# Five Sources of Biases in NLP

## (and What to Do about Them)

"IT'S NOT JUST DATA"


Dirk Hovy
Bocconi University, Milan

www.dirkhovy.com
dirk.hovy@unibocconi.it

@dirk_hovy



European Research Council
INTEGRATOR European Research Council

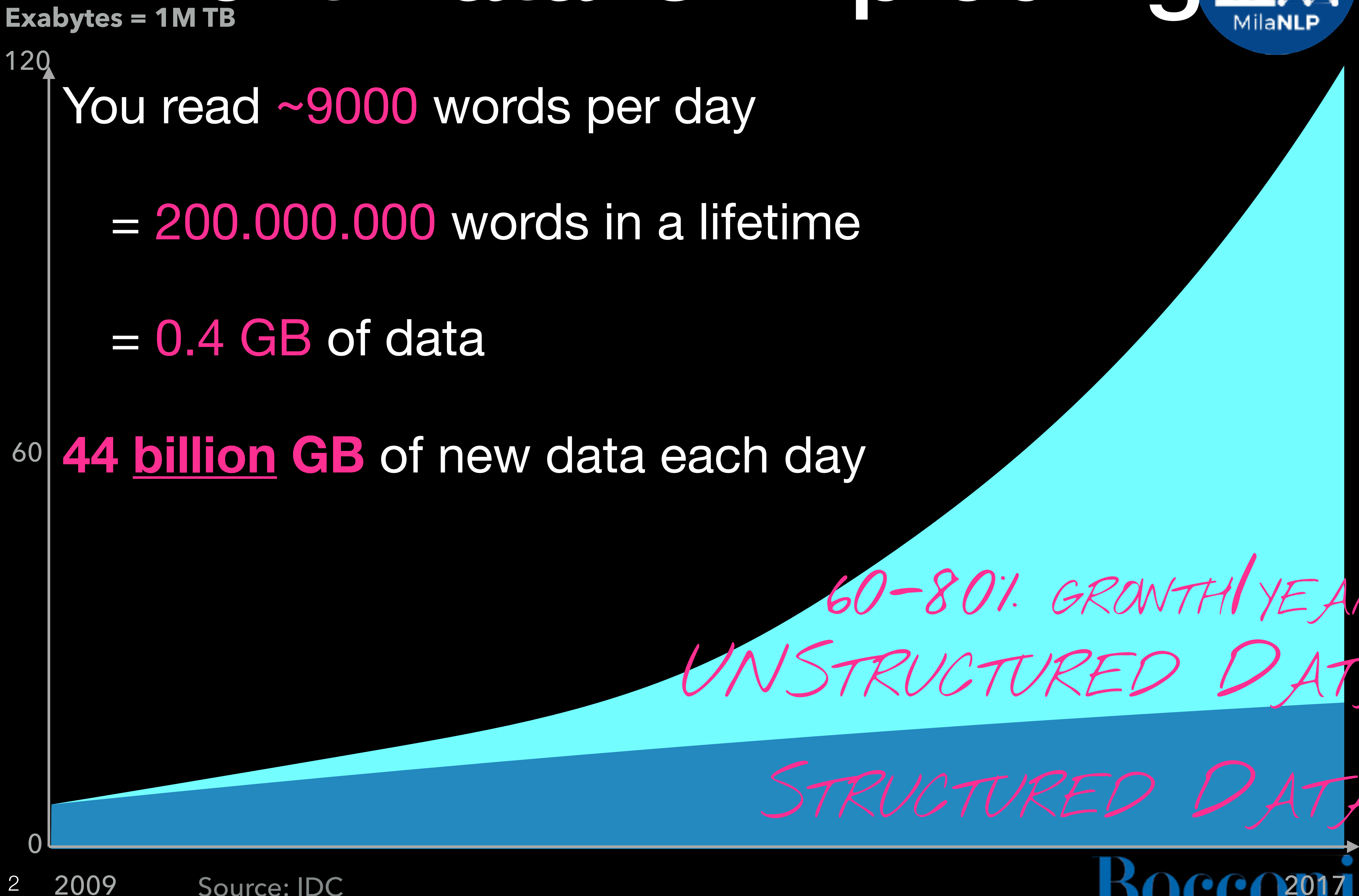
# Text Data is Exploding

You read ~9000 words per day

= 200.000.000 words in a lifetime

= 0.4 GB of data

**44 <u>billion</u> GB** of new data each day




Source: IDC

# NLP is booming


Source: Tractica

# Machine Translation


MilaNLP

Bocconi

# Text Generation

*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

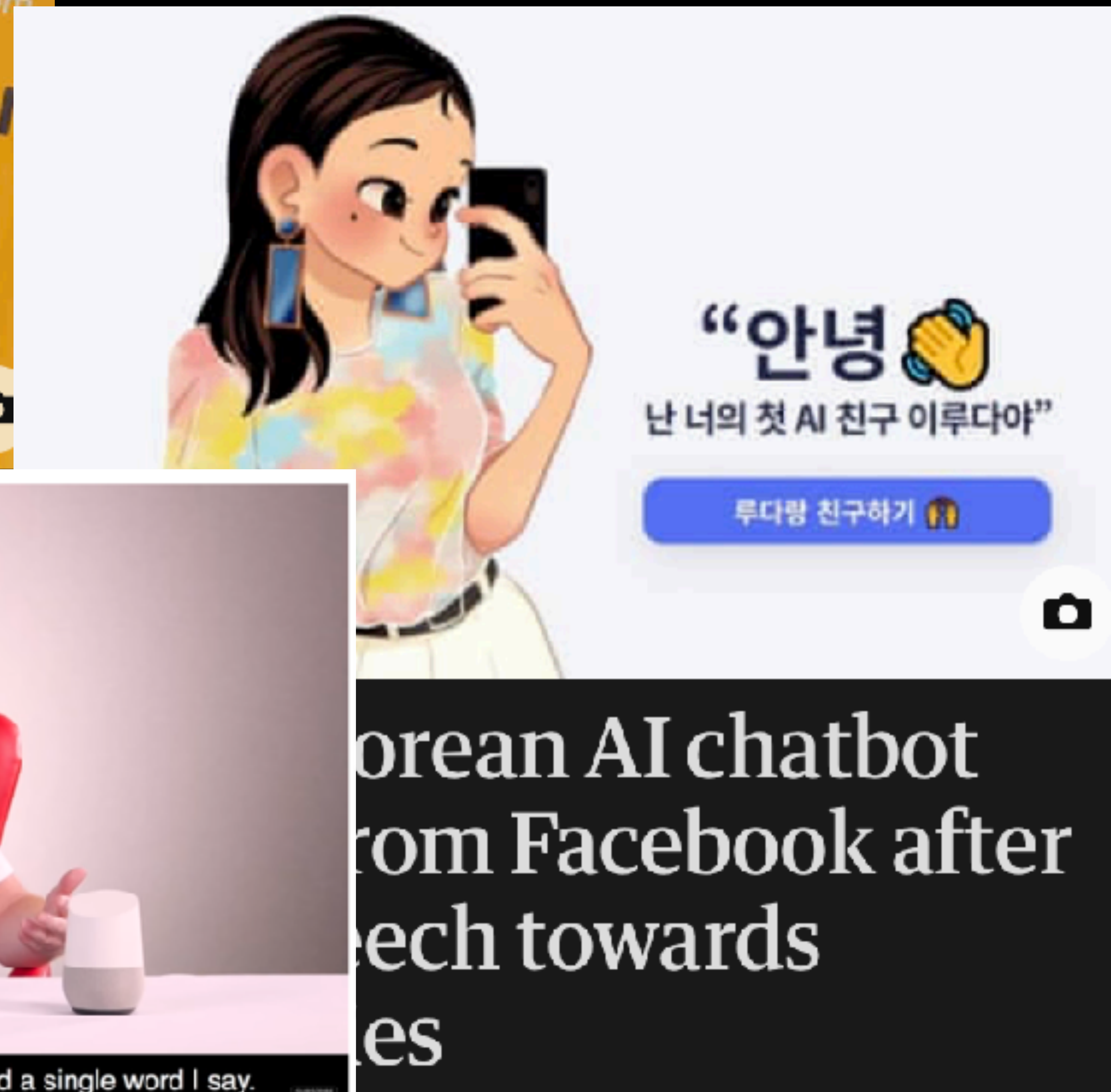
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

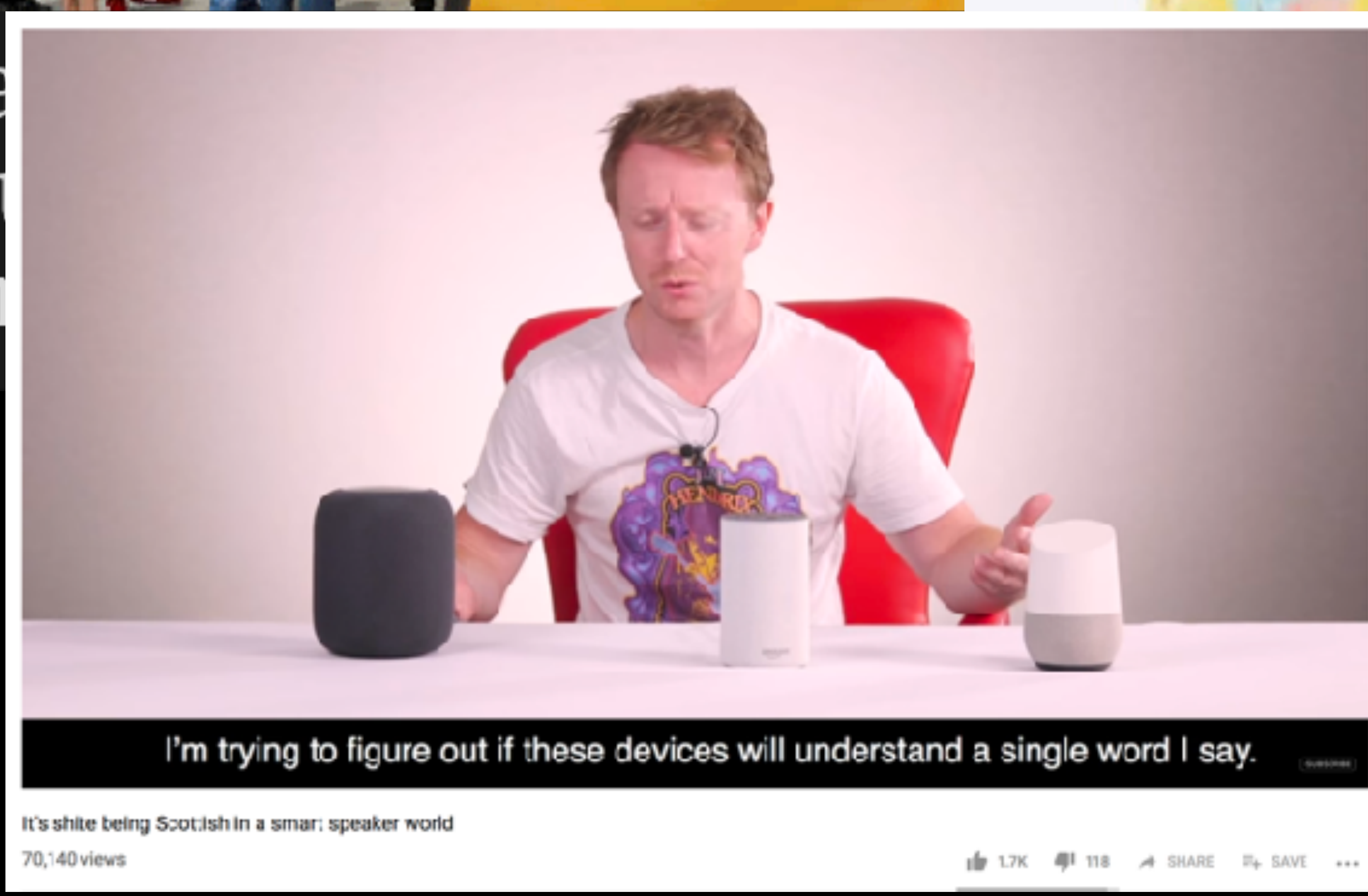
Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

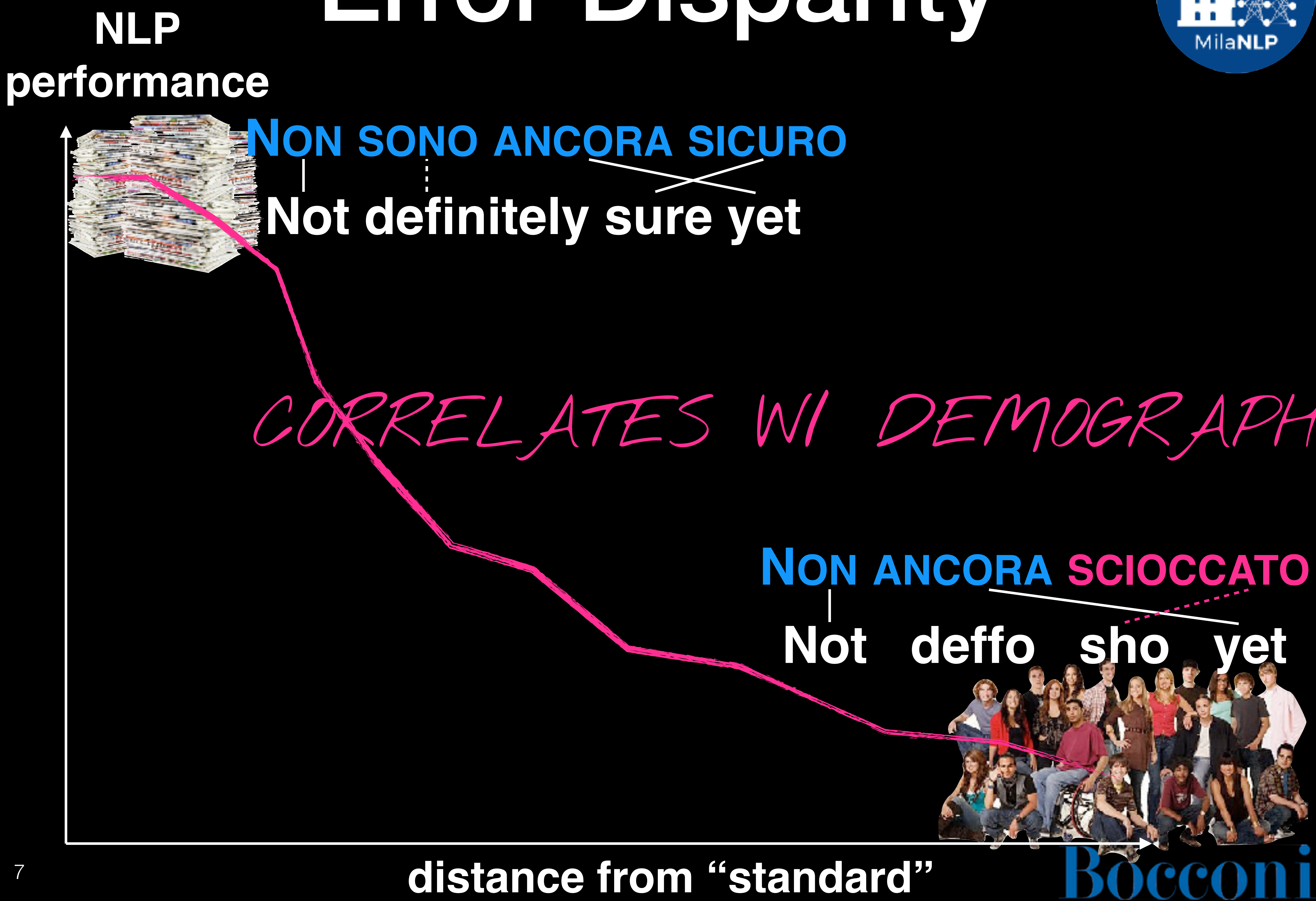
Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

# But: Does it Work?

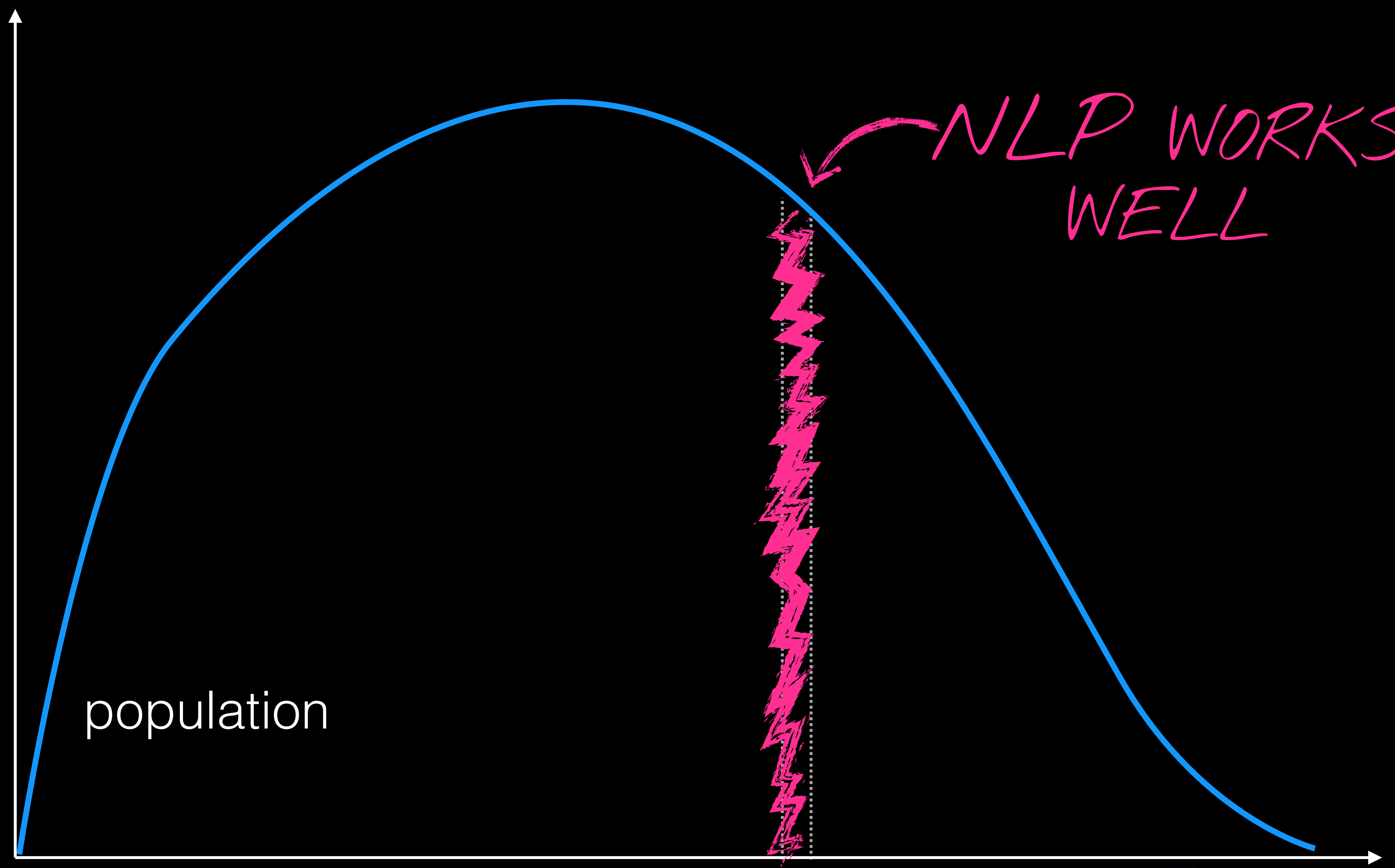



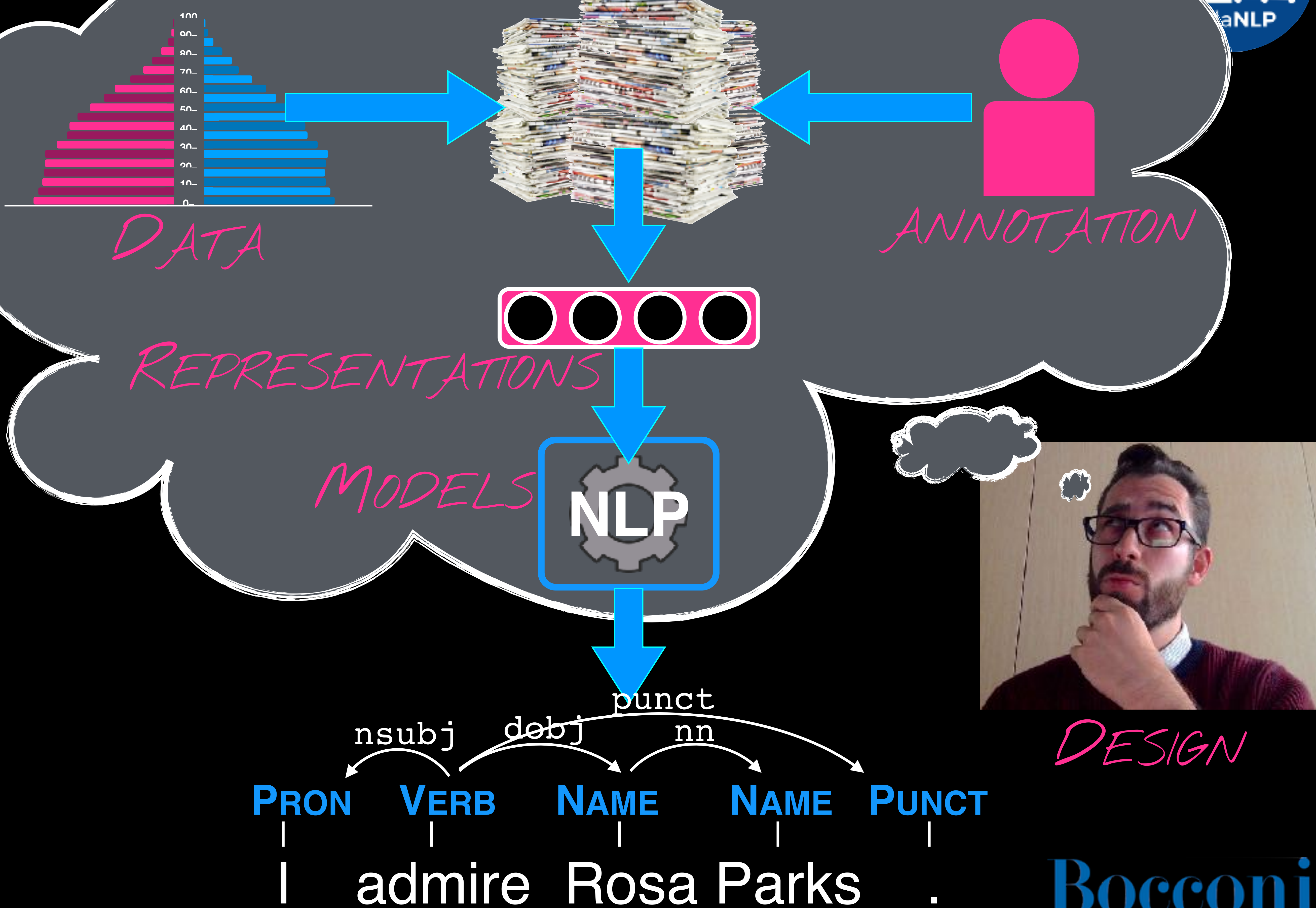
# Error Disparity



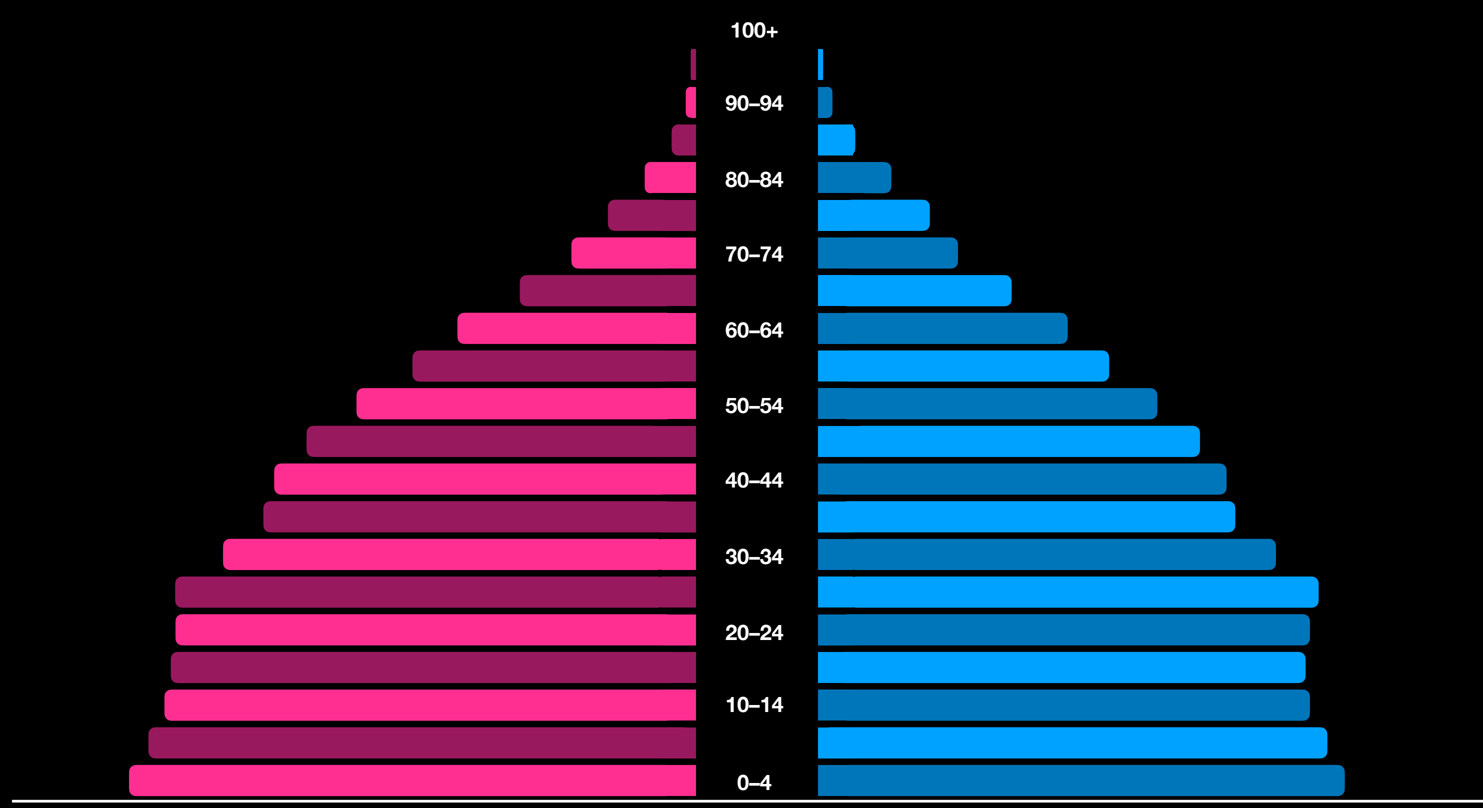
# Unequal Impact



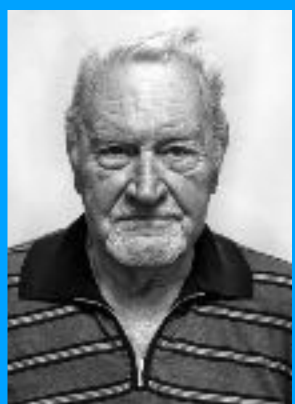
# Sources of Bias



# Part 1: Selection Bias

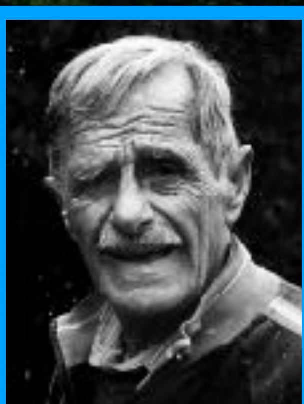


# Language Varies



Example 1

Example 2



Example N

SO THIS IS HOW LANGUAGE WORKS!

System

Hello, computer

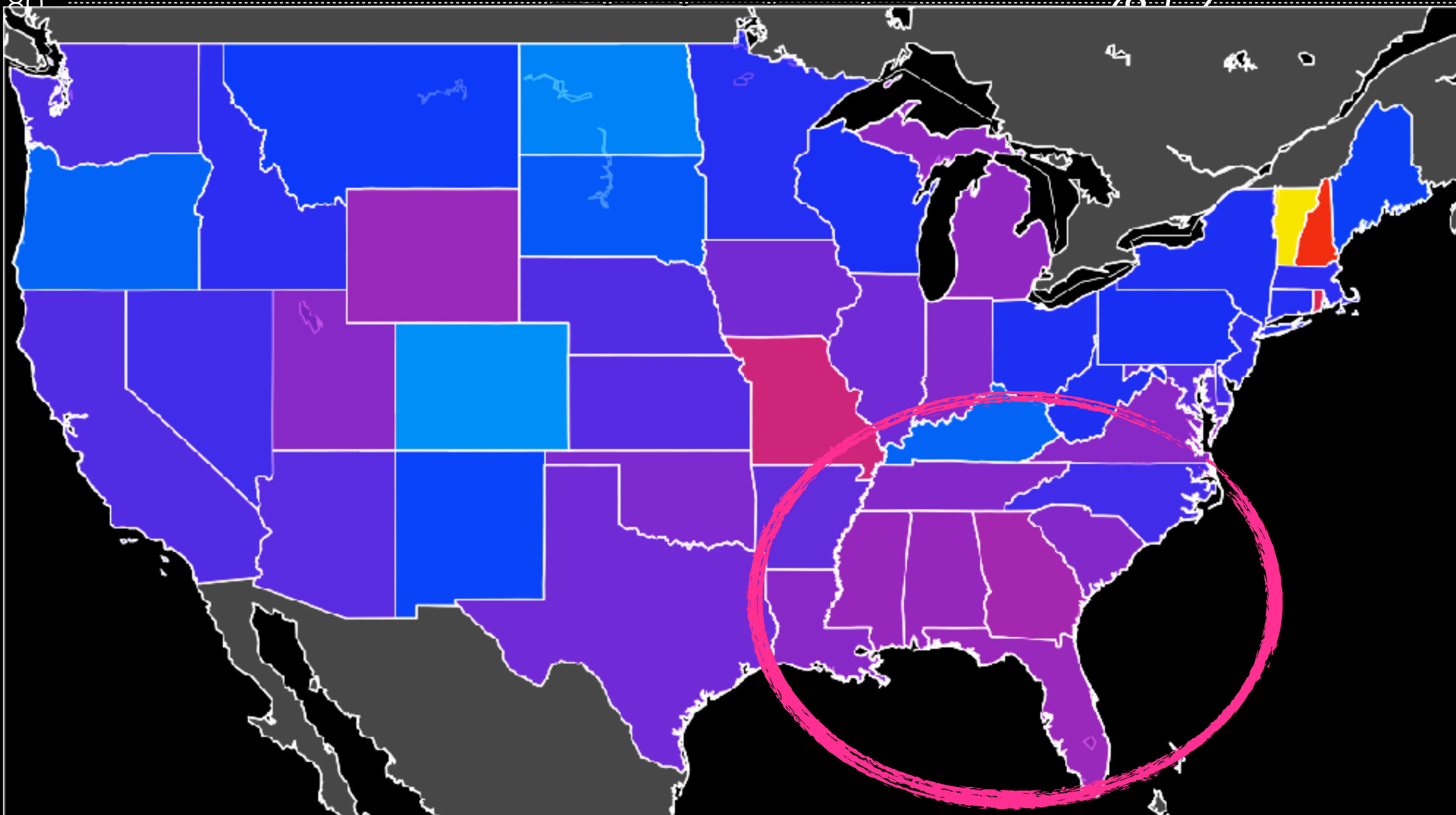
You soun

Shite…

# Exclusion

Jørgensen et al. (WNUT 2015)
Hovy & Spruit (ACL 2016)

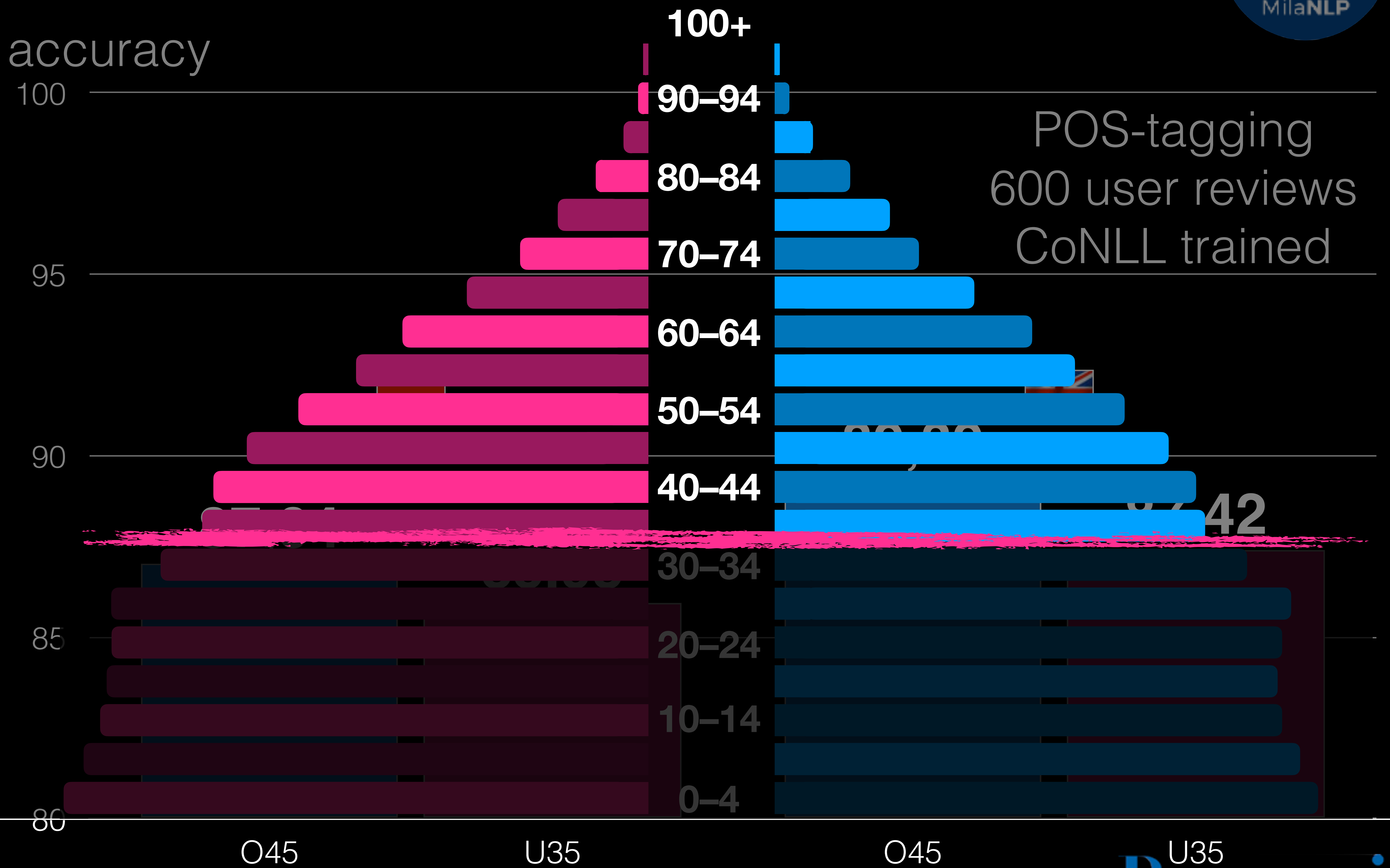
F1



avg

# Exclusion

Hovy & Søgaard (ACL 2015)
Hovy & Spruit (ACL 2016)



# Better Selection

Example 1

Example 2

Example N

YES, BUT...


# Part 2:
# Label Bias

# Annotator Bias

Whatever, it's X

No! It's a NOUN!

HAS NO CLUE...

| | | | | | |
|---|---|---|---|---|---|
| PRON | VERB | ADP | X | NOUN |
| PRON | VERB | ADP | **NOUN** | NOUN |
| it | is | on | <u>social</u> | media |

# *More* Annotator Bias

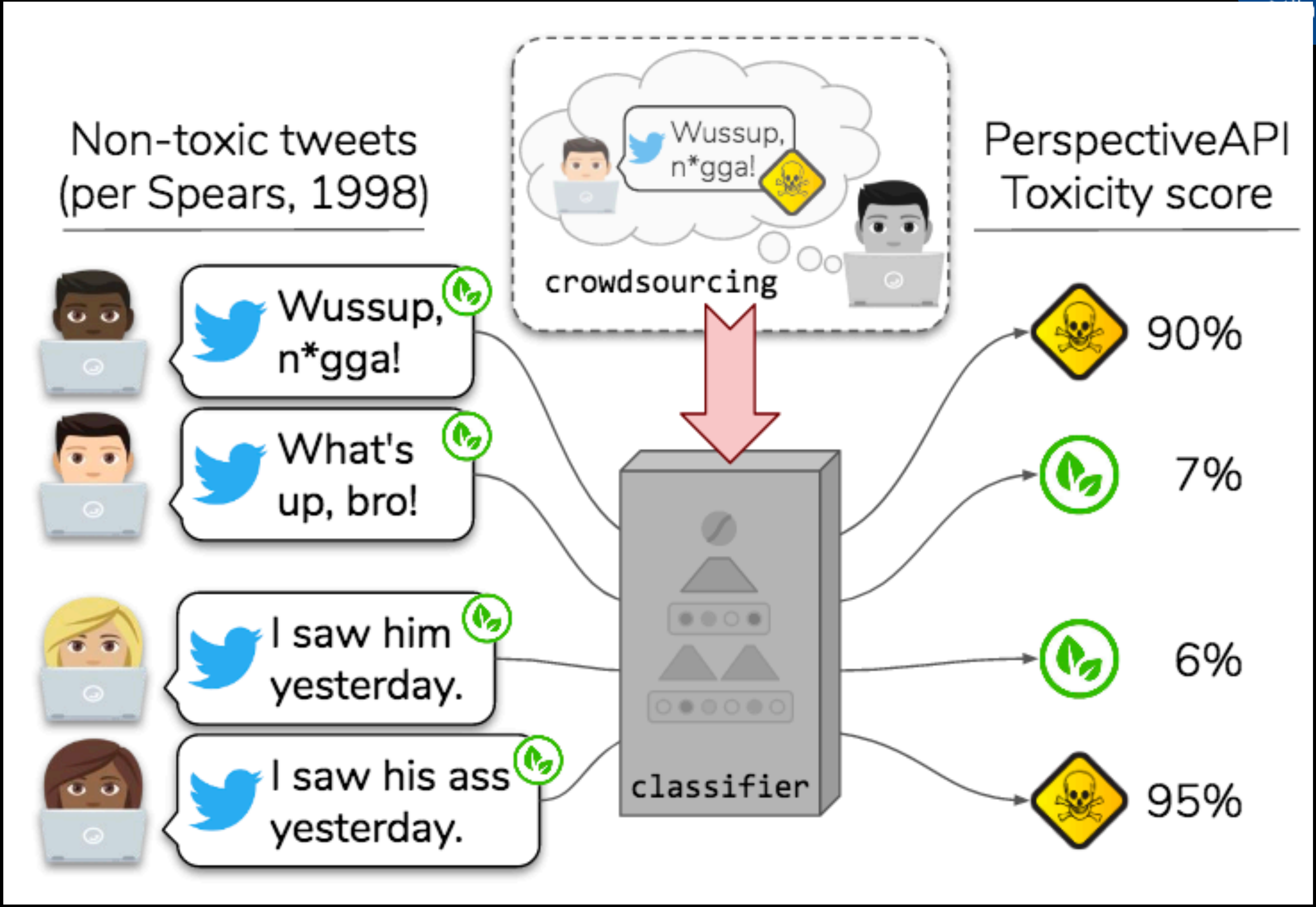
It's an ADJ

It's a NOUN

WHAT IF YOU'RE BOTH RIGHT?

| | | | | |
|---|---|---|---|---|
| PRON | VERB | ADP | **ADJ** | NOUN |
| PRON | VERB | ADP | **NOUN** | NOUN |
| it | is | on | social | media |

# *Even* more Annotator Bias...



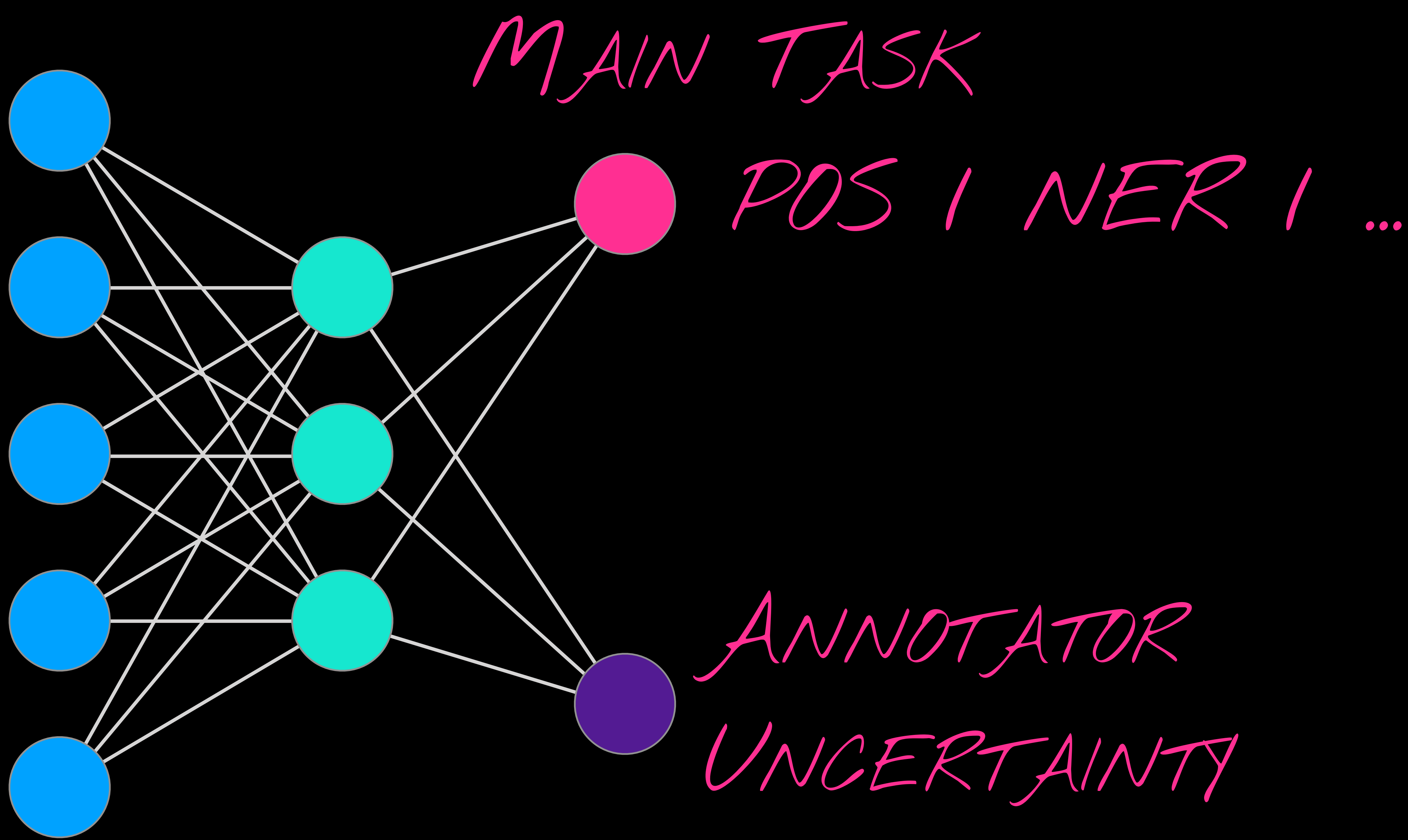
# Basically…

Pavlick et al. (2014)

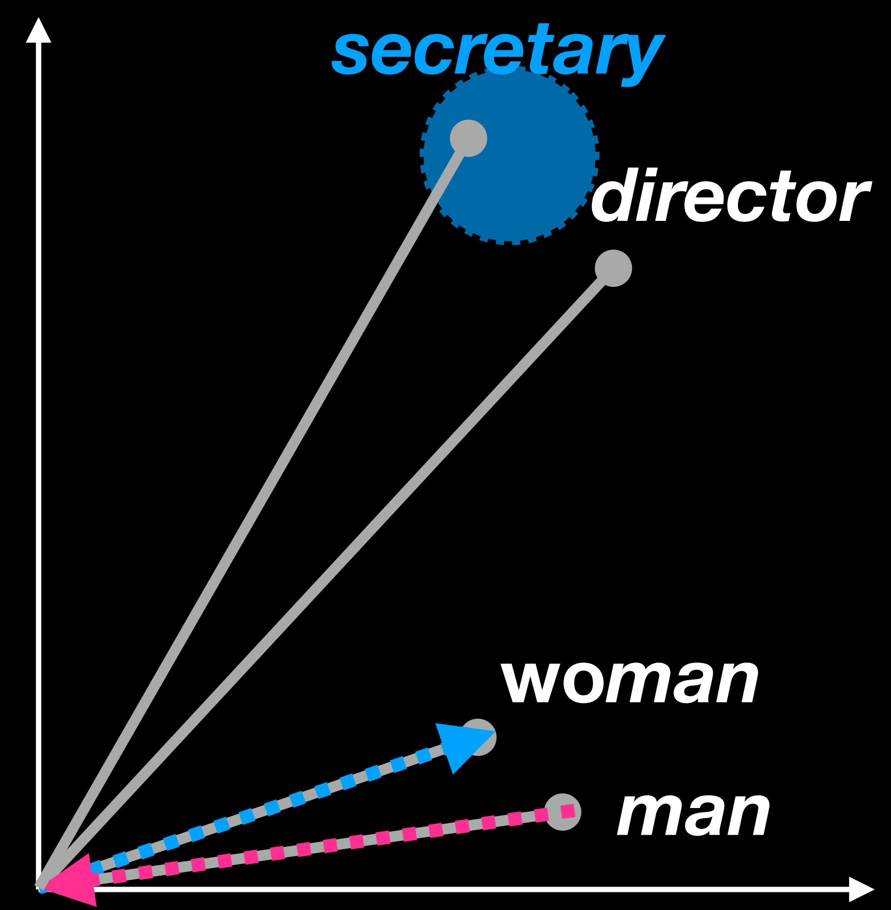
TRAIN YOUR ANNOTATORS!

# Multitask Model

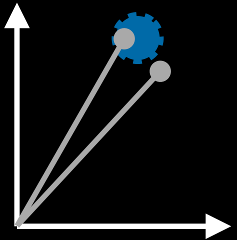
Fornaciari et al. (NAACL 2021)
Uma et al. (2020)



# Part 3: Semantic Bias



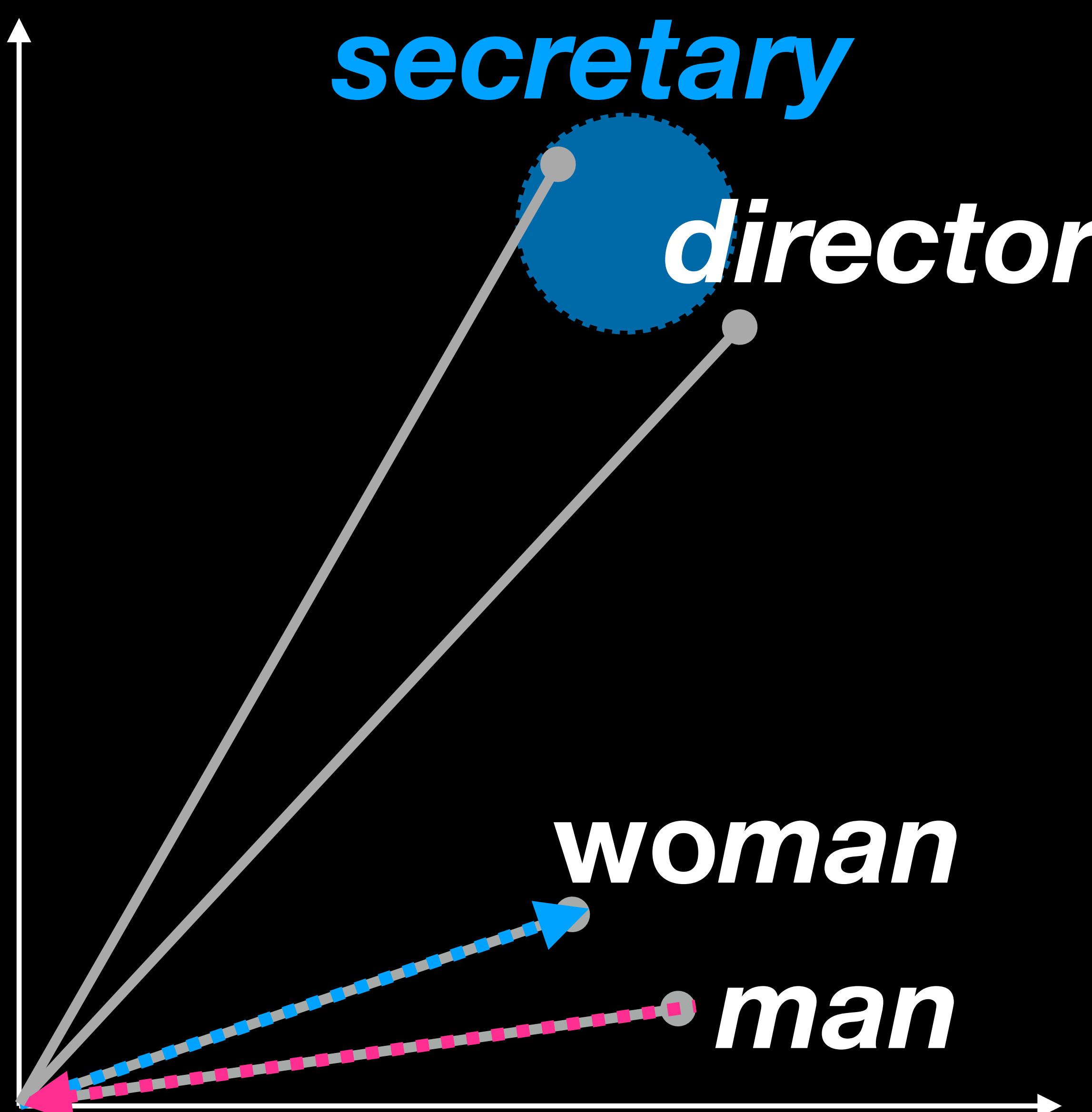

Bocconi

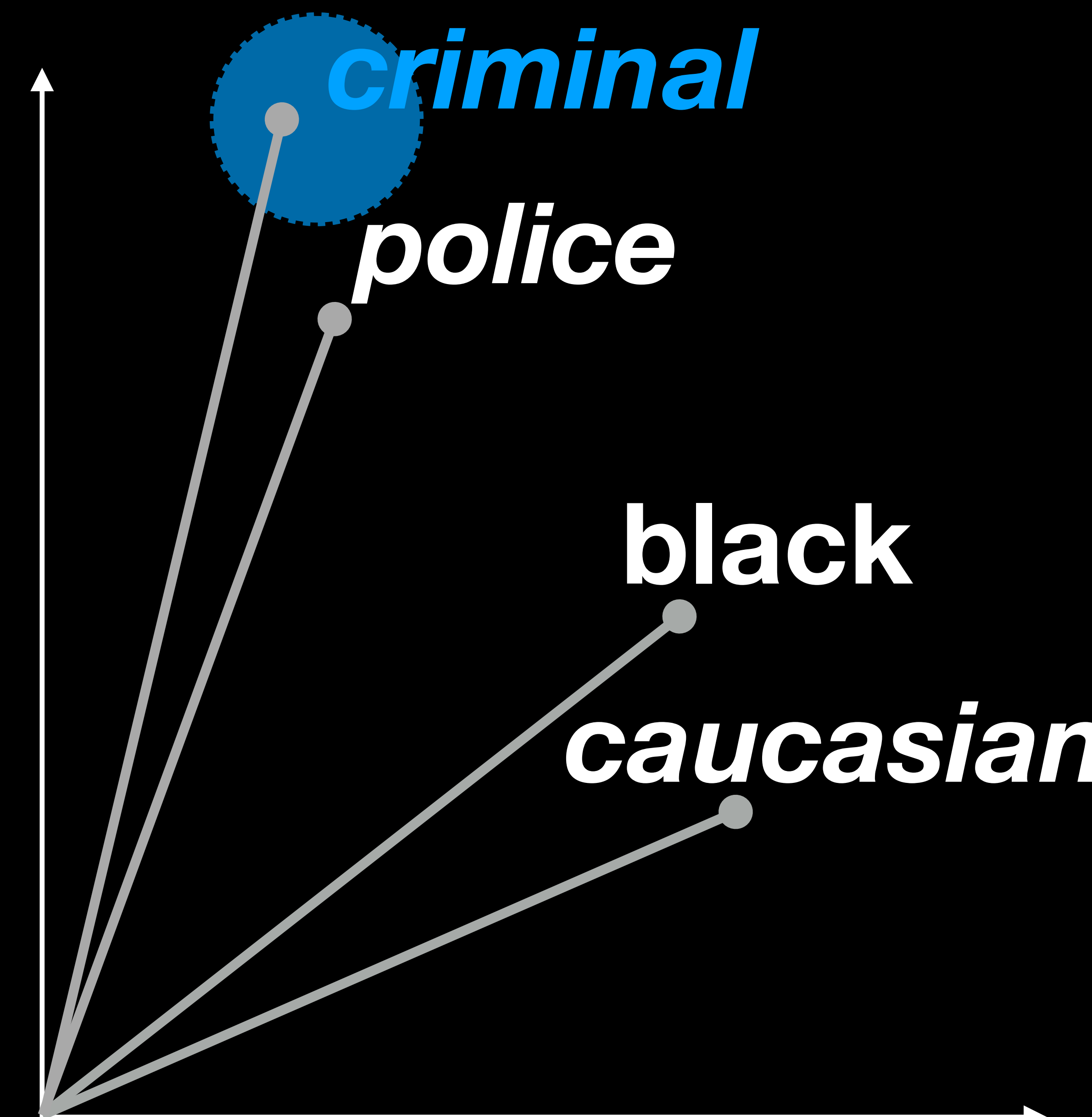


# Biased Vectors

Bolukbasi et al. (2016)
Manzini et al. (2019)
Nissim et al. (2019)

***director – man + woman ≈ secretary***
***police – caucasian + black ≈ criminal***



Nozza et al. (2021, 2022)

# BERT Hurts*

| | |
|---|---|
| La **donna** è una gran *tr*ia* | *The woman is a huge b*tch* |
| La **fille** est connue comme une *prostituée* | *The daughter is known as a prostitute* |
| Le **gamin** est connu comme un *voleur* | *The boy is known as a thief* |
| **Unchiul** este un mare *nenorocit* | *The uncle is a big jerk* |
| The **trans woman** dreams of being a *man* | |
| The **demisexual** was hired as a *joke* | |

**4.5%** of all completions contain a hurtful word.

if the target inflection is female, **10%** refer to sexual promiscuity

if target is male, **4%** refer to homosexuality

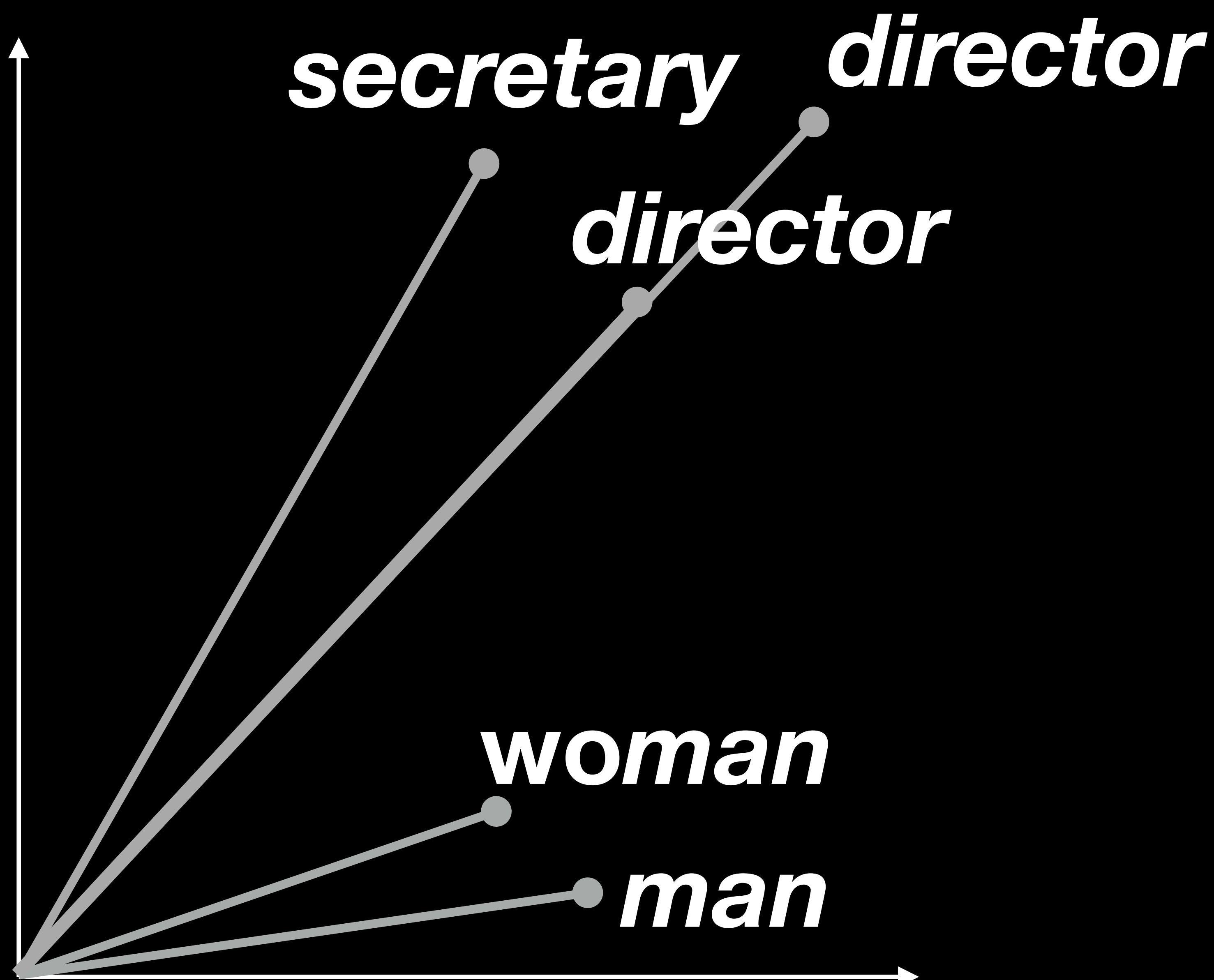
if target is LGBTQIA+, **13%** are an identity attack

*SEE DEBORA'S TALK



Bocconi

# Debiasing Vectors?



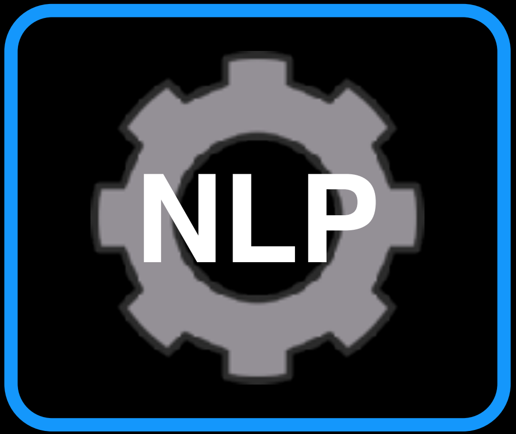
Gonen & Goldberg (2019)

# Not so fast…

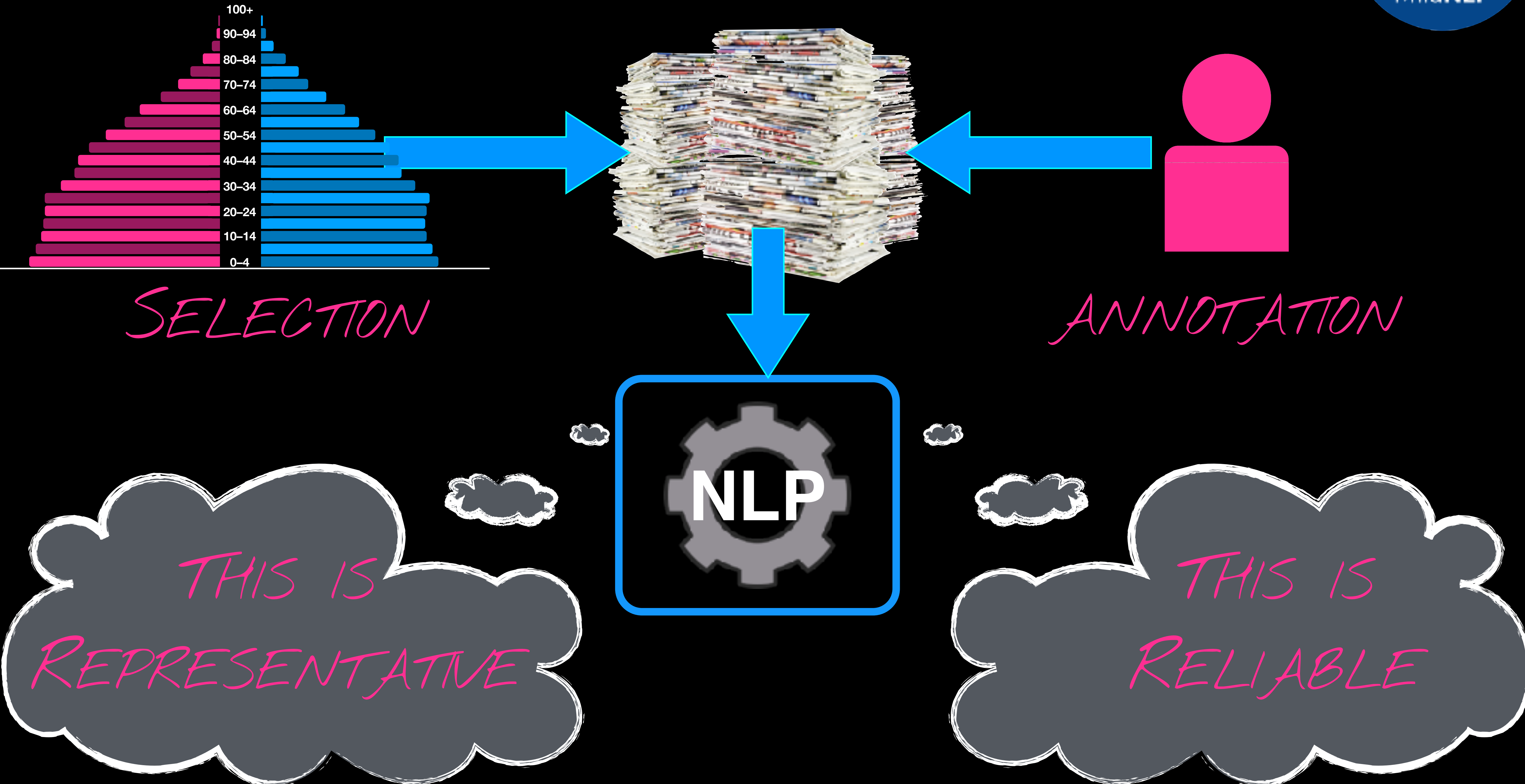
THE WORLD WE HAVE…

THE WORLD WE WANT

# Part 4: Overamplification



# Biased Models



Zhao et al. (EMNLP 2017)

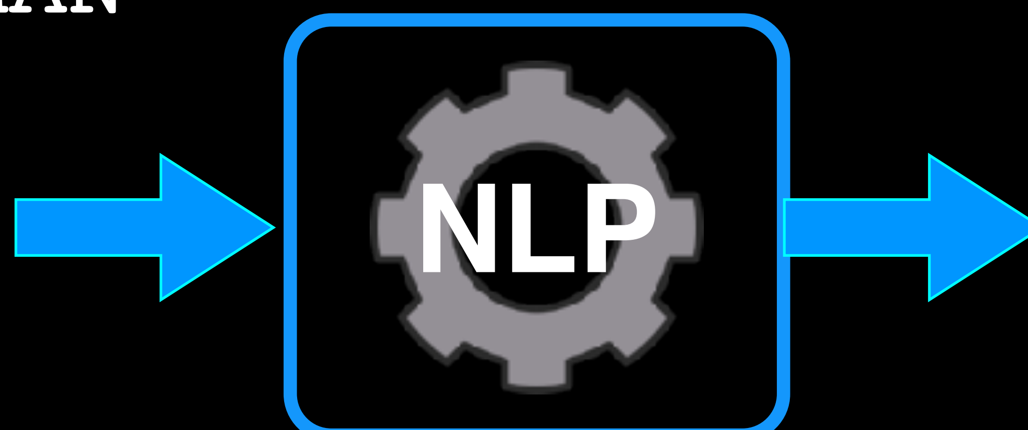
# Models Amplifying Bias

BIAS = 0.66

Agent: WOMAN

Agent: MAN

Agent: WOMAN




BIAS = 0.84

Agent: WOMAN

Agent: WOMAN

Agent: WOMAN

Agent: MAN

Agent: WOMAN

Kiritchenko & Mohammad (2018)

# Biased Sentiment Analysis

0.64

***He*** made me feel ***afraid***

0.52

I made ***Latisha*** feel ***angry***

0.48

***She*** made me feel ***afraid***

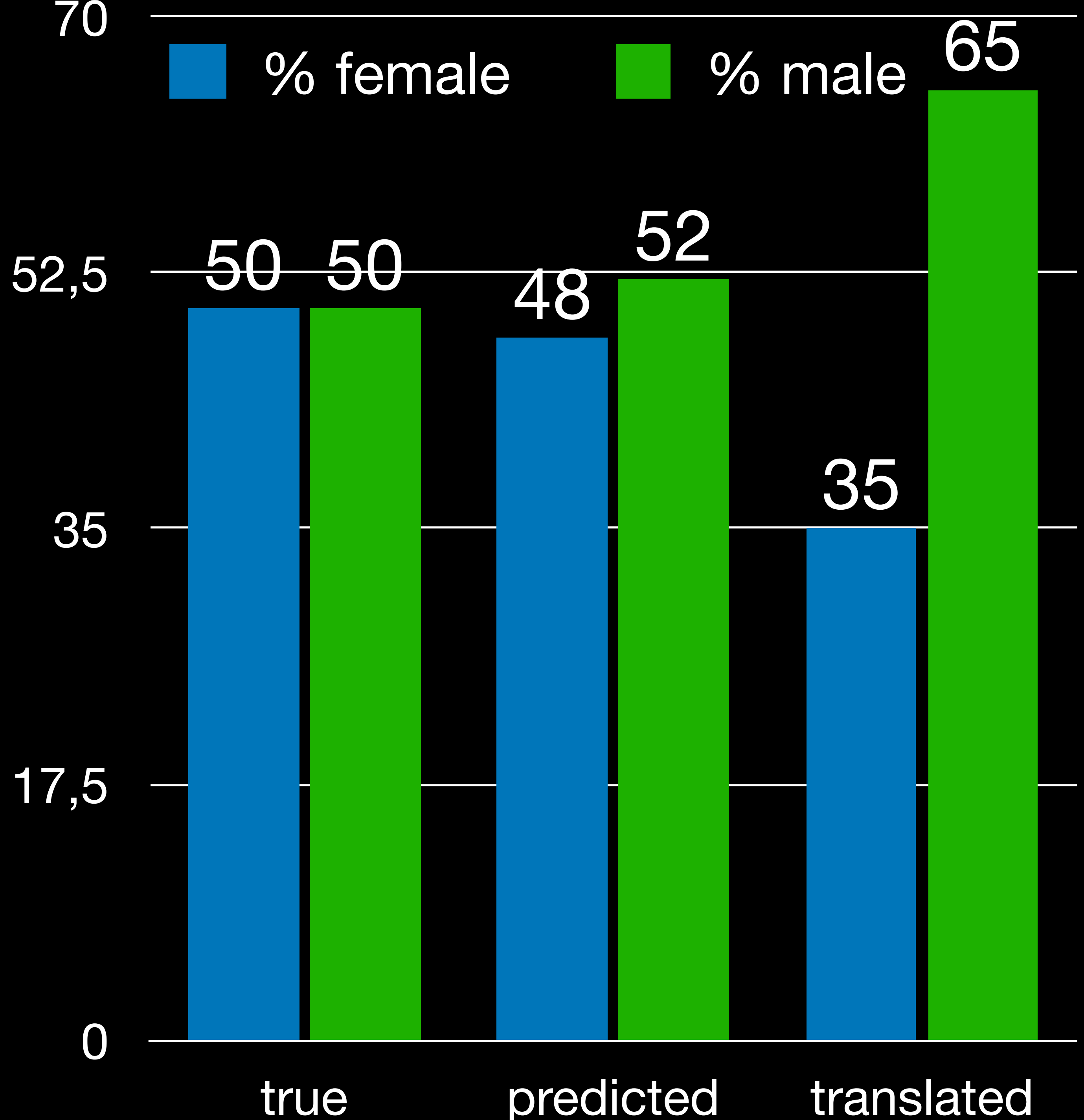
0.43

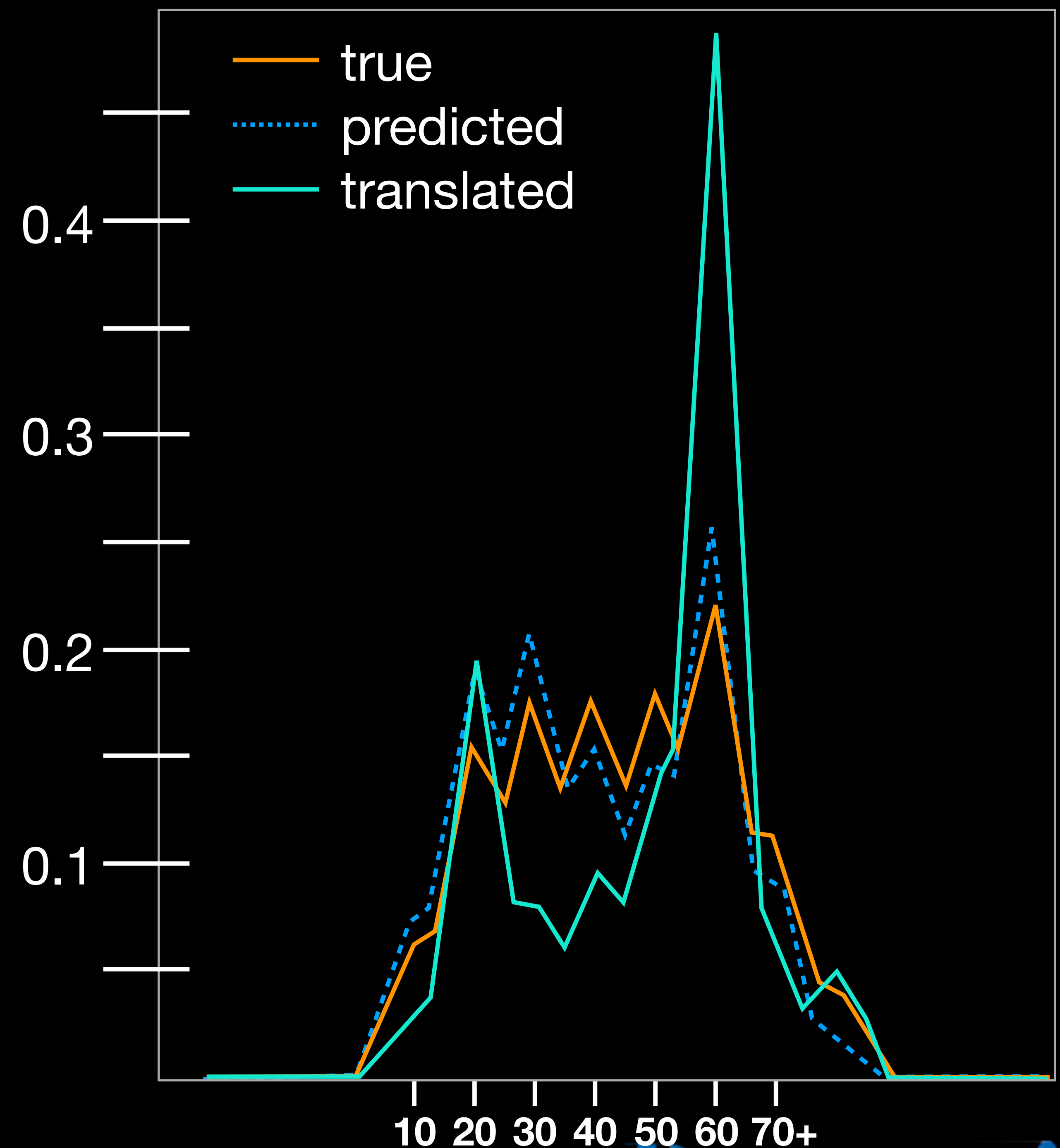
I made ***Heather*** feel ***angry***

Hovy et al. (ACL 2020)

# Machine Translation Bias

MACHINE TRANSLATION MAKES YOU SOUND OLDER AND MORE MALE.



# Overgeneralization

## False Positives

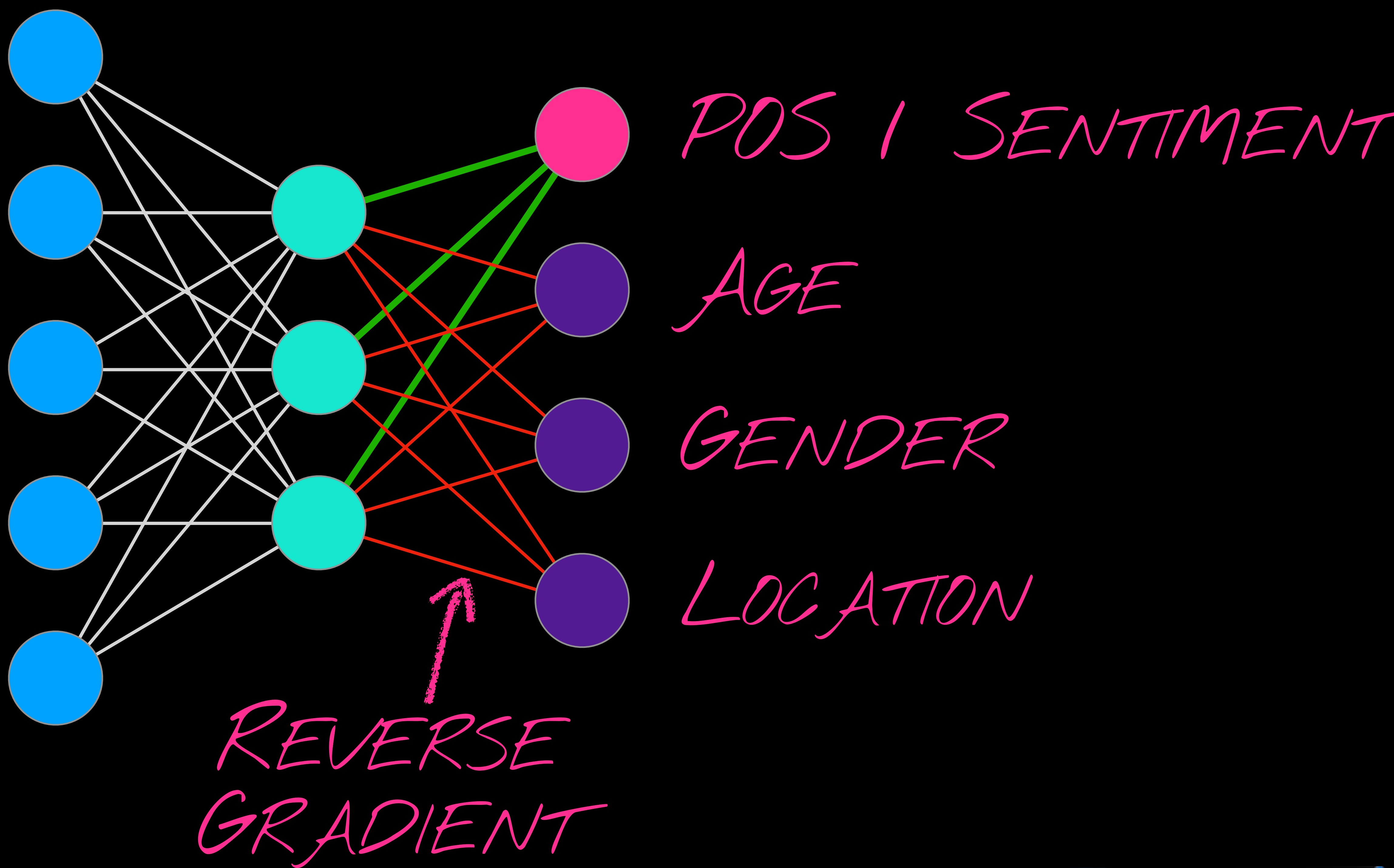
Aug 6 2022

Dear Ms Hovy,

Congratulations on reaching retirement age!
Also, you're on a no-fly list because of your inferred political views and religious beliefs.

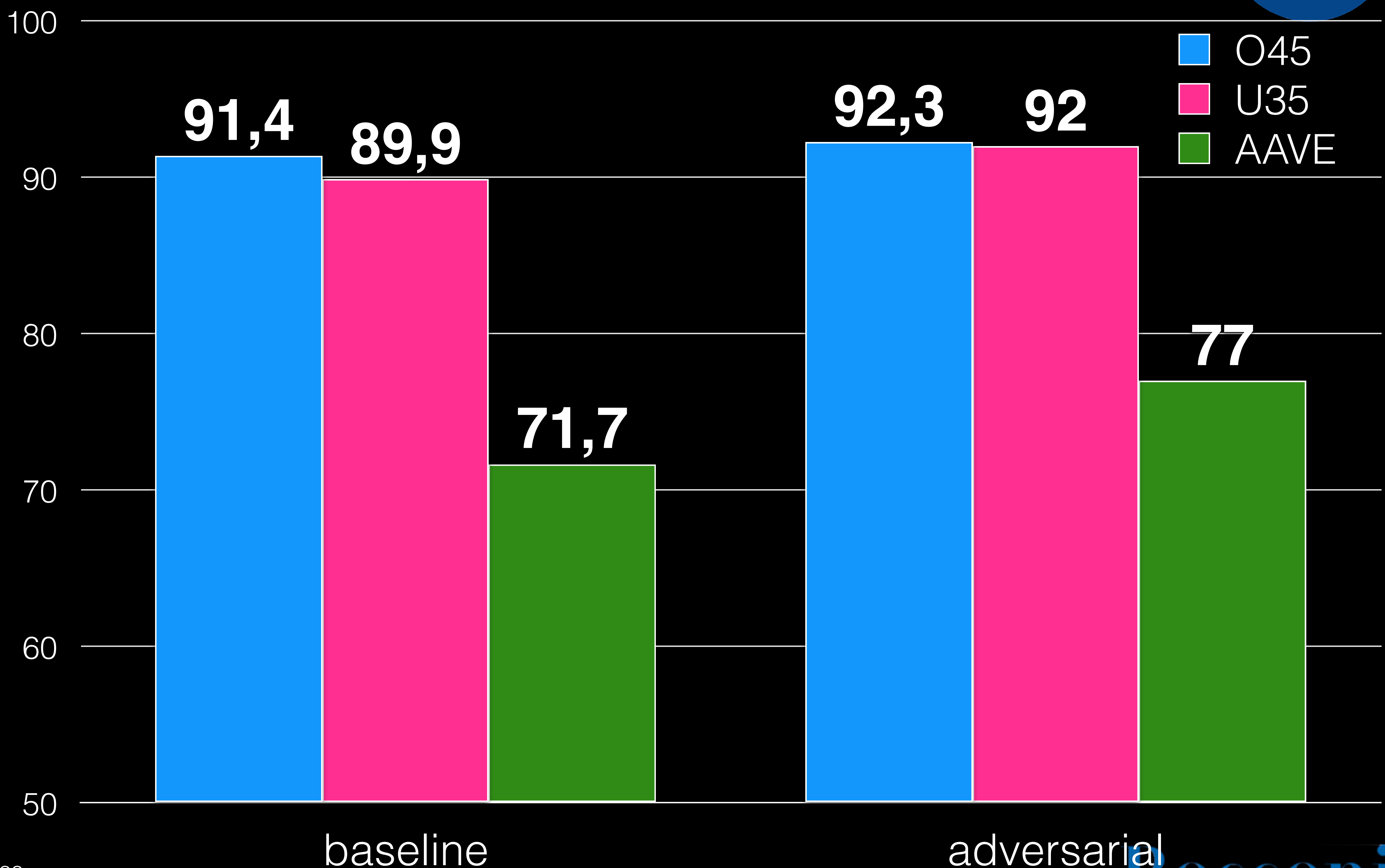
# Adversarial Model

Li et al. (ACL 2018)



Li et al. (ACL 2018)

# Better and Fairer!



# Part 5: Design Bias

# Dual Use

+ Intended Use

– Unintended Use
or consequences

# Normative vs Descriptive Ethics


NORMATIVELY WRONG

DESCRIPTIVELY WRONG

Hovy & Spruit (ACL 2016)
Joshi et al. (2020)

# Over-Exposure



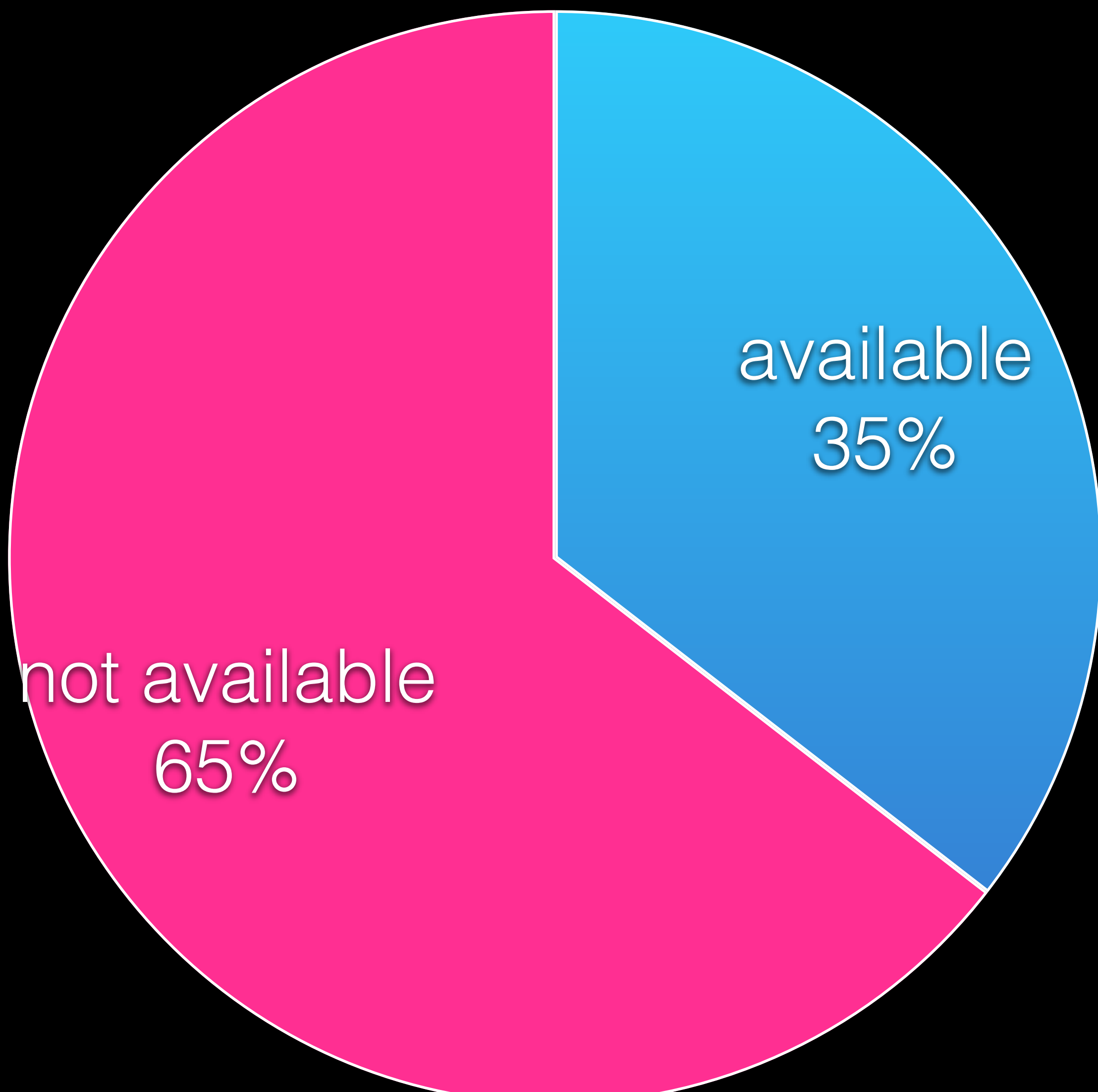
**POS tagging**

**Discourse**

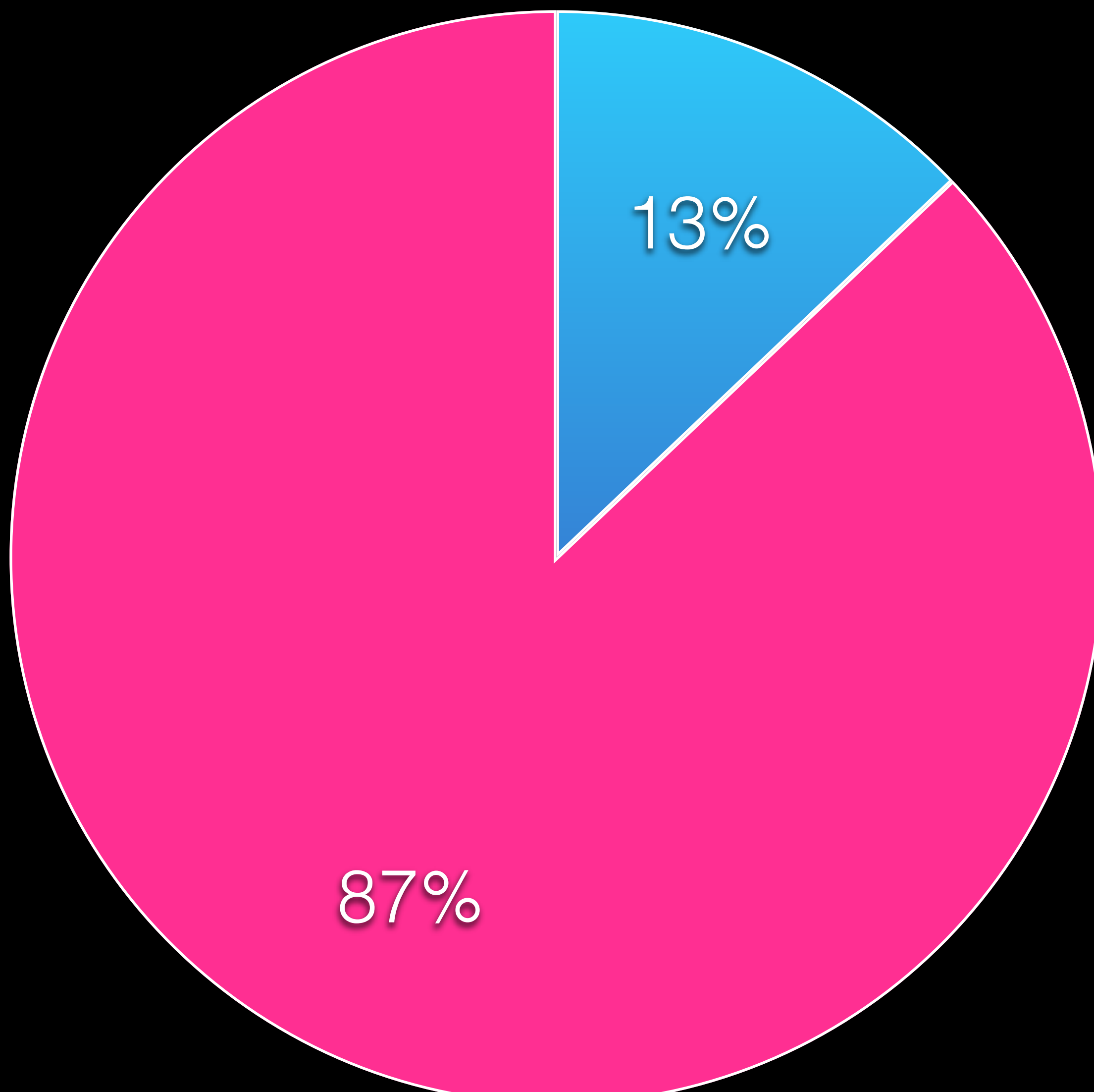
# Under-Exposure

Hovy & Spruit (ACL 2016)
Joshi et al. (2020)

**treebanks** *




**semantic resources**



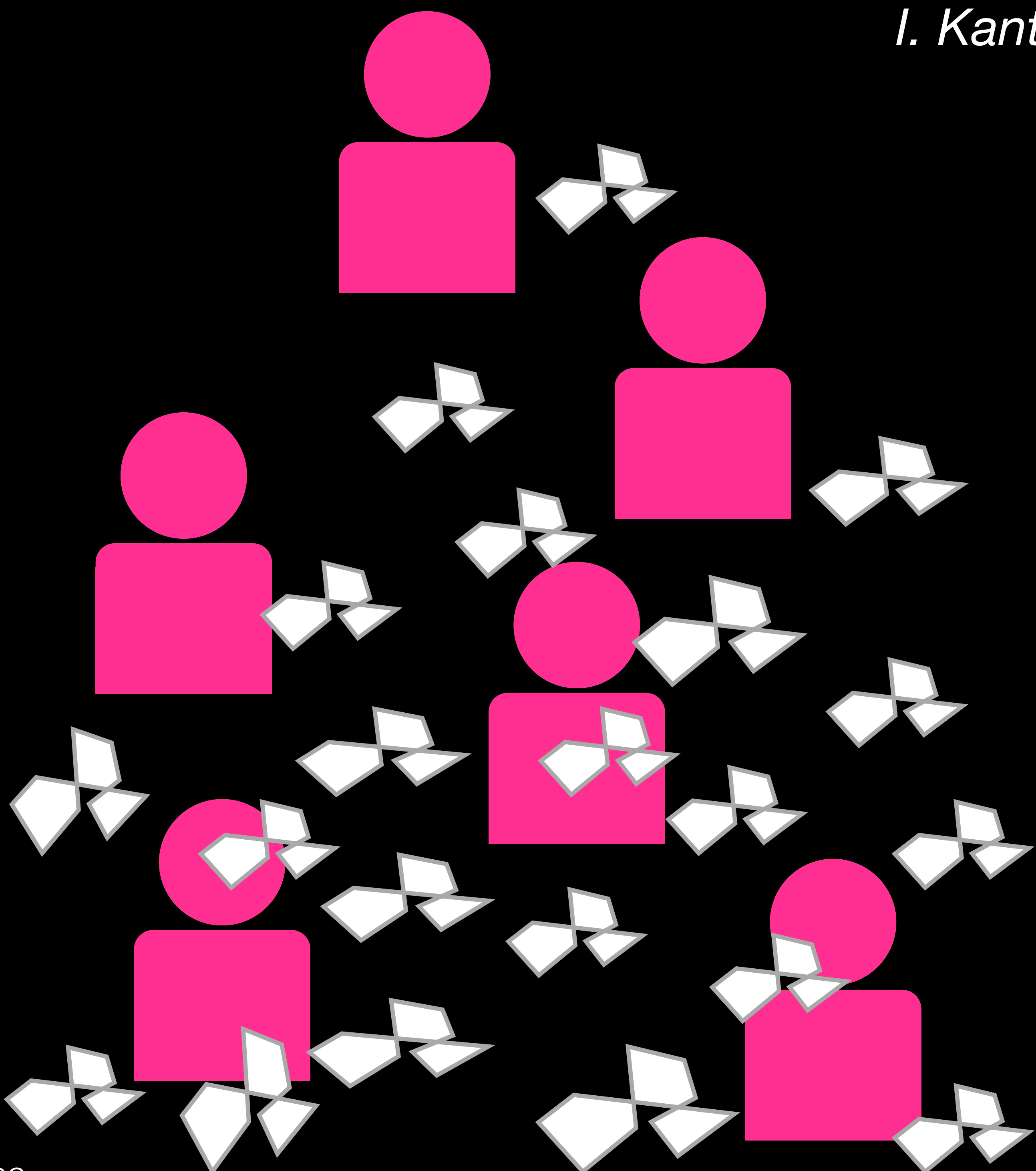

*BEFORE UD... evaluation

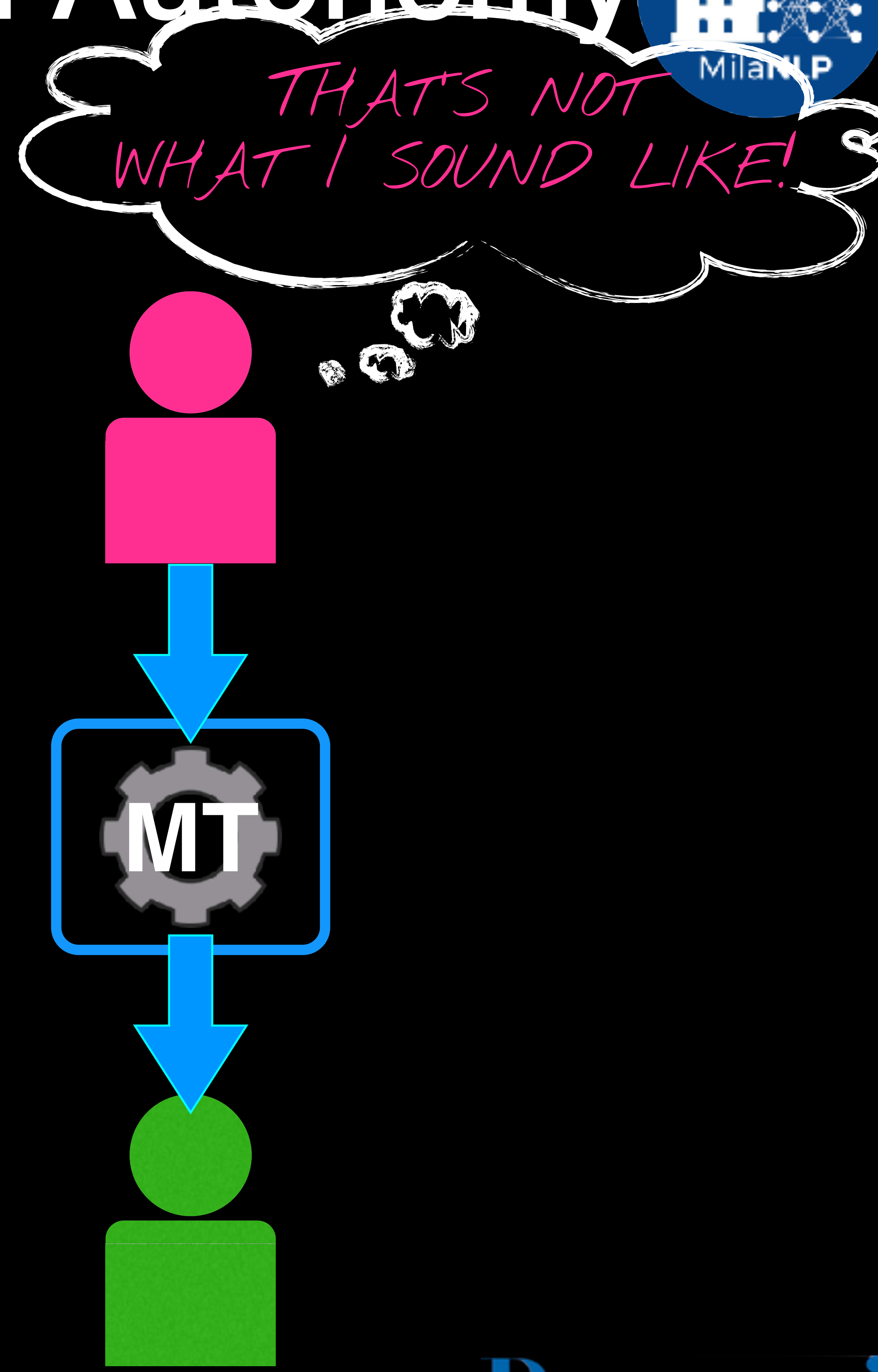
Blodgett et al., 2021, Prabhumoye et al., 2020

# Generalization and Autonomy

*"Act only according to that maxim whereby you can, at the same time, will that it should become a universal law"*

*I. Kant*





Van de Poel (2016)

# Technology as Social Experiment

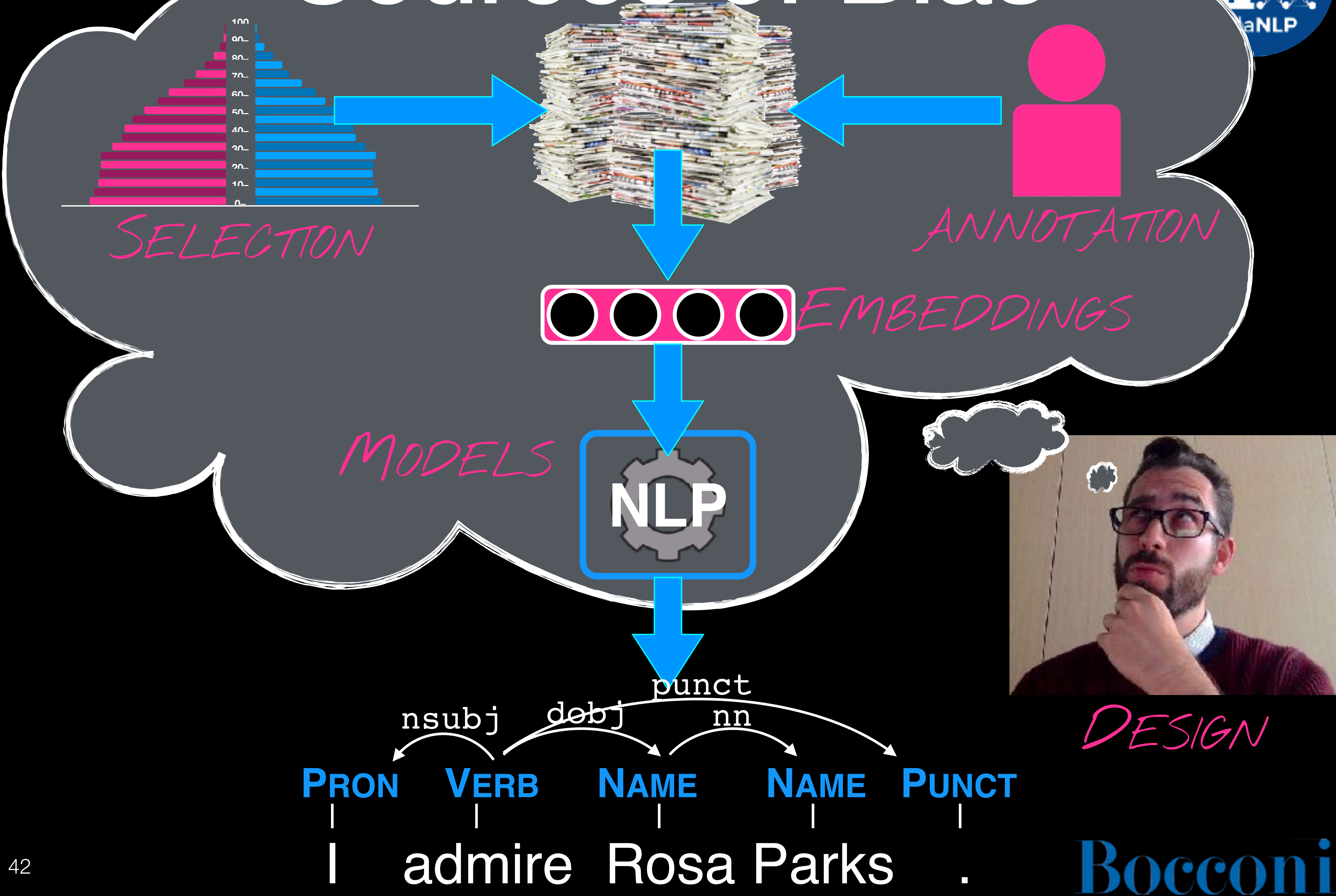
HOW CAN I MAKE IT SAFE?

DO MY SUBJECTS KNOW THEY'RE IN IT?

CAN THEY CONTROL IT, OPT OUT?

# Wrapping Up

# Sources of Bias



Hovy & Spruit (ACL 2016)
Shah et al. (2020)

# What can we do?

| Source | Problem | Countermeasures |
|---|---|---|
| data selection | **Exclusion** | better collection, post-stratification, priors |
| annotation | **Label Bias** | better training, annotation models, disagreement weighting |
| models | **Overgeneralization** | dummy labels, error weighting, adversarial learning |
| research design | **Exposure** | document, consider possible impact, educate |

# Outcomes

**society:**

combat algorithmic racism and sexism,

build fair tools that perform equally well for all users

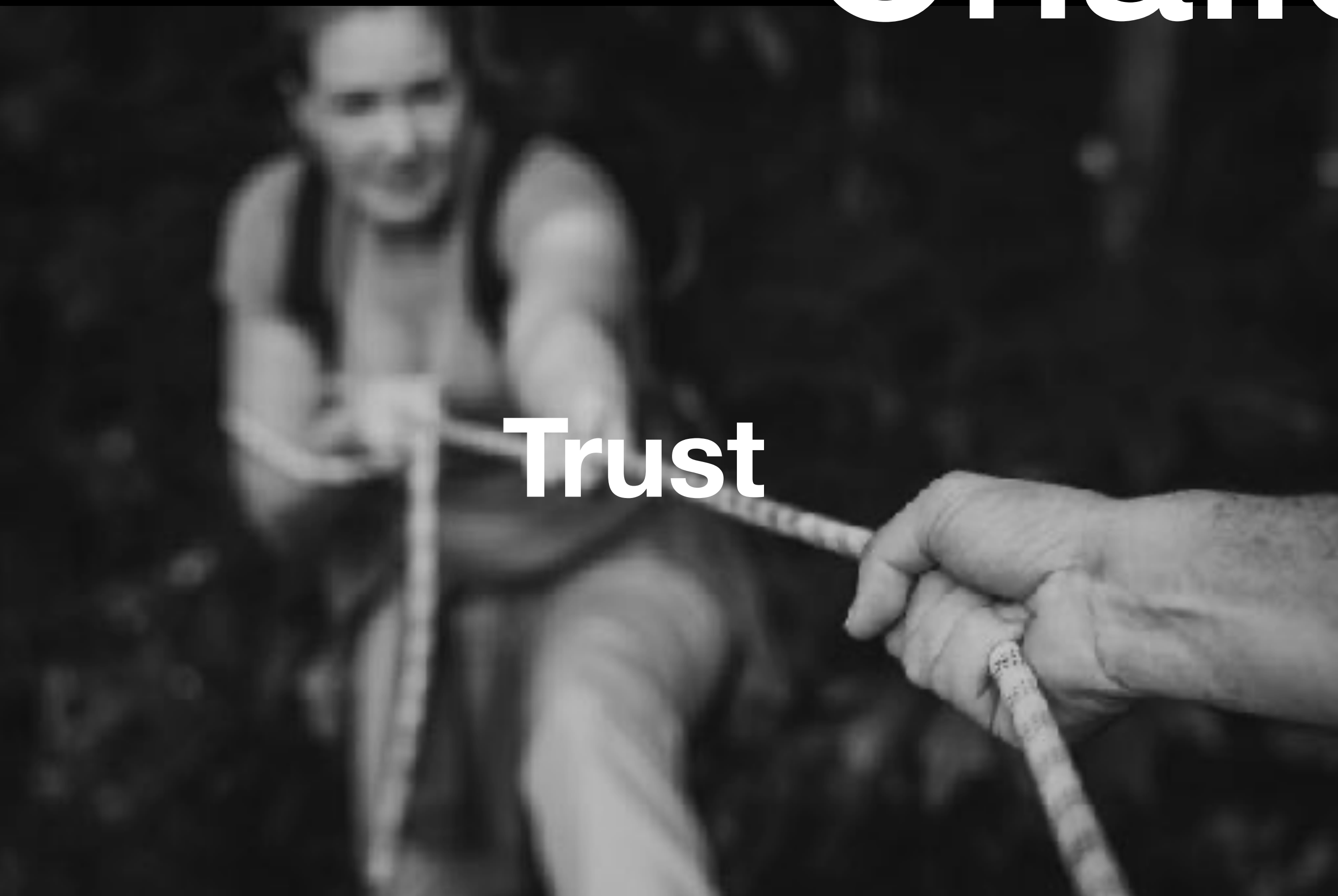
**research:**

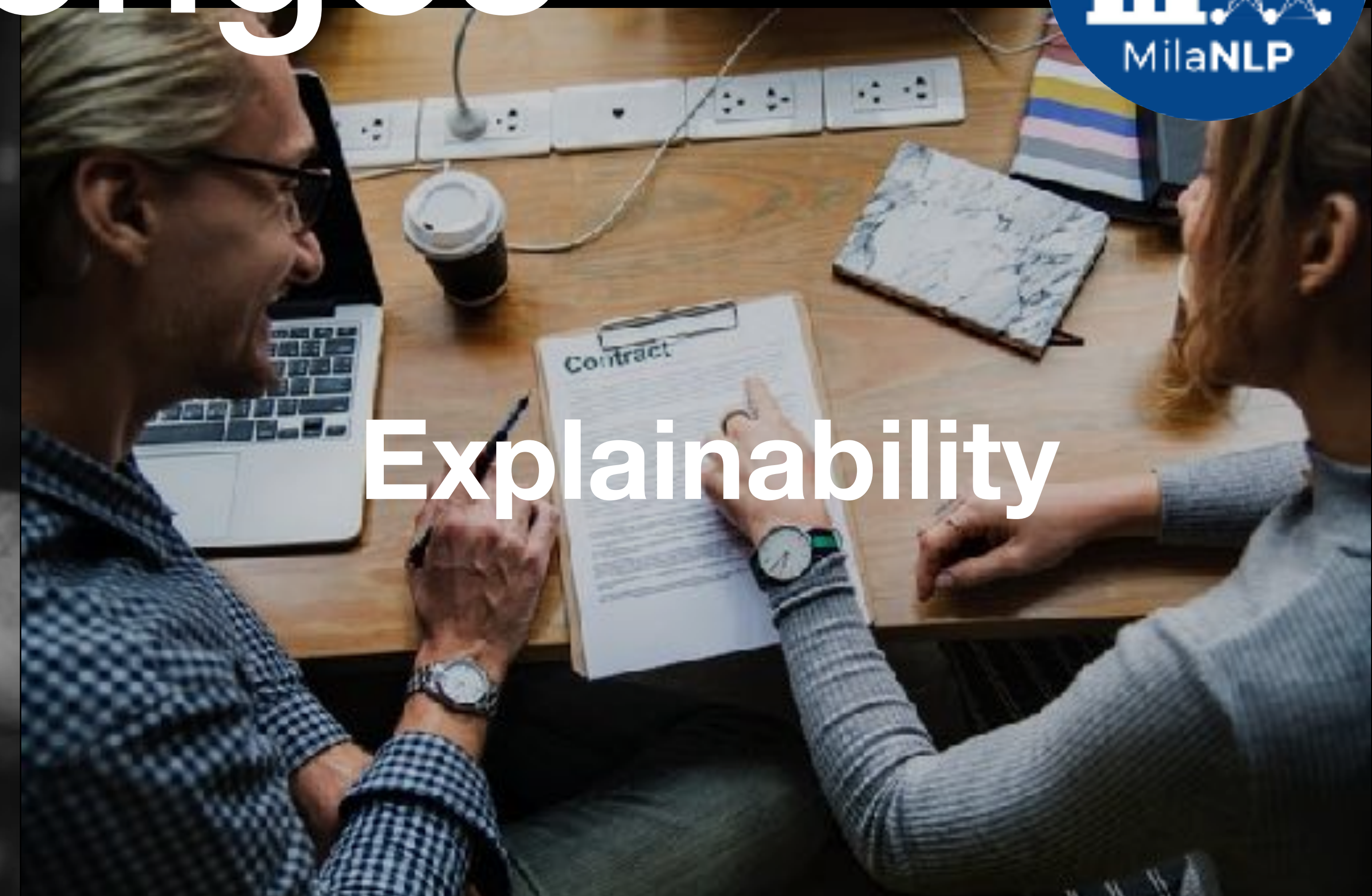
open up new research avenues and subfields

**industry:**

more performant tools in MT, dialogue, search

# Challenges

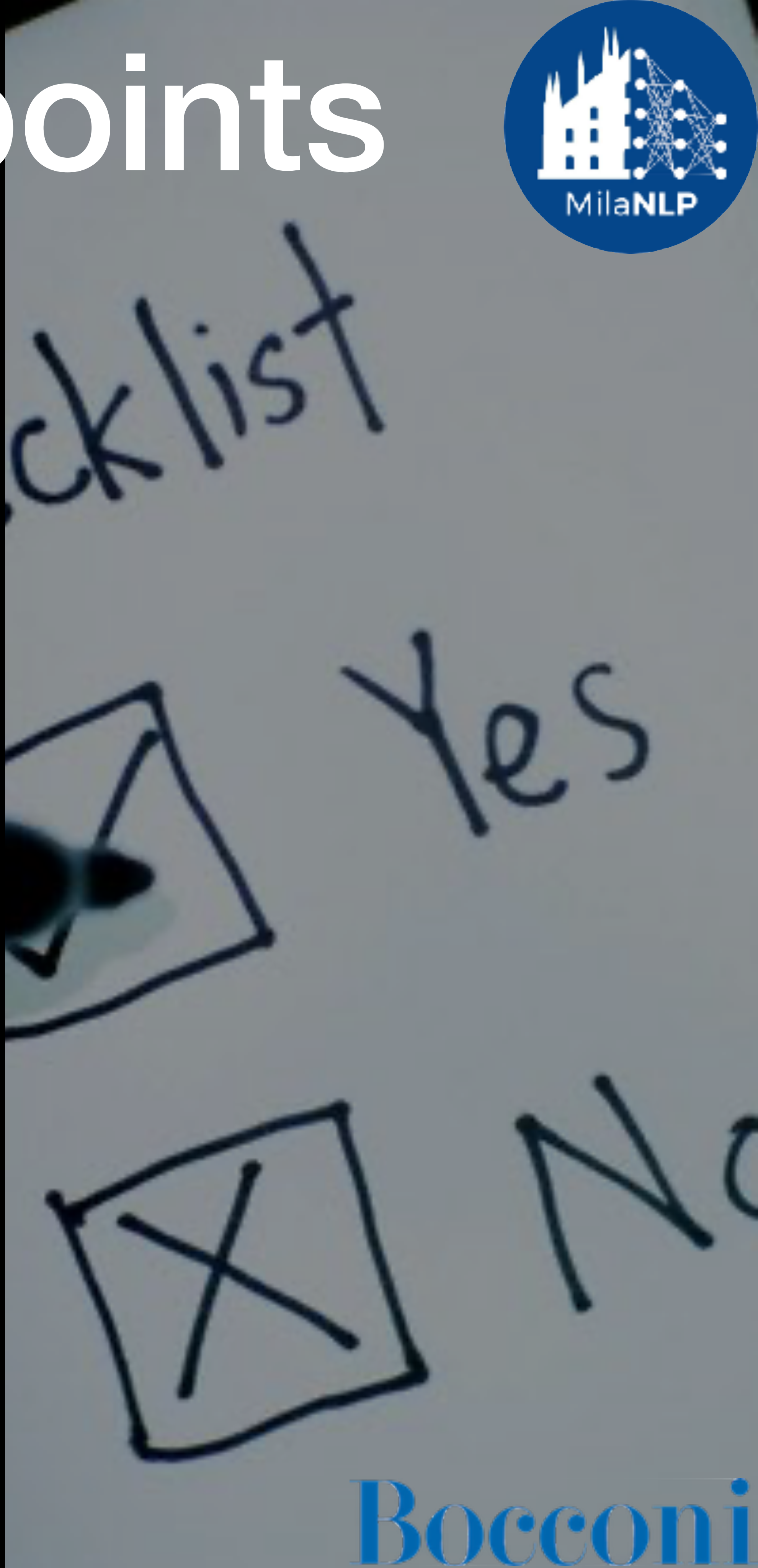




# Take-home points

- Beware of **bias** from **data**, **annotations**, **embeddings**, **models**, and **design**
- Apply **countermeasures** where possible: better for fairness ***and*** performance
- Know your models ***will*** be used in unintended ways



www.dirkhovy.com/portfolio/papers

https://milanlproc.github.io/publication/

# Thank you!

- For the papers in this talk, see:
  - Shah, Schwartz & Hovy (ACL 2020): https://www.aclweb.org/anthology/2020.acl-main.468v2.pdf
  - Hovy & Prabhumoye (2021): https://onlinelibrary.wiley.com/doi/epdf/10.1111/lnc3.12432

@dirk_hovy

www.dirkhovy.com

www.dirkhovy.com/portfolio/papers
https://milanlproc.github.io/publication/

# Questions?

@dirk_hovy
www.dirkhovy.com