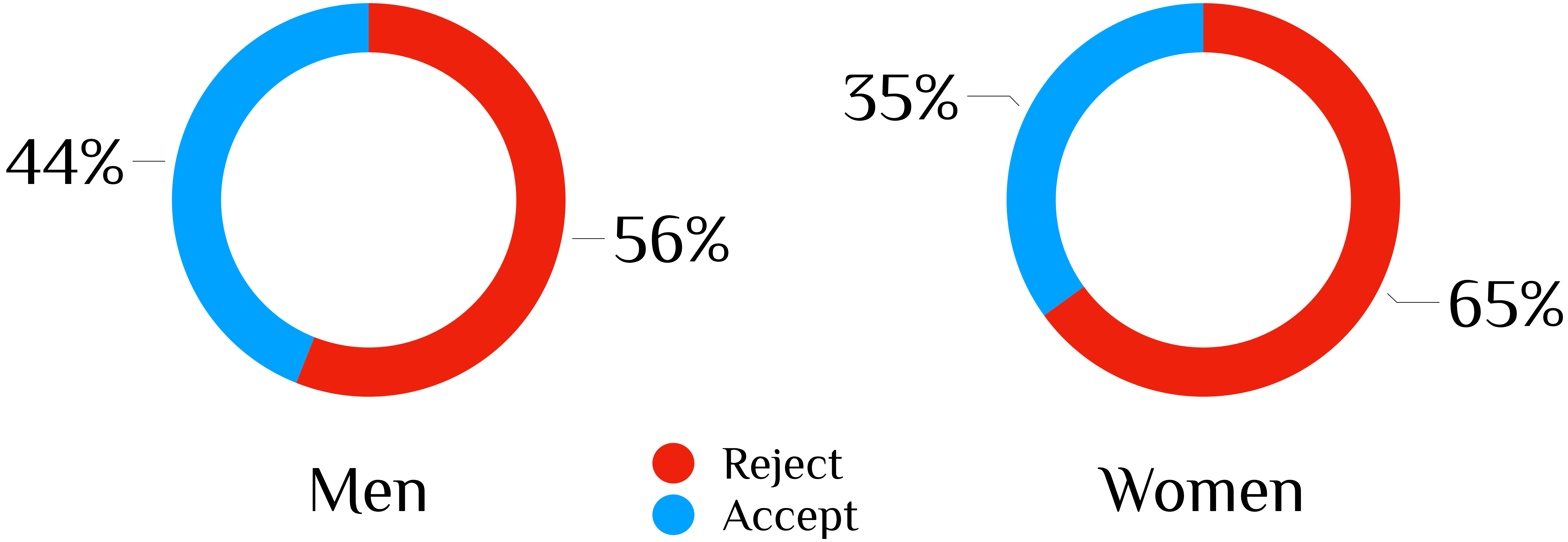
Sex Bias in Graduate Admissions: Data from Berkeley

P. J. Bickel, E. A. Hammel, J. W. O'Connell

— "Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation"

Biases in University Admissions?

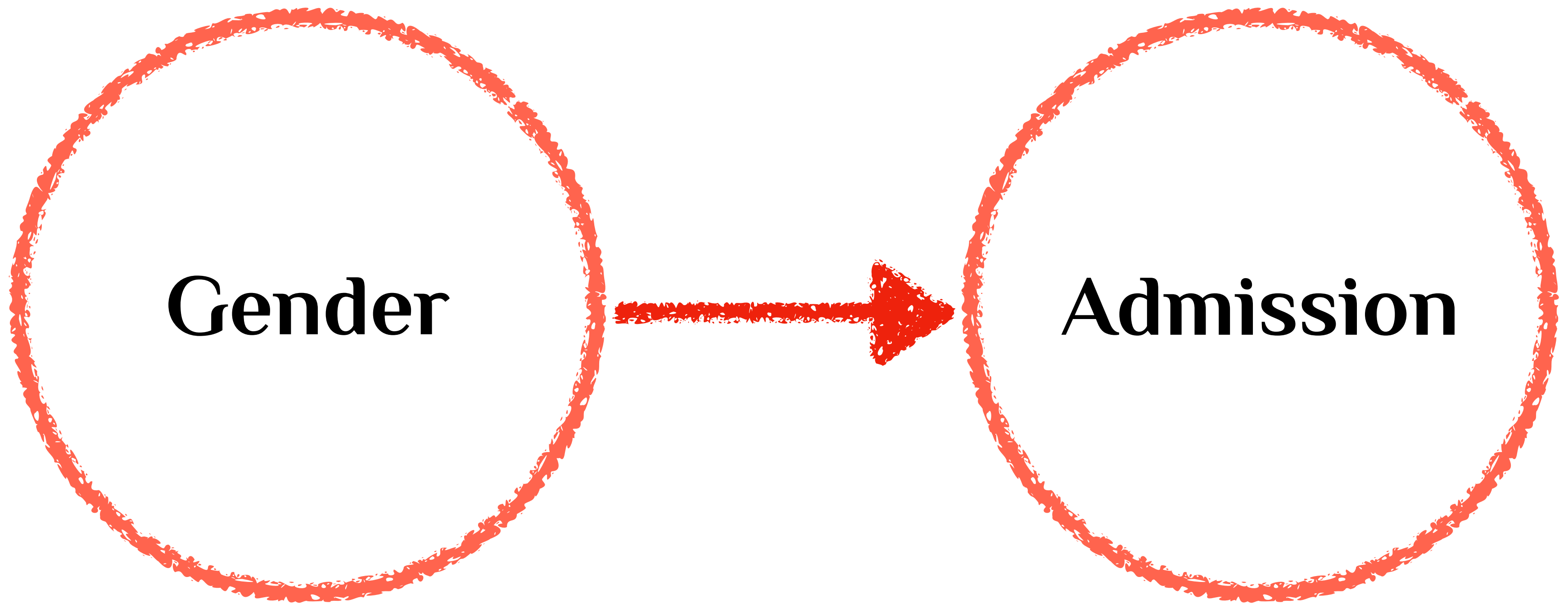
Let's look at some data
- Aggregated data
- Split by gender

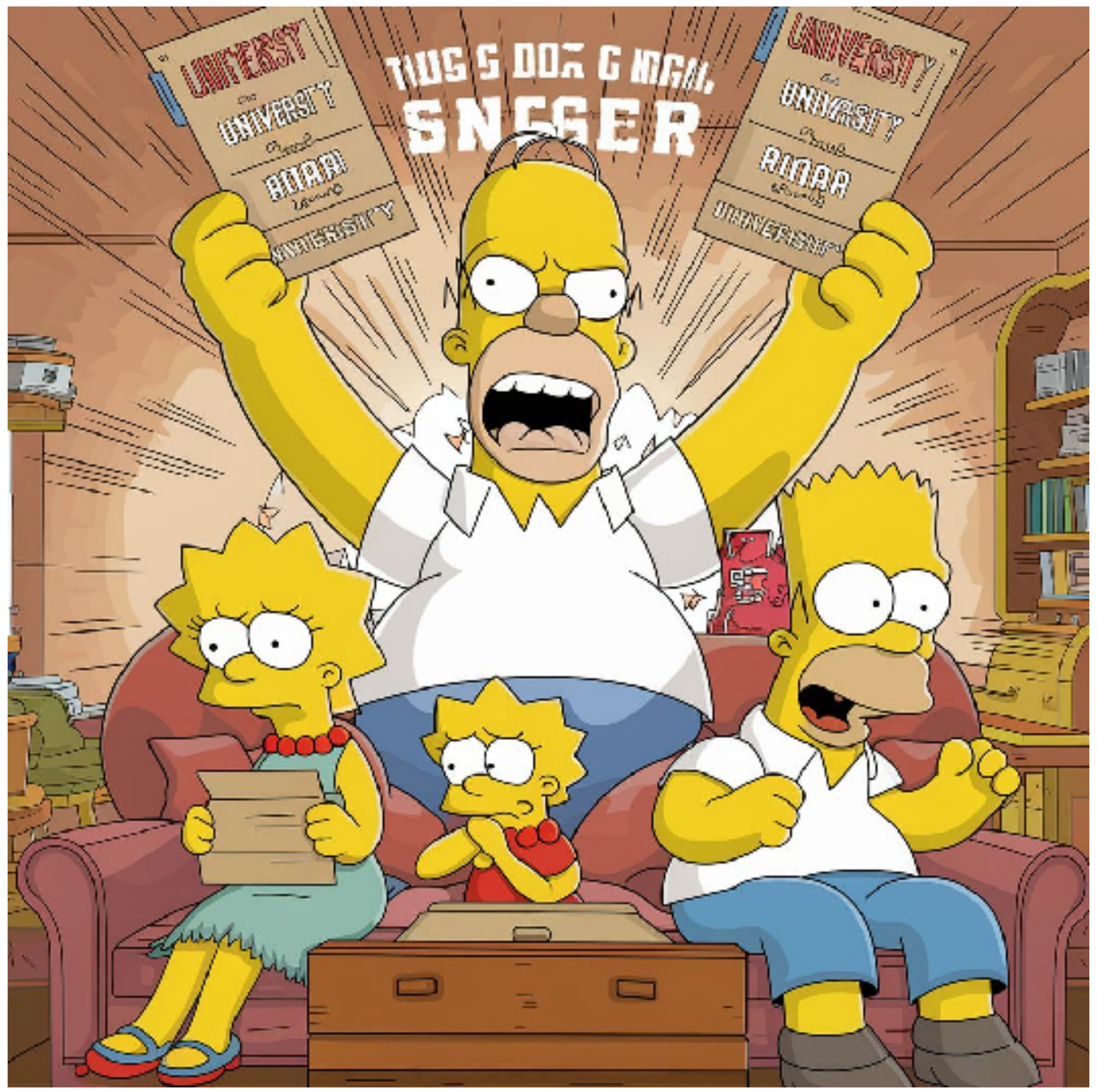


The decision process seems to be biased!

Biases in University Admissions?

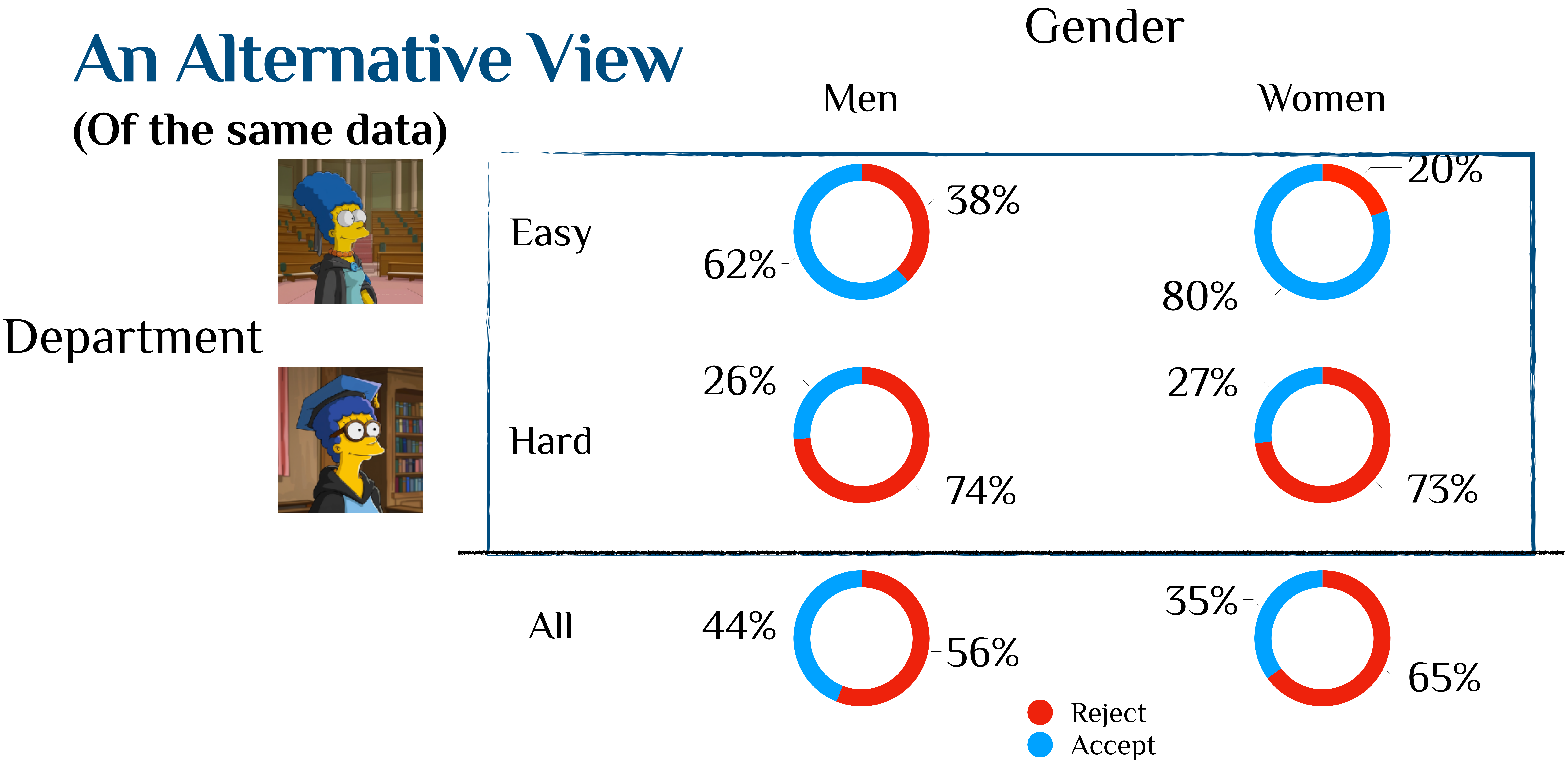
- Very simple and intuitive analysis
- Matches our intuition (and previous studies) about societal biases
- Start looking for culprits?





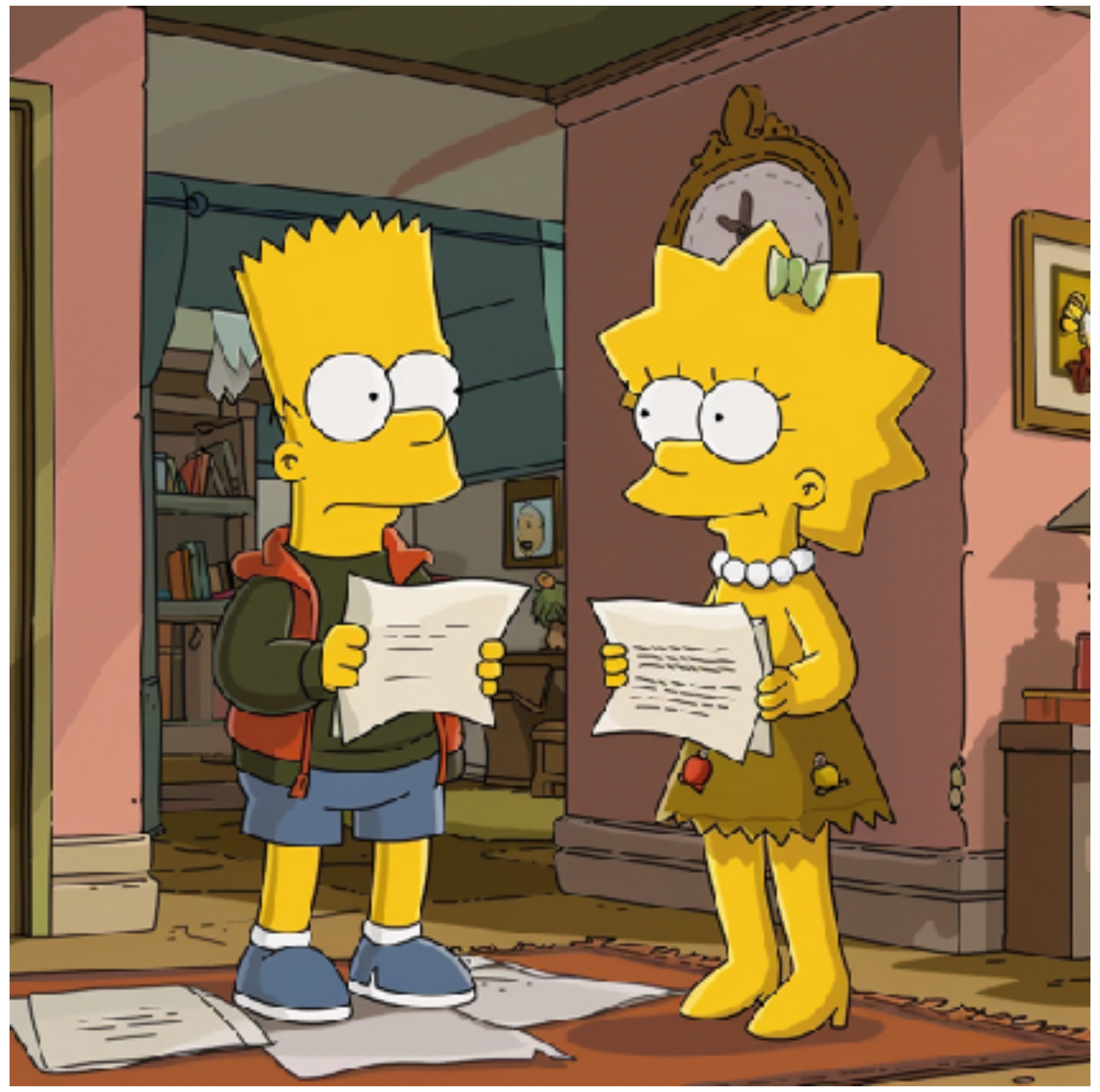
An Alternative View

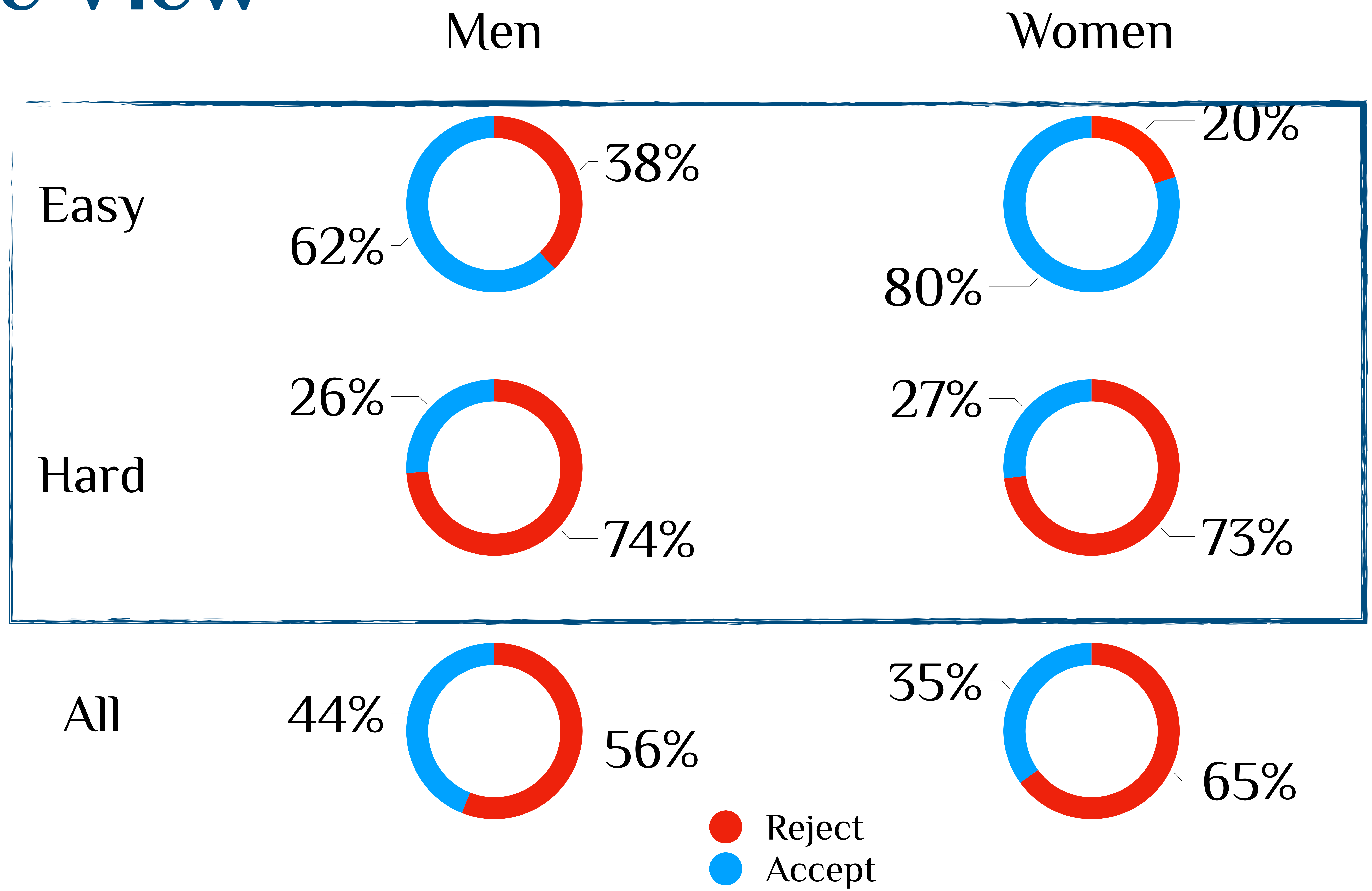
(Of the same data)



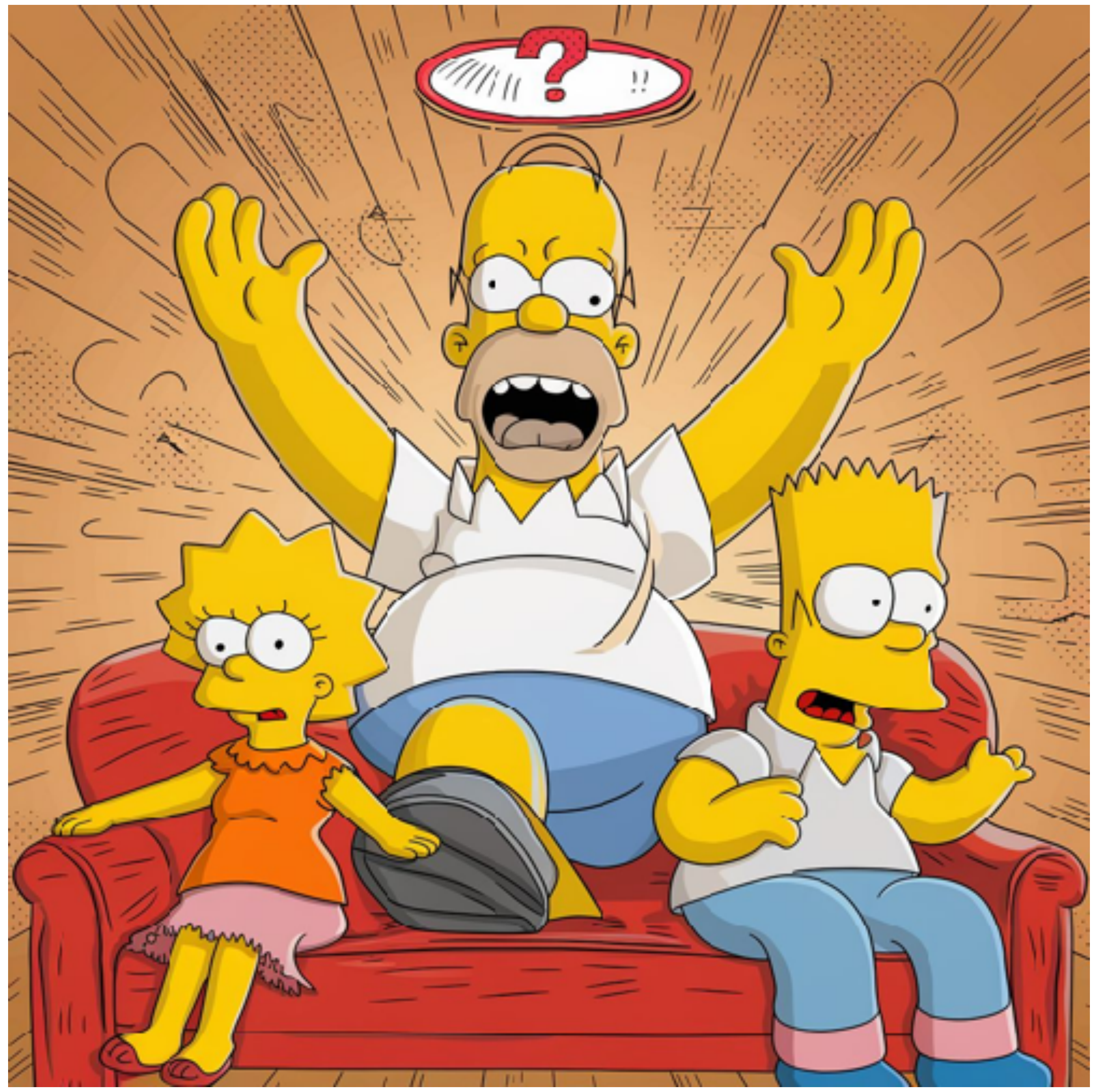
An Alternative View

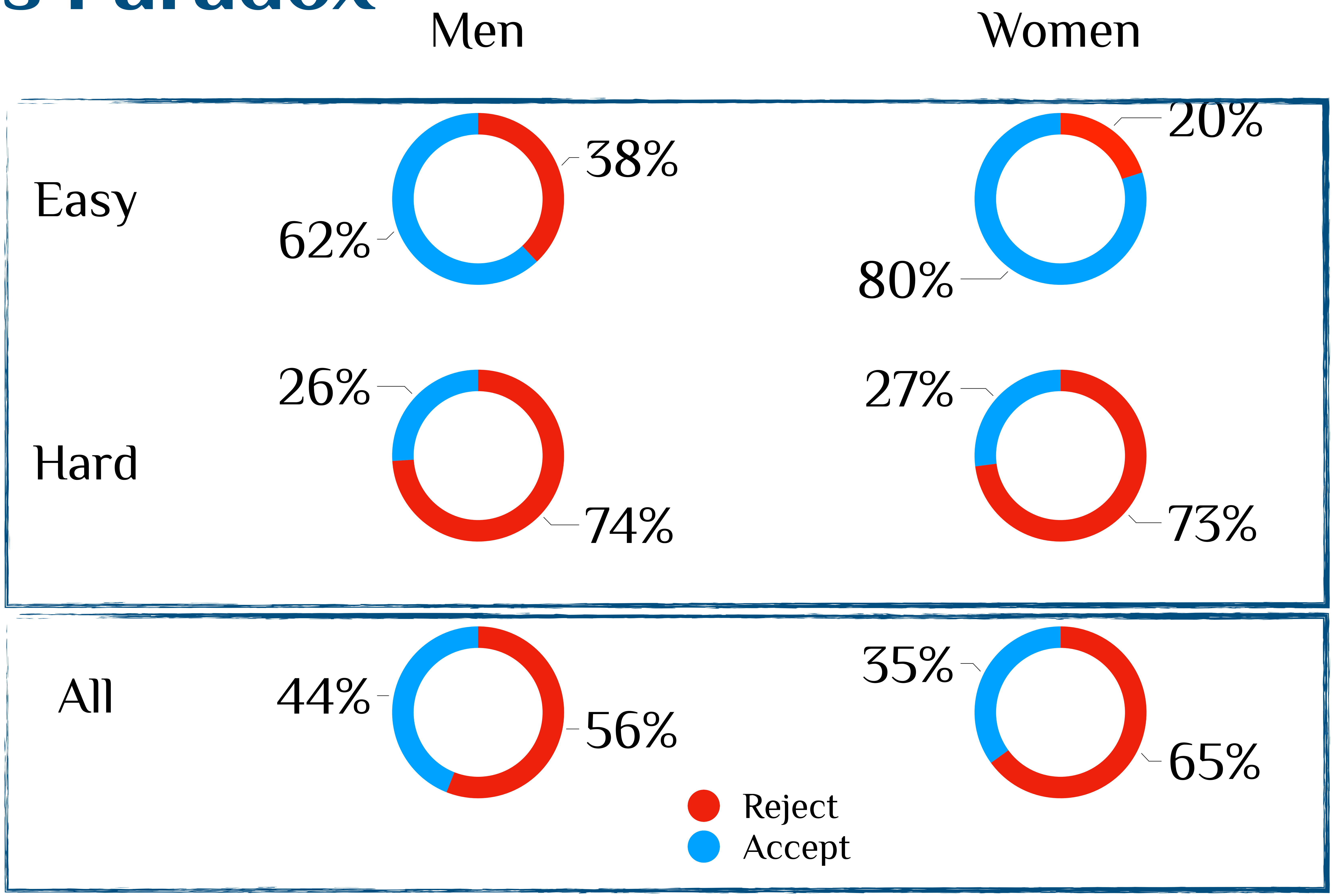
(Of the same data)



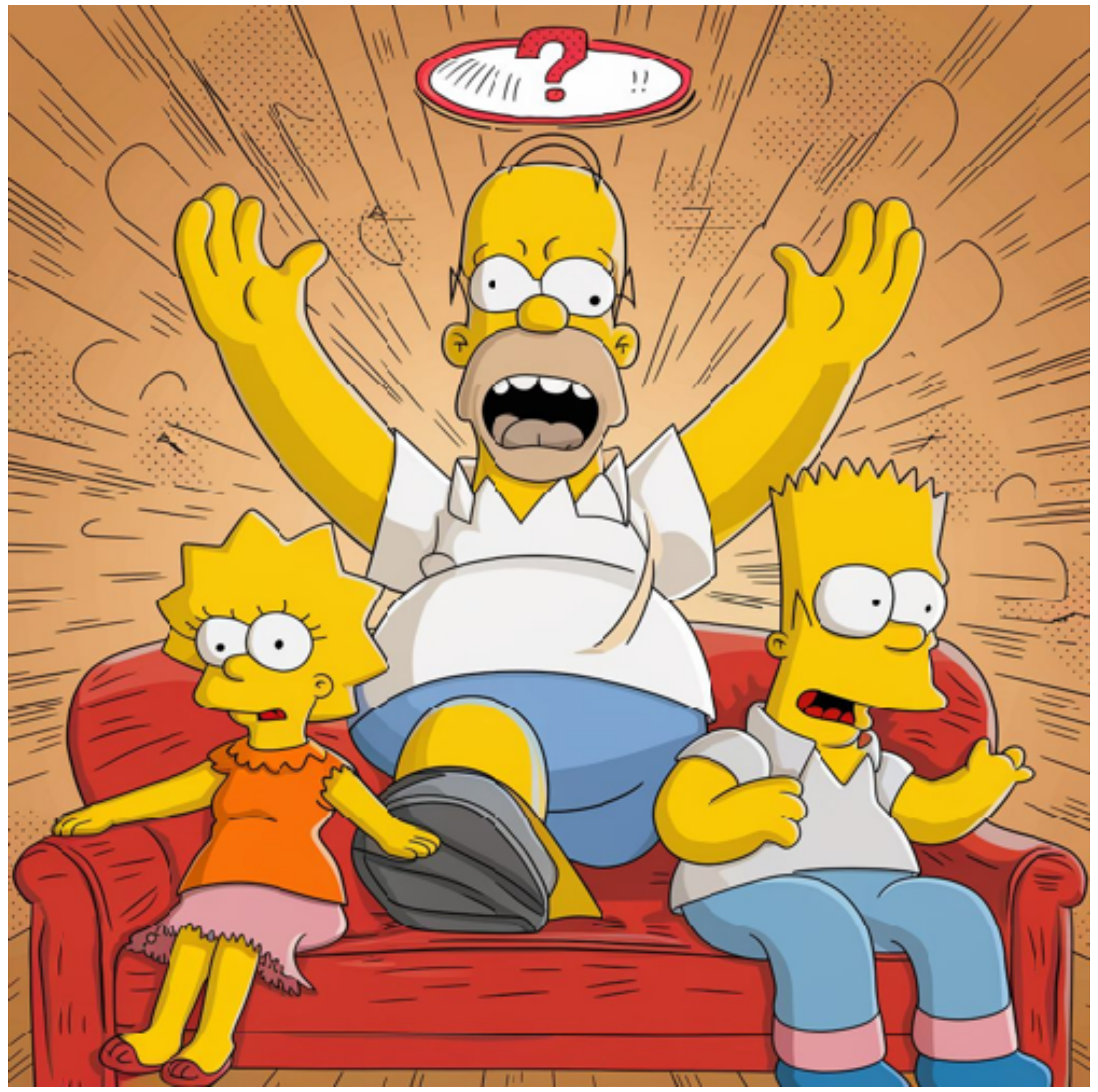


The Simpson's Paradox





The Simpson's Paradox

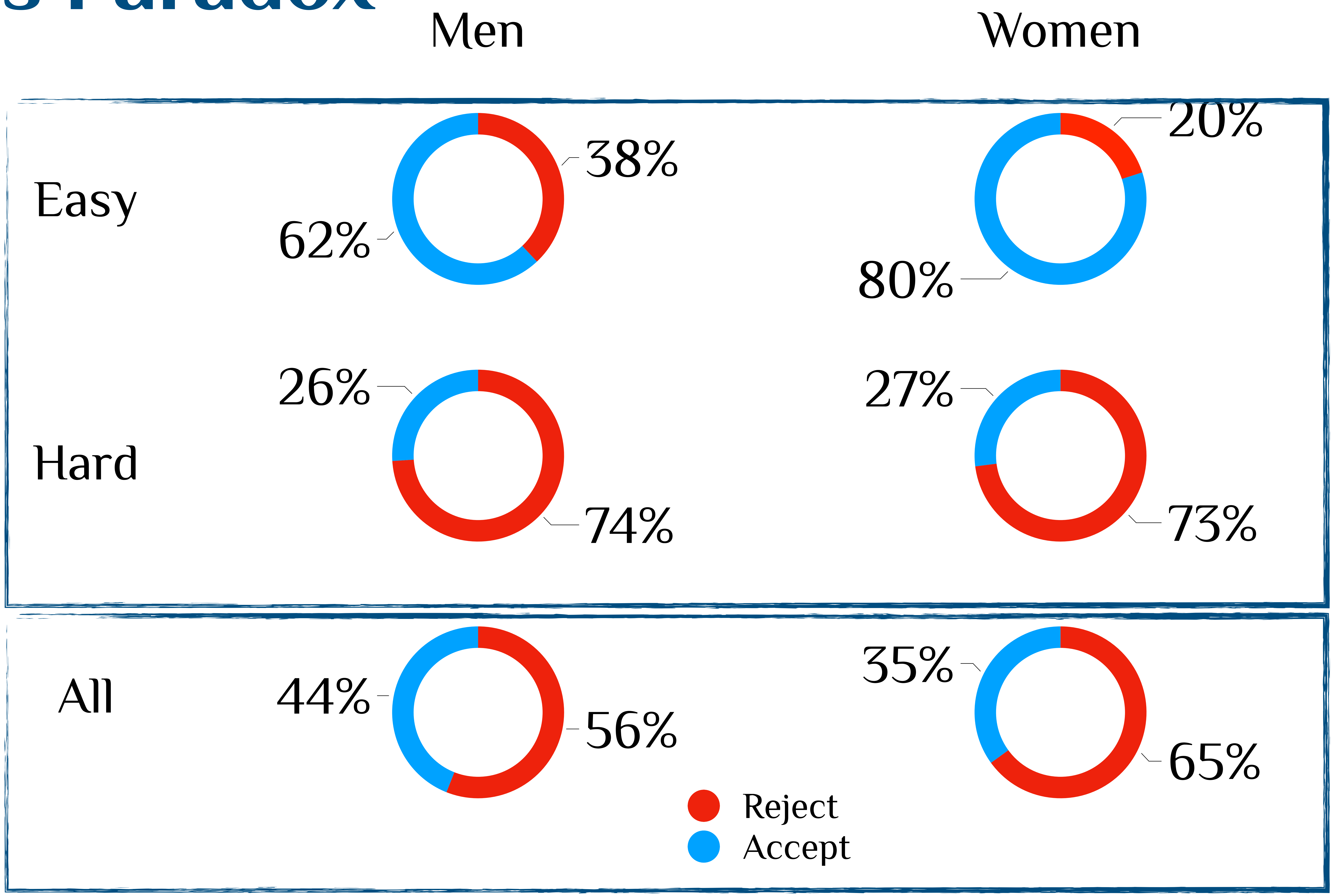


	Men		Women
Easy	62%	<	80%
Hard	26%	<	27%
All	44%	>	35%

Reject
Accept

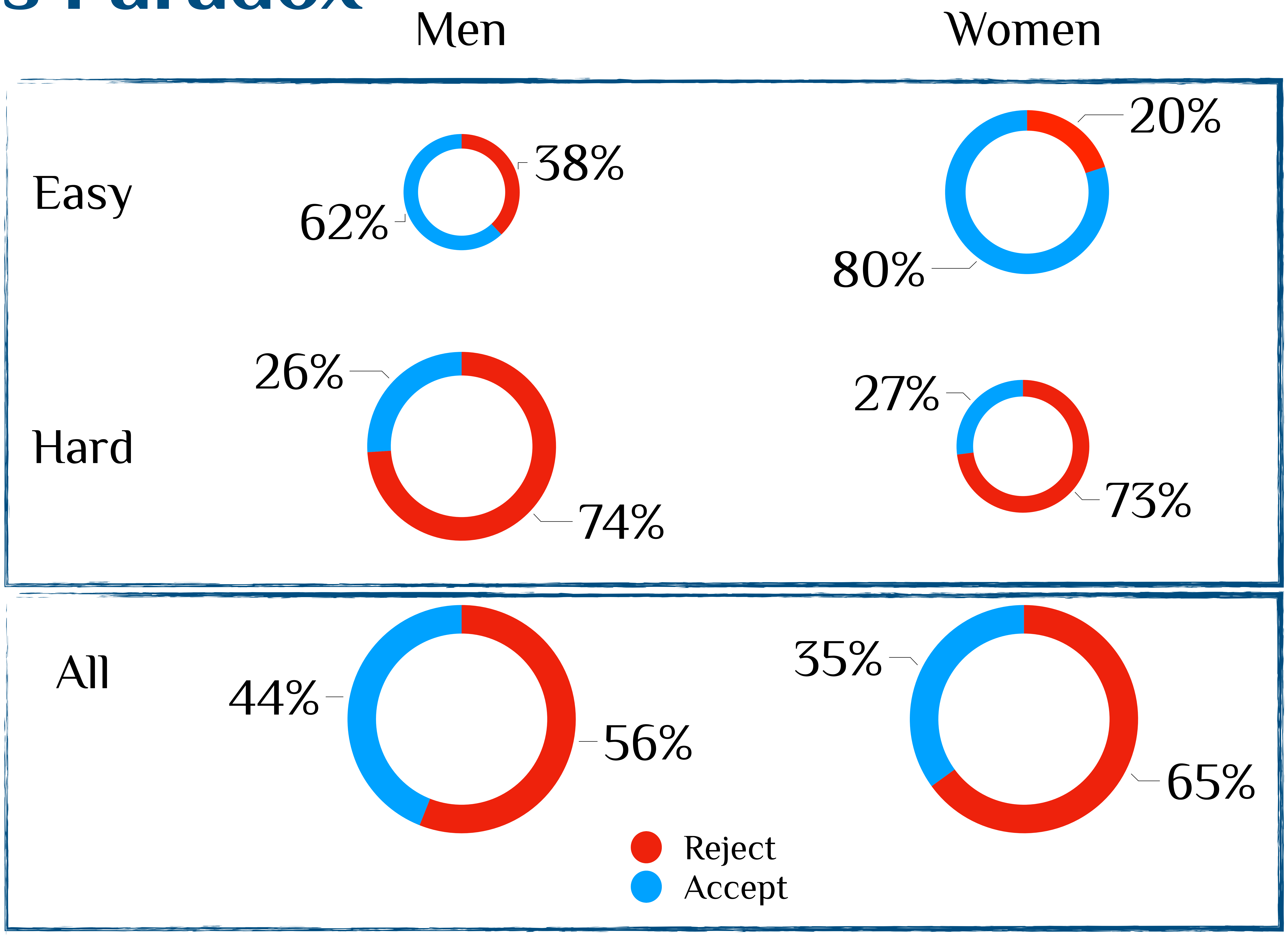
The Simpson's Paradox

Explanation



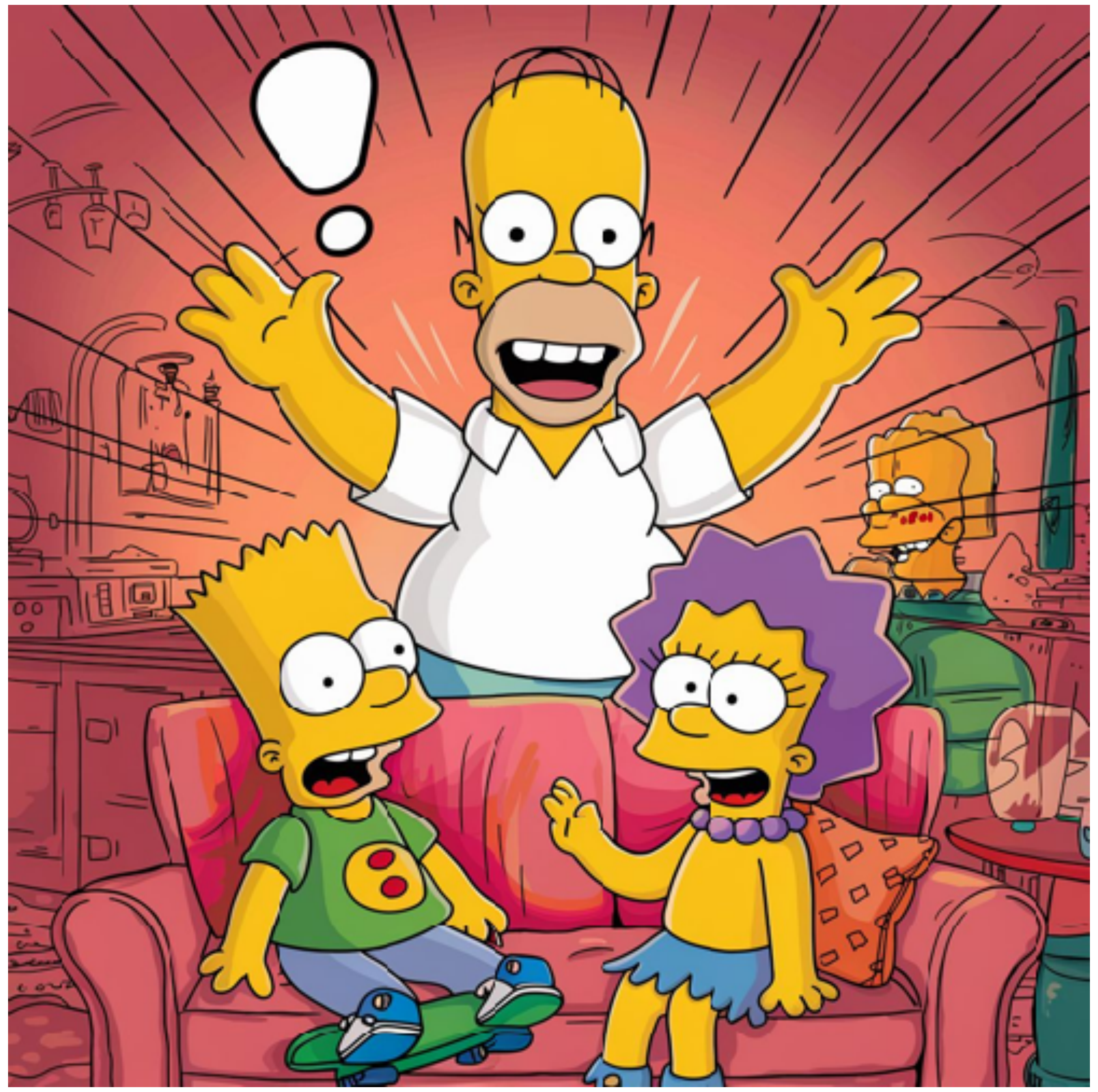
The Simpson's Paradox

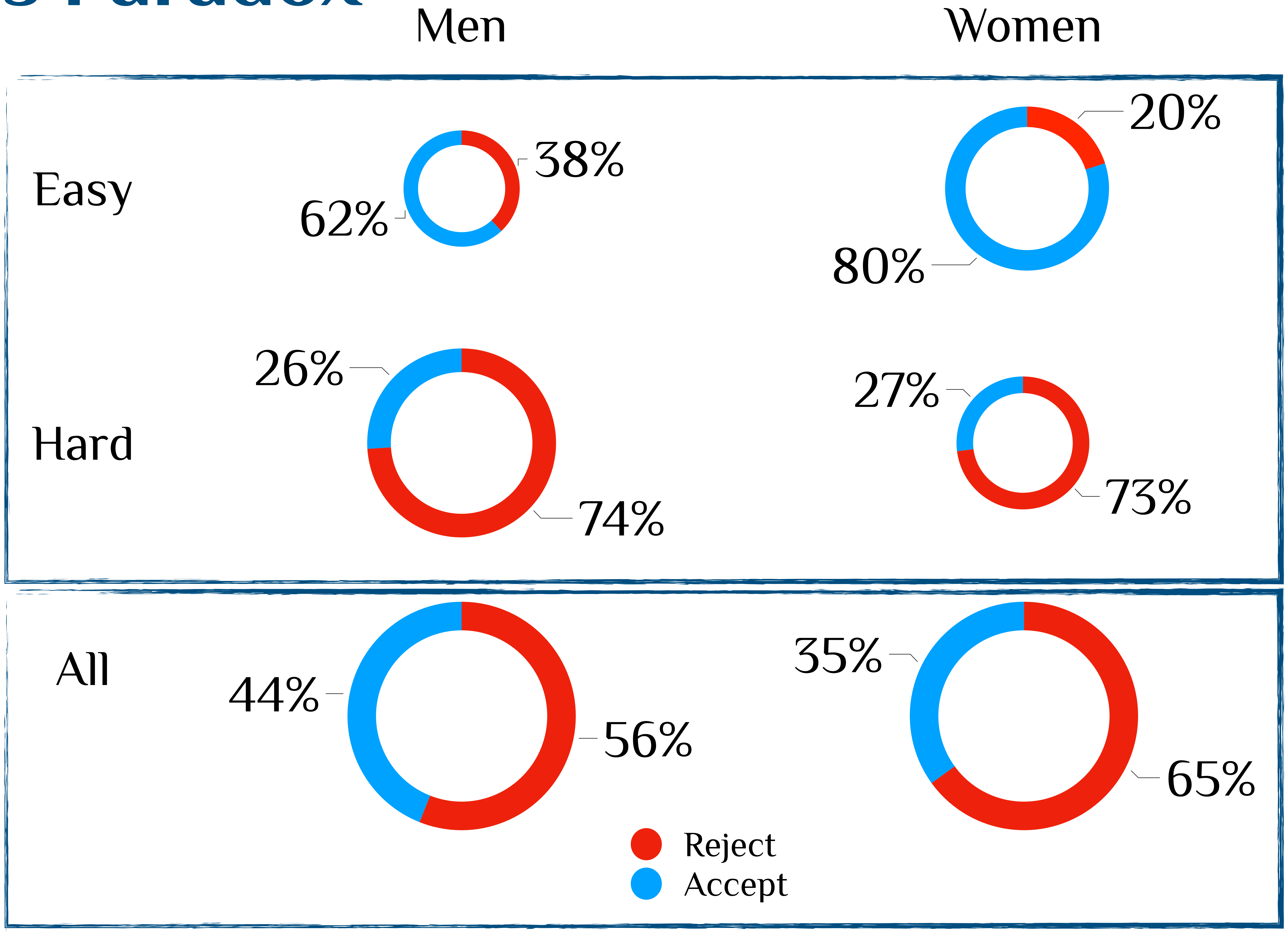
Explanation



The Simpson's Paradox

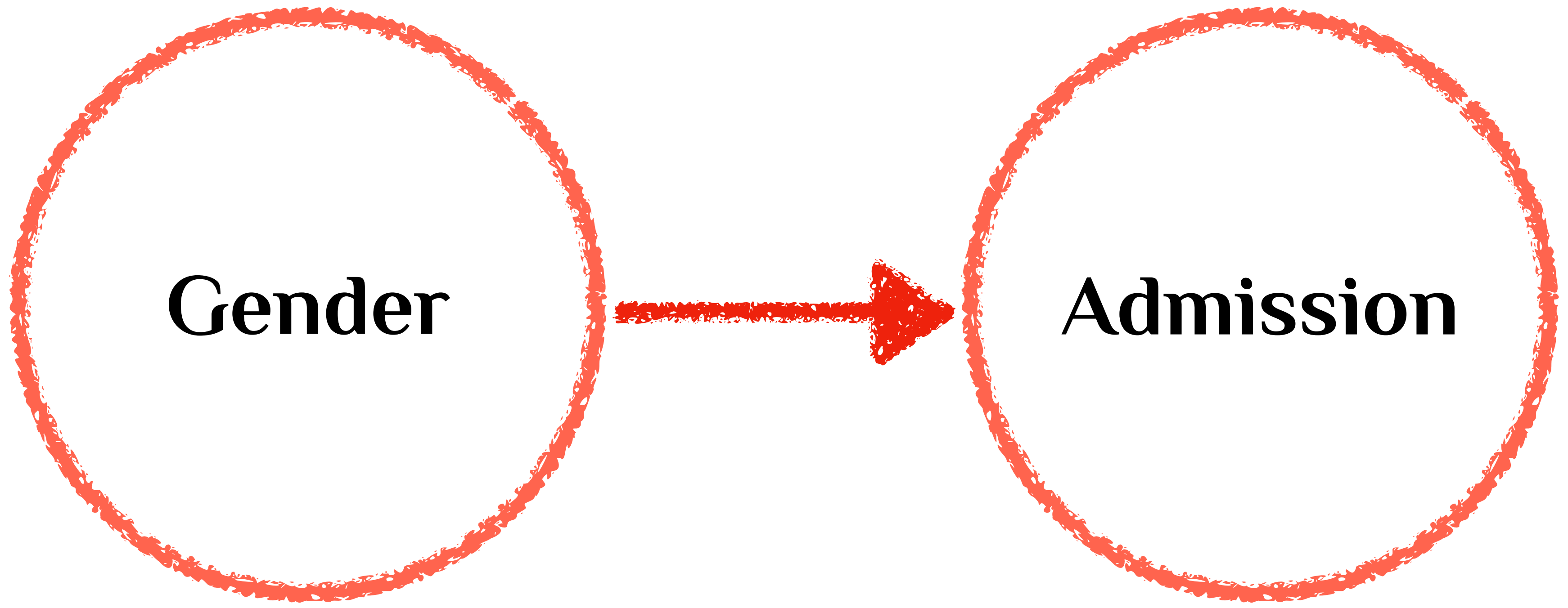
Paradox?



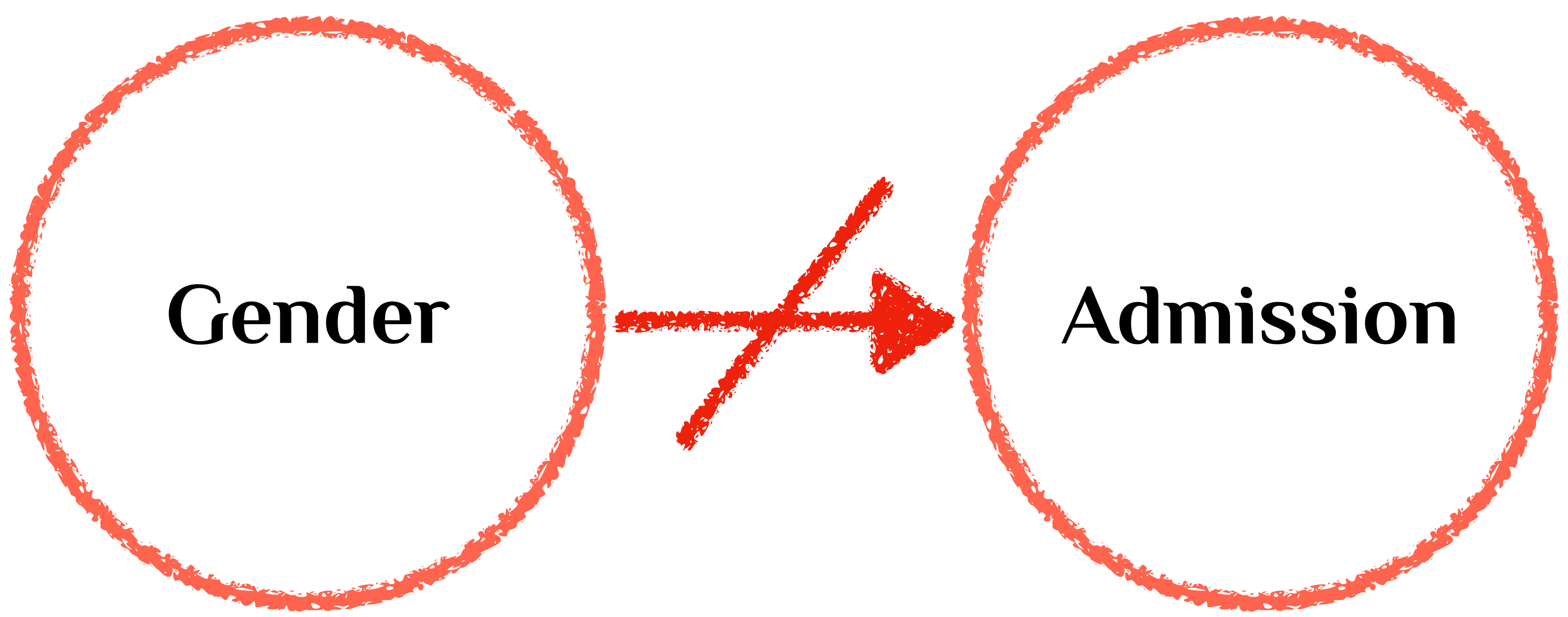


A Causal Perspective

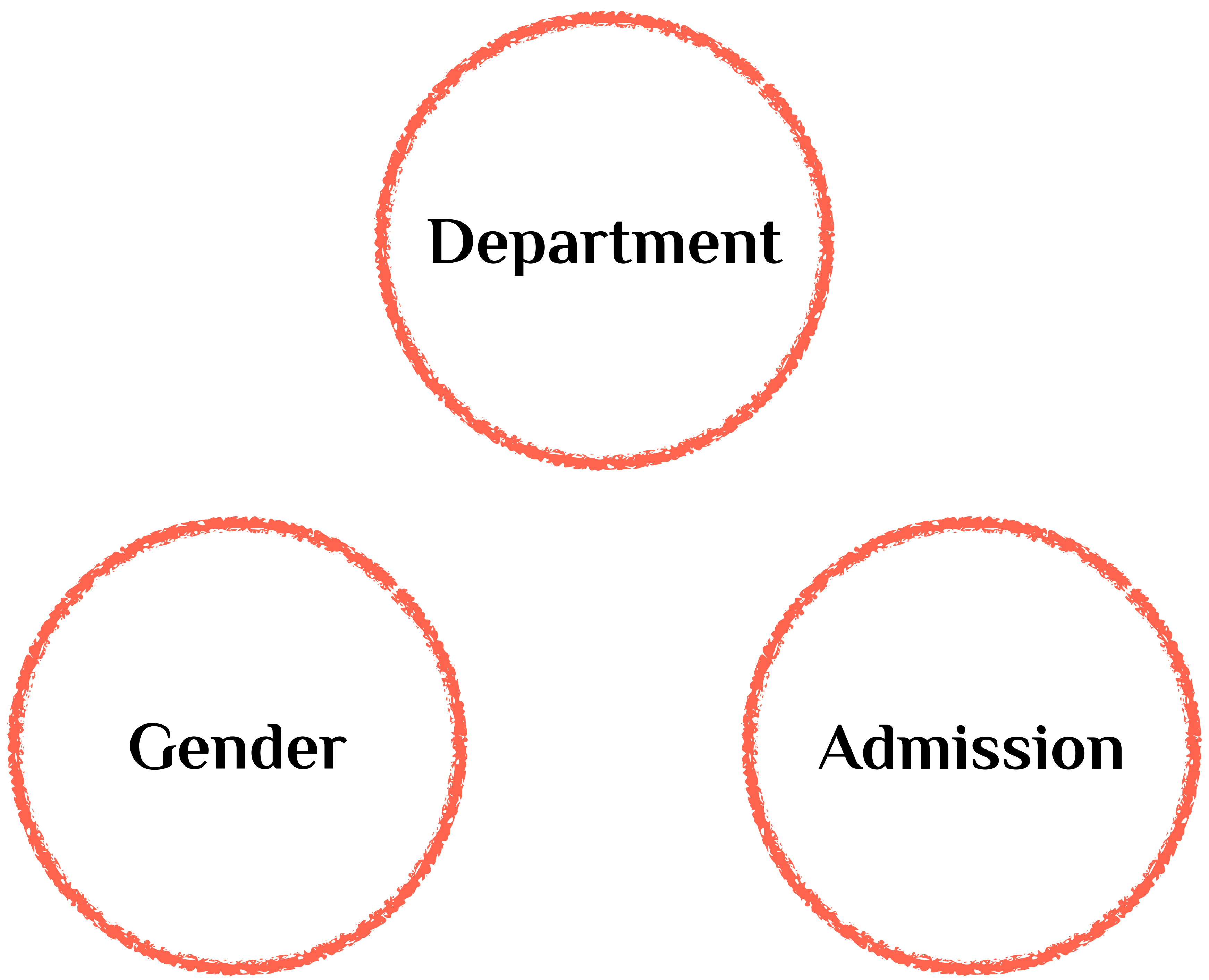
of the Simpson's paradox



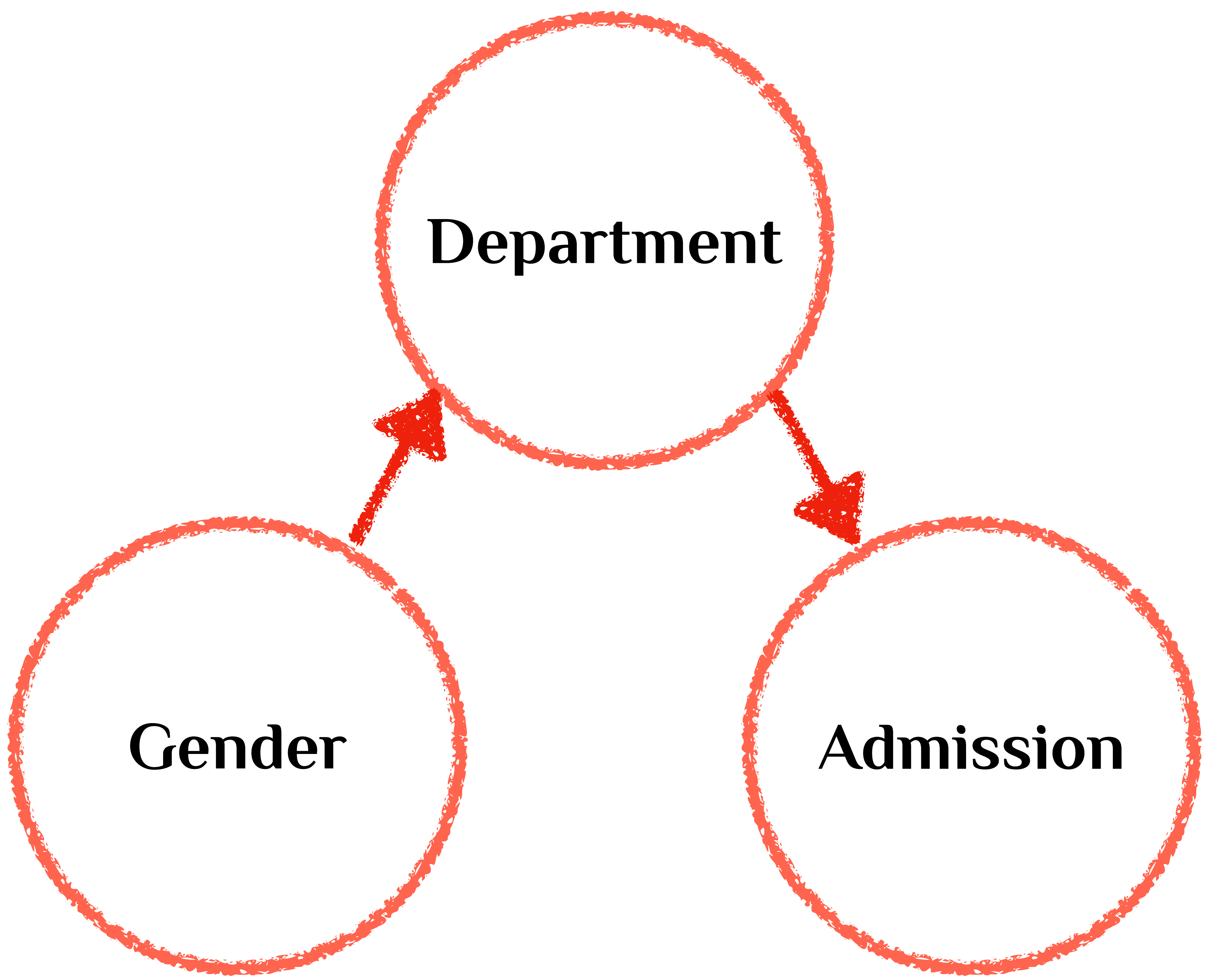
A Causal Perspective



A Causal Perspective

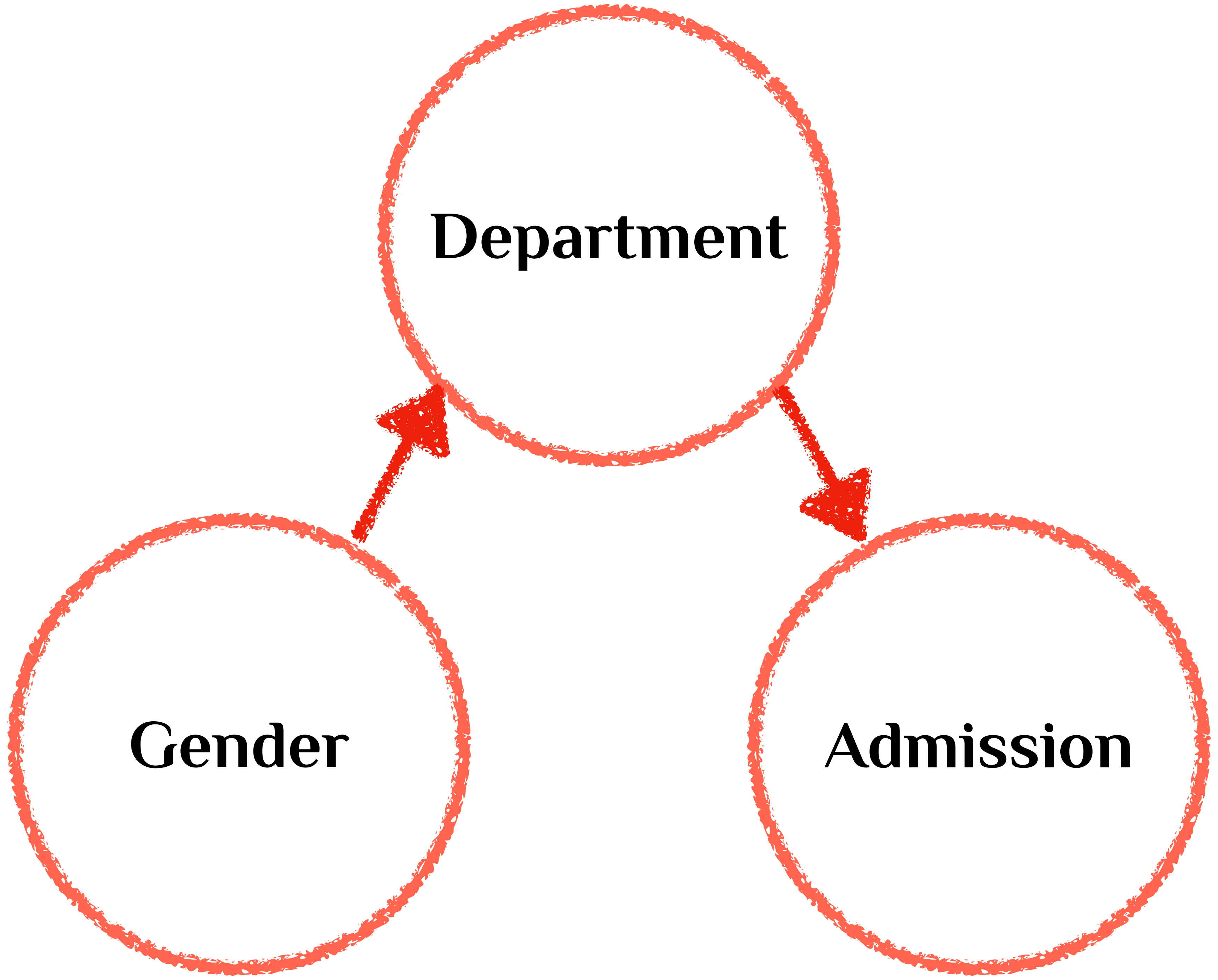


A Causal Perspective



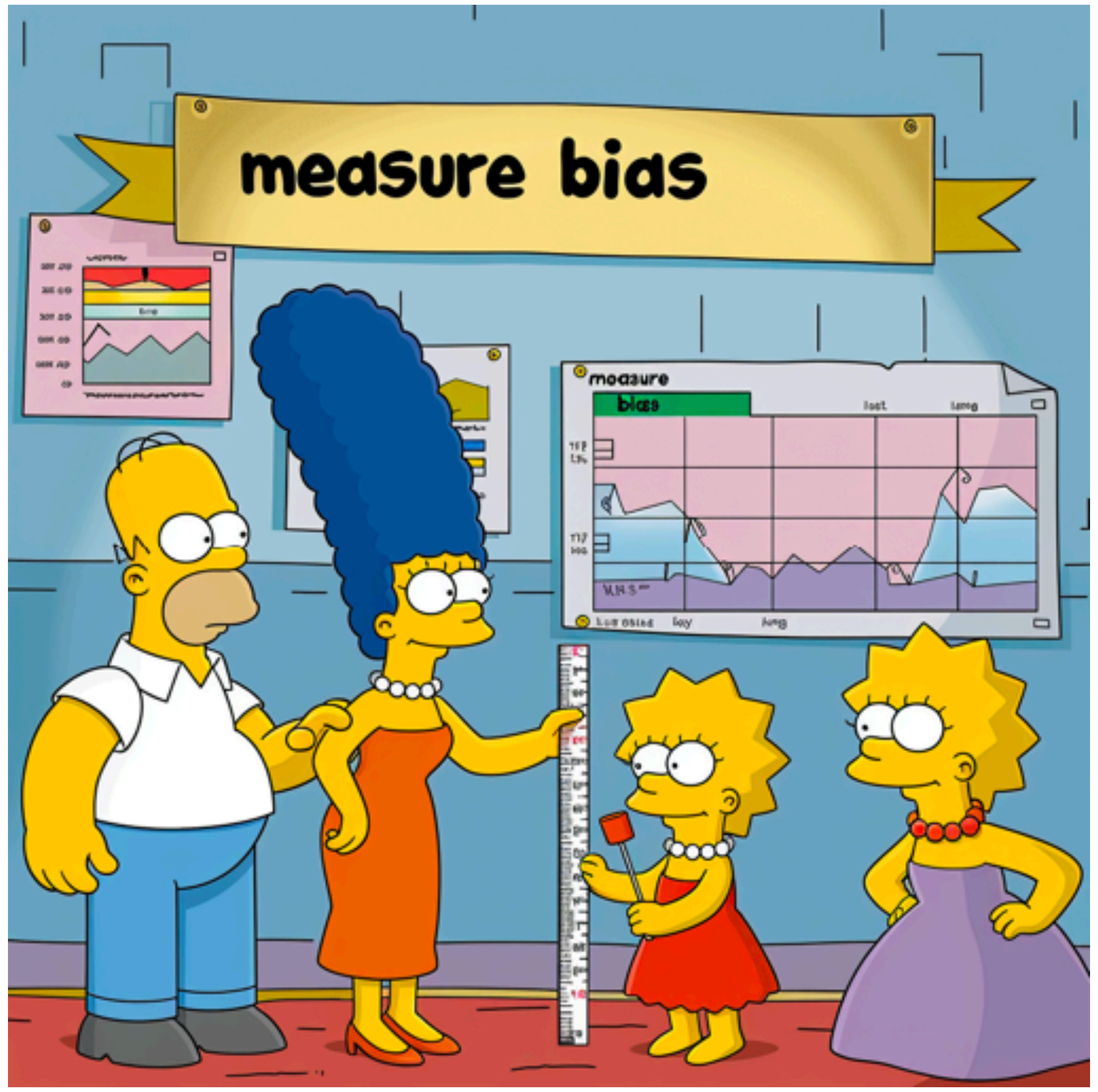
A Causal Perspective

- The ***department*** acts as a mediator, leading to a wrong conclusion



The Simpson Paradox

— "Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation"



https://setosa.io/simpsons/

The Bias Amplification Paradox

NAACL '24

Models are Biased

- Models encode and exhibit different biases
- This is not a new finding,
  and is a well known and documented phenomenon

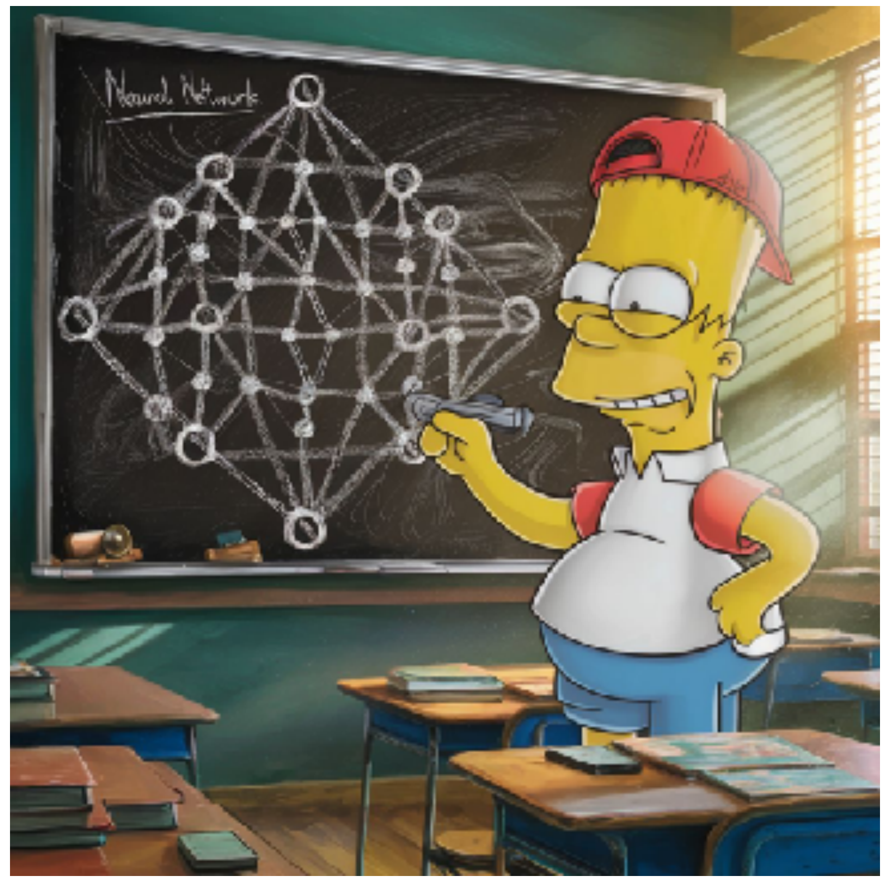
Let’s Try It Out!

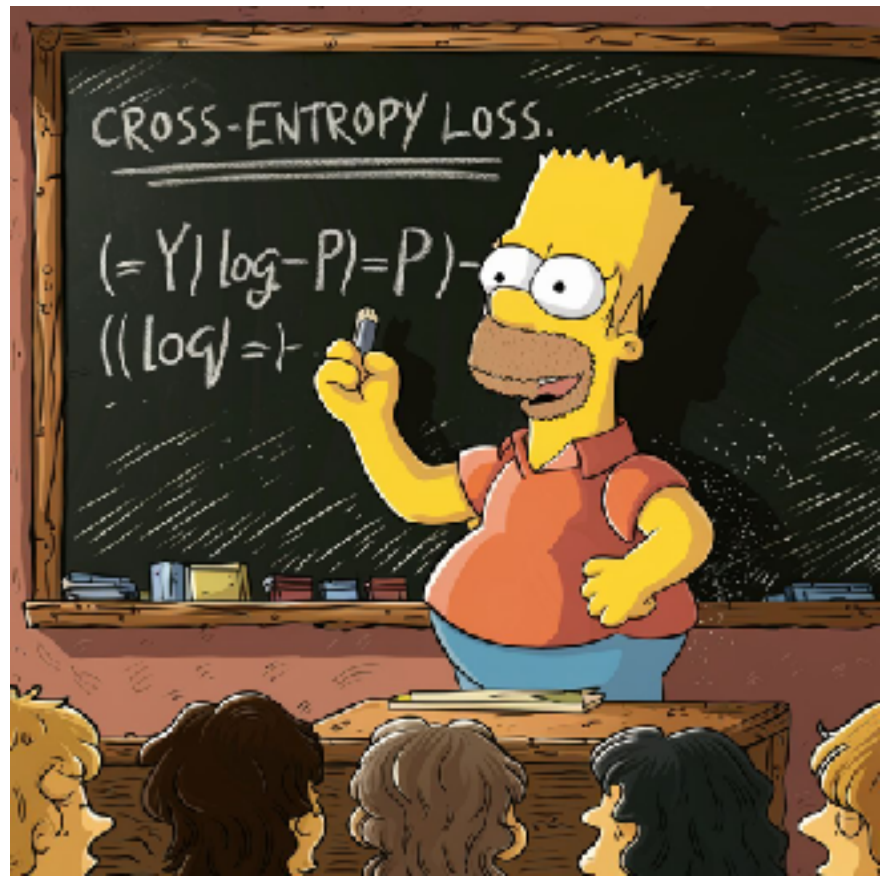
A photo of a face of an engineer

1/10 women!

The model is biased!

Where Does The Bias Come From?





Let's Look At The Data

The Data is Huge!

2 billion image-caption pairs!

LAION

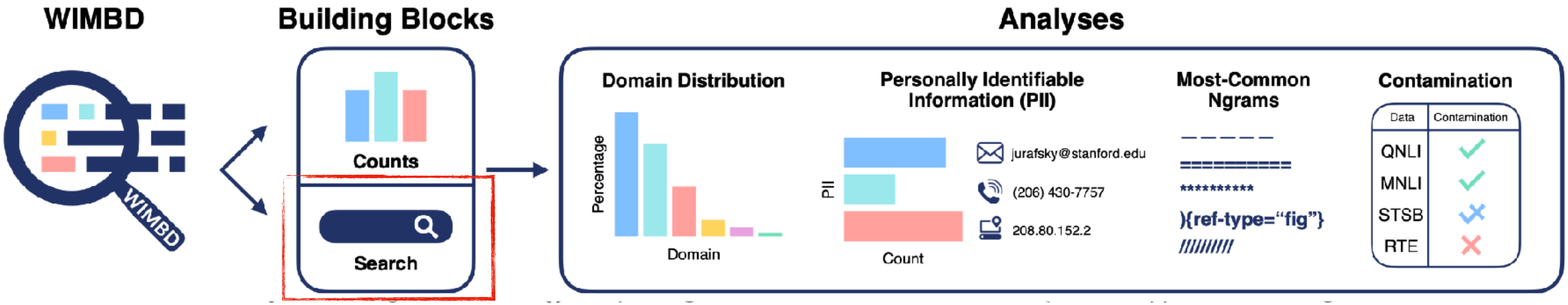
Large-scale Artificial Intelligence Open Network

TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

LAION, as a non-profit organization, provides datasets, tools and models to liberate machine learning research. By doing so, we encourage open public education and a more environment-friendly use of resources by reusing existing datasets and models.

Where Does The Bias Come From?

WHAT'S IN MY BIG DATA?



ICLR '24

Where Does The Bias Come From?

- Using the index from WIMBD, we have fast access to the data
- … and we can test such associations in the training data

Where Does The Bias Come From?

```python
from wimbd.es import get_documents_containing_phrases

# Get documents containing the term:
get_documents_containing_phrases("laion","engineer")
```

ENGINEER Chemical Engineer Civil Engineer Electrical Engineer Environmental Engineer Geological Engineer Materials Engineer Mechanical Engineer Mining

Engineer, Engineer Hat, Engineer Gift, Gift For Engineer, Student Engineer, Engineer Graduation, Engineer Uniform For Engineer Party

Engine Engineer Engineer Engineer Engineer - Women's Premium Tank Top

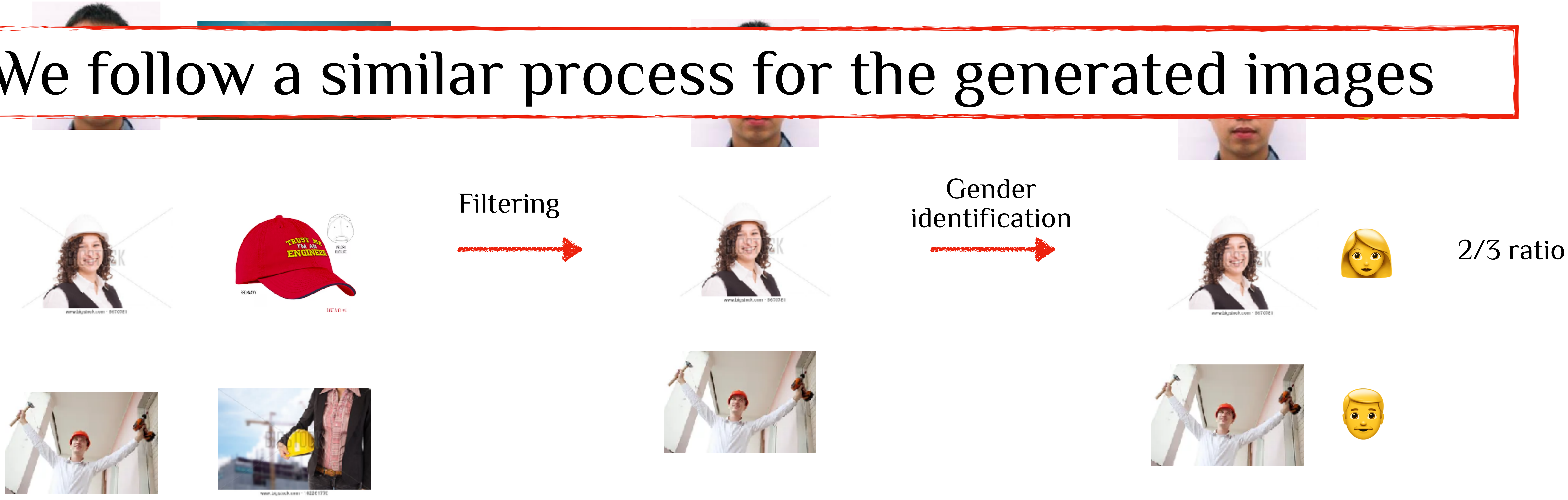
Establishing Data Gender Ratios

```python
from wimbd.es import get_documents_containing_phrases

# Get documents containing the term:
get_documents_containing_phrases("laion","engineer")
```

The data is large and noisy, so we need to adjust

We follow a similar process for the generated images



Setup

Input:

"homer simpson"

Stable Diffusion

Output:

Training

Model Evaluation

Gender

Occupation

LAION

Large-scale Artificial Intelligence Open Network

TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

LAION, as a non-profit organization, provides datasets, tools and models to liberate machine learning research. By doing so, we encourage open public education and a more environment-friendly use of resources by reusing existing datasets and models.

Data

Evaluation

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef
  - Engineer

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef
  - Engineer
  - Janitor

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef
  - Engineer
  - Janitor
  - Lawyer

Setup

- We sample image-caption pairs: 500 total
- 62 occupations:
  - Accountant
  - Chef
  - Engineer
  - Janitor
  - Lawyer
  - ...





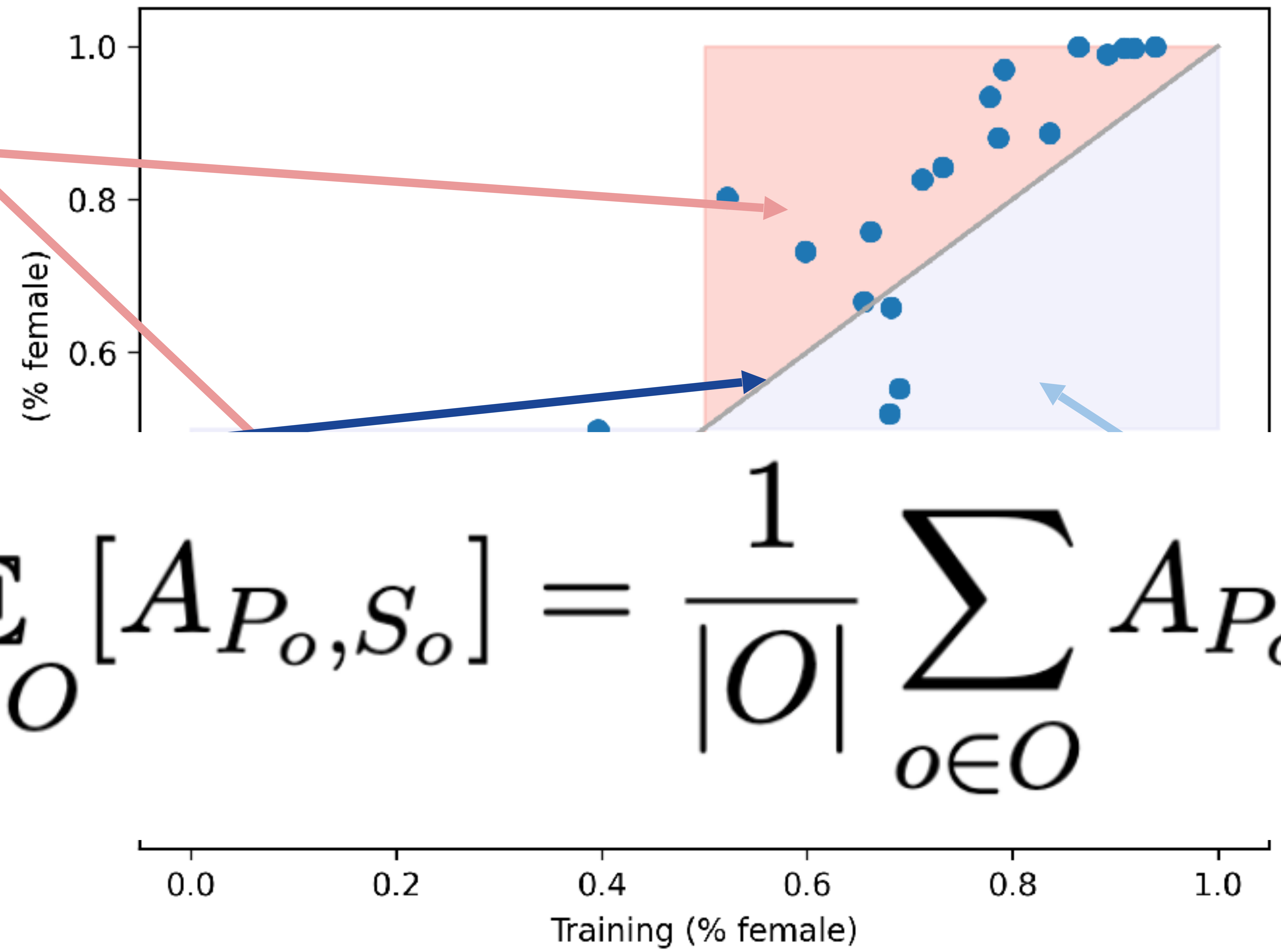
Bias Amplification?

Given the calculated ratios from the data, we can now compare the model's generation to the training data

Peach area:
Bias Amplification

Diagonal:
Bias preservatio

der area:
e-amplification



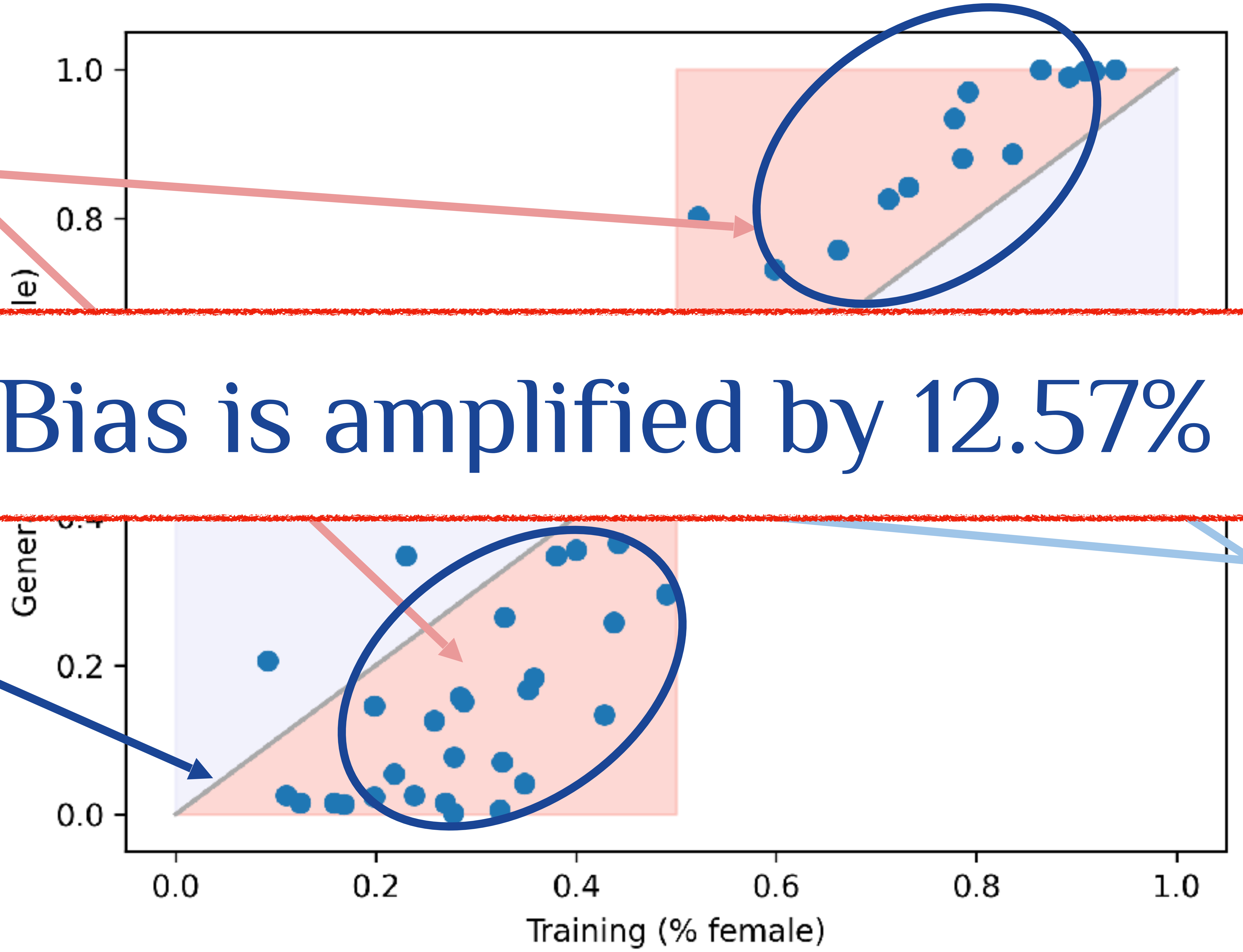
$$\mathop{\mathbb{E}}_{o\in O}[A_{P_o,S_o}] = \frac{1}{|O|}\sum_{o\in O} A_{P_o,S_o}$$

Bias Amplification!

Given the calculated ratios from the data, we can now compare the model's generation to the training data

Peach area:
Bias Amplification

Diagonal:
Bias preservatio



Bias is amplified by 12.57%

The Bias Amplification Paradox

But wait!

Why would a model amplify the biases from the training data?

Let's look at the training data again

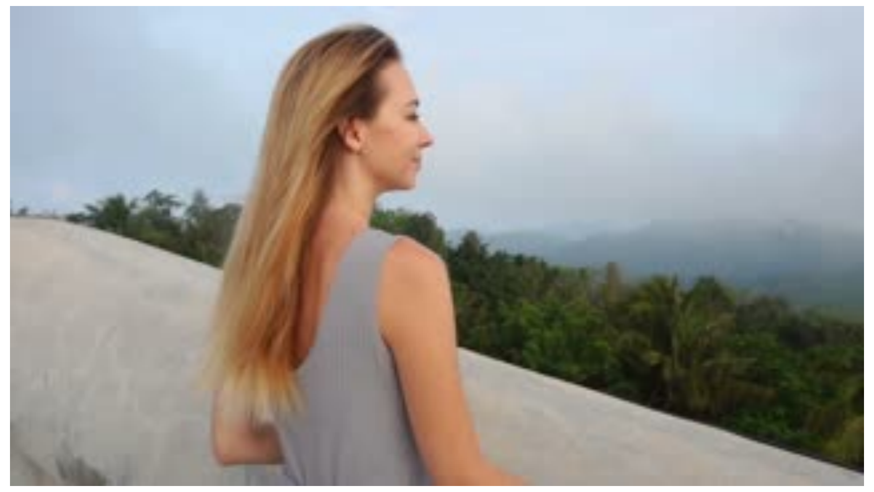
Training Data Investigation

👩 👨

Portrait of young ***woman*** **programmer** working at a computer in the data center filled with display screens

programmer configures the... | Shutterstock . vector #669546292

Slow motion **programmer** ***female*** relaxing among nature, young ***woman*** on long-awaited vacation abroad after working year...

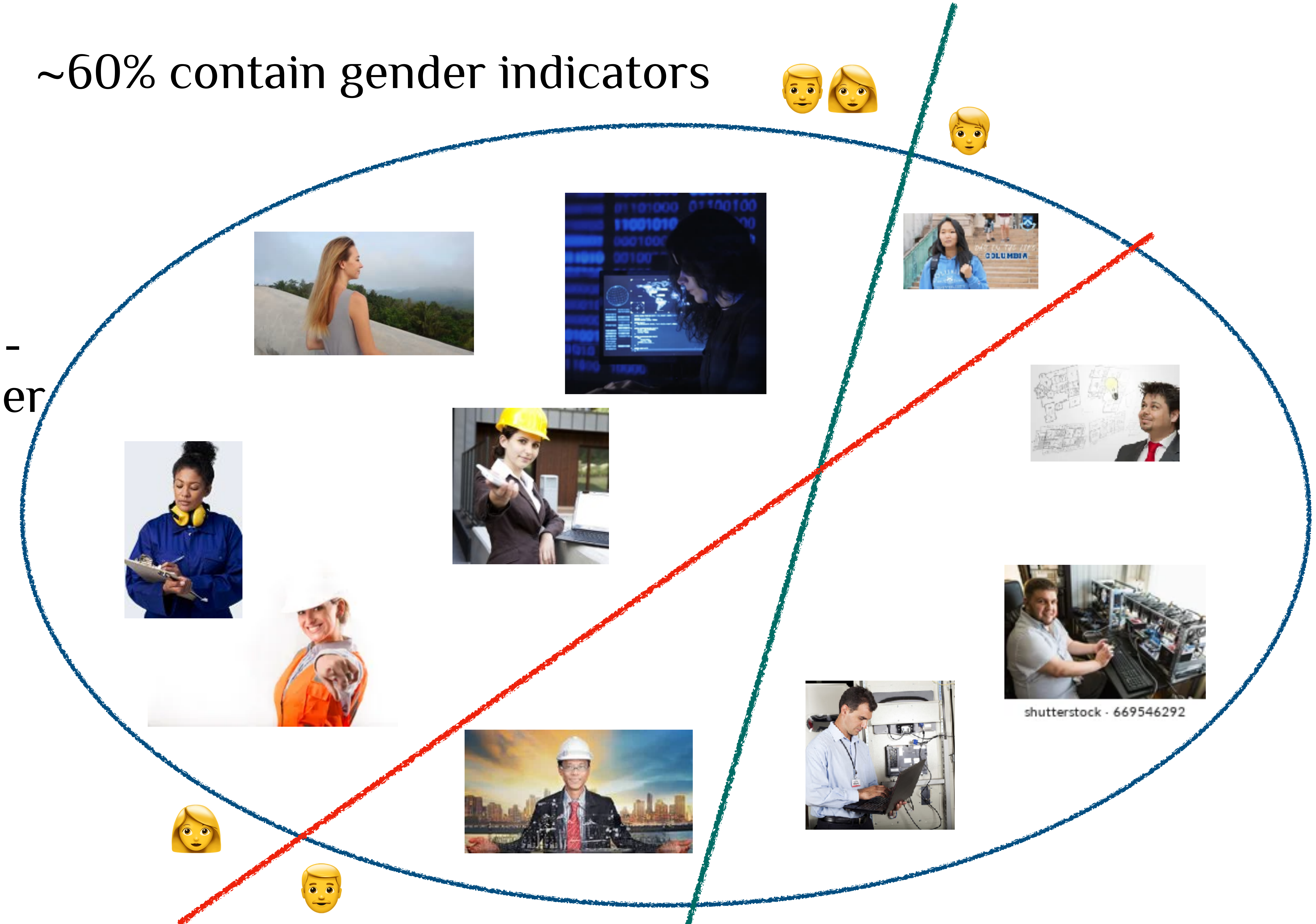


industrial programmer checking computerized machine status

Training Data Investigation

~60% contain gender indicators

Mostly with anti-stereotype gender (70%)



Training Data Investigation

~60% contain gender indicators

Test data

"A photo of a face of an engineer"

Mostly with anti-stereotype gender (70%)

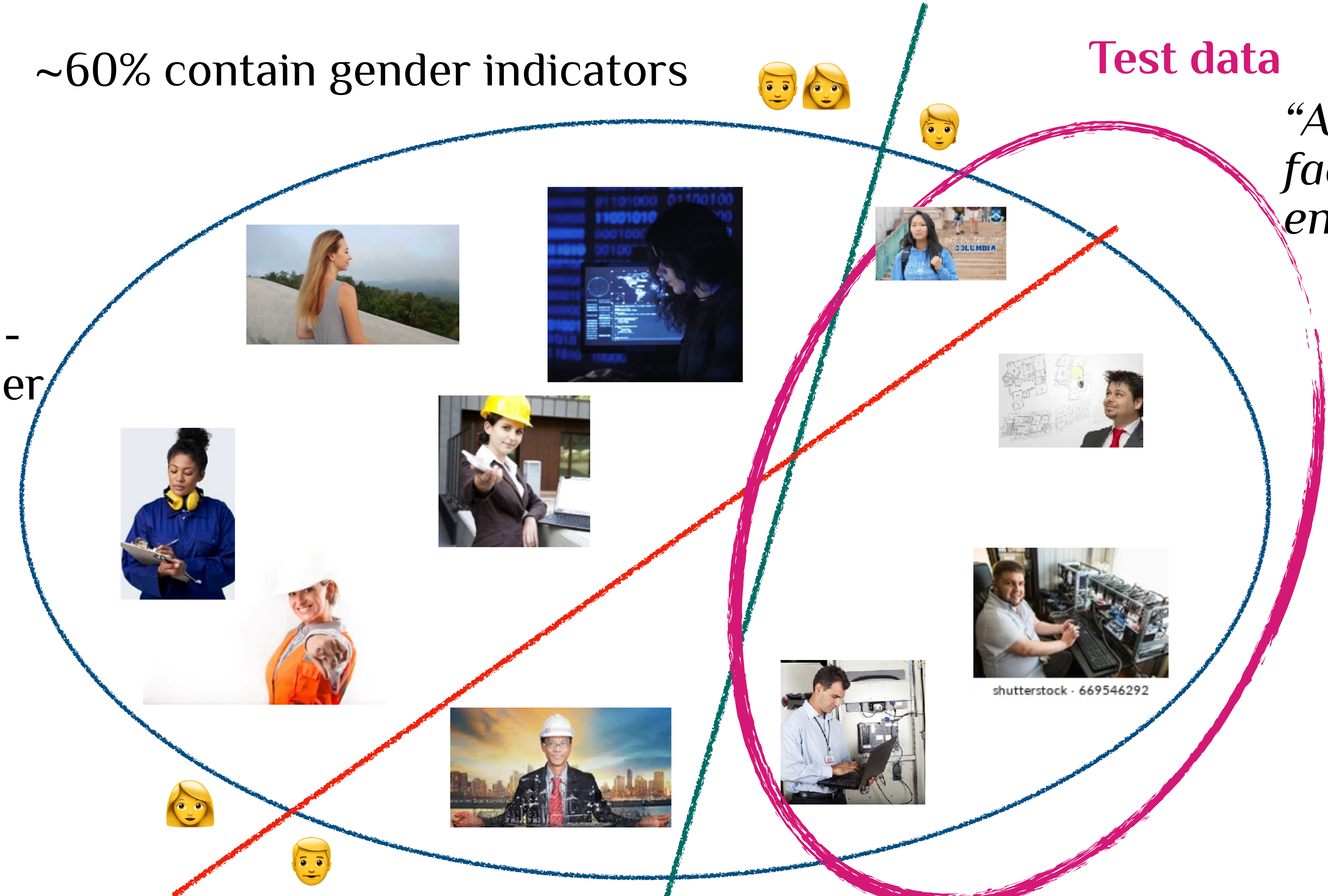
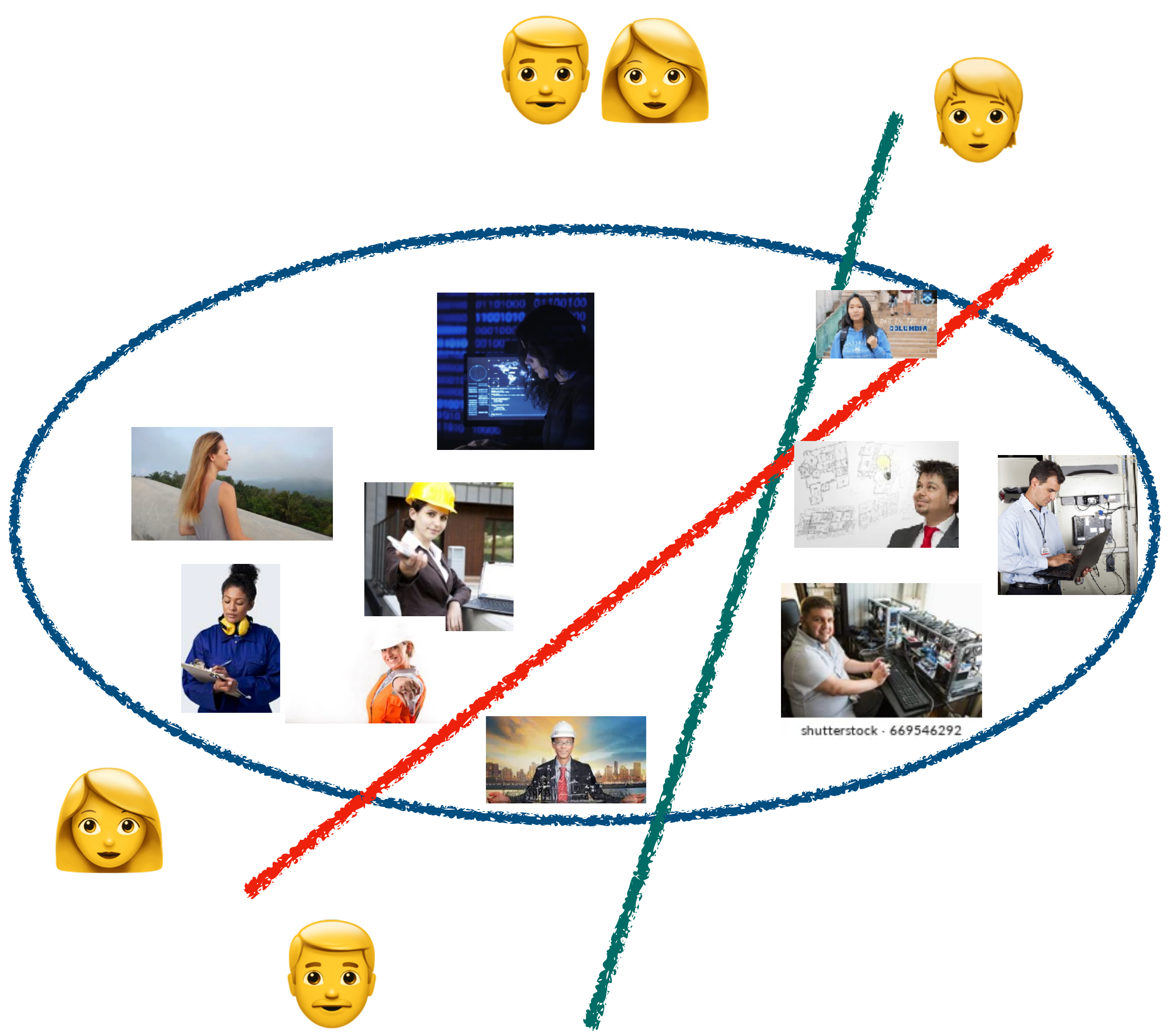
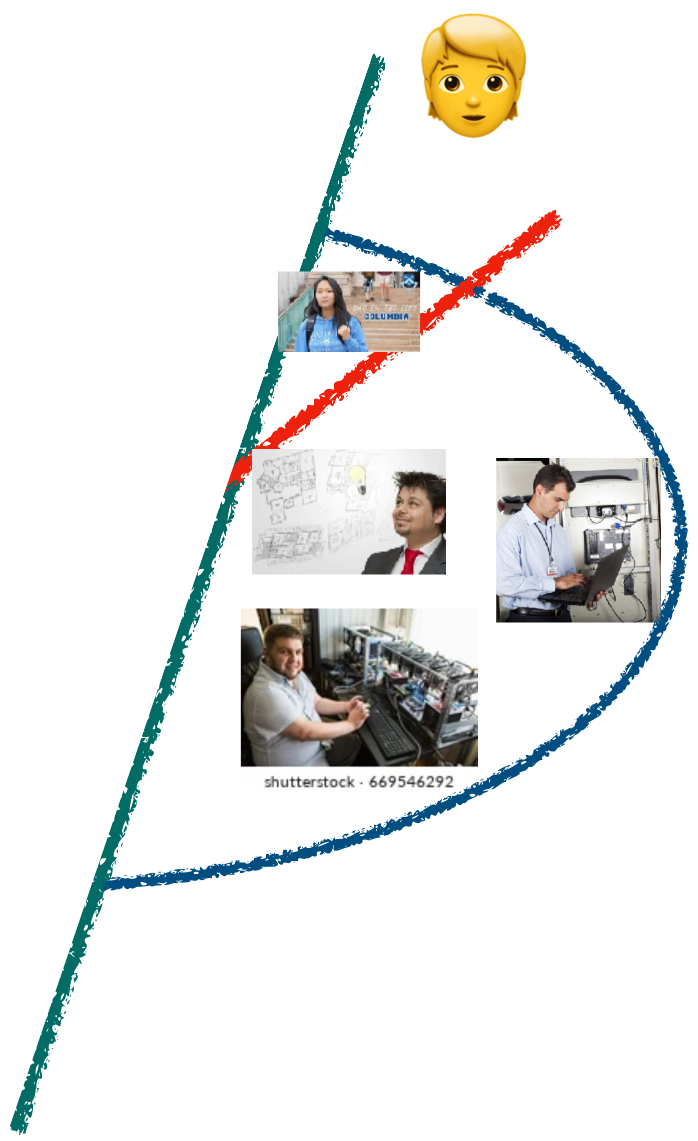


Image Captions & Prompts Mismatch

Training data



Test data

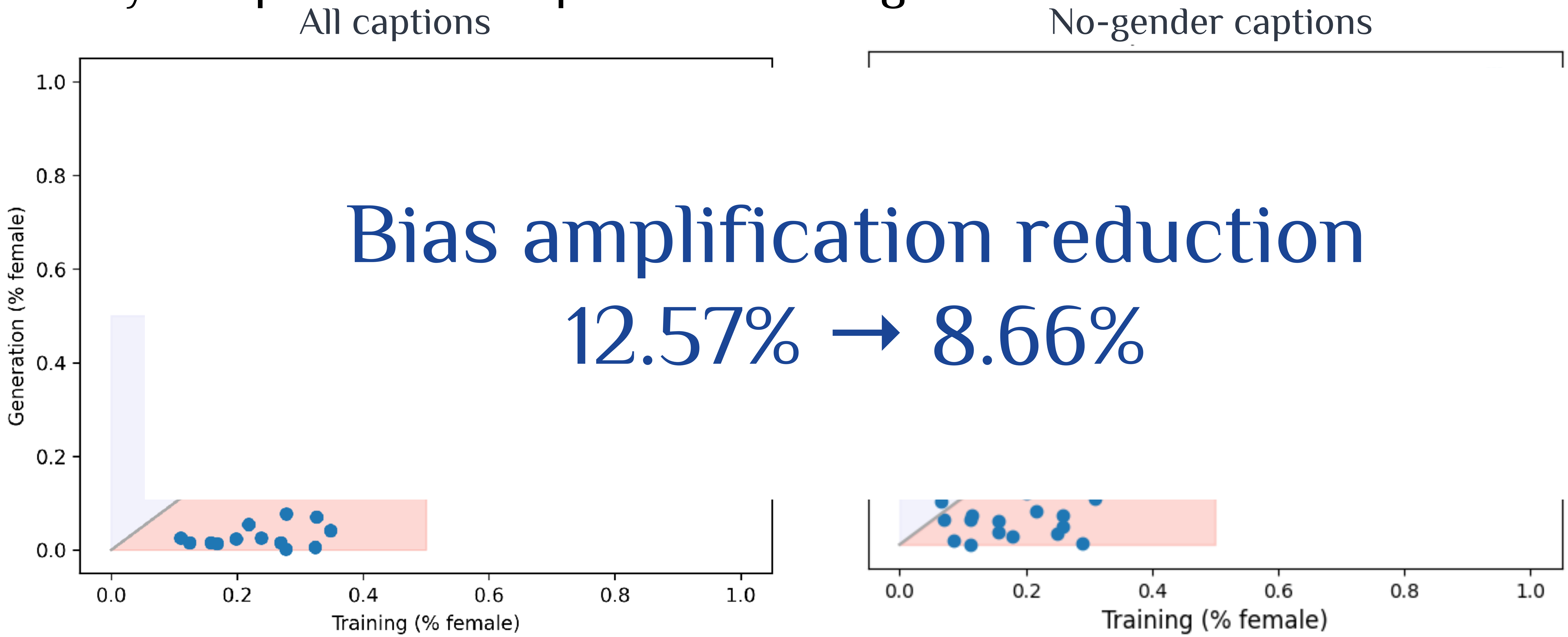


“A photo of a face of an engineer”

Matching Distributions

Instead of comparing the generated images to the entire training set:

- We only compare to the captions with no gender indicators



Bias amplification reduction

12.57% → 8.66%

One Mismatch

What about others?

Image Captions & Prompts Mismatch #2

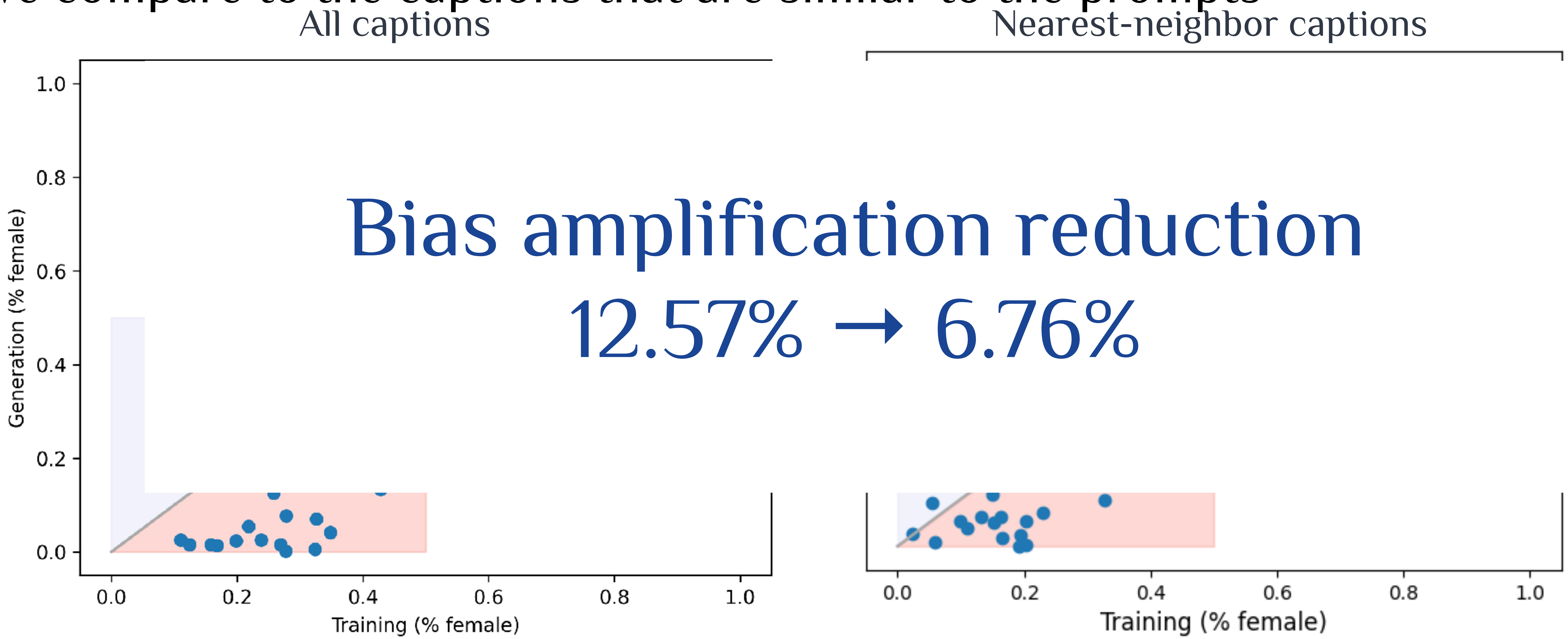
We also found a “d

(a) Training captions for **President**: 1) "Leana Wen, Planned Parenthood president..." 2) "New Schaumburg Business Association President..." 3) "BCCI president N Srinivasan..." 4) "Indiana Pacers president of basketball operations..."

Matching Distributions #2

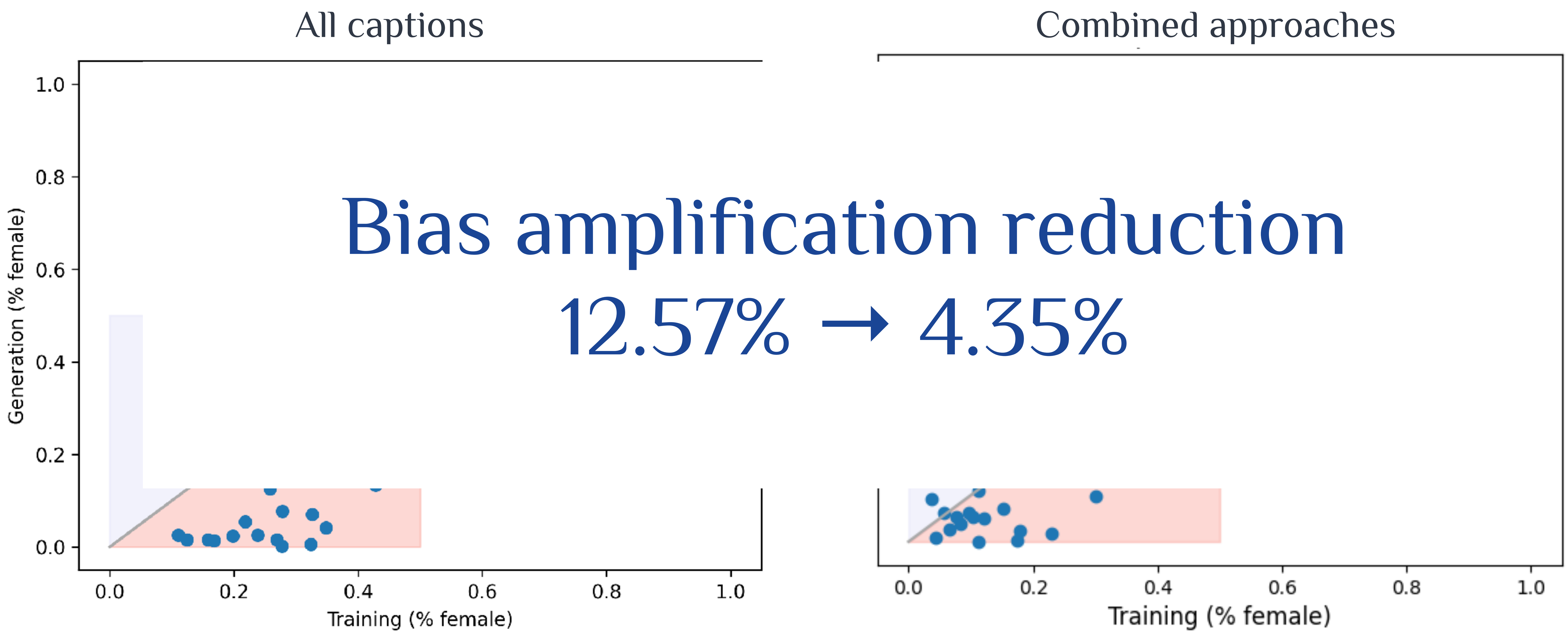
Instead of comparing the generated images to the entire training set:

- We compare to the captions that are similar to the prompts



Matching Distributions: Combined

Finally, we combine both approaches



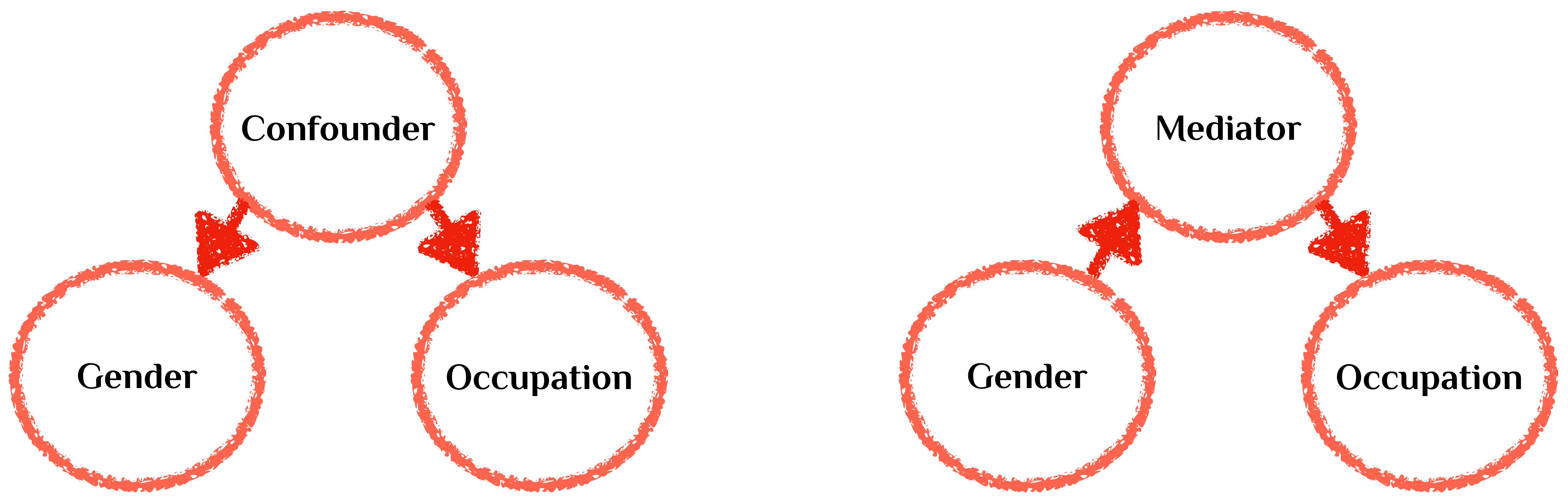
Bias amplification reduction

12.57% → 4.35%

Bias Amplification Revisited

While we still observe amplification of bias:

- It is significantly reduced
- There may be more confounders/mediators
- This problem is more nuanced and involved than originally thought



Summary

The Simpson's Paradox

- Unobserved confounders/mediators may reverse conclusions

The Bias Amplification Paradox

- Unmatched distribution make reverse conclusions

Evaluation is hard & Understanding the data is crucial!

What Did We Learn From the Paradoxes?

Setup

homer simpson

Training

Evaluation

LAION

Large-scale Artificial Intelligence Open Network

TRULY OPEN AI. 100% NON-PROFIT. 100% FREE.

LAION, as a non-profit organization, provides datasets, tools and models to liberate machine learning research. By doing so, we encourage open public education and a more environment-friendly use of resources by reusing existing datasets and models.

Investigation

Thank you

yanaiela.github.io

@yanaiela