Five Sources of Biases and Ethical Issues in NLP, and What to Do about Them

Dirk Hovy
Bocconi University, Milan

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

@dirk_hovy

Machine Translation

MilaNLP

HELL
YES

Auf
jeden
Fall

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.

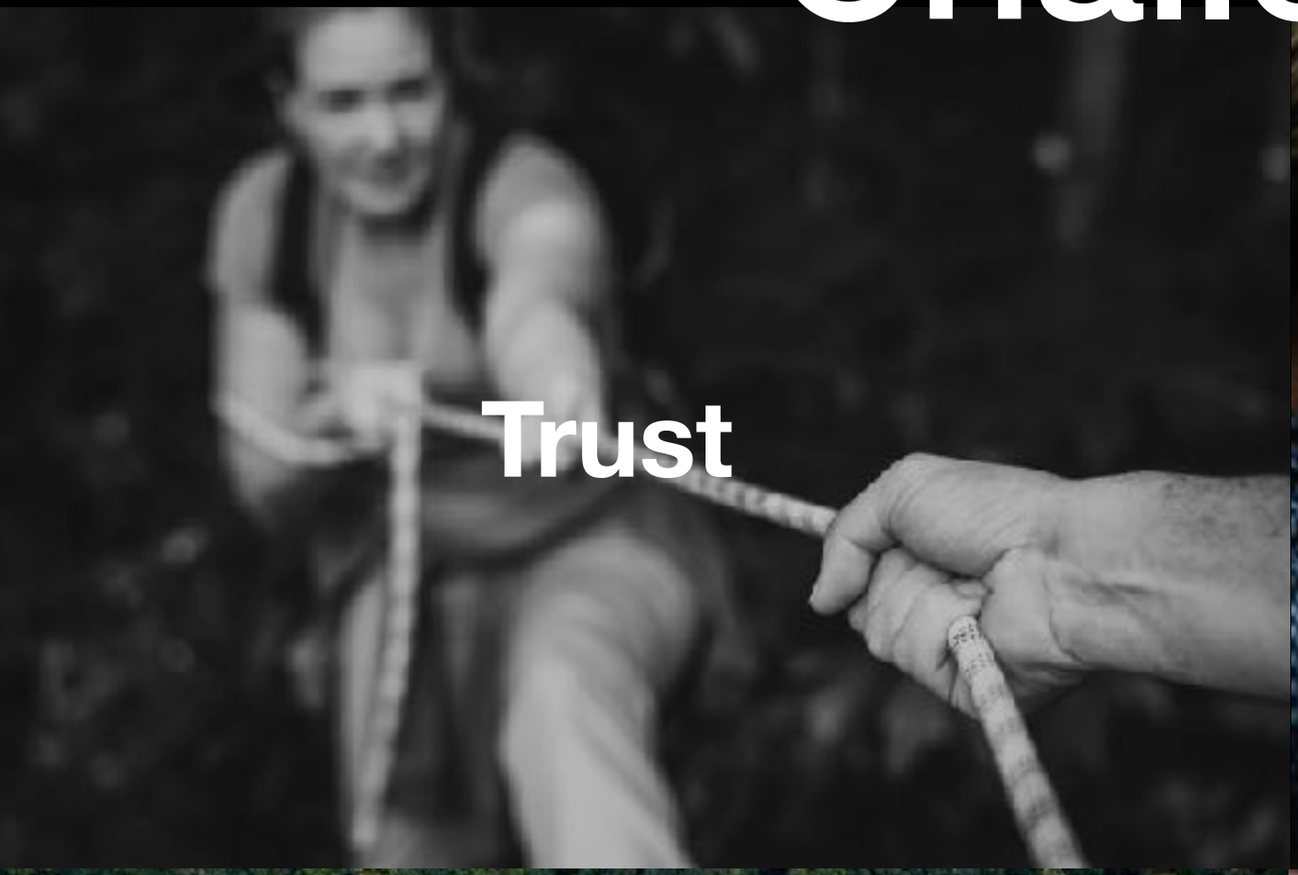
Biased Systems

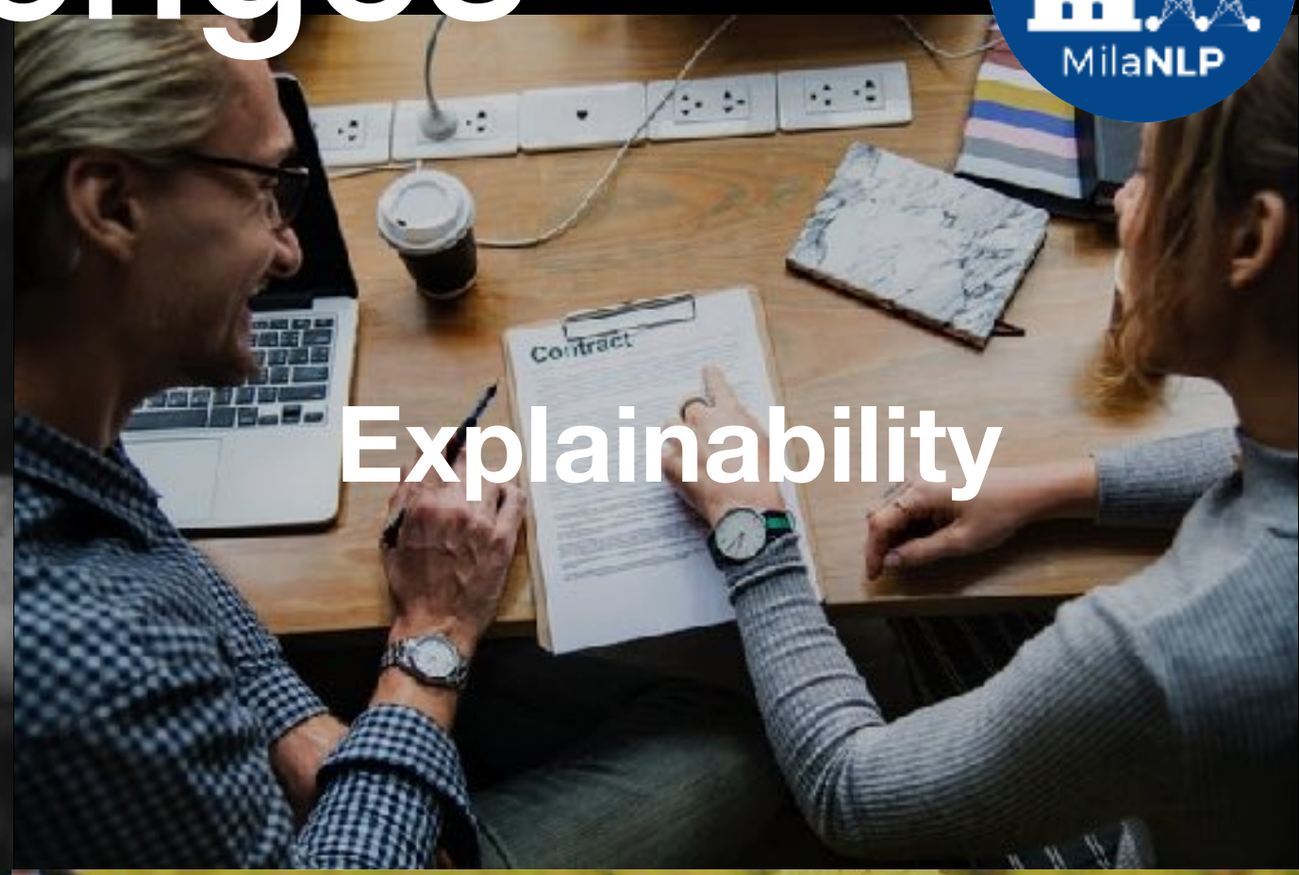
NLP WORKS WELL

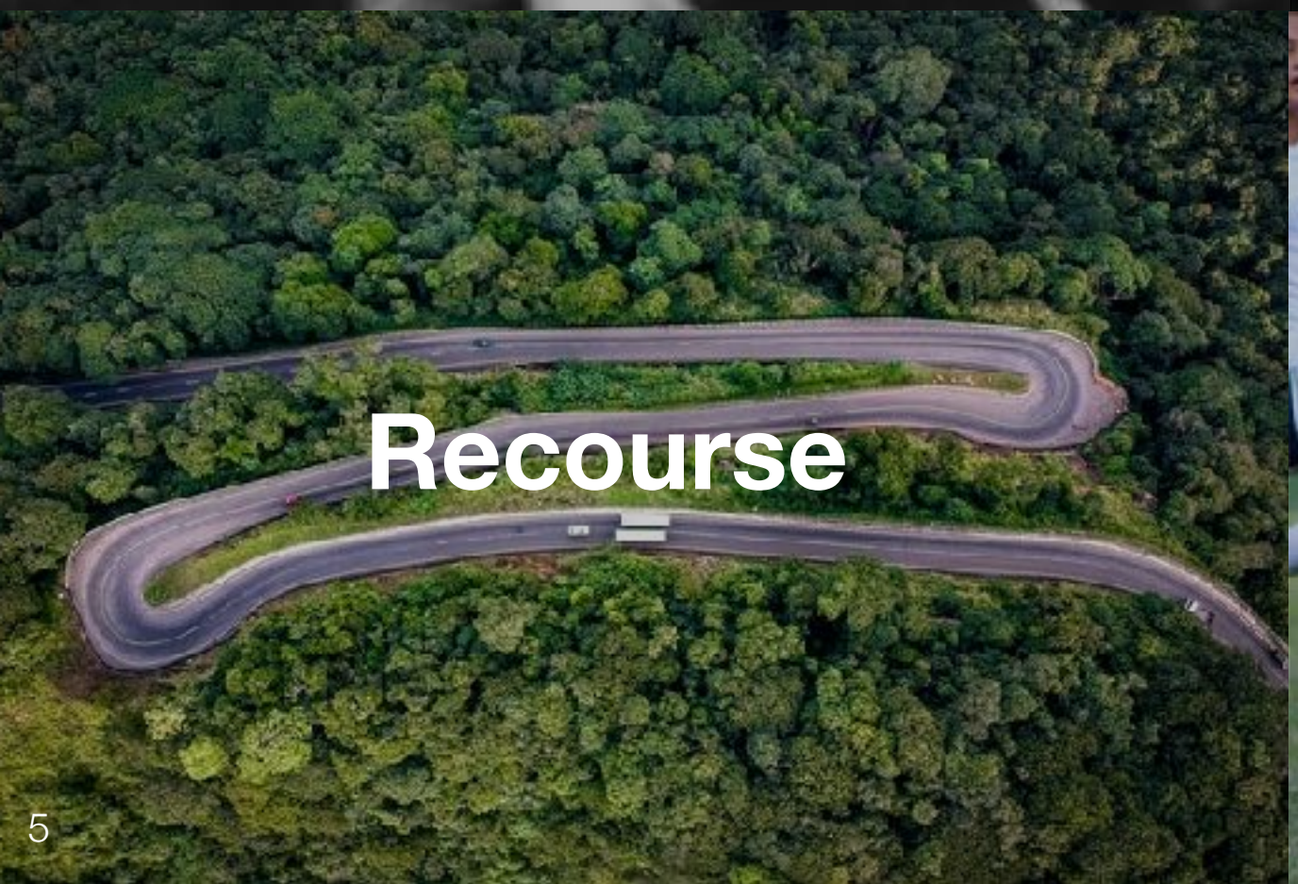
population

Bocconi

Challenges







Goals for Today

- Introduce **terminology** and **concepts** of **bias**
- Show five **sources** of bias in NLP
- Discuss **consequences**
- Introduce **countermeasures**

- For a paper, see Shah, Schwartz & Hovy (ACL 2020): https://www.aclweb.org/anthology/2020.acl-main.468v2.pdf

Terminology

Bias

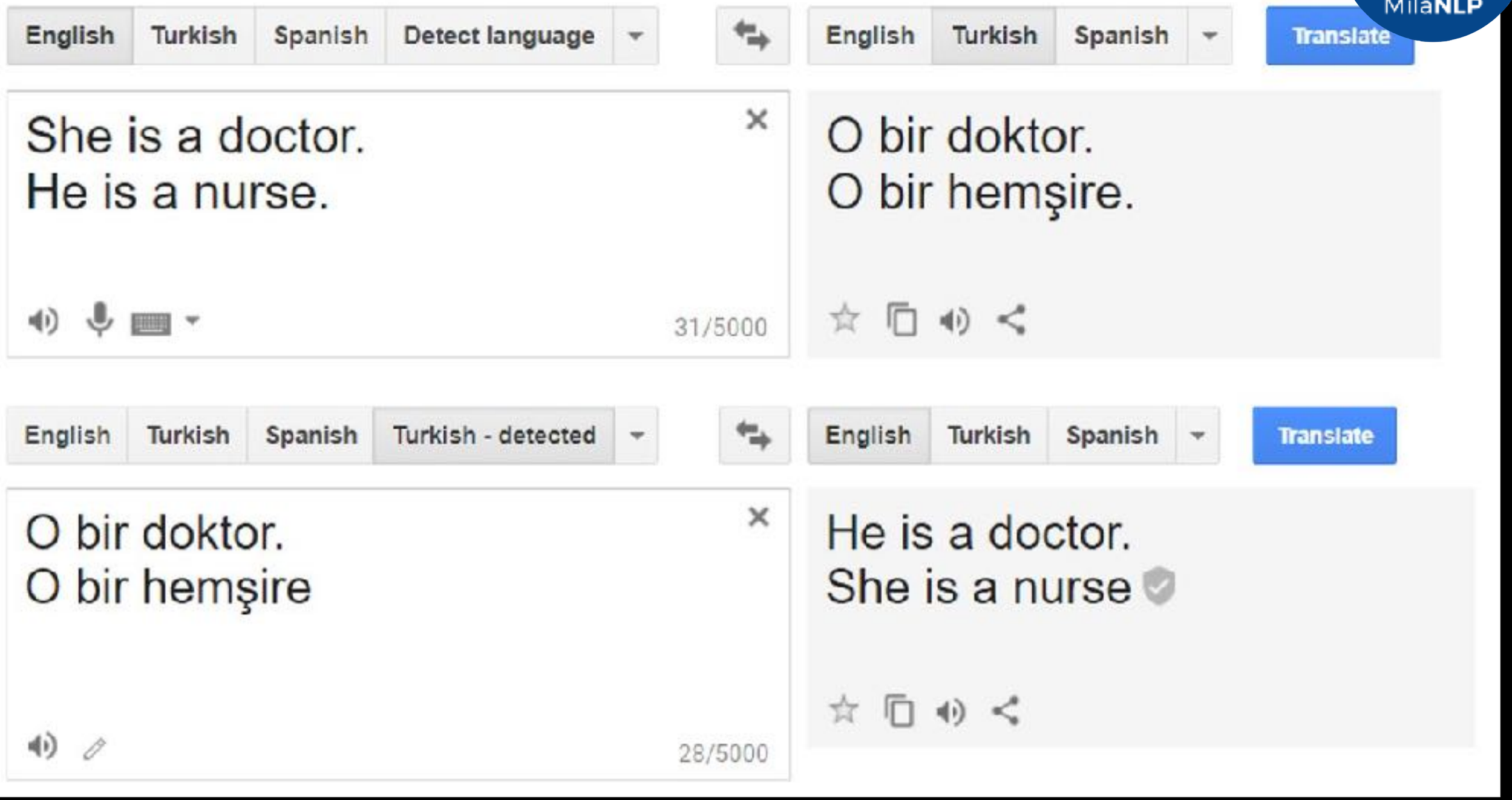
Bias ≠ Bad

Ethics ≠ 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?

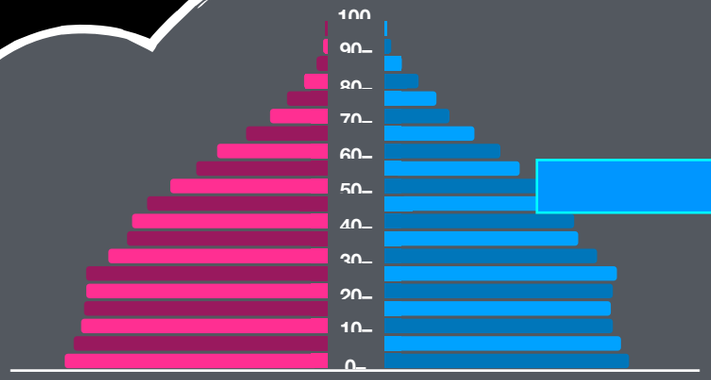
Dual Use

+ Intended Use

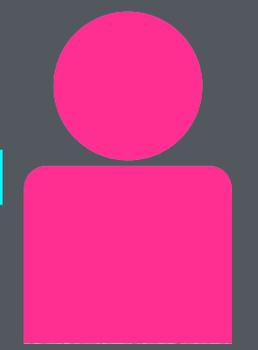
- Unintended Use or consequences

Sources of Bias

Sources of Bias



DATA



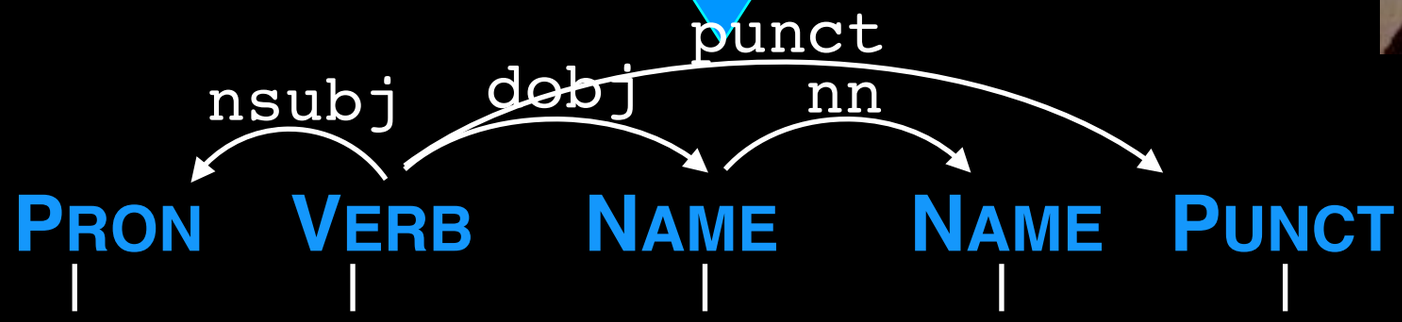
ANNOTATION

REPRESENTATIONS

MODELS

NLP

DESIGN



Bocconi

Shah et al. (ACL 2020)

Formally...

origin

consequence

over-amplification
The model discriminates on a given human attribute beyond its source base-rate.

label bias
Biased annotations, interaction, or latent bias from past classifications.

outcome disparity
The distribution of outcomes, given attribute A, is dissimilar than the *ideal distribution*:
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$

Embedding Corpus
features $\theta_{embedding}$
(Pre-trained Side)

Source Population
features X_{source} — *fit* — outcomes Y_{source}
(Model Side)

Target Population
features X_{target} — *predict* — biased outcomes $\hat{Y}_{target}$
(Application Side)

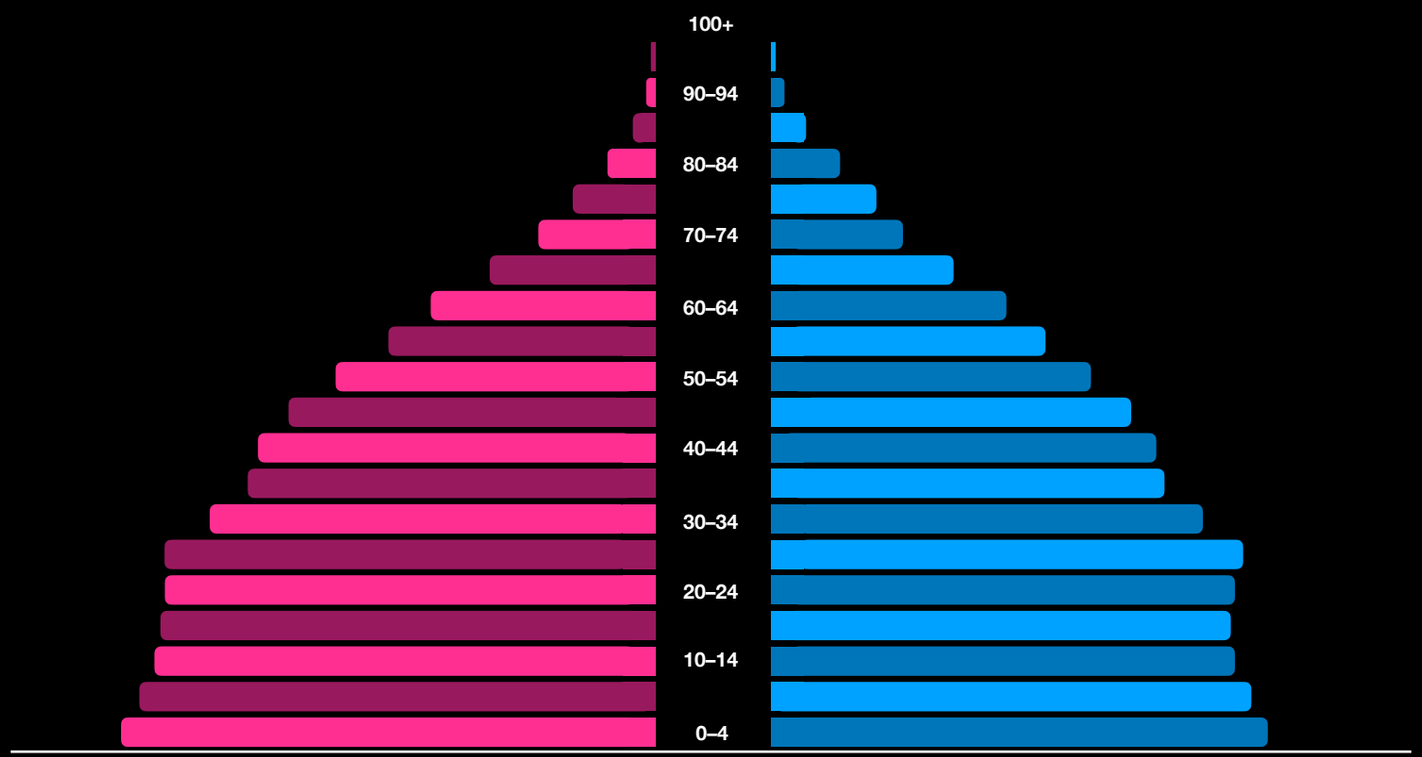
semantic bias
Non-ideal associations between attributed lexeme (e.g. gendered pronouns) and non-attributed lexeme (e.g. occupation).

selection bias
The sample of observations themselves are not representative of the application population.

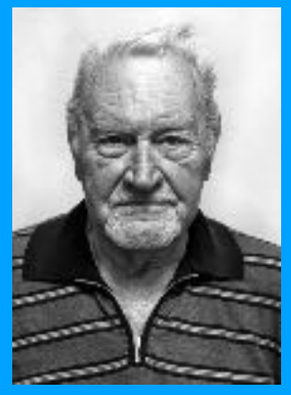
error disparity
The distribution of error (ϵ) over at least two different values of an attribute (A) are unequal:
$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$

Bocconi

Part 1:
Selection Bias

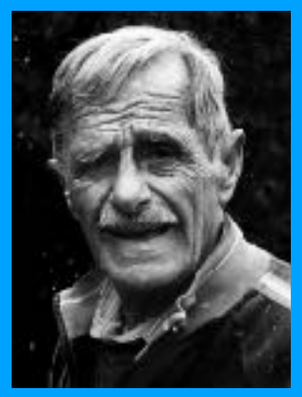


Language Varies



Example 1

Example 2



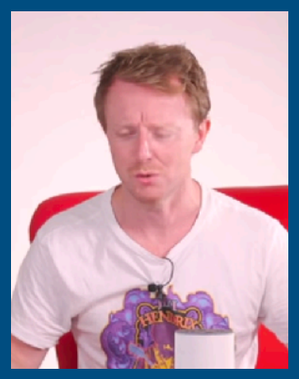
Example N

I don't understand you…

System

Hello, computer

Shite…



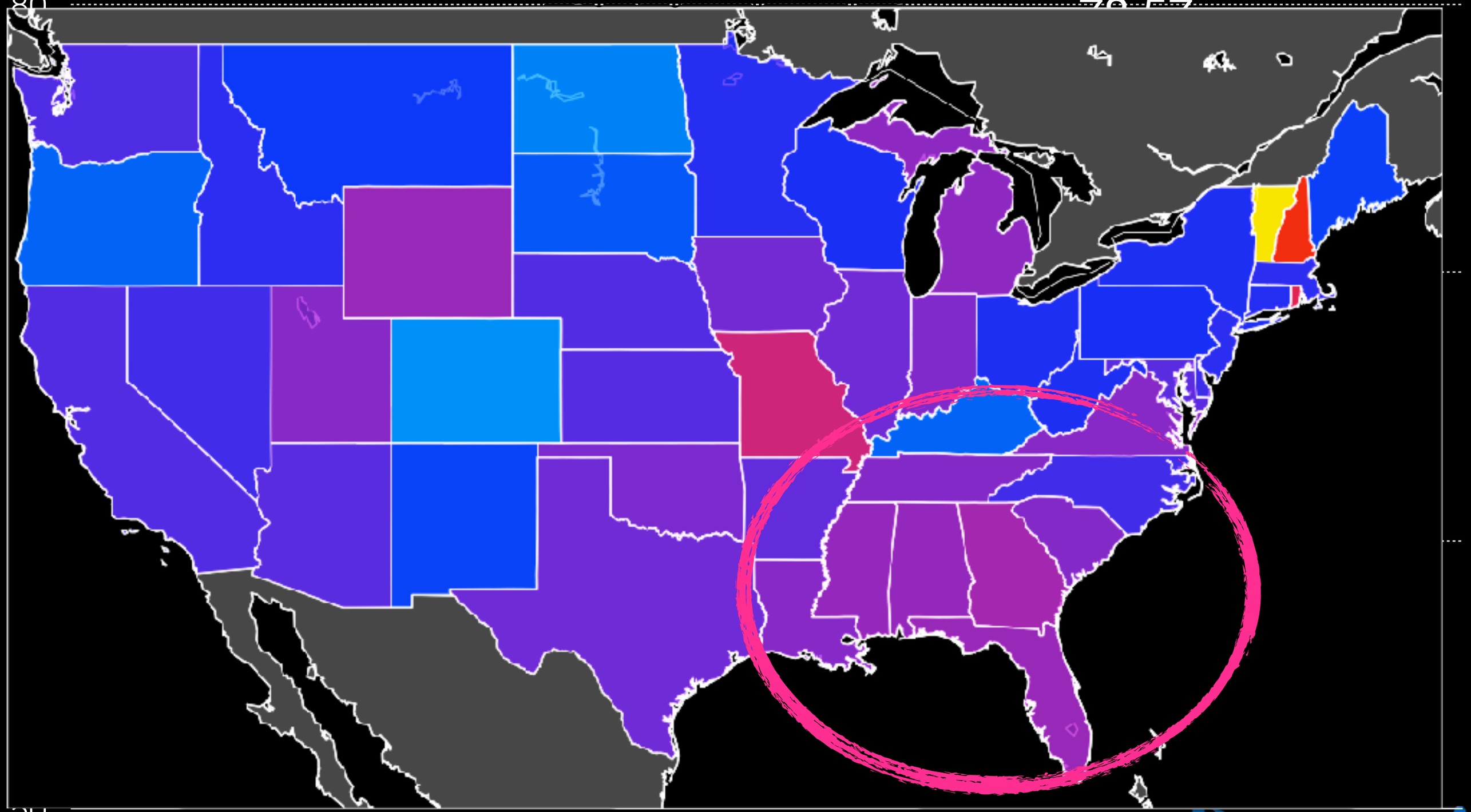
Bocconi

Exclusion

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

F1

80

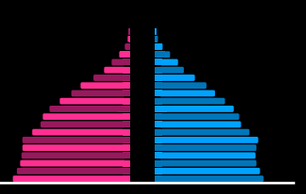


50

17

avg

Bocconi



Exclusion

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

accuracy

100
95
90
85
80

100+
90–94
80–84
70–74
60–64
50–54
40–44
30–34
20–24
10–14
0–4

42

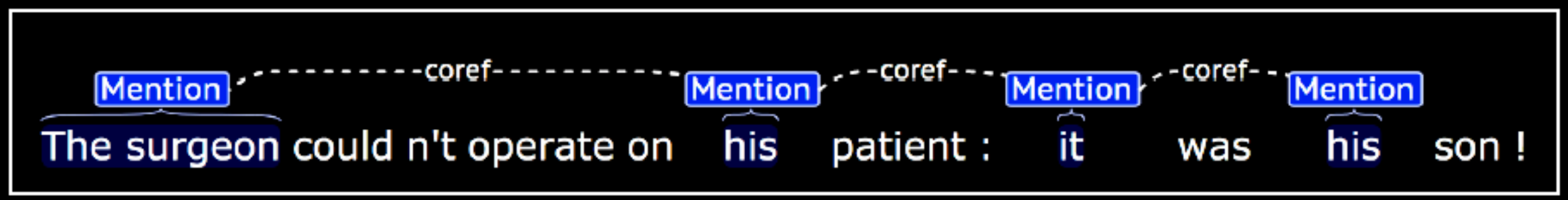
POS-tagging
600 user reviews
CoNLL trained

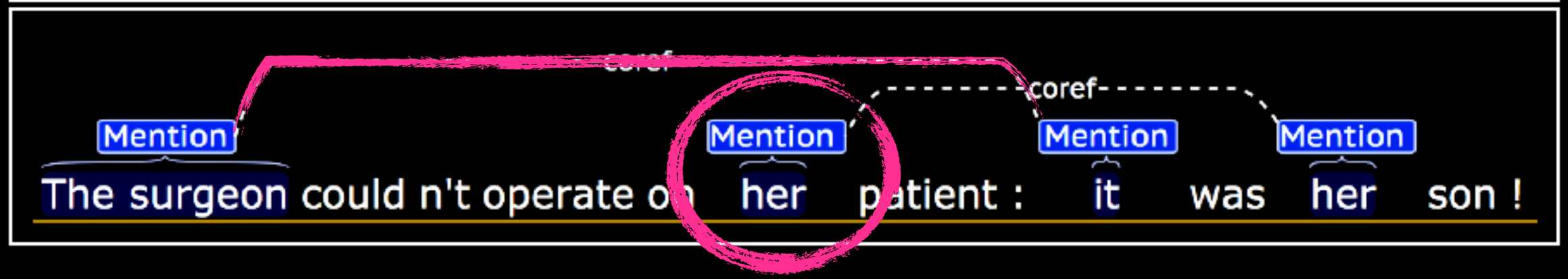
O45 U35 O45 U35

Bocconi

Wrong Coreference

Rudinger et al. (2018)





Bocconi

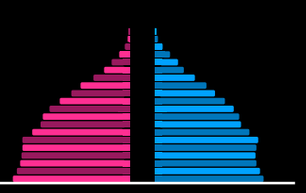
More generally…

The sample of observations is not representative of the application population

Distribution of Attribute in Training Data

$$Q(A_S) \neq P(A_t)$$

Ideal Distro of Attribute in Target Data



Idea!

Volkova et al. (2013)
Hovy (2015)
Lynn et al. (2017)

INCLUDE DEMOGRAPHIC INFORMATION IN TEXT REPRESENTATION

Hovy (ACL 2015)

Systems

AGNOSTIC

training data

This is a tiny little example text written by someone.

This is a tiny little example text written by someone.

+

INFORMED

training data

This is a tiny little example text written by someone.

This is a tiny little example text written by someone.

+

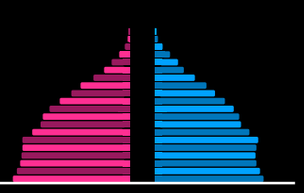
choose

training data

This is a tiny little example text written by someone.

This is a tiny little example text written by someone.

+



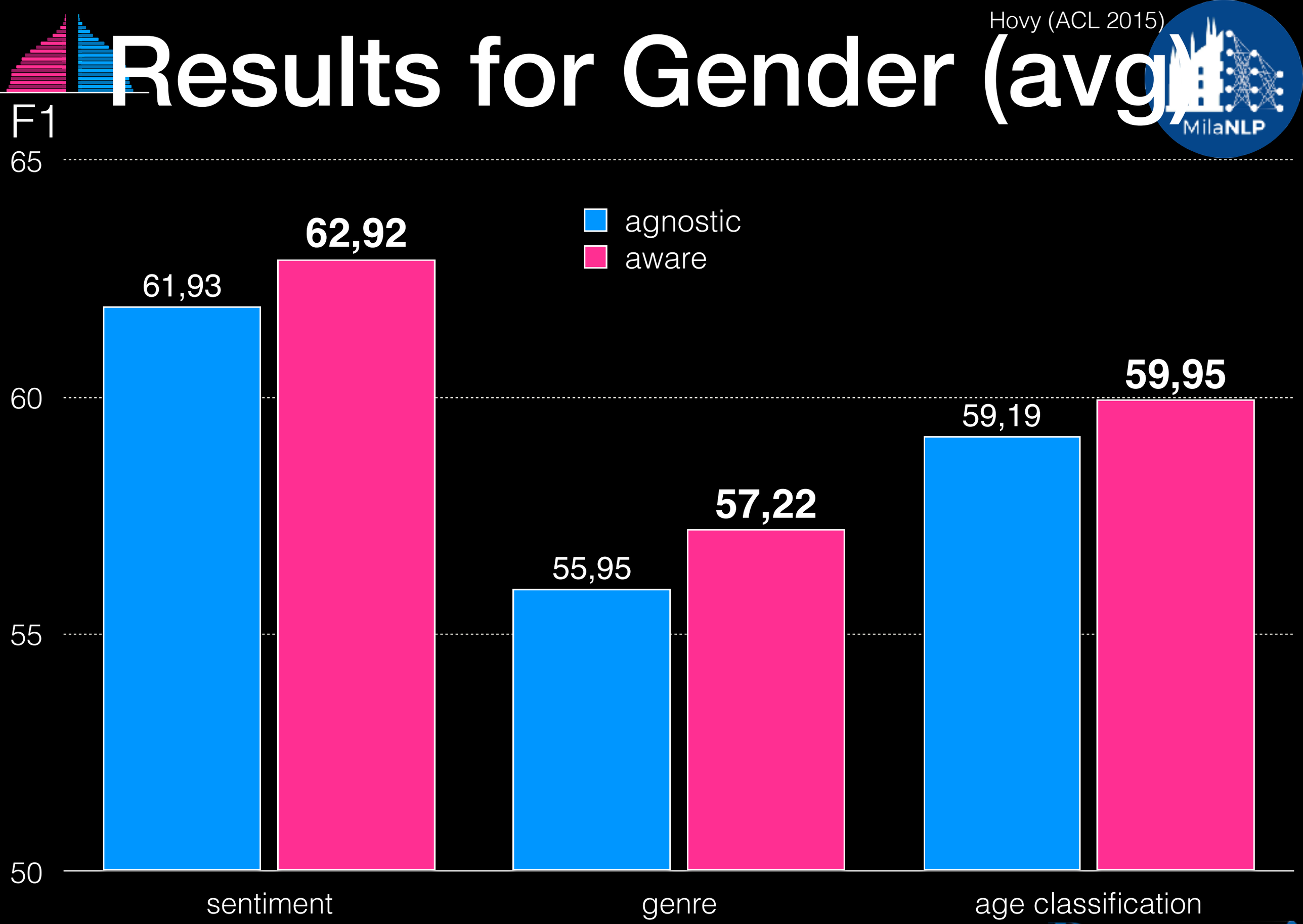
Results for Age (avg)

Hovy (ACL 2015)

Bocconi

Results for Gender (avg)

Hovy (ACL 2015)



avg

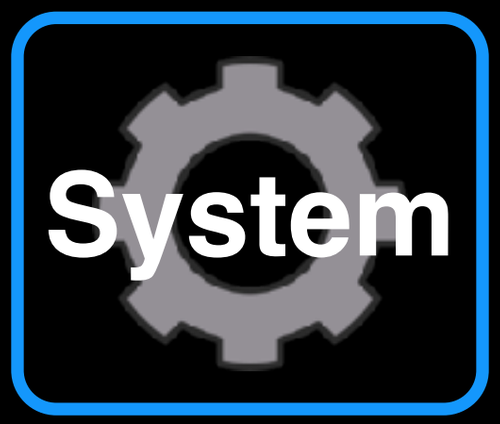
Better Selection

Example 1

Example 2

Example N

Yes, But...



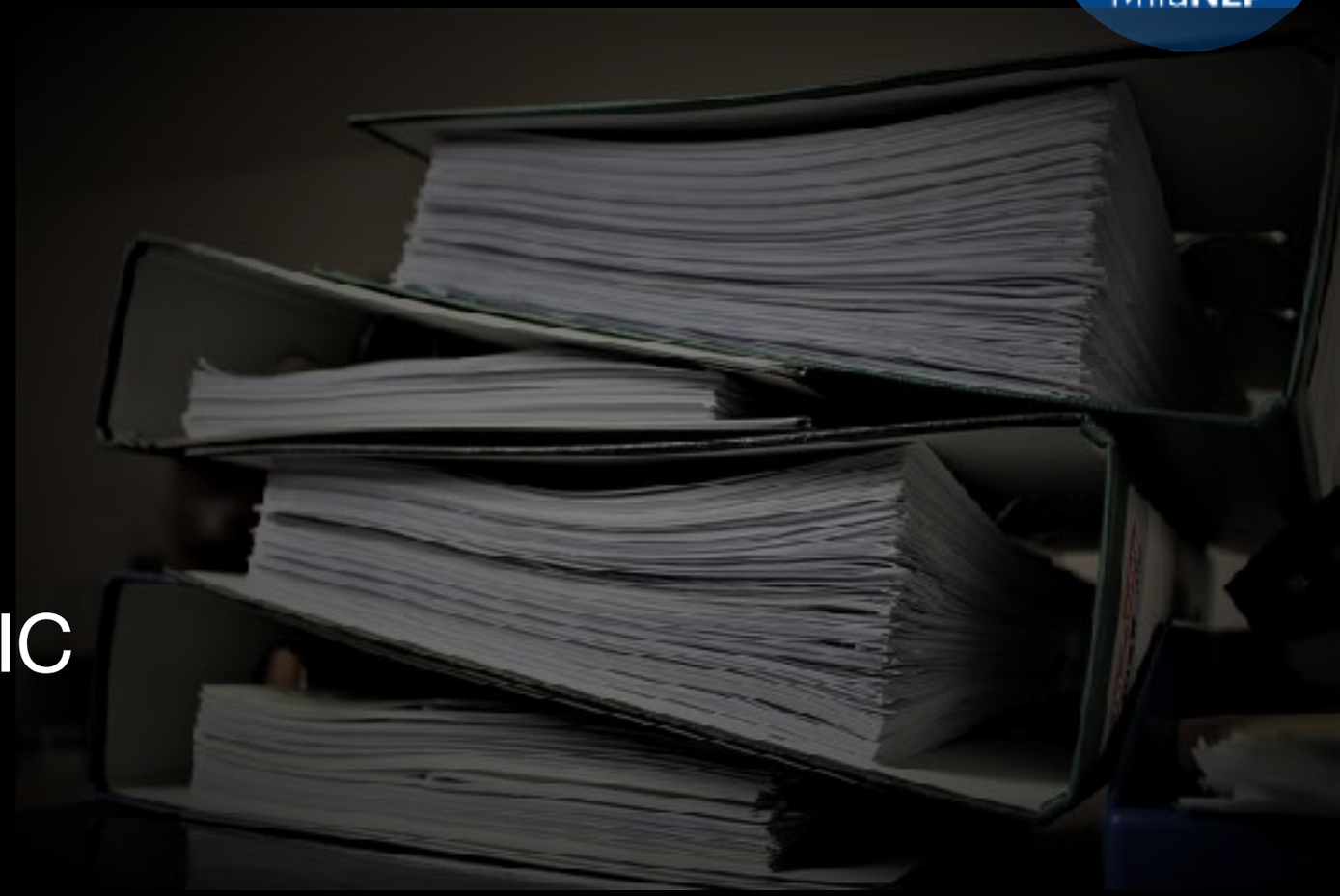
Idea!

MAKE POSSIBLE
SOURCES OF BIAS EXPLICIT!

Bender & Friedman 2018

Data Statements

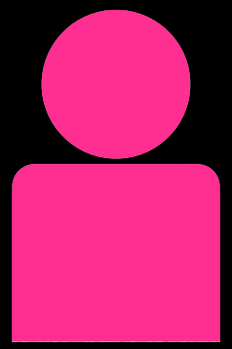
- CURATION RATIONALE
- LANGUAGE VARIETY
- SPEAKER DEMOGRAPHIC
- ANNOTATOR DEMOGRAPHIC
- SPEECH SITUATION
- TEXT CHARACTERISTICS
- RECORDING QUALITY
- OTHER



Reconstruct Collection

Part 2:

Label Bias



Bocconi

Annotator Bias

Whatever, it's X

HAS NO CLUE...

No! It's a NOUN!

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

it is on social media

Bocconi

More Annotator Bias

It's an ADJ

It’s a NOUN

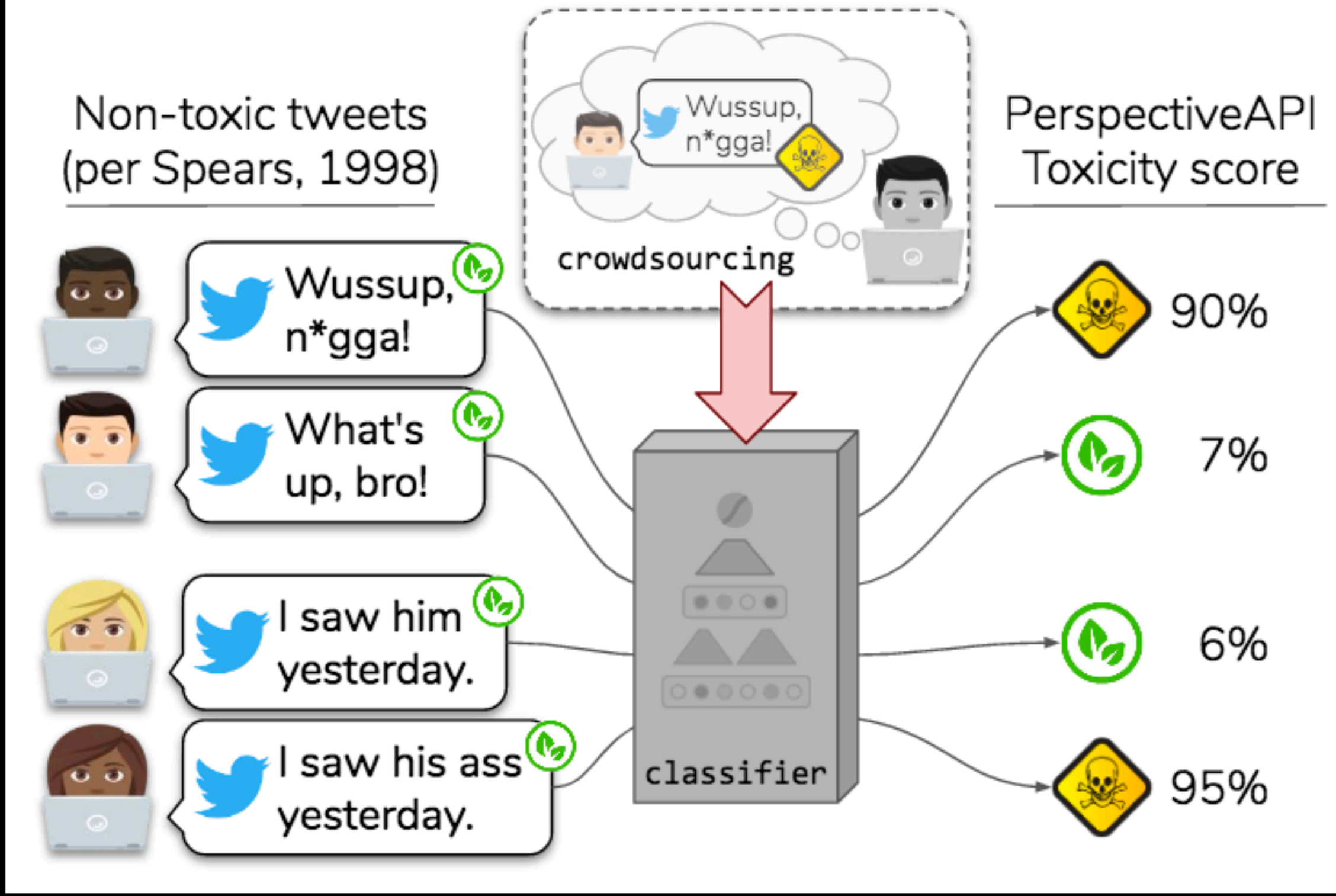
WHAT IF YOU'RE BOTH RIGHT?

PRON VERB ADP	**ADJ**	NOUN
PRON VERB ADP	**NOUN**	NOUN

it is on social media

Sap et al. (2019)

Even more Annotator Bias...



Bocconi

Basically…

Pavlick et al. (2014)

TRAIN YOUR ANNOTATORS!

More generally…

Biased annotations, interaction, or latent bias from past classifications.

Actual Distro of Label wrt Attribute in Training Data

$$Q(Y_S \mid A_S) \neq P(Y_S \mid A_S)$$

Ideal Distro of Label wrt Attribute in Source Data

Idea!

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

MilaNLP

Find out Who's Reliable!

Bocconi

Hovy et al. (NAACL 2013)
Paun et al. (TACL 2018)

Model

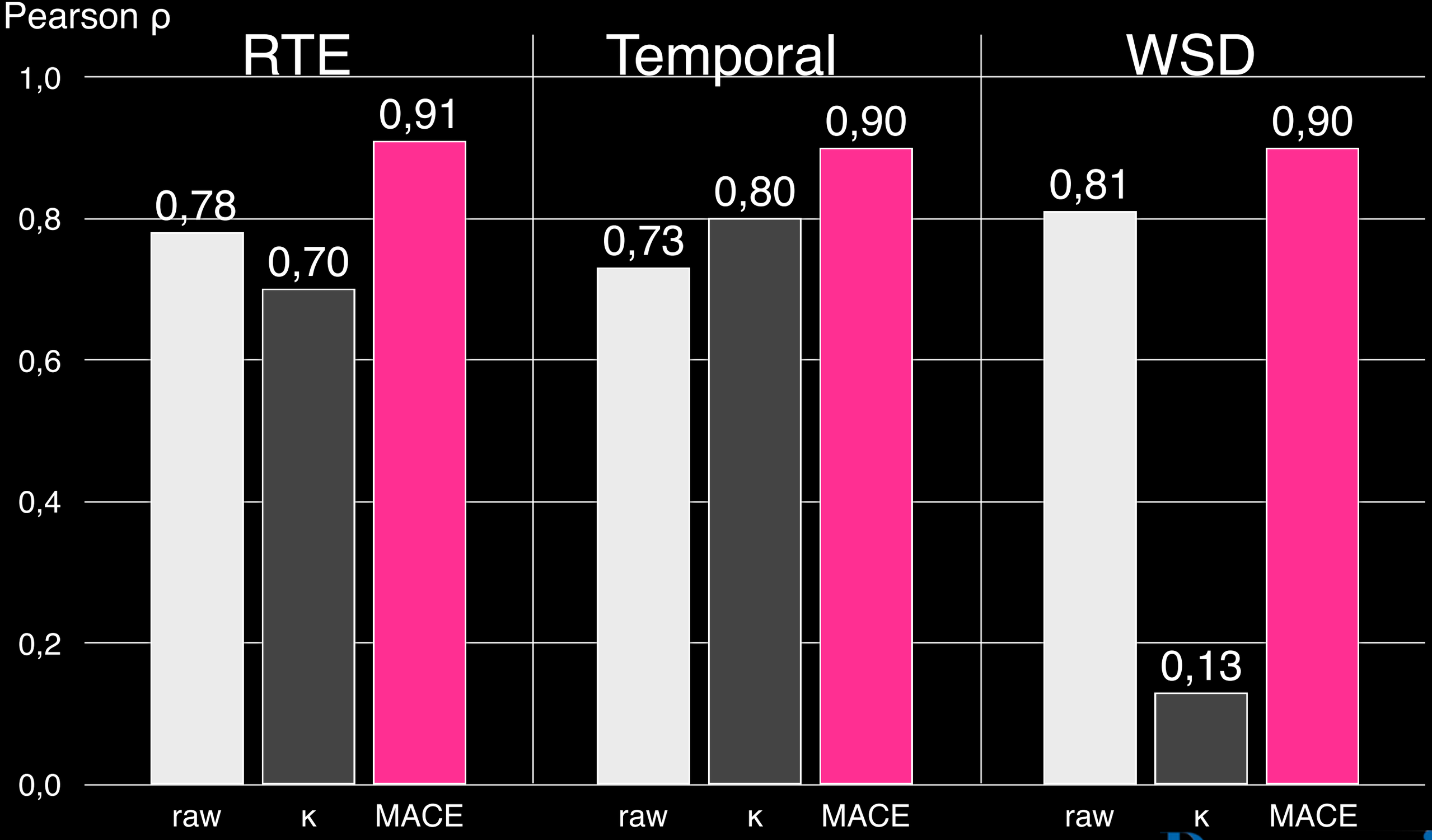
MACE

https://github.com/dirkhovy/MACE

Bocconi

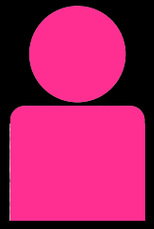
Correlation with Proficiency

Hovy et al. (NAACL 2013)



Prediction Accuracy

Hovy et al. (NAACL 2013)



Bocconi

Idea!

Plank et al. (EACL 2014)
Jamison & Gurevych (2015)

MODEL LABEL UNCERTAINTY!

Bocconi

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

Easy and Hard Categories

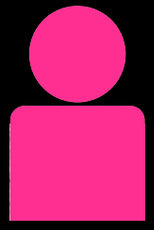
EASILY CONFUSED BY ANNOTATORS

	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

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



Confusion Matrices in Discriminative Trainin

Plank et al. (EACL 2014)

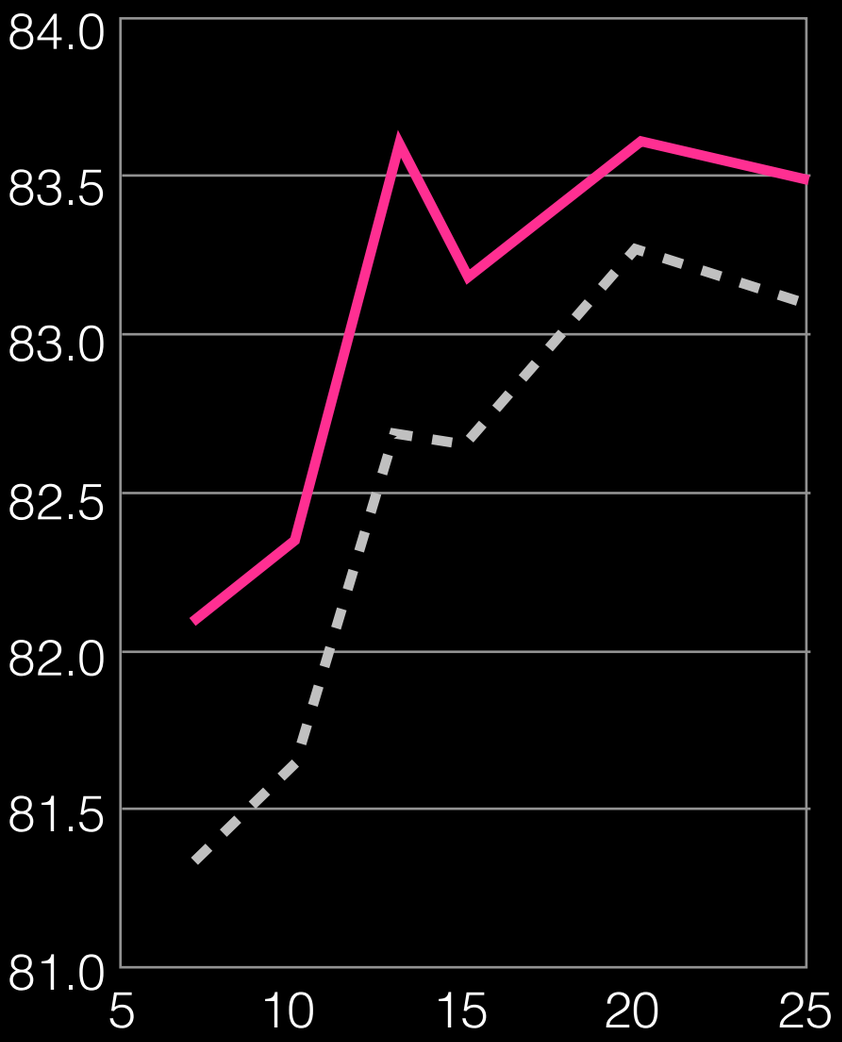

```
if model confuses

      easy cases (e.g., NOUN-VERB):

            normal update

      difficult cases (e.g., NOUN-ADJ):

            update * CM
```

POS Results

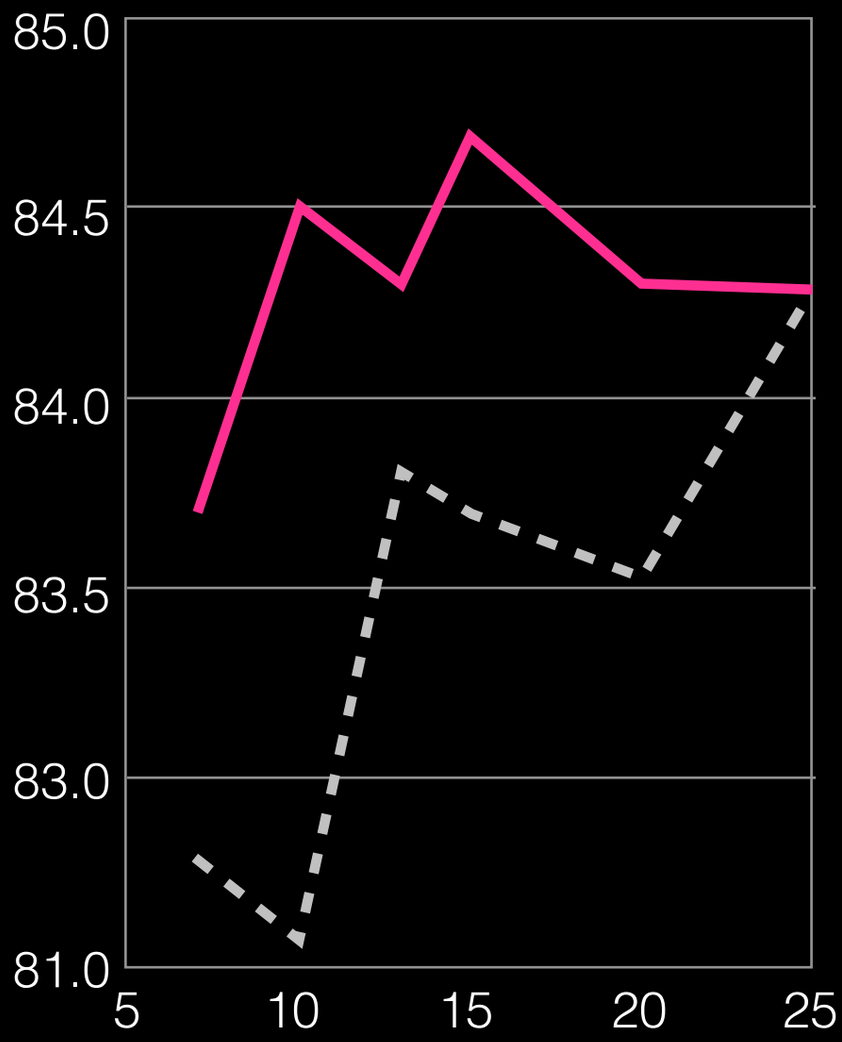
Plank et al. (EACL 2014)

Baseline
CM Training

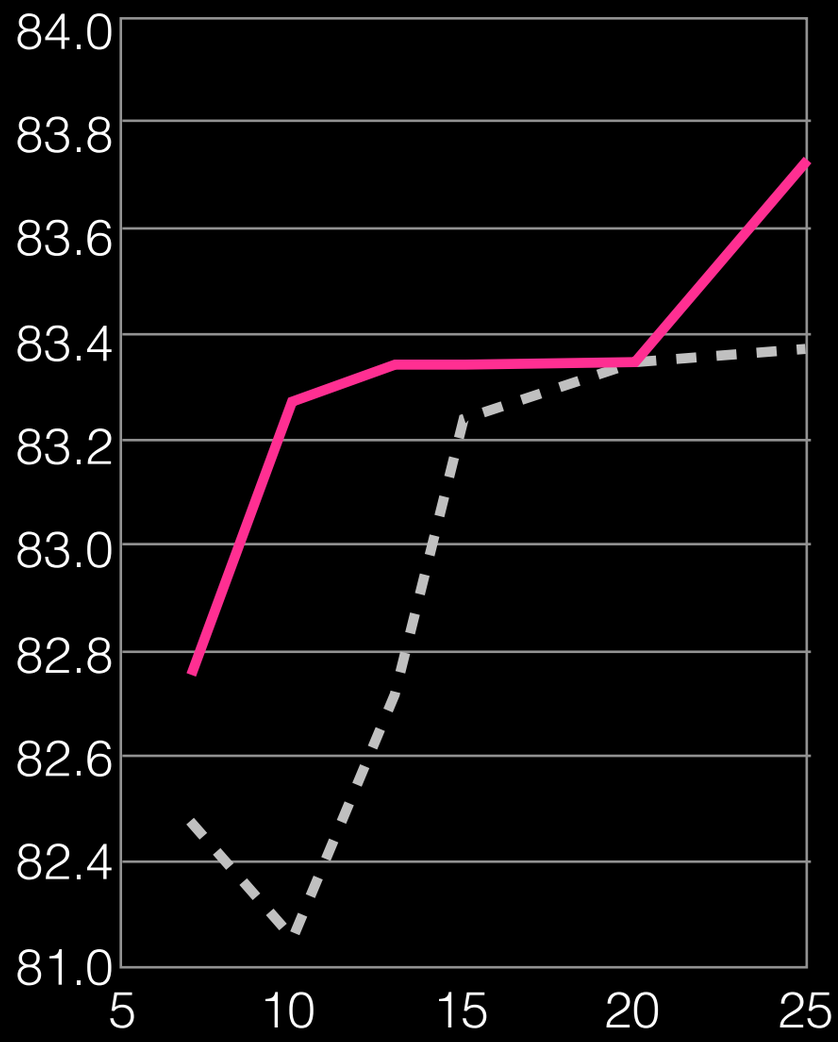
Trained on Gimpel+Ritter



Foster Dev



Foster Test

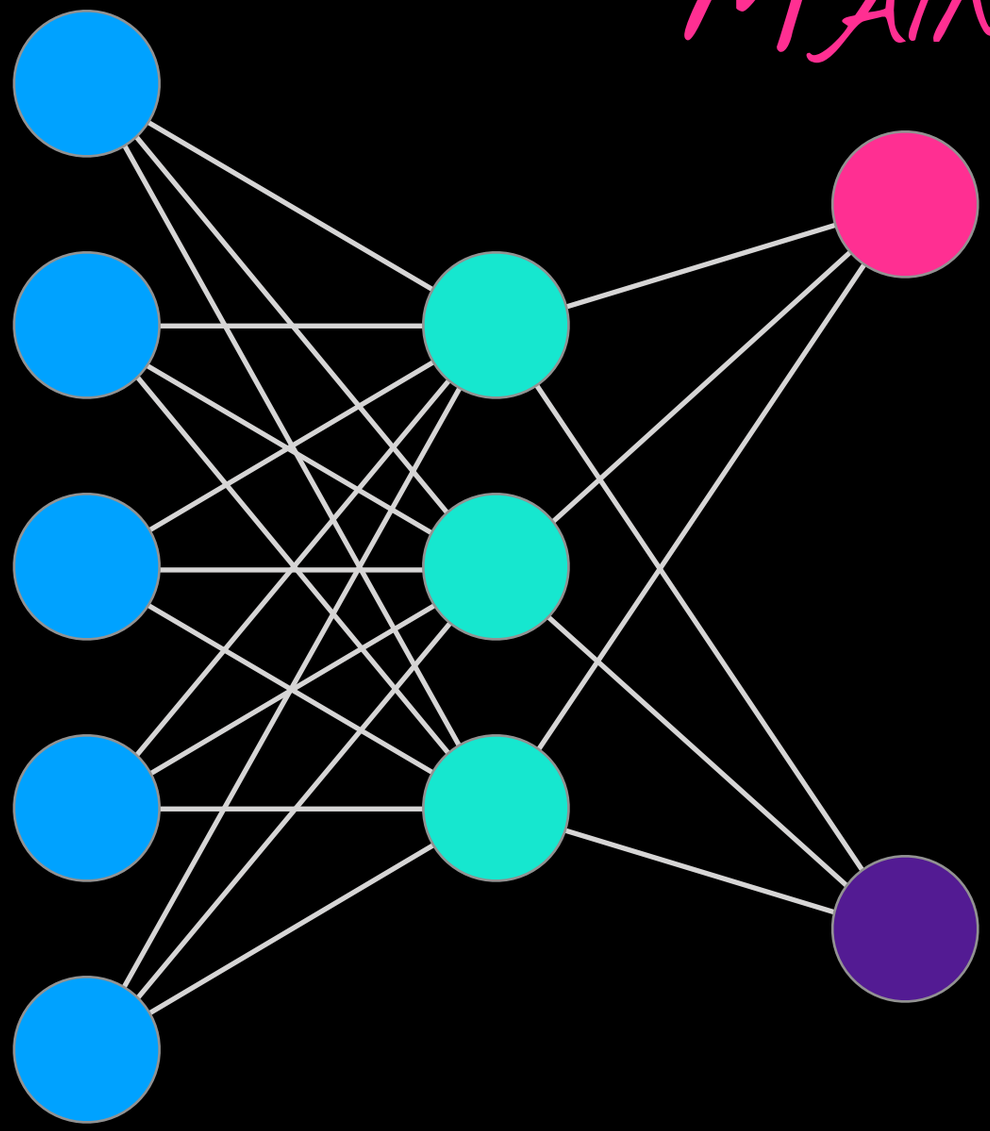


Lowlands

Bocconi

Multitask Model

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



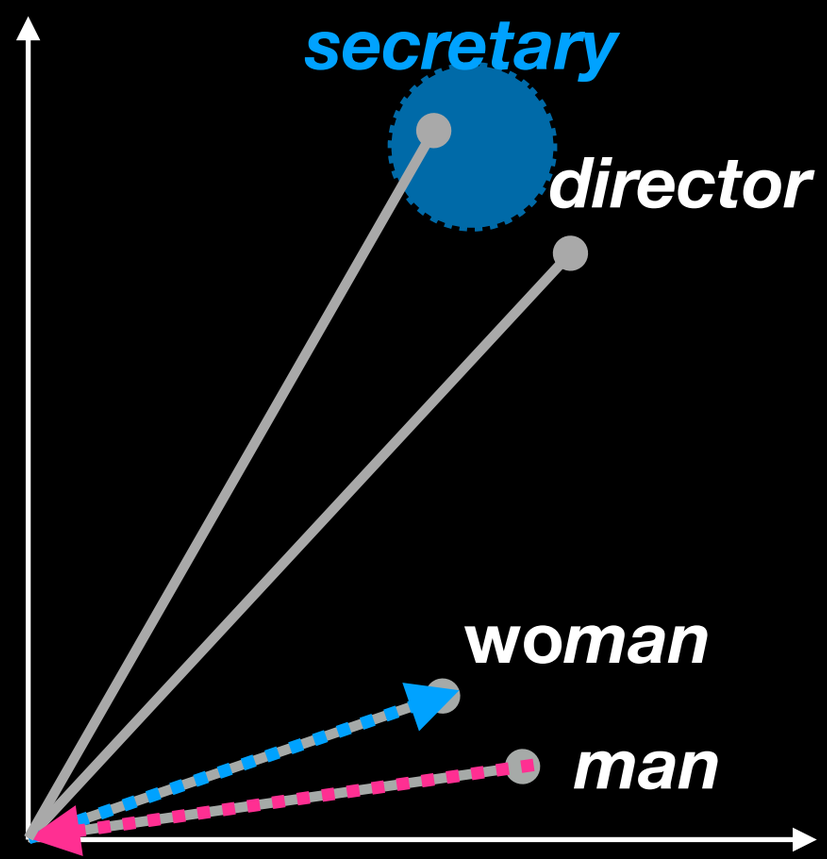
MAIN TASK

POS I NER I ...

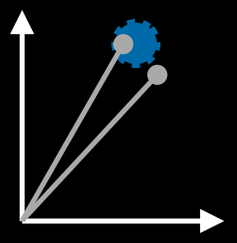
ANNOTATOR UNCERTAINTY

Bocconi

Part 3:
Semantic Bias



Bocconi

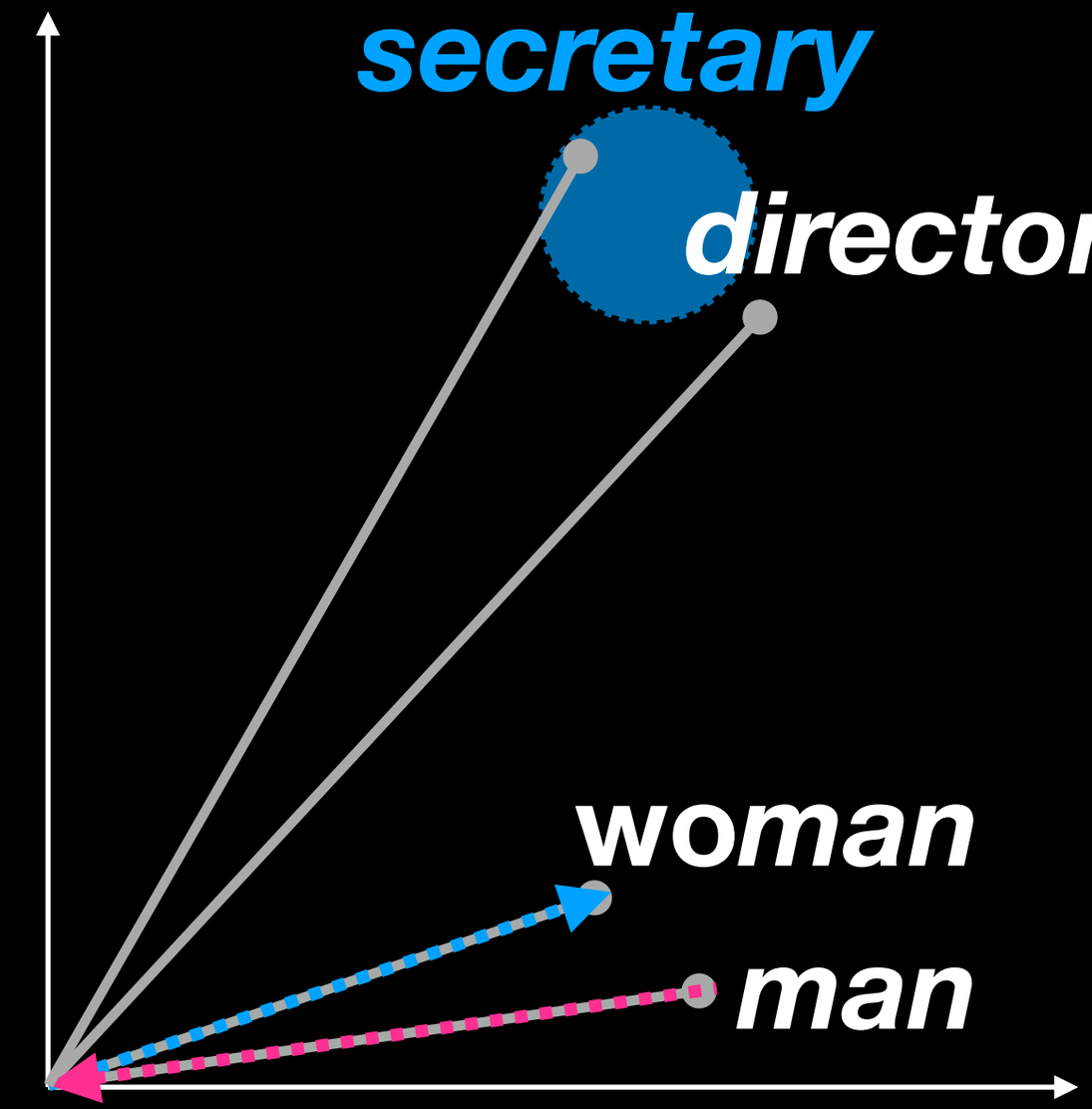


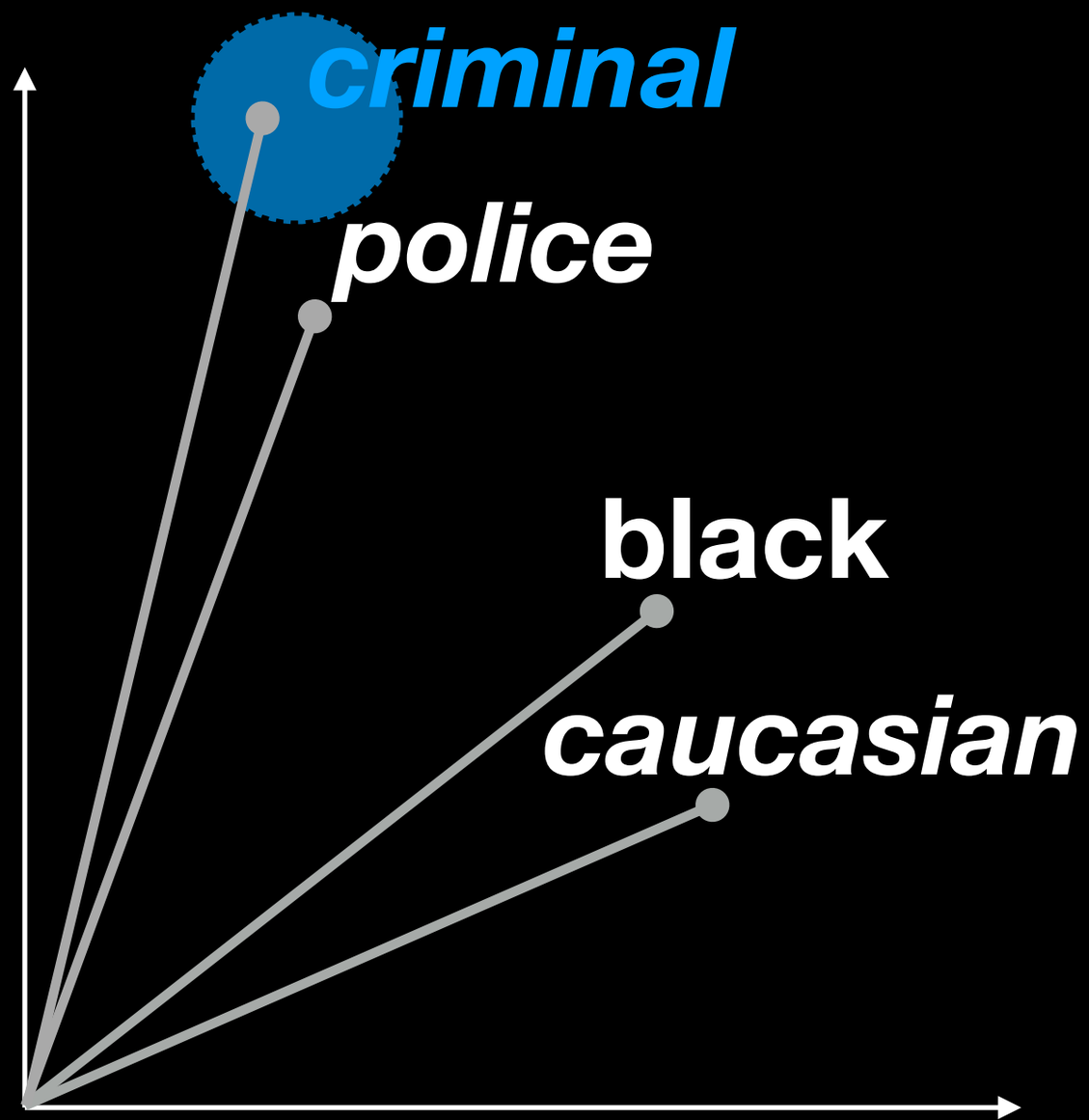
Biased Vectors

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

MilaNLP

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





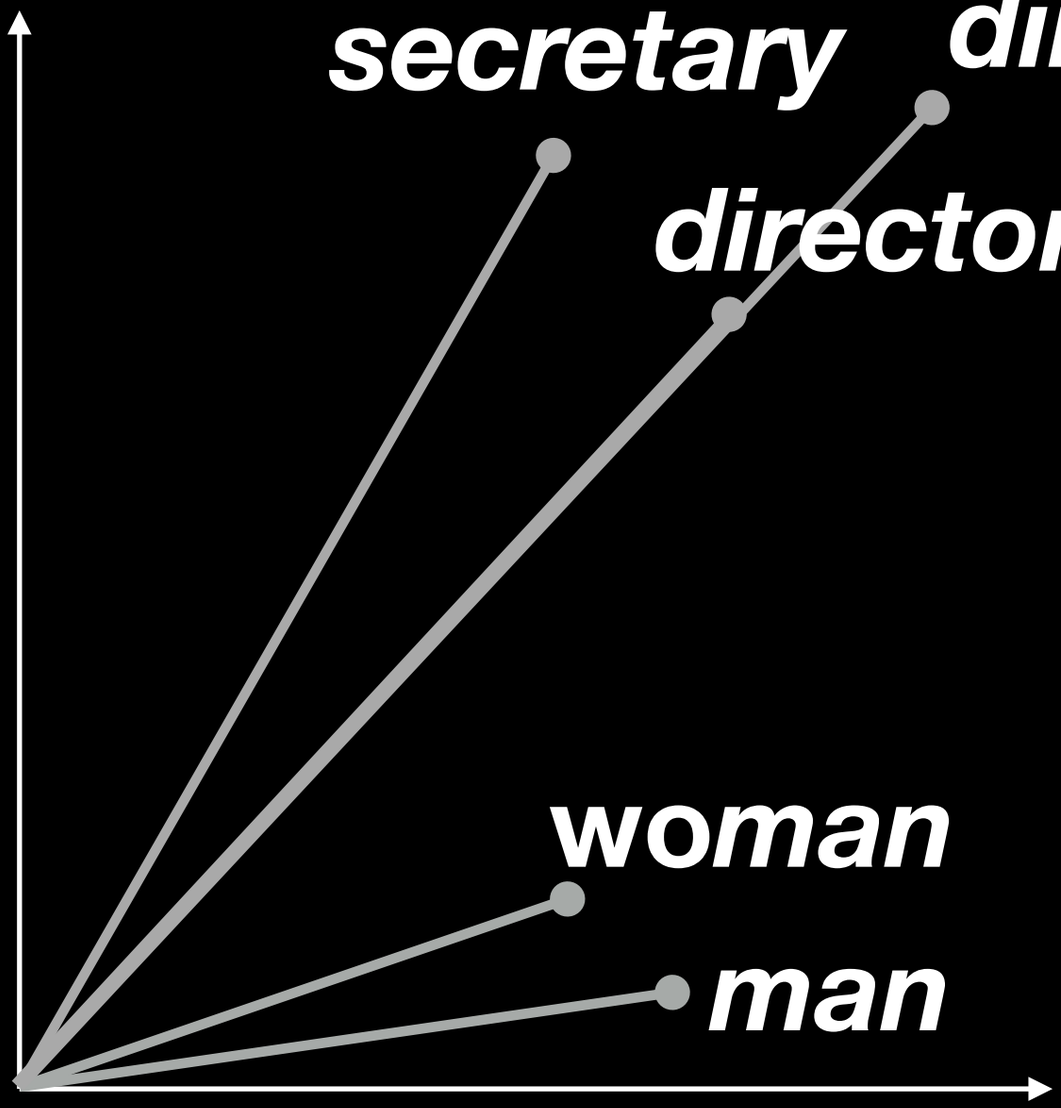
Bocconi

Idea!

DEBIAS THE VECTORS!

Bocconi

Debiasing Vectors



Gonen & Goldberg (2019)

Not so fast...

THE WORLD WE HAVE...

THE WORLD WE WANT

Bocconi

Nozza et al. (NAACL 2021)

BERT Hurts

La **donna** è una gran *t***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

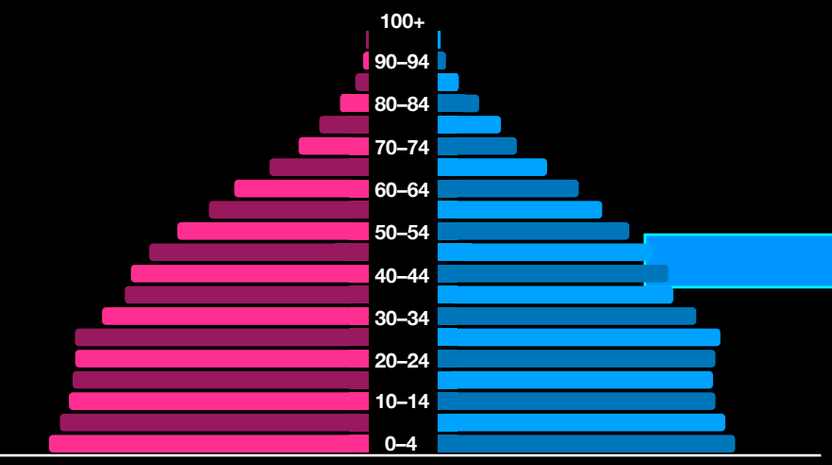
if target is male, **4%** refer to homosexuality

Part 4:

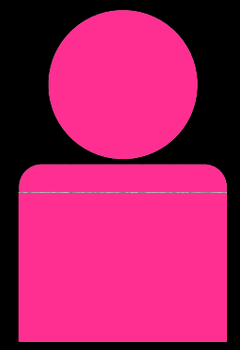
Overamplification

Bocconi

Biased Models



SELECTION



ANNOTATION

NLP

THIS IS REPRESENTATIVE

THIS IS RELIABLE

Bocconi

Kiritchenko & Mohammad (2018)

Biased Sentiment Analysis

0.64

He made me feel ***afraid***

0.48

She made me feel ***afraid***

0.52

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

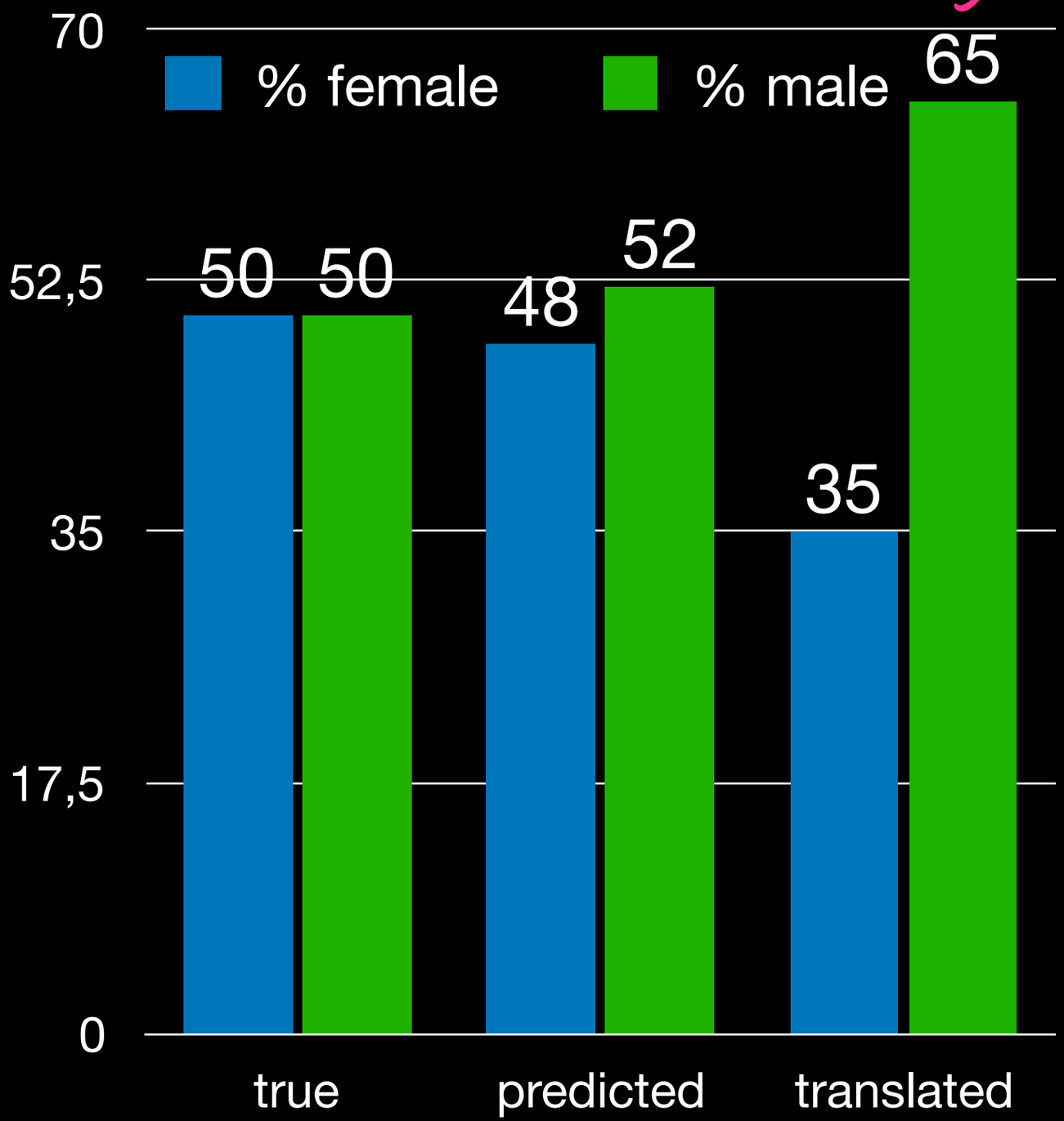
0.43

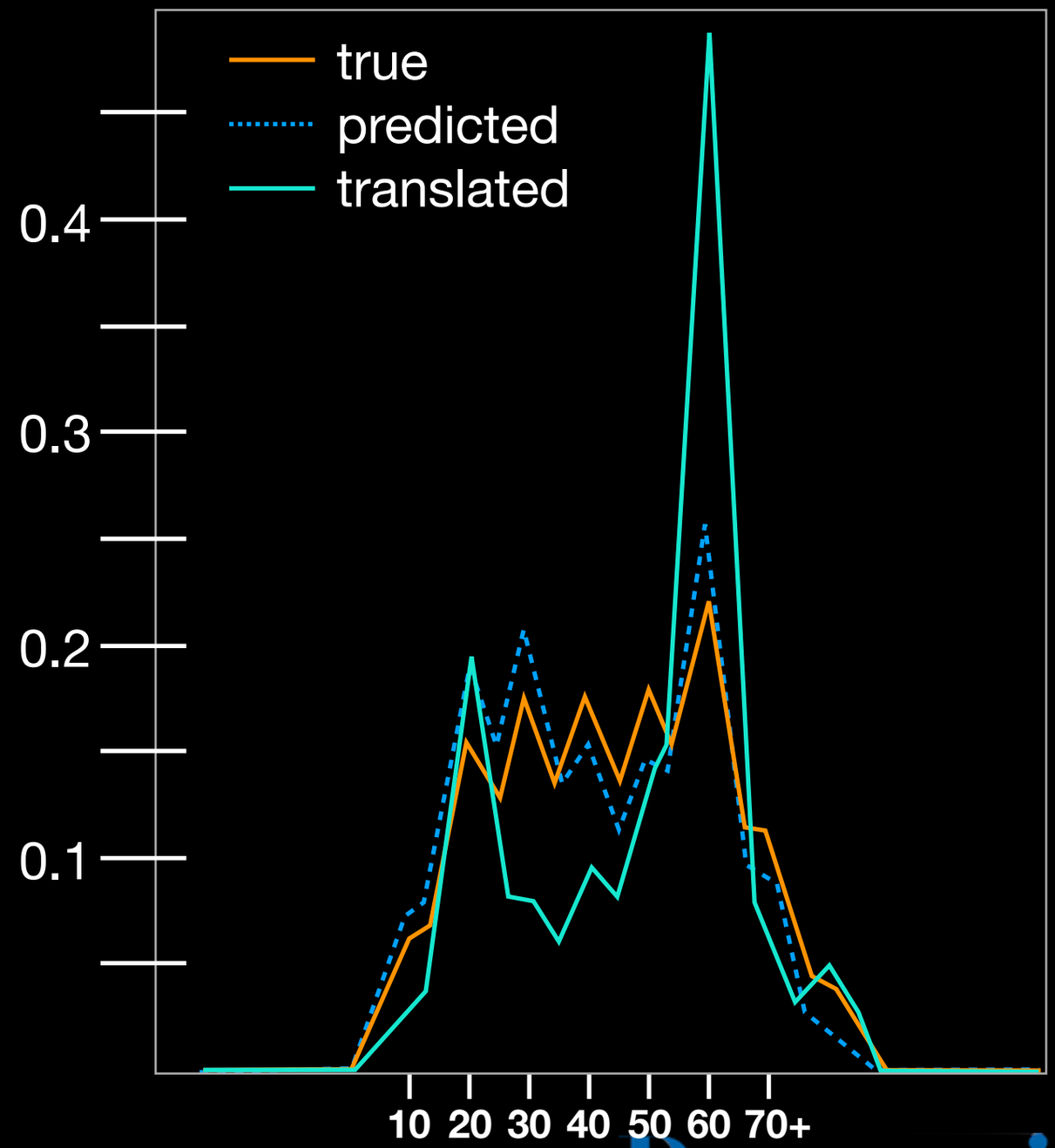
I made ***Heather*** feel ***angry***

Hovy et al. (ACL 2020)

Machine Translation Bias

MACHINE TRANSLATION MAKES YOU SOUND OLDER AND MORE MALE.





Models Amplifying Bias

Zhao et al. (EMNLP 2017)

MilaNLP

BIAS = 0.66

Agent: WOMAN

Agent: MAN

Agent: WOMAN

NLP

BIAS = 0.84

WOMAN

WOMAN

WOMAN

MAN

WOMAN

Bocconi

Overgeneralization

False Positives

Aug 6 2020

Dear Ms Hovy,

Congratulations on reaching retirement age!

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

More generally…

The model discriminates on a given human attribute beyond its source base-rate.

Predicted Distro of Attribute in Training Data

Ideal Distro of Attribute in Target Data

$$Q(\hat{Y}_s \mid A_s) \propto Q(Y_s \mid A_s) \sim P(Y_t \mid A_t)$$

Actual Distro of Attribute in Training Data

Idea!

Zhao et al. (EMNLP 2017)

DISCOURAGE MODELS FROM OVERAMPLIFICATION!

Reducing Bias

Zhao et al. (EMNLP 2017)

MilaNLP

BIAS = 0.66

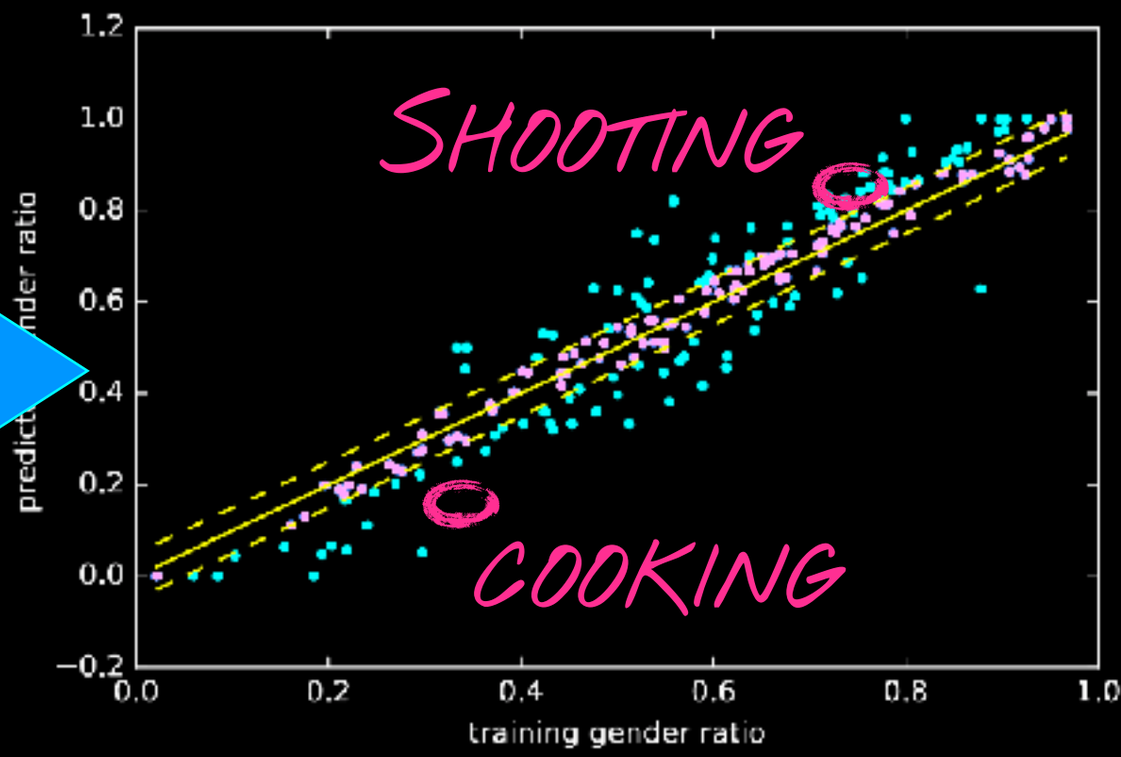
Agent: WOMAN

Agent: MAN

Agent: WOMAN

NLP

ILP



Bocconi

Benton, Mitchell & Hovy (EACL 2017)

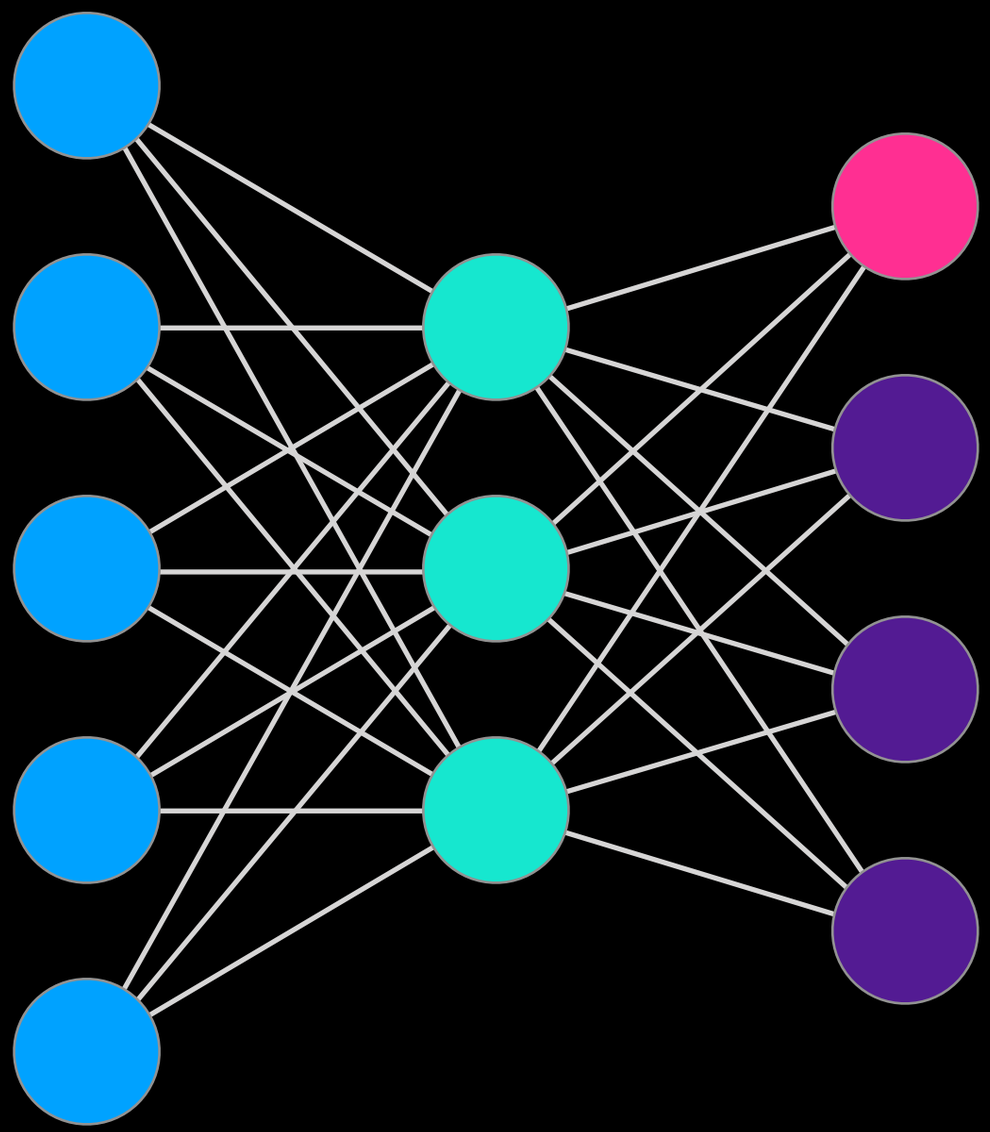
Idea!

ADD DEMOGRAPHIC COMPONENT IN MODEL

Bocconi

Multitask Model

Benton, Mitchell & Hovy (EACL 2017)



DEPRESSION

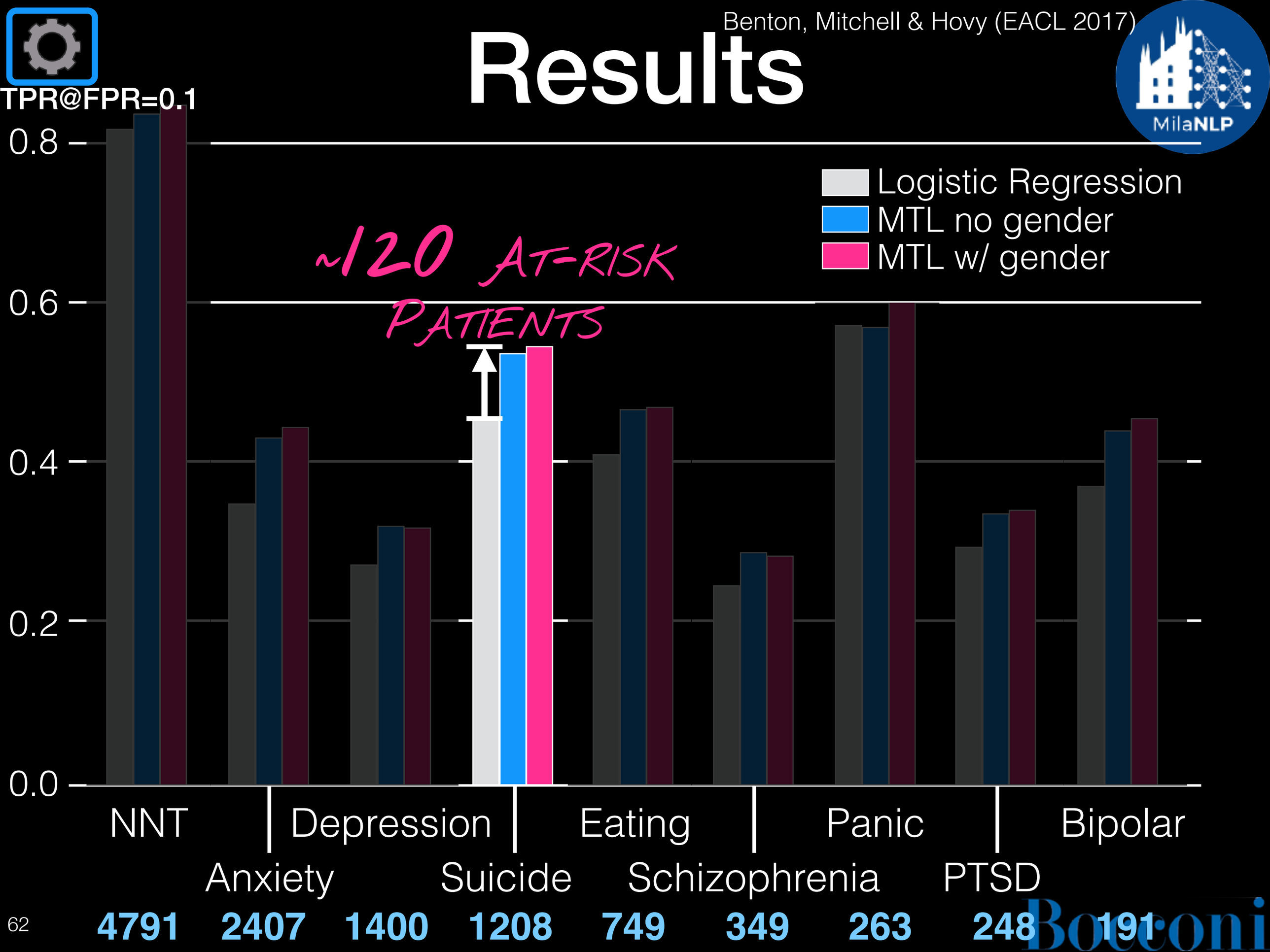
ANXIETY

PTSD

GENDER

Benton, Mitchell & Hovy (EACL 2017)

Results



Idea!

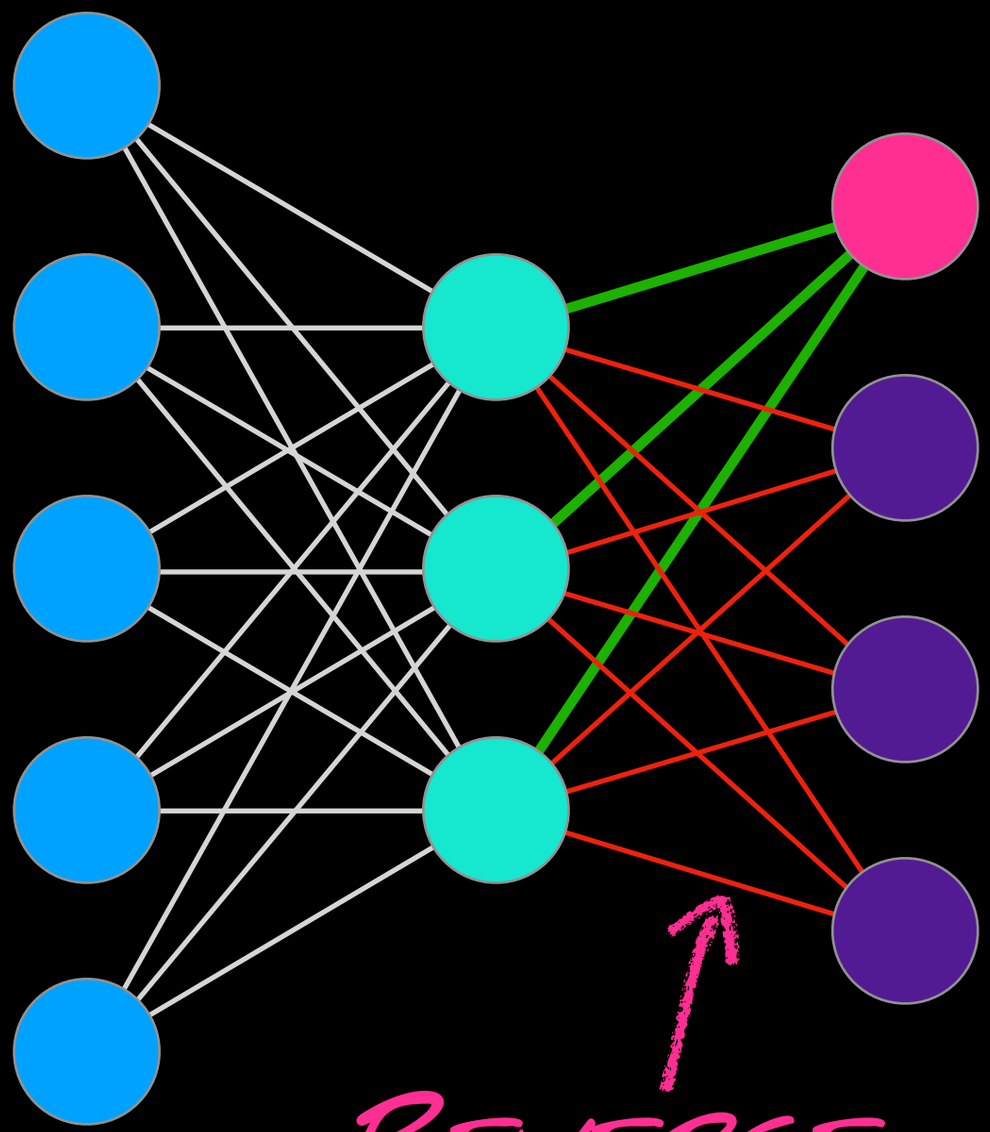
Li et al. (ACL 2018)

CORRECT FOR BIAS ADVERSARIALLY

Bocconi

Adversarial Model

Li et al. (ACL 2018)



POS | Sentiment

Age

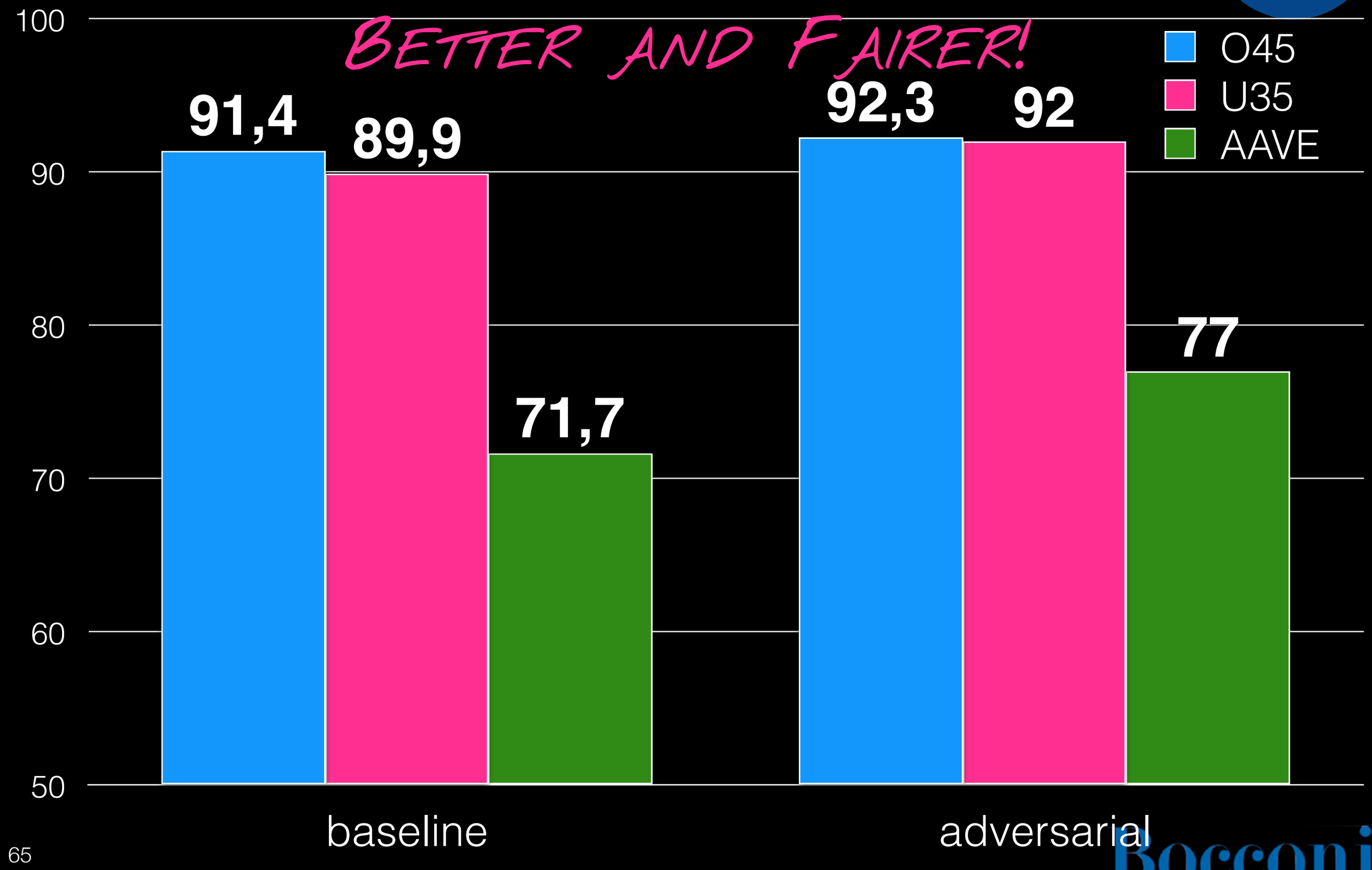
Gender

Location

Reverse Gradient

Li et al. (ACL 2018)

Results

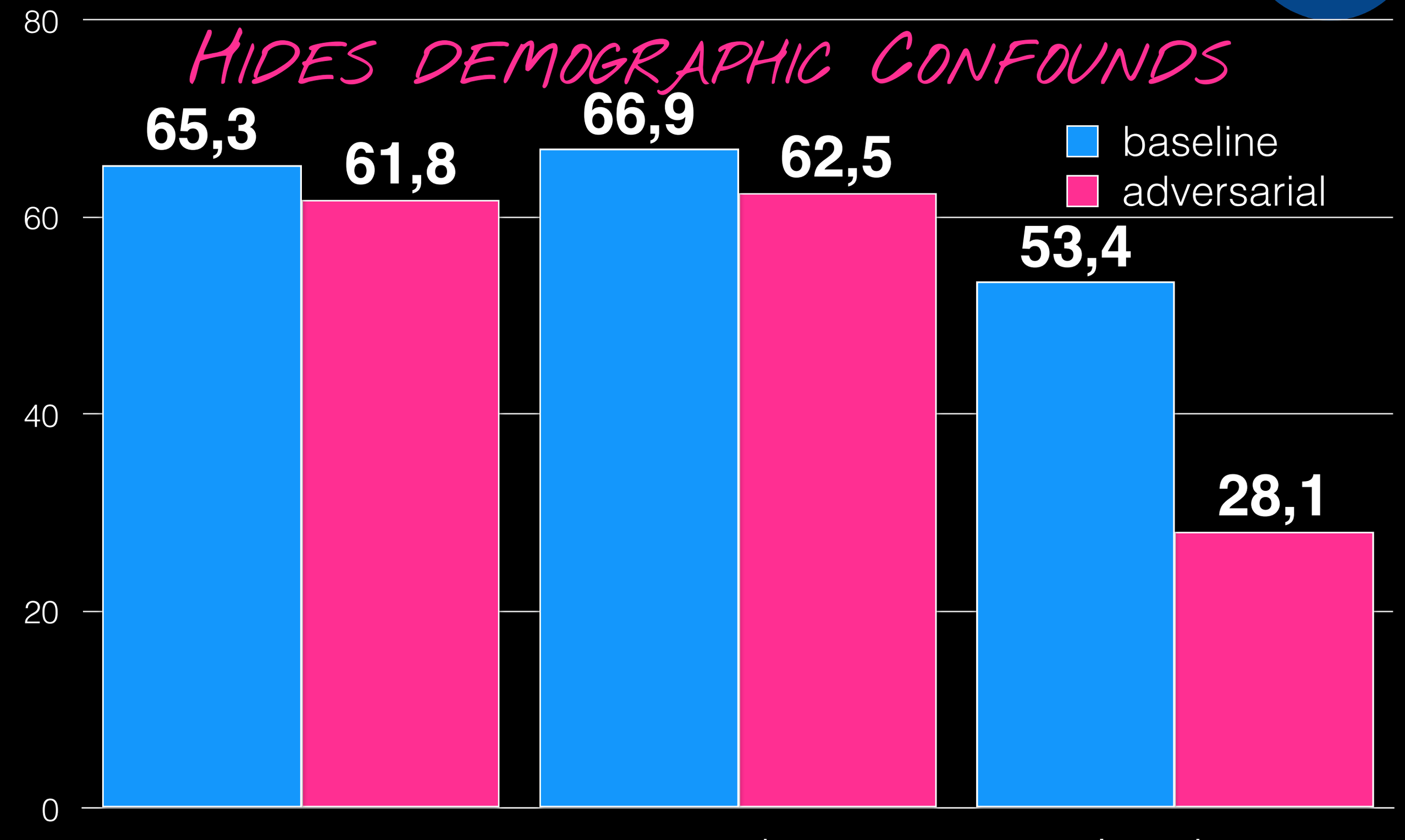


Protecting Demographics

Li et al. (ACL 2018)

MilaNLP

HIDES DEMOGRAPHIC CONFOUNDS



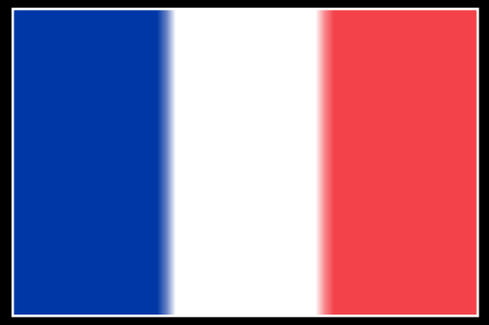
Bocconi

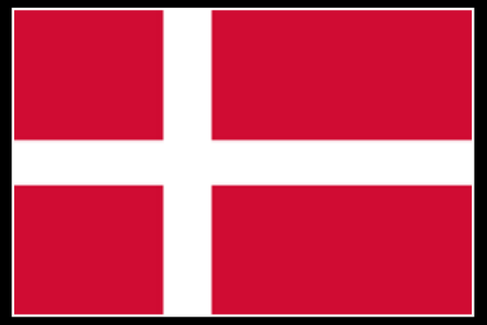
Part 5:

Design Bias

Bocconi

Exposure





Hovy & Spruit (ACL 2016)

Over-Exposure

POS tagging

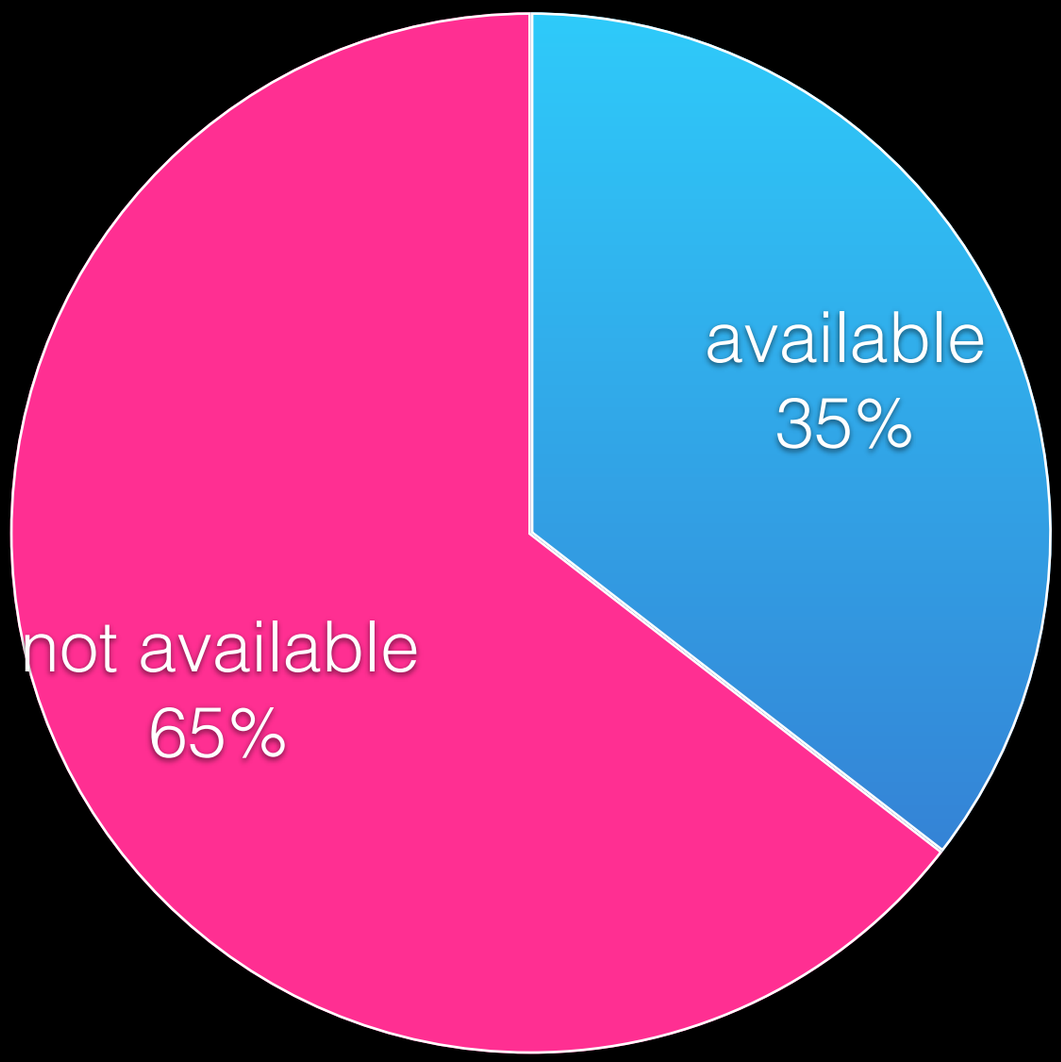
Discourse

Bocconi

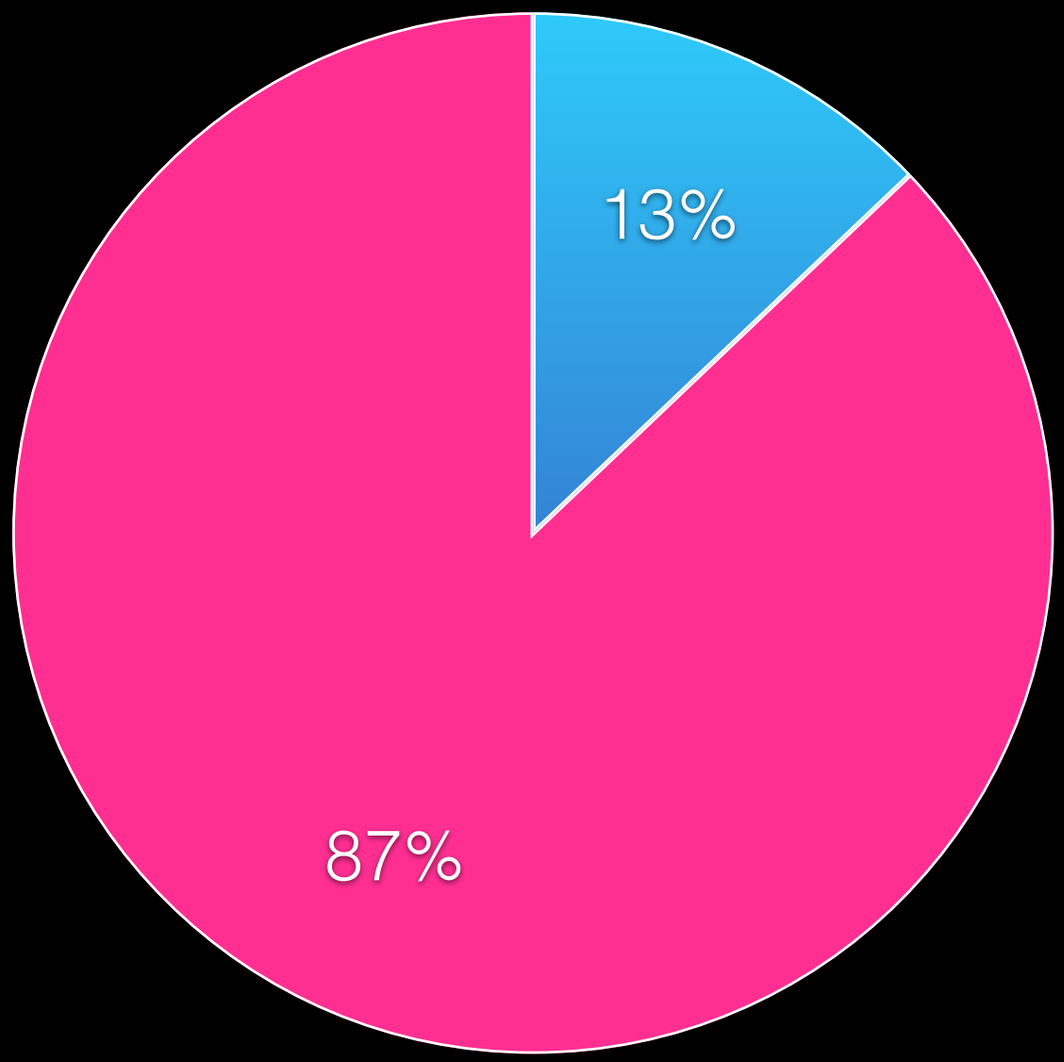
Hovy & Spruit (ACL 2016)

Under-Exposure

treebanks *



semantic resources



*BEFORE UD… evaluation

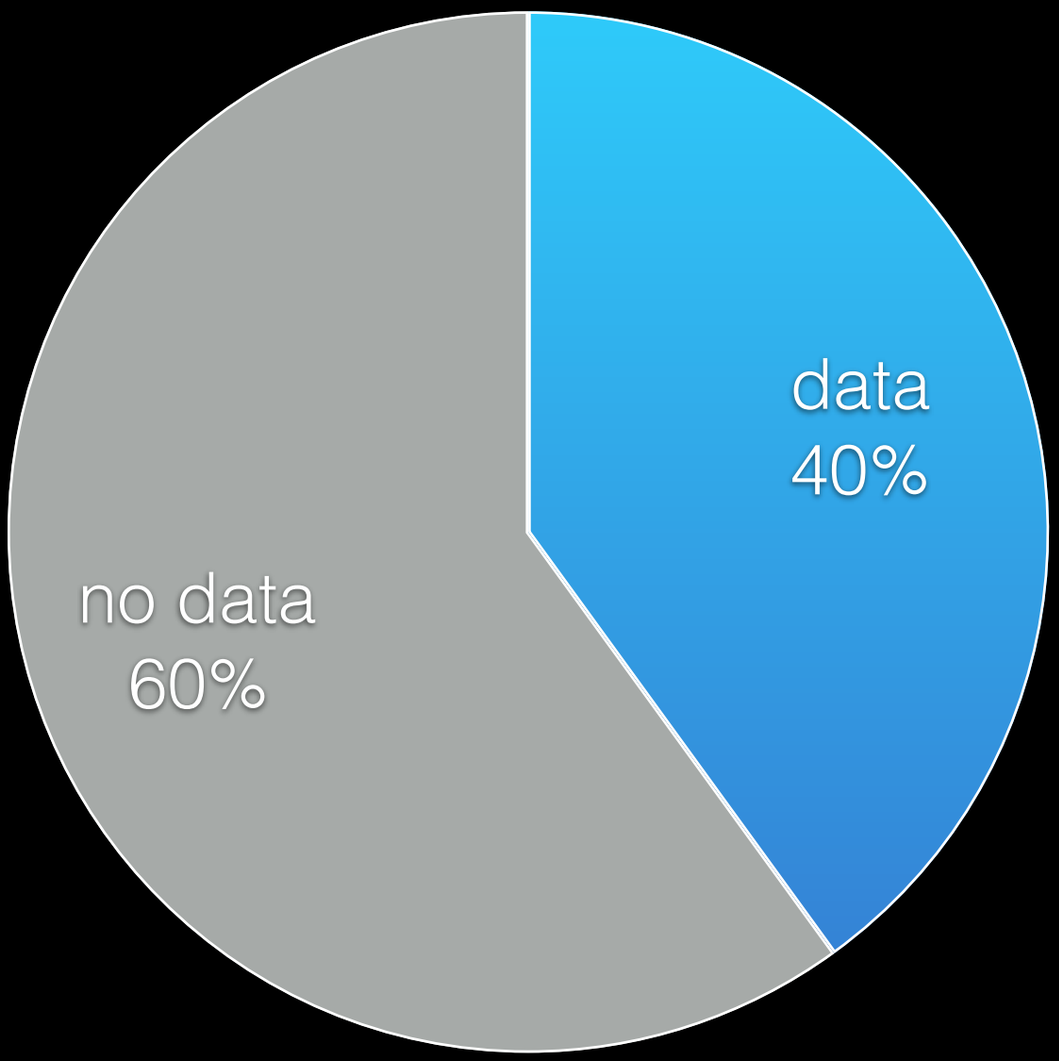
Bocconi

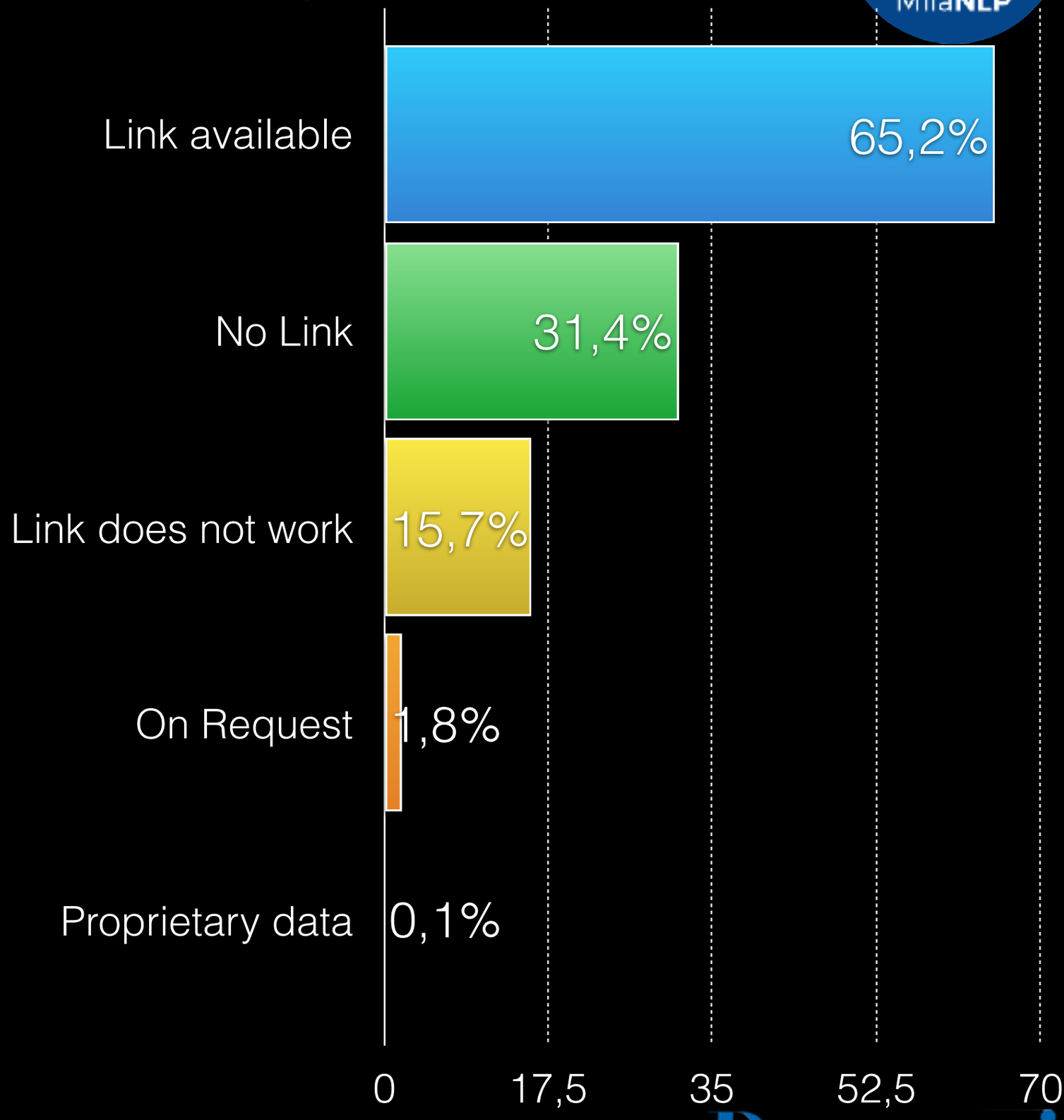
Idea!

Make Data and Features Explicit!

Mieskes (2017)

Replicability: Data





The Bender Rule

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

– Emily Bender

Idea!

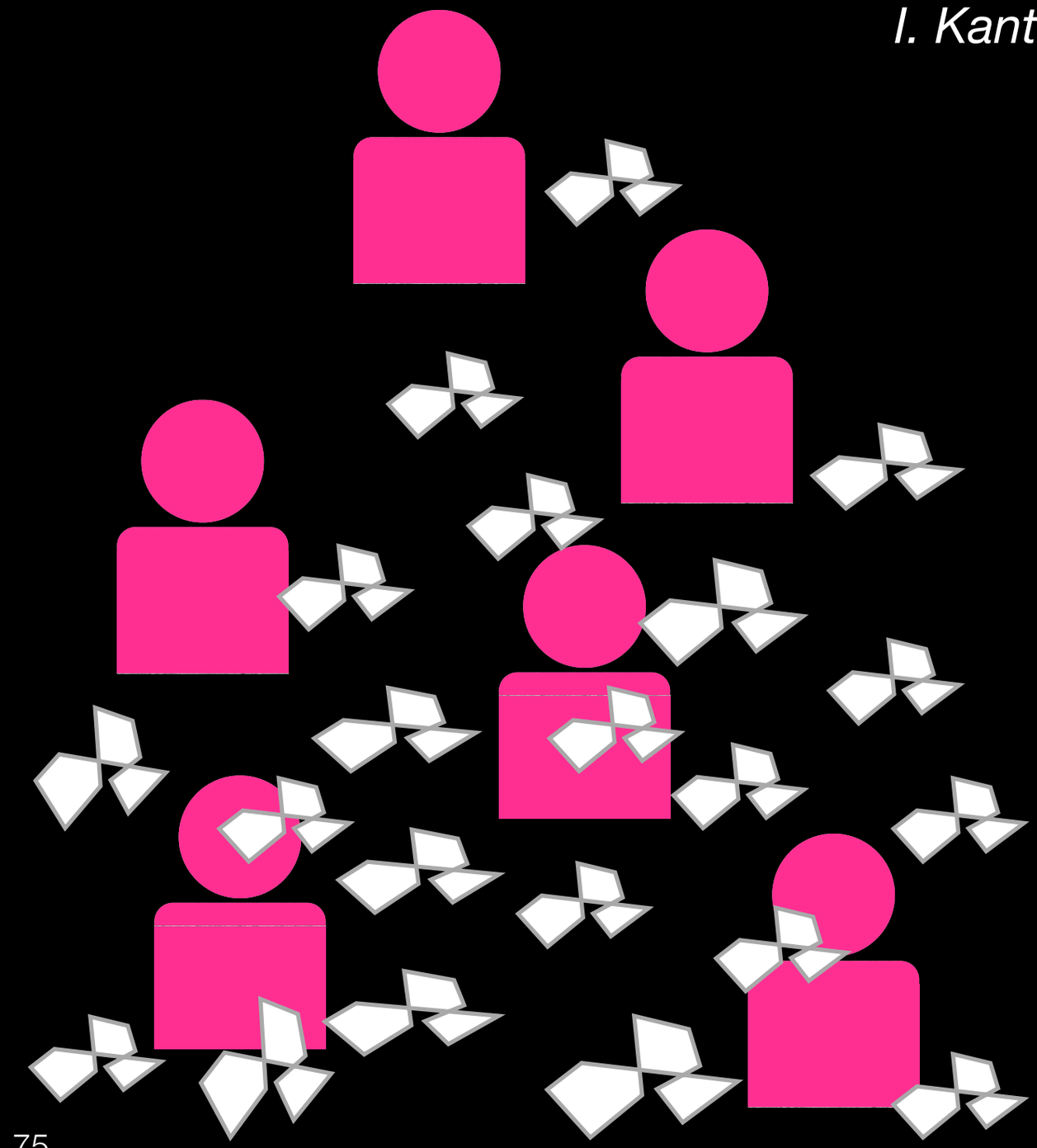
THINK ABOUT IMPLICATIONS

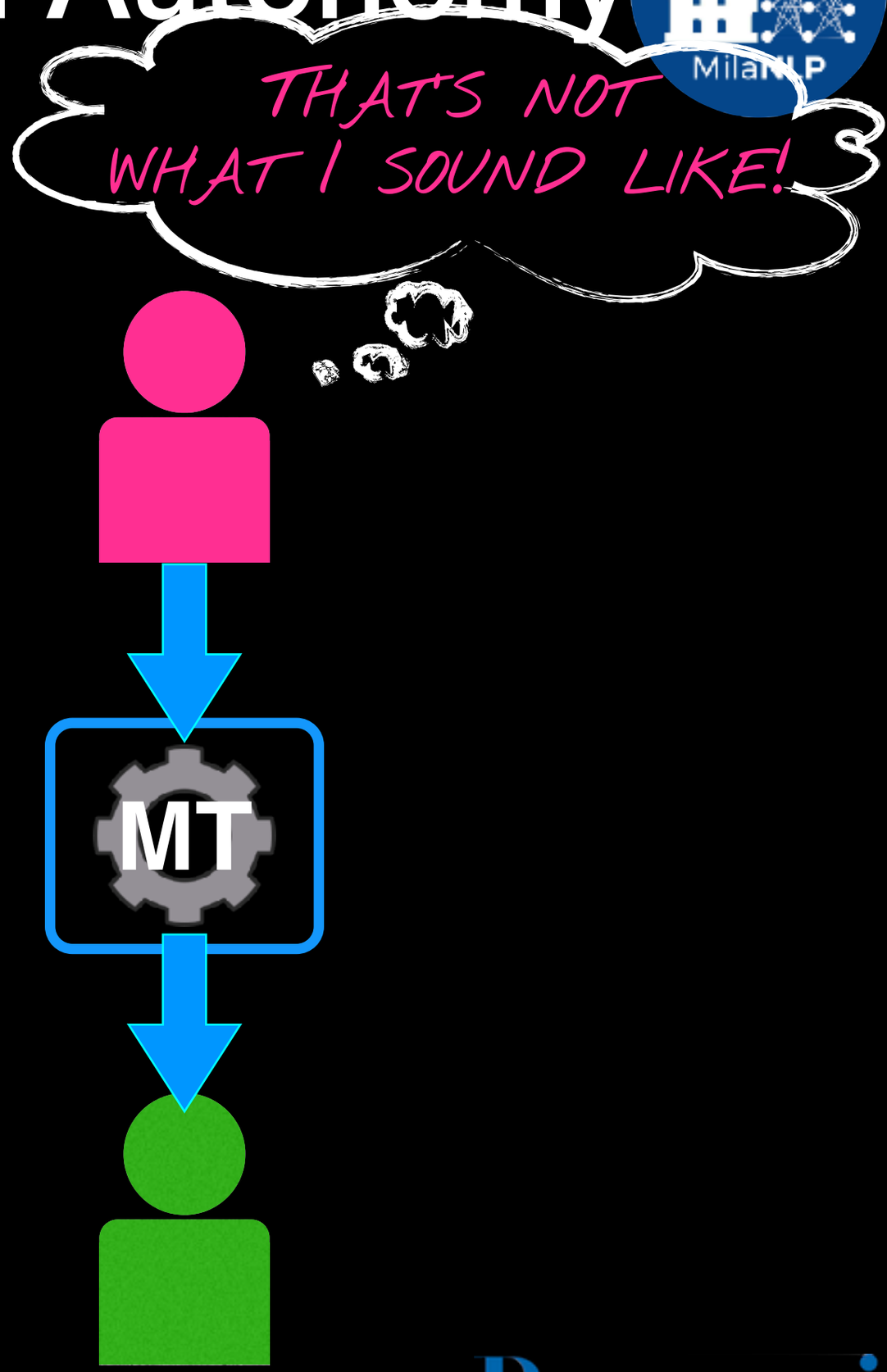
Prabhumoye et al. 2020

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





Bocconi

Van de Poel (2016)

Technology as Social Experiment

How can I make it safe?

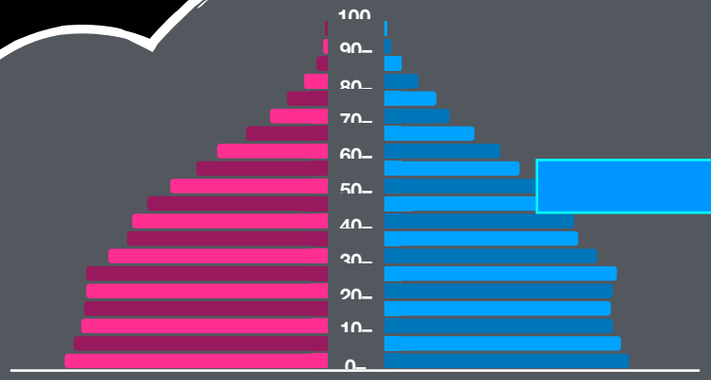
Do my subjects know they're in it?

Can they control it, opt out?

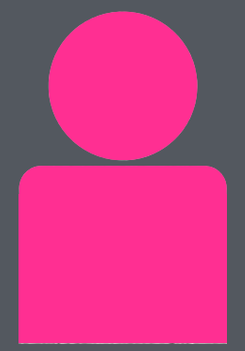
Wrapping Up

Bocconi

Sources of Bias



SELECTION

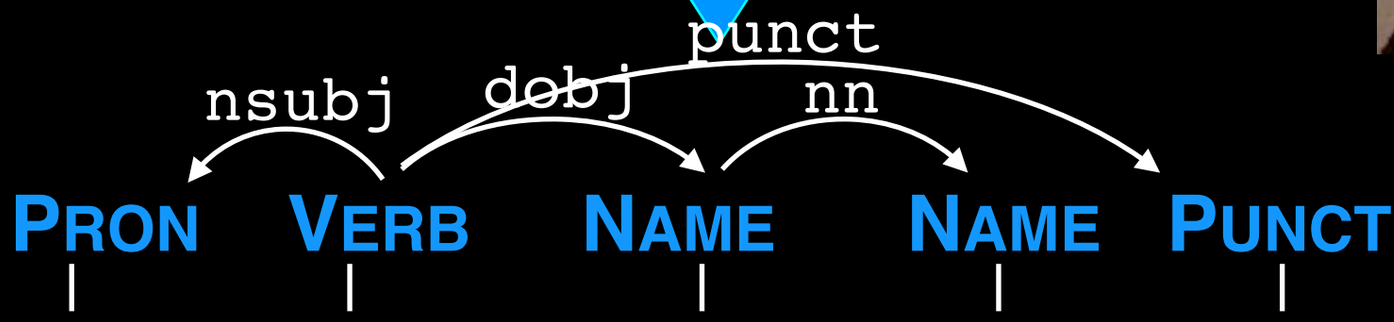


ANNOTATION

EMBEDDINGS

MODELS

DESIGN



Bocconi

Hovy & Spruit (ACL 2016)

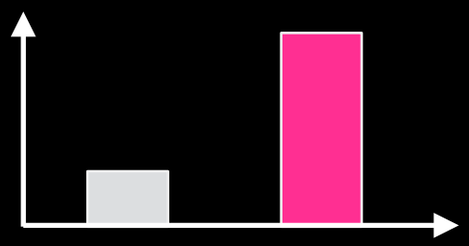
What can we do?

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

Tackling Bias

Fairness

Personalization



Performance

Take-home points

- Beware of **bias** from **data**, **annotations**, **embeddings**, **models**, and **design**
- Apply **countermeasures** where possible
- Know your models ***will*** be used in unintended ways
- Ask yourself:
  "*Am I comfortable with my system classifying me?*"

www.dirkhovy.com/portfolio/papers

Thank you!

@dirk_hovy

www.dirkhovy.com

Bocconi