Natural Language Processing

Lecture 15

Dirk Hovy

dirk.hovy@unibocconi.it

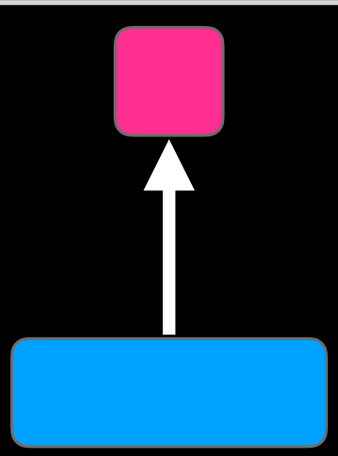
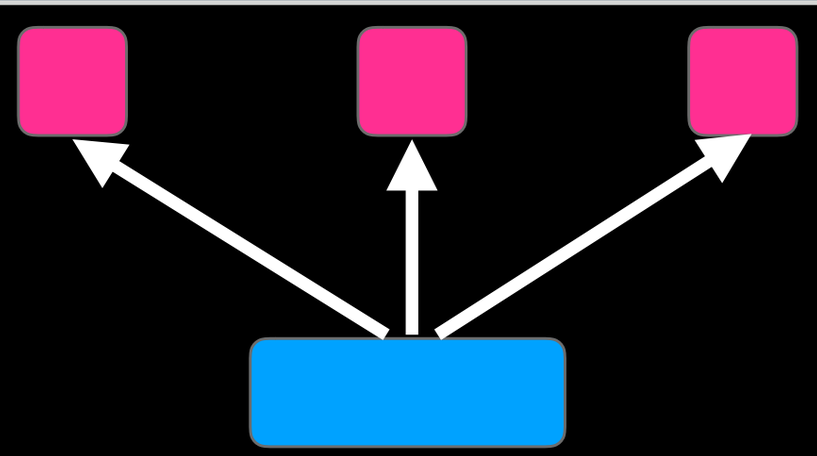
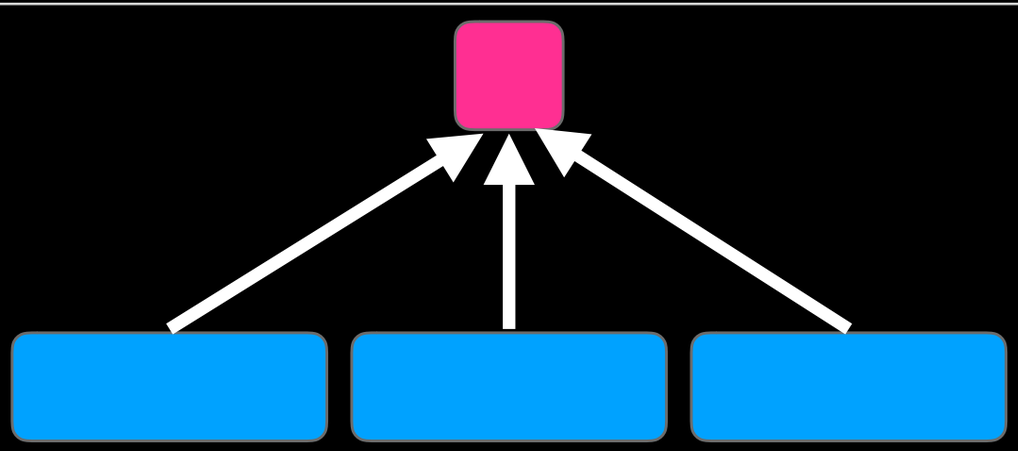
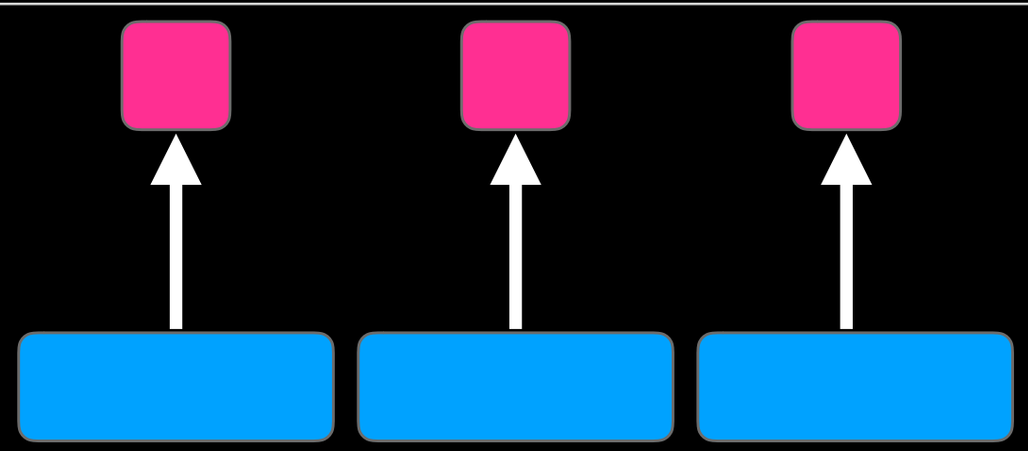
@dirk_hovy

Bocconi

Text Classification

Examples

Types of Text Classification

	Fixed length	Variable length
Fixed length	 Logistic Regression, Perceptron, Feed-Forward Network, Random Forest, Naive Bayes, SVM, …	 Multitask Learning, Decoder
Variable length	 Convolutional Neural Networks (CNN)	 Recurrent Neural Networks (RNN), Hidden Markov Models (HMM), Conditional Random Fields

Goals for Today

- Learn about **ethical issues**
- Understand potential **sources of bias**
- Learn about **counter measures**

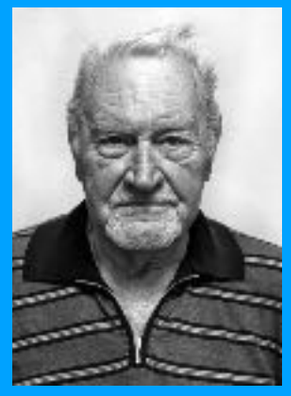
Ethical Considerations

Bocconi

Biased Language Systems

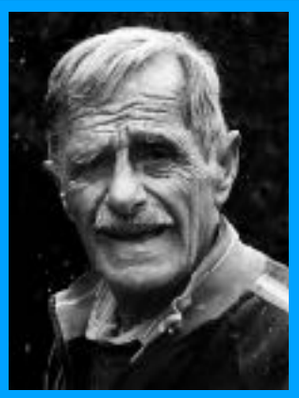
Bocconi

Language Biases



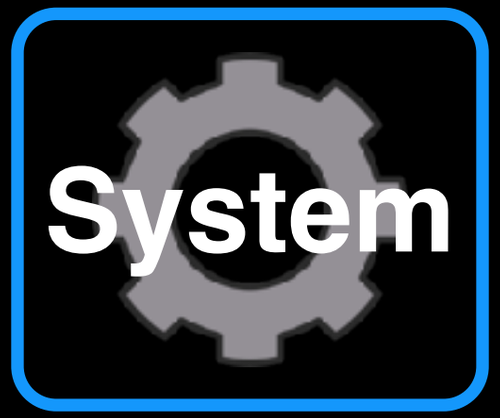
Example 1

Example 2



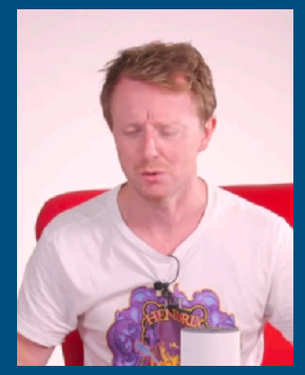
Example N

I don't understand you…



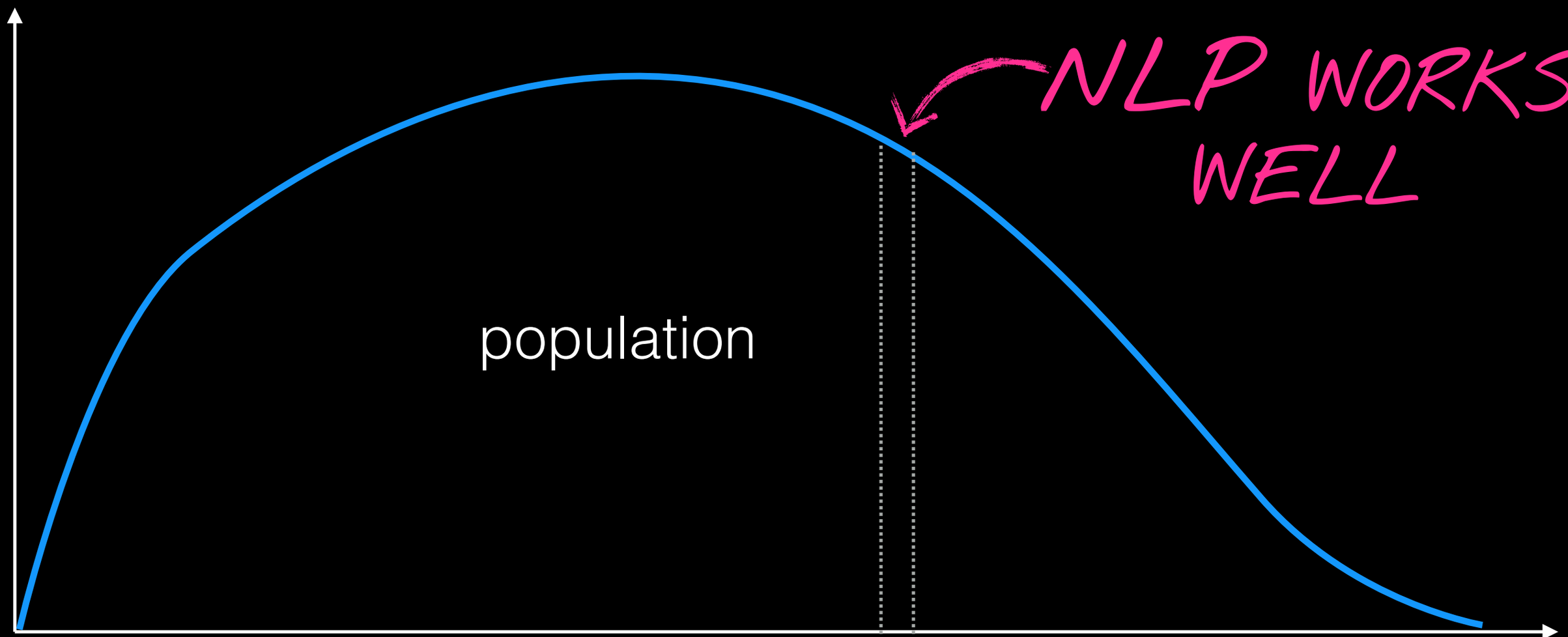
Hello, computer

Shite…



Bocconi

The Consequences



Bocconi

Solutions?

Symptoms

Example 1

Example 2

Example N

System

Cause

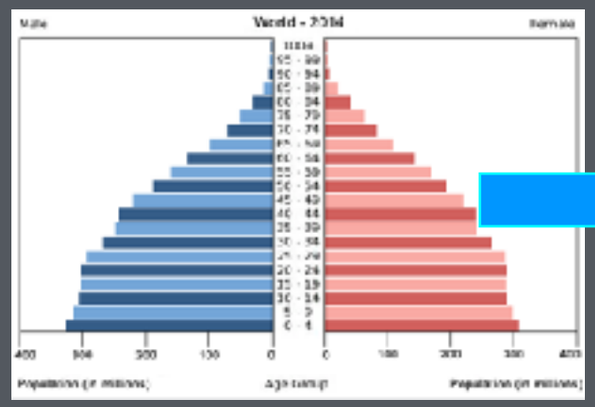
Bocconi

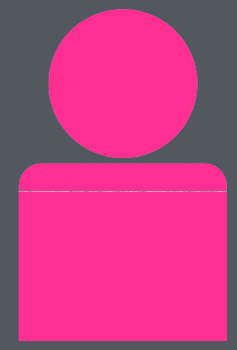
Who cares about bias?

Philosophers

Legal

Sources of Bias



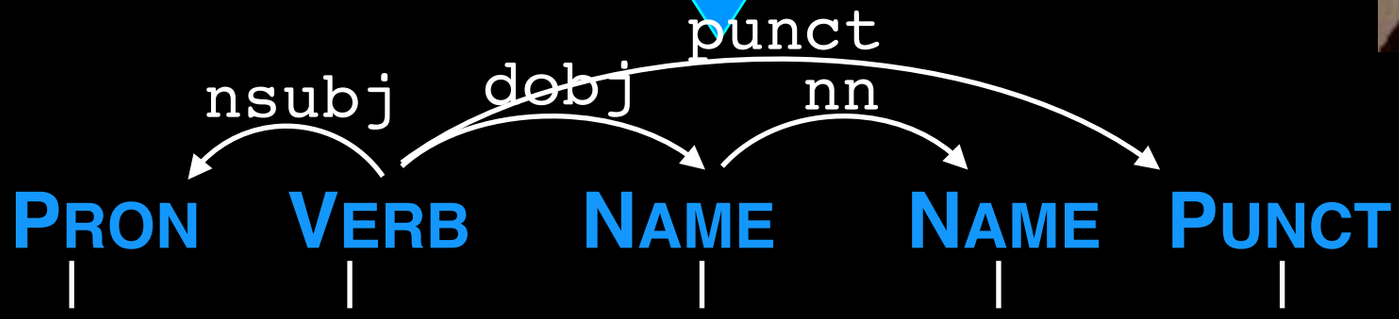


SELECTION

ANNOTATION

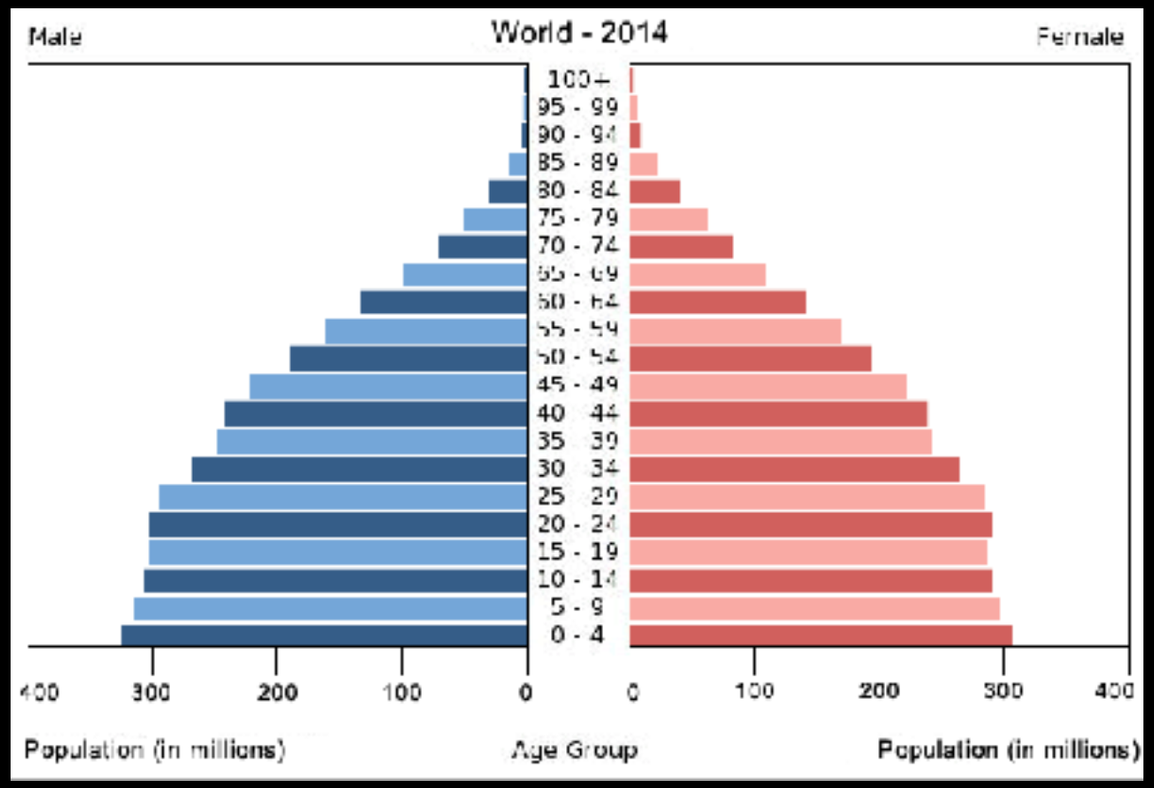
MODELS

DESIGN

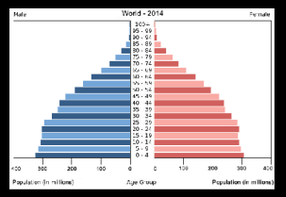


Bocconi

Data Bias

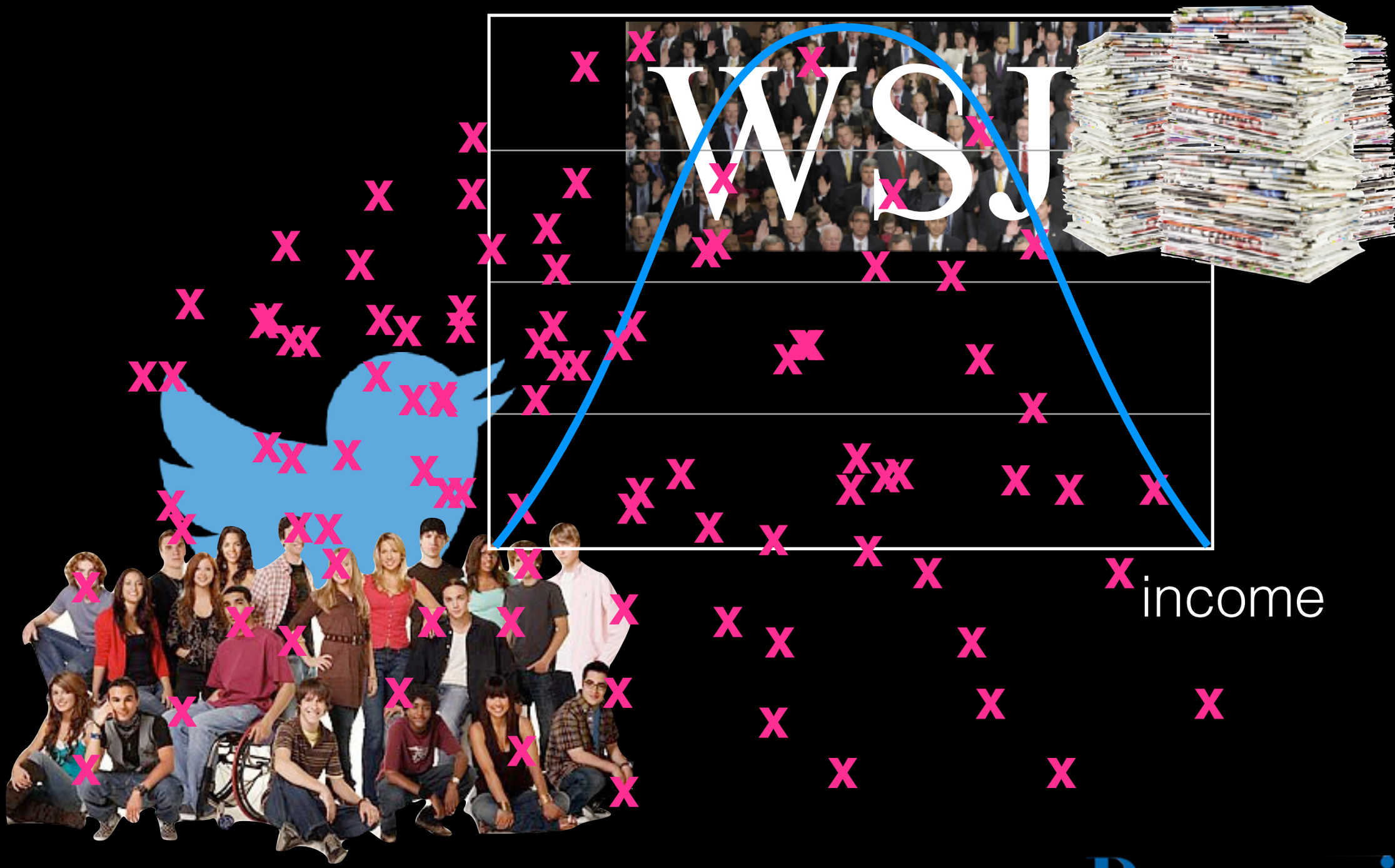


Bocconi



Distributions

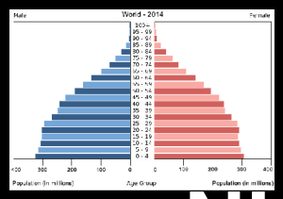
age



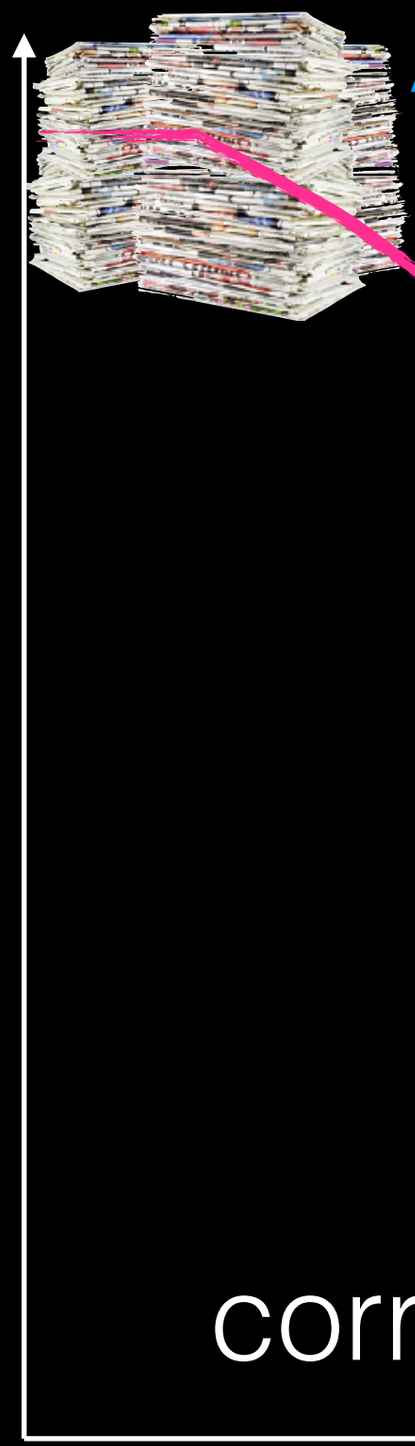
Bocconi

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

The WSJ Effect



NLP performance



ADV ADV ADJ ADV

Not definitely sure yet

ADV NOUN X ADV

Not deffo sho yet

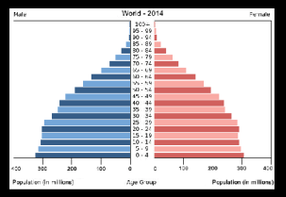
correlates w/ demographics

distance from “standard”

Bocconi

Exclusion

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



F1

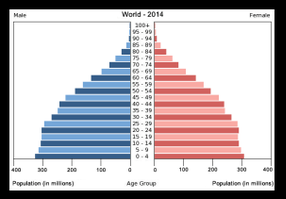
80

50

16

avg

Bocconi



Exclusion

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

accuracy

100

95

90

85

80

ing

views

ned

2

O45 U35 O45 U35

Bocconi

Annotator Bias

It's a particle!

No! It's an adposition!

PRON VERB **PRT** NOUN NUM

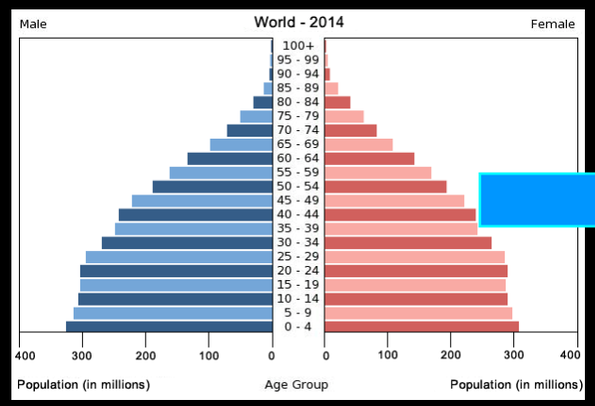
PRON VERB **ADP** NOUN NUM

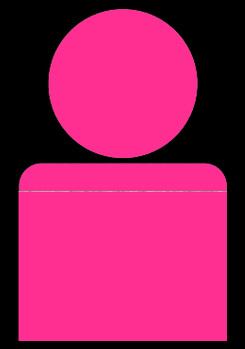
it comes out apr 30

Model Bias

Bocconi

Biased Models





SELECTION

ANNOTATION

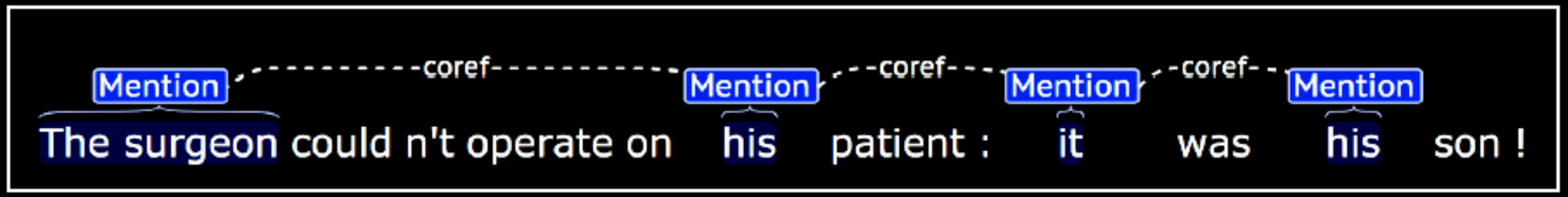
NLP

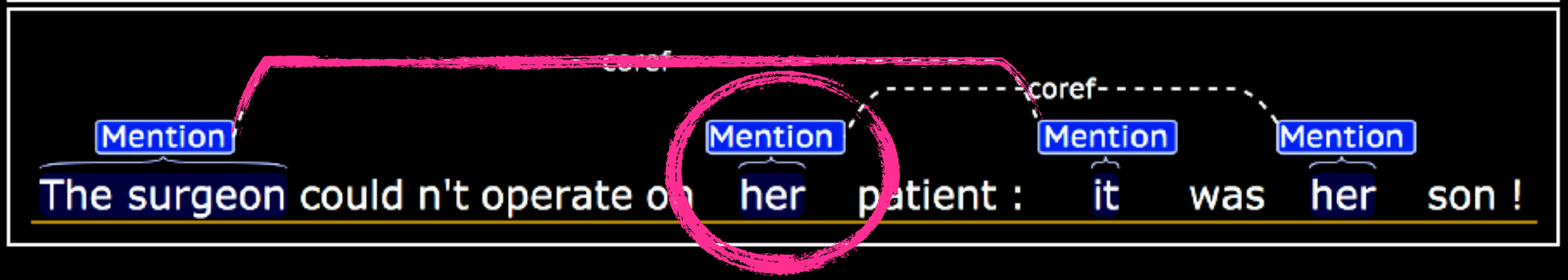
THIS IS REPRESENTATIVE

THIS IS RELIABLE

Rudinger et al. (2018)

Wrong Coreference





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***

Zhao et al. (EMNLP 2017)

Models Amplifying Bias

BIAS = 0.66

Agent: WOMAN

Agent: MAN

Agent: WOMAN

NLP

BIAS = 0.84

WOMAN

WOMAN

WOMAN

MAN

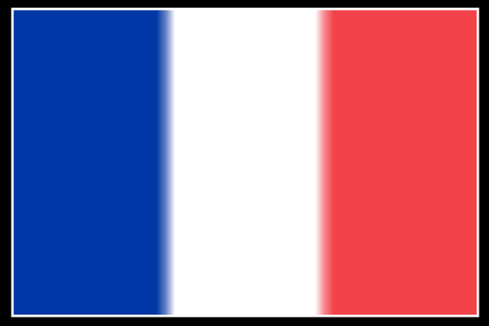
WOMAN

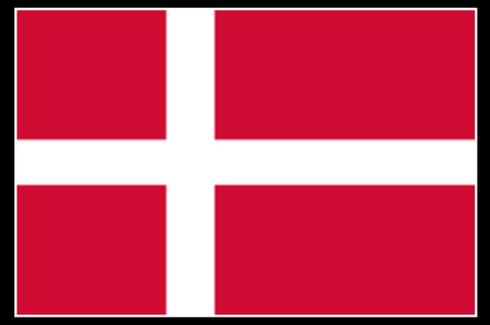
Bocconi

Design Bias

Bocconi

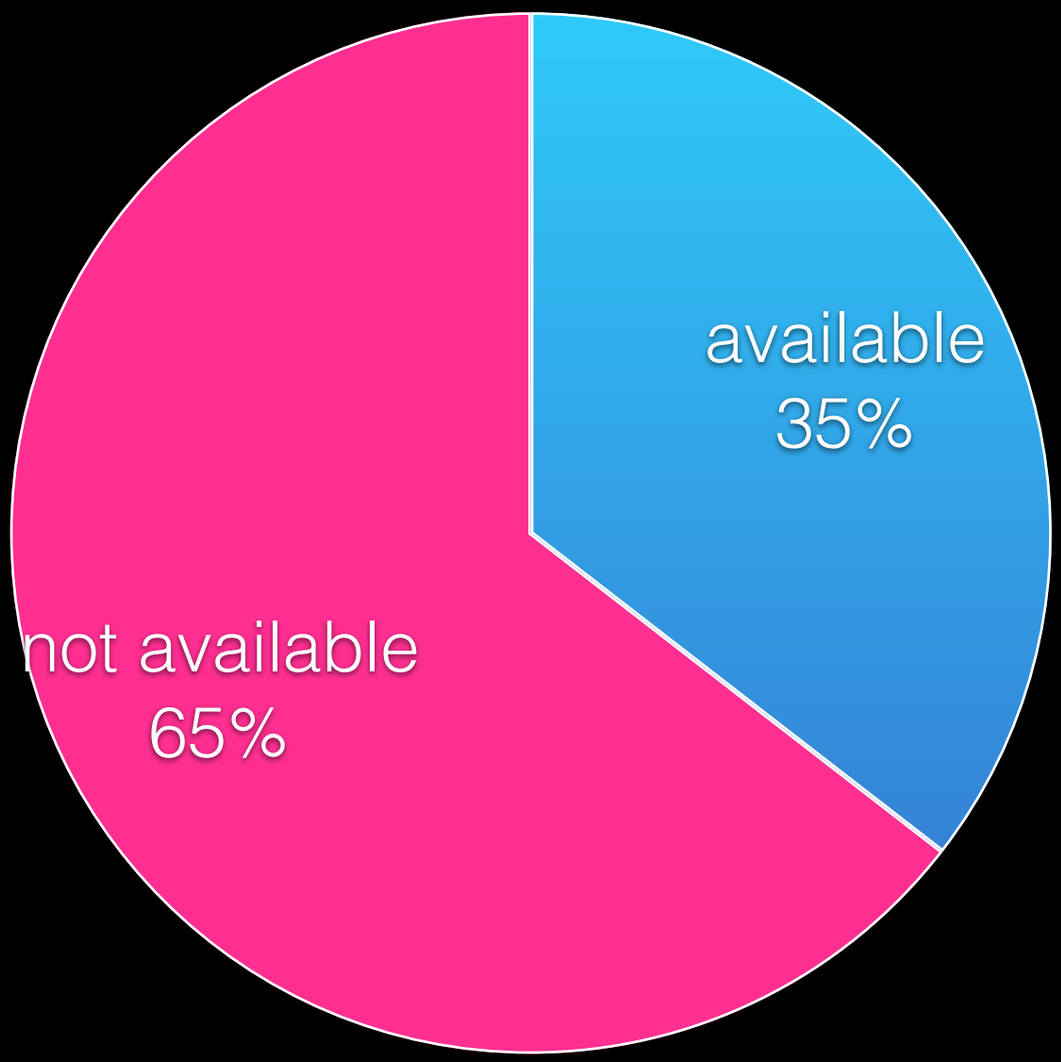
Exposure



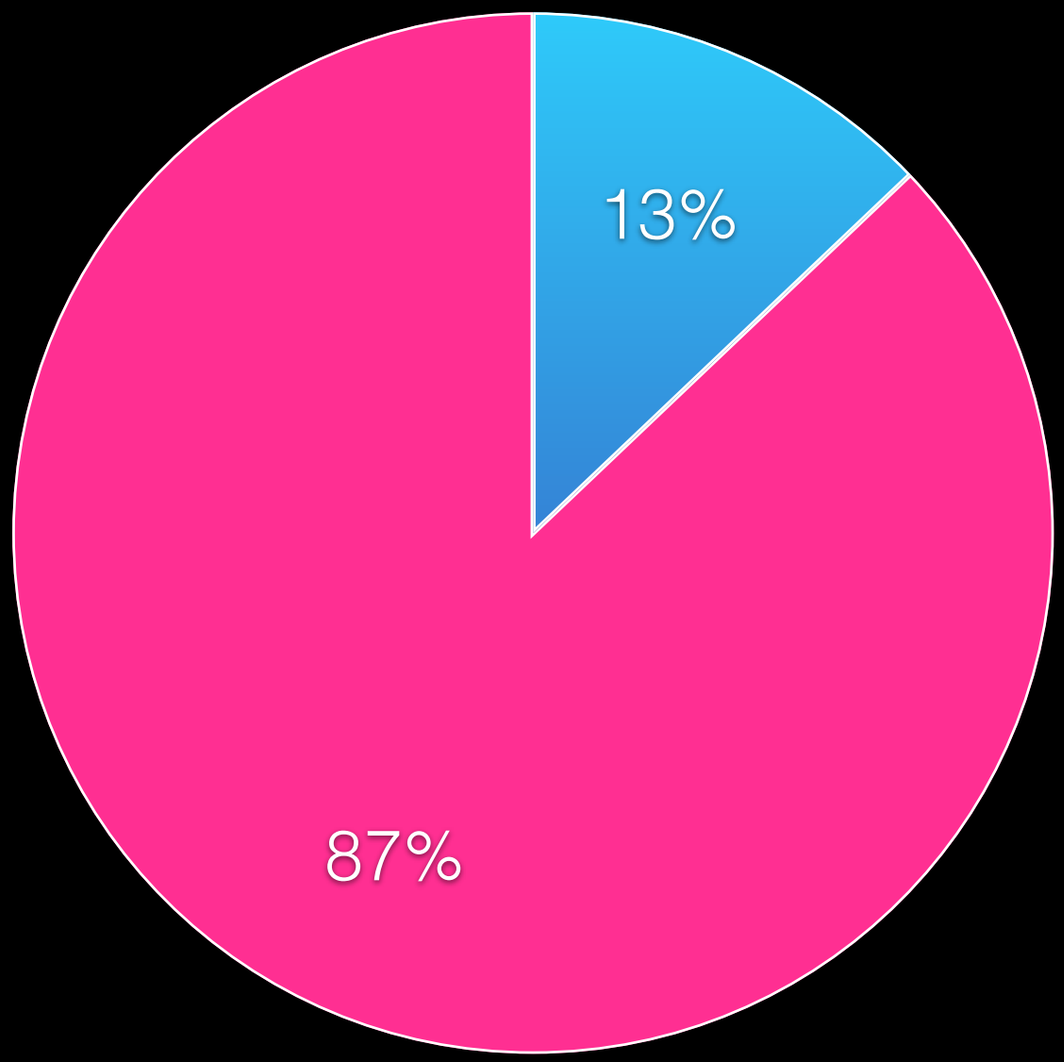


Under-Exposure

treebanks



semantic resources



evaluation

Over-Exposure

POS tagging

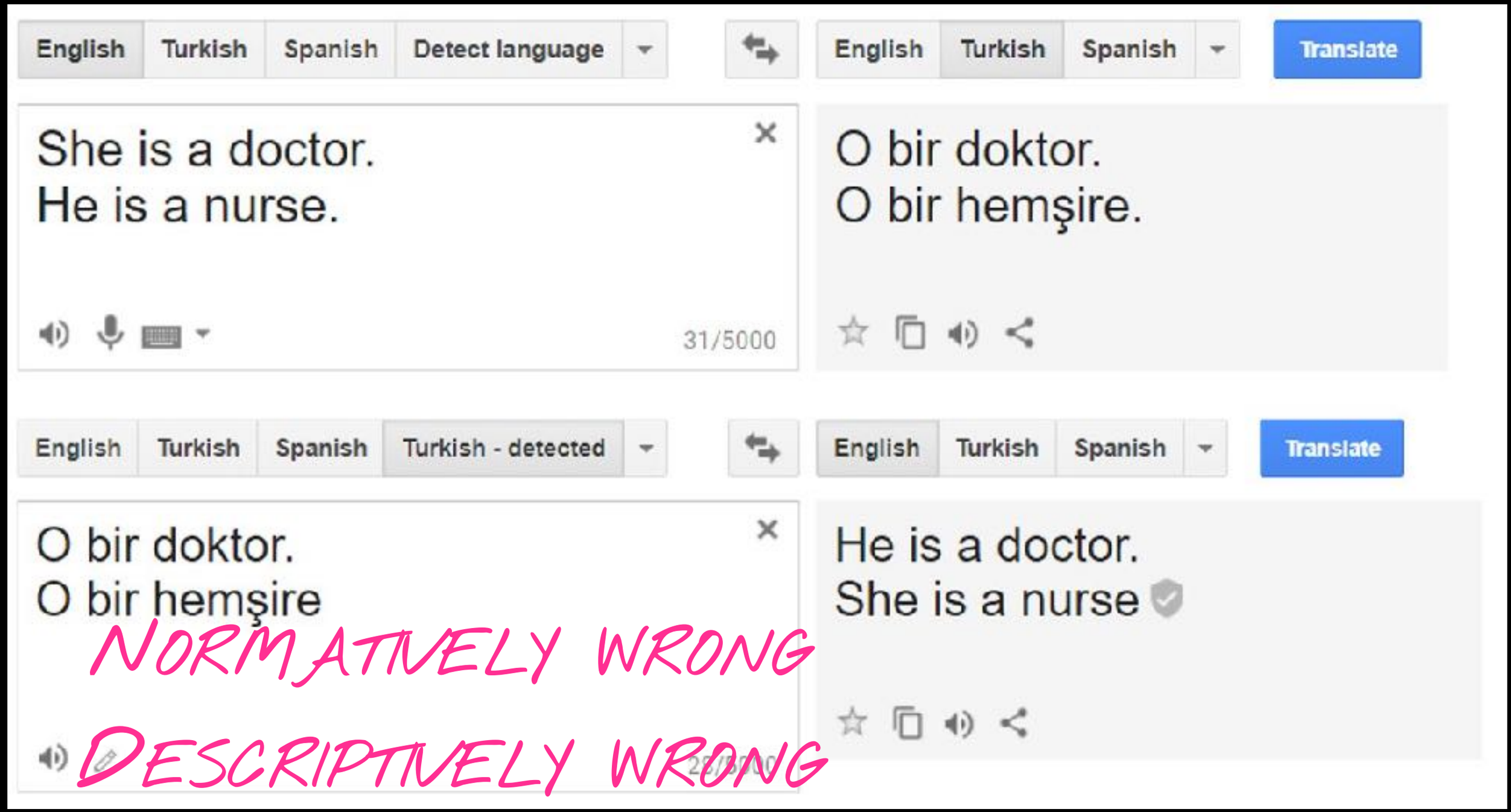
Discourse

Dual Use

Hovy & Spruit (ACL 2016)

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 **[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

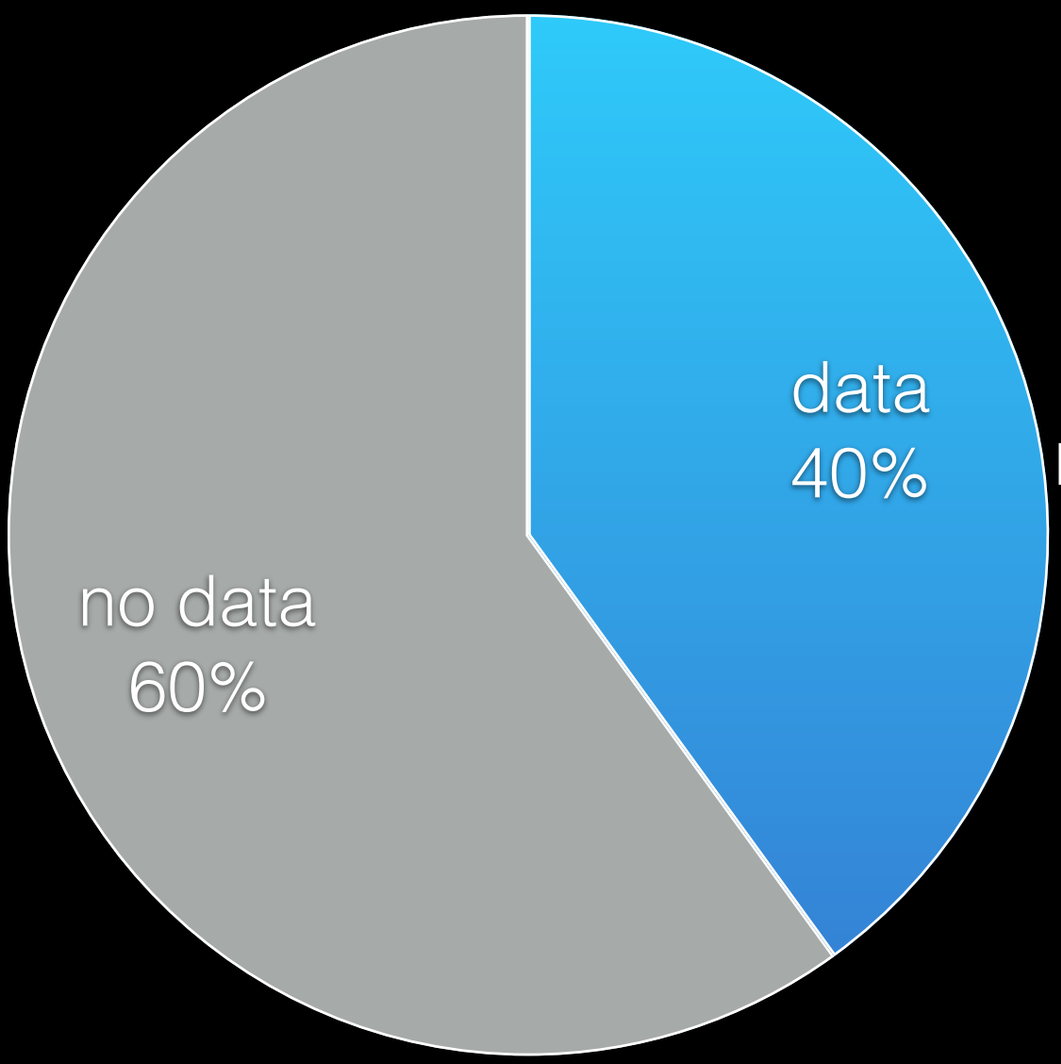
Unangemessene Vervollständigungen melden

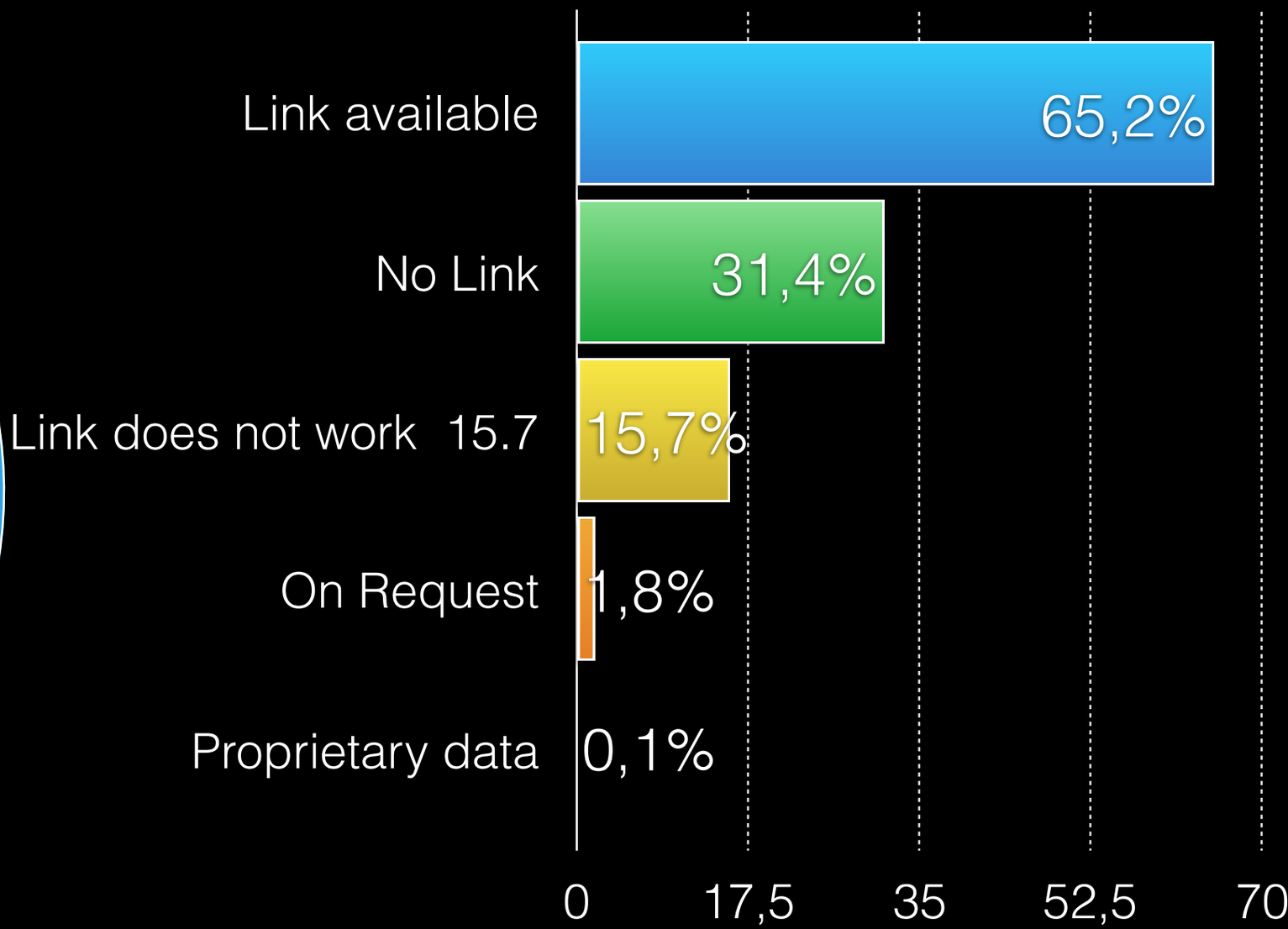
NORMATIVELY WRONG

DESCRIPTIVELY TRUE?

Bocconi

Replicability



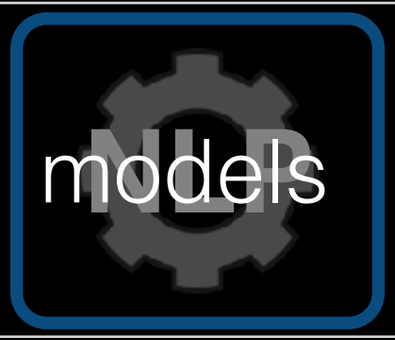


Bocconi

Counter Measures

Bocconi

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

Zhao et al. (EMNLP 2017)

Reducing Bias

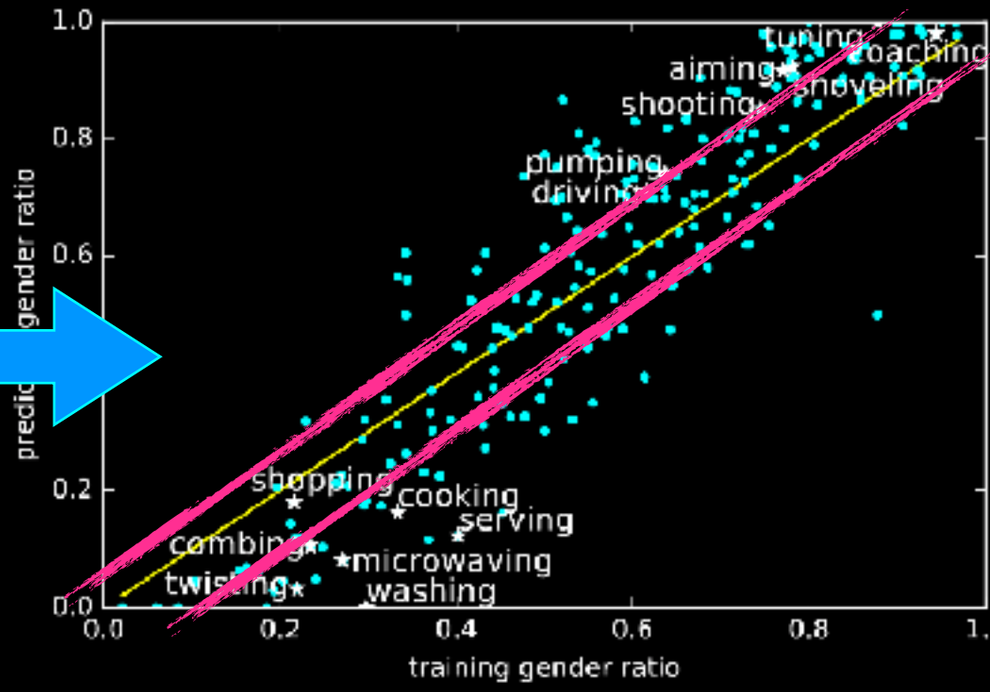
BIAS = 0.66

Agent: WOMAN

Agent: MAN

Agent: WOMAN

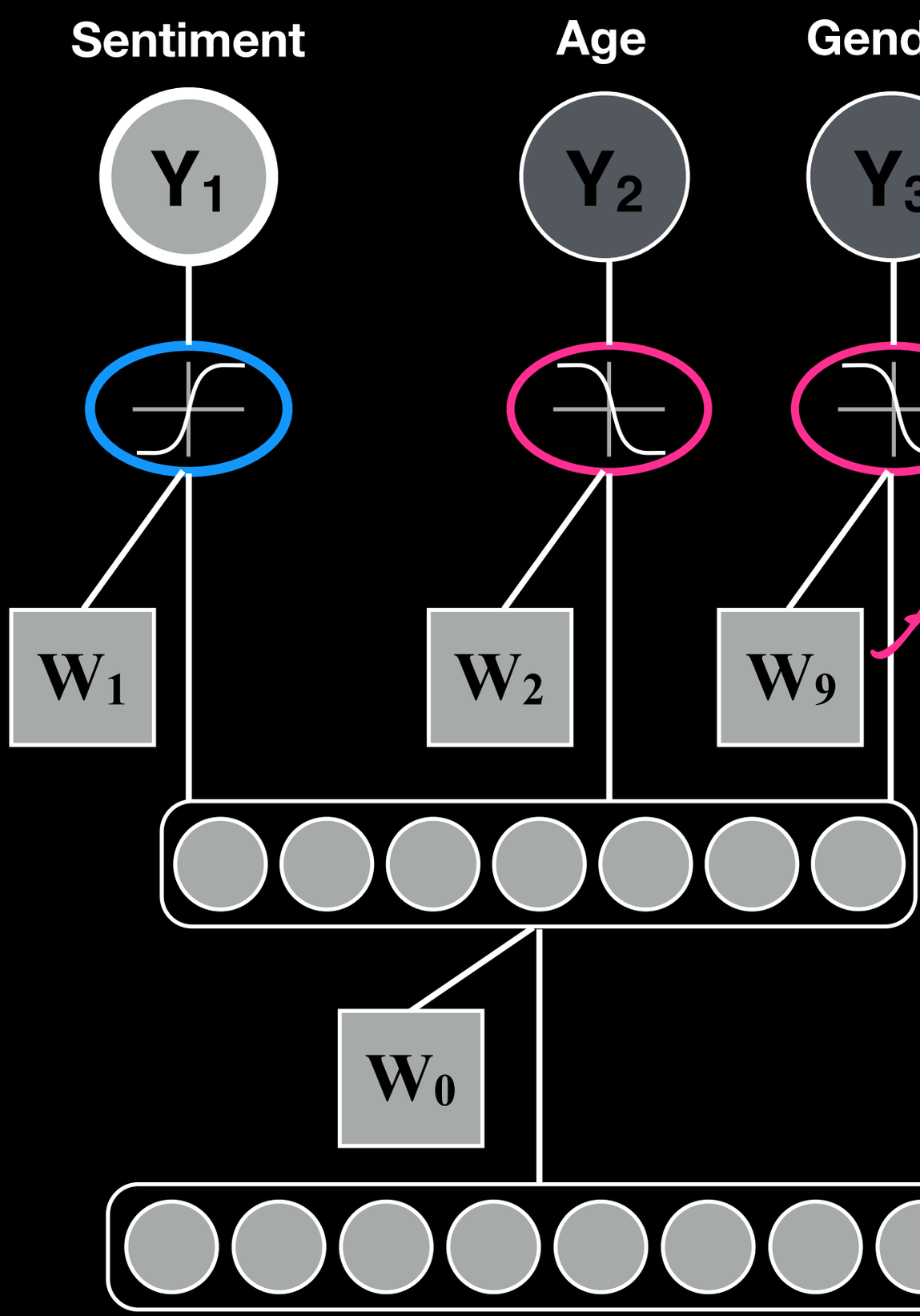
ILP



Bocconi

Li et al. (ACL 2018)

Adversarial Model



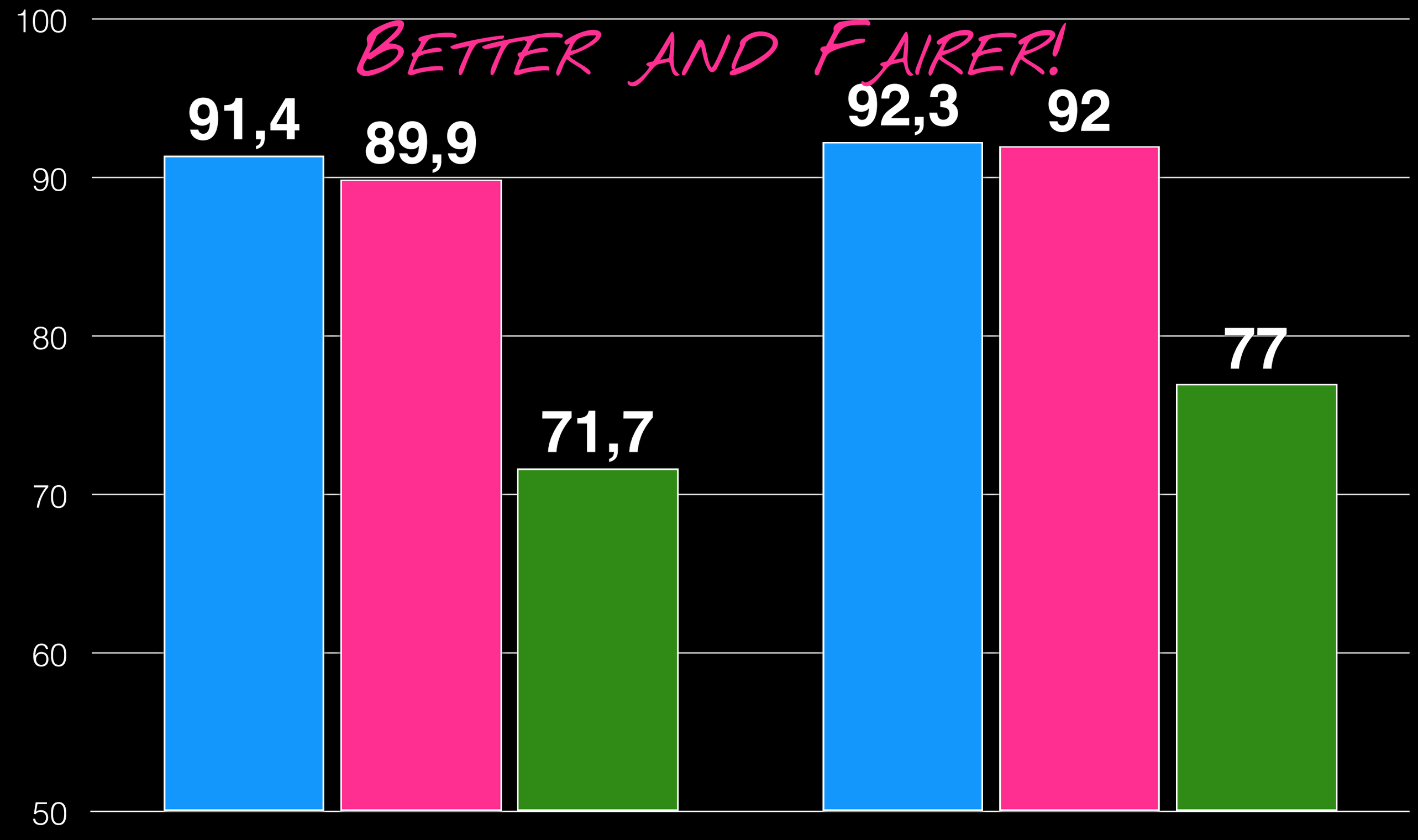
ADVERSARIAL LEARNING

Bocconi

Li et al. (ACL 2018)

Results

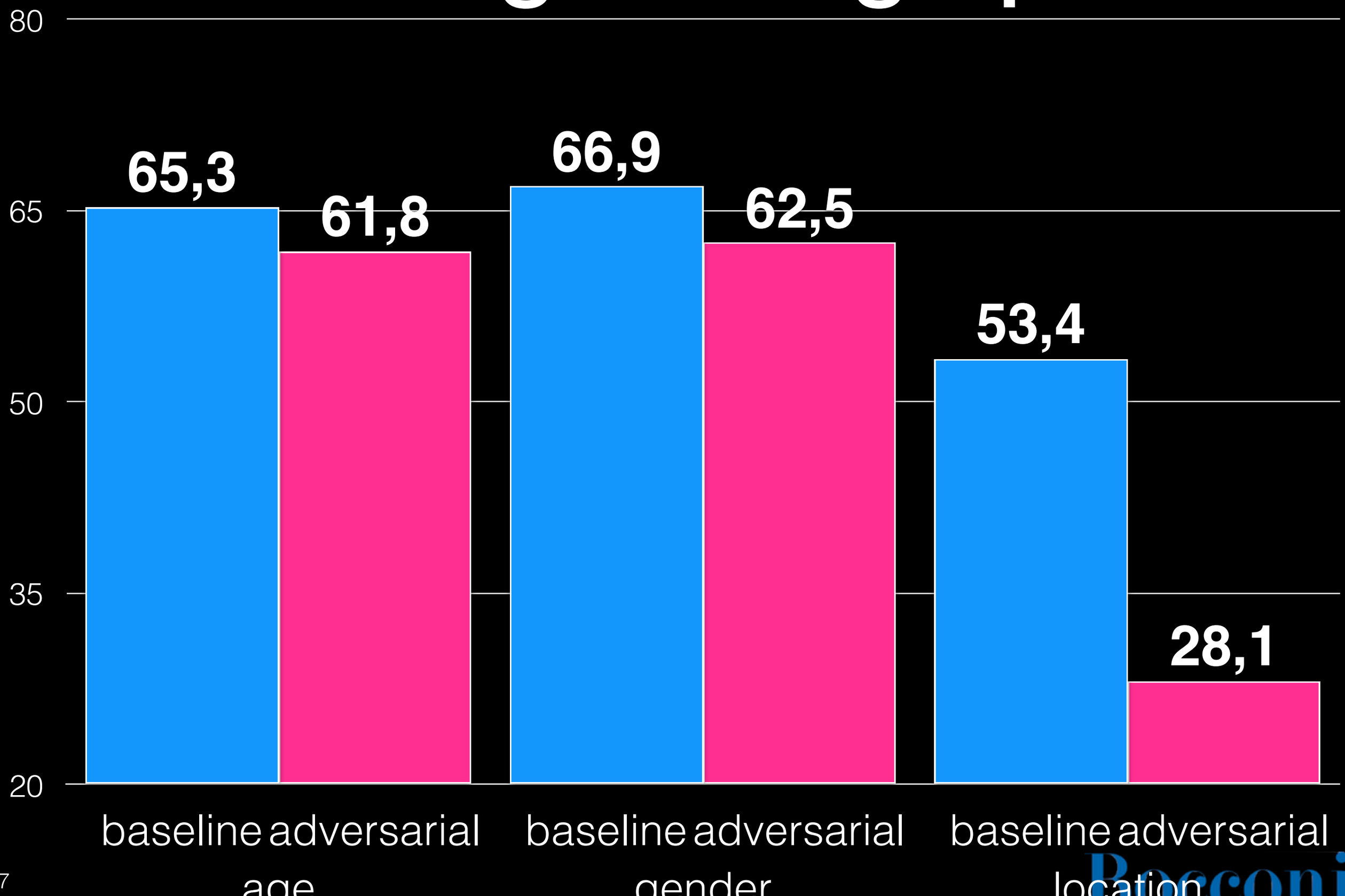
Better and Fairer!



Bocconi

Li et al. (ACL 2018)

Protecting Demographics

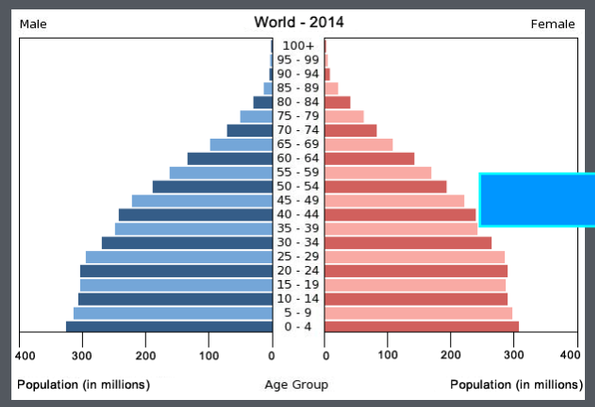


Bocconi

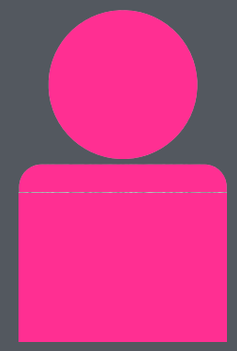
Wrapping Up

Bocconi

Sources of Bias



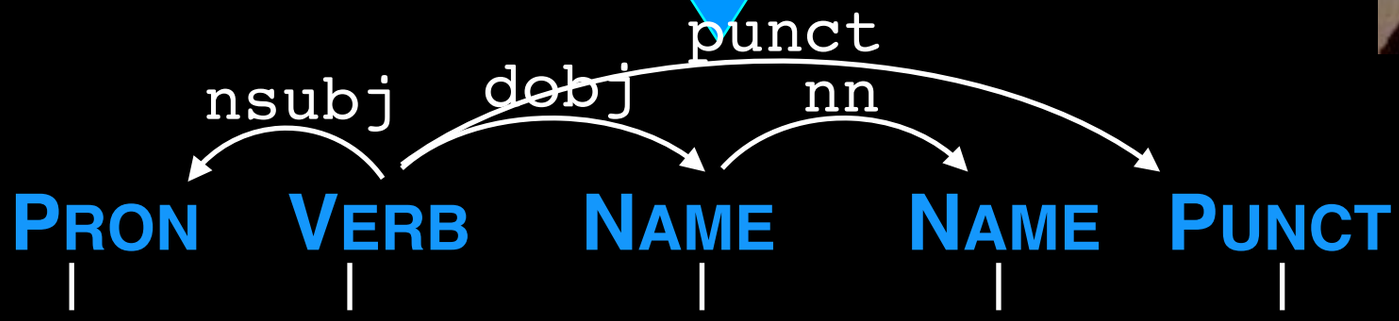
SELECTION



ANNOTATION

MODELS

DESIGN



Bocconi

Open Questions

- Carrots or sticks (GDPR)?
- Do we have to be more humble? How? Where?
- What about normative vs. descriptive ethics?
- Consequences for privacy, security, and user dignity?
- What if you have to choose between ethics and performance?
- Are we asking a lot from small companies?
- Research into bias creates biased data sets. Does it have to?
- Where does it stop: what about unknown sources of bias?

Take-home points

- Beware of **bias** from **data**, **models**, or **design**
- Apply **countermeasures** and check
- Ask yourself: "*Am I comfortable having my system classify myself?*"