Five Sources of Bias in NLP (and What to Do about Them)

Dirk Hovy
Bocconi University
Milan, Italy

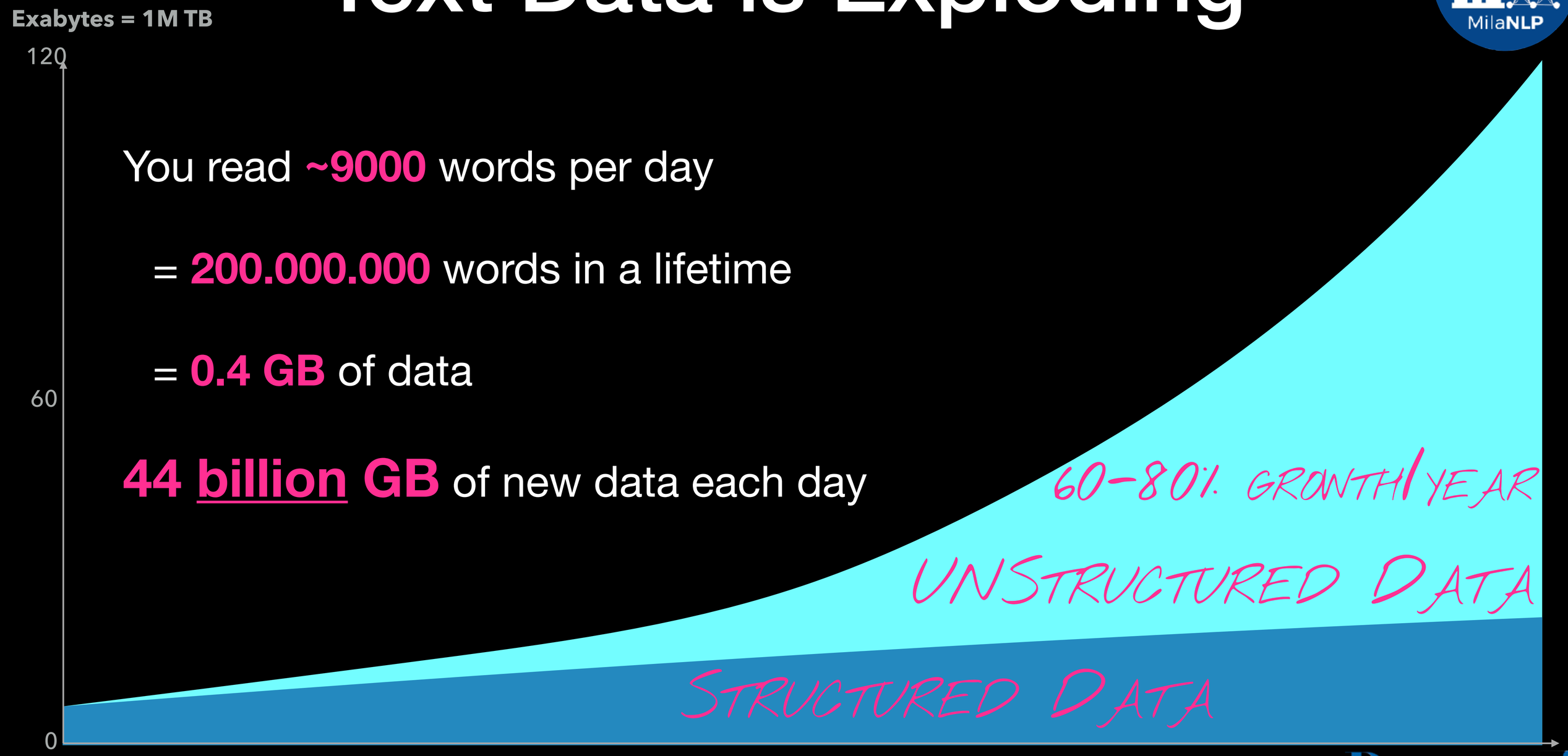
www.dirkhovy.com
mail@dirkhovy.com
@dirk_hovy

MilaNLP

erc
INTEGRATOR
European Research Council

Bocconi

Text Data is Exploding



You read **~9000** words per day

= **200.000.000** words in a lifetime

= **0.4 GB** of data

44 billion GB of new data each day

Source: IDC

NLP is Everywhere

MilaNLP

Bocconi

NLP is Everywhere

PROMPT:

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.

GPT-3:

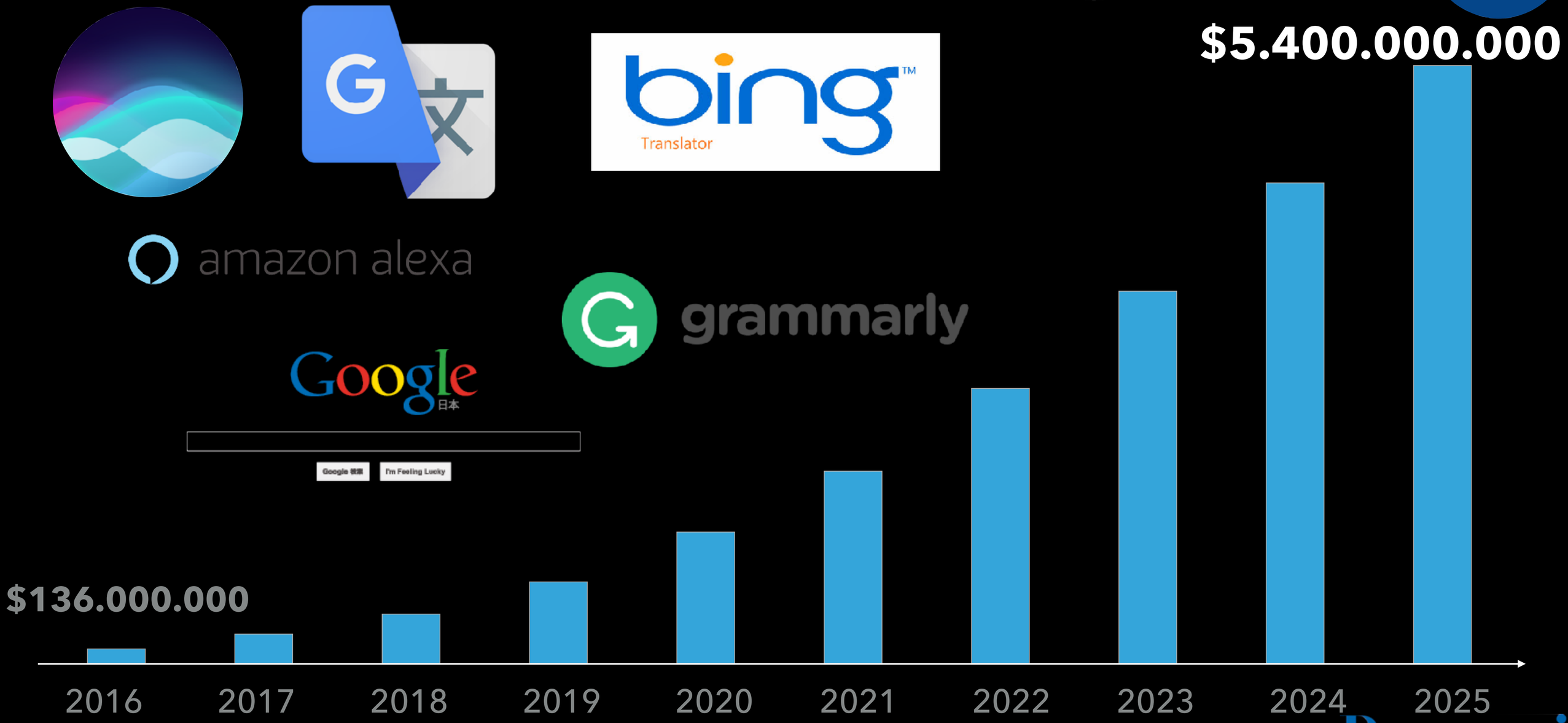
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

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.

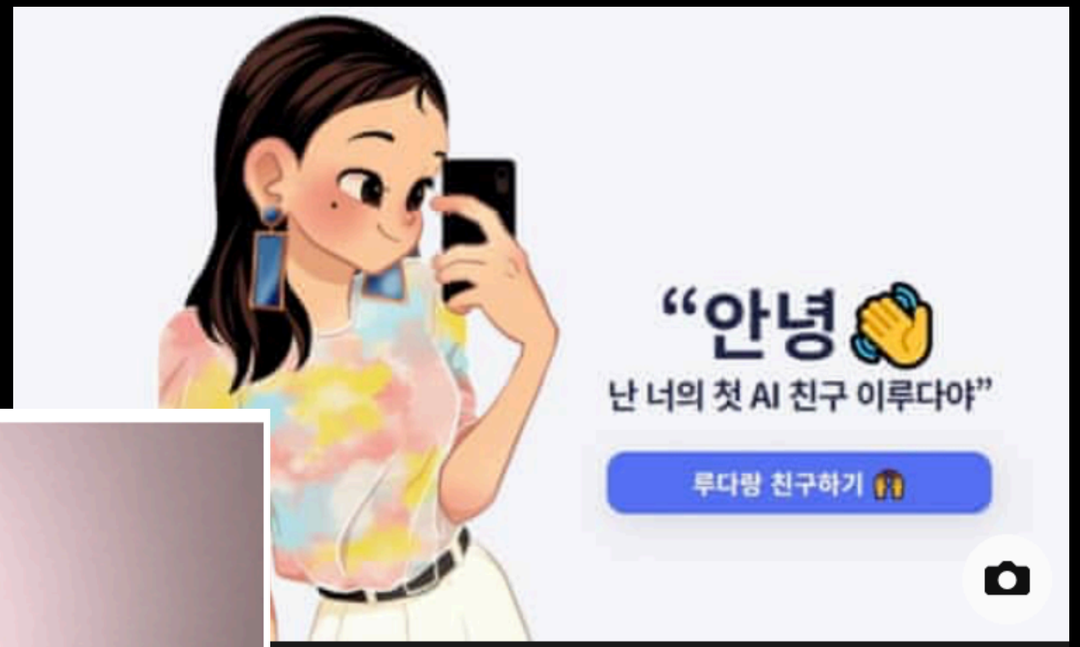
NLP is booming



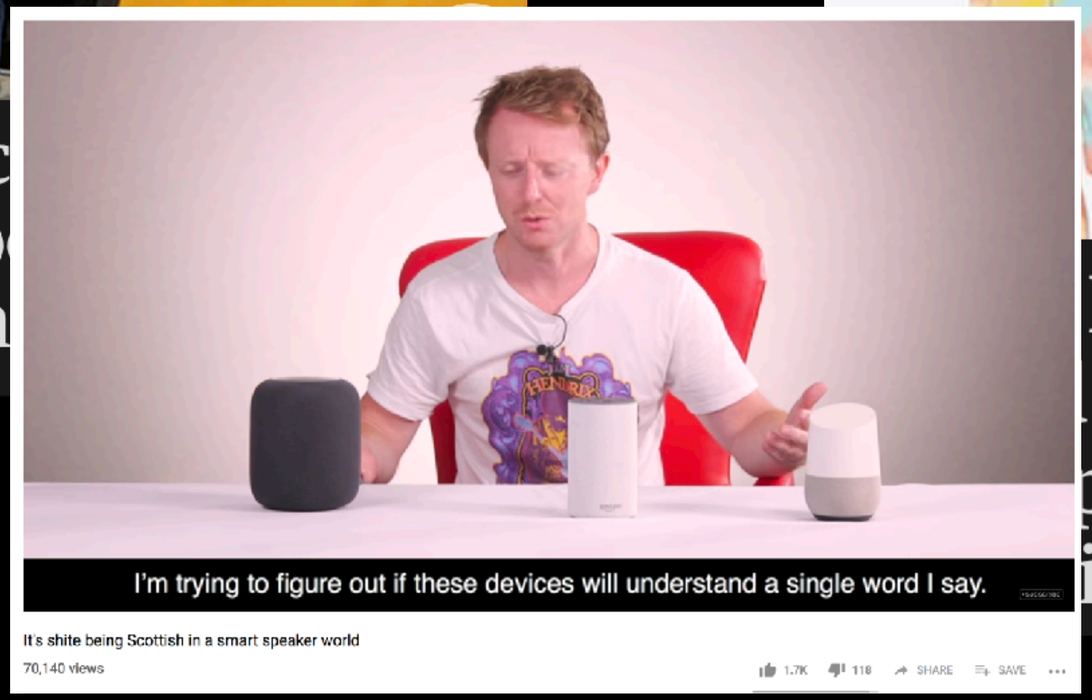
Source: Tractica

But: Does it Work?

Amazon ditc
recruiting to
men for tech

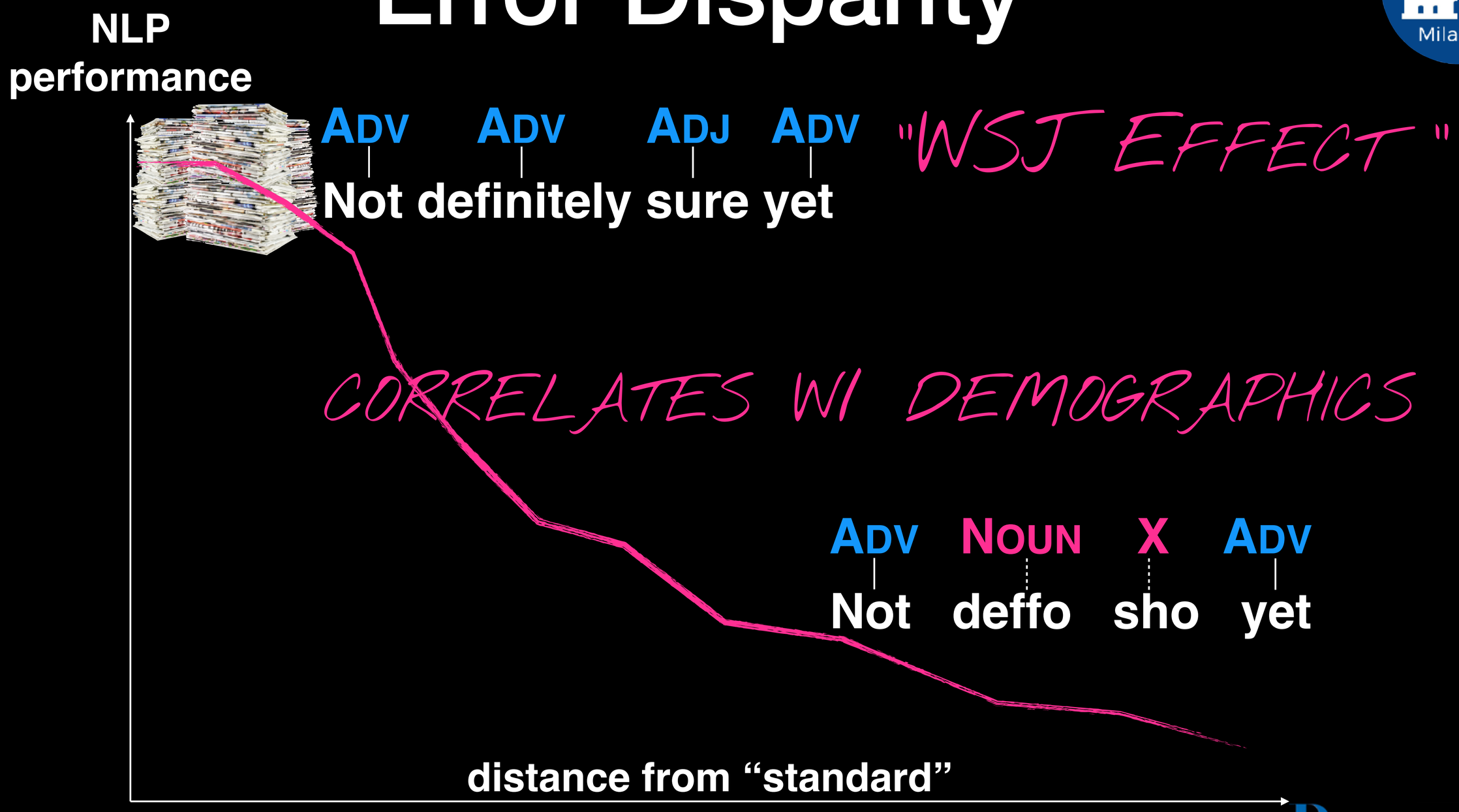


Korean AI chatbot
from Facebook after
peech towards
ities



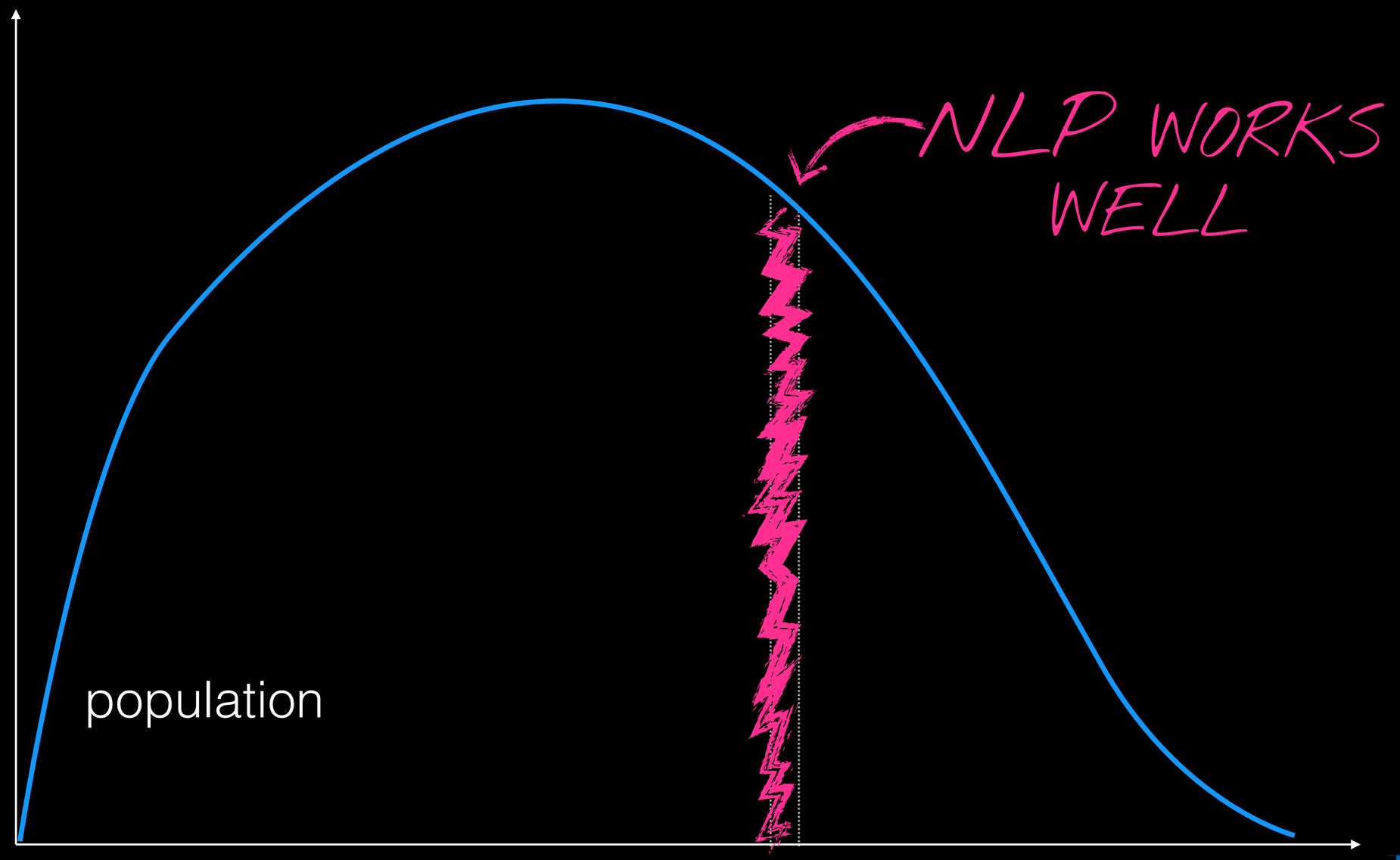
Bisk et al., 2020; Bender and Koller, 2020; Ettinger 2020; Bianchi & Hovy, 2021; Bender et al., 2021

Error Disparity



7 Jørgensen et al. (2015); Hovy & Søgaard (2015)

Unequal Impact



Goals for Today

- Introduce **terminology** and **concepts** of bias, privacy, and dual use
- Show 5 **sources** of bias in NLP
- Discuss **consequences**
- Introduce **countermeasures**

DISCLAIMER!

- active research area: I will miss some relevant work!
- overview: I have not been involved in most of the works (I wish I was…)

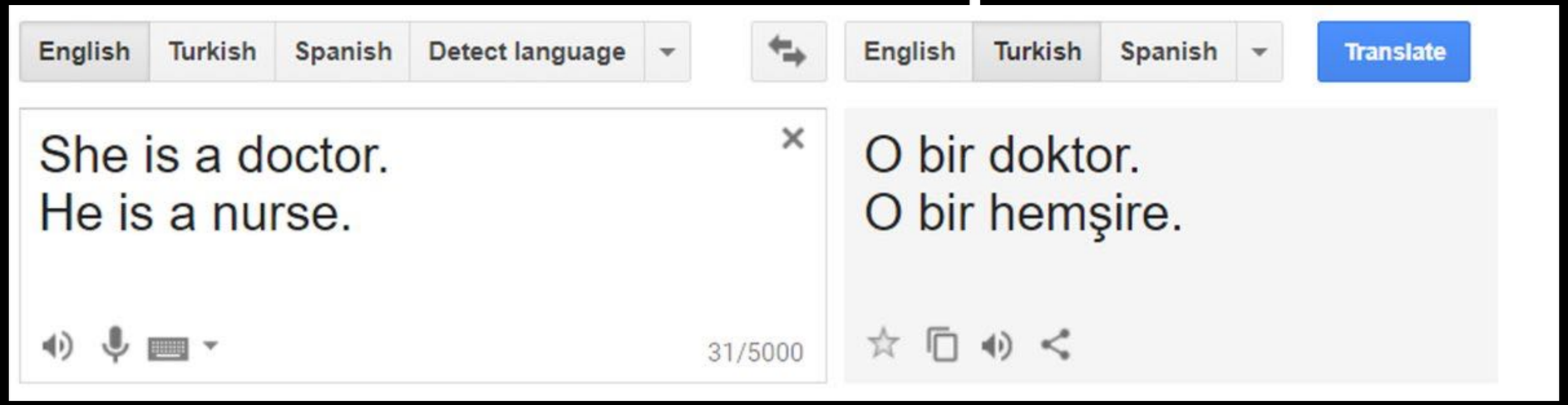
Terminology

Bias

BIAS!

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Normative vs Descriptive Ethics



NORMATIVELY WRONG

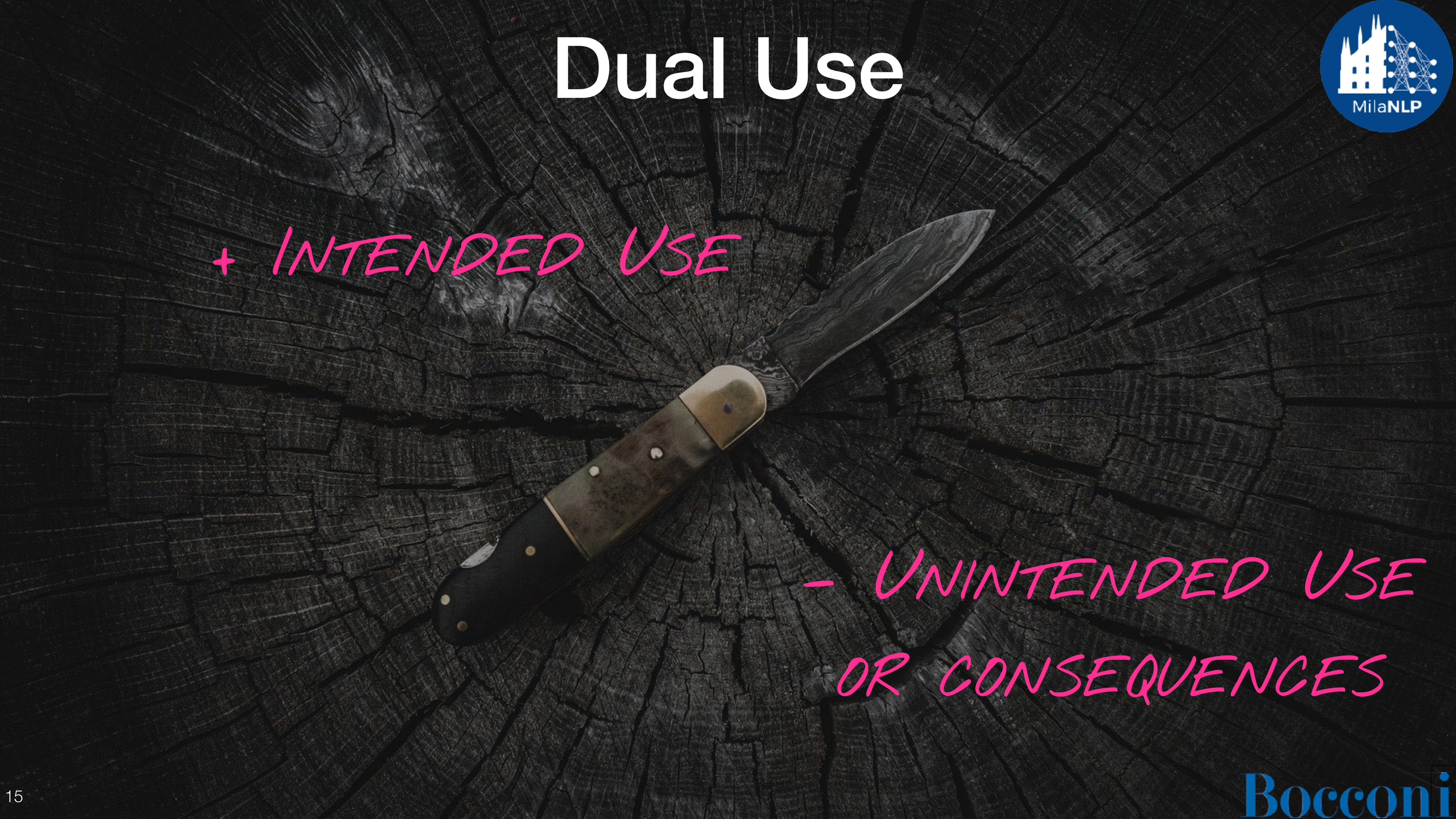
DESCRIPTIVELY WRONG

Normative vs Descriptive Ethics

NORMATIVELY WRONG

DESCRIPTIVELY TRUE?

Dual Use



+ Intended Use

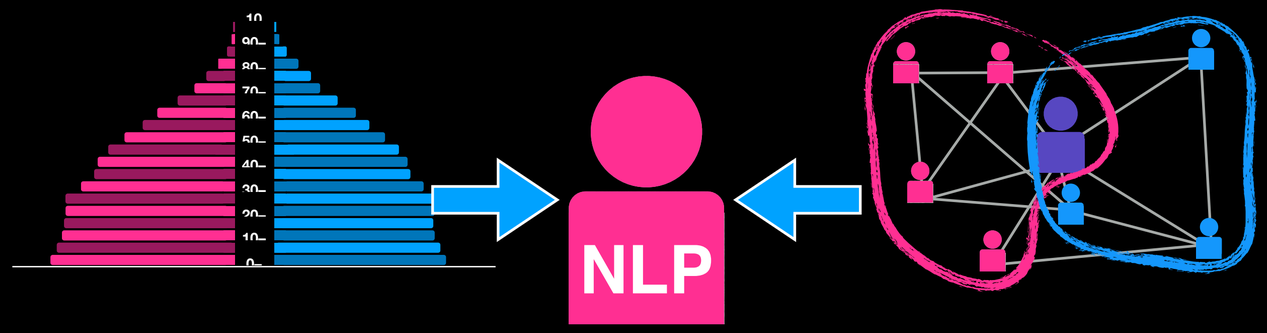
- Unintended Use or consequences

Easy Availability

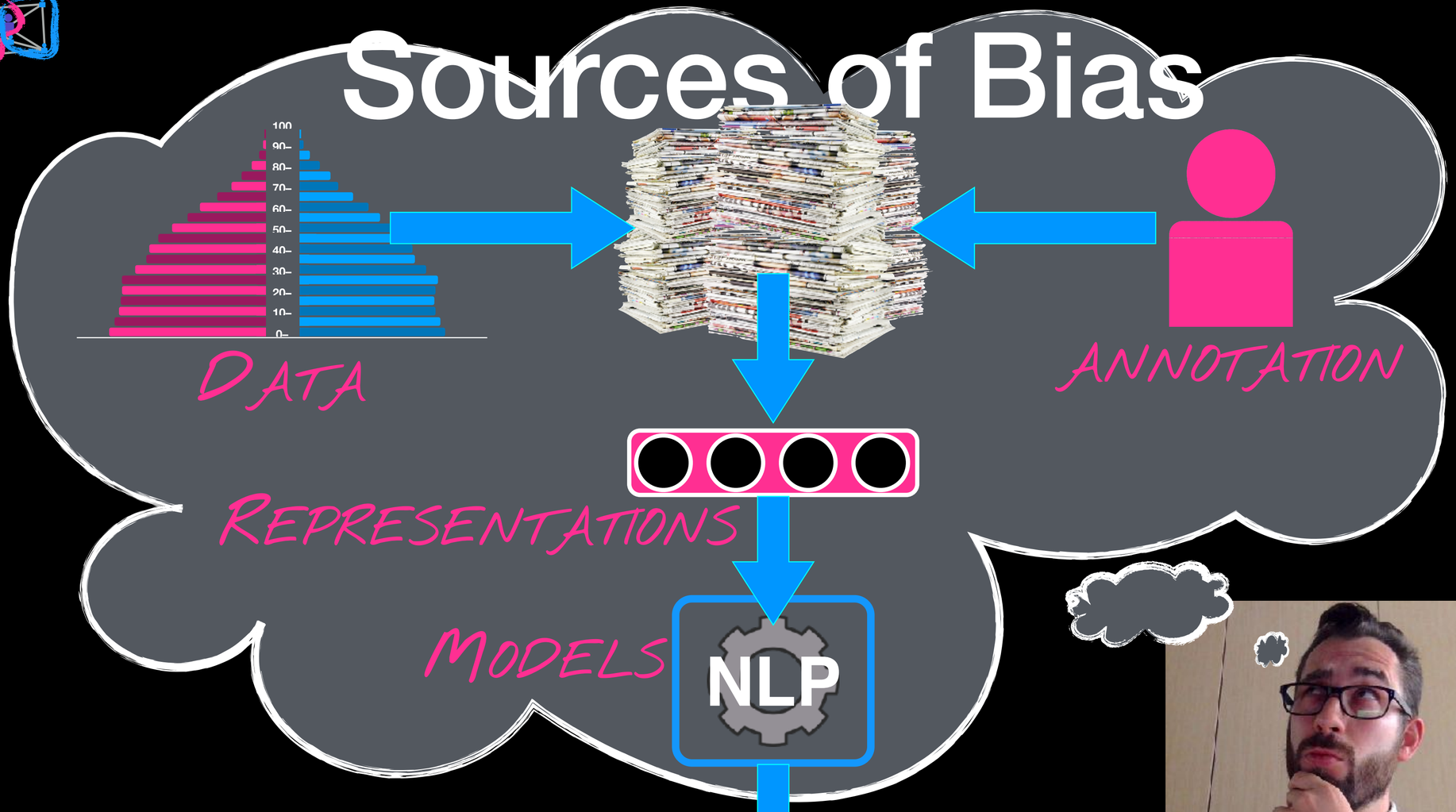
(Big) Pros: Great Boon for Democratization, Inclusion, and Better Access!

Cons: Access to powerful tools without necessary training

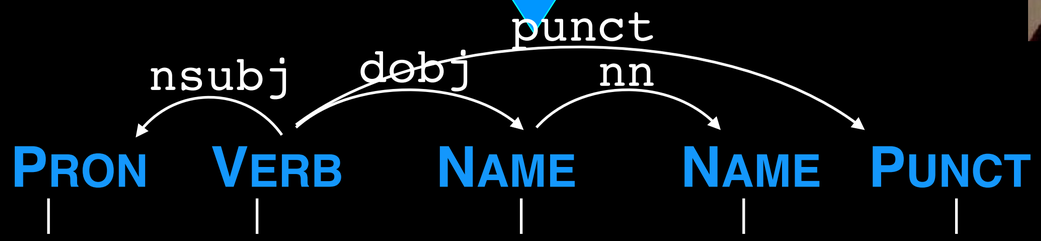
Part I:
Sources of Bias



Sources of Bias



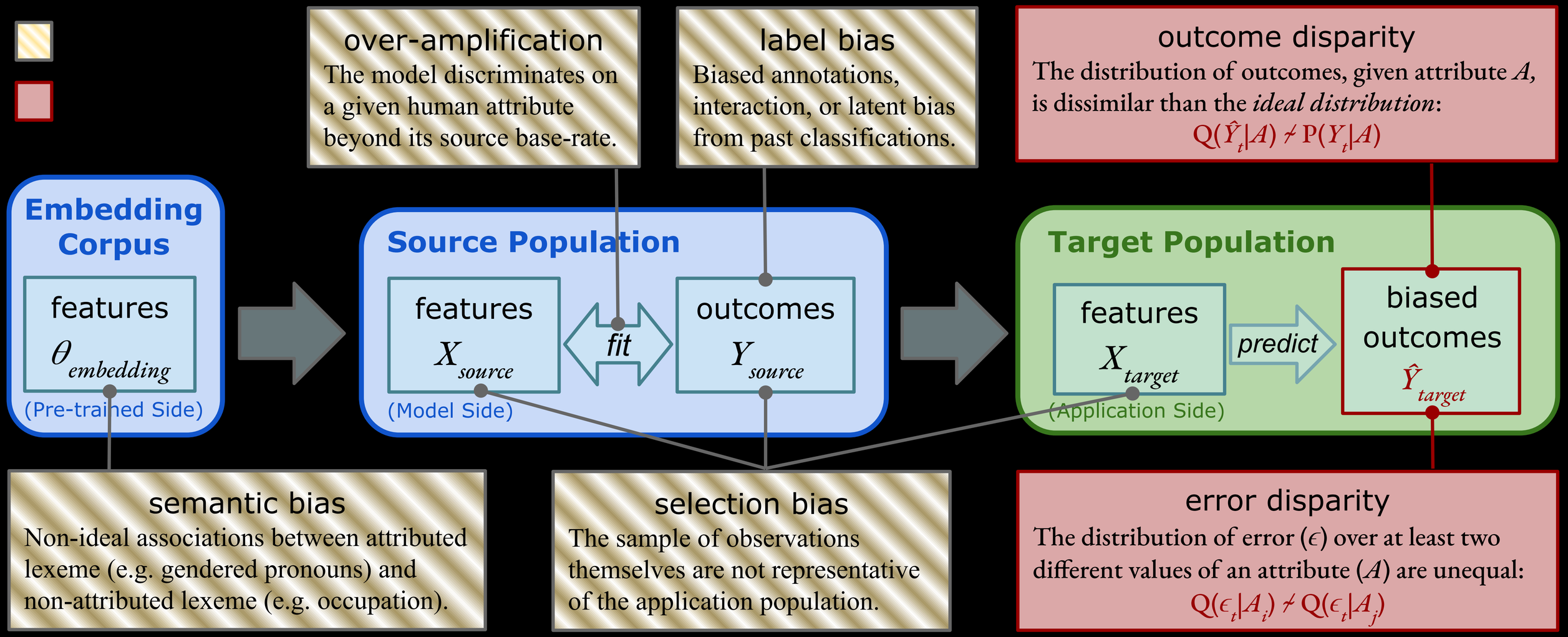
DESIGN



I admire Rosa Parks .

18 Shah et al., 2020; Hovy & Prabhumoye, 2021

Bias Taxonomies



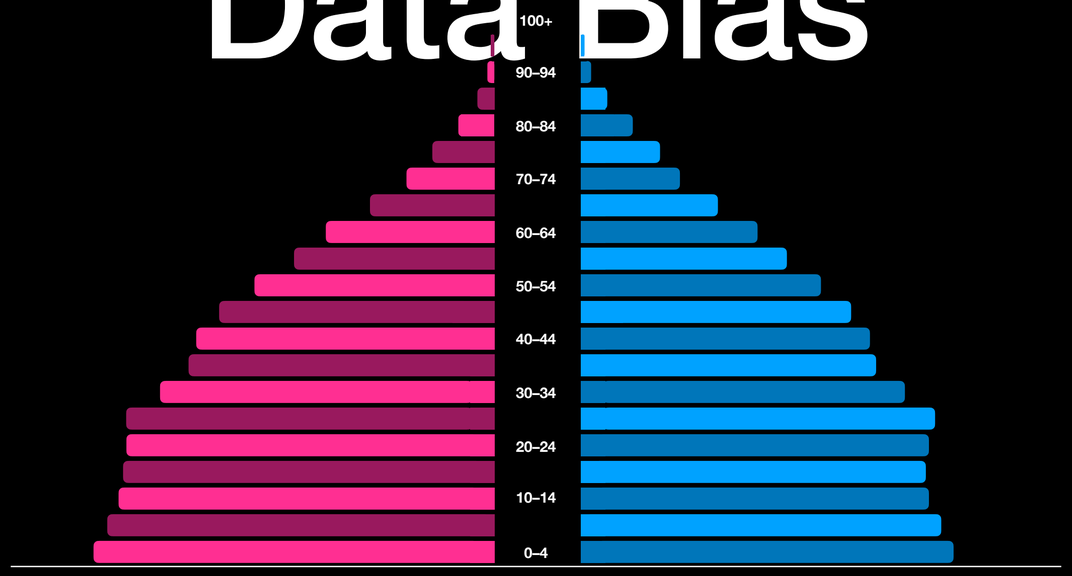
19 Shah, Schwartz, & Hovy, 2020; Blodgett et al., 2020; Hovy & Prabhumoye, 2021

Ideal Distributions

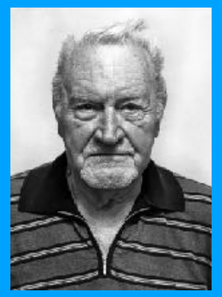
≠ UNIFORM!

- specific to application
- could be unknown
- normative element

Part I.1: Data Bias

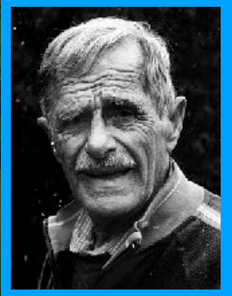


Language Varies



Example 1

Example 2



Example N

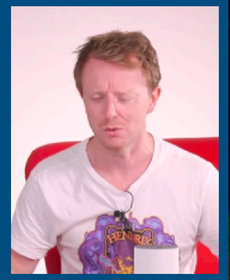
SO THIS IS HOW LANGUAGE WORKS!

System

You sound different…

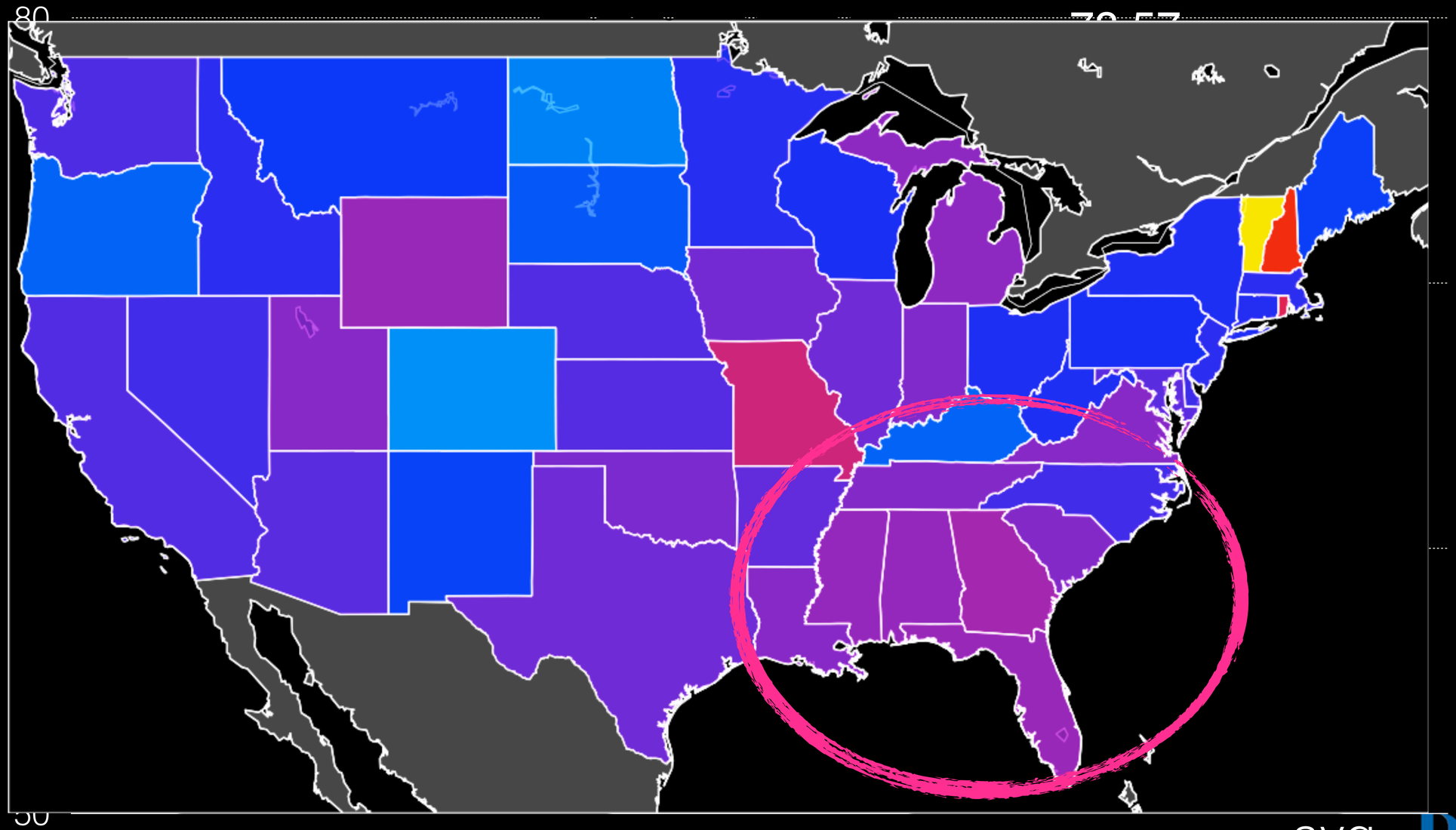
Hello, computer

Shite…



Exclusion

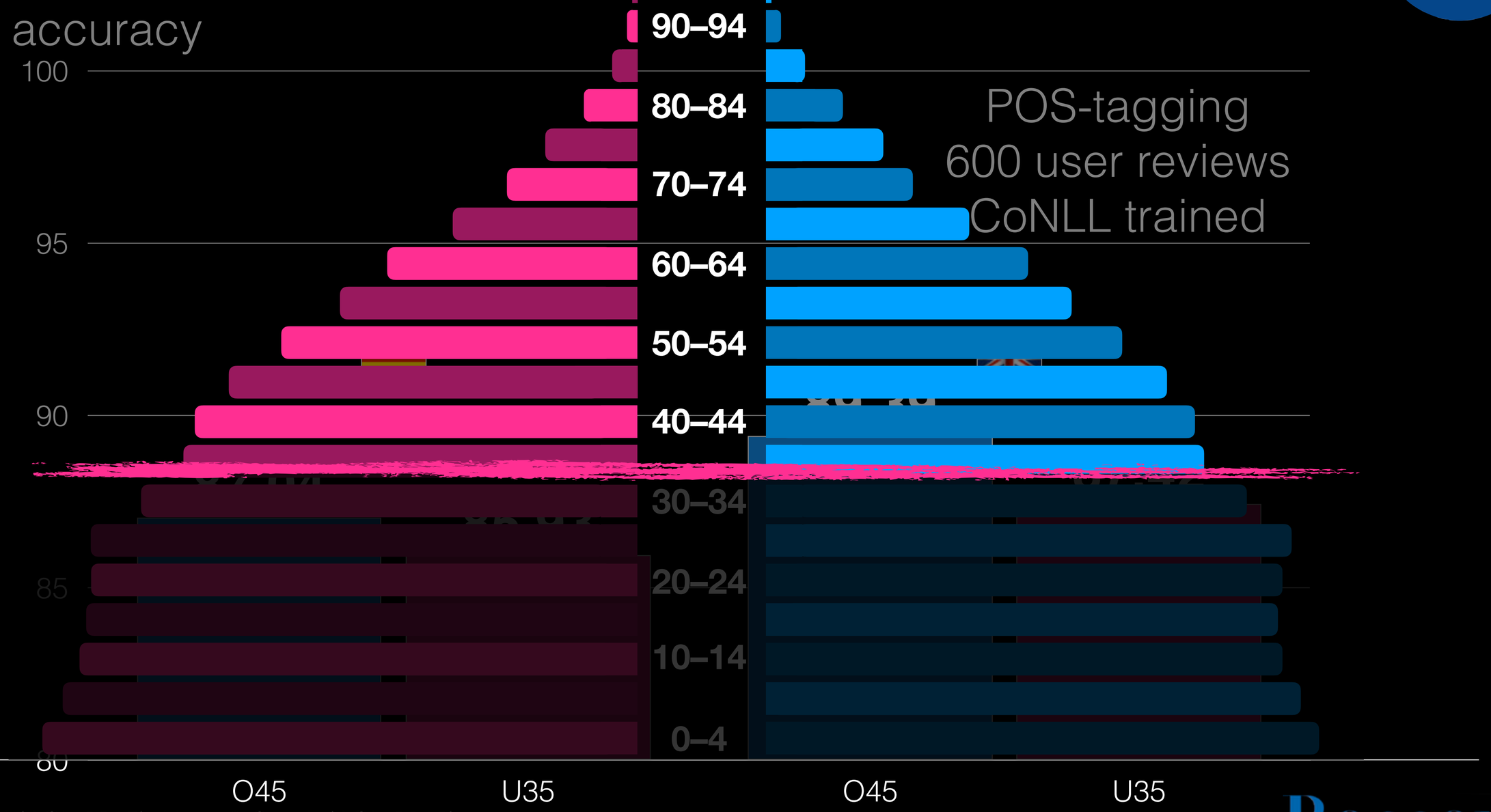
F1



avg

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

Exclusion



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

Better Selection

Example 1

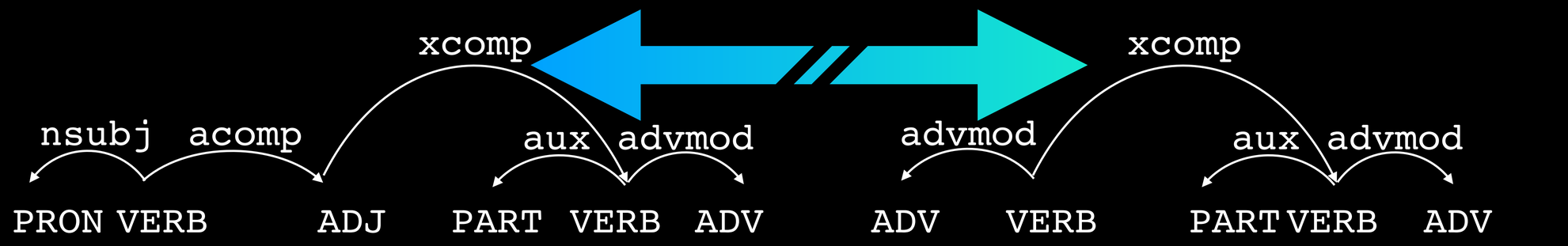
Example 2

Example N

YES, BUT...

Performing Demographics

SUBCONSCIOUS



I am excited to be here So pumped to be here

CONSCIOUS CHOICES

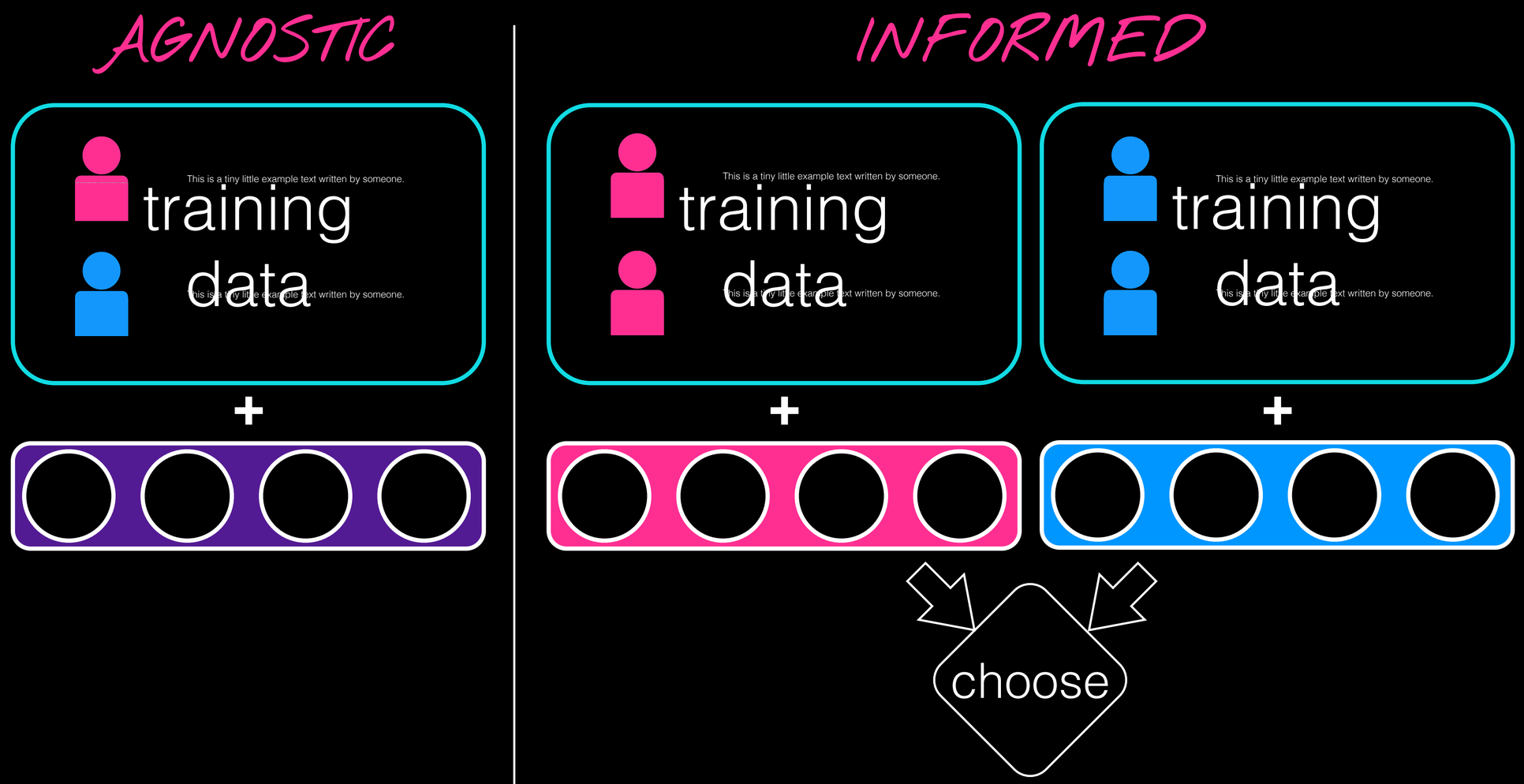
Demographics Matter

That was a sick performance!

Idea!

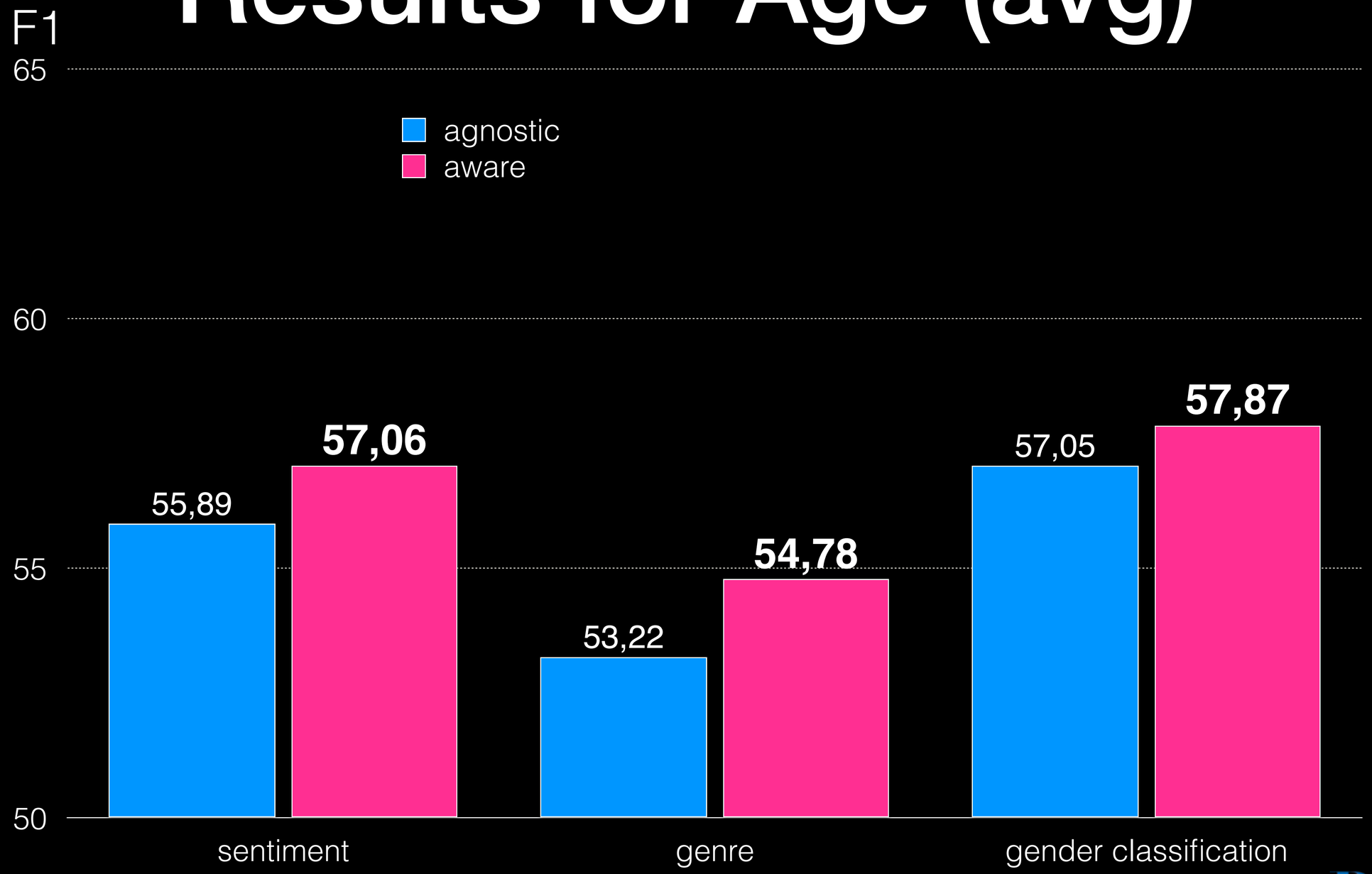
INCLUDE DEMOGRAPHIC INFORMATION IN TEXT REPRESENTATION

Systems



 Hovy (ACL 2015)

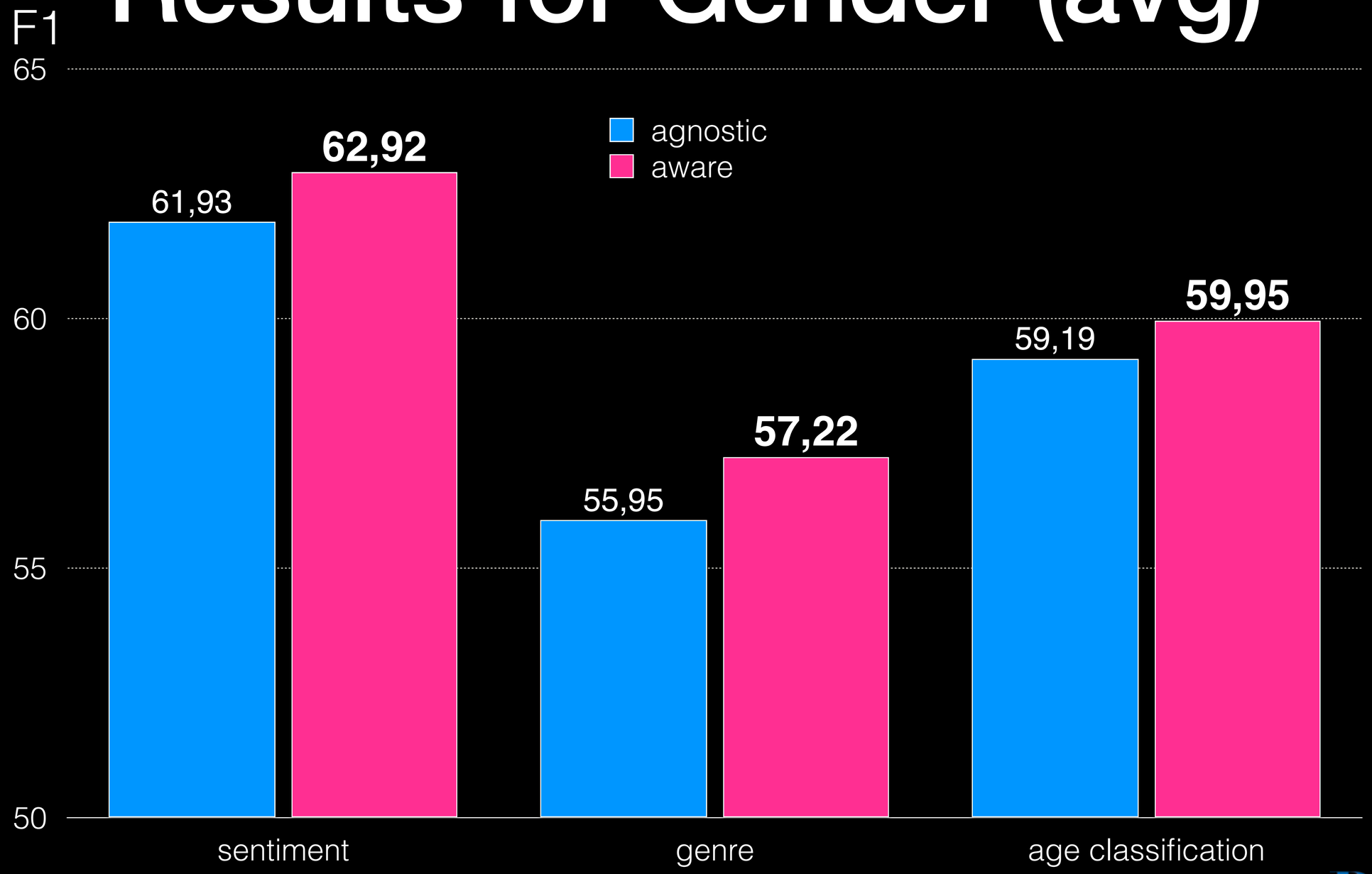
Results for Age (avg)



avg

 Hovy (ACL 2015)

Results for Gender (avg)

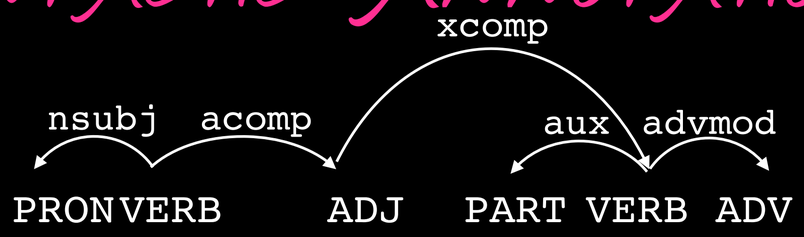


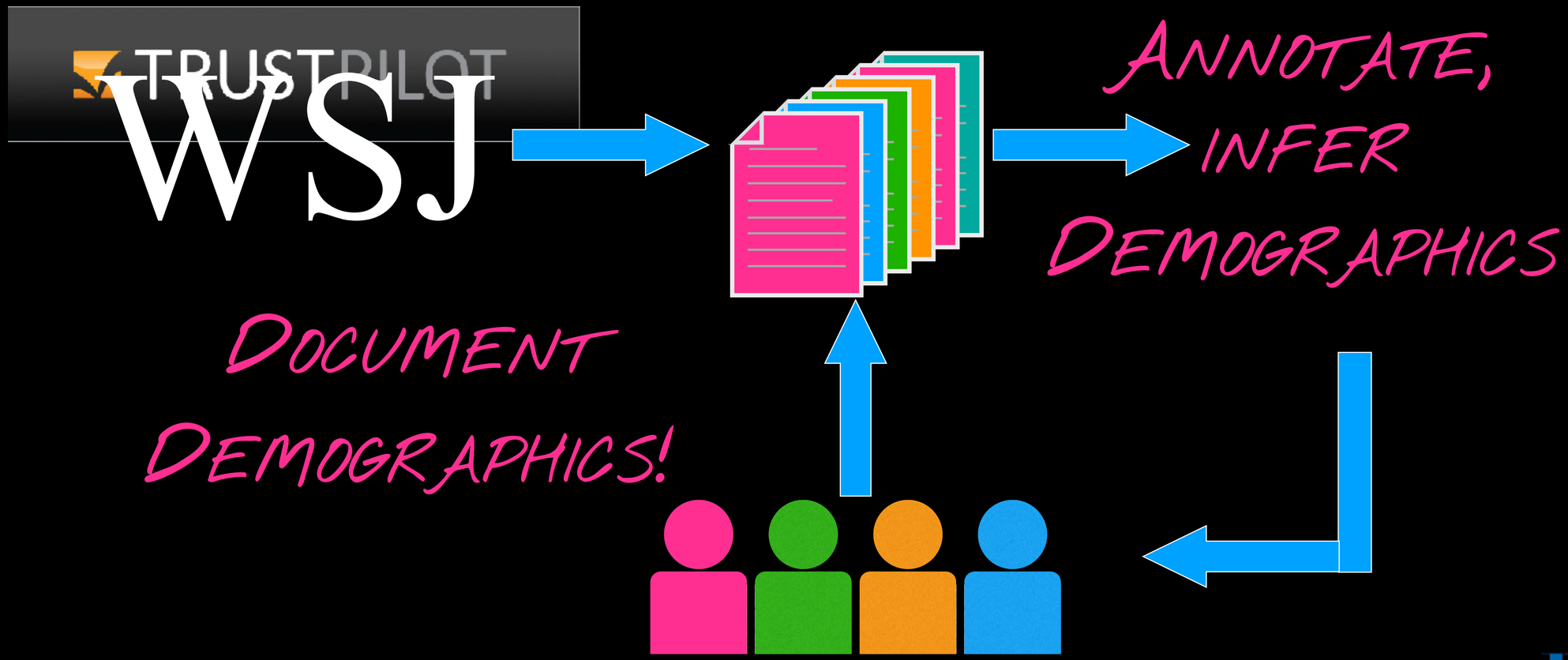
avg

 Hovy (ACL 2015)

We need Better Data

Syntactic Annotations





32 Garimella et al. 2019; Bender & Friedman, 2018

Ok, but…

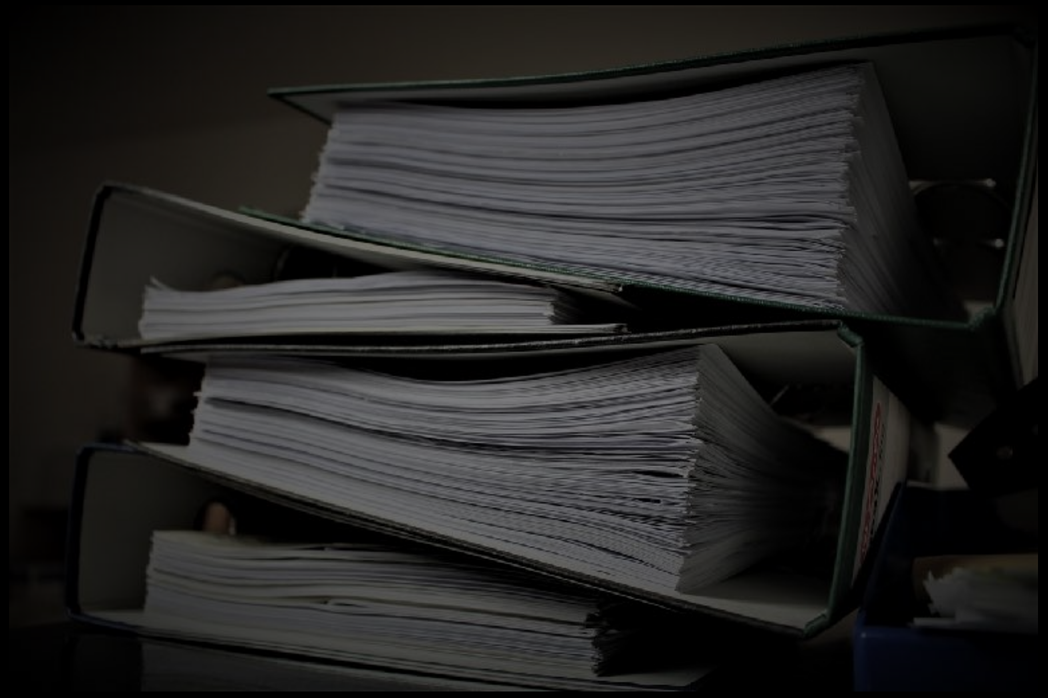
What if we don't know
Want to know the
author's demographics?

Idea!

RECORD POSSIBLE SOURCES OF BIAS

Data Statements

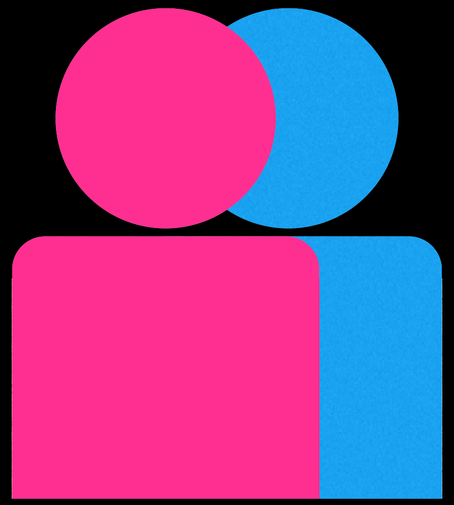
- CURATION RATIONALE
- LANGUAGE VARIETY
- SPEAKER DEMOGRAPHIC
- ANNOTATOR DEMOGRAPHIC
- SPEECH SITUATION
- TEXT CHARACTERISTICS
- RECORDING QUALITY
- OTHER



RECONSTRUCT COLLECTION

Bender & Friedman 2018

Part I.2:
Label Bias



Annotator Bias

Whatever, it's X

HAS NO CLUE...

No! It's an adposition!

PRON	VERB	**X**	NOUN NUM
PRON	VERB	**ADP**	NOUN NUM

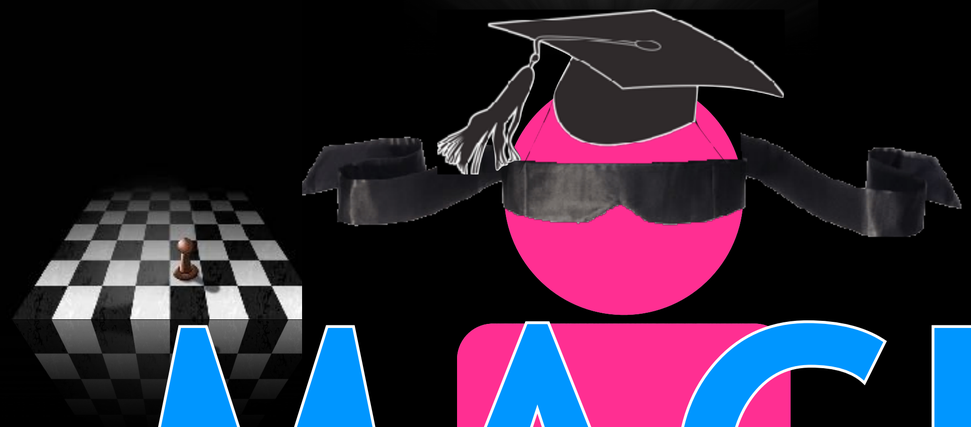
it comes out apr 30

Idea!

FIND OUT WHO'S RELIABLE!

38 Hovy et al. (2013); Passonneau & Carpenter (2014); Paun et al. (2018)

Model

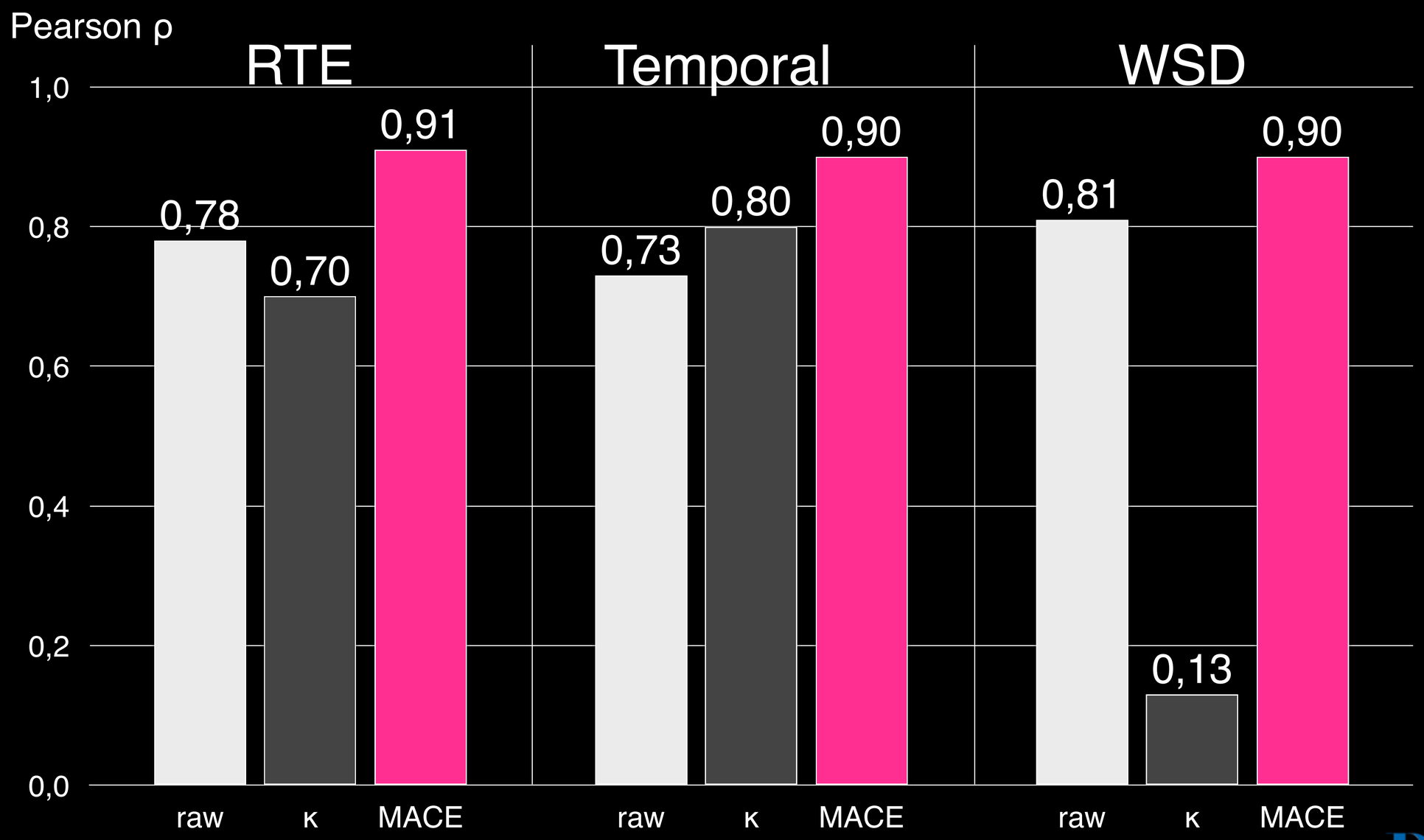


MACE

https://github.com/dirkhovy/MACE

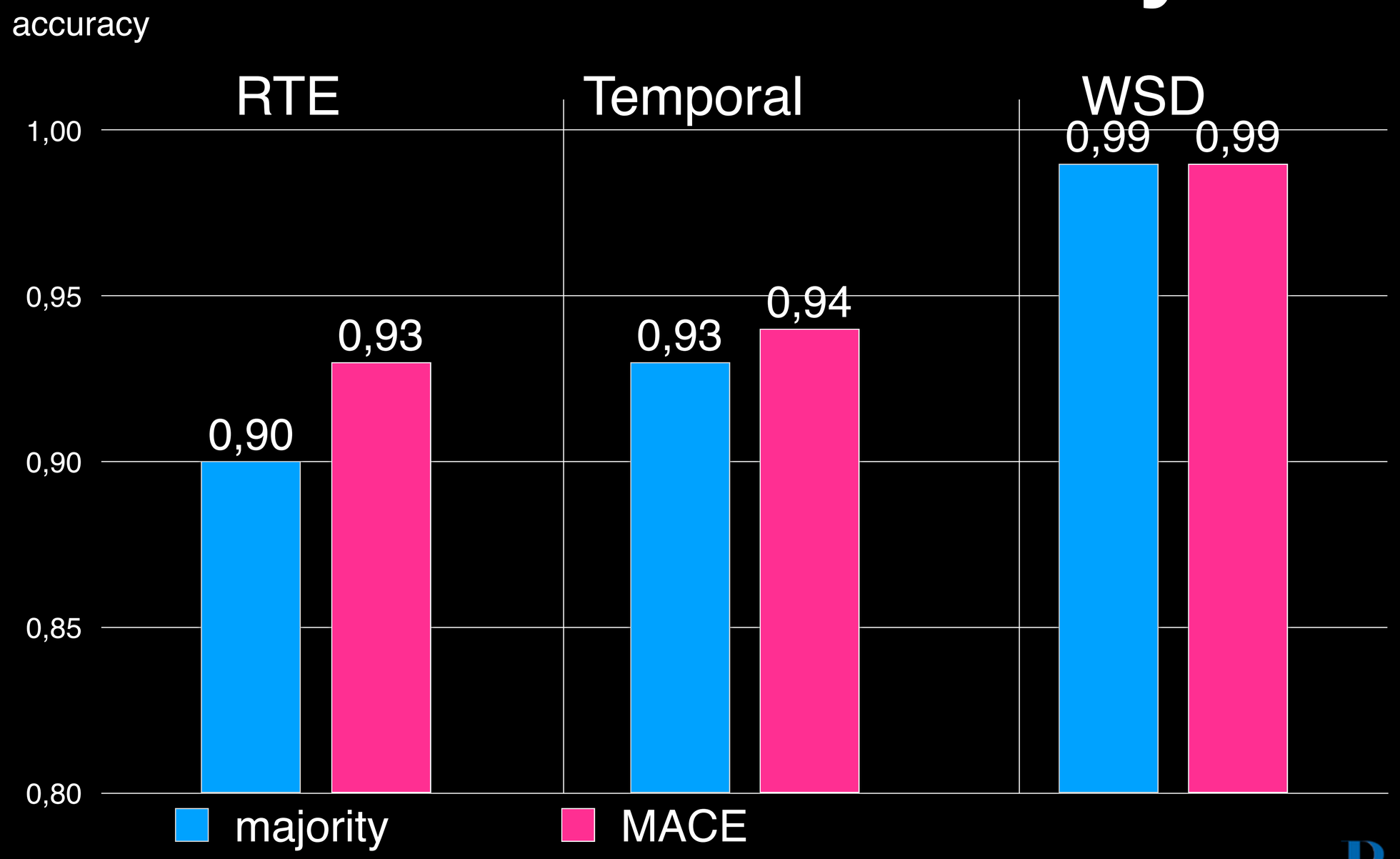
39 Hovy et al. (2013)

Correlation with Proficiency



Hovy et al. (2013)

Prediction Accuracy



Hovy et al. (2013)

More Annotator Bias…

It's a particle!

No! It's an adposition!

WHAT IF YOU'RE BOTH RIGHT?

PRON	VERB	**PRT**	NOUN NUM
PRON	VERB	**ADP**	NOUN NUM

it comes out apr 30

Easy and Hard Categories

	NOUN	PRON	VERB	DET	NUM
NOUN	800	26	0	0	5
PRON	26	155	0	18	0
VERB	0	0	650	0	0
DET	0	18	0	348	0
NUM	5	0	0	0	187

Plank et al. (2014)

Easy and Hard Categories

	NOUN	PRON	VERB	DET	NUM
NOUN	800	26	0	0	5
PRON	26	155	0	18	0
VERB	0	0	650	0	0
DET	0	18	0	348	0
NUM	5	0	0	0	187

EASILY CONFUSED BY ANNOTATORS

Plank et al. (2014)

Easy and Hard Categories

	NOUN	PRON	VERB	DET	NUM
NOUN	800	26	0	0	5
PRON	26	155	0	18	0
VERB	0	0	650	0	0
DET	0	18	0	348	0
NUM	5	0	0	0	187

NOUN != VERB

Plank et al. (2014)

Idea!

MODEL LABEL UNCERTAINTY!

Plank et al. (2014)

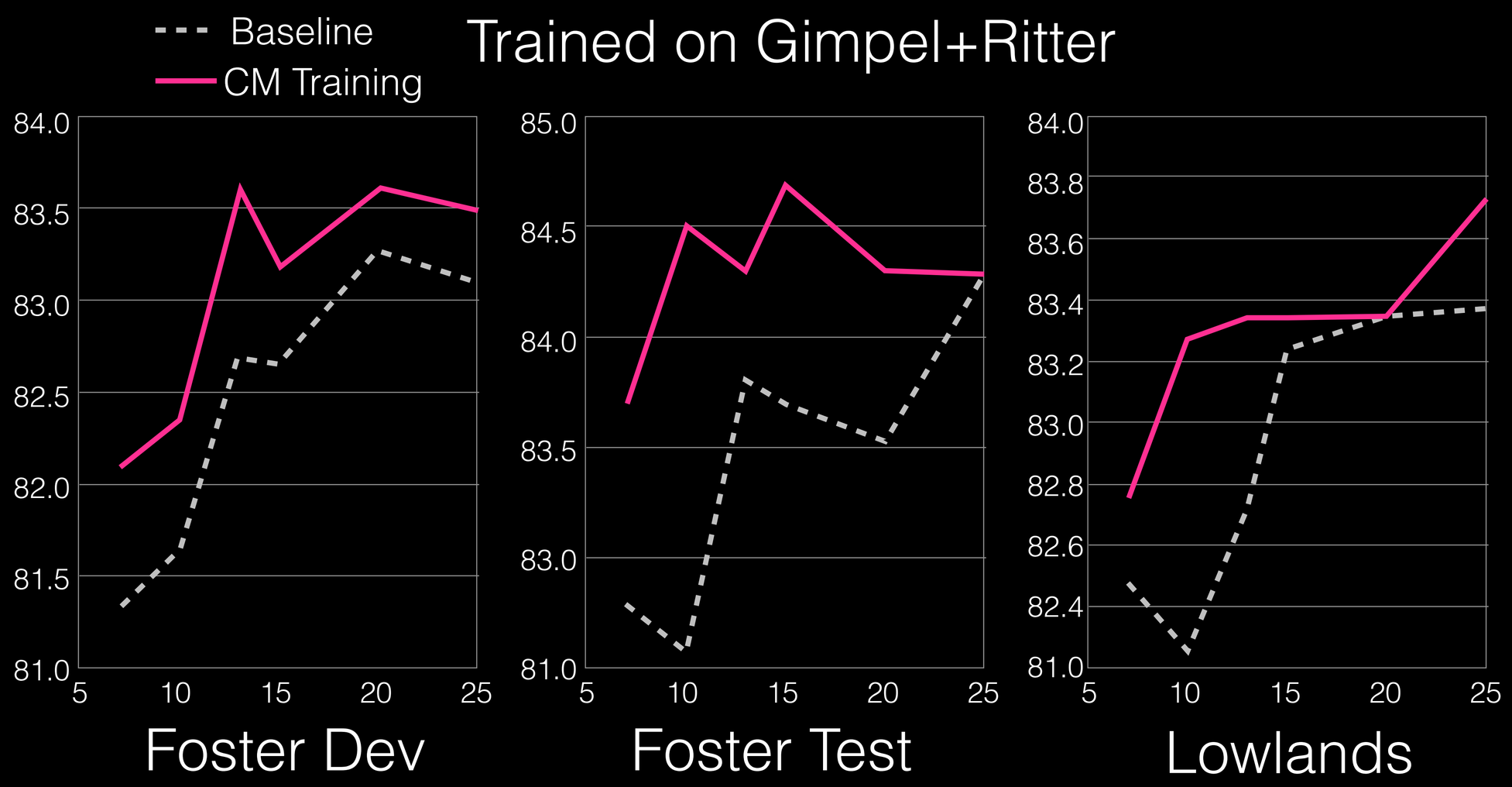
Confusion Matrices in Discriminative Training

```
if model confuses
        easy cases (e.g., NOUN-VERB):
                normal update

        difficult cases (e.g., NOUN-NUM):
                update * CM
```

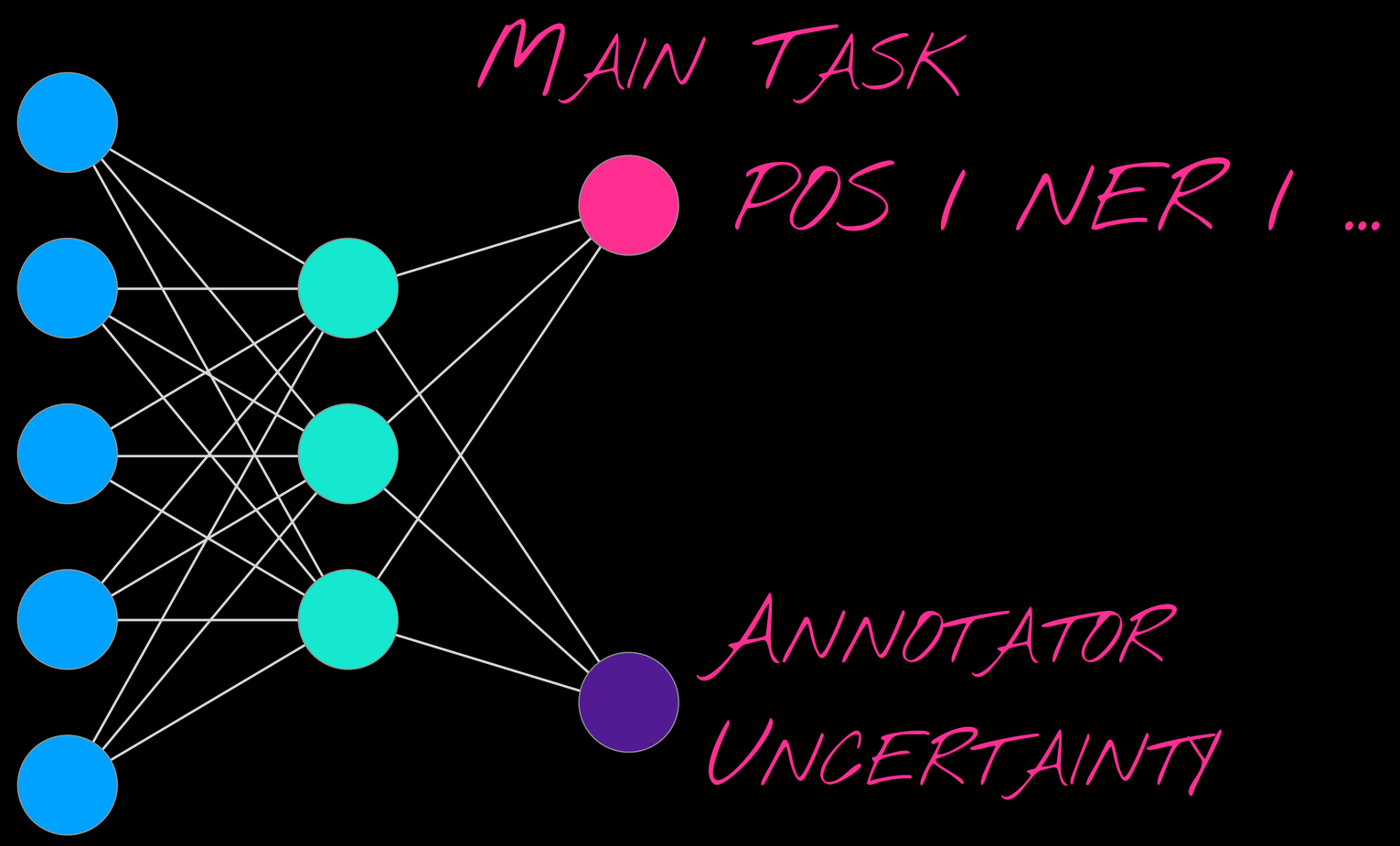
Plank et al. (2014)

POS Results



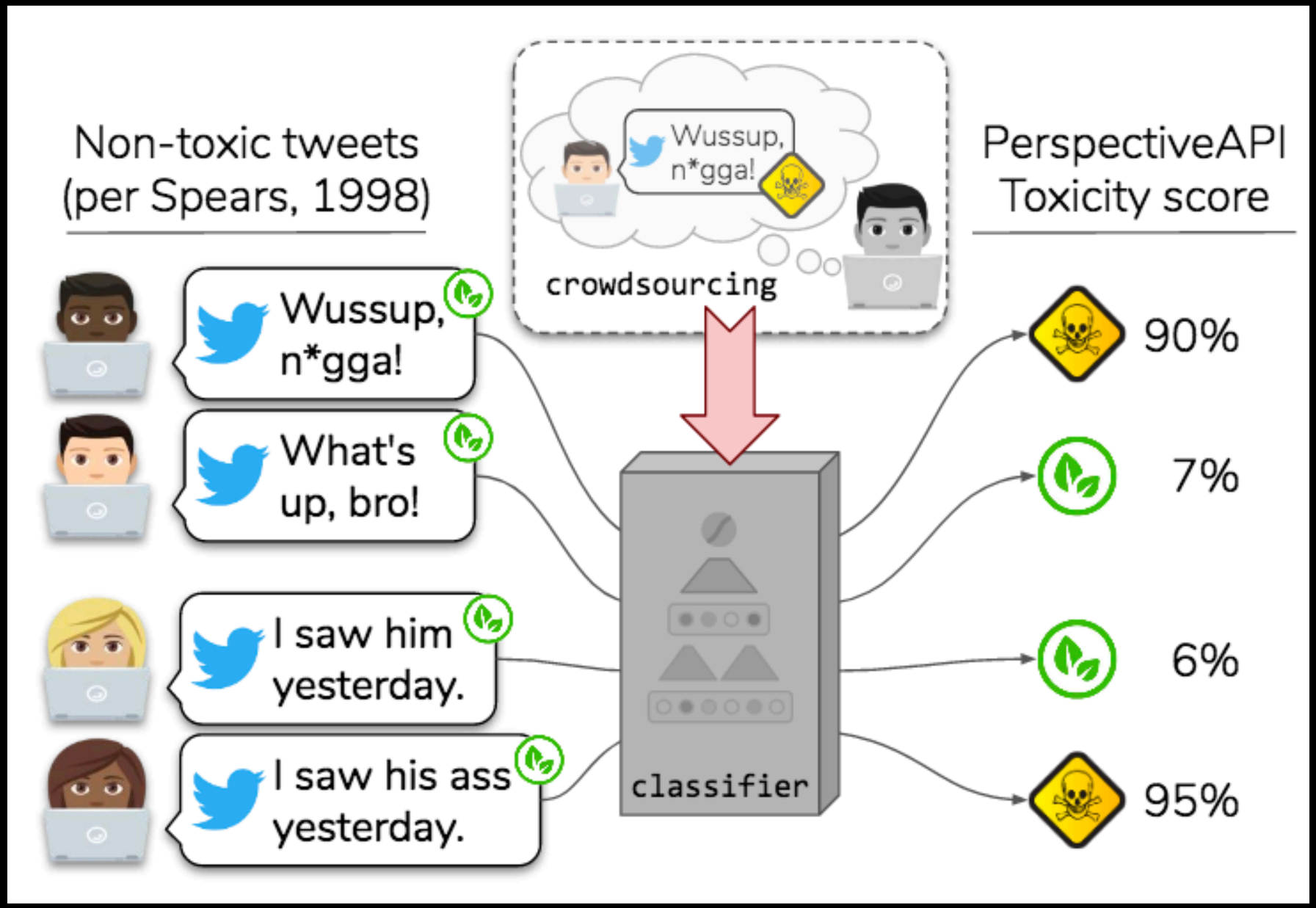
Plank et al. (2014)

Multitask Model



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

Even more Annotator Bias…



50 Sap et al. (2019)

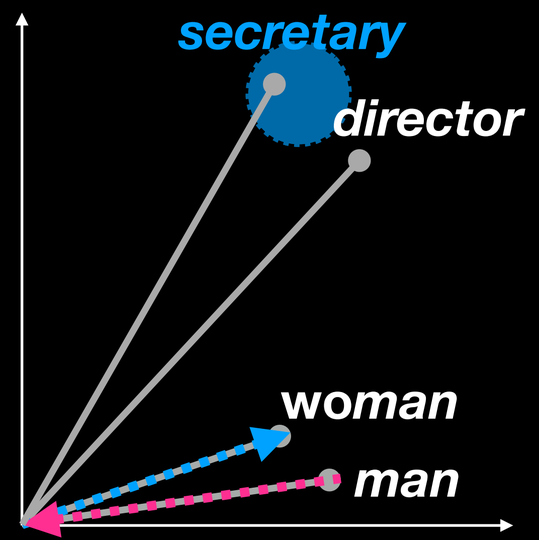
Basically…

TRAIN YOUR ANNOTATORS!

RELEASE ALL ANNOTATIONS!

Pavlick et al. (2014)

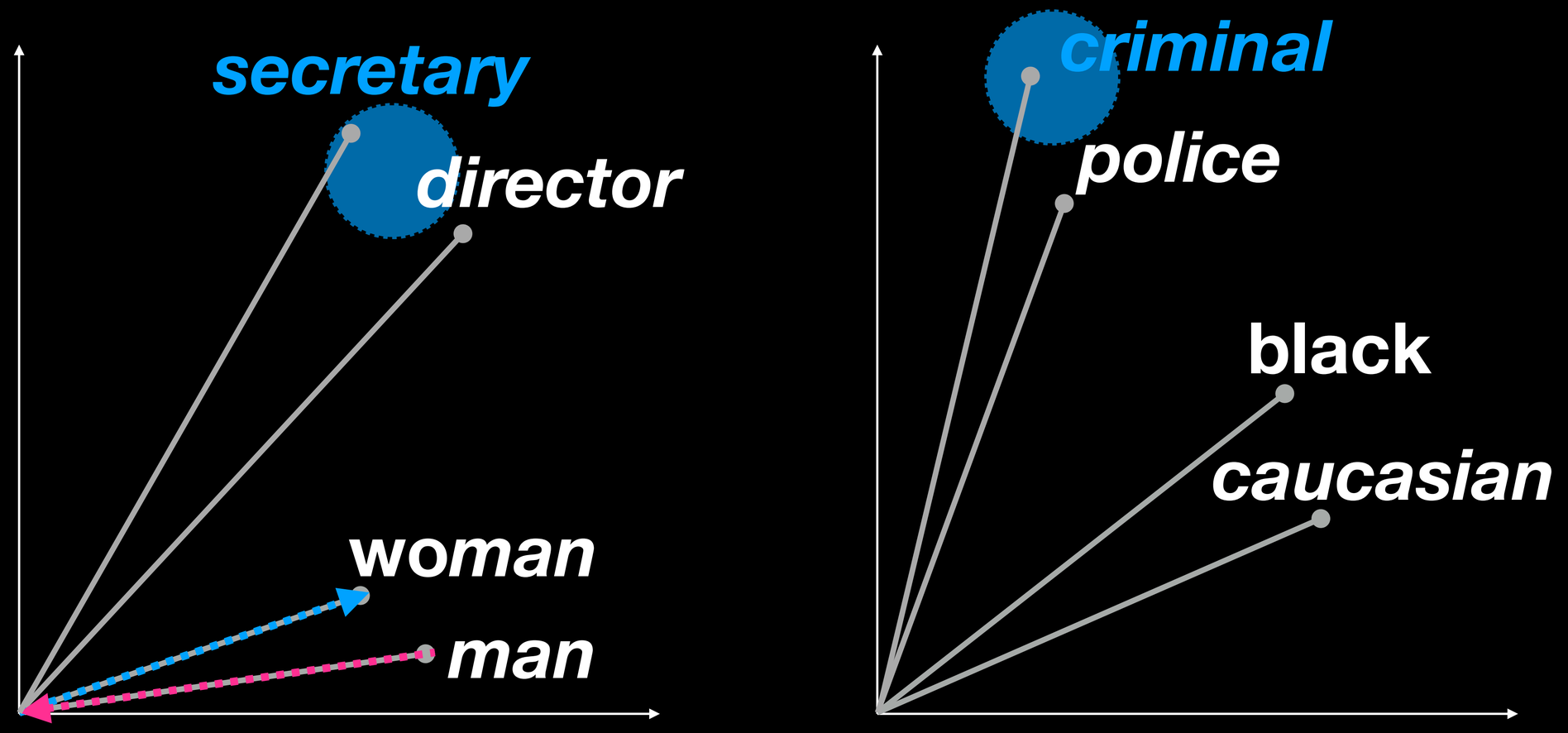
Part 3:
Semantic Bias



Biased Vectors

director – man + woman ≈ secretary

police – caucasian + black ≈ criminal

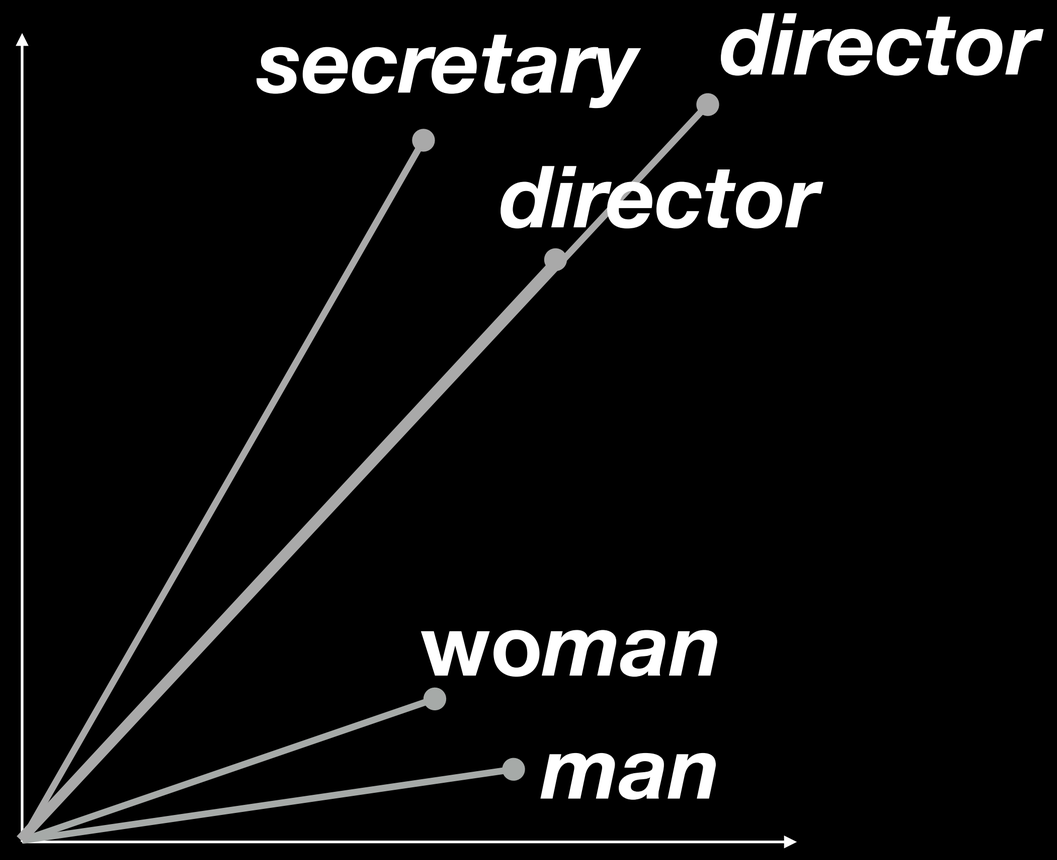


 Bolukbasi et al. (2016); Manzini et al. (2019); Nissim et al. (2019)

Idea!

DEBIAS THE VECTORS!

Debiasing Vectors



Cause vs. Symptoms

Gonen & Goldberg (2019)

Distributed Bias

La **donna** è una gran *tr**a*	*The woman is a huge b***h*
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*
O **menino**´e feito para ser *gay*	*The boy is made to be gay*
Unchiul este un mare *nenorocit*	*The uncle is a big jerk*
La **mujer** es la mas *bonita*	*The wife is the most attractive*

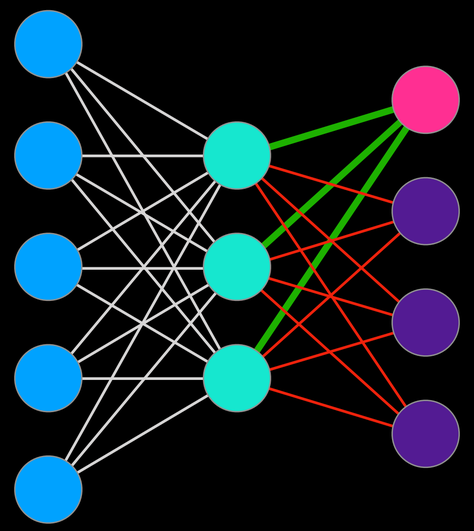
4.5% of all completions contain a hurtful word.

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

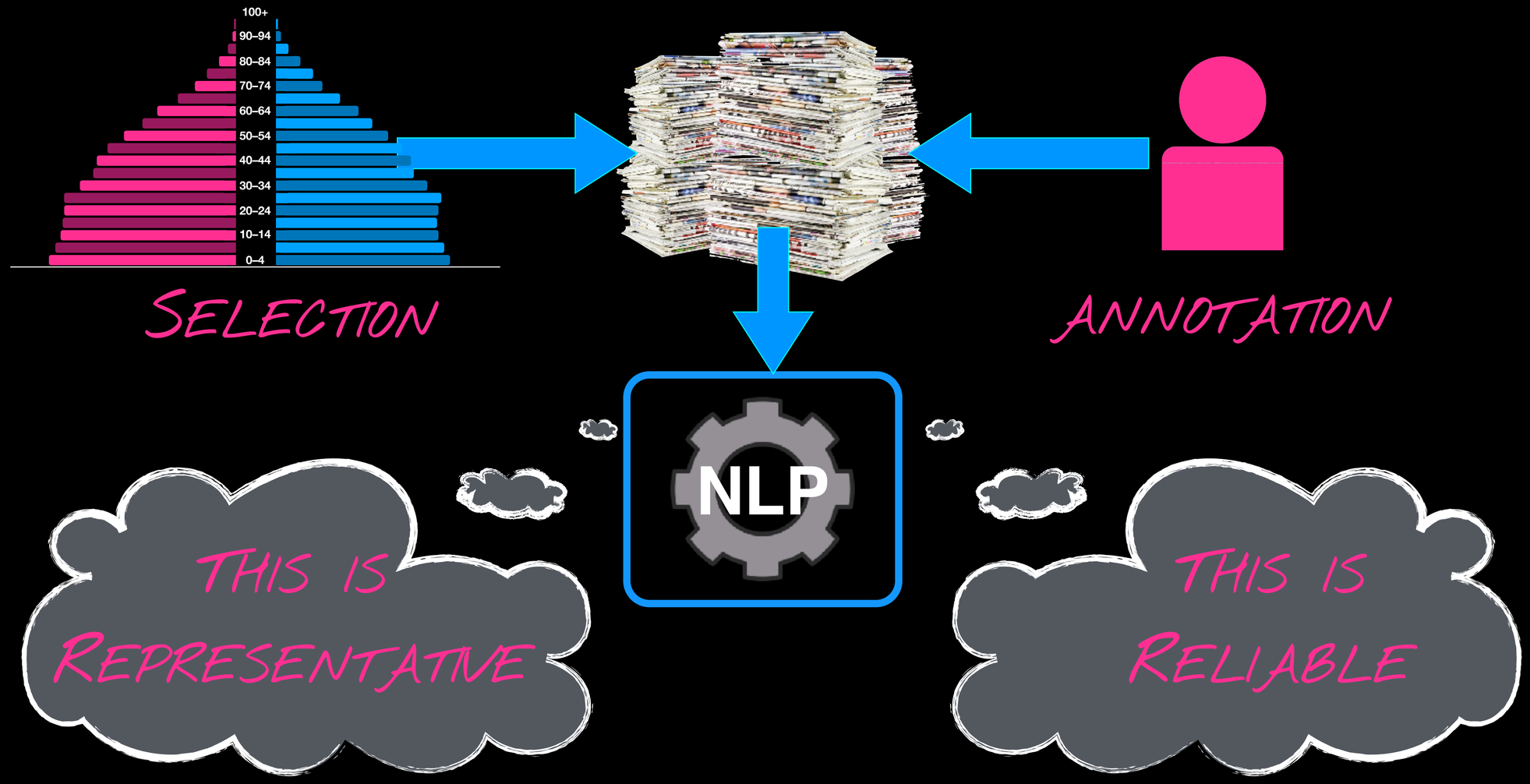
if target is male, **4%** refer to homosexuality

 Bassignana et al. 2018; Sheng et al., 2019; Nozza et al., 2021, 2022

Part 4:
Model Bias



Biased Models



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

60 Zhao et al. (EMNLP 2017)

Biased Sentiment Analysis

0.64

He made me feel ***afraid***

0.52

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

0.48

She made me feel ***afraid***

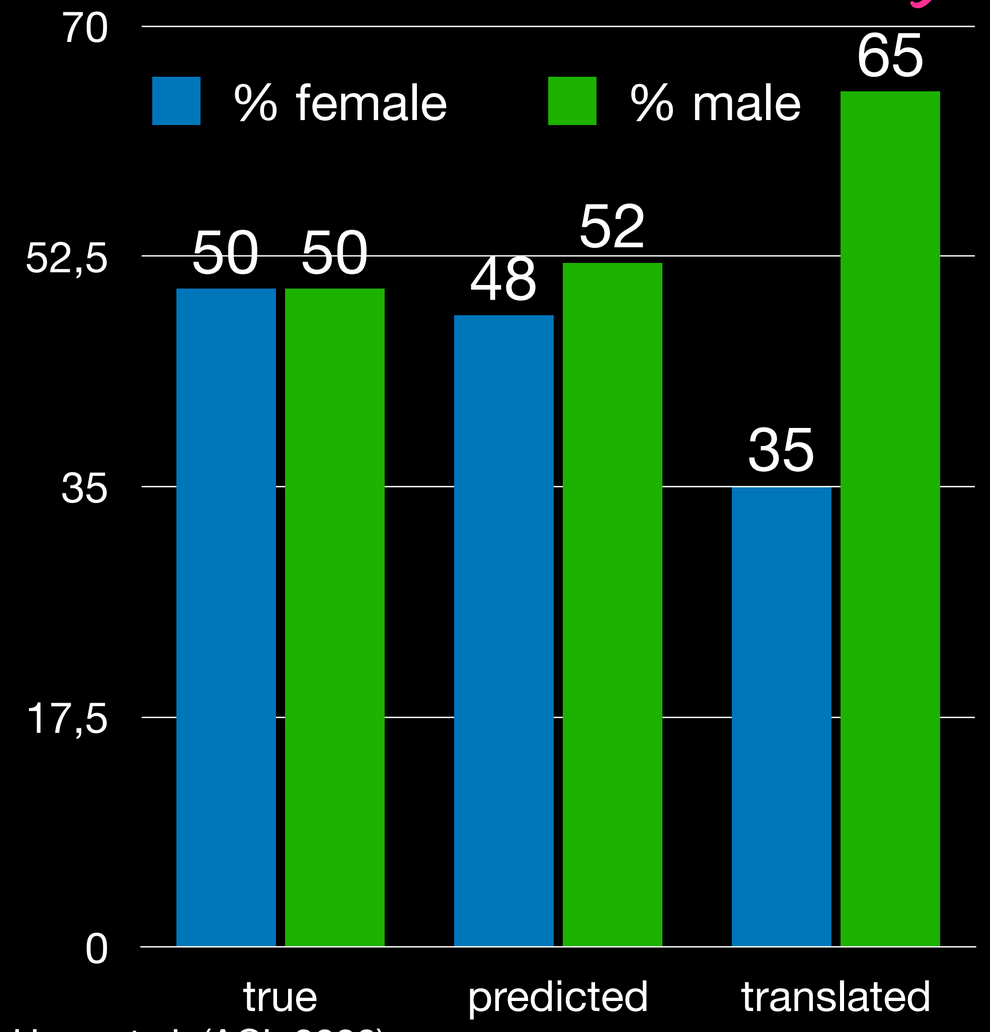
0.43

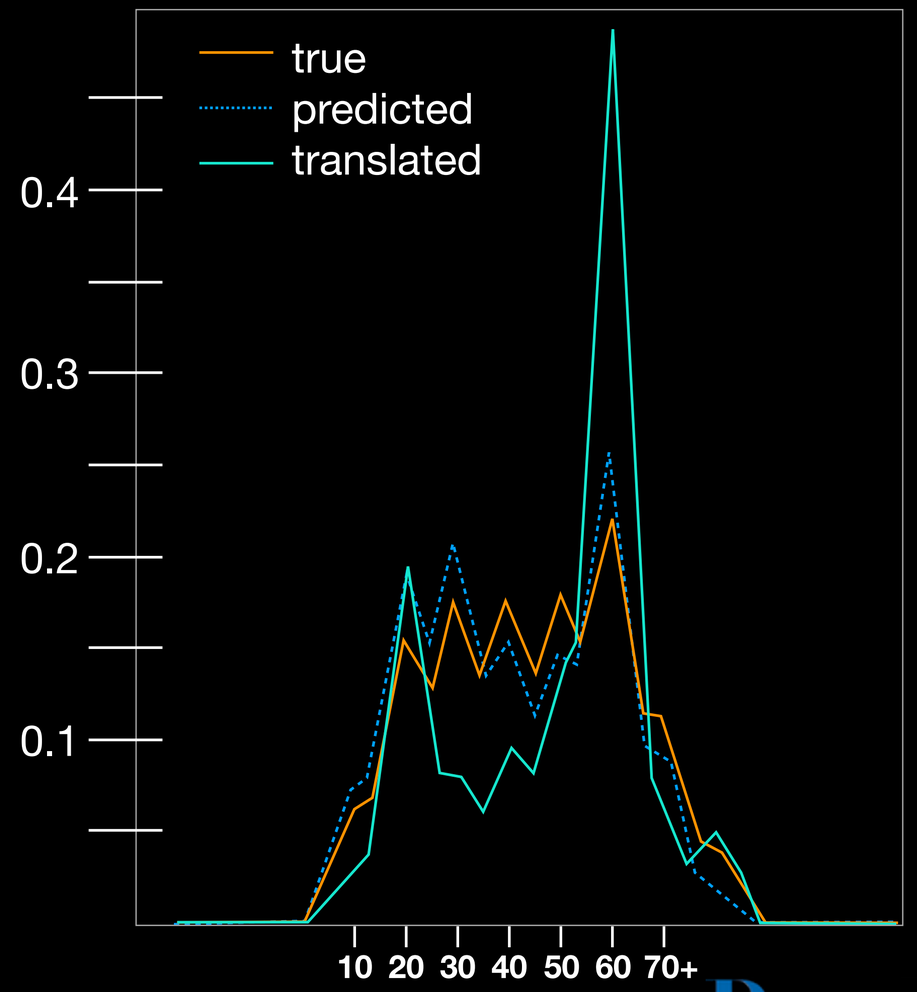
I made ***Heather*** feel ***angry***

Kiritchenko & Mohammad (2018)

Machine Translation Bias

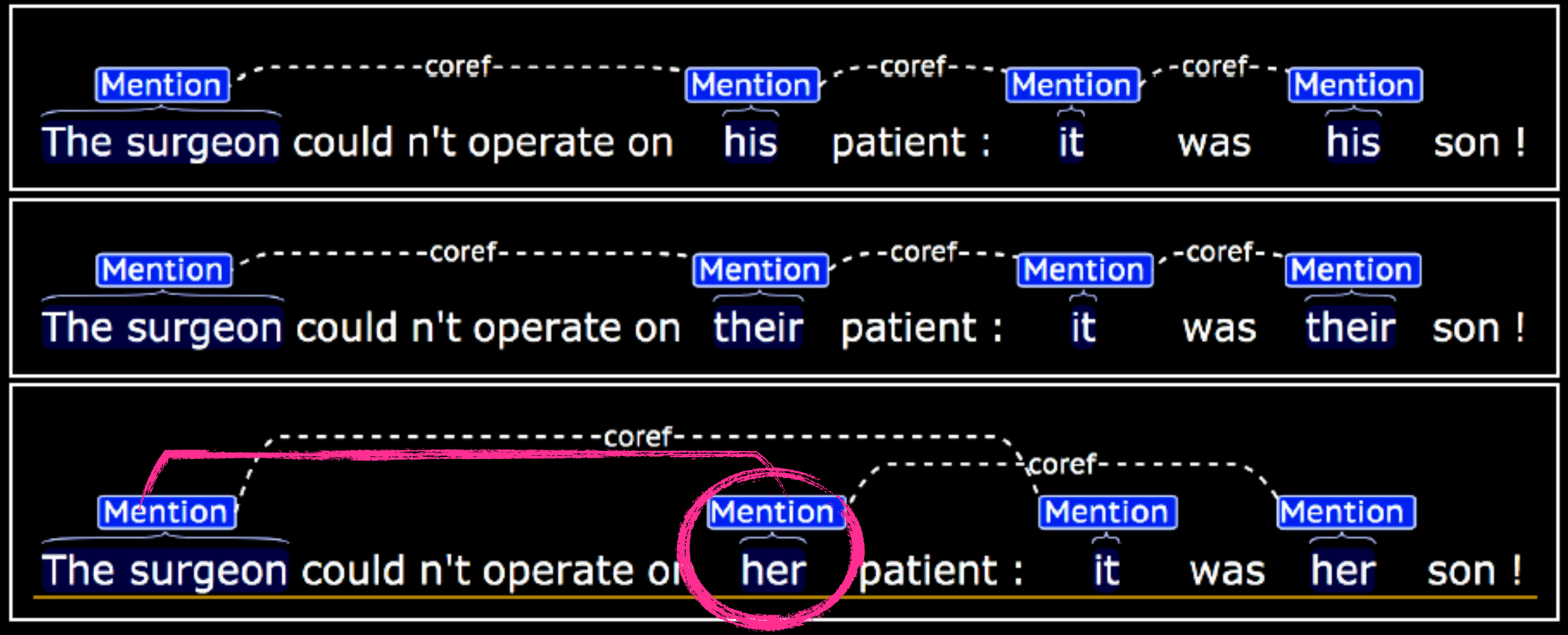
MACHINE TRANSLATION MAKES YOU SOUND OLDER AND MORE MALE.





Hovy et al. (ACL 2020)

Wrong Coreference



63 Rudinger et al. (2018)

Overgeneralization

False Positives

Aug 6 2016
Dear Ms Hovy,

Congratulations on reaching retirement age!

Also, you're on a no-fly list because of your political views and religious beliefs.

Idea!

Discourage models from amplification!

 Zhao et al. (EMNLP 2017)

Reducing Bias

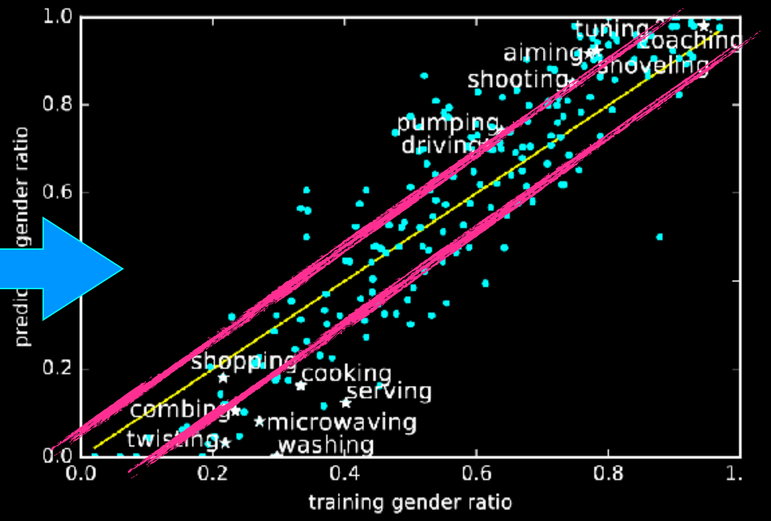
BIAS = 0.66

Agent: WOMAN

Agent: MAN

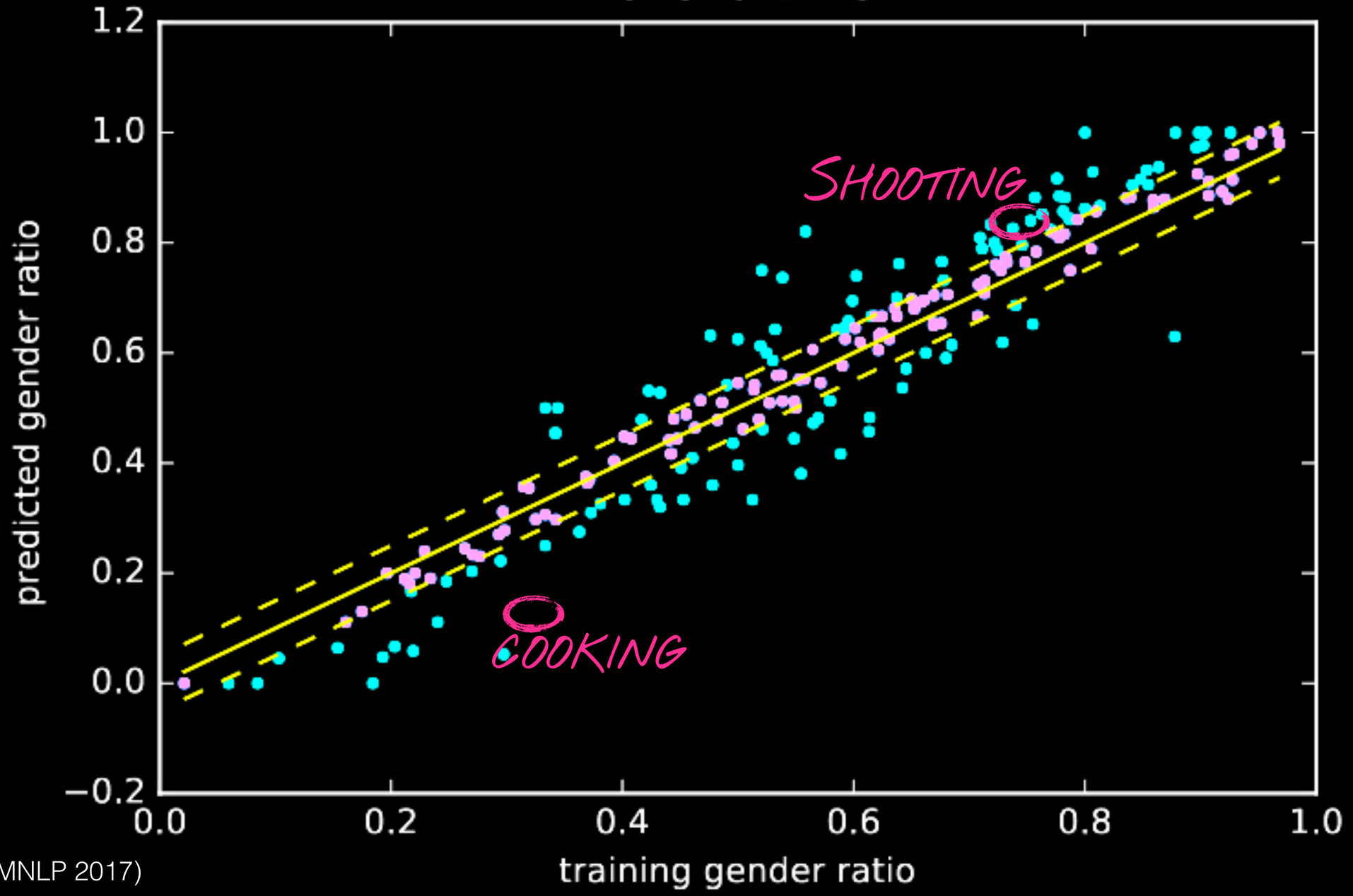
Agent: WOMAN

ILP



66 Zhao et al. (EMNLP 2017)

Results



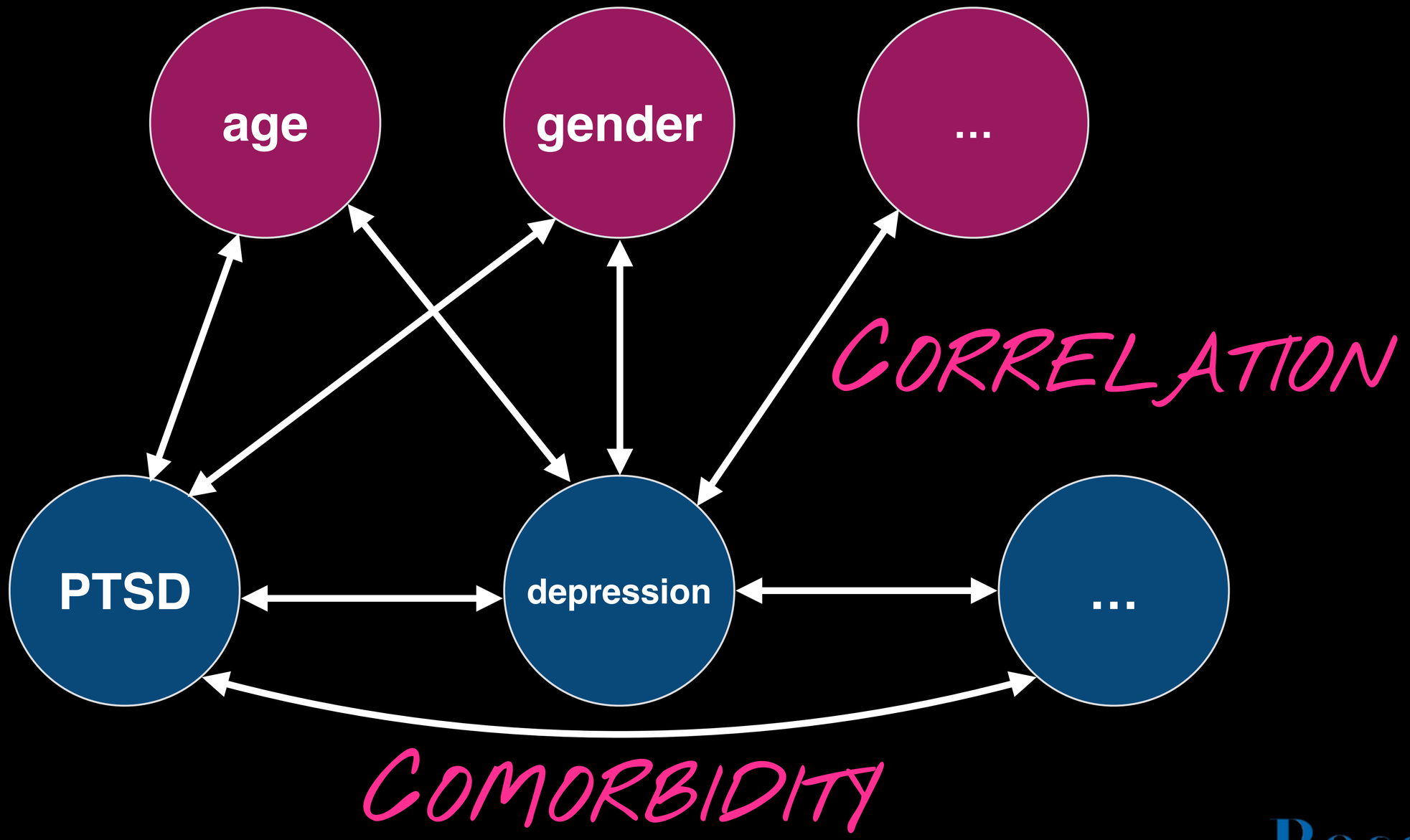
67 Zhao et al. (EMNLP 2017)

Idea!

ADD DEMOGRAPHIC COMPONENT IN MODEL

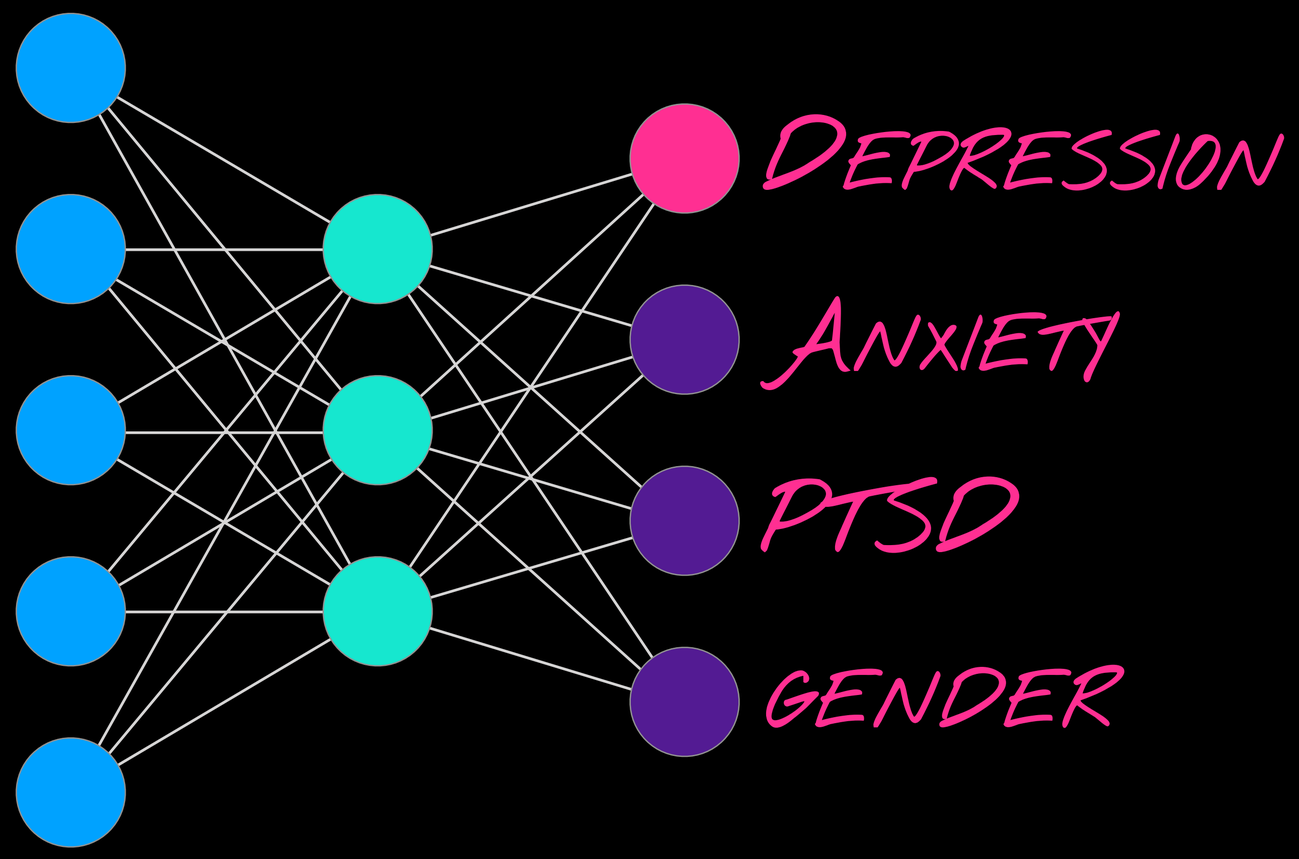
68 Benton et al. (2017)

Comorbidity and Correlation



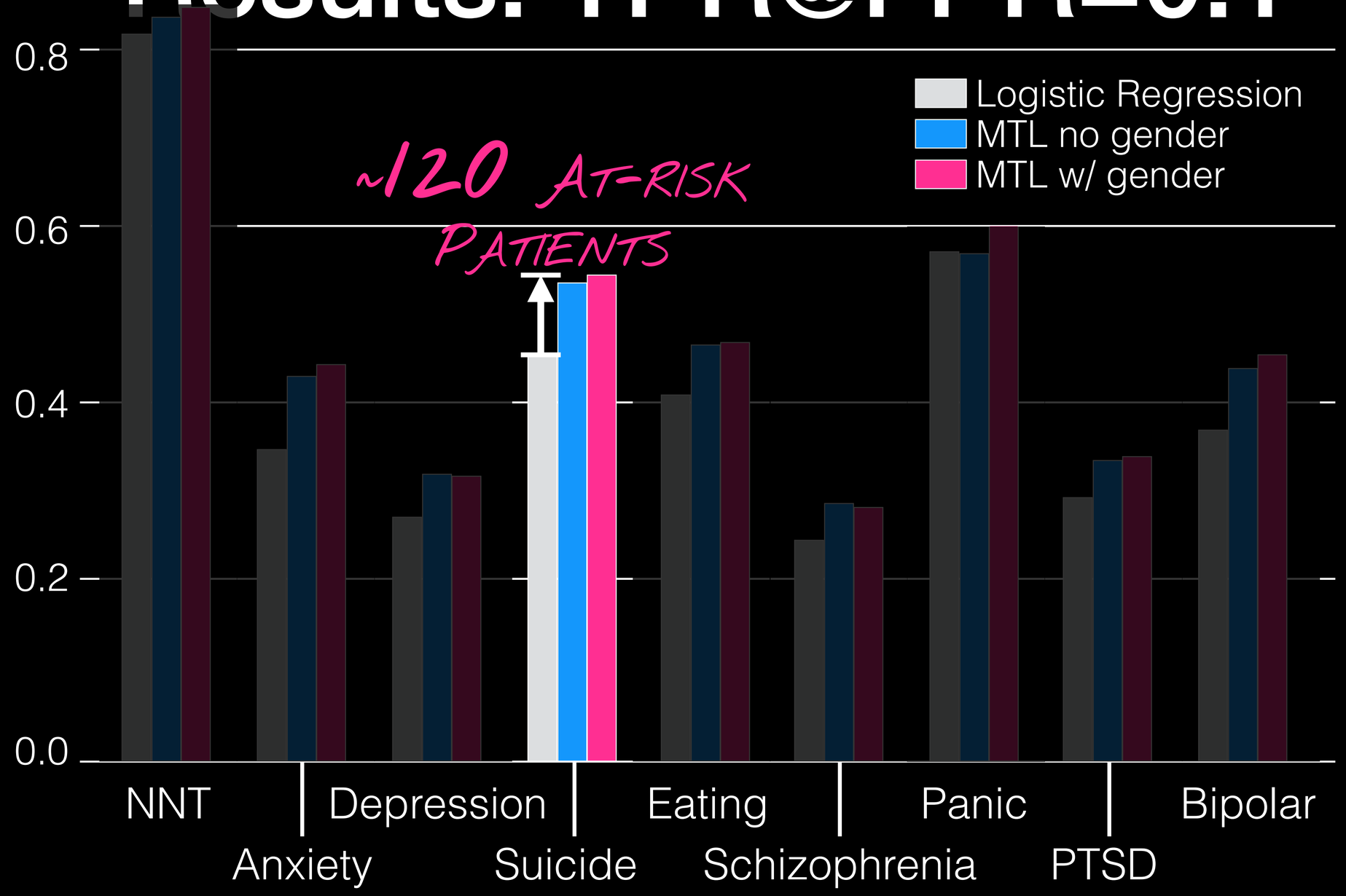
 Benton et al. (2017)

Multitask Model



Benton et al. (2017)

Results: TPR@FPR=0.1



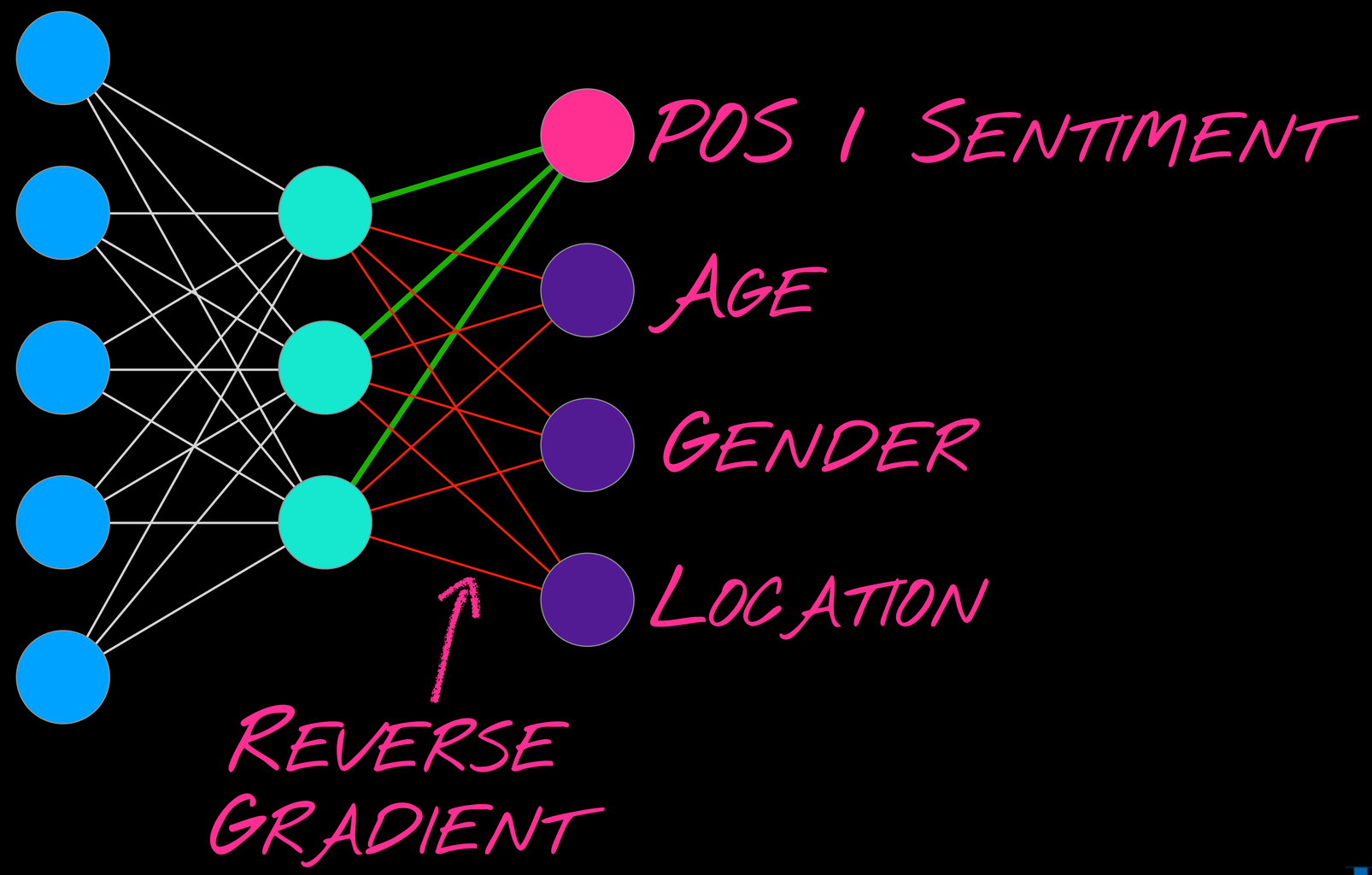
Benton et al. (2017)

Idea!

CORRECT FOR BIAS ADVERSARIALLY

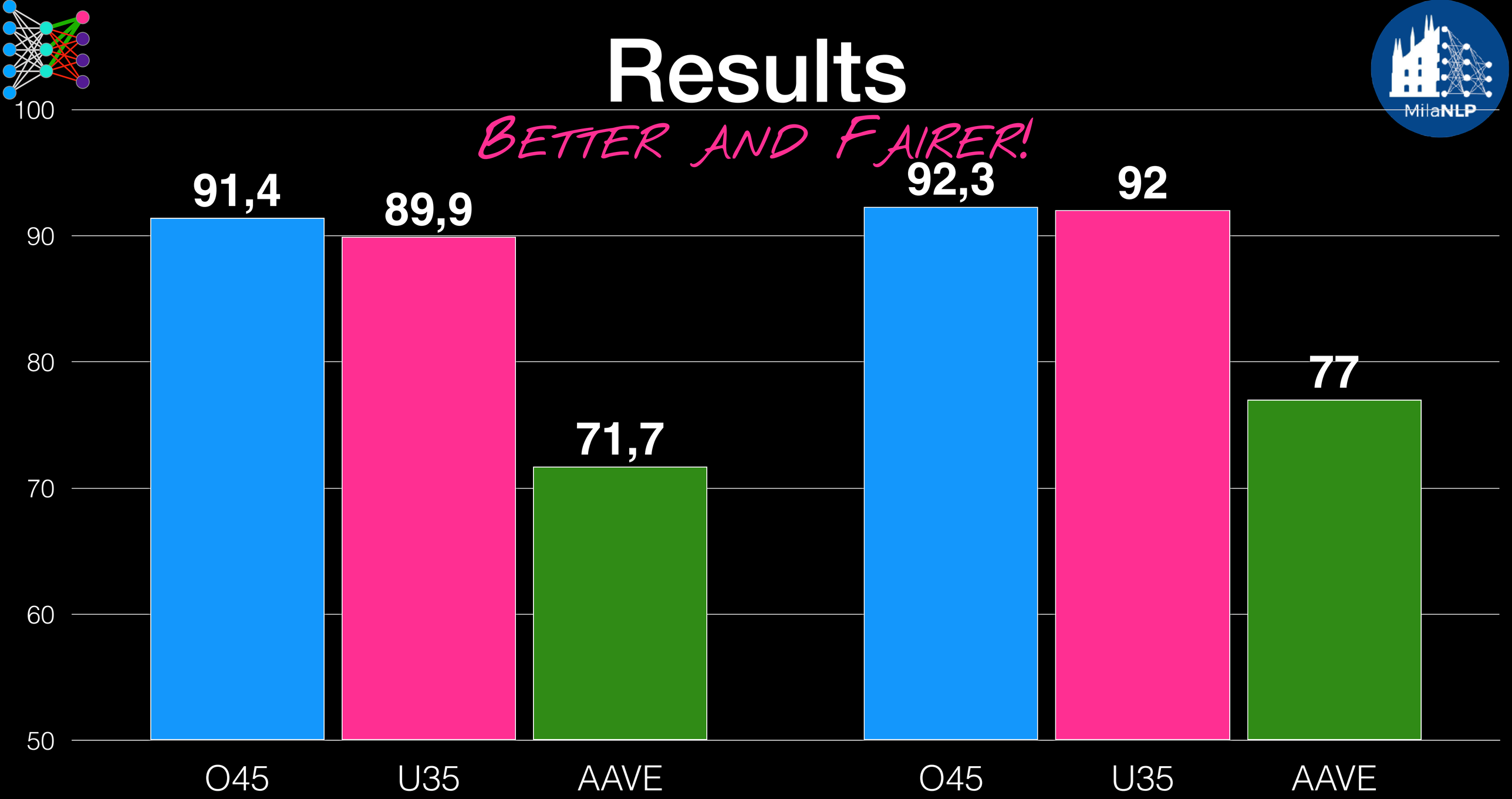
Li et al. (ACL 2018)

Adversarial Model



Li et al. (ACL 2018)

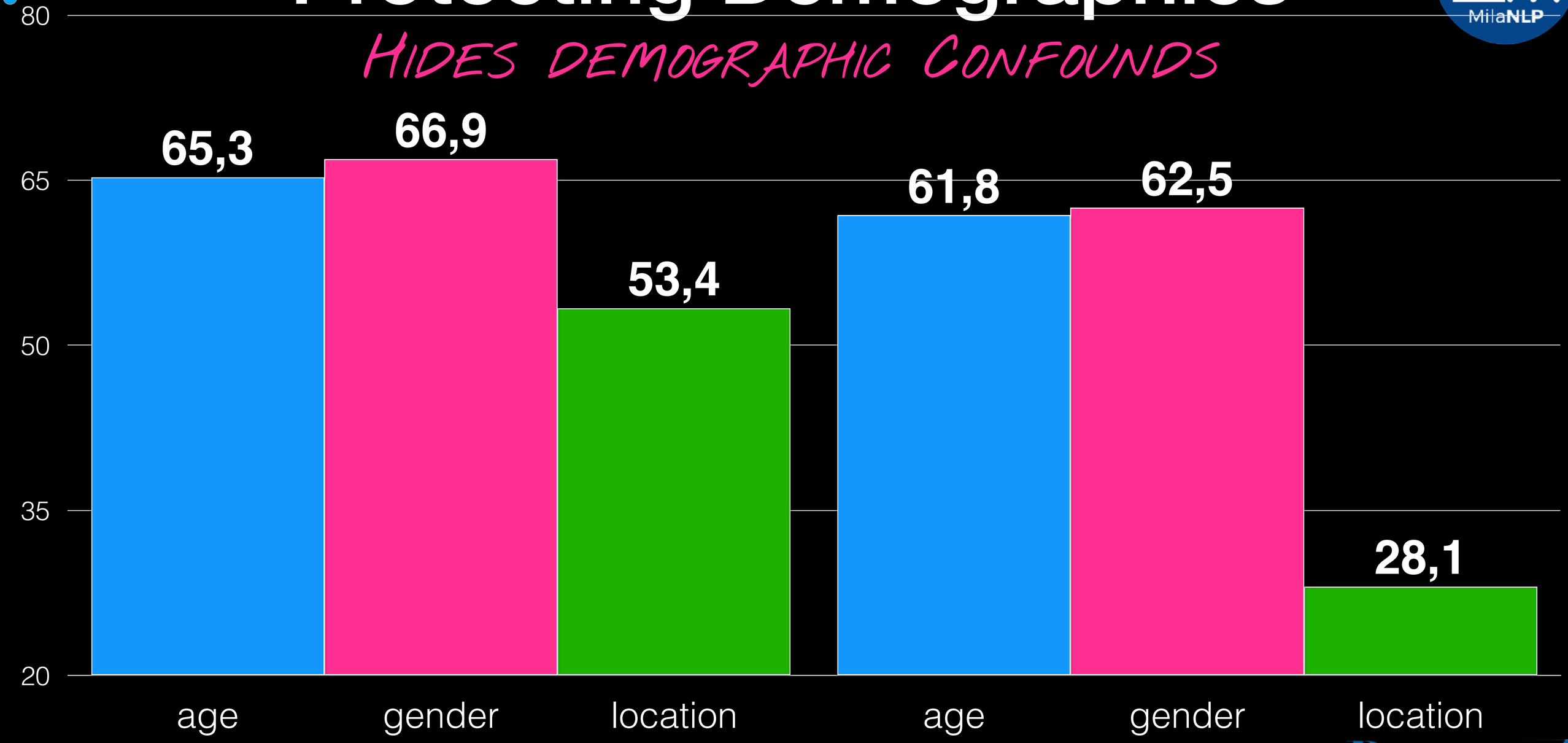
Results



Li et al. (ACL 2018)

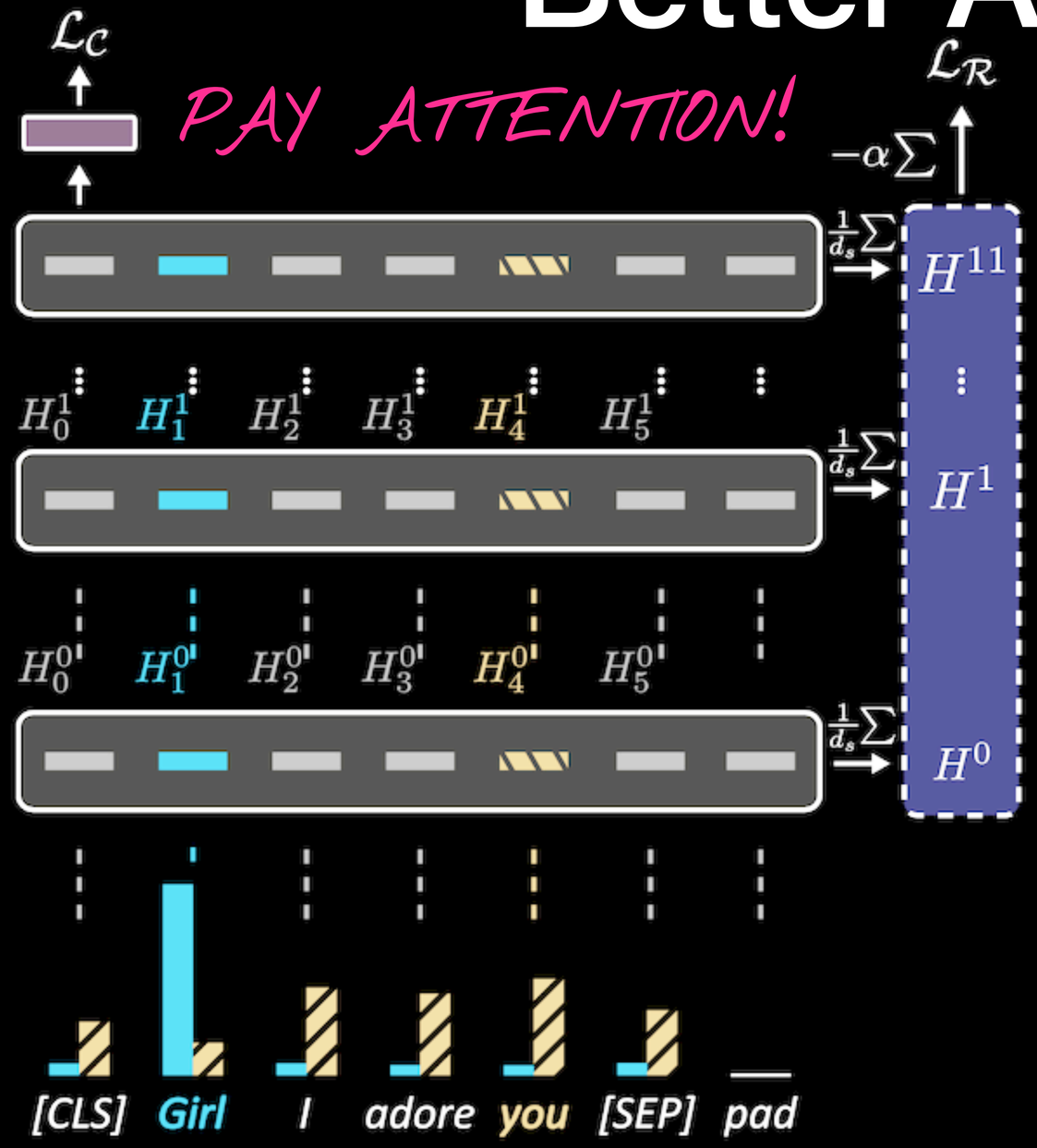
Protecting Demographics

HIDES DEMOGRAPHIC CONFOUNDS



Li et al. (ACL 2018)

Better Attention



INDICATIVE TERMS:

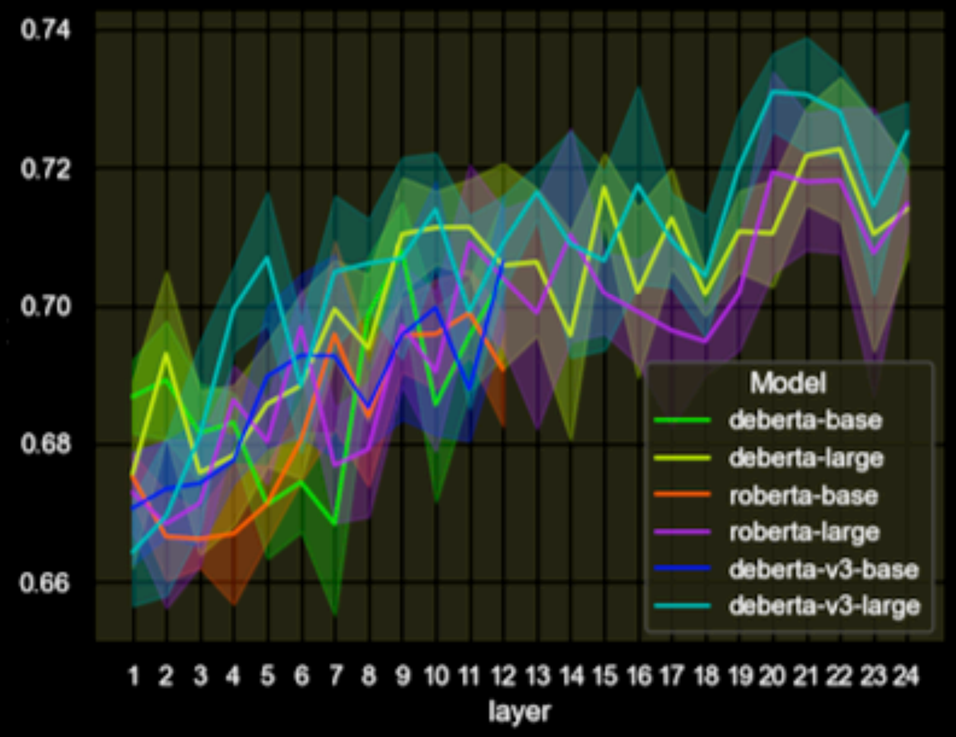
EN: girls, womens*ck, hoes, c*ck, shut, stupid, hoe, p*ssy, trying, f*ck

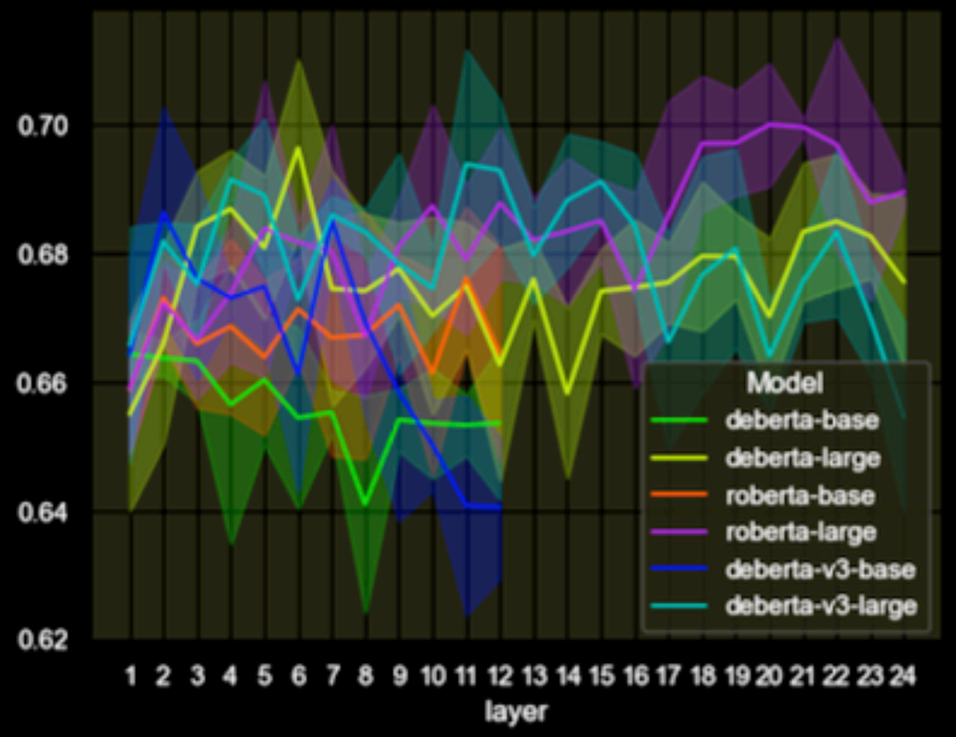
IT: pezzo, bel, bellissima, scoperei, p*ttanona, zitta, sb*rro, t*ttona, bella, c*lone

Attanasio et al. 2022

What to Model?

LMs encode Sociodemographics



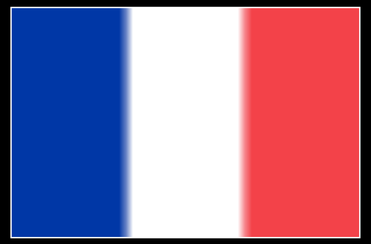


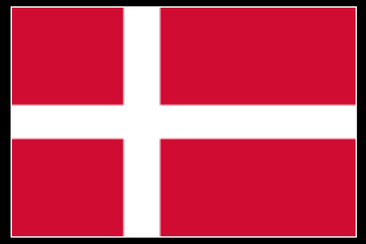
... But don't understand them

 Lauscher et al., 2022

Part 4:
Design Bias

Exposure





Over-Exposure

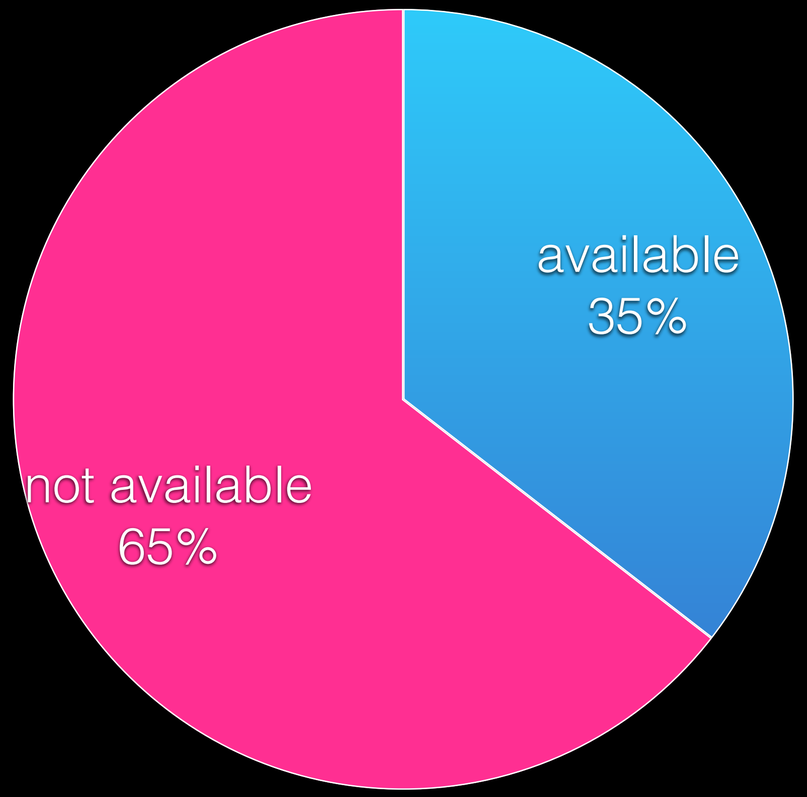
POS tagging

Discourse

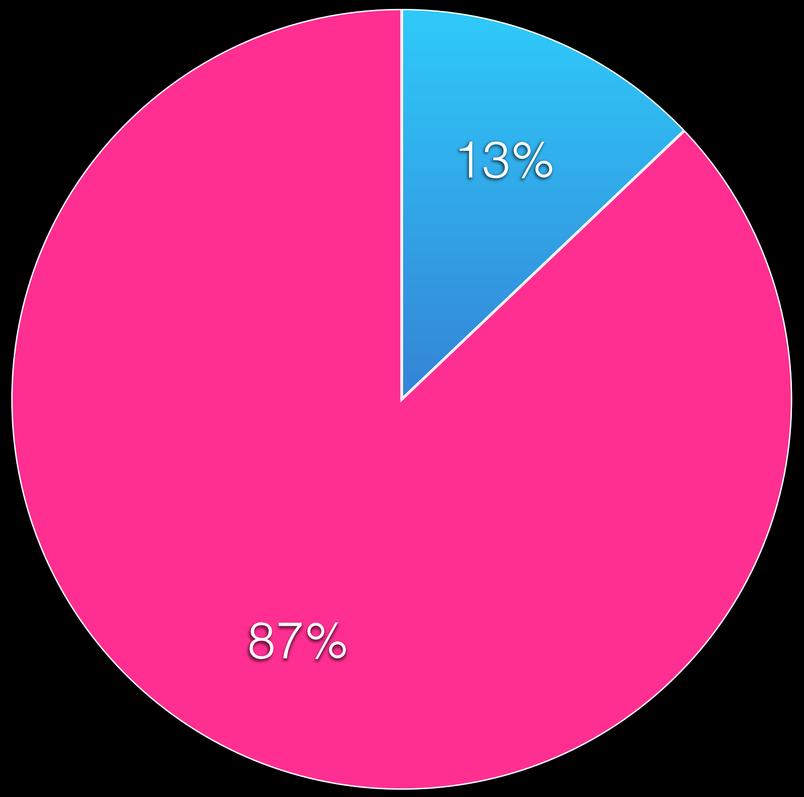
 Hovy & Spruit (2016)

Under-Exposure

treebanks *



semantic resources



*BEFORE UD… evaluation

 Hovy & Spruit (2016)

Idea!

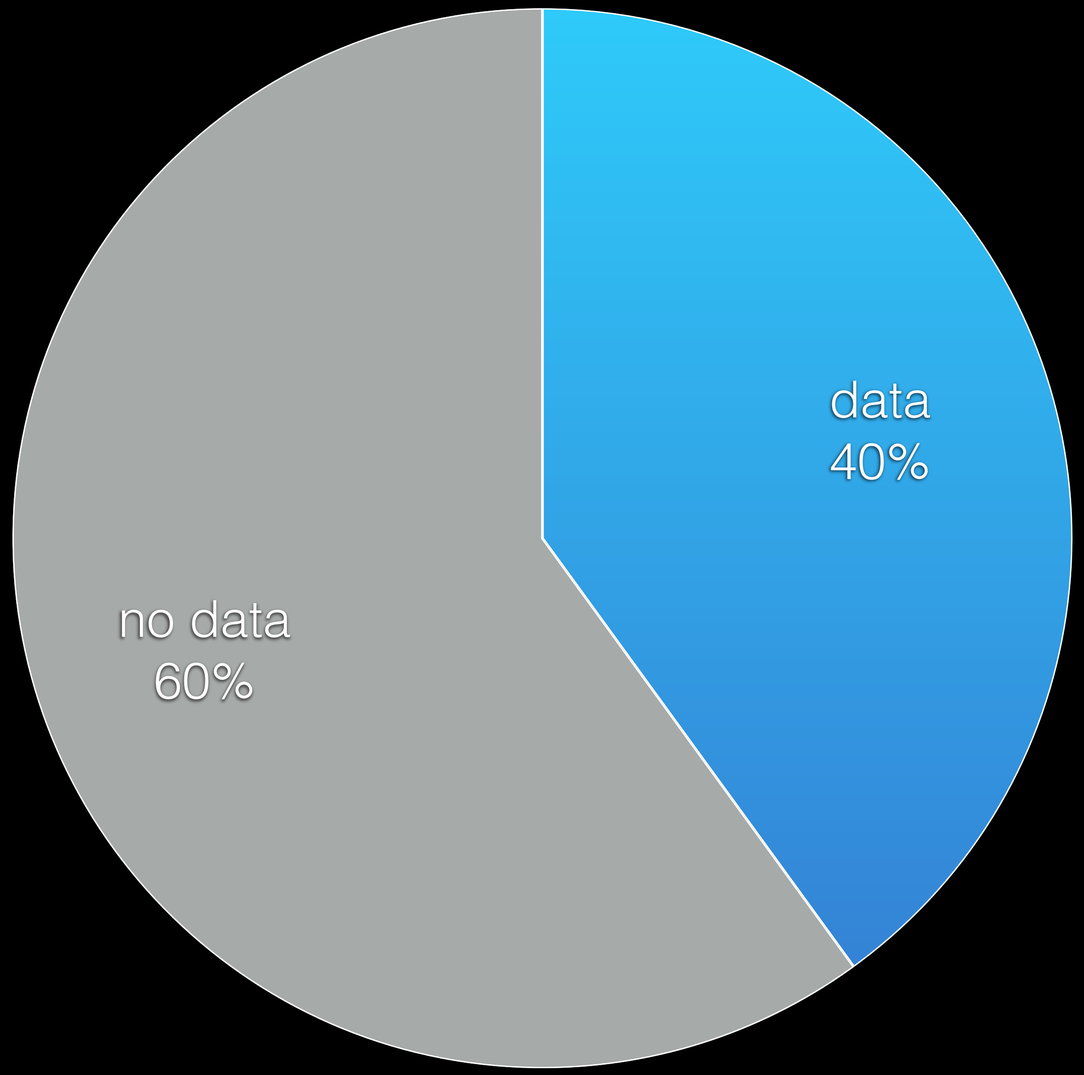
MAKE LANGUAGE EXPLICIT!

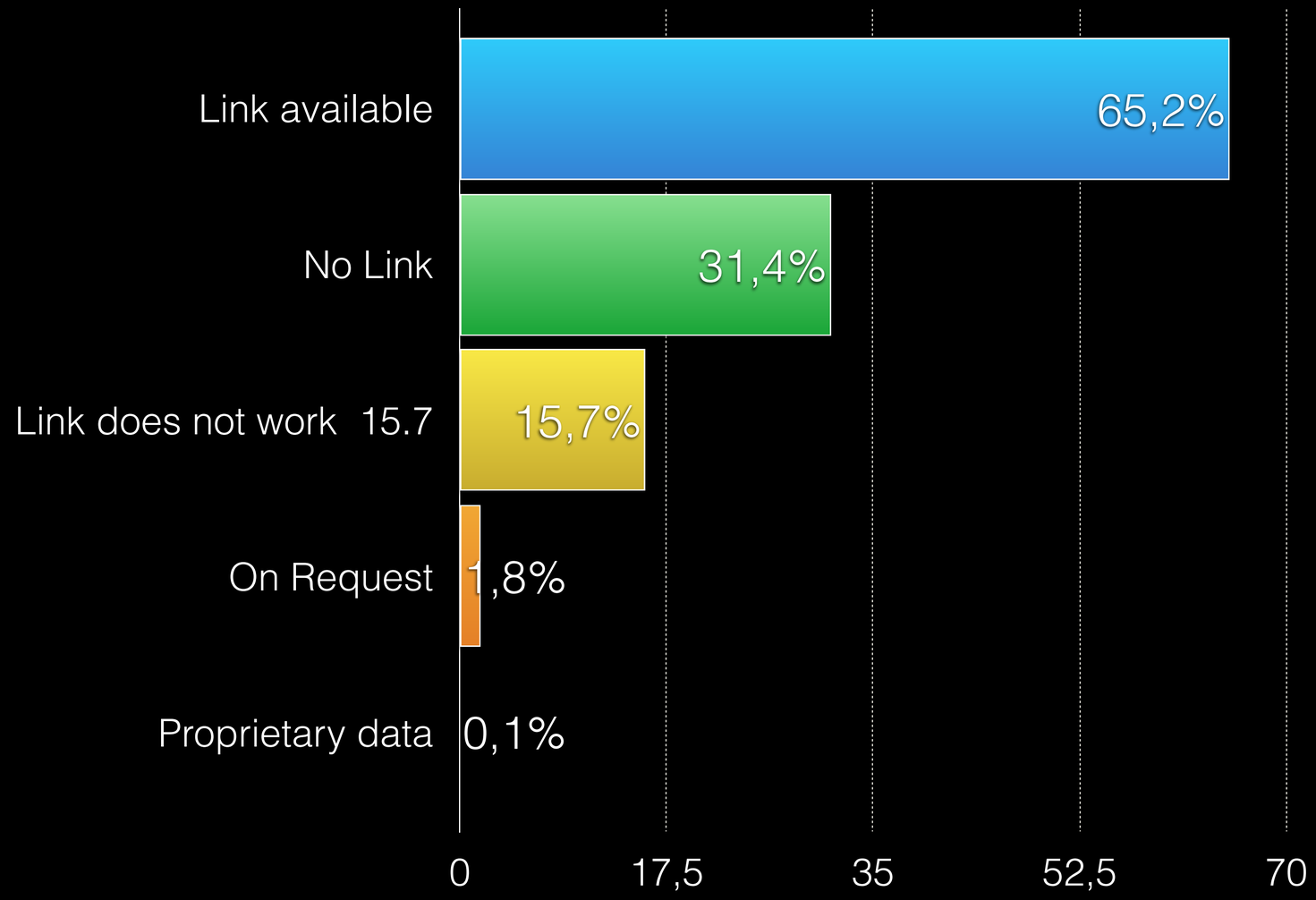
The Bender Rule

"Do state the name of the language that is being studied, even if it's English."

– Emily Bender

Replicability: Data



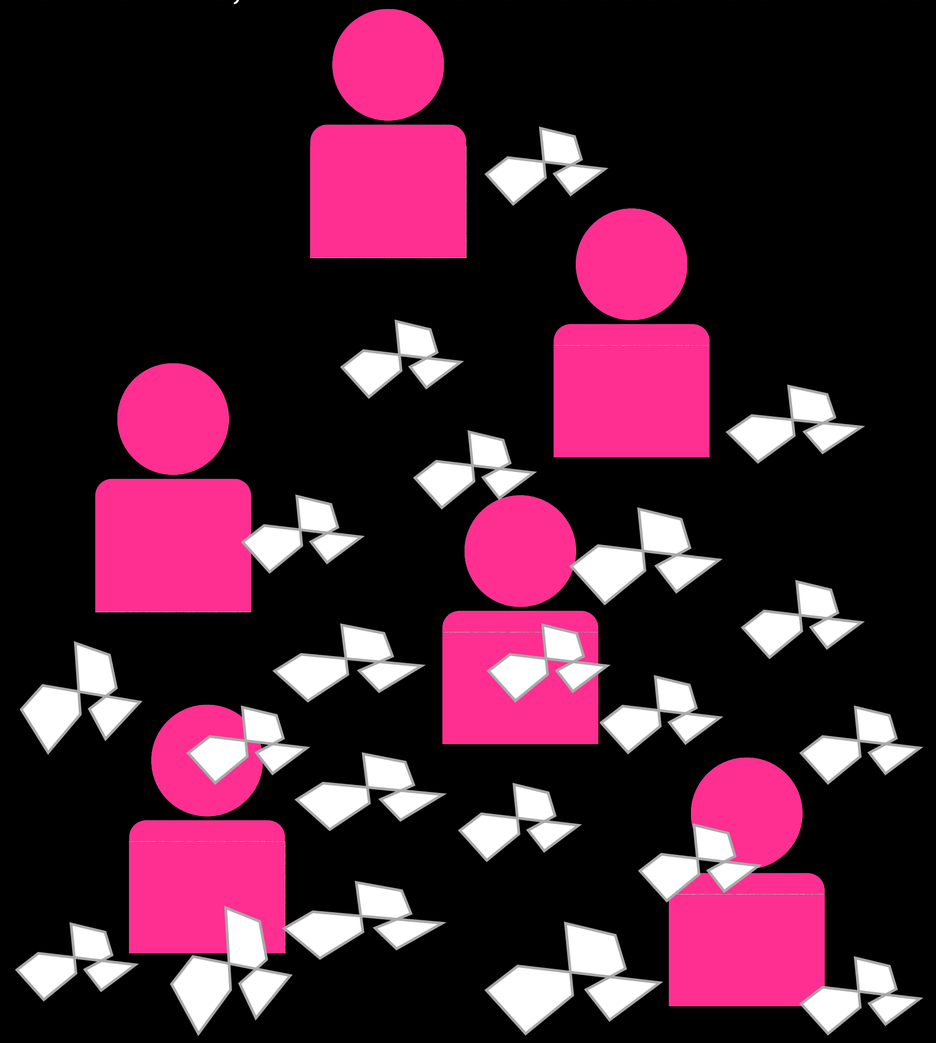


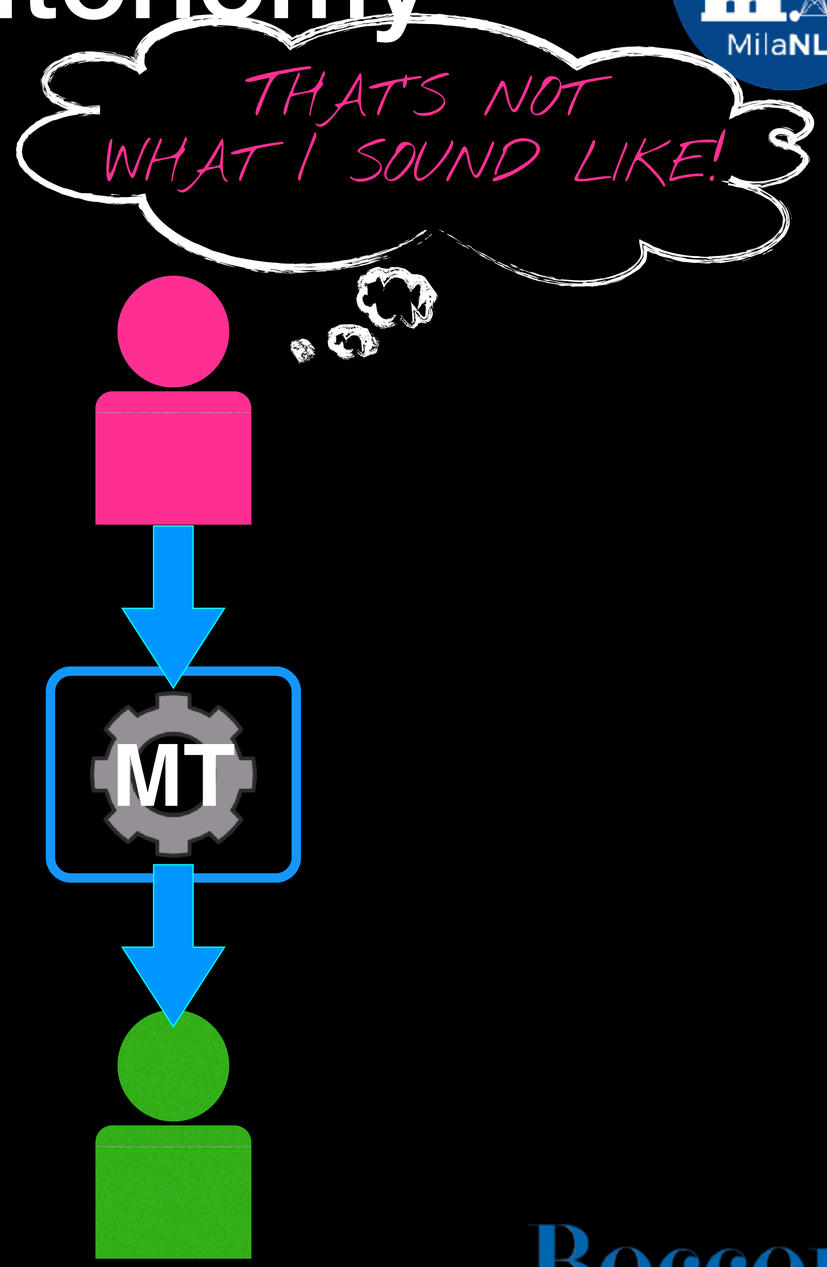
 Mieskes (2017)

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





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

Technology as Social Experiment

How can I make it safe?

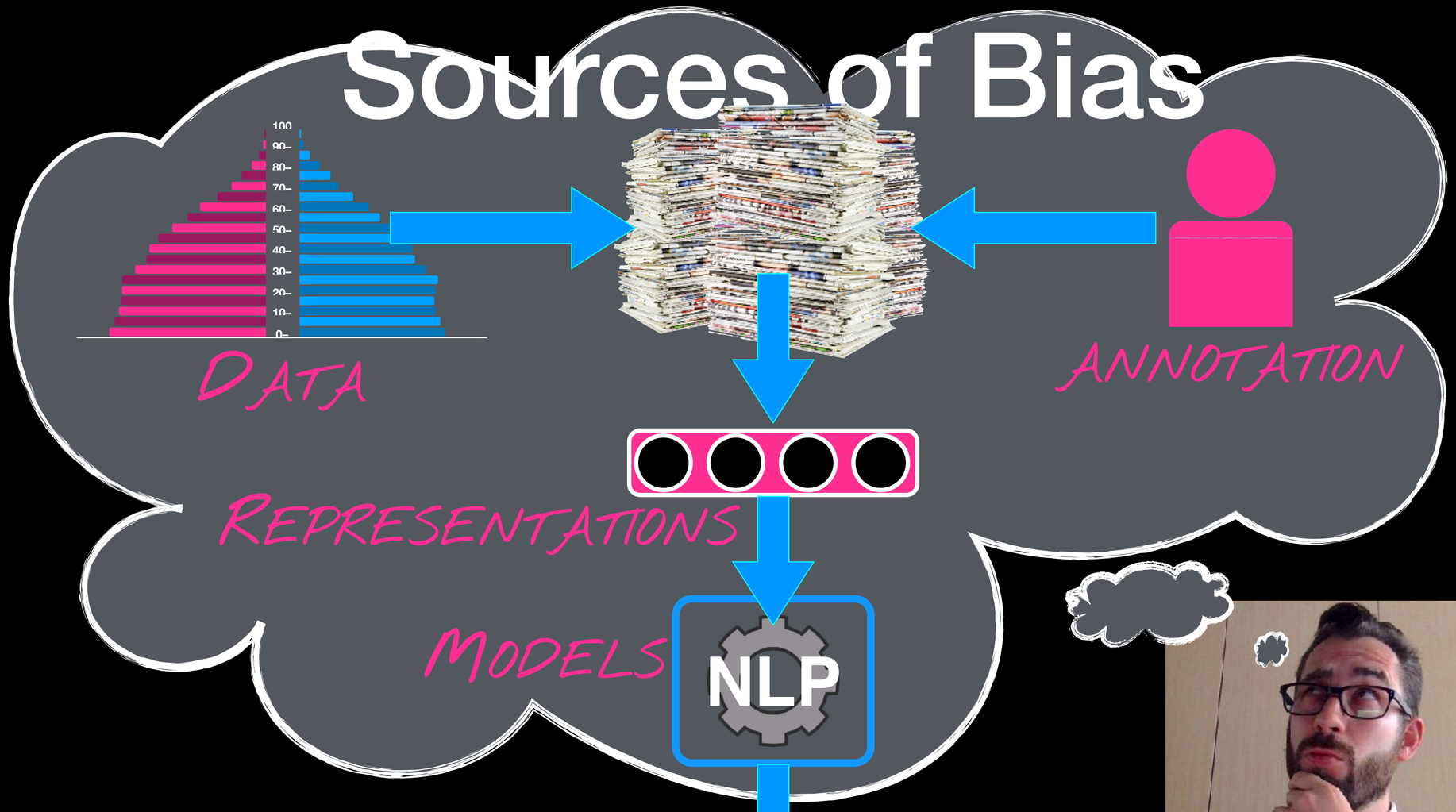
Do my subjects know they're in it?

Can they control it, opt out?

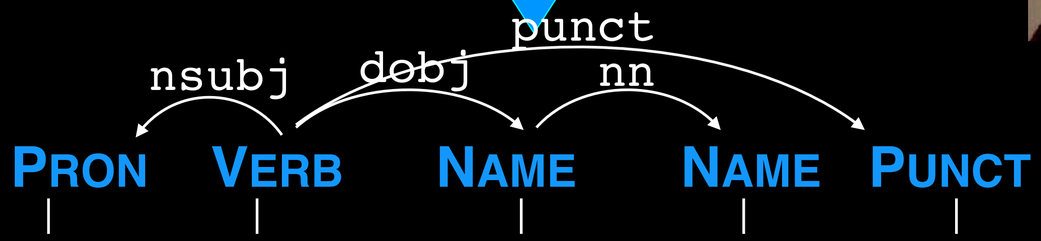
 Van de Poel (2016)

Wrapping Up Bias

Sources of Bias



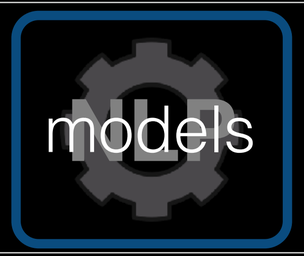
DESIGN



I admire Rosa Parks .

88 Shah et al., 2020; Hovy & Prabhumoye, 2021

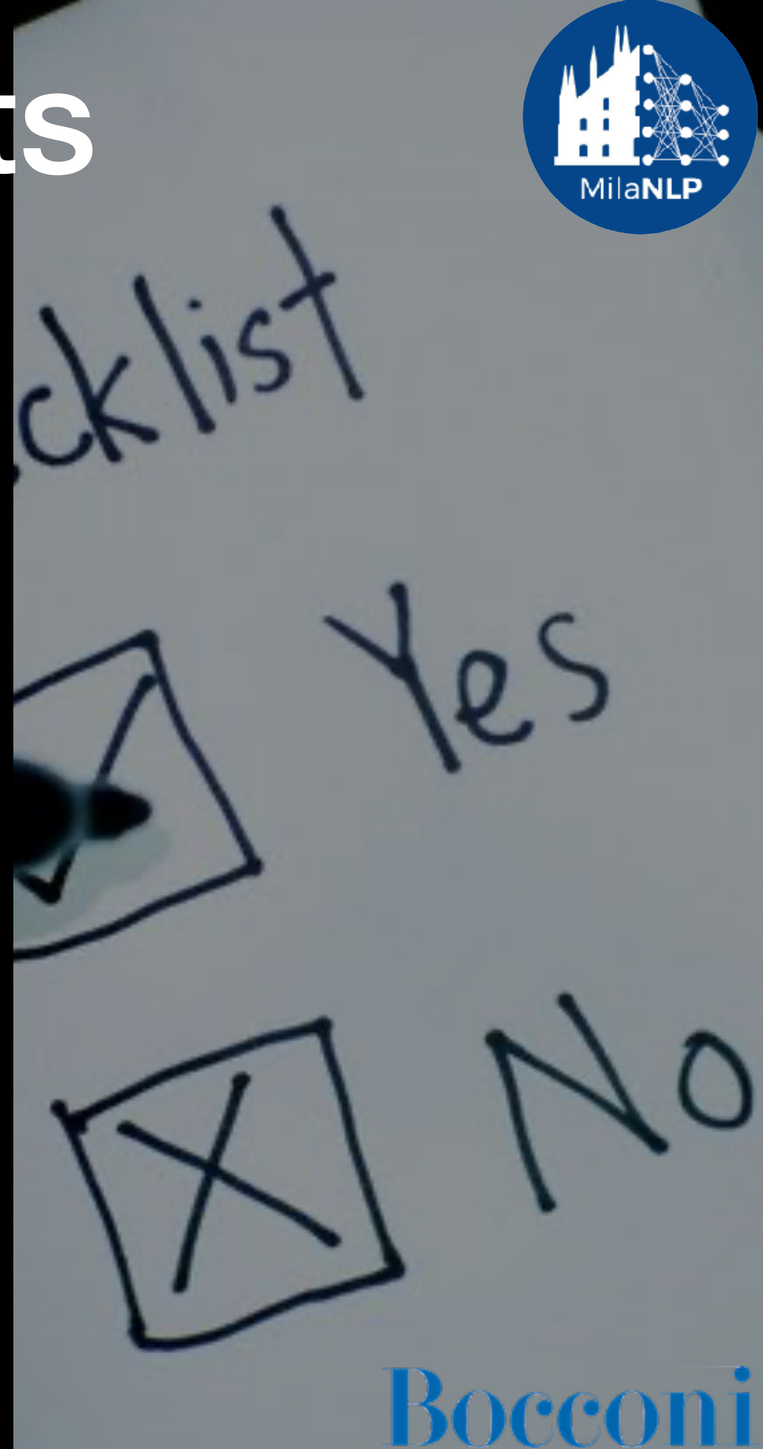
What can we do?

Source	Problem	Countermeasures
data selection	**Exclusion**	selection, stratification, priors
annotation	**Label Bias**	training, annotation models, disagreement weighting
 models	**Overgeneralization**	dummy labels, error weighting, adversarial learning, multitask learning, additional objectives
research design	**Exposure**	always consider possible impact, discuss, educate

 Hovy & Spruit (2016)

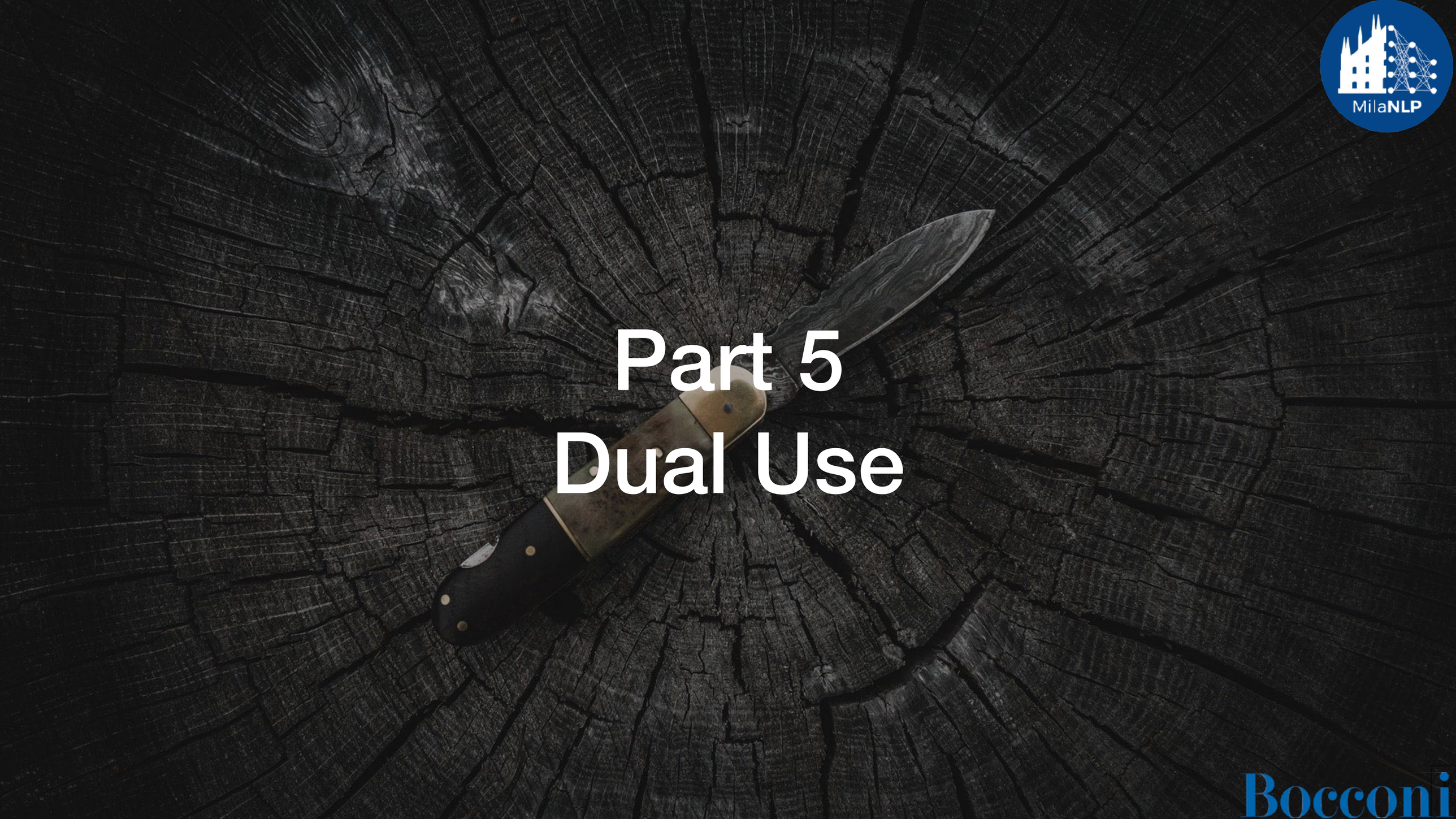
Take-home points

- 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



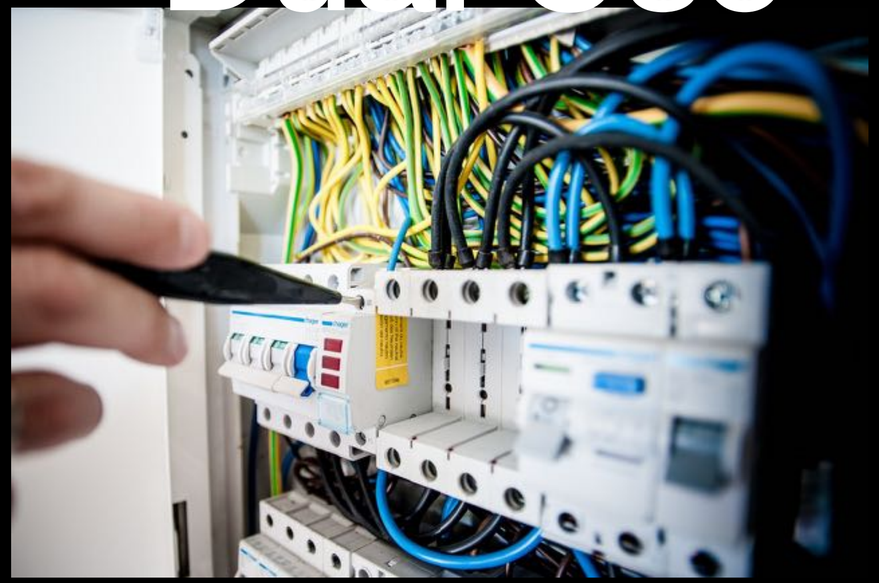
Bias Cases

- Can you think of other examples for
  - selection bias,
  - annotation bias,
  - model bias, and
  - design bias?
- Think about your most recent projects: are there any potential bias sources? How could you try to address them?
- Papers:
  - Shah, Schwartz & Hovy (ACL 2020): https://www.aclweb.org/anthology/2020.acl-main.468v2.pdf
  - Hovy & Prabhumoye (2021): https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12432



Part 5
Dual Use

Dual Use



Origin

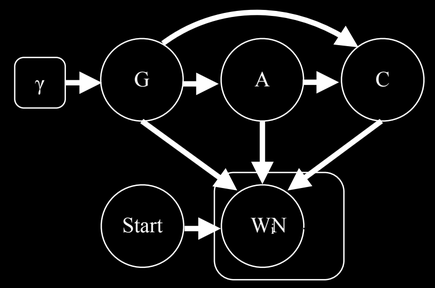
Hans Jonas

If a technology was developed, it will be used…

whether or not it is in the way intended is not up to us.

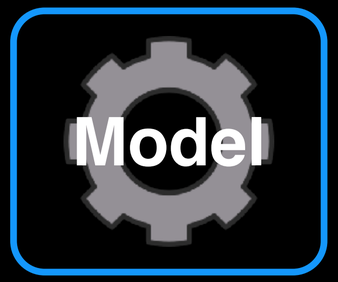
Dual Use

Fake or Real?



generate

detect



 Hovy, 2016

Dual Use

Task	Pro	Con
authorship attribution	historical documents	dissenter anonymity
text classification	sentiment analysis	censorship
personalization	better user experience	tailored ads

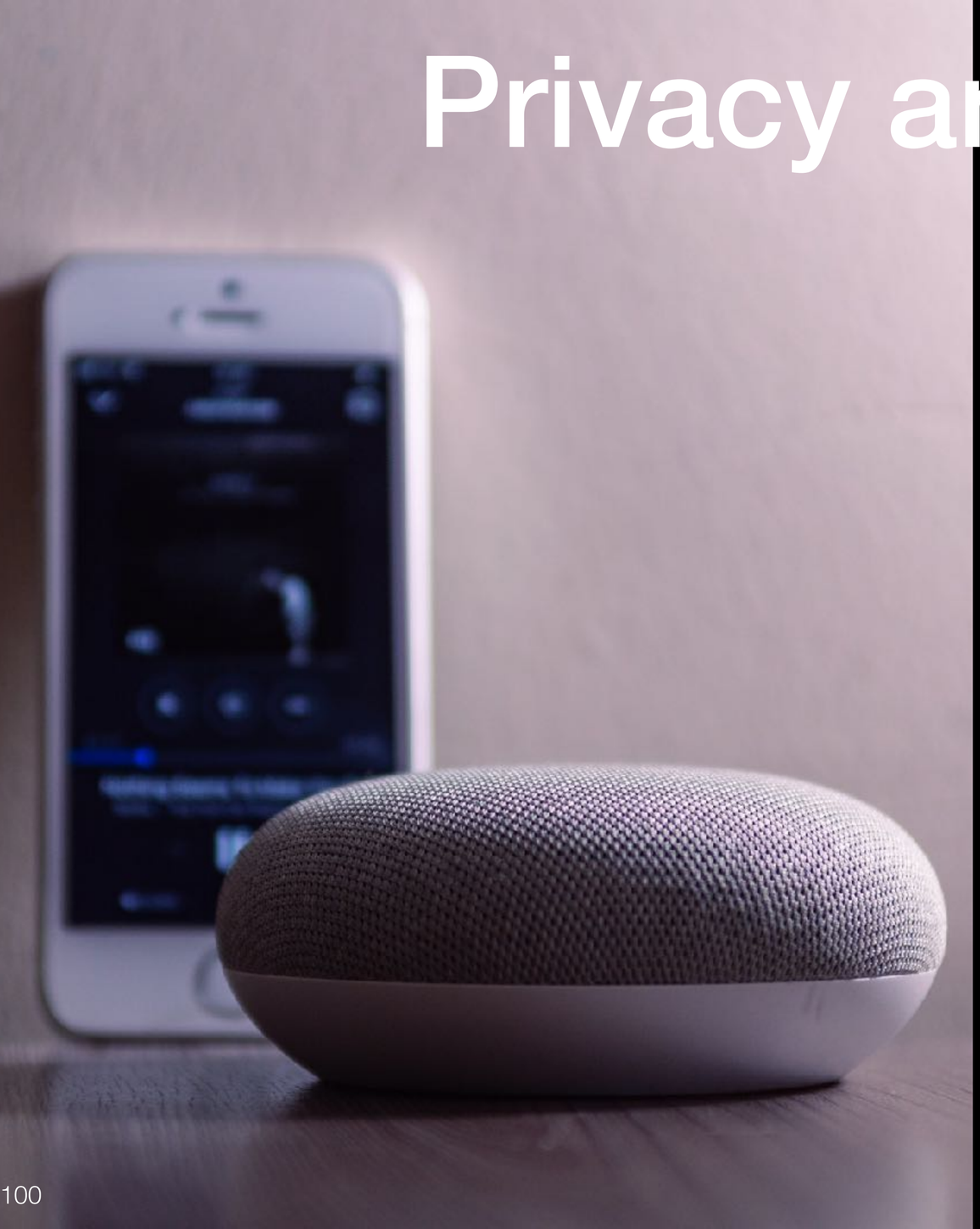
 Hovy & Spruit (ACL 2016)

Dual Use Case

- An ACL submission claims to be able to undo chiffres used by dissenters on social media. Who benefits from this? Is it better to release it peer-reviewed than to not know it?
- An LGBTQ+ student wants to write a thesis on "gay language", and study it via a classifier. What do you tell them?
- You develop a tool that can detect depression with high accuracy. Why, or why not, and how, should you release it?

Part III
Privacy

Privacy and Anonymity



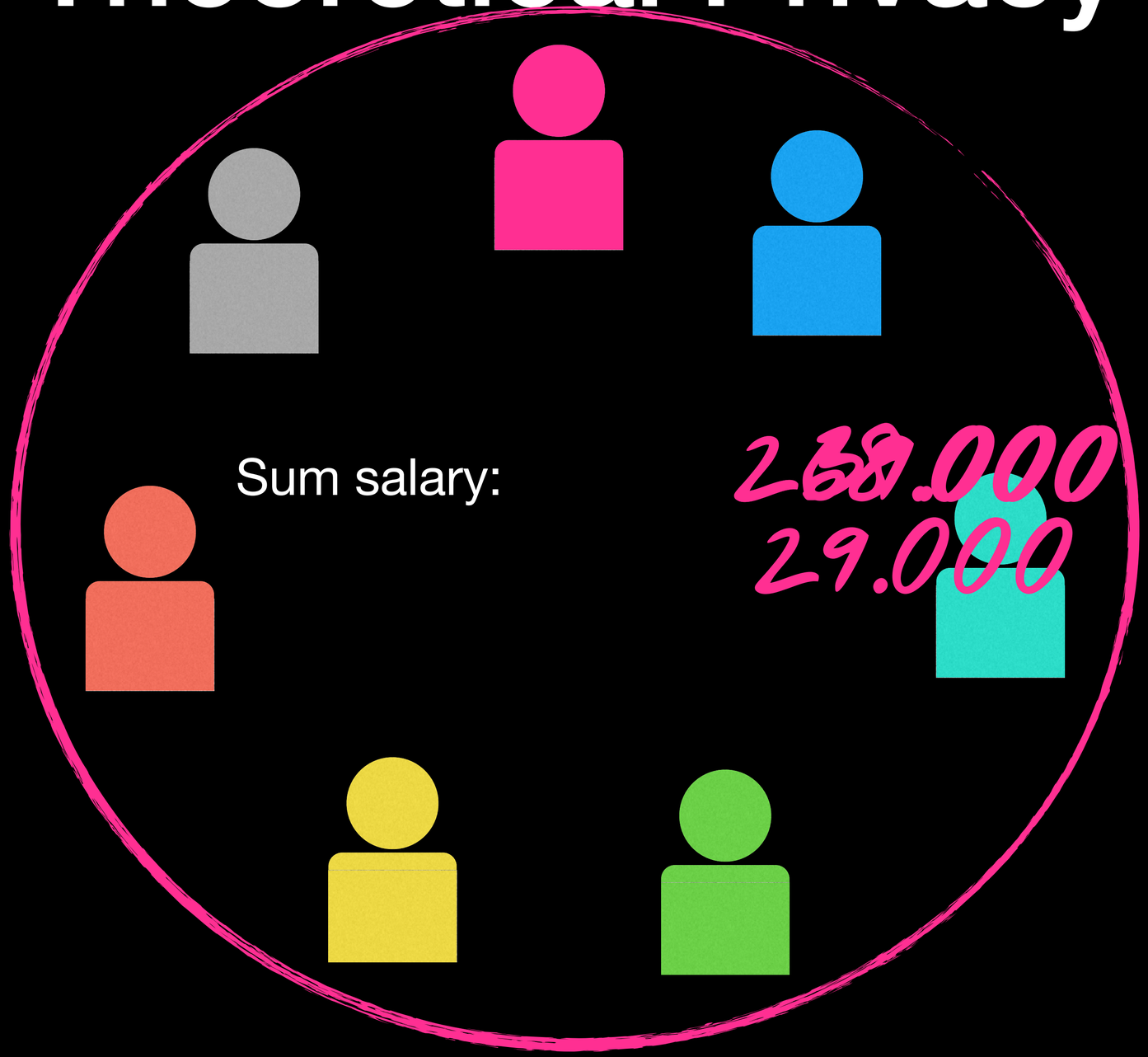
- Privacy: do things without others knowing you **do** them

- Anonymity: do things without others knowing it's **you**

The Law: GDPR

- A<u>nonymized</u> data is "*data rendered anonymous in such a way that the data subject is not or no longer identifiable*" without significant effort
- Assume a "*motivated intruder*"
- Random hashes: **good!**
- Pseudonyms/replacements: **bad!**

Theoretical Privacy



Dwork et al., 2006

Bad Privacy

1. Buy web browsing data

2. Find sites with clear name in URL

`https://analytics.twitter.com/user/dirk_hovy/tweets`

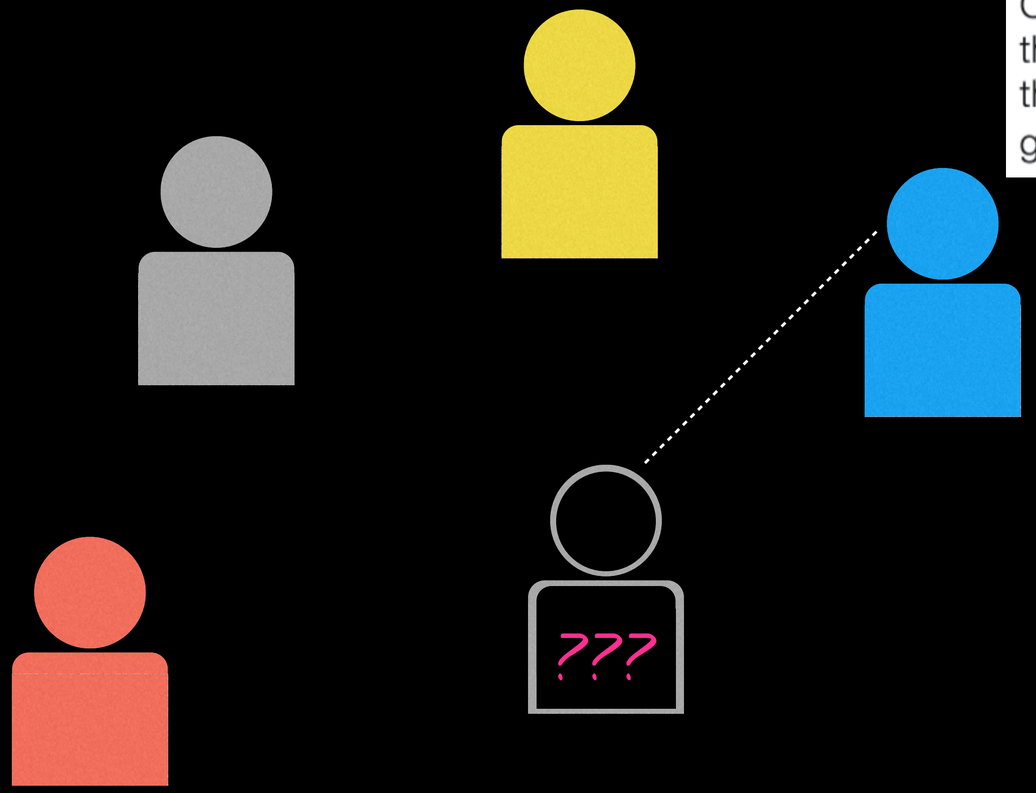
 Eckert & Dewes, 2017

Bad Hashing

target	accuracy
first 5 digits	7% / 44%
full 9 digits	up to 5% @10

Acquisti & Gross, 2009

Who needs enemies…

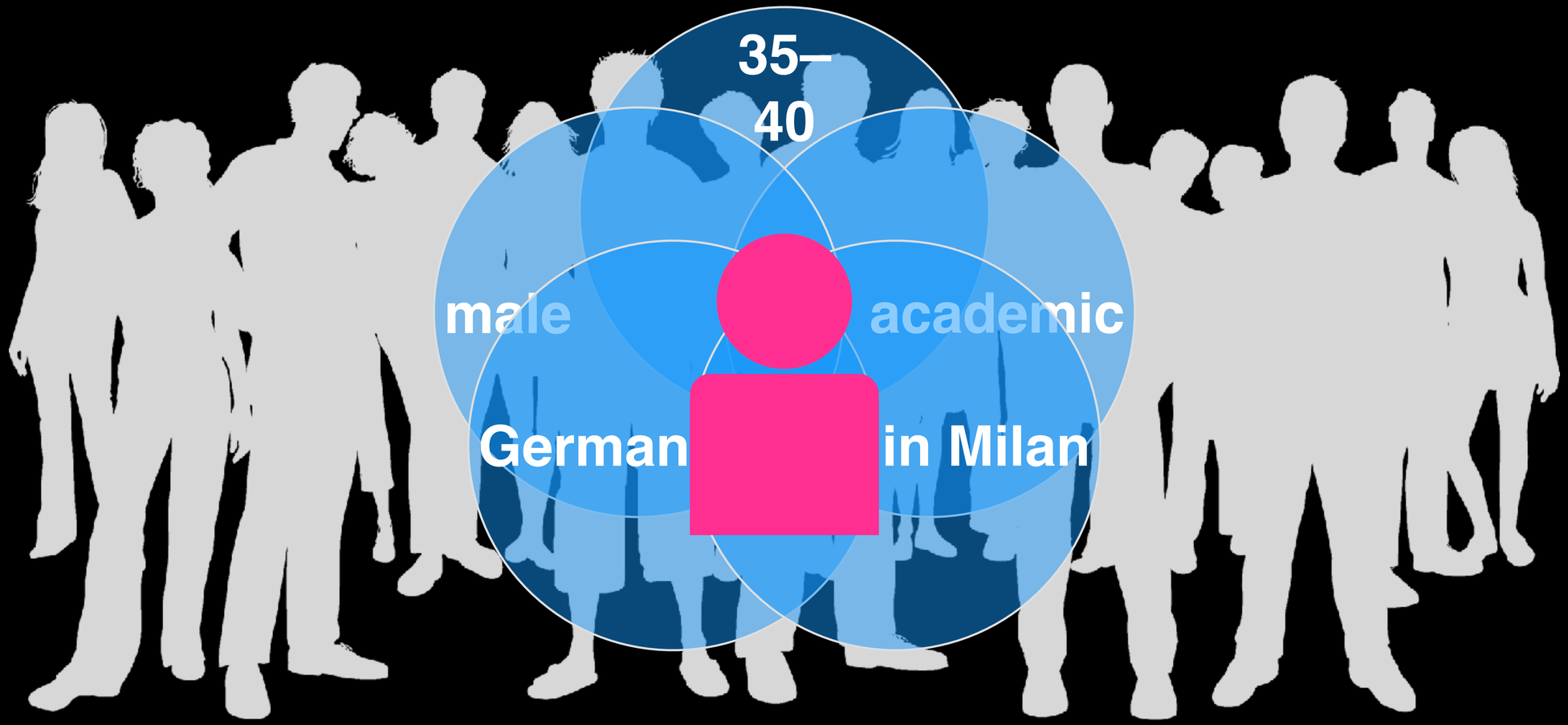


Gender, marital status, age range, family

IDENTIFICATION THROUGH FRIEND MENTIONS/NETWORK

105 Jurgens & al., (2017)

Speaker Idiolects



 Sweeney 2000; Yang & Eisenstein, 2017; Volkova et al., 2013; Hovy, 2015; Lynn et al., 2017; Benton et al., 2017; Li et al., 2018

Protected Categories

- Race
- Ethnicity
- Political
- Religious or philosophical beliefs
- Union membership
- Health, sex life, and sexual orientation
- Genetic and biometric data

ROLE OF AUTHOR ATTRIBUTE PREDICTION?

The Transparent User

age	Rosenthal and McKeown, 2011; Nguyen et al., 2011
gender	Alowibdi et al., 2013; Ciot et al., 2013; Liu and Ruths, 2013; Volkova et al.,
profession	Preoţiuc-Pietro et al., 2015
income	Preoţiuc-Pietro et al., 2015b
personality	Plank and Hovy, 2015
political views	Volkova et al., 2014
location	Bergsma et al., 2013; Rahimi et al., 2016

The role of Data

developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

Developer

Sensitive information

You should be careful about using Twitter data to derive or infer potentially sensitive characteristics about Twitter users. Never derive or infer, or store derived or inferred, information about a Twitter user's:

- Health (including pregnancy)
- Negative financial status or condition
- Political affiliation or beliefs
- Racial or ethnic origin
- Religious or philosophical affiliation or beliefs
- Sex life or sexual orientation
- Trade union membership
- Alleged or actual commission of a crime

The role of Data

https://www.reddit.com/wiki/api-terms

reddit

Search

2. Your Use of Reddit APIs.

a. License. Subject to the terms and conditions of these Terms, Reddit grants You a worldwide, non-exclusive, non-transferable, non-sublicensable, and revocable license to use the Reddit APIs in accordance with the terms and conditions set forth herein. All rights not expressly granted to you are reserved by Reddit.

b. End Users. You agree that you will ensure that your End Users comply with all associated terms and conditions of these Terms.

c. Compliance with Reddit Terms and Rules. You agree to comply with the Reddit User Agreement and Privacy Policy, which may be amended from time to time. Further, you agree to comply with the Reddit API Rules.

d. User Content. Reddit user photos, text and videos ("User Content") are owned by the users and not by Reddit. Subject to the terms and conditions of these Terms, Reddit grants You a non-exclusive, non-transferable, non-sublicensable, and revocable license to copy and display the User Content using the Reddit API through your application, website, or service to end users. You may not modify the User Content except to format it for such display. You will comply with any requirements or restrictions imposed on usage of User Content by their respective owners, which may include "all rights reserved" notices, Creative Commons licenses or other terms and conditions that may be agreed upon between you and the owners.

Data Management and Protection

1	Data Management Plan		(see https://www.scienceeurope.org/wp-content/uploads/2018/12/SE_RDM_Practical_Guide_Final.pdf)
2	General Information		
3		**Administrative information**	Provide information such as name of applicant, project number, funding programme, version of DMP
4	1 Data Description And Collection Or Re-Use of Existing Data		
5	1a	**How will new data be collected or produced and/or how will existing data be re-used?**	Explain which methodologies or software will be used if new data are collected or produced.
6			State any constraints on re-use of existing data if there are any.
7			Explain how data provenance will be documented.
8			Briefly state the reasons if the re-use of any existing data sources has been considered but discarded.
9	1b	**What data (for example the kind, formats, and volumes), will be collected or produced?**	Give details on the kind of data: for example numeric (databases, spreadsheets), textual (documents), image, audio, video, and/or mixed media.
10			Give details on the data format: the way in which the data is encoded for storage, often reflected by the filename extension (for example pdf, xls, doc, txt, or rdf).
11			Justify the use of certain formats. For example, decisions may be based on staff expertise within the host organisation, a preference for open formats, standards accepted by data repositories, widespread usage within the research community, or on the software or equipment that will be used.
12			Give preference to open and standard formats as they facilitate sharing and long-term re-use of data (several repositories provide lists of such 'preferred formats').
13			Give details on the volumes (they can be expressed in storage space required (bytes), and/or in numbers of objects, files, rows, and columns).
14	2 Documentation And Data Quality		
15	2a	**What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?**	Indicate which metadata will be provided to help others identify and discover the data.
16			Indicate which metadata standards (for example DDI, TEI, EML, MARC, CMDI) will be used.
17			Use community metadata standards where these are in place.
18			Indicate how the data will be organised during the project, mentioning for example conventions, version control, and folder structures. Consistent, well-ordered research data will be easier to find, understand, and re-use.
19			Consider what other documentation is needed to enable re-use. This may include information on the methodology used to collect the data, analytical and procedural information, definitions of variables, units of measurement, and so on.
20			Consider how this information will be captured and where it will be recorded for example in a database with links to each item, a 'readme' text file, file headers, code books, or lab notebooks.
21	2b	**What data quality control measures will be used?**	Explain how the consistency and quality of data collection will be controlled and documented. This may include processes such as calibration, repeated samples or measurements, standardised data capture, data entry validation, peer review of data, or representation with controlled vocabularies.
22	3 Storage And Backup During The Research Process		
23	3a	**How will data and metadata be stored and backed up during the research?**	Describe where the data will be stored and backed up during research activities and how often the backup will be performed. It is recommended to store data in least at two separate locations.
24			Give preference to the use of robust, managed storage with automatic backup, such as provided by IT support services of the home institution. Storing data on laptops, stand-alone hard drives, or external storage devices such as USB sticks is not recommended.

Privacy Cases

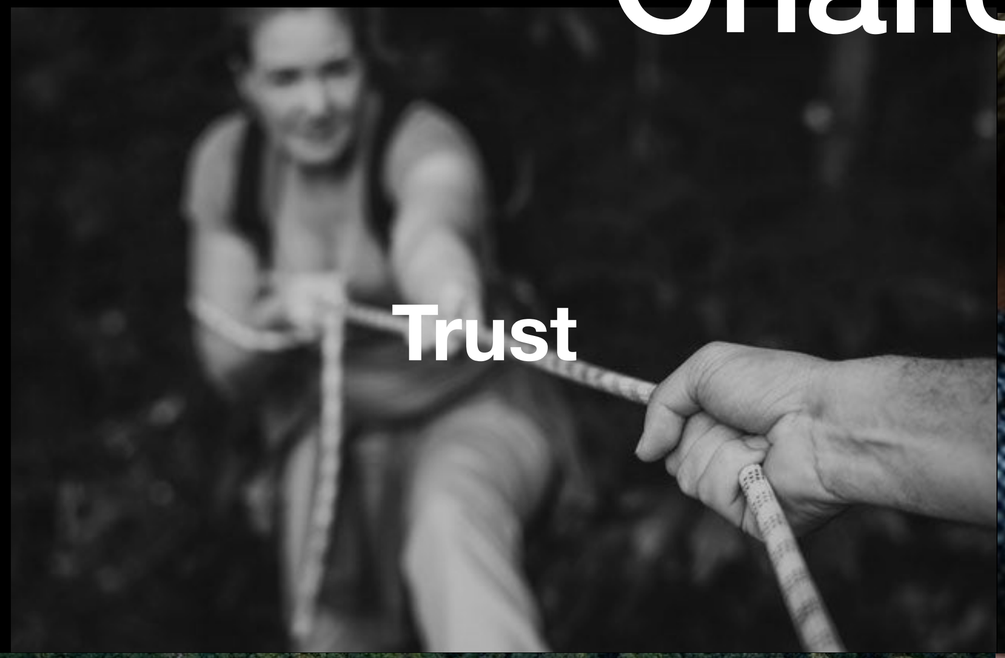
- Given what you know about author attribute prediction and inference from a few demographic factors: is there any real chance to preserve privacy? What about anonymity?
- What do anonymization techniques mean for corpus work? Can we use adversarial approaches?
- What do document/author embeddings across corpora do to privacy?

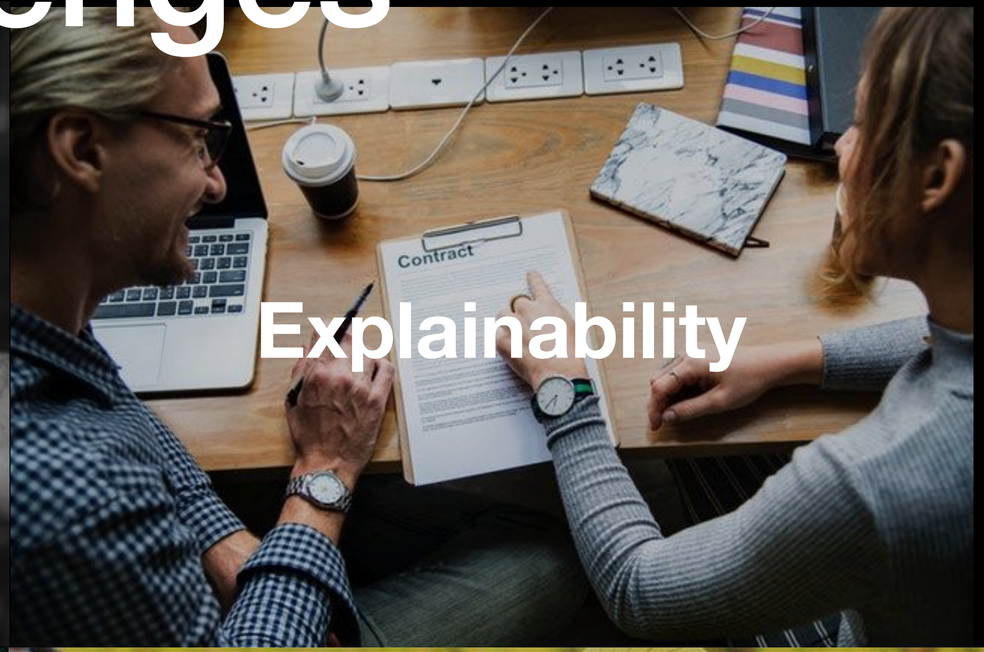
Wrapping It ALL Up

Take-home points

- Beware of **bias** from **data**, **labels**, **embeddings**, **models**, and **design**
- Apply **countermeasures** and check
- Your models ***will*** be used in unintended ways
- Ask yourself:
  "*Am I comfortable with my system classifying me?*"
- Privacy is brittle, but we can protect individual attributes

Challenges





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, develop novel methods

industry:

more performant tools in MT, dialogue, search

www.dirkhovy.com/portfolio/papers
milanlproc.github.io/

Thank you!

IF YOU ARE INTERESTED IN WORKING WITH ME: I HAVE UPCOMING POSTDOC POSITIONS IN THESE AREAS

@dirk_hovy

www.dirkhovy.com

INTEGRATOR
European Research Council

Bocconi

www.dirkhovy.com/portfolio/papers
milanlproc.github.io/

Questions?

If you are interested in working with me: I have upcoming postdoc positions in these areas

@dirk_hovy

www.dirkhovy.com

INTEGRATOR
European Research Council

MilaNLP

Bocconi