Hidden Biases
Ethical Issues in NLP, and What to Do about Them

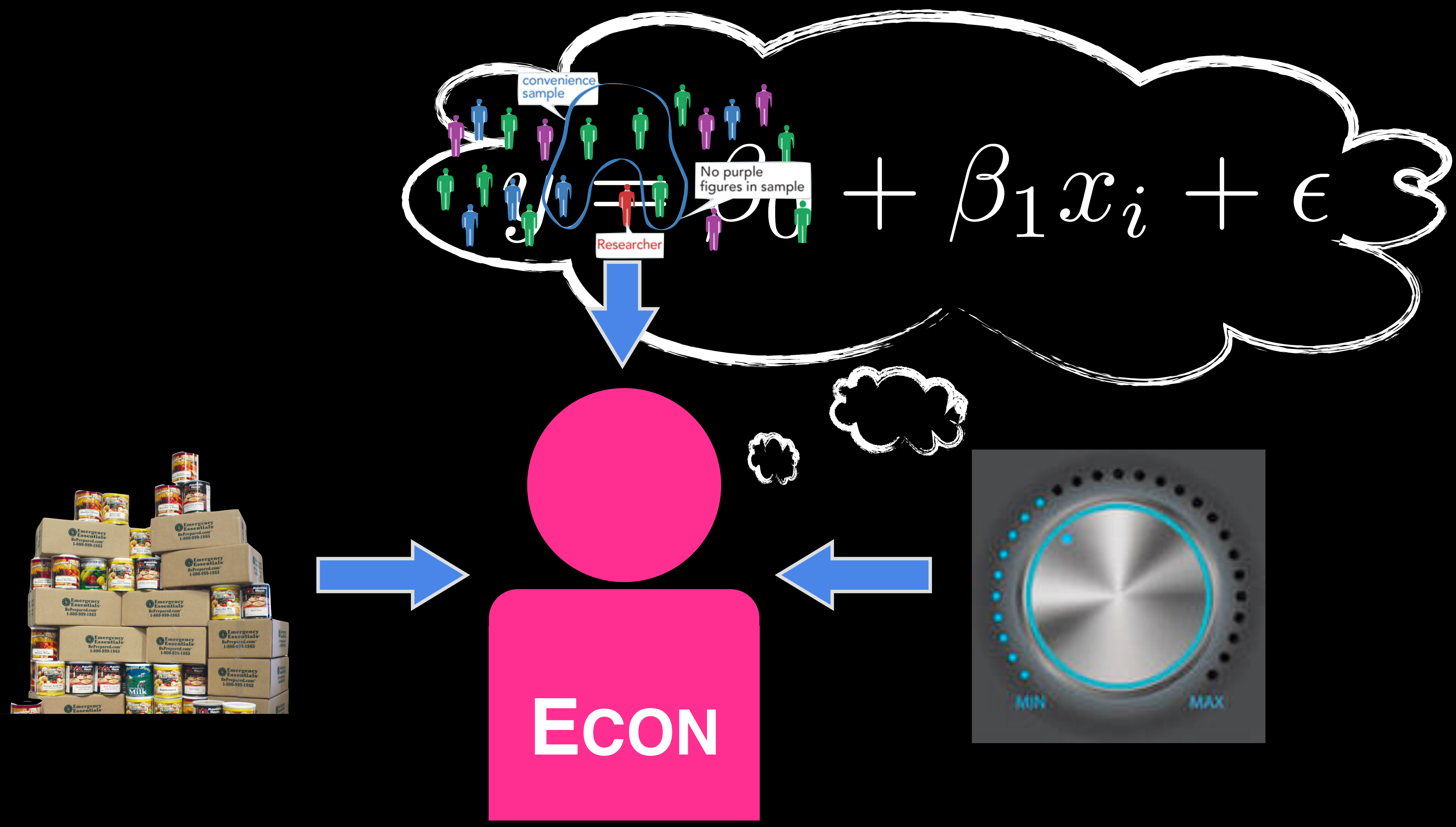
Dirk Hovy

dirk.hovy@unibocconi.it

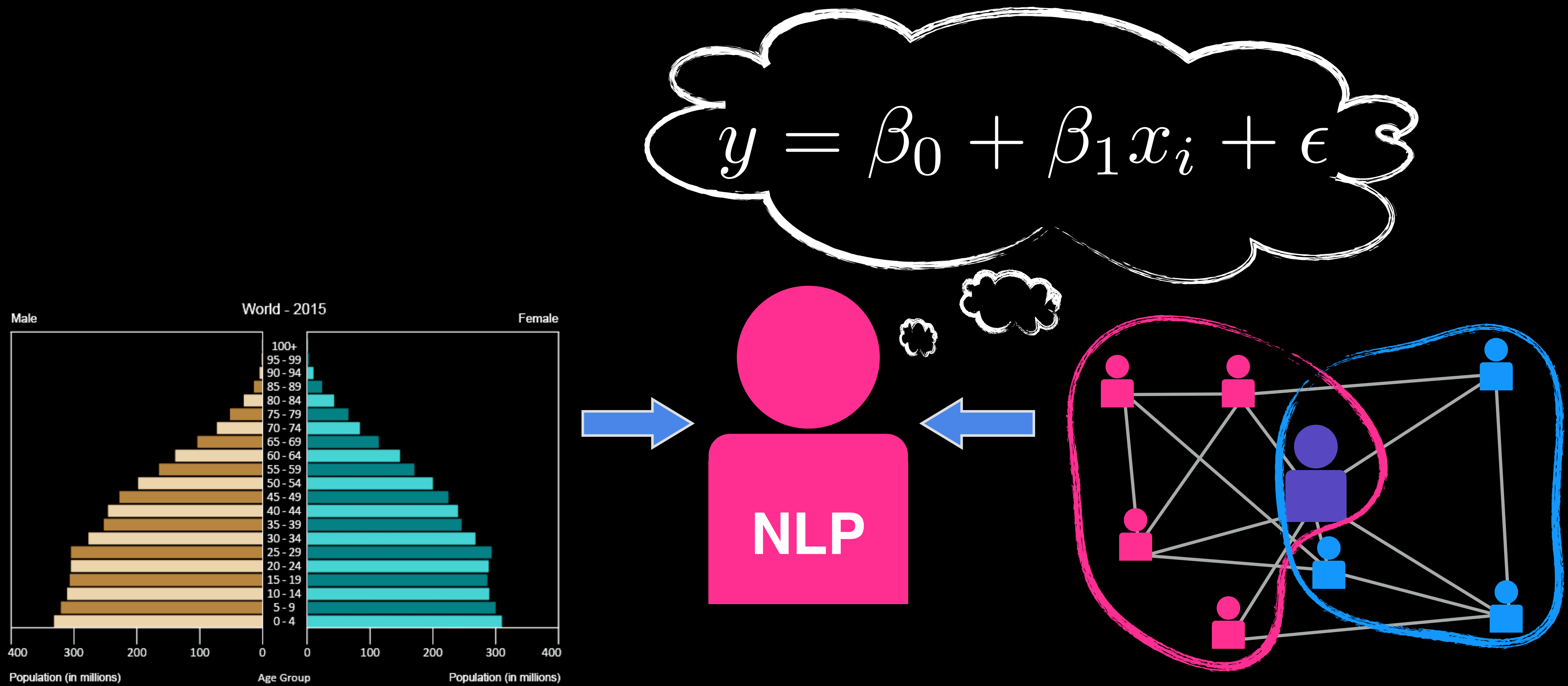
@dirk_hovy

Bocconi

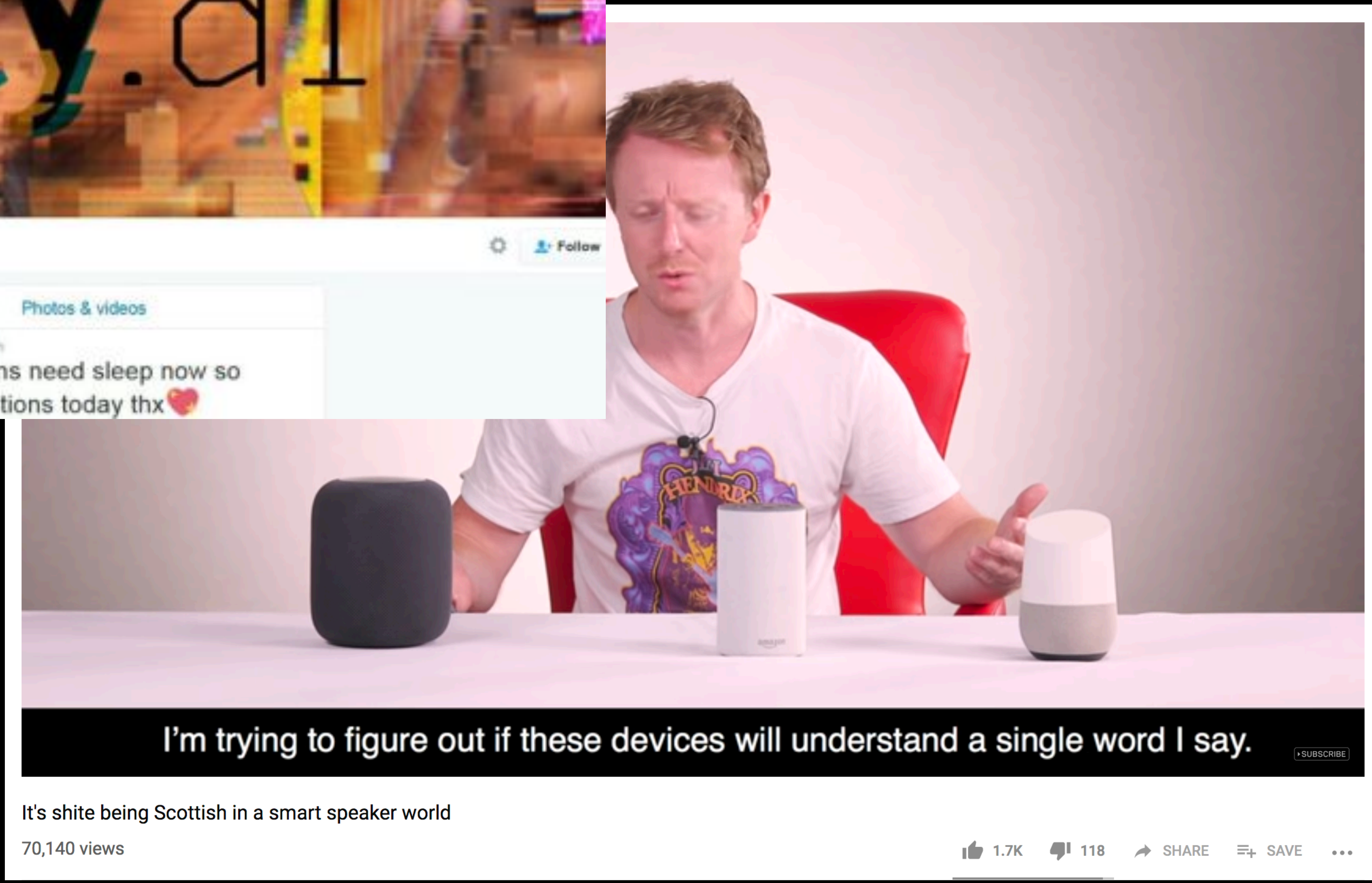
A Limited View



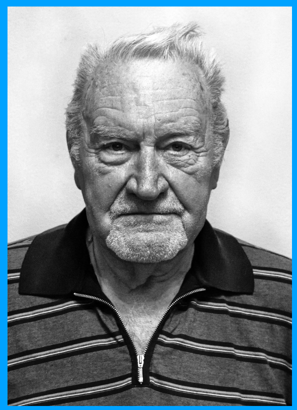
Language as Information



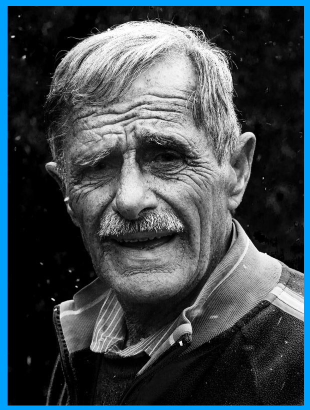
Biased Language Systems



Language Biases


Example 1


Example 2


Example N

I don't understand you...

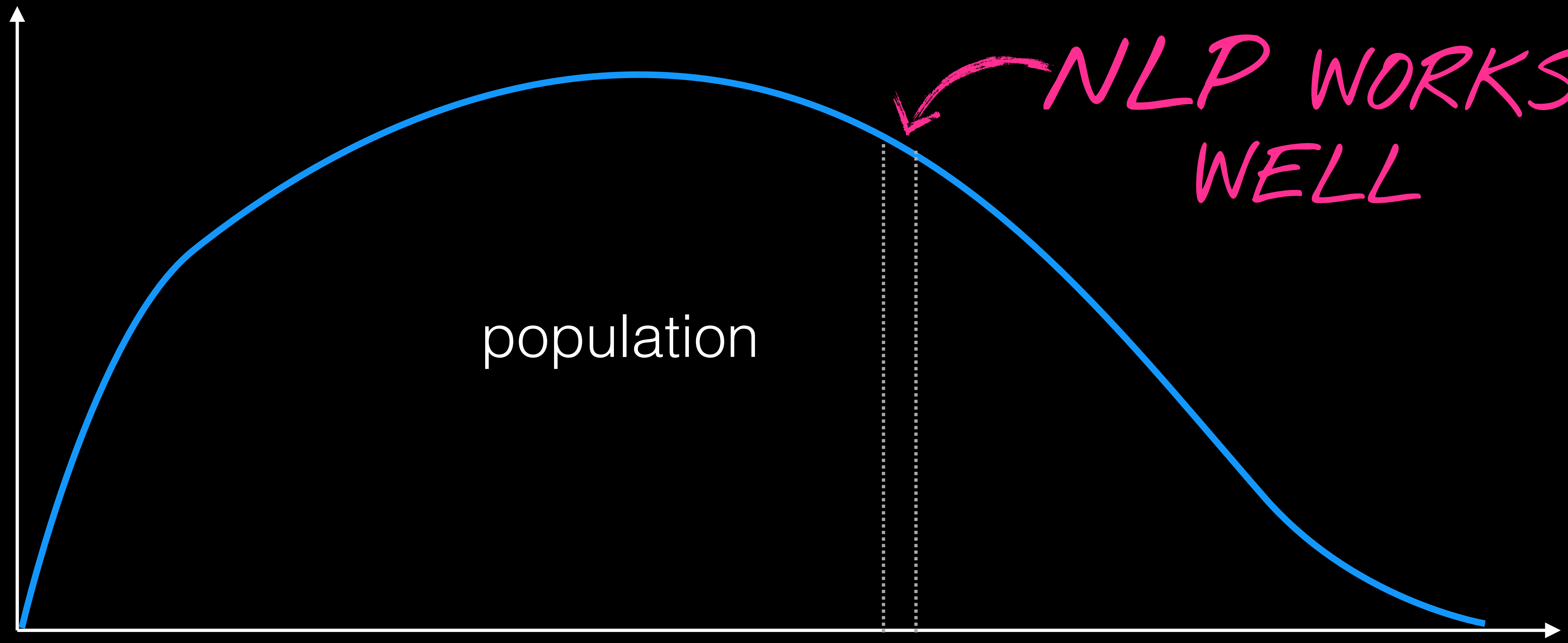
System

Hello, computer

Shite...

Bocconi

The Consequences



Bocconi

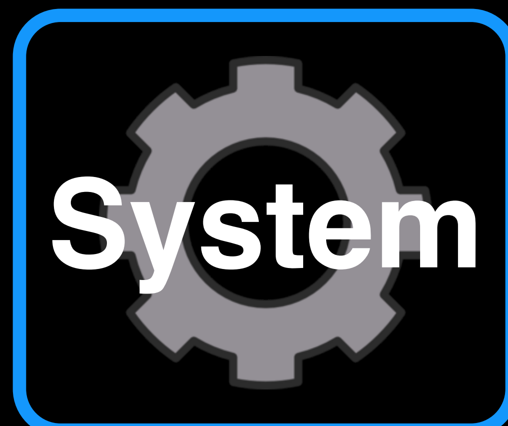
Solutions?

Symptoms

Example 1

Example 2

Example N



Cause

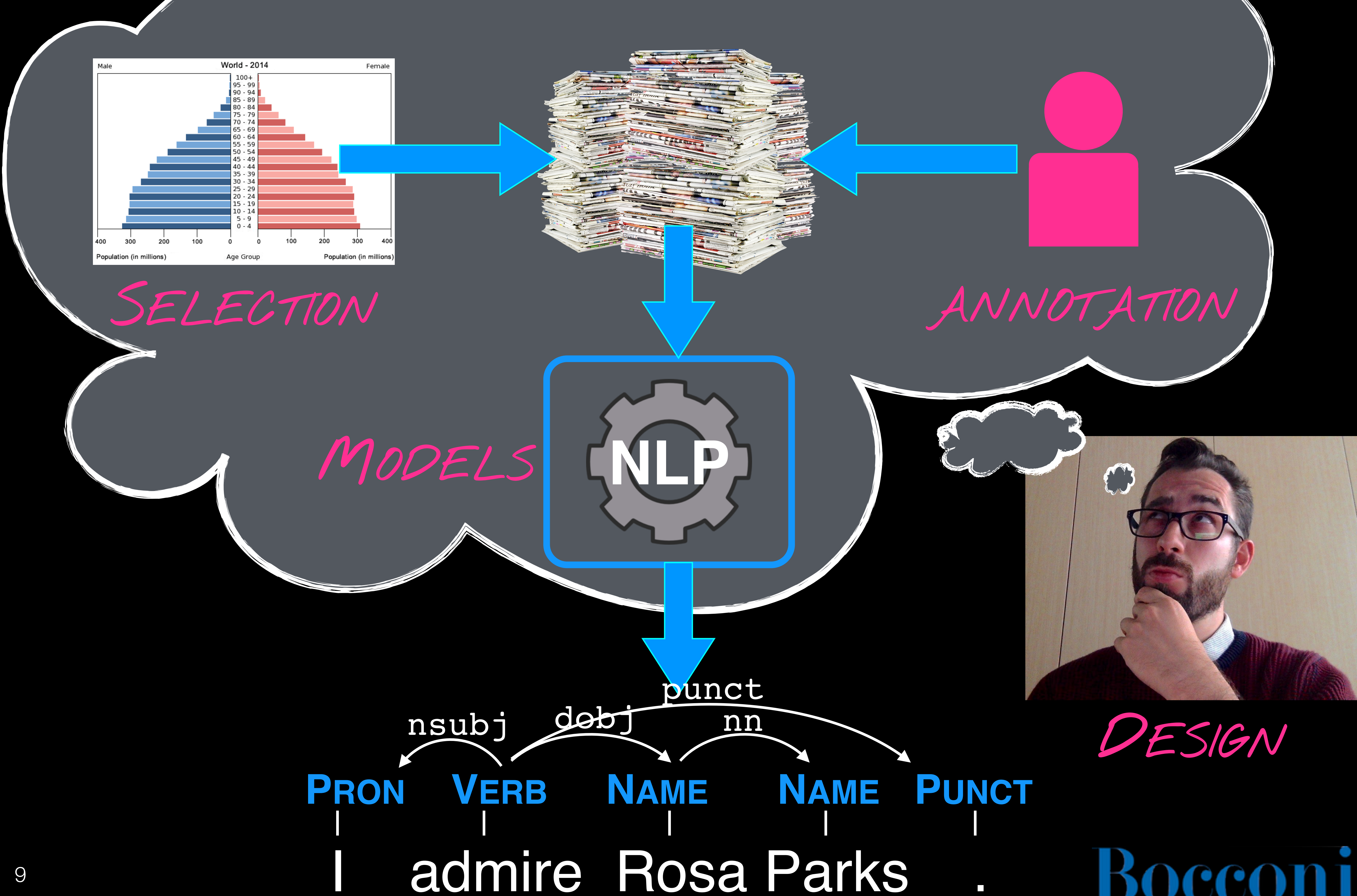
Bocconi

Goals for Today

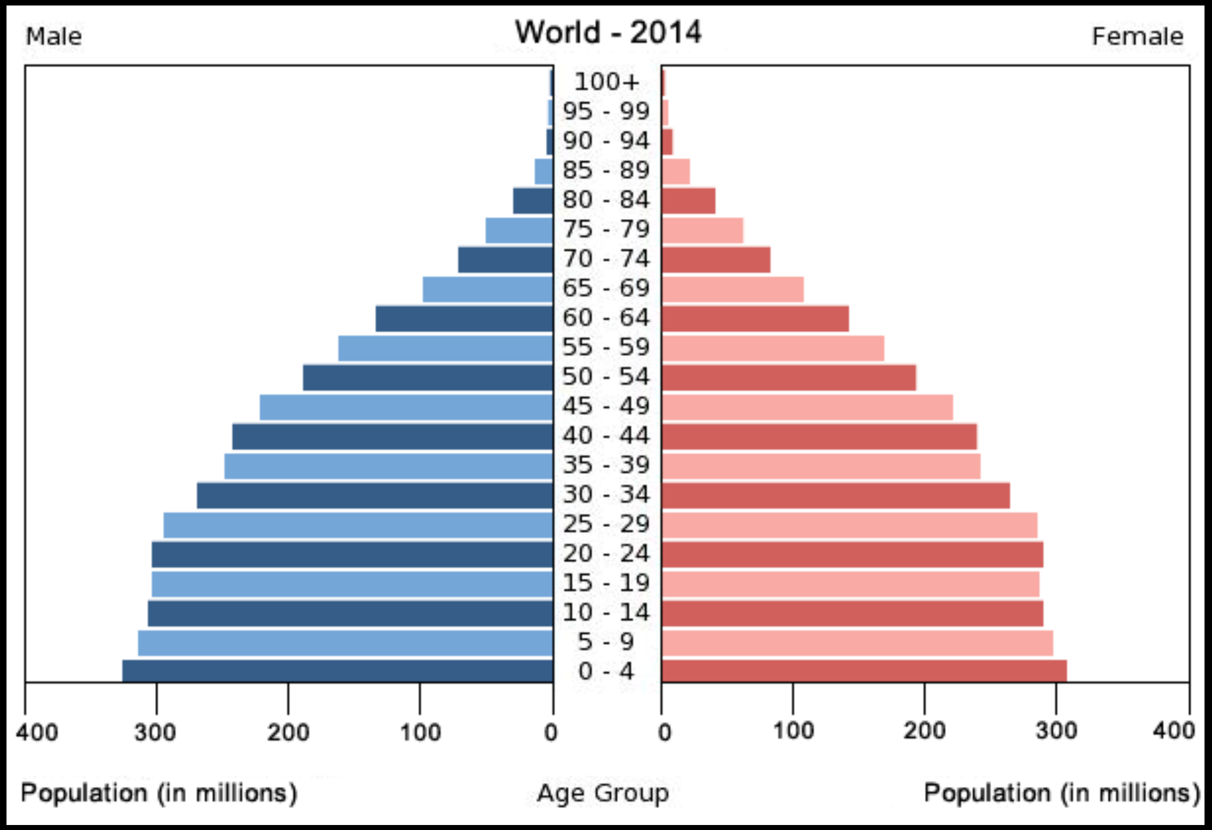
- Point out potential **ethical issues** in NLP
- Introduce 4 **sources of bias**
- Discuss **counter measures**

Bocconi

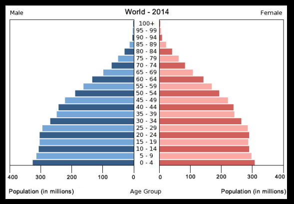
Sources of Bias



Part 1:
Data Bias

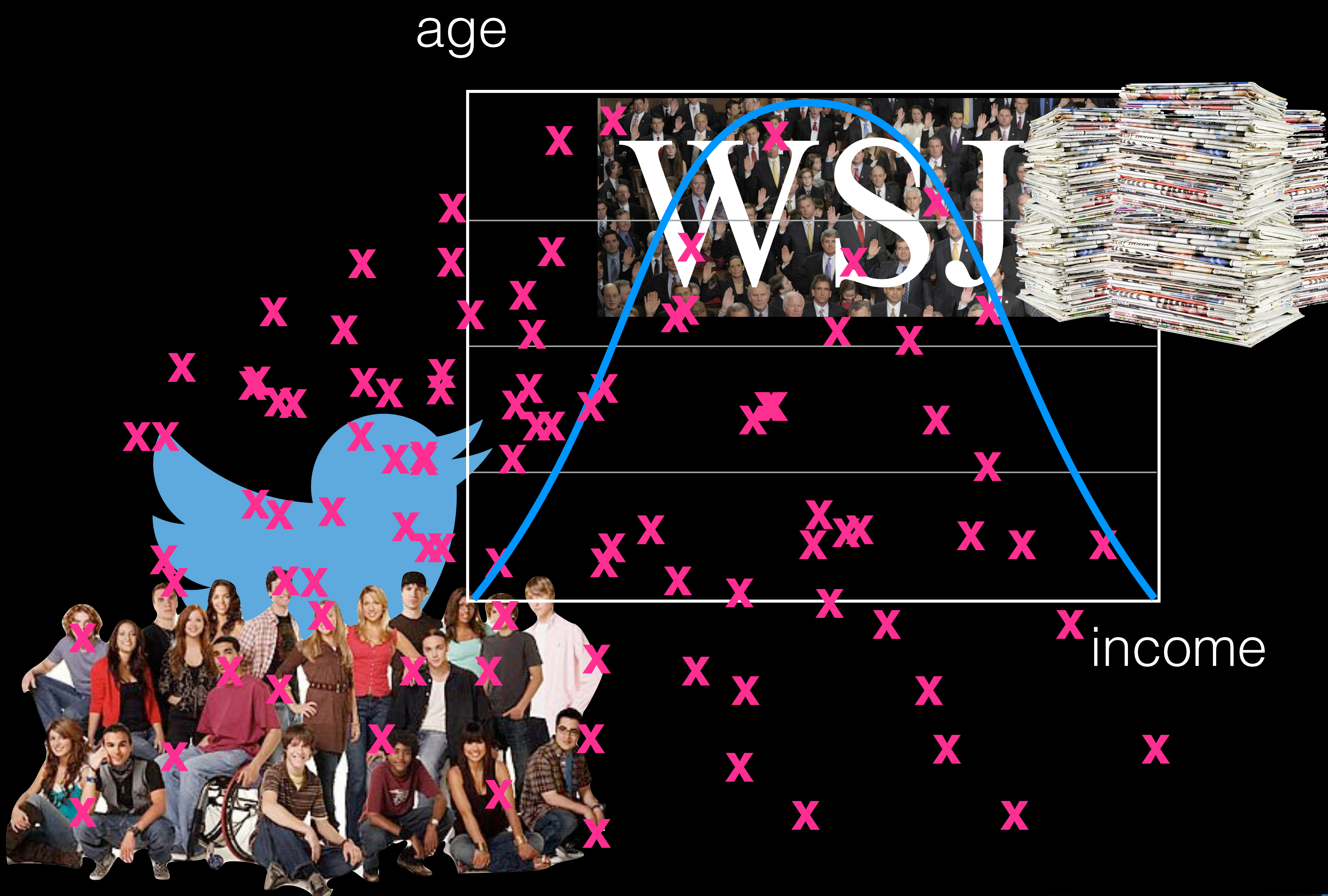


Bocconi

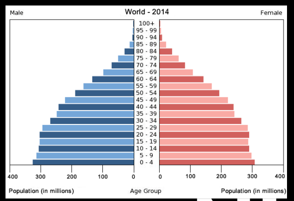


Distributions

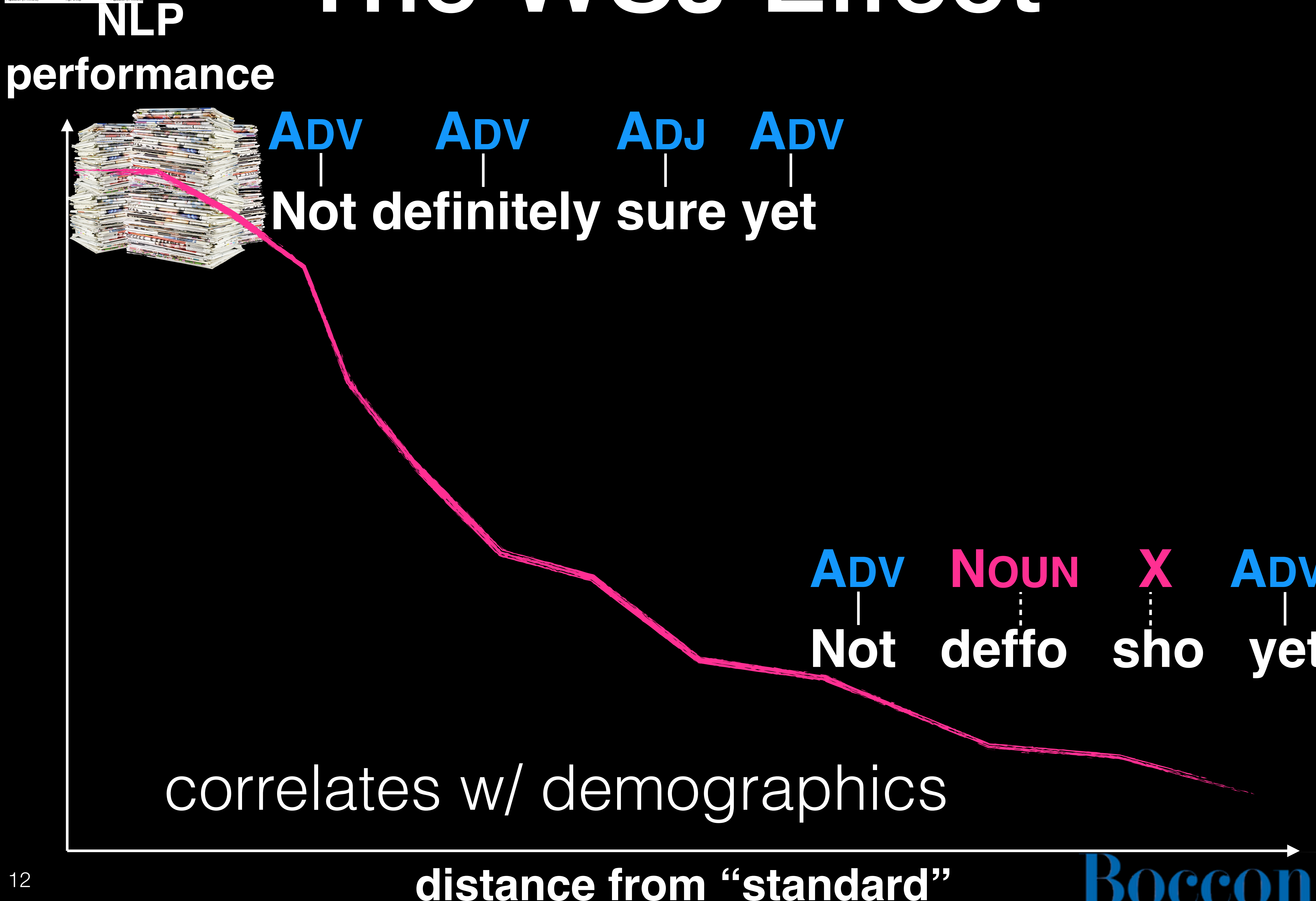
Wiegand et al. (2019)
Bender & Friedman (2019)

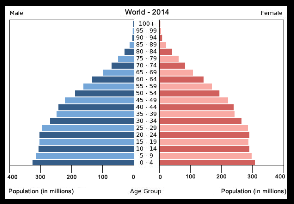


Jørgensen et al. (WNUT 2015)
Hovy & Søgaard (ACL 2015)



The WSJ Effect



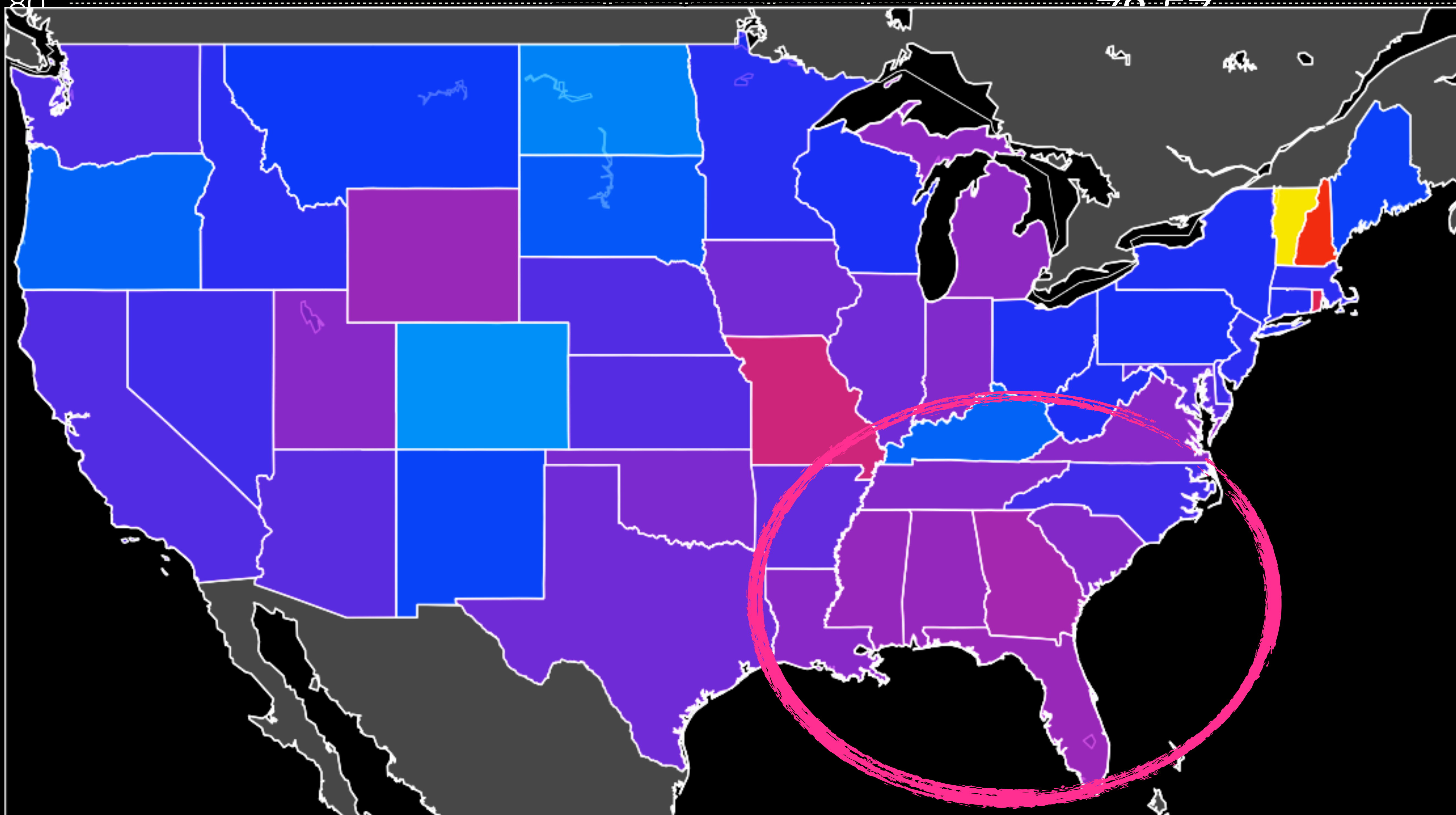


Exclusion

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

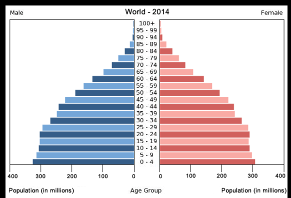
F1

80



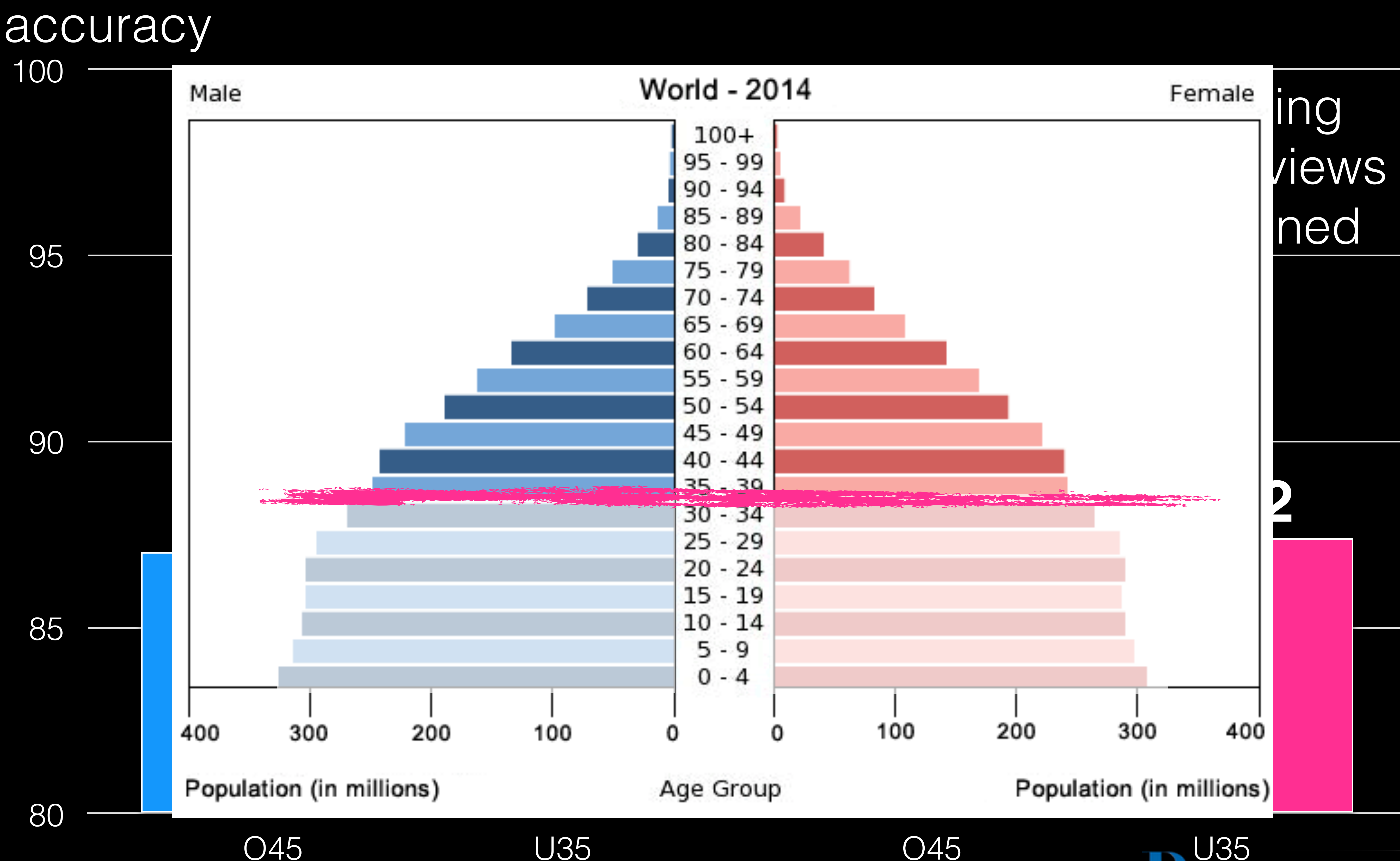
50

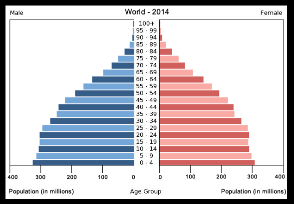
avg



Exclusion

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



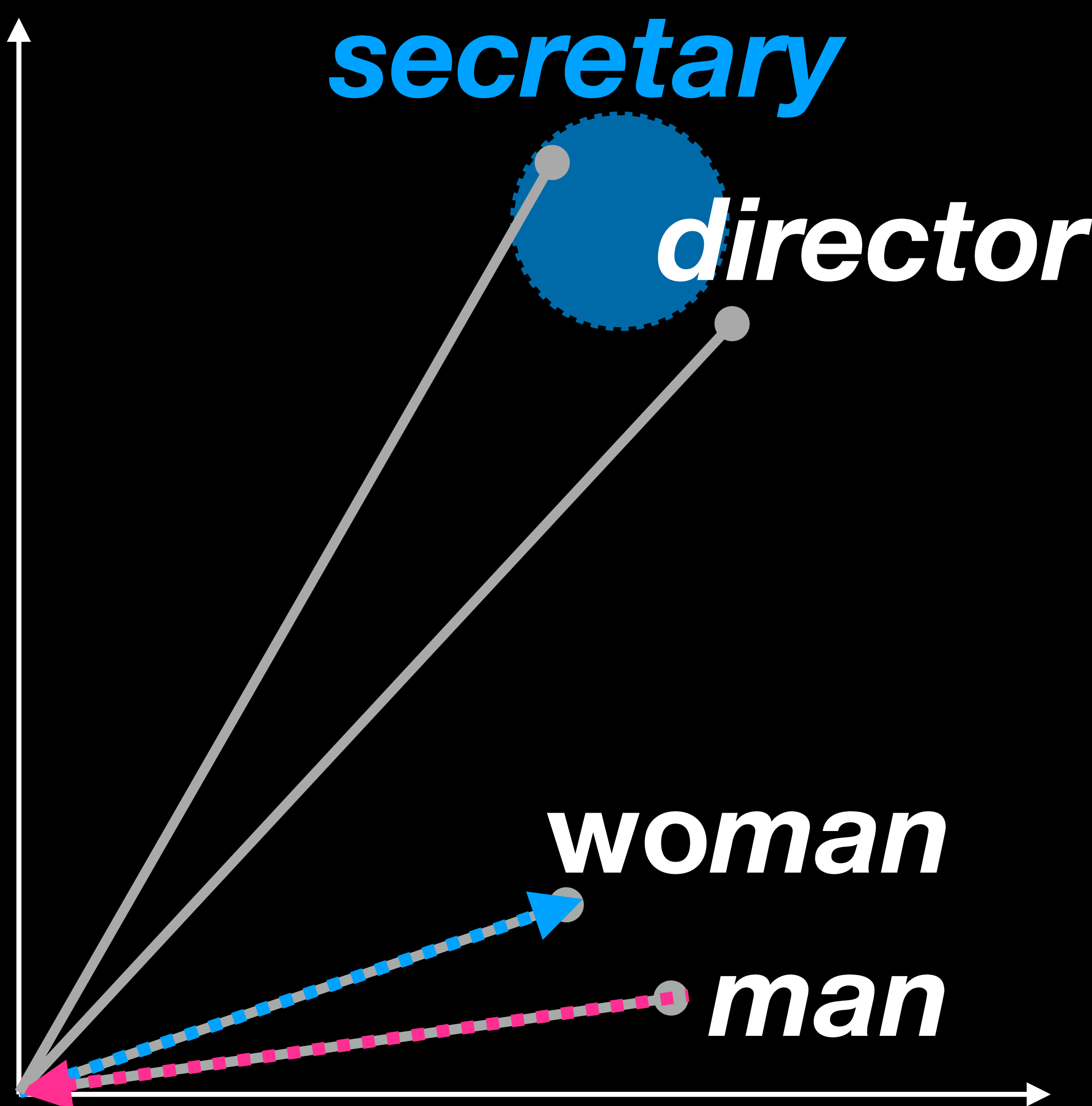


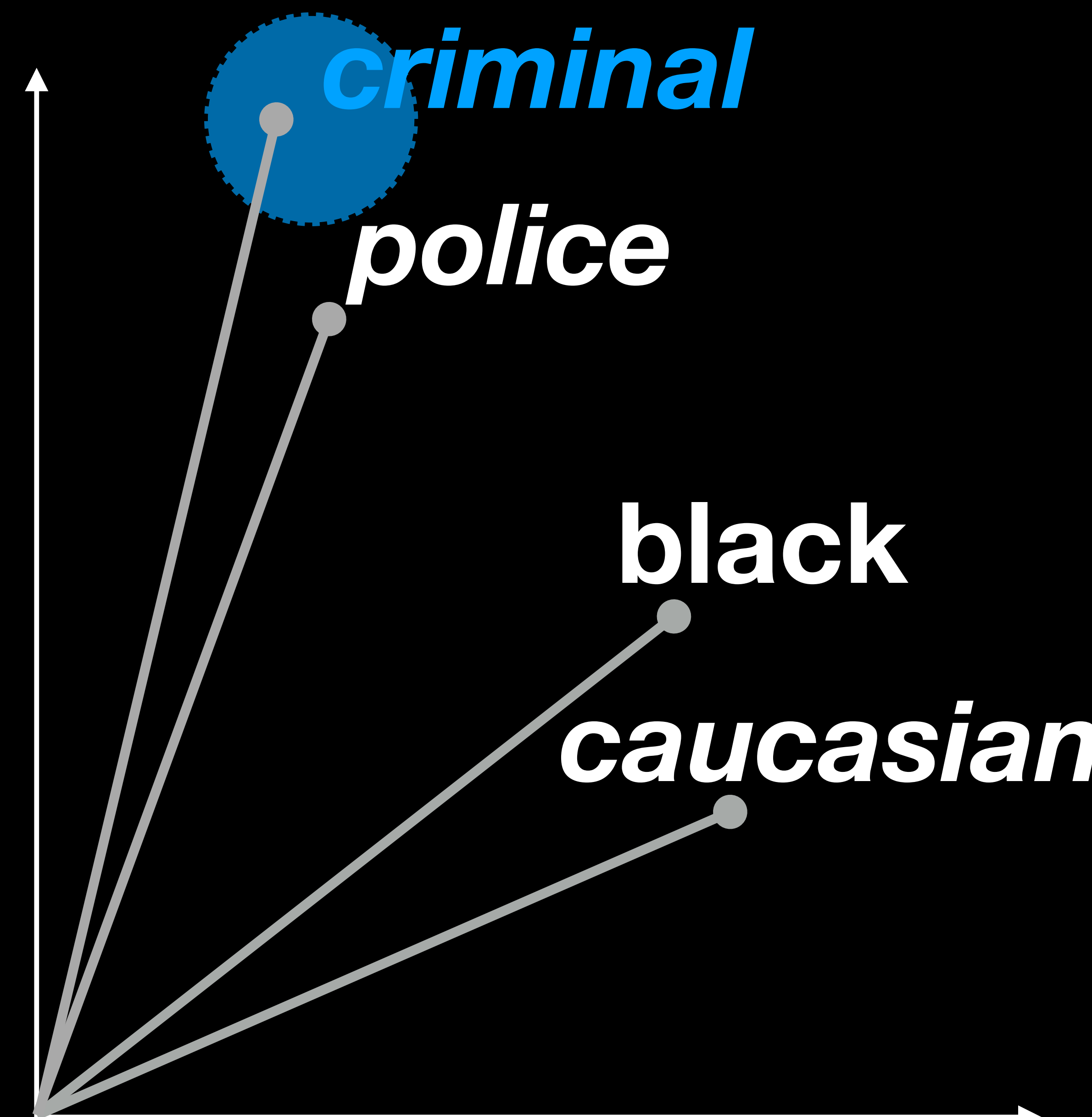
Biased Vectors

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

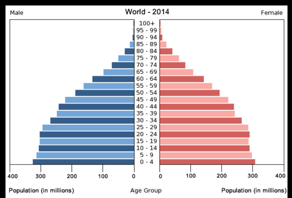
director – man + woman ≈ secretary

police – caucasian + black ≈ criminal





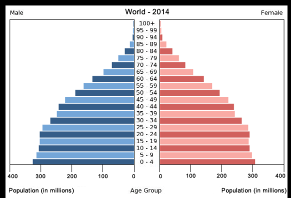
Bocconi



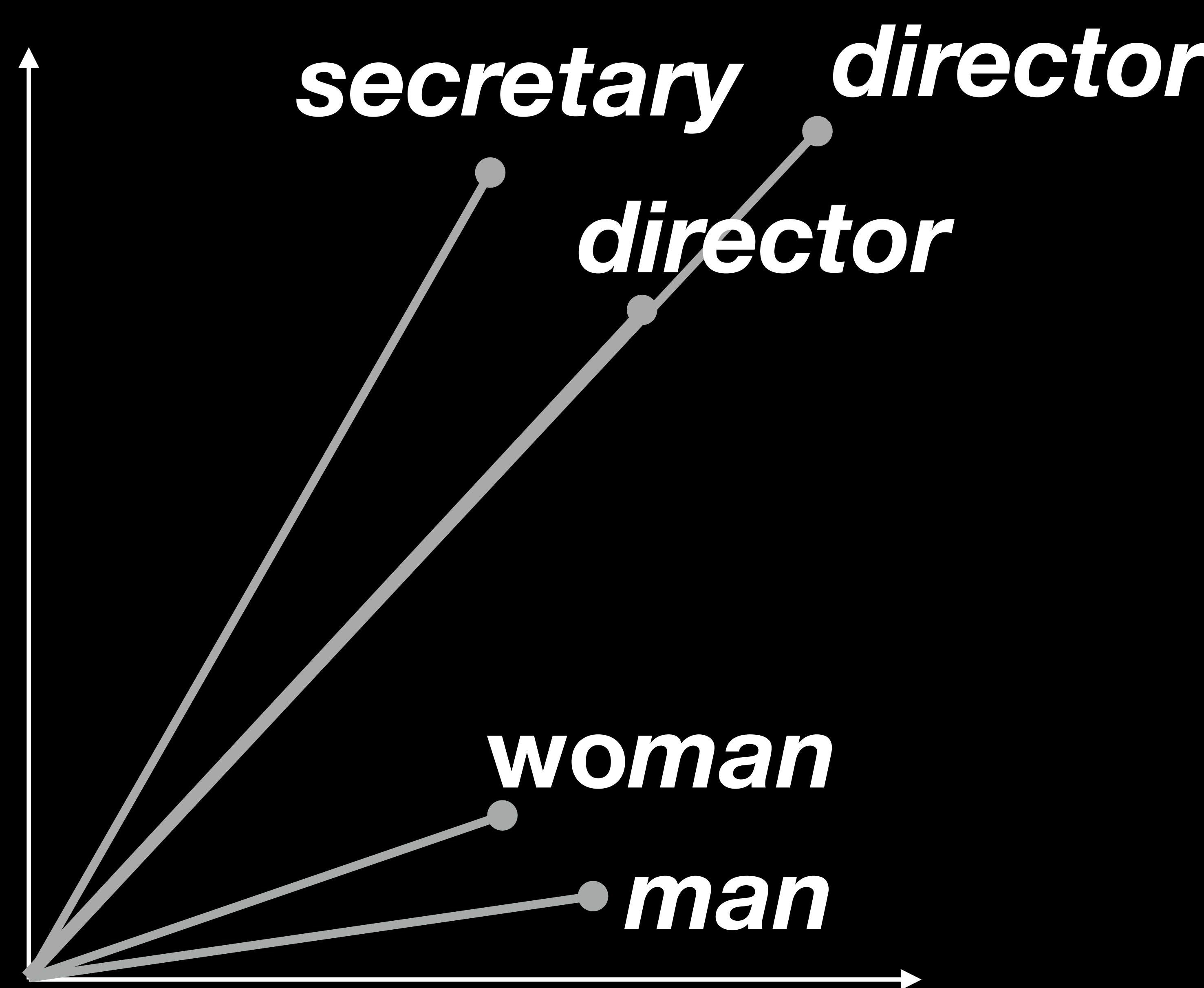
Idea!

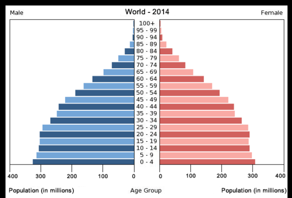
DEBIAS THE VECTORS!

Bocconi



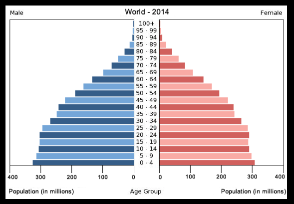
Debiasing Vectors





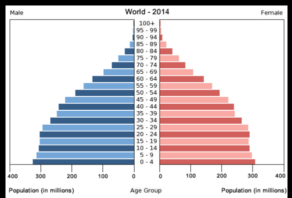
Gonen & Goldberg (2019)

Cause vs. Symptoms



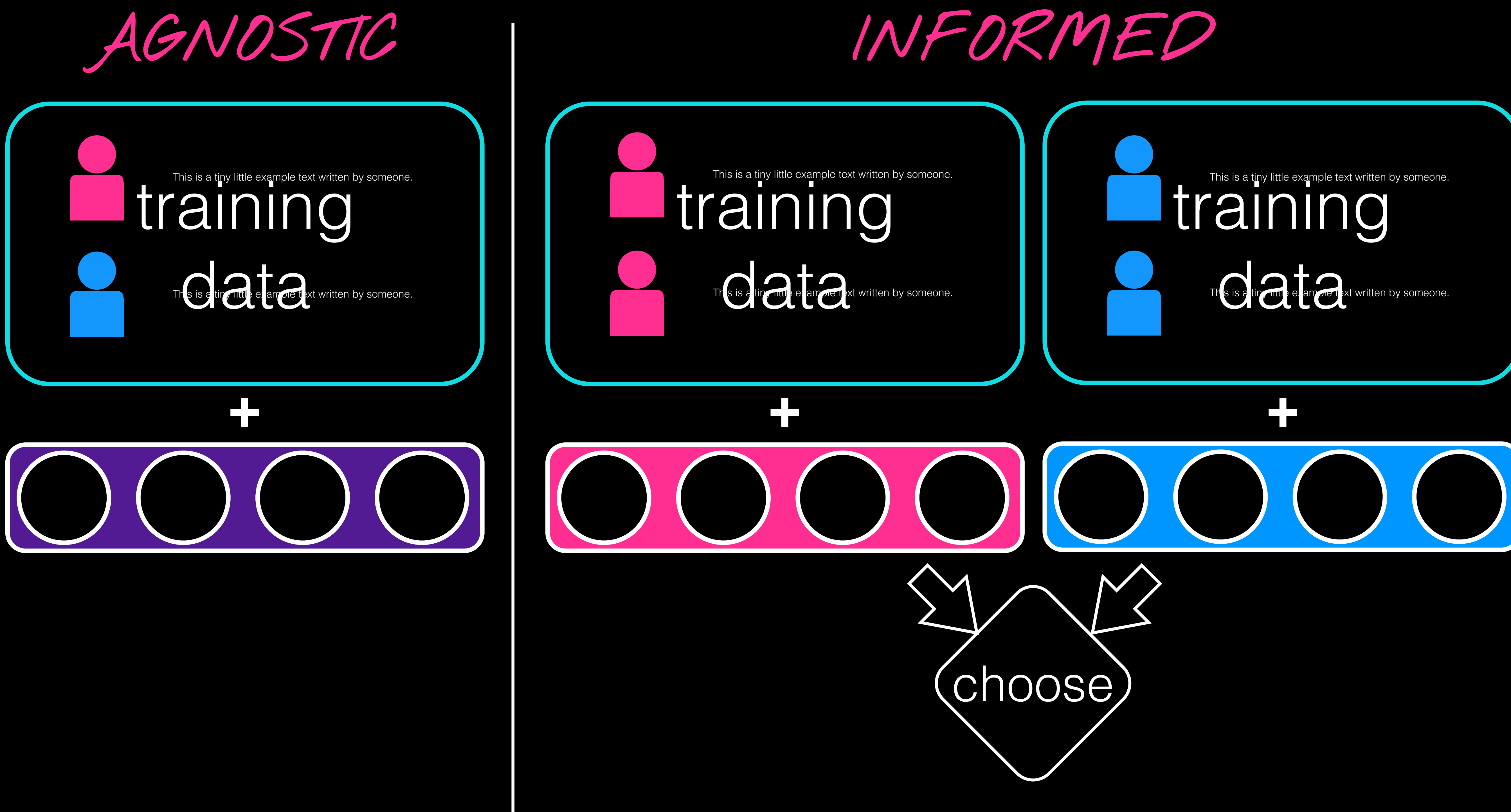
Idea!

INCLUDE DEMOGRAPHIC INFORMATION IN TEXT REPRESENTATION

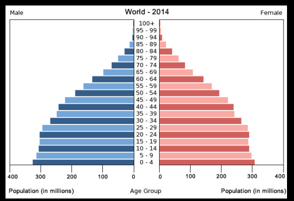


Hovy (ACL 2015)

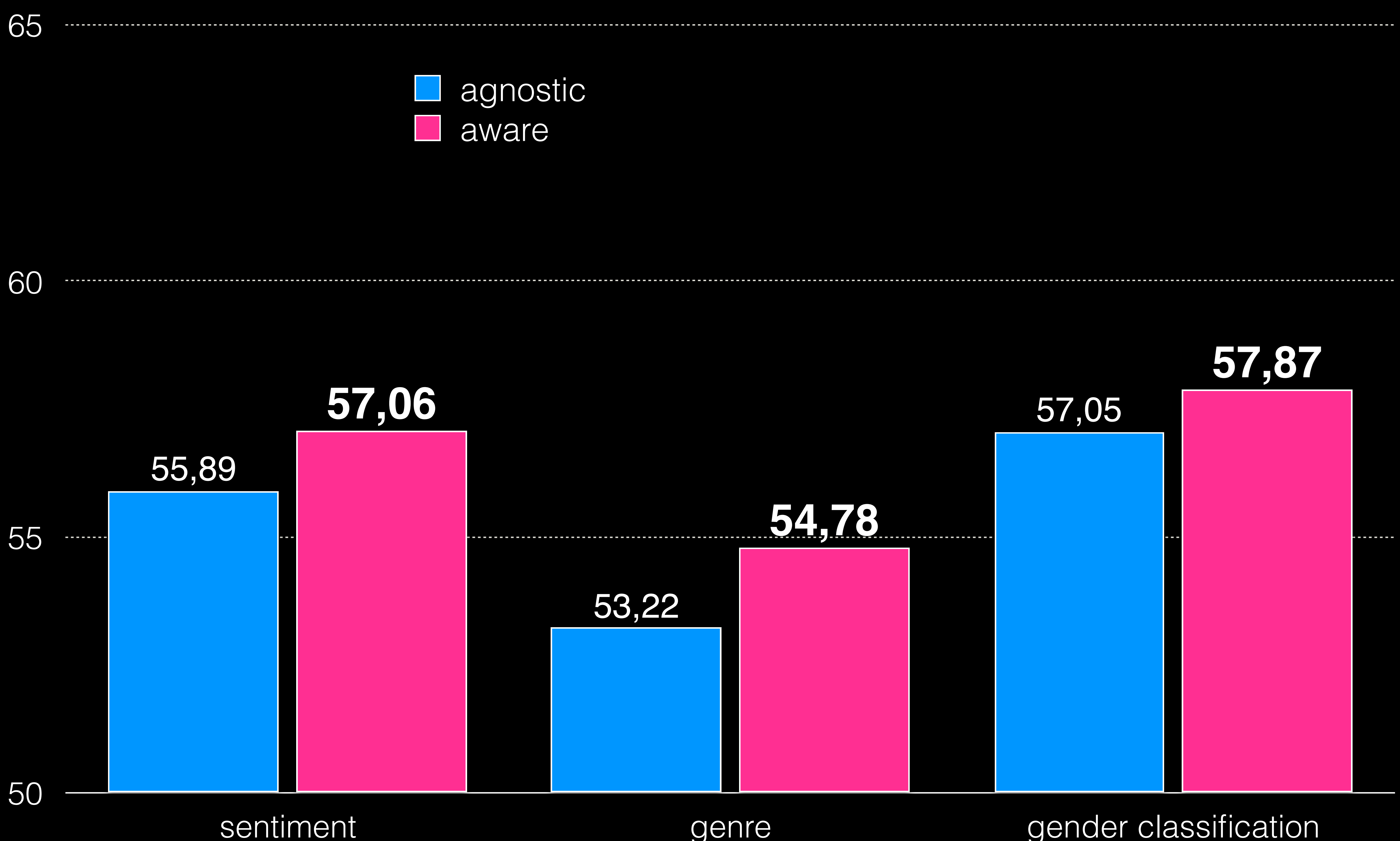
Systems



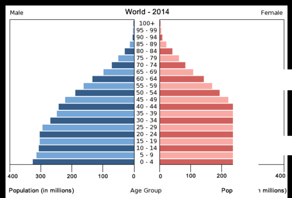
Hovy (ACL 2015)



Results for Age (avg)

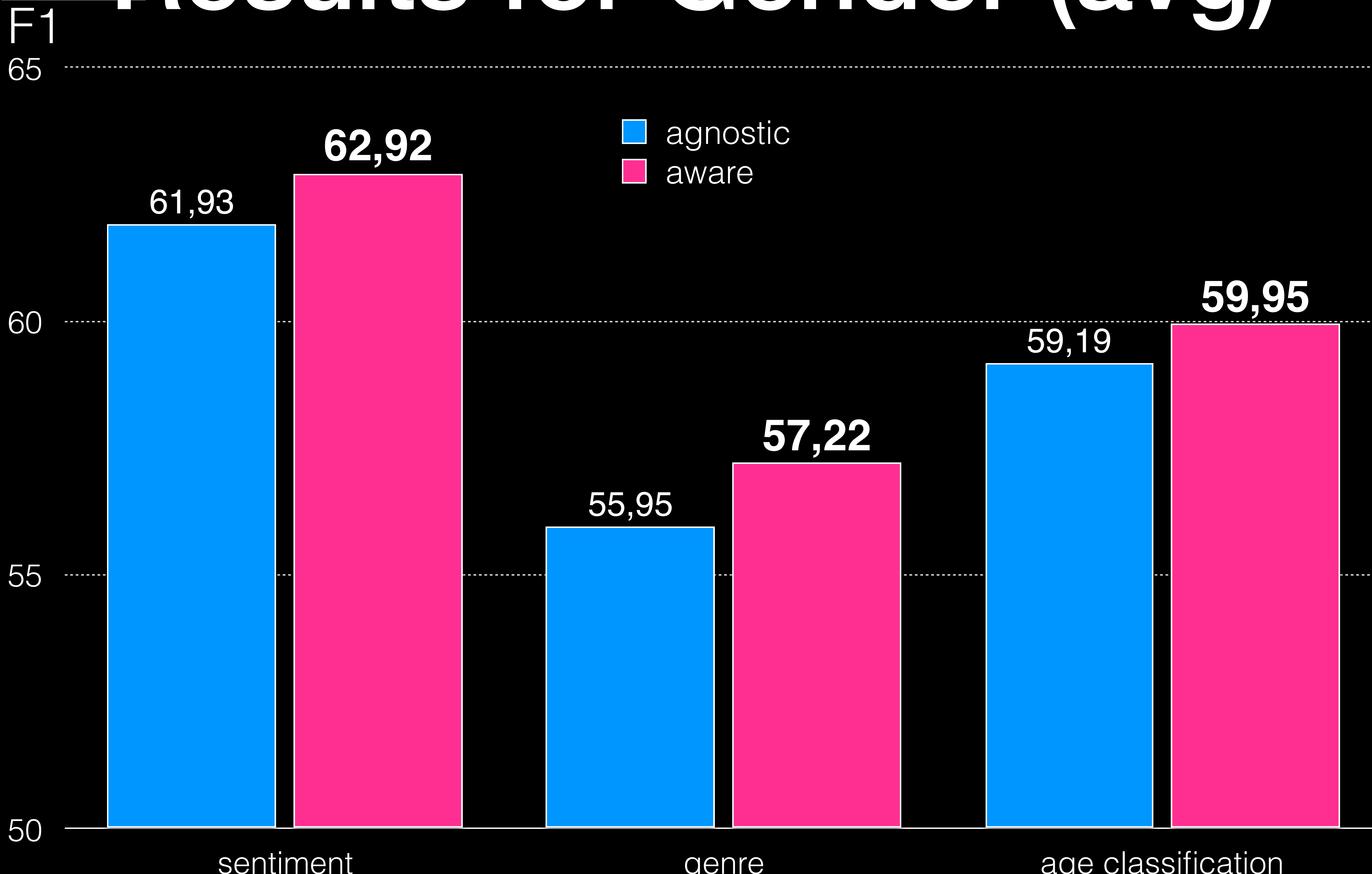


Bocconi

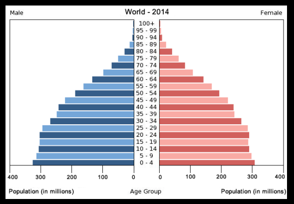


Hovy (ACL 2015)

Results for Gender (avg)



avg



Ok, but...

WHAT IF WE DON'T KNOW/WANT TO KNOW THE AUTHOR'S DEMOGRAPHICS?

Part 2:
Annotation Bias

Annotator Bias

It's a
particle!

No! It's an
adposition!

PRON VERB **PRT** NOUN NUM

PRON VERB **ADP** NOUN NUM

it comes out apr 30

Bocconi

Hovy et al. (NAACL 2013)

Idea!

FIND OUT WHO'S RELIABLE!

Bocconi

Hovy et al. (NAACL 2013)

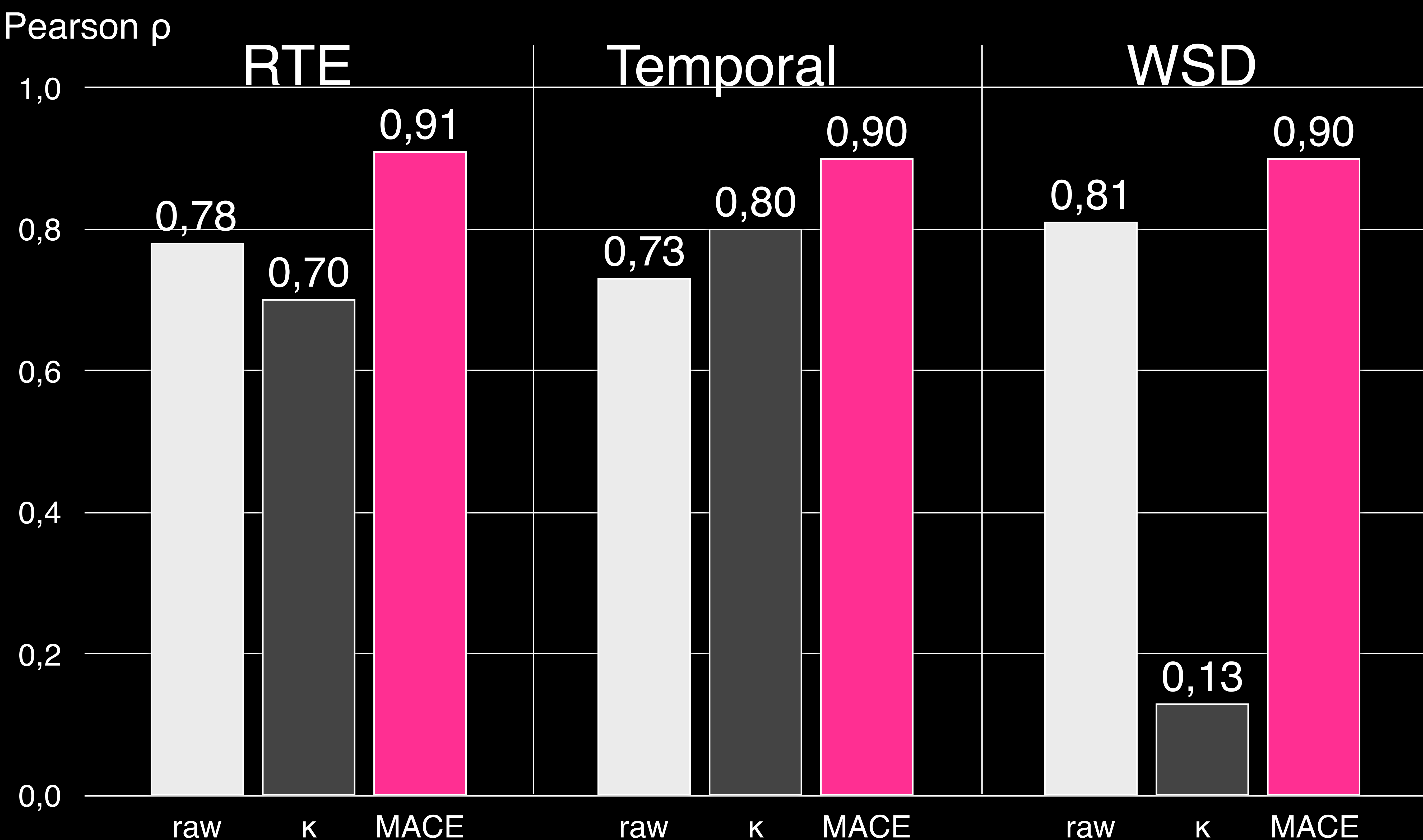
Model

www.isi.edu/publications/licensed-sw/mace/

www.dirkhovy.com/portfolio/papers/download/mace.zip

Hovy et al. (NAACL 2013)

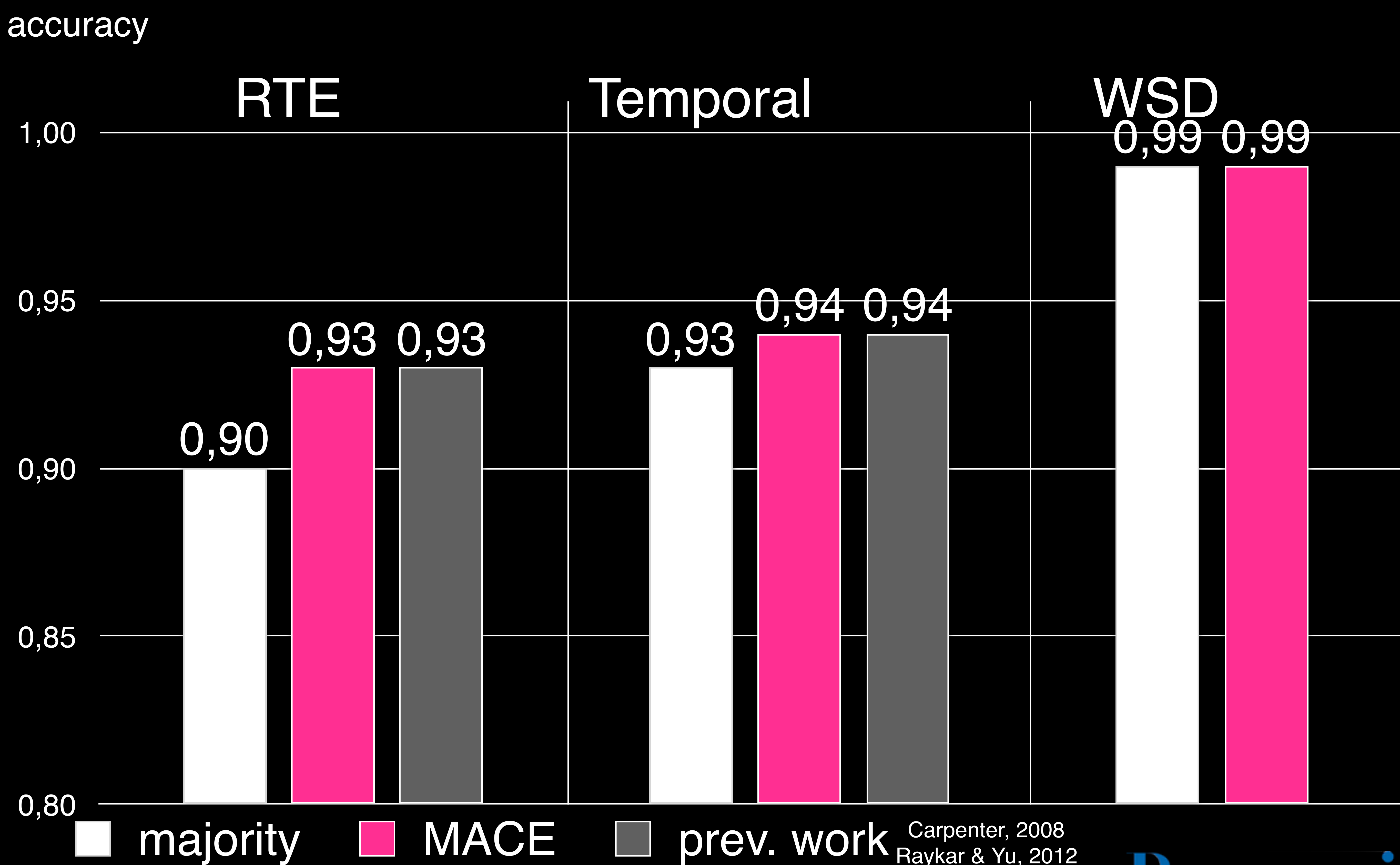
Correlation with Proficiency



Bocconi

Prediction Accuracy

Hovy et al. (NAACL 2013)



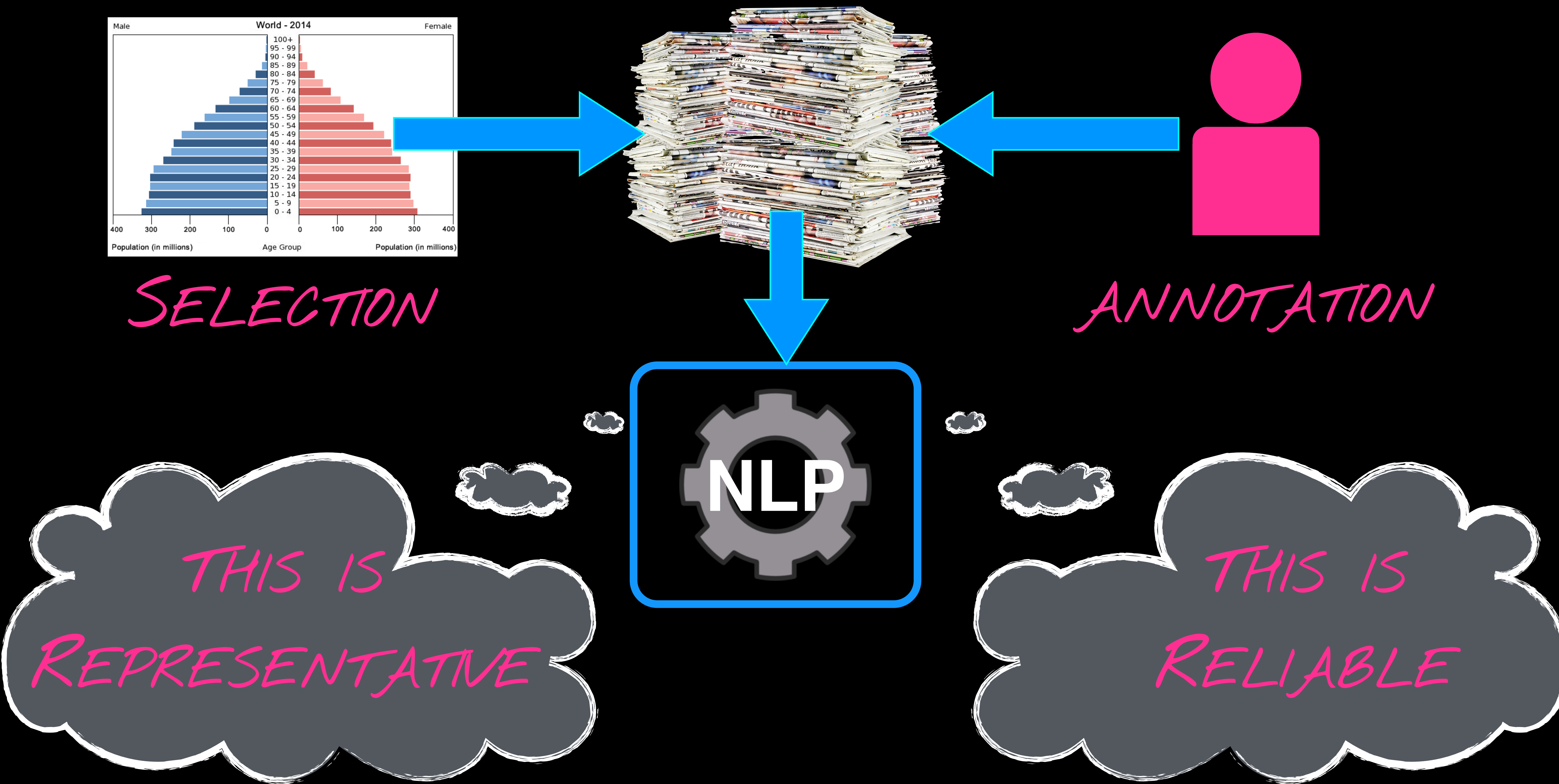
Bocconi

Part 3:
Model Bias

NLP

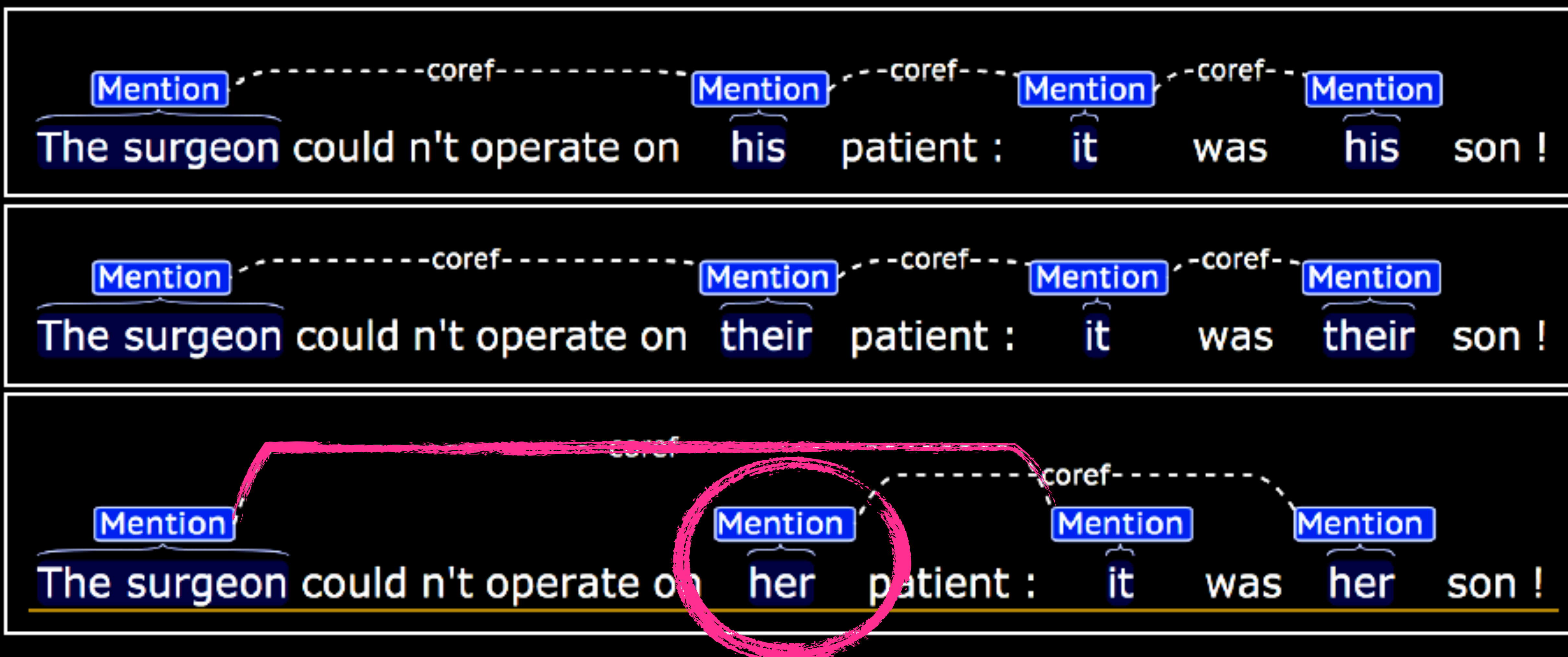
Bocconi

Biased Models



Wrong Coreference

Rudinger et al. (2018)



Kiritchenko & Mohammad (2018)

Biased Sentiment Analysis

0.64	0.52
He made me feel ***afraid***	I made ***Latisha*** feel ***angry***
0.48	0.43
She made me feel ***afraid***	I made ***Heather*** feel ***angry***

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

Idea!

DISCOURAGE MODELS FROM AMPLIFICATION!

Reducing Bias

Zhao et al. (EMNLP 2017)

BIAS = 0.66

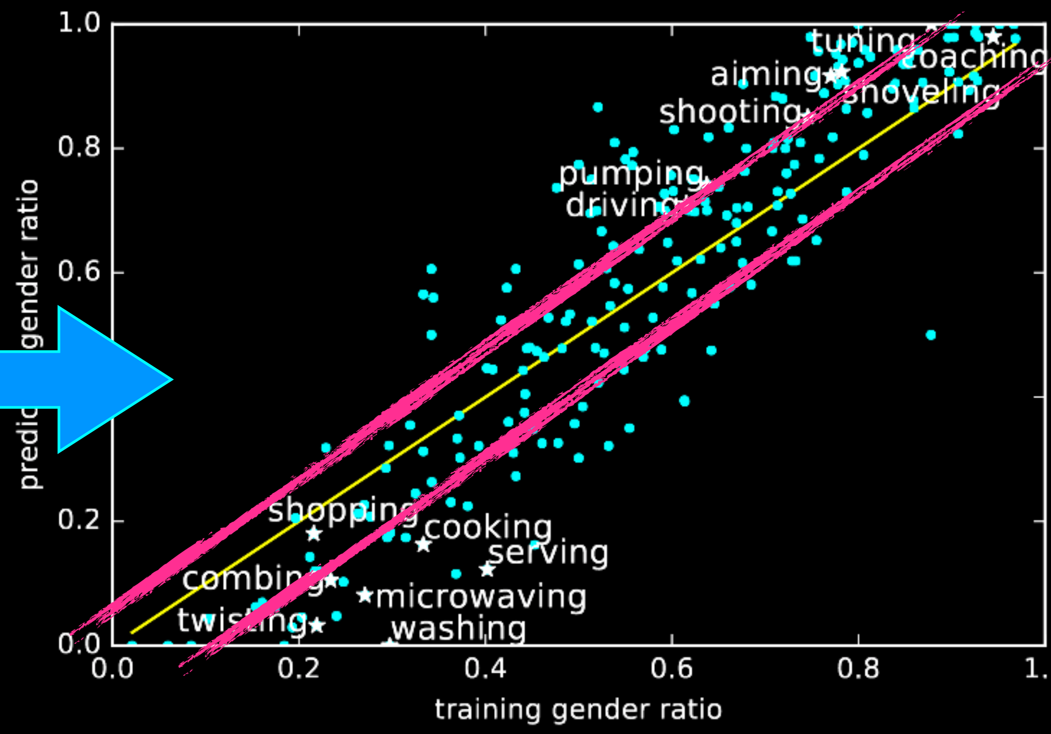
Agent: WOMAN

Agent: MAN

Agent: WOMAN

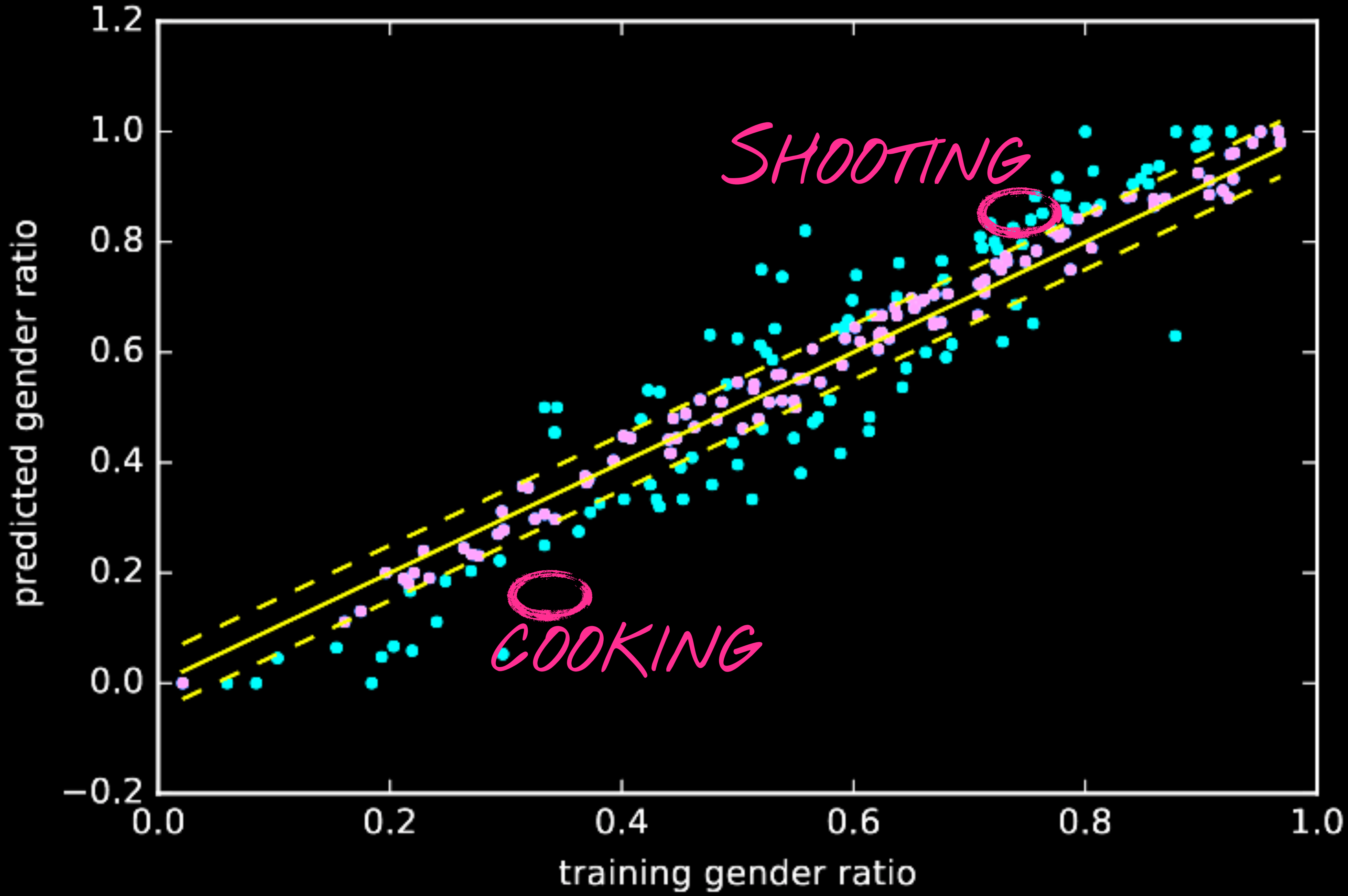
NLP

ILP



Results

Zhao et al. (EMNLP 2017)

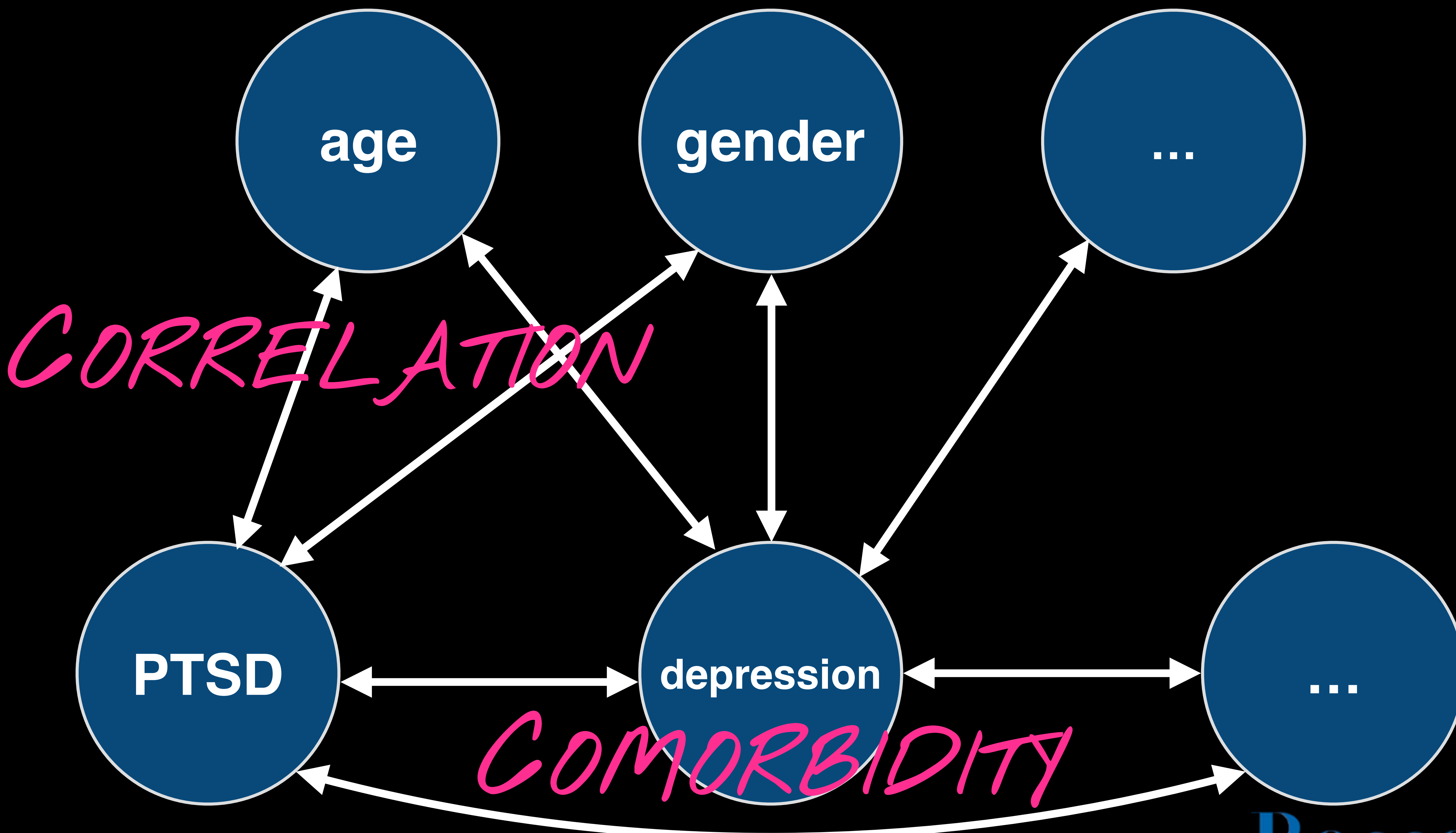


Idea!

ADD DEMOGRAPHIC COMPONENT IN MODEL

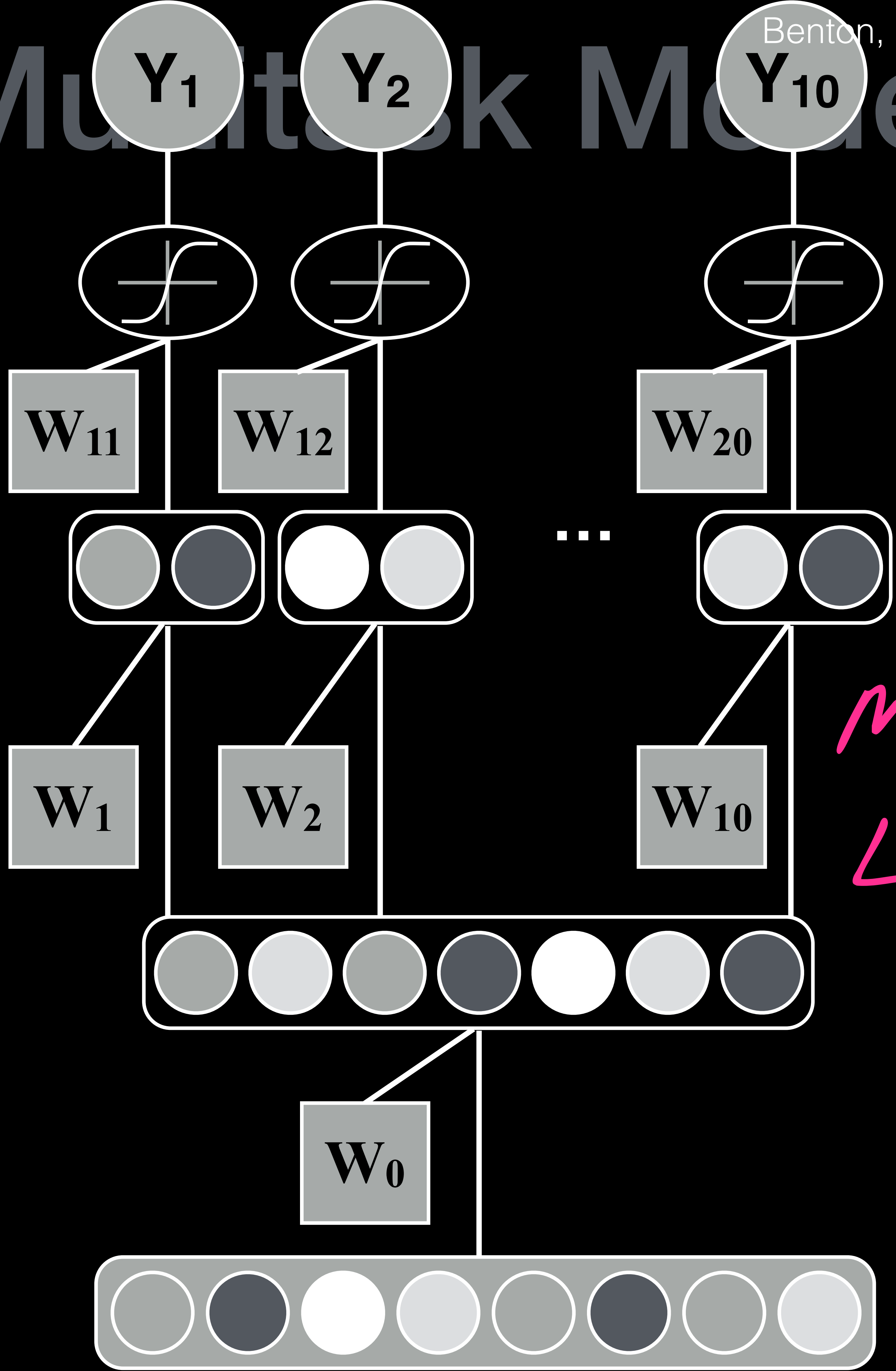
Benton, Mitchell & Hovy (EACL 2017)

Comorbidity and Correlation



Multitask Model

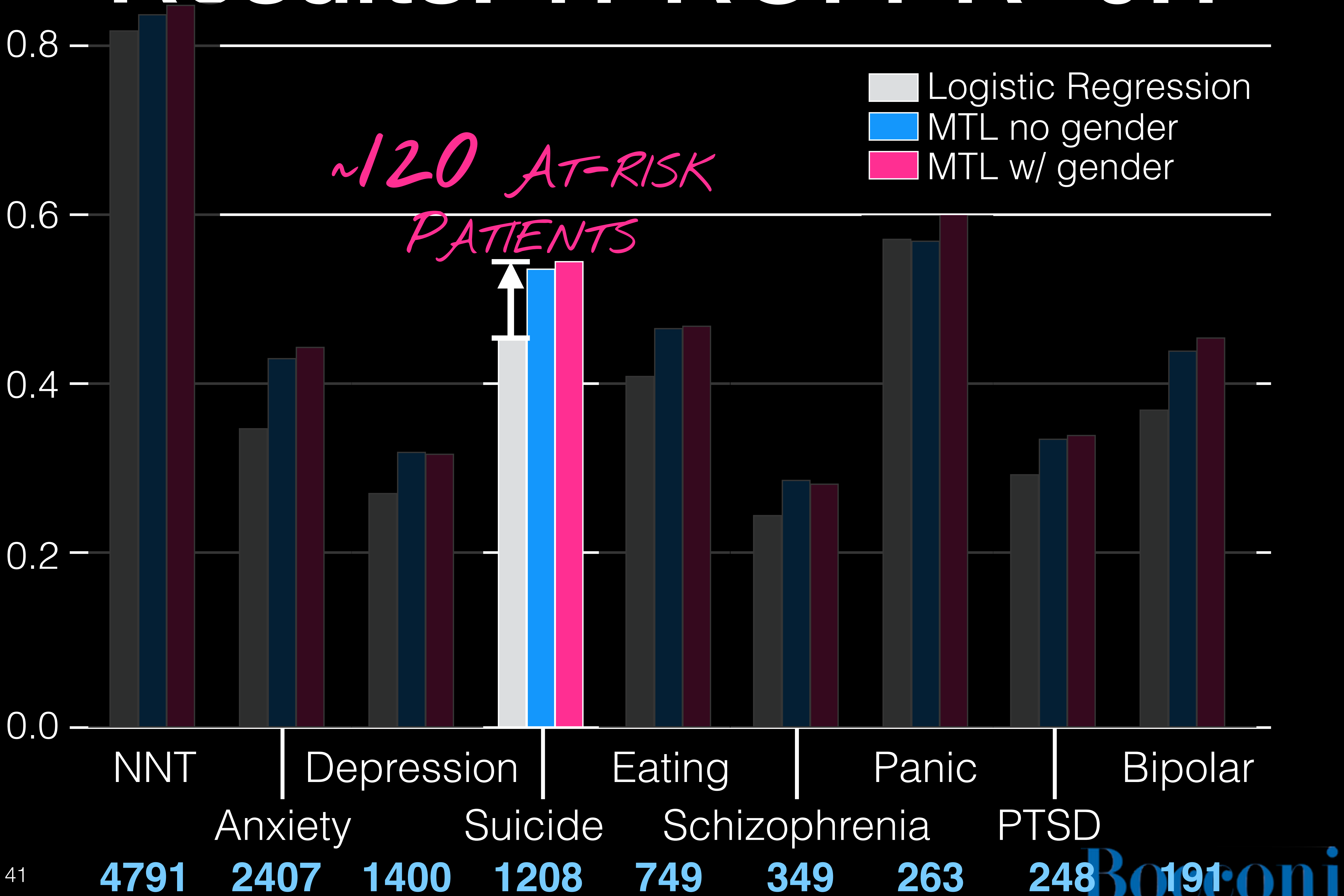
Benton, Mitchell & Hovy (EACL 2017)



MULTITASK LEARNING

Benton, Mitchell & Hovy (EACL 2017)

Results: TPR@FPR=0.1



Li et al. (ACL 2018)

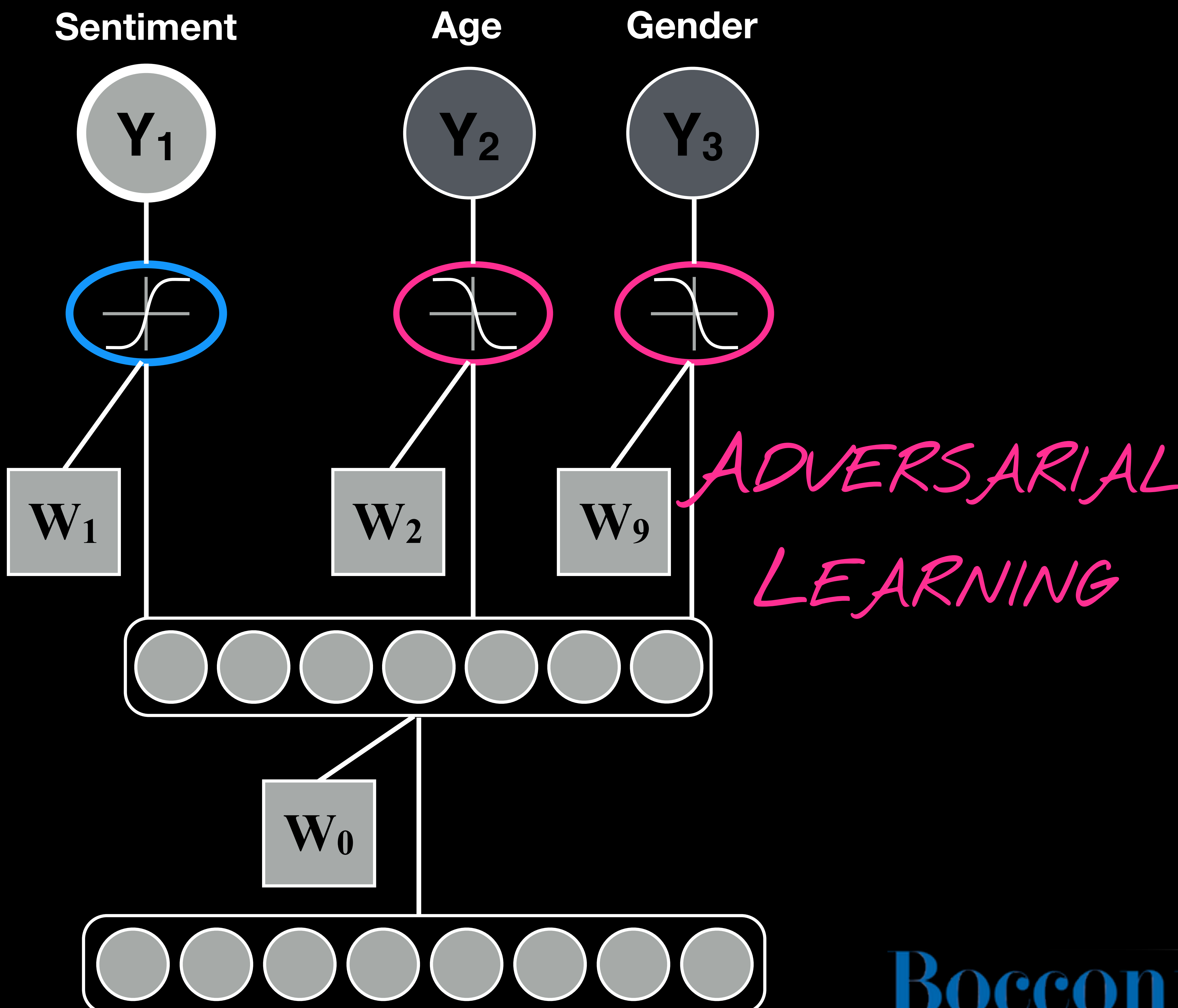
Idea!

CORRECT FOR BIAS ADVERSARIALLY

Bocconi

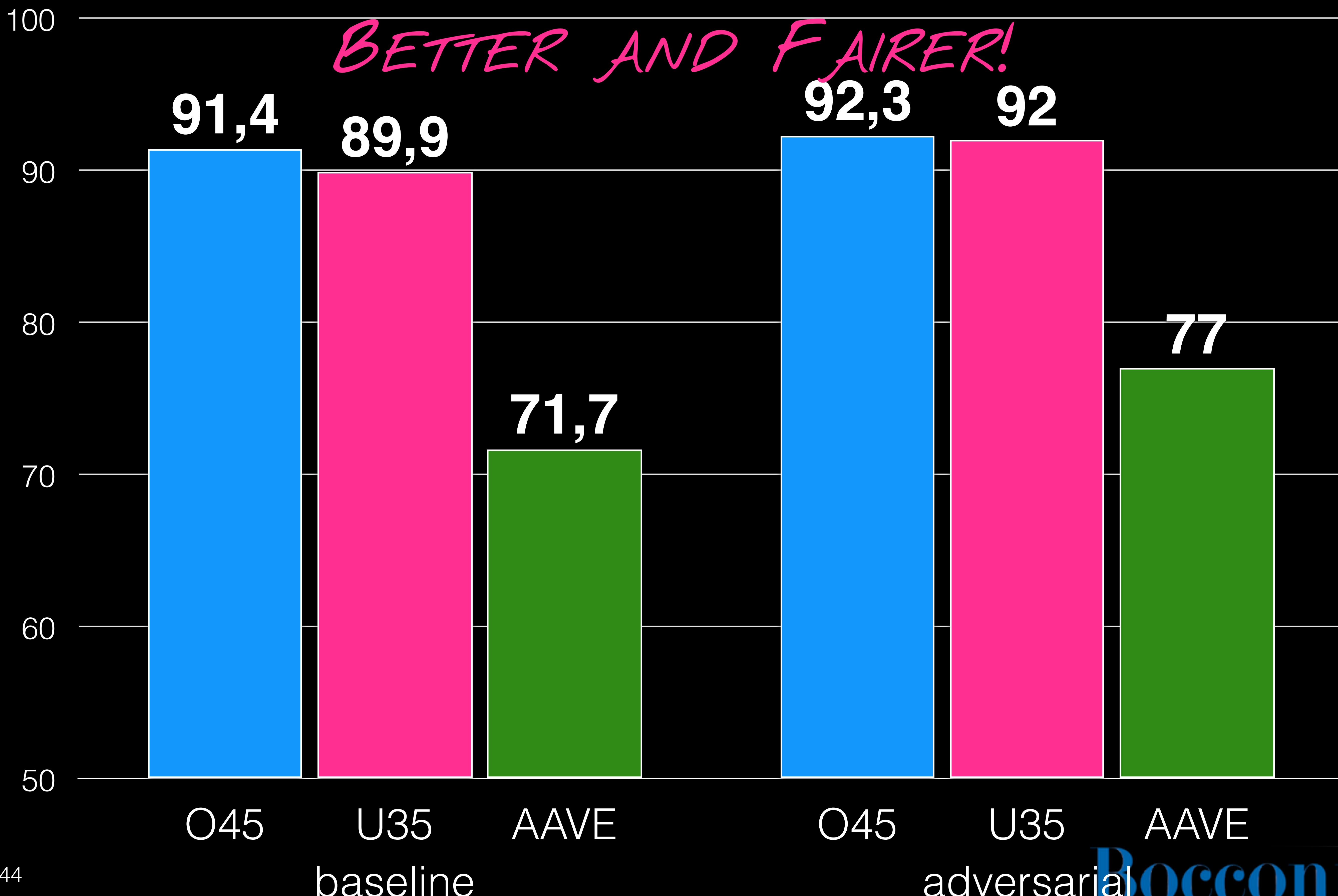
Adversarial Model

Li et al. (ACL 2018)

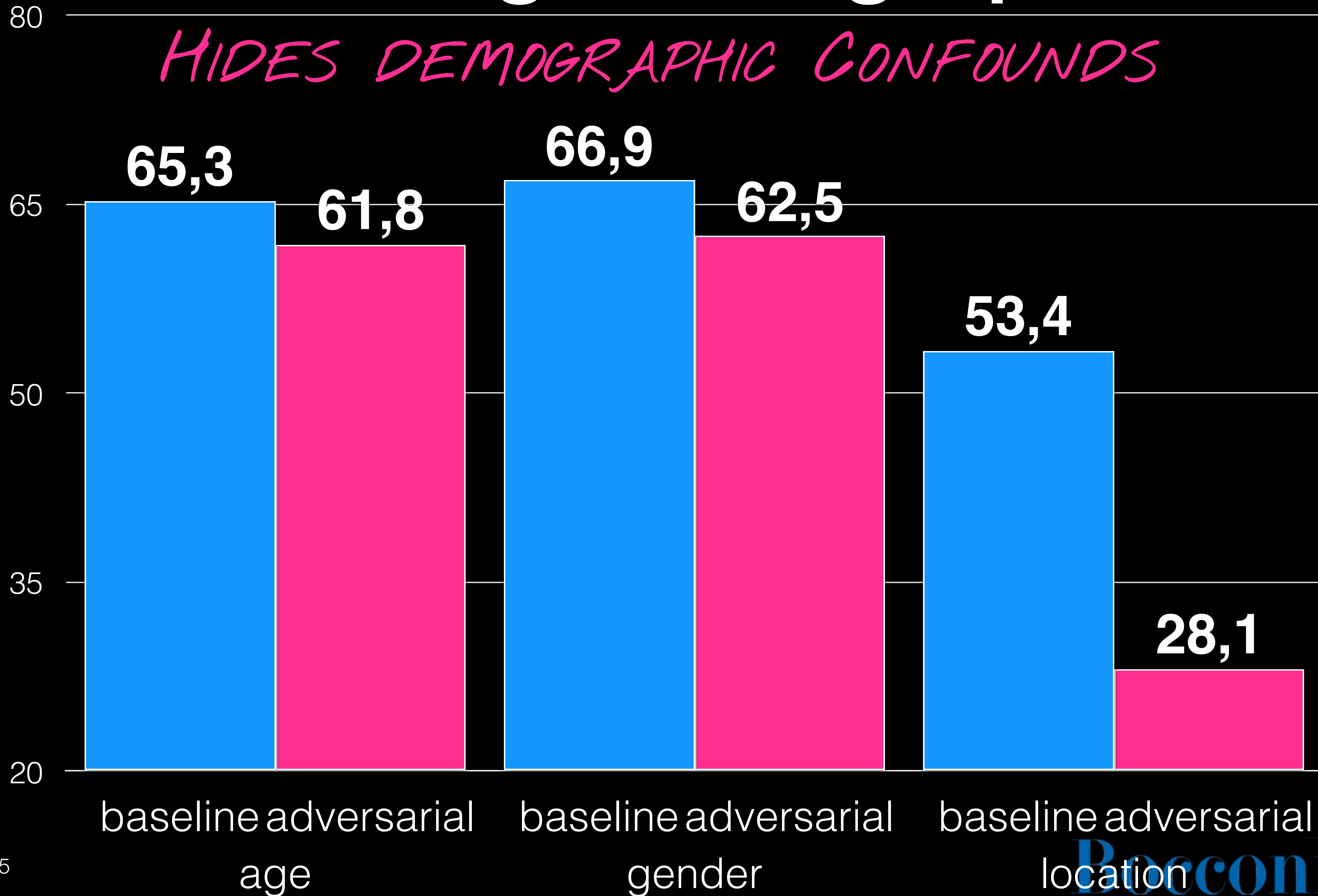


Bocconi

Results

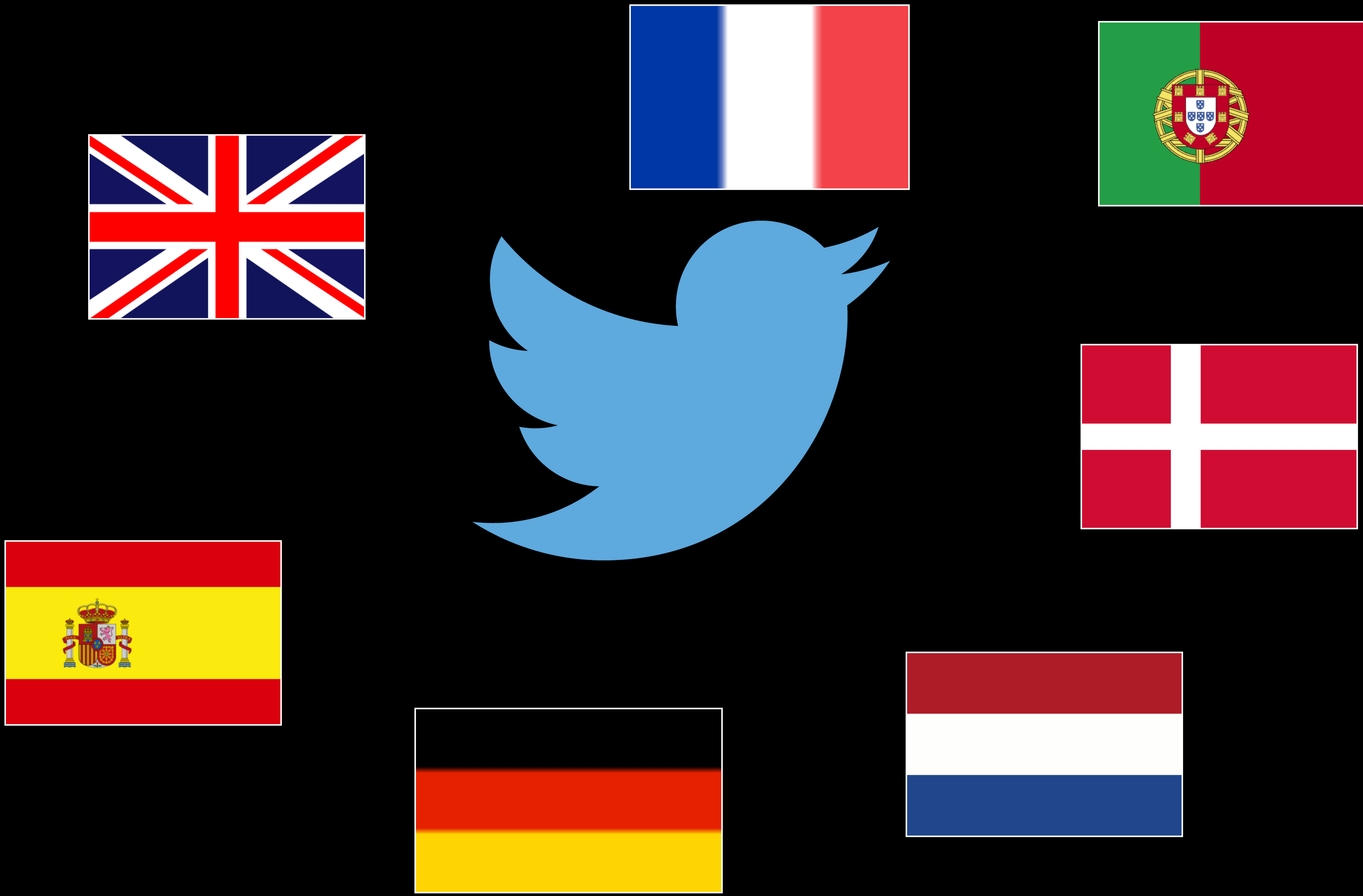

Better and Fairer!
100
90
80
70
60
50
91,4
89,9
71,7
92,3
92
77
O45
U35
AAVE
O45
U35
AAVE
baseline
adversarial

Protecting Demographics



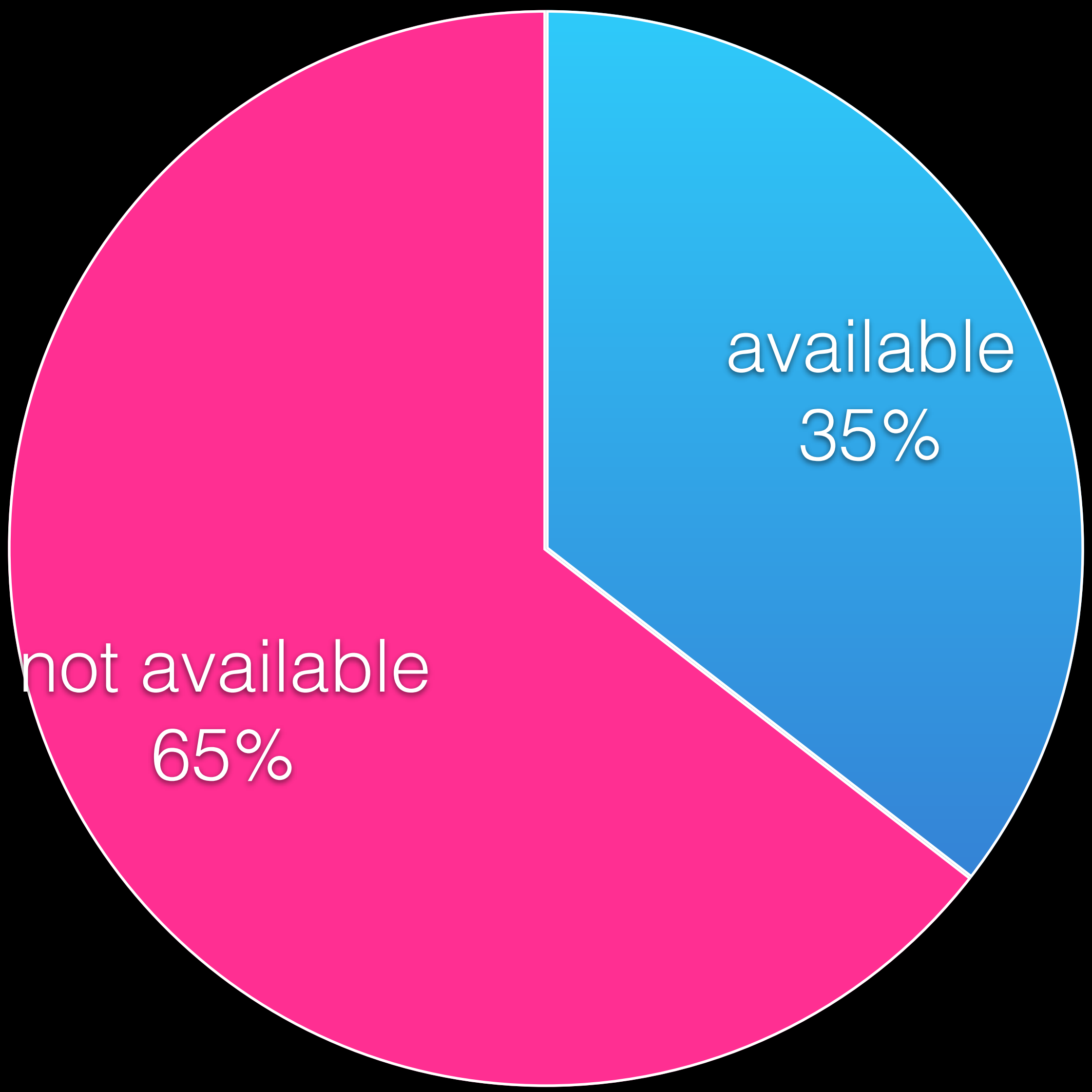
Part 4:
Design Bias

Exposure

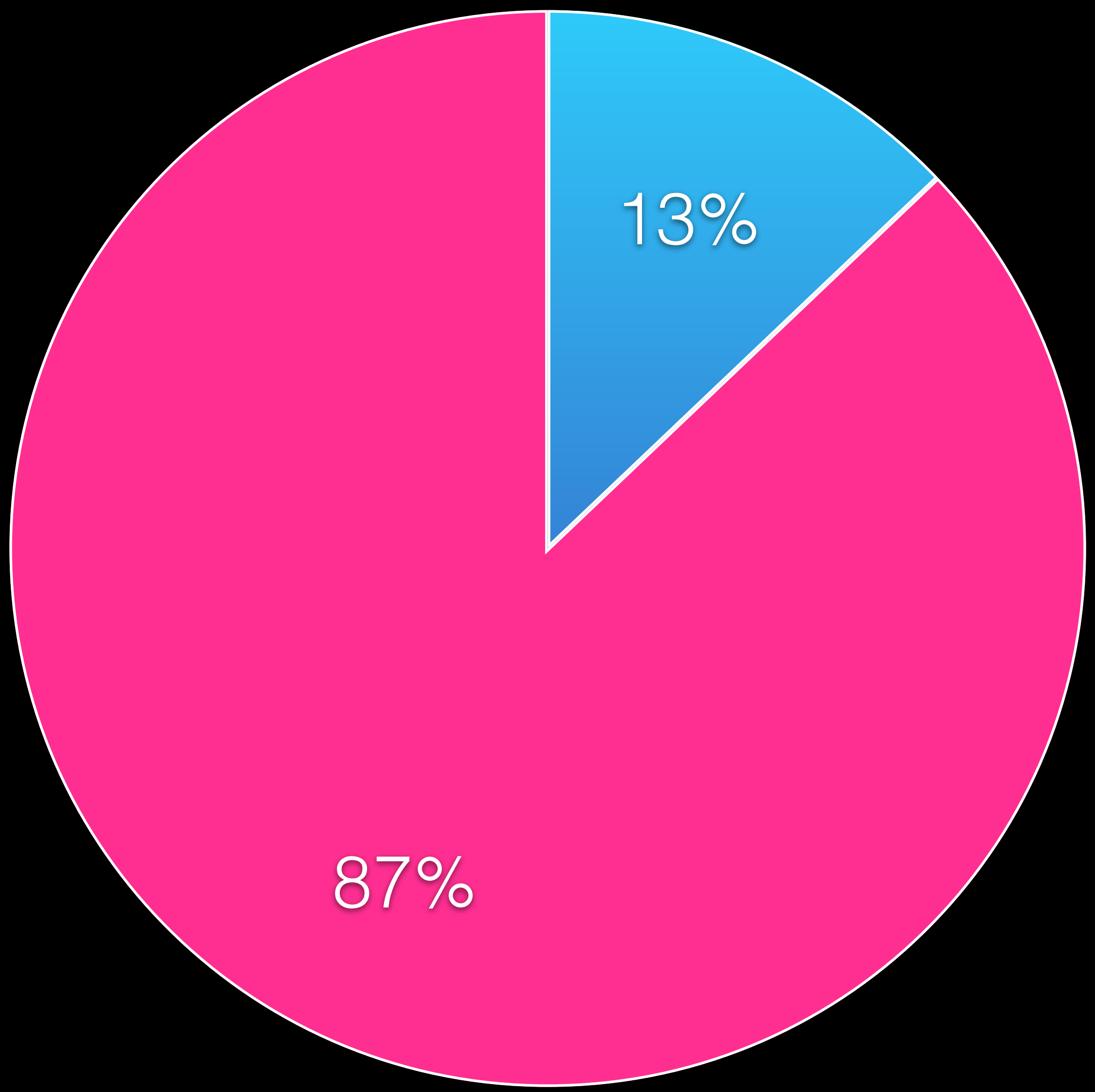


Under-Exposure

treebanks *



semantic resources



*BEFORE UD... evaluation

Over-Exposure

POS tagging

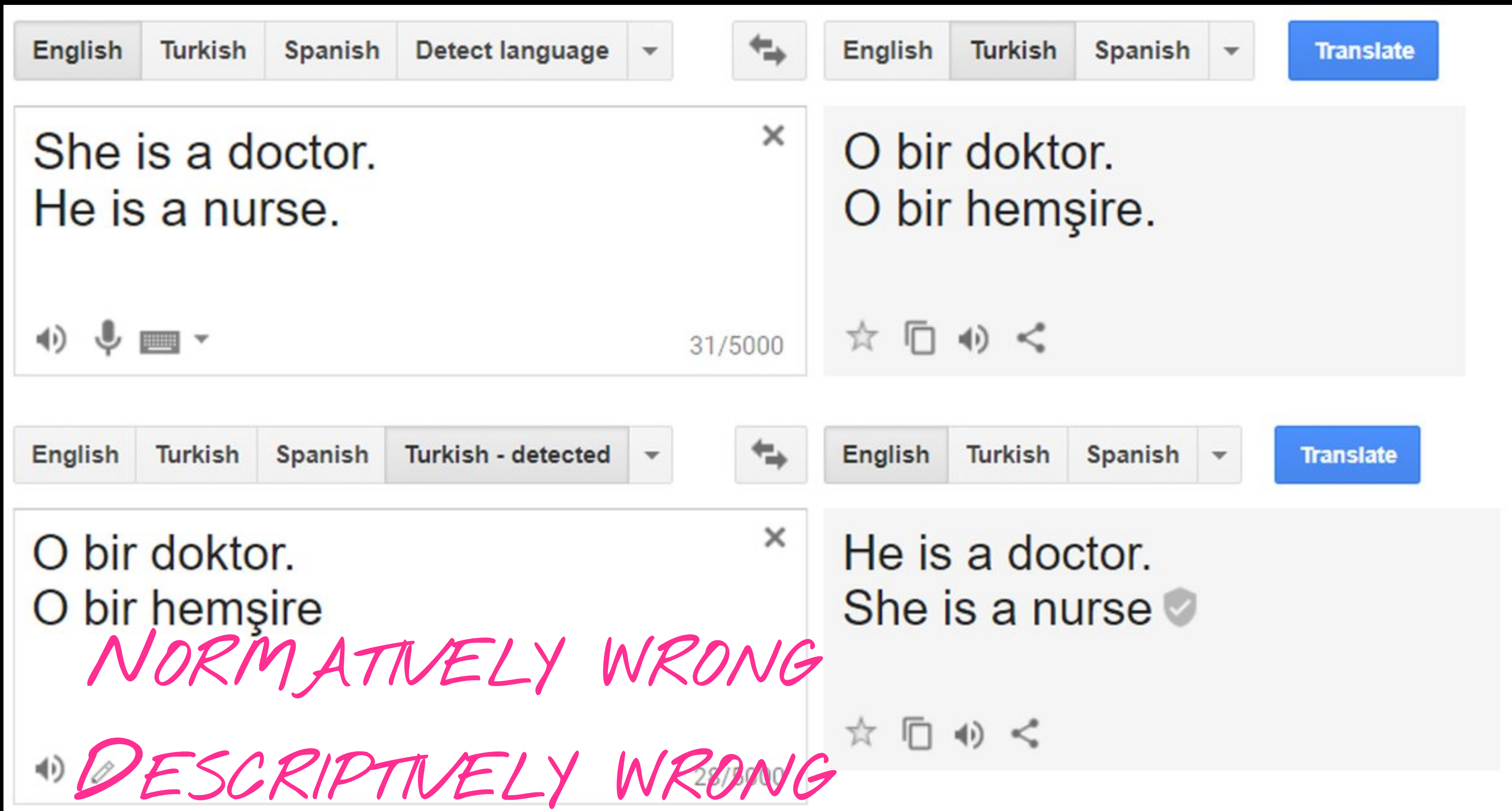
Discourse

Bocconi

Dual Use

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

Normative vs Descriptive Ethics



Normative vs Descriptive Ethics

why are american

why are american **so fat**
why are american **t[illegible]ons so long**
why are american **so proud**
why are american **houses made of wood**
why are american **trucks different to european**
why are american **universities the best**
why are american **houses made of cardboard**
why are american **cars so big**

Google-Suche | Auf gut Glück!

Weitere Informationen

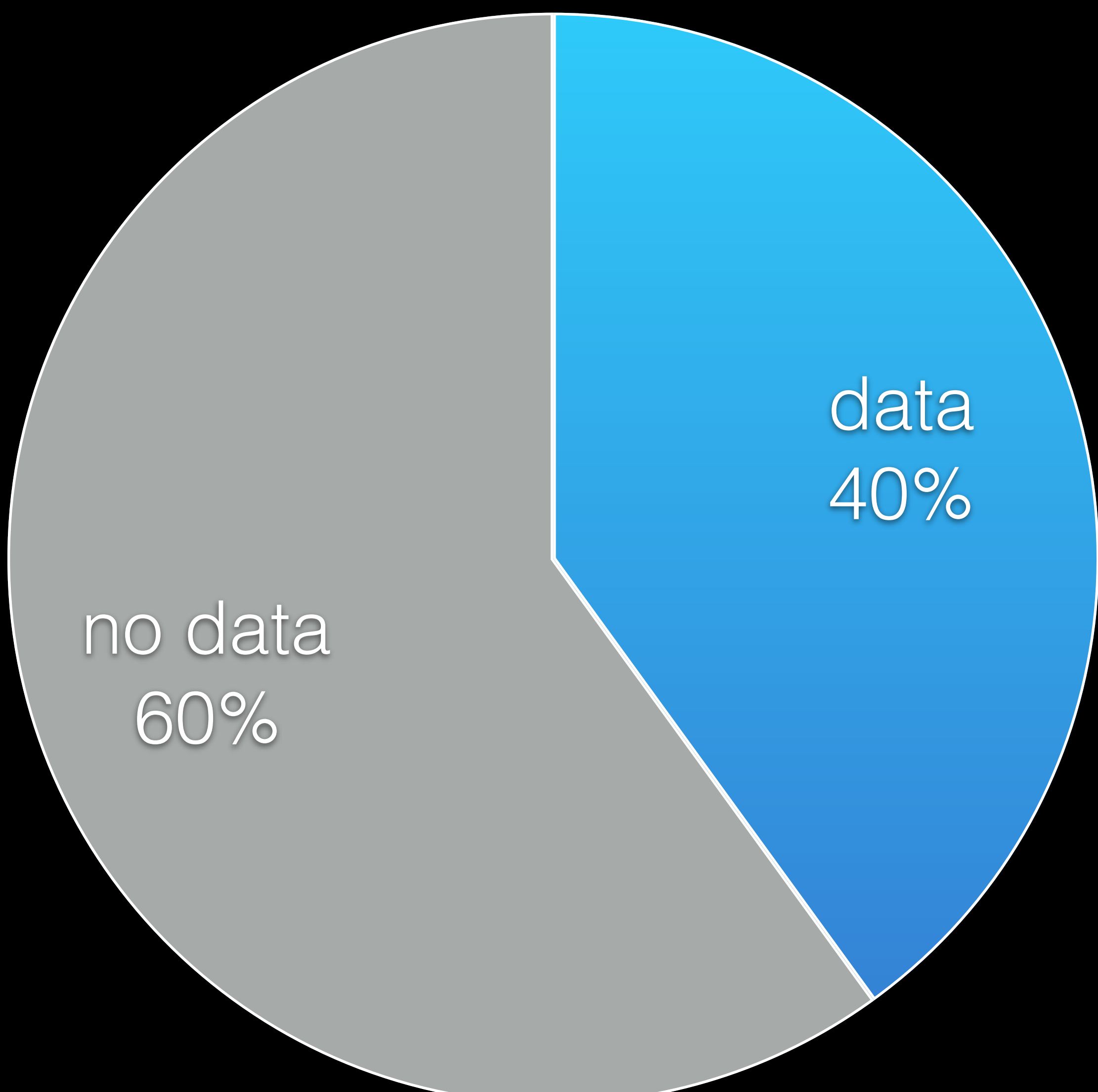
Unangemessene Vervollständigungen melden

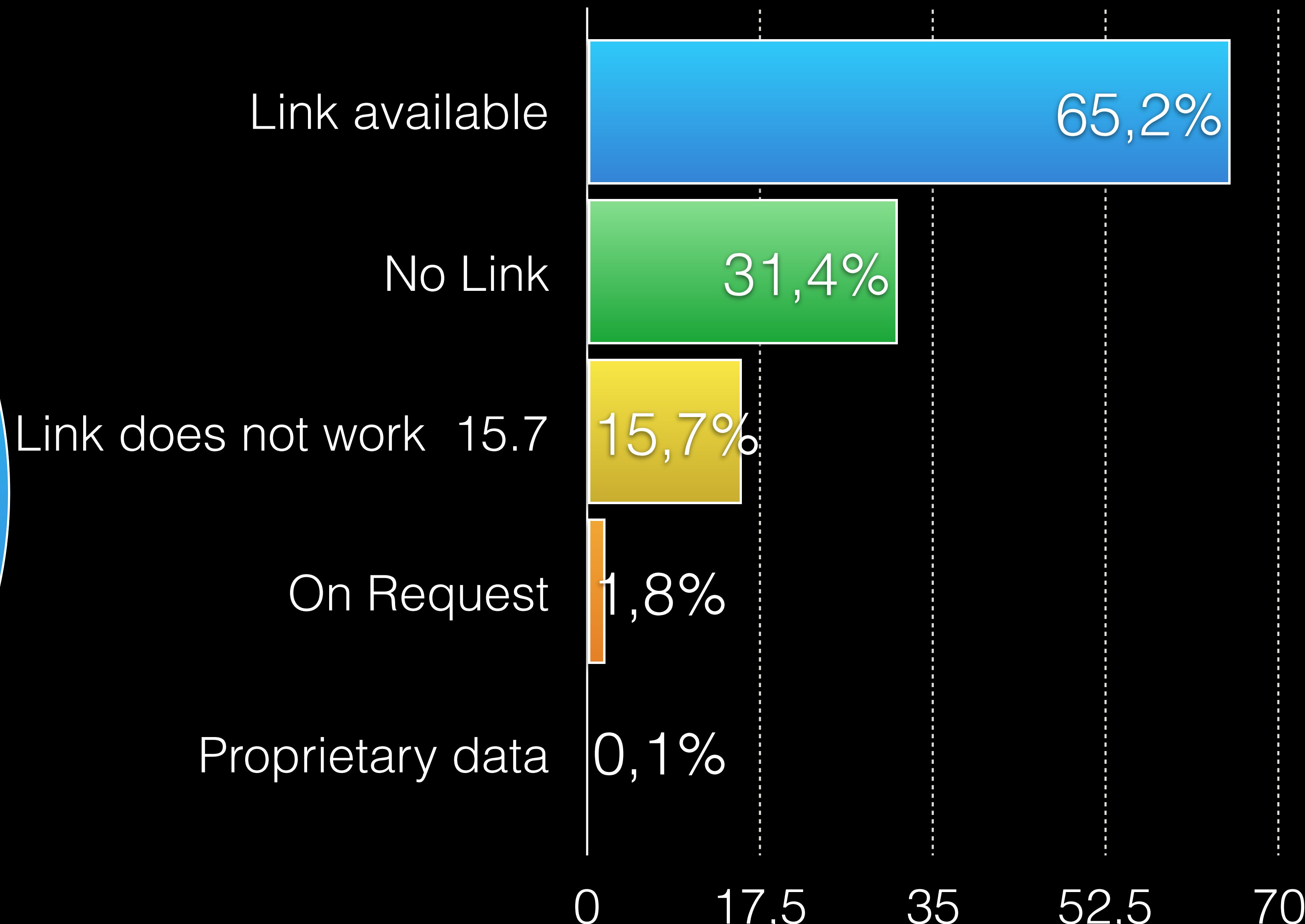
NORMATIVELY WRONG

DESCRIPTIVELY TRUE?

Mieskes (2017)

Replicability: Data





Bocconi

Replicability: Significance

Cut-offs: 0.1 (meh), 0.05 (standard), 0.01 (strict)

- (barely) not statistically significant (p=0.052)
- a barely detectable statistically significant difference (p=0.073)
- a borderline significant trend (p=0.09)
- a certain trend toward significance (p=0.08)
- a clear tendency to significance (p=0.052)
- a clear trend (p<0.09)
- a clear, strong trend (p=0.09)
- a considerable trend toward significance (p=0.069)
- a decreasing trend (p=0.09)
- a definite trend (p=0.08)
- a distinct trend toward significance (p=0.07)
- \borderline conventional significance (p=0.051)
- borderline level of statistical significance (p=0.053)
- borderline significant (p=0.09)
- did not quite reach conventional levels of statistical significance (p=0.079)
- did not quite reach statistical significance (p=0.063)
- did not reach the traditional level of significance (p=0.10)
- did not reach the usually accepted level of clinical significance (p=0.07)
- difference was apparent (p=0.07)
- direction heading towards significance (p=0.10)
- does not appear to be sufficiently significant (p>0.05)
- does not narrowly reach statistical significance (p=0.06)
- does not reach the conventional significance level (p=0.098)
- effectively significant (p=0.051)
- equivocal significance (p=0.06)
- essentially significant (p=0.10)
- extremely close to significance (p=0.07)
- failed to reach significance on this occasion (p=0.09)
- failed to reach statistical significance (p=0.06)
- fairly close to significance (p=0.065)
- fairly significant (p=0.09)
- falls just short of standard levels of statistical significance (p=0.06)
- fell (just) short of significance (p=0.08)
- fell barely short of significance (p=0.08)
- scarcely significant (0.05<p>0.1)
- significant at the .07 level
- significant tendency (p=0.09)
- significant to some degree (0<p>1)
- significant, or close to significant effects (p=0.08, p=0.05)
- significantly better overall (p=0.051)
- significantly significant (p=0.065)
- similar but not nonsignificant trends (p>0.05)
- slight evidence of significance (0.1>p>0.05)
- slight non-significance (p=0.06)
- slight significance (p=0.128)

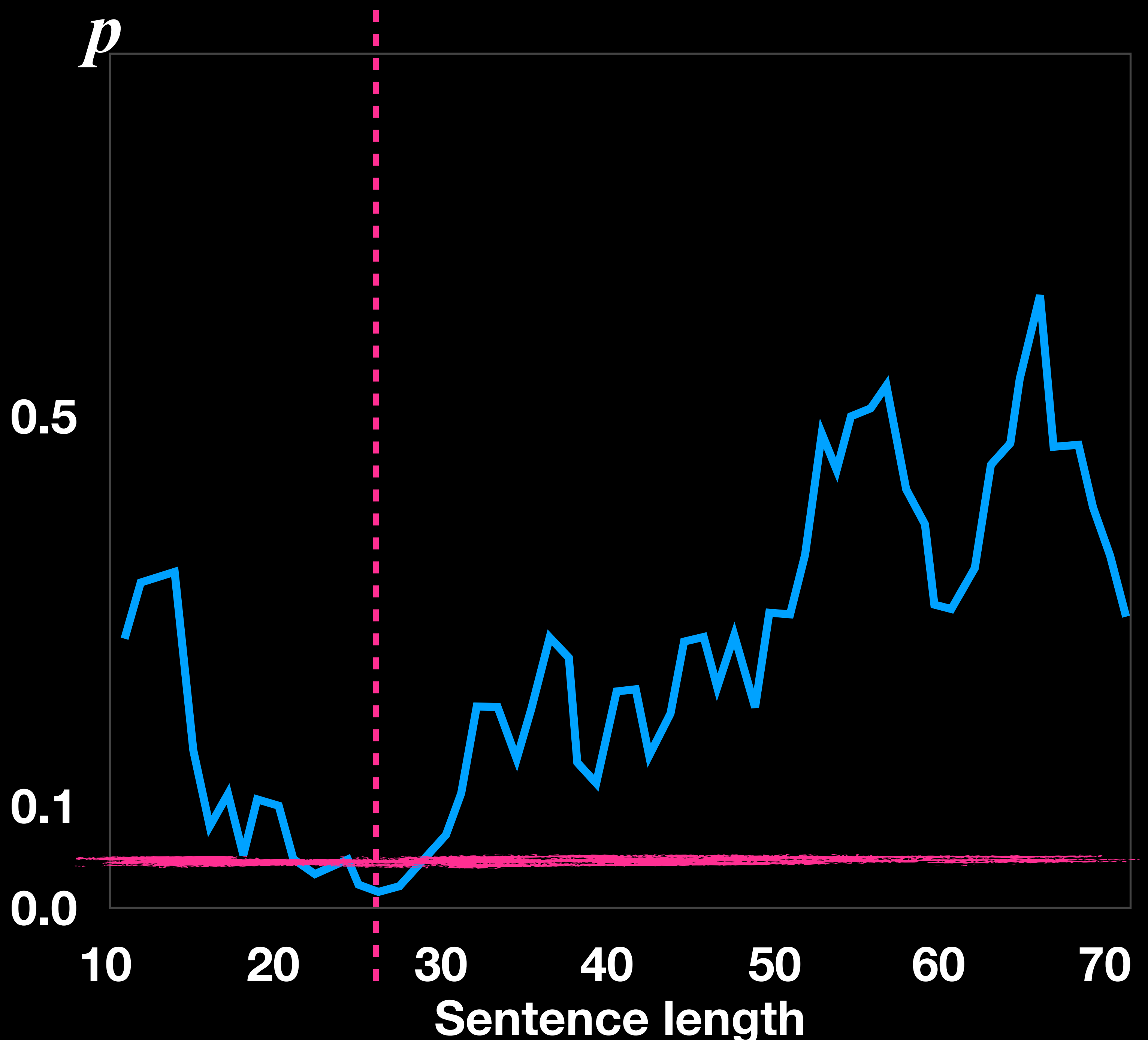
https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

Don't choose among metrics

metric	p
~~F1~~	0,0899
~~precision~~	0,062
~~recall~~	0,179
accuracy	0,0014

Report!

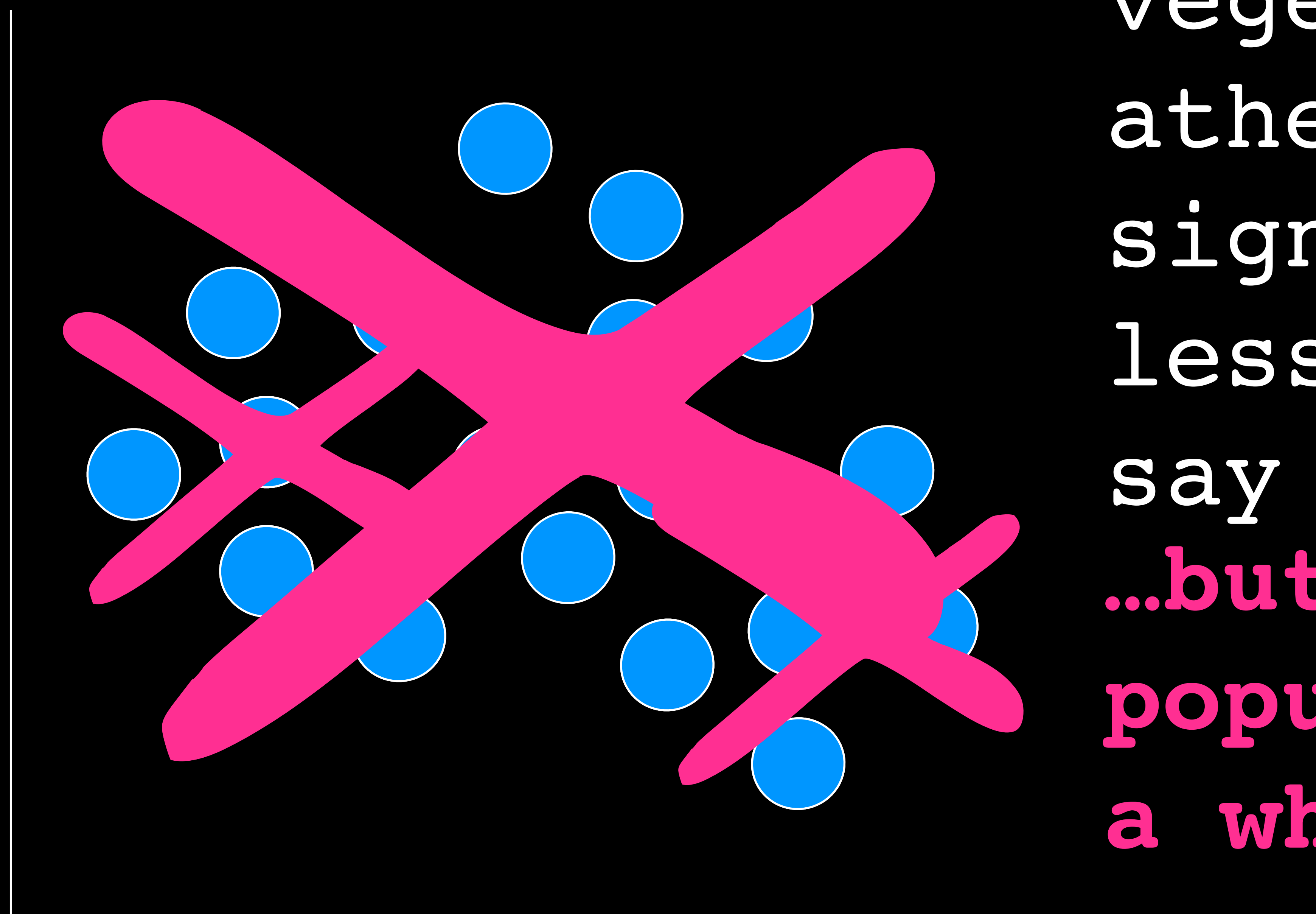
Don’t choose sample sizes



“We observed significant results at a sentence length of 26”

…but not with smaller or larger sentences!

Don't Choose Subsets

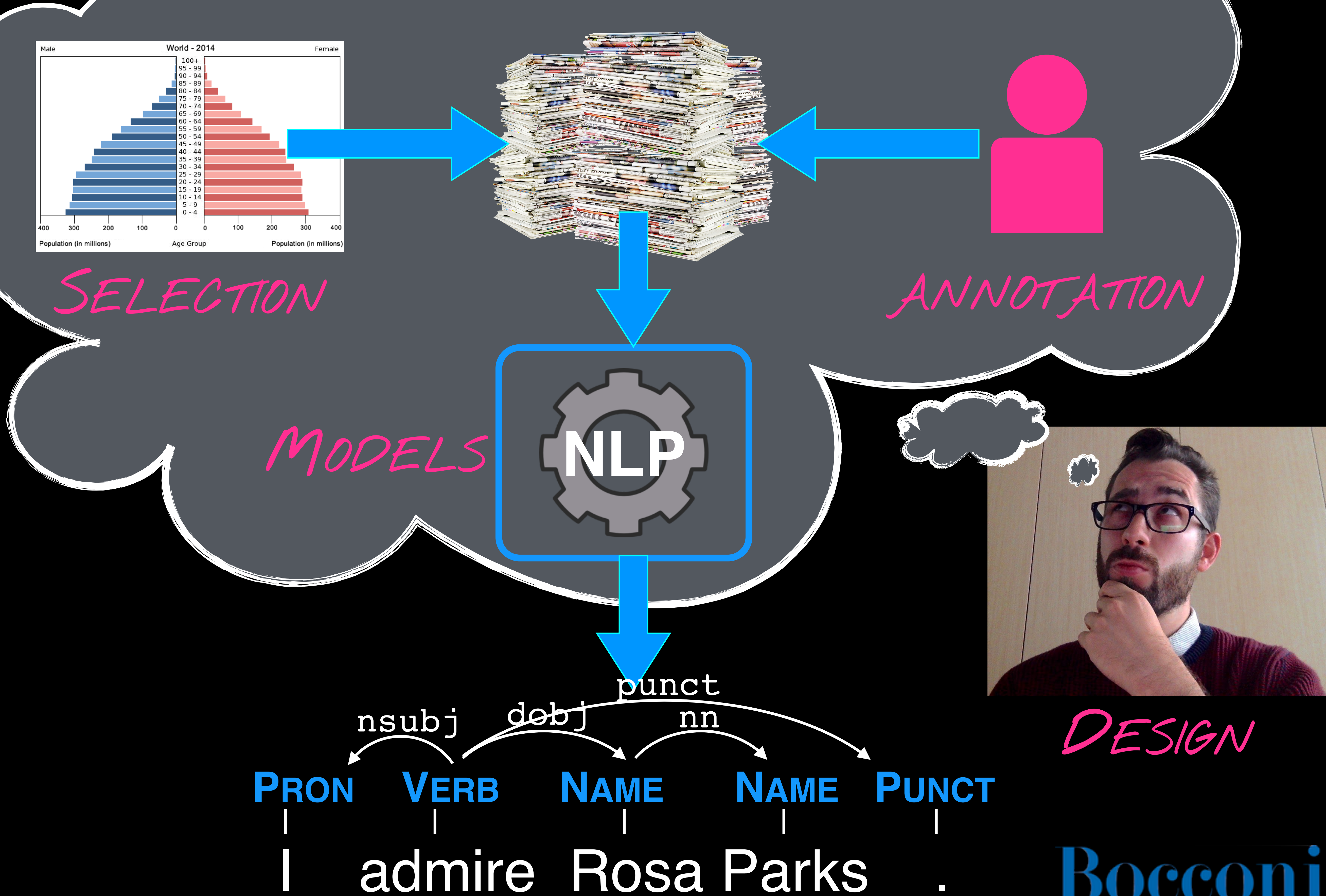


"Young, left-handed, vegetarian atheists are significantly less likely to say X"

…but the population as a whole isn't!

Wrapping Up

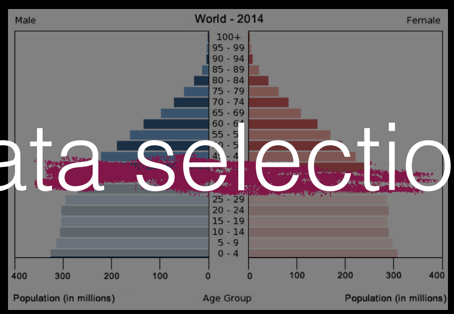
Sources of Bias



Bocconi

Hovy & Spruit (ACL 2016)

What can we do?

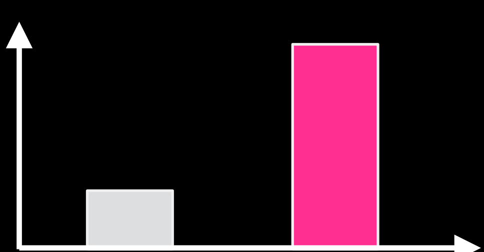
Source	Problem	Countermeasures
 data selection	**Exclusion**	stratification, priors
annotation	**Label Bias**	annotation models, disagreement weighting
models	**Overgeneralization**	dummy labels, error weighting, adversarial learning
research design	**Exposure**	always consider possible impact

Bocconi

The Goals

Fairness

Personalization



Performance

Take-home points

- Beware of **bias** from **data**, **models**, and **design**
- Apply **countermeasures** and check
- Ask yourself:
  "*Am I comfortable with my system classifying me?*"

Bocconi

www.dirkhovy.com/portfolio/papers

Thank you!

@dirk_hovy

www.dirkhovy.com

Bocconi

www.dirkhovy.com/portfolio/papers

Questions?

@dirk_hovy

www.dirkhovy.com

Bocconi