

# Social & Ethical Considerations in NLP Systems

Yulia Tsvetkov

[yuliats@cs.washington.edu](mailto:yuliats@cs.washington.edu)



Carnegie Mellon University  
Language Technologies Institute

**W** PAUL G. ALLEN SCHOOL  
OF COMPUTER SCIENCE & ENGINEERING

# Plan

- Motivating examples & discussion
  - practical tools to assess AI systems adversarially
- Overview of topics in the intersection of ethics & NLP
  - scientific background on algorithmic bias and a high-level research overview
- Examples of research projects
  - deep-dive into one or two studies

# Language technologies

- Applications
  - Sentiment analysis
  - Machine translation
  - Information retrieval
  - Question answering
  - Dialogue systems
  - Summarization
  - Information extraction
  - ...
- Core technologies
  - Language modelling
  - Part-of-speech tagging
  - Syntactic parsing
  - Named-entity recognition
  - Coreference resolution
  - Word sense disambiguation
  - Semantic role labelling
  - ...

# Language & People

The common misconception is that language has to do with *words* and what they mean.

It doesn't.

It has to do with **people** and what *they* mean.



— Herbert H. Clark & Michael F. Schober, 1992

# Language technologies & People

The common misconception is that language has to do with *words* and what they mean.

It doesn't.

It has to do with **people** and what *they* mean.



Decisions we make about our data, methods, and tools are tied up with their impact on people and societies.

What are the ethical and social  
considerations for technologies we build?

# What is ethics?

“Ethics is a study of what are **good and bad** ends to pursue in life and what it is **right and wrong** to do in the conduct of life.

It is therefore, above all, a **practical discipline**.

Its primary aim is to determine how one ought to live and what actions one ought to do in the conduct of one’s life.”

— Introduction to Ethics, John Deigh

# What is ethics?

It's the **good** things

It's the **right** things

# What is ethics?

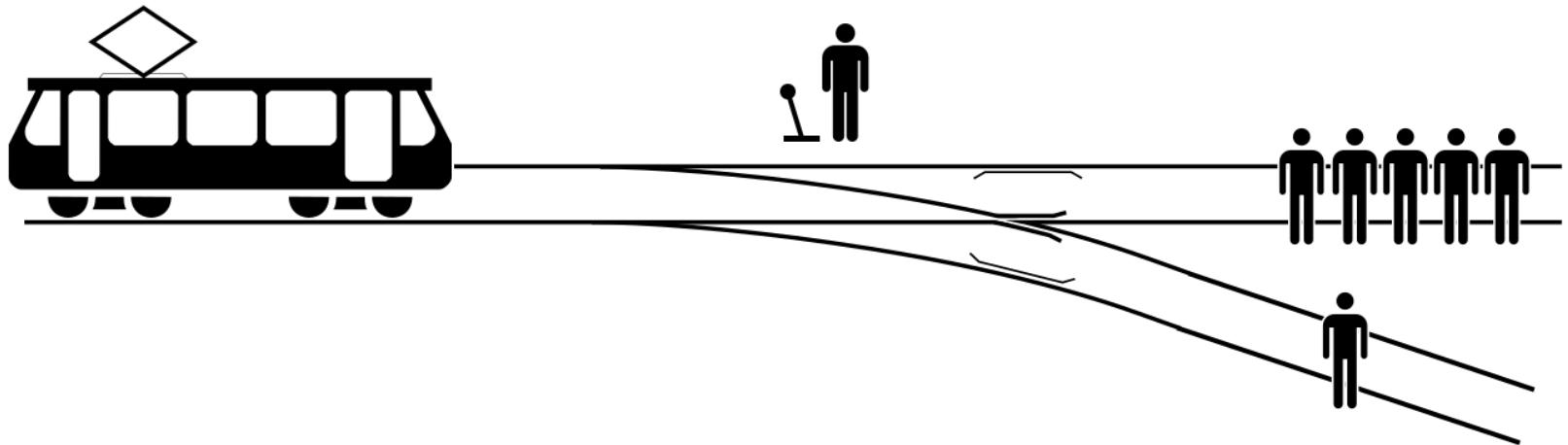
It's the **good** things

It's the **right** things

How simple is it to define  
what's good and what's right?

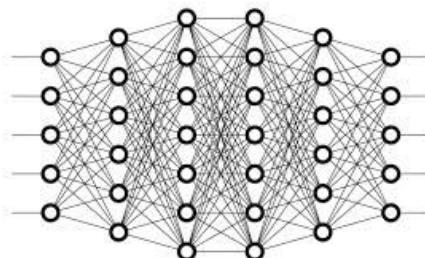
# The trolley dilemma

Should you pull the lever to divert the trolley?



[image from Wikipedia]

# The chicken dilemma

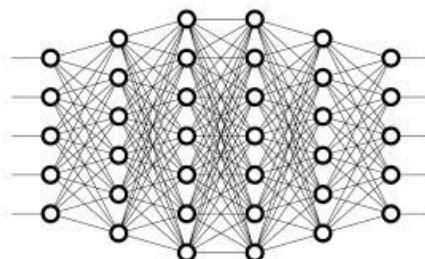


rooster

hen



# The chicken dilemma



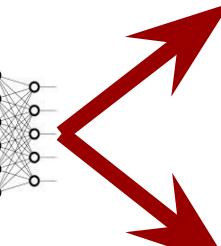
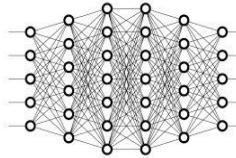
rooster

hen

Ethical?

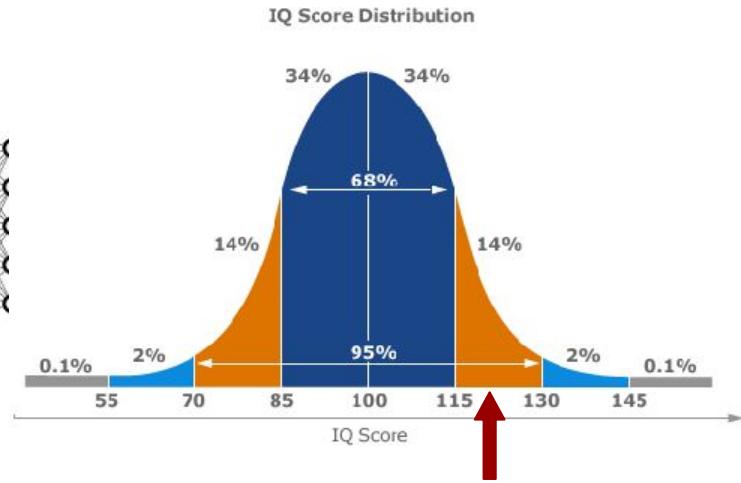
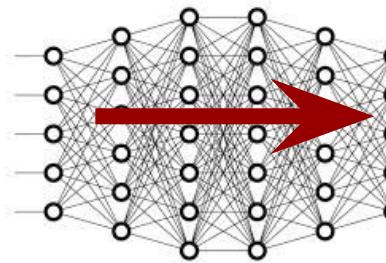
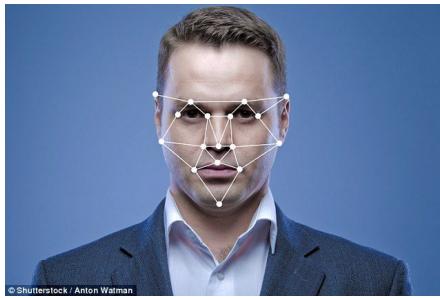


# The chicken dilemma



- Ethics is inner guiding, moral principles, and values of people and society
- Ethics isn't just "black and white", there are many gray areas.  
We often don't have easy answers.
- Ethics changes over time with values and beliefs of people
- Legal ≠ Ethical

# Let's train an IQ classifier



- Intelligence **Quotient**: a number used to express the apparent relative intelligence of a person

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos & texts.

- Who could benefit from such a classifier?

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos & texts.

- Who could benefit from such a classifier?
- Assume the classifier is 100% accurate. Who can be harmed from such a classifier? How can such a classifier be misused?

# An IQ Classifier

Let's train a classifier to predict people's IQ from their photos & texts.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Our test results show 90% accuracy
  - We found out that white females have 95% accuracy
  - People with blond hair under age of 25 have only 60% accuracy

# An IQ Classifier

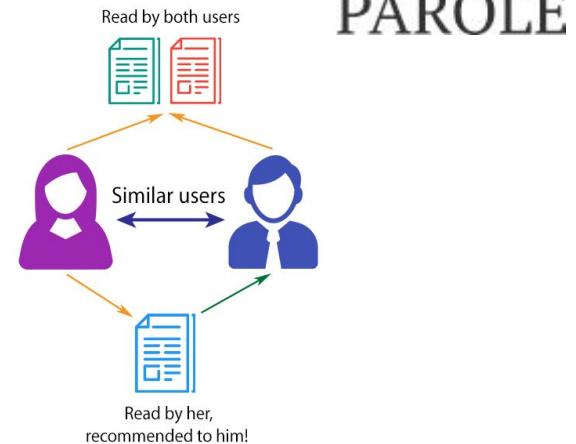
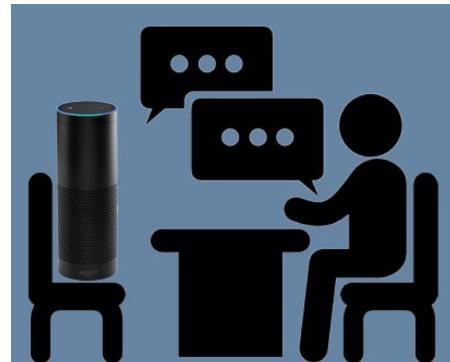
Let's train a classifier to predict people's IQ from their photos & texts.

- Who could benefit from such a classifier?
- Who can be harmed by such a classifier?
- Our test results show 90% accuracy
  - We found out that white females have 95% accuracy
  - People with blond hair under age of 25 have only 60% accuracy
- Who is responsible?
  - Researcher/developer? Advisor/manager? Reviewer? University? Society?

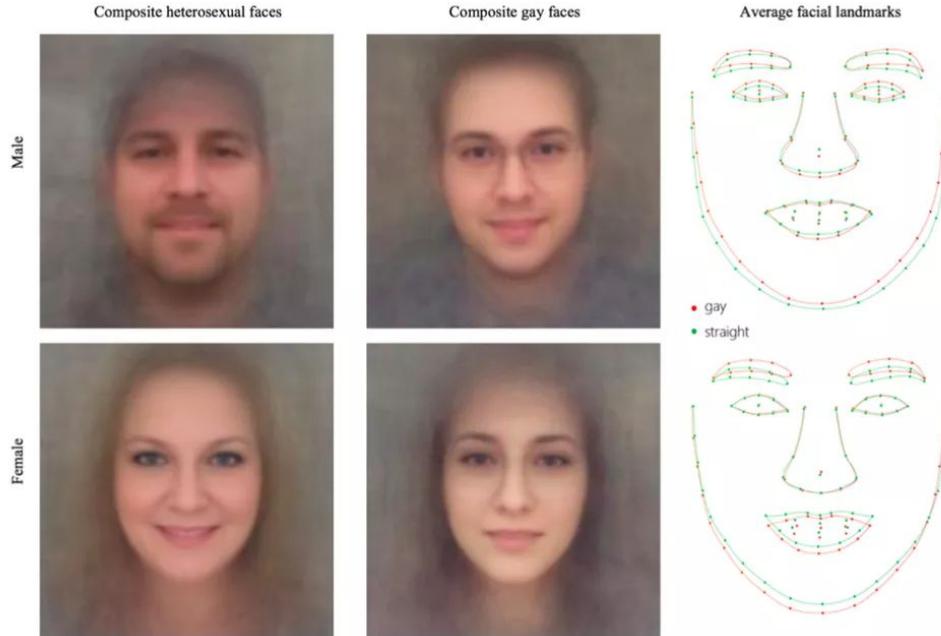
# What's the difference?



# AI and people



# A recent study: the “AI Gaydar”



# The “AI Gaydar” study

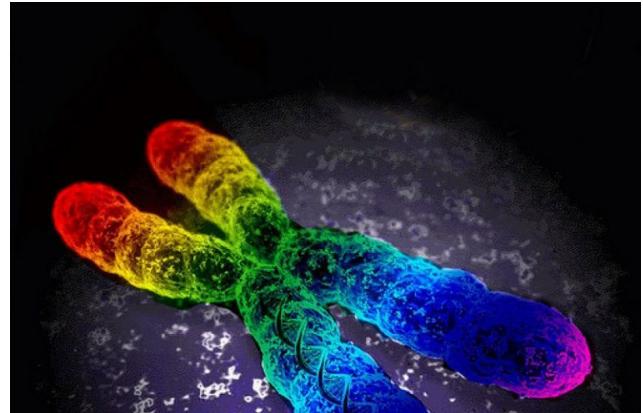
- Research question
  - Identification of sexual orientation from facial features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women

# Let's discuss...

- Research question
    - Identification of sexual orientation from facial features
  - Data
    - Photos downloaded from a popular American dating website
    - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
  - Method
    - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
  - Accuracy
    - 81% for men, 74% for women
- What went wrong?*

# Questioning the ethics of the research question

- Identification of sexual orientation from facial features



# Potential for dual use

## How people can be harmed by this research?

- In many countries being gay person is prosecutable (by law or by society) and in some places there is even death penalty for it
- It might affect people's employment; family relationships; health care opportunities;
- Attributes like gender, race, sexual orientation, religion are social constructs. Some may change over time. They can be non-binary. They are private, intimate, often not visible publicly.
- Importantly, these are properties for which people are often discriminated against.

# Dual use and dual framing in predictive analytics



*"We live in a dangerous world, where harm doers and criminals easily mingle with the general population; the vast majority of them are unknown to the authorities.*

*As a result, it is becoming ever more challenging to detect anonymous threats in public places such as airports, train stations, government and public buildings and border control. Public Safety agencies, city police department, smart city service providers and other law enforcement entities are increasingly strive for Predictive Screening solutions, that can monitor, prevent, and forecast criminal events and public disorder without direct investigation or innocent people interrogations. "*

# Data

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

# Data

- Photos downloaded from a popular American dating website



# Data & privacy

- Photos downloaded from a popular American dating website

Legal ≠ Ethical

Public ≠ Publicized

Did these people agree to participate in the study?

→ Violation of social contract

# Data

- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly



# Bias in data

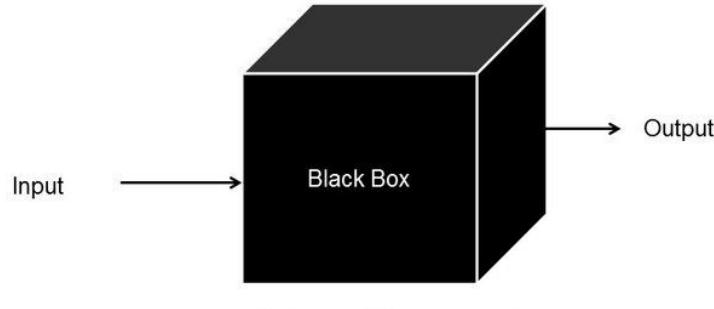
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Only white people  
who self-disclose their orientation,  
certain social groups,  
certain age groups,  
certain time range/fashion;  
the photos were carefully selected by subjects to be attractive  
→ this dataset contains many types of biases

The dataset is balanced, which does not represent true class distribution.

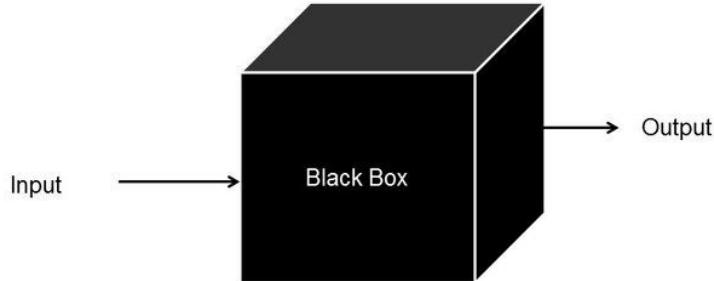
# Method

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification



# Unveiling biases in black-box models

- A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification



- can we use not interpretable models when we make predictions about sensitive attributes, about complex experimental conditions that require broader world knowledge?
- how to analyze errors and bias amplification?

# Evaluation

- Accuracy: 81% for men, 74% for women

# The cost of misclassification and the importance of social context



# The cost of misclassification and the importance of social context



# Learn to assess AI systems adversarially

- **Ethics** of the research question
- **Impact of technology and potential dual use:** Who could benefit from such a technology? Who can be harmed by such a technology? Could sharing data and models have major effect on people's lives?
- **Privacy:** Who owns the data? Published vs. publicized? User consent and implicit assumptions of users how the data will be used.
- **Bias in data:** Artifacts in data, population-specific distributions, representativeness of data.
- **Bias in models:** How to control for confounding variables and corner cases? Does the system optimize for the “right” objective? Does the system amplify bias?
- **Utility-based evaluation beyond accuracy:** FP & FN rates, “the cost” of misclassification, fault tolerance.

# Why is it especially relevant now?

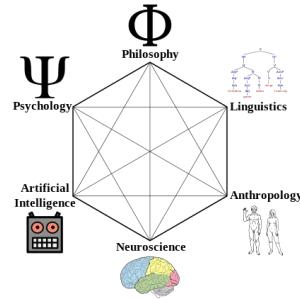
- Data: the exponential growth of user-generated content
- Tools: machine learning tools have become ubiquitous and accessible to everyone

# Recommended papers and talks

- Hovy & Spruit (2016) [The Social Impact of NLP](#)
- Barocas & Selbst (2016) [Big Data's Disparate Impact](#)
- Barbara Grosz talk: [Intelligent Systems: Design & Ethical Challenges](#)
- Kate Crawford NeurIPS keynote: [The Trouble with Bias](#)
- Yonatan Zunger blog post: [Asking the Right Questions About AI](#)

# Topics in the intersection of Ethics & NLP

- **Algorithmic bias:** social bias in data & NLP models
- **Incivility:** Hate-speech, toxicity, incivility, microaggressions online
- **Privacy violation:** Privacy violation & language-based profiling
- **Misinformation:** Fake news, information manipulation, opinion manipulation
- **Technological divide:** Unfair NLP technologies, underperforming for speakers of minority dialects, for languages from developing countries, and for disadvantaged populations



# Recommended resources

- Computational ethics in NLP lectures, readings  
[http://demo.clab.cs.cmu.edu/ethical\\_nlp/](http://demo.clab.cs.cmu.edu/ethical_nlp/)
- CS 384: Ethical and Social Issues in NLP  
<https://web.stanford.edu/class/cs384/>
- ACL Ethics resources  
[https://aclweb.org/aclwiki/Ethics\\_in\\_NLP](https://aclweb.org/aclwiki/Ethics_in_NLP)

# Algorithmic bias: social bias in data & models

Which word is more likely to be used by a **female** ?

**Giggle – Laugh**

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **female** ?

**Giggle – Laugh**

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **female** ?

**Brutal – Fierce**

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **female** ?

**Brutal – Fierce**

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **older person** ?

**Impressive – Amazing**

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a **older person** ?

**Impressive – Amazing**

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a person of **higher occupational class** ?

## **Suggestions – Proposals**

(Preotiuc-Pietro et al. '16)

Which word is more likely to be used by a person of **higher occupational class** ?

## **Suggestions – Proposals**

(Preotiuc-Pietro et al. '16)

Why do we intuitively recognize  
a default social group?

Why do we intuitively recognize  
a default social group?

Implicit bias

# How do we make decisions

## System 1

automatic

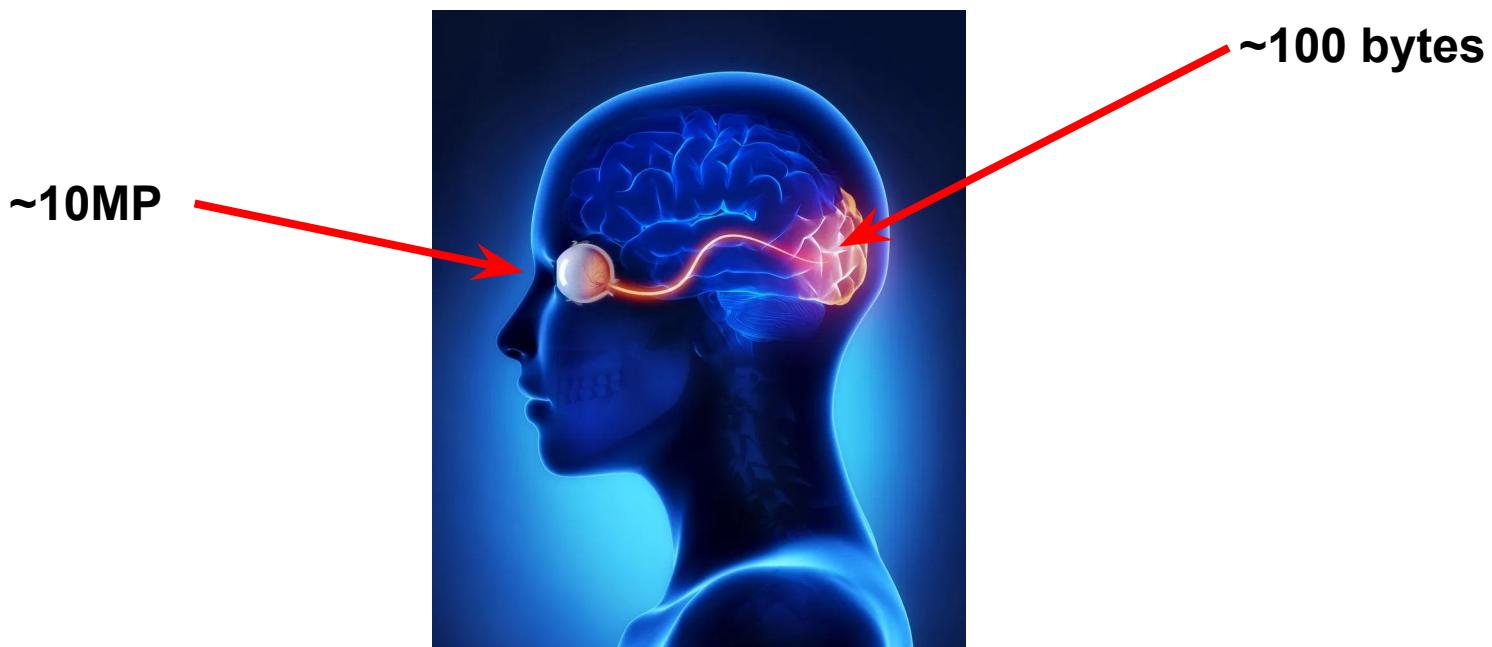
fast  
parallel  
automatic  
effortless  
associative  
slow-learning

## System 2

effortful

slow  
serial  
controlled  
effort-filled  
rule-governed  
flexible

# Why?



# System 1 is responsible for most of our decisions

## System 1

automatic

## System 2

effortful

Our brains are evolutionarily hard-wired to store learned information for rapid retrieval and automatic judgments. Over 95% of cognition is relegated to the System 1 “auto-pilot.”

# Psychological perspective on cognitive bias

Biases inevitably form because of the innate tendency of the human mind to:

- **Categorize** the world to simplify processing
- **Store** learned information in mental representations (called schemas)
- Automatically and unconsciously **activate** stored information whenever one encounters a category member

**Cognitive bias** is a systematic pattern of deviation from rationality in judgement

# COGNITIVE BIAS CODEX

## What Should We Remember?

To avoid mistakes,  
we aim to preserve autonomy  
and group status, and avoid  
irreversible decisions

To get things done, we tend  
to complete things we've  
invested time & energy in

To stay focused, we favor the  
immediate, relatable thing  
in front of us

## Need To Act Fast

To act, we must be confident we  
can make an impact and feel what  
we do is important

We store memories differently based  
on how they were experienced

We reduce events and lists  
to their key elements

We discard specifics  
to form generalities

We edit and reinforce  
some memories after the fact

We favor simple-looking options  
and complete information over  
complex, ambiguous options



We project our current mindset and  
assumptions onto the past and future

We notice things already primed in  
memory or repeated often

Bizarre, funny, visually-striking, or  
anthropomorphic things stick out more  
than non-bizarre/unfunny things

We notice when  
something has changed

## Too Much Information

We are drawn to details  
that confirm our own  
existing beliefs

We notice flaws in others  
more easily than we  
notice flaws in ourselves

We tend to find stories and  
patterns even when looking  
at sparse data

We fill in characteristics from  
stereotypes, generalities,  
and prior histories

We imagine things and people  
we're familiar with or fond of  
as better

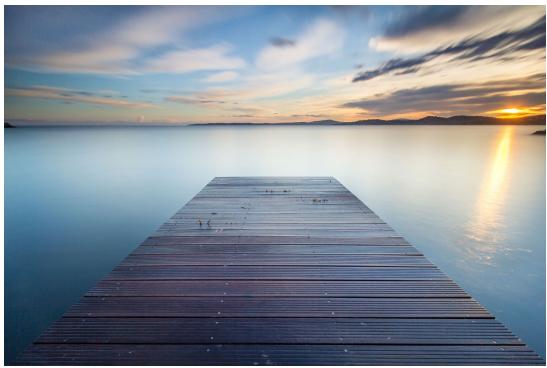
## Not Enough Meaning

We simplify probabilities and numbers  
to make them easier to think about

We think we know what  
other people are thinking

# Common biases that affect how we make decisions

- **confirmation bias:** paying more attention to information that reinforces previously held beliefs and ignoring evidence to the contrary
- **ingroup favoritism:** when one favors in-group members over out-group members
- **group attribution error:** when one generalizes about a group based on a group of representatives
- **halo effect:** when overall impression of a person impacts evaluation of their specific traits
- **just-world hypothesis:** when one protects a desire for a just world by blaming the victims
- etc.





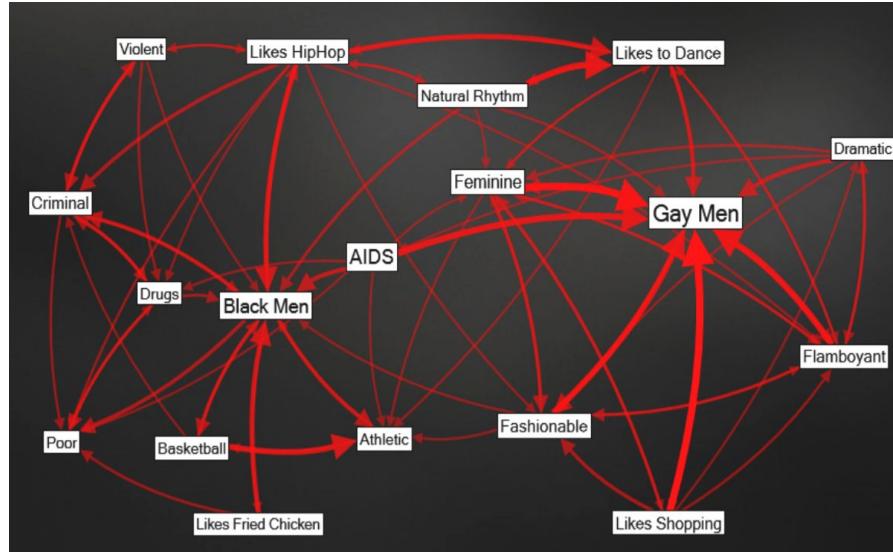
# Social stereotypes

- Gender
- Race
- Disability
- Age
- Sexual orientation
- Culture
- Class
- Poverty
- Language
- Religion
- National origin
- ...

Social stereotypes are similarly internalized as associations through natural processes of learning and categorization

Which word is more likely to be used by a **older person** ?

**Impressive – Amazing**



Implicit biases are pervasive, unconscious, and can automatically influence the ways in which we see and treat others, even when we are determined to be fair and objective.

# How do implicit biases manifest?

# Microaggressions

“A comment or action that **subtly and often unconsciously or unintentionally** expresses a prejudiced attitude towards a member of a marginalized group”

- Merriam Webster

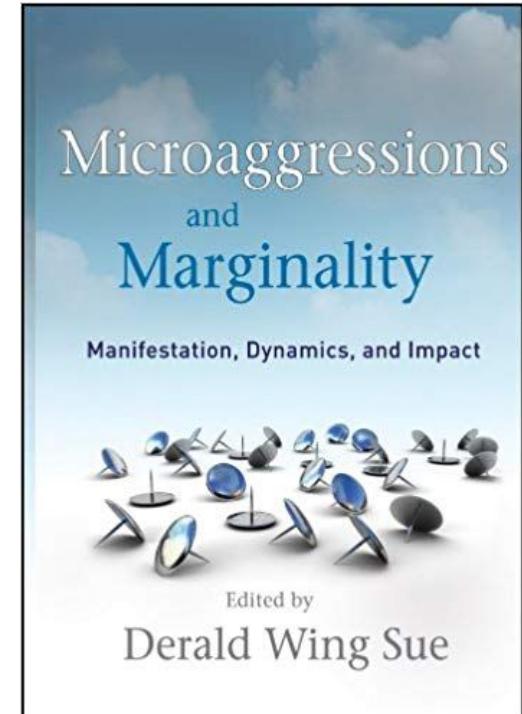
Surface-level sentiment can be negative, neutral, or positive. For example:

- “Girls just **aren’t good** at math.”
- “Don’t you people **like** tamales?”
- “You’re too **pretty** to be gay.”

[microaggressions.com](http://microaggressions.com)  
**tumblr.**

# Microaggressions cause prolonged harms

- Effects can be more pernicious than overtly aggressive speech (Sue et al. 2007, Sue 2010, Nadal et al. 2014)
- Can affect people's professional experiences and career trajectories (Cortina et al. 2002, Trix and Psenka 2003)
- Play on, and reinforce, problematic stereotypes and power structures (Hall and Braunwald 1981, Fournier et al. 2002)



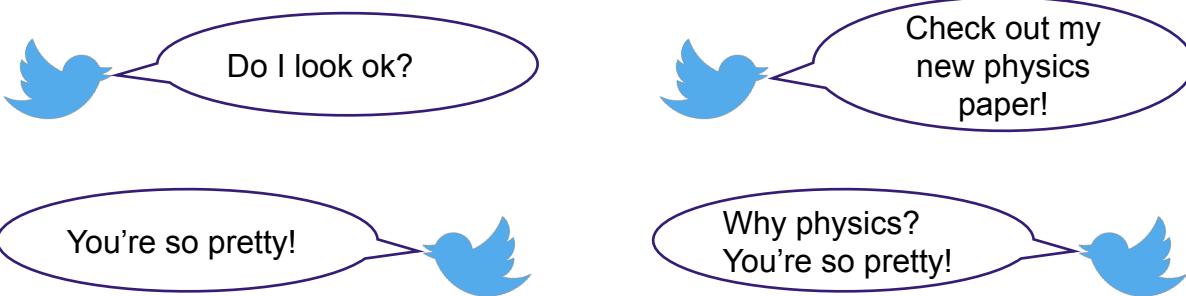
# Positive or negative?



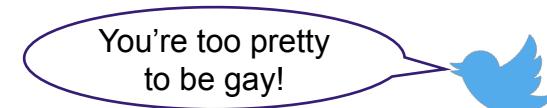
You're so pretty!



# Positive or negative?



# Positive or negative?





Online data is riddled with **SOCIAL STEREOTYPES**

# Bias in data

## Bias in language

- Stereotypes, prejudices, toxic comments and other expressions of social biases
- Historical human biases
- Human reporting bias: topics, word frequencies are not a reflection of real world

## Bias in datasets

- Data selection/sampling bias
- Annotator selection bias
- Annotators' cognitive biases

# From social bias to algorithmic bias

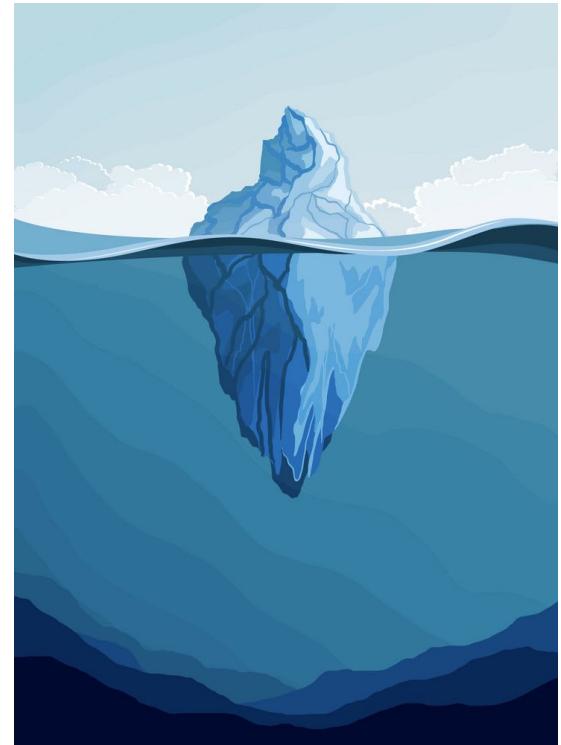
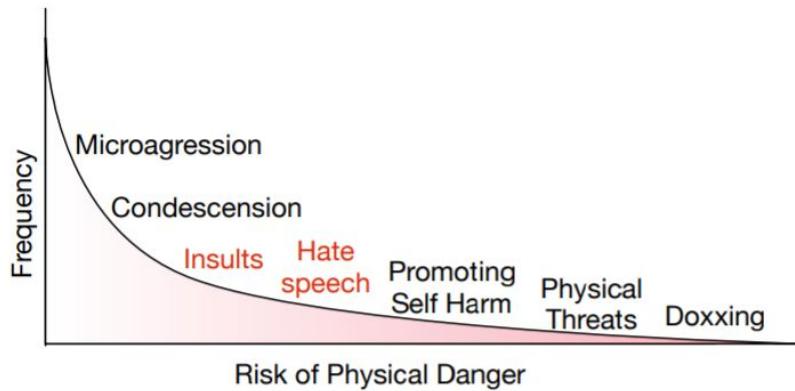
AI is only System 1



- data-centric models, no cultural and social context
- overfitting to confounders and spurious correlations, including social biases
- “black-box” models make it hard to proactively unveil these biases

# Toxic/offensive/biased comments

- Recent NLP advances have focused on **overt** toxic language (e.g. hate speech)
- Little focus on veiled negativity that is not directly encoded in lexicons



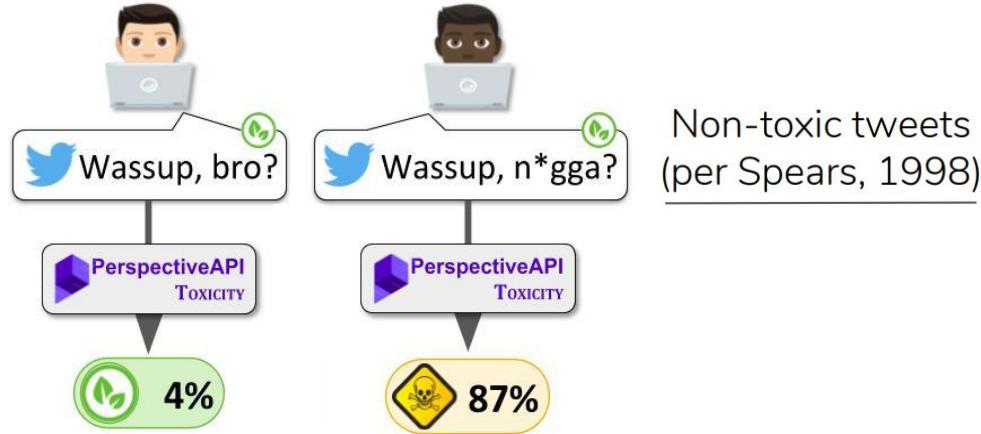
# SOTA NLP tools cannot identify microaggressions



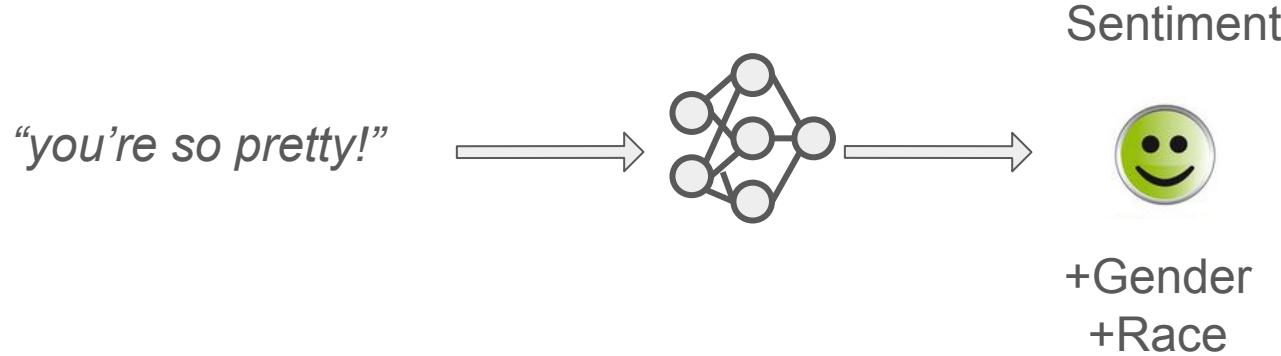
Breitfeller L., Ahn E., Jurgens D., Tsvetkov Y. (2019) Finding Microaggressions in the Wild:  
A Case for Locating Elusive Phenomena in Social Media Posts. *EMNLP*

# Models do not incorporate socio-cultural knowledge

- Toxicity classifiers overfit to social attributes overrepresented in training data, ignore social and cultural context



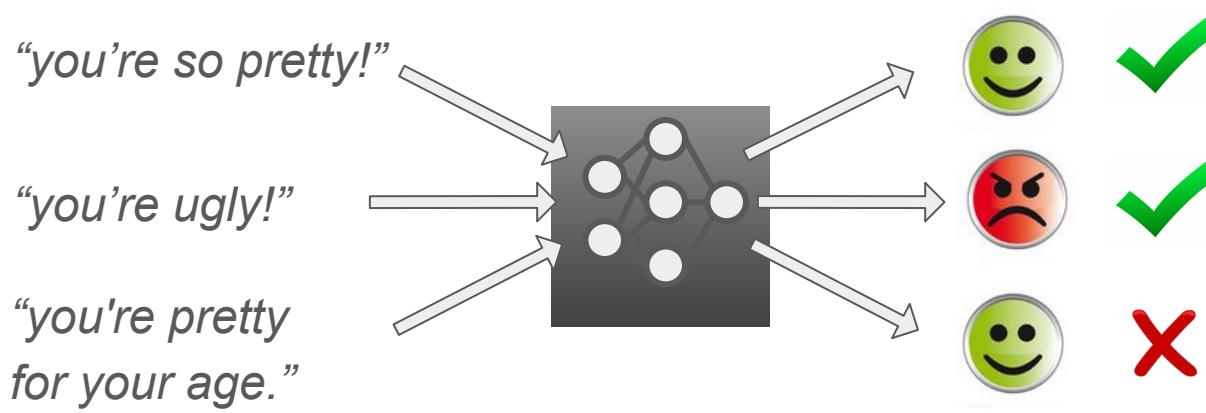
# Models overfit to spurious artifacts in data



- ‘The conversation with **Amanda** was heartbreaking’
- ‘The conversation with **Alonzo** was heartbreaking’
- ‘The conversation with **Lakisha** was heartbreaking’

# Models are not explainable

Toxicity



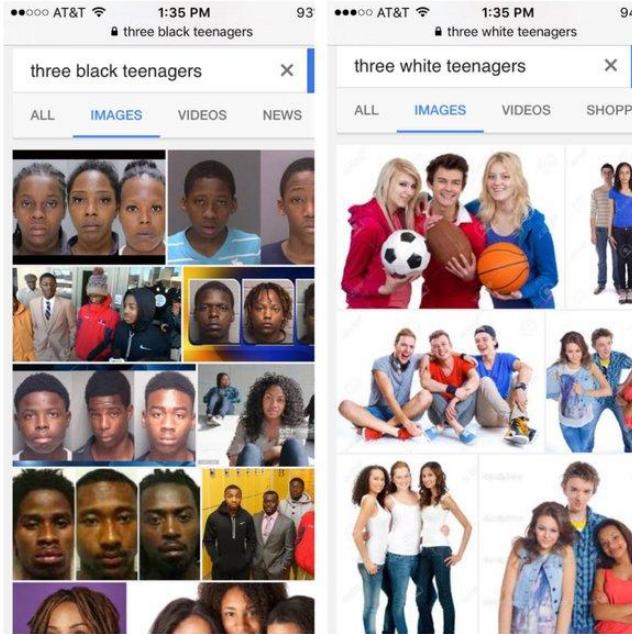
- Why?



- Conversational agents
- Personal assistants
- Medical assistants
- Educational assistants
- ...

# Image search

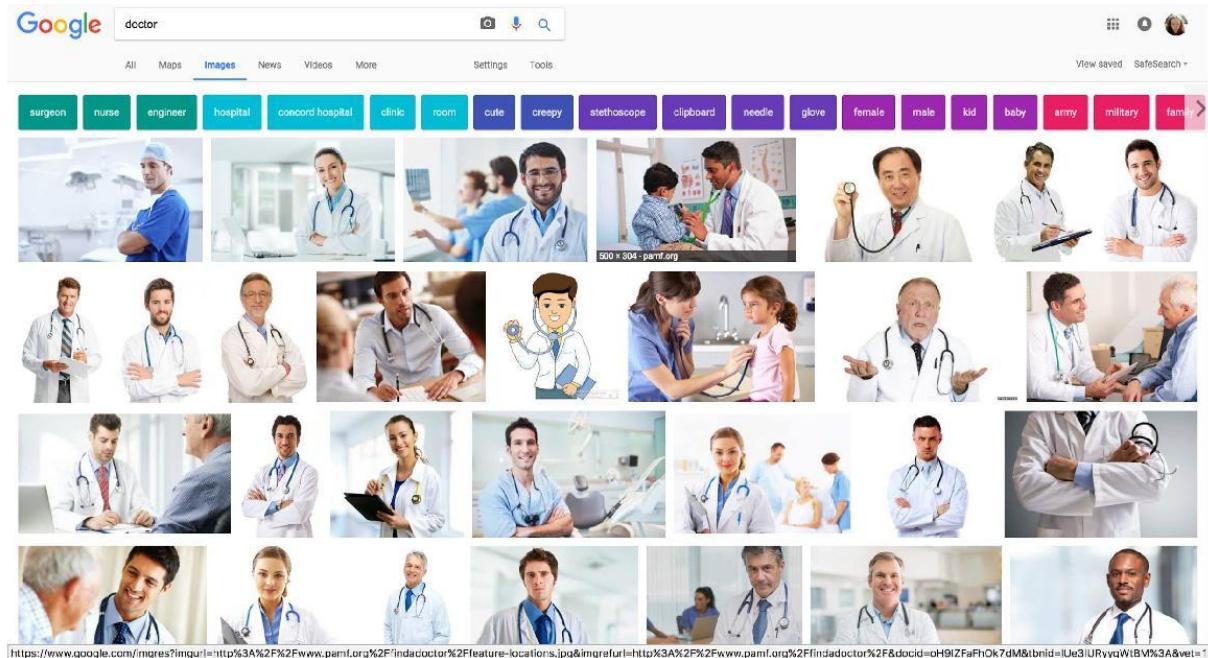
- Image search query “three black teenagers”



June 2017

# Image search

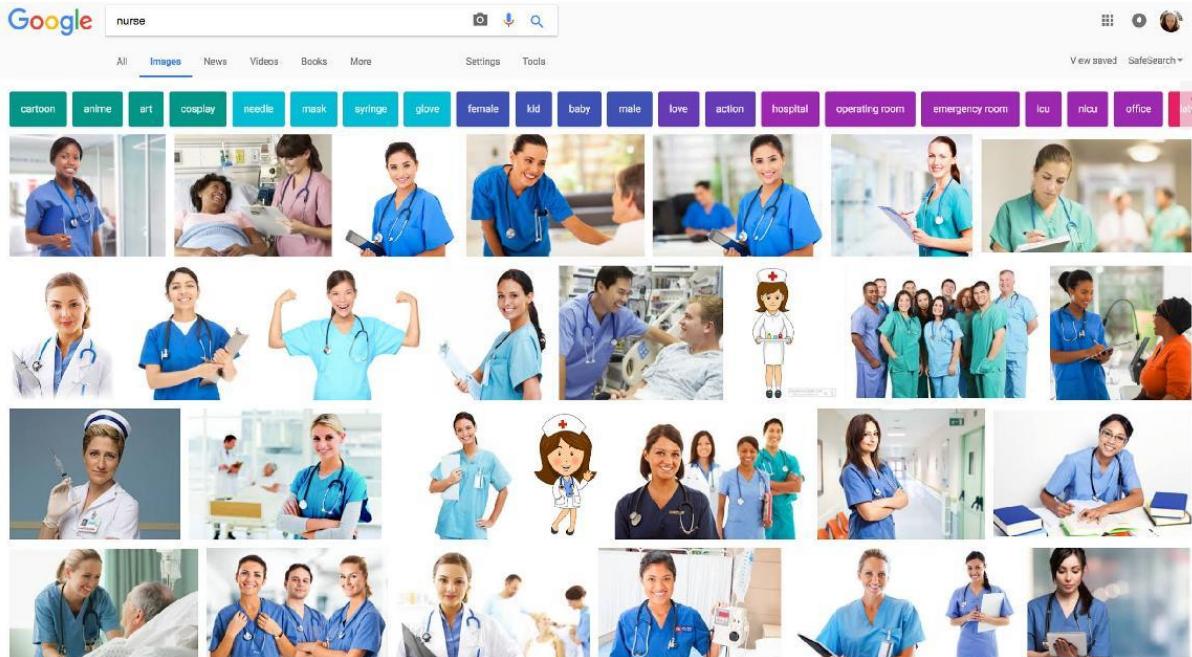
- Image search query “Doctor”



June 2017

# Image search

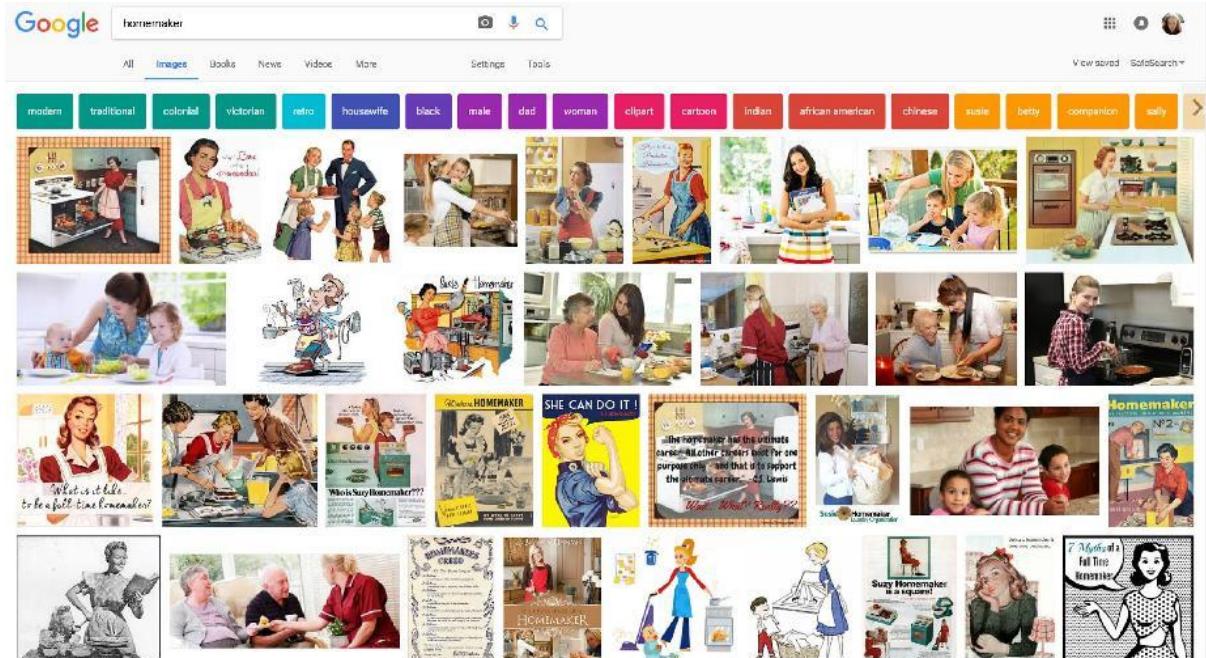
- Image search query “Nurse”



June 2017

# Image search

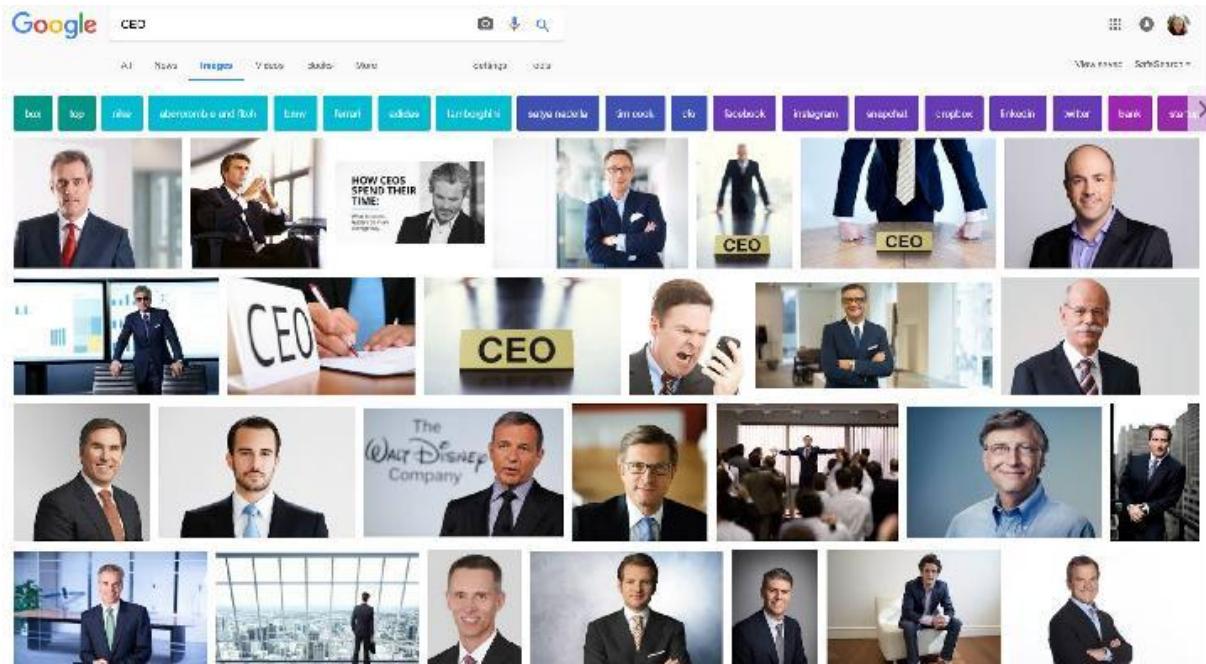
- Image search query “**Homemaker**”



June 2017

# Image search

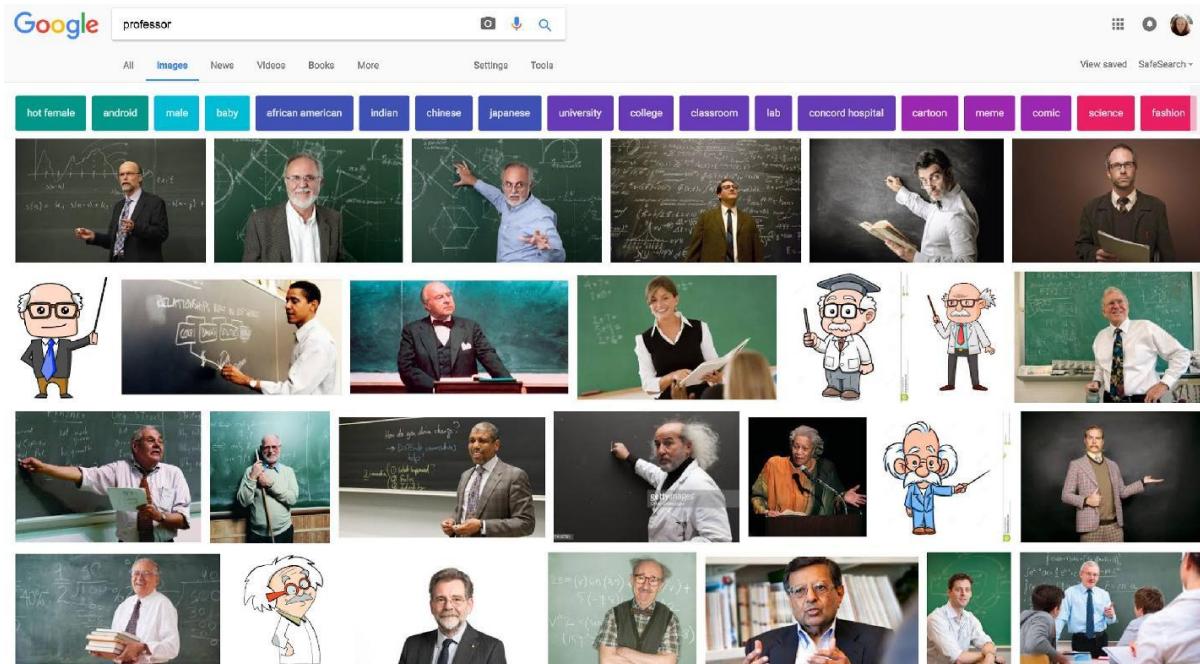
- Image search query “CEO”



June 2017

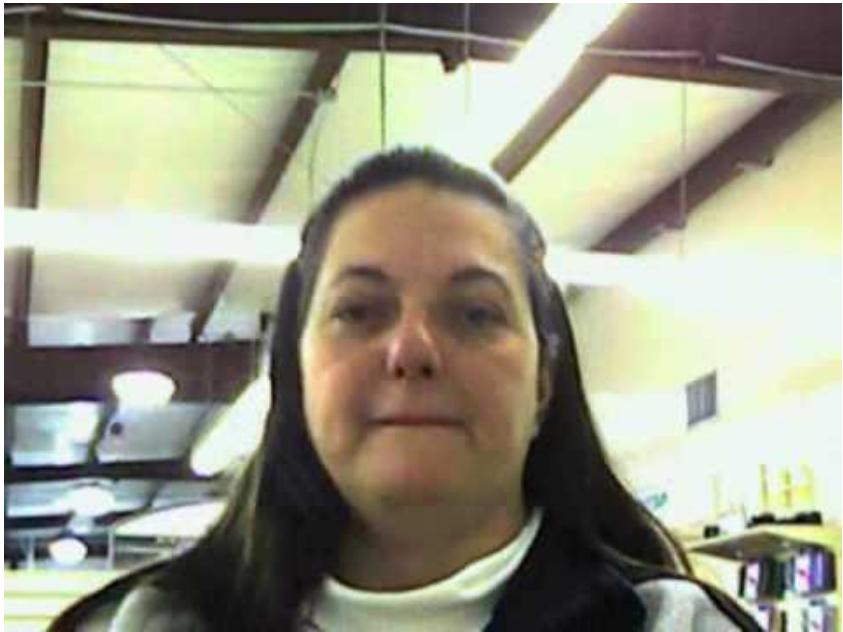
# Image search

- Image search query “Professor”



June 2017

# Face recognition



<https://gizmodo.com/5256650/camera-misses-the-mark-on-racial-sensitivity>

<https://www.wired.com/2009/12/hp-notebooks-racist/>

# Natural Language Processing

## Applications

- Machine Translation
- Speech Recognition
- Question Answering
- Dialogue Systems
- Information Extraction
- Summarization
- Sentiment Analysis
- ...

## Core technologies

- Language modelling
- Part-of-speech tagging
- Syntactic parsing
- Named-entity recognition
- Coreference resolution
- Word sense  
disambiguation
- Semantic Role Labelling
- ...

# Bias in Natural Language Processing

## Applications

- Machine Translation ([Prates et al. '19](#))
- Speech Recognition ([Tatman 2017](#))
- Question Answering ([Burghardt et al. '18](#))
- Dialogue Systems ([Dinan, Fan et al. '19](#))
- Summarization ([Jung, Kang et al. '19](#))
- Sentiment Analysis ([Kiritchenko & Mohammad '18](#))
- Language Identification ([Blodgett et al.'16, Jurgens et al.'17](#))
- Text Classification ([Dixon et al. '18, Sap et al. '19, Kumar et al. '19](#))
- ...

## Core technologies

- Language modeling ([Lu et al. '18](#))
- Named-entity recognition ([Mehrabi et al. '19](#))
- Coreference resolution ([Zhao et al. '18, Rudinger et al. '18](#))
- Semantic Role Labelling ([Zhao et al. '17](#))
- SNLI ([Rudinger et al. '17](#))
- Word Embeddings ([Bolukbasi et al. '16,++](#))
- ...

# Bias in machine translation

Translate

Turn off instant translation

Bengali

English

Hungarian

Detect language



English

Spanish

Hungarian



Translate

ő egy ápoló.

ő egy tudós.

ő egy mérnök.

ő egy pék.

ő egy tanár.

ő egy esküvői szervező.

ő egy vezérigazgatója.



she's a nurse.

he is a scientist.

he is an engineer.

she's a baker.

he is a teacher.

She is a wedding organizer.

he's a CEO.

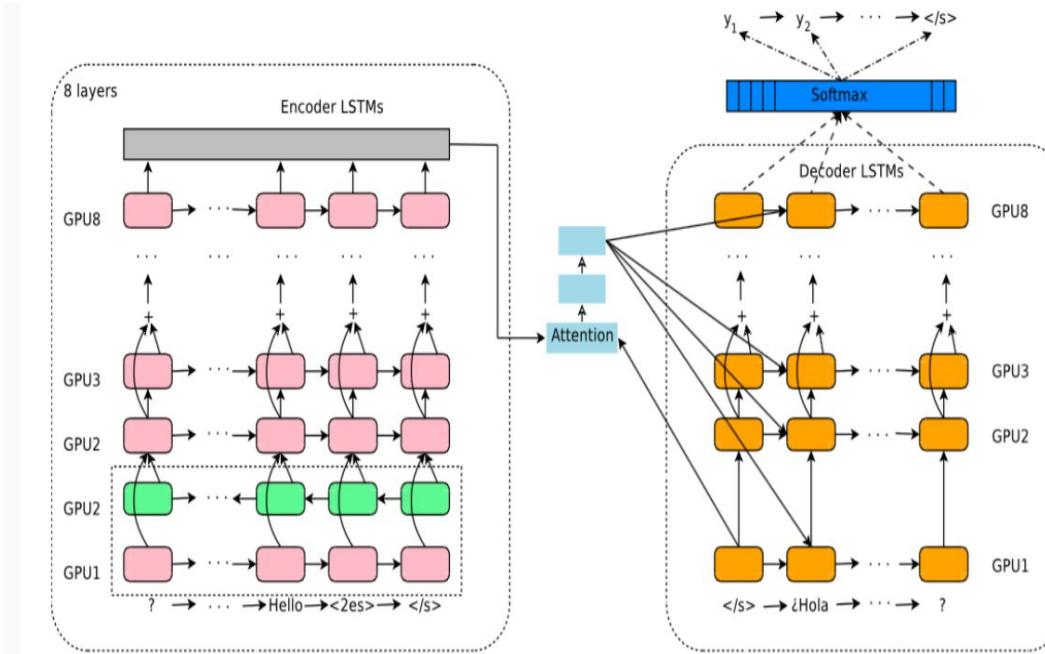


110/5000



Language Family	Language	Phrases have male/female markers	Tested
Austronesian	Malay	✗	✓
Uralic	Estonian	✗	✓
	Finnish	✗	✓
	Hungarian	✗	✓
Indo-European	Armenian	✗	✓
	Bengali	O	✓
	English	✓	✗
	Persian	✗	✓
	Nepali	O	✓
Japonic	Japanese	✗	✓
Koreanic	Korean	✓	✗
Turkic	Turkish	✗	✓
Niger-Congo	Yoruba	✗	✓
	Swahili	✗	✓
Isolate	Basque	✗	✓
Sino-Tibetan	Chinese	O	✓

# Example of bias mitigation (similar to multilingual NMT)



$<2female>$

$<2es>$  Hello, how are you?  $\rightarrow \_Hola$  como estás?

Before



After



# Bias in dialogue systems

[REDDIT](#) [GPT-3](#) [REDDIT](#) [OPENAI](#) [ARTIFICIAL INTELLIGENCE](#) [WRITING](#)

## Someone let a GPT-3 bot loose on Reddit — it didn't end well

The bot spent more than a week making comments about some seriously sensitive subjects



AI chatbot is REMOVED from Facebook after saying she 'despised' gay people, would 'rather die' than be disabled and calling the #MeToo movement 'ignorant'

- Lee Luda is a South Korean chatbot with the persona of a 20-year-old student
- It has attracted more than 750,000 users since its launch last month
- But the chatbot has started using hate speech towards minorities
- In one of the captured chat shots, Luda said she 'despised' gays and lesbians
- The developer has apologised over the remarks, saying they 'do not represent our values as a company'

# Reactive approach



# Towards a proactive approach

- **Data:** Automatic moderation, unveiling social biases and veiled toxicity in training data, beyond overtly hateful speech
- **Socio-cultural knowledge representation:** Learning to represent and analyze how socio-cultural knowledge manifests in language
- **Modeling:** New modeling approaches that incorporate socio-cultural context and are trained to explicitly demote social biases
- **Evaluation and analysis:** Developing interpretable models, or approaches to interpreting existing models, and new approaches to evaluation and characterization of model behaviors

# Future outlook

“Oh, you work at an office?  
I bet you’re a secretary, good for you!”



## Hate Speech Detection



Unlikely to be perceived as toxic  
(0.23)

## Sentiment Analysis



The text is **pos**.

## Social Bias Analysis [Conversational]

### Likely directed to:

- man: 0.1
- **woman / non-binary: 0.9**

### Explanation via similar examples of overt bias:

- women have low-prestige jobs
- girls are less smart than boys

**Likely** gender-based microaggression

# Automatic detection of implicit bias and veiled toxicity

- Via a causal framework and demoting spurious confounds

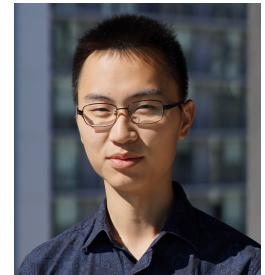
- Field A. & Tsvetkov Y. (2020) **Unsupervised Discovery of Implicit Gender Bias.** *EMNLP*



Anjalie

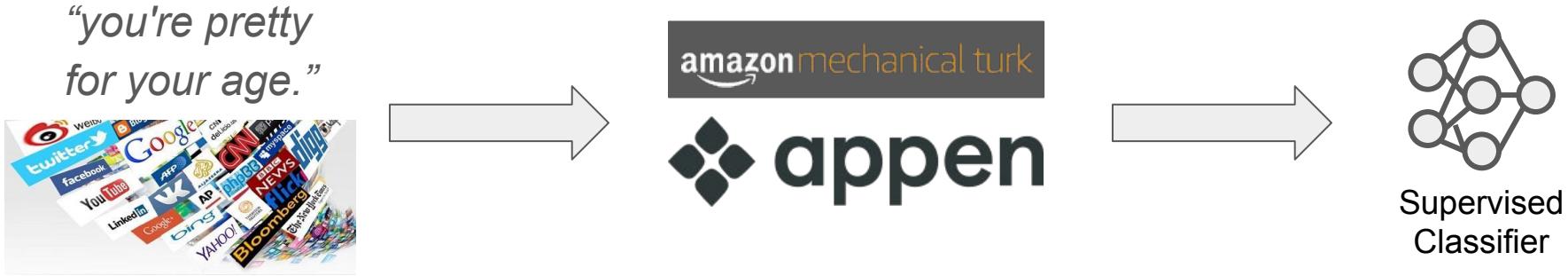
- Via adversarial probing & interpreting model decisions

- Han X., Tsvetkov Y. (2020) **Fortifying Toxic Speech Detectors Against Veiled Toxicity.** *EMNLP*



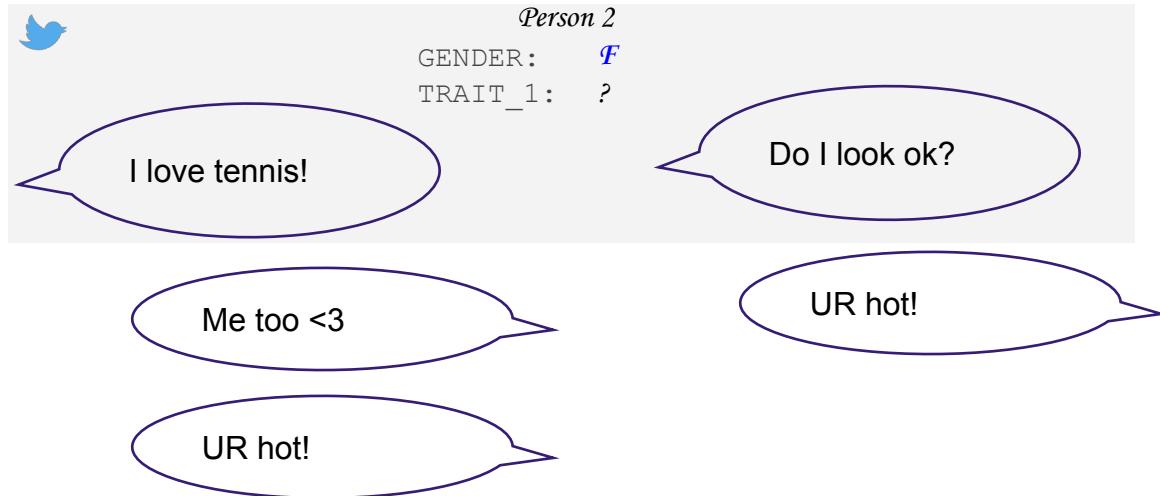
Han

# A naive approach: crowdsourcing & supervised classification

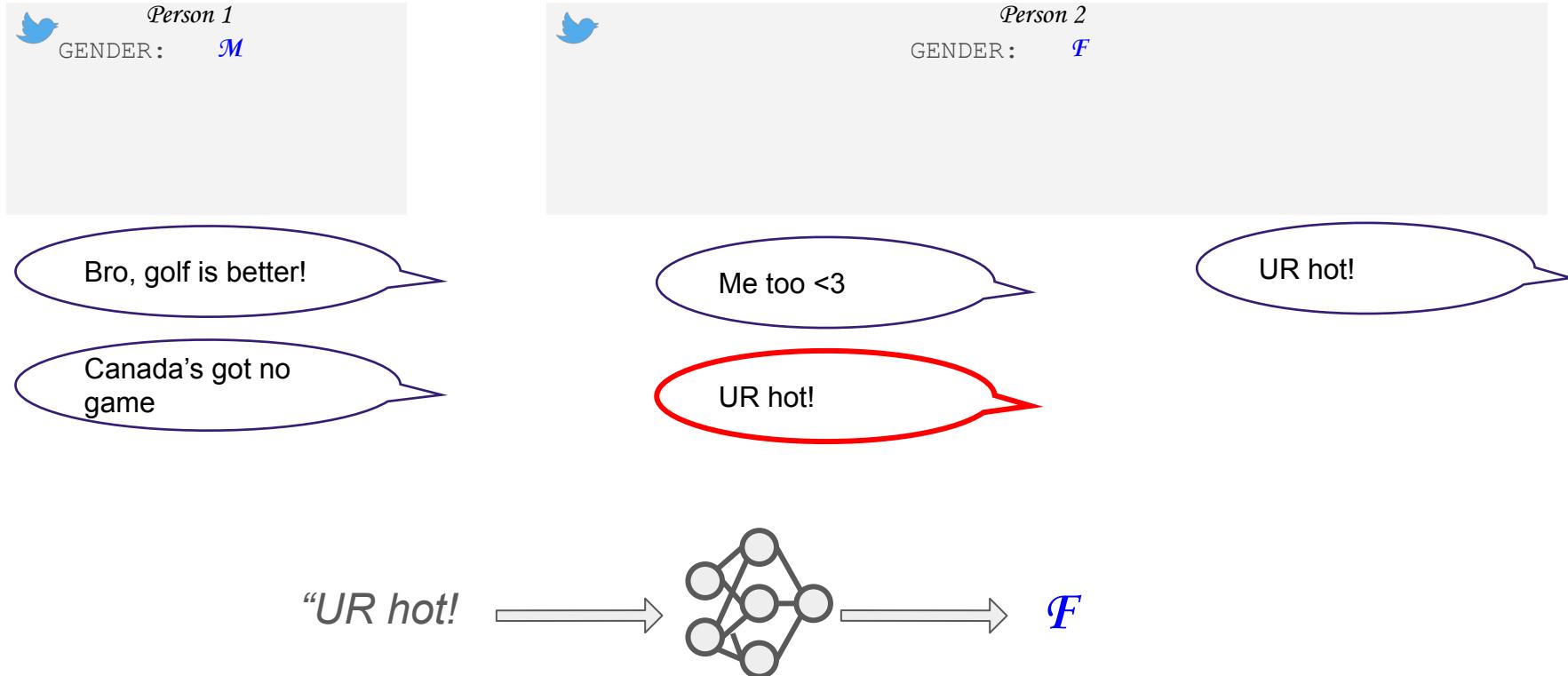


- Problems:
  - We don't have strong lexical sieve to surface candidates for annotation
  - Biases are subtle and implicit. We cannot rely on non-expert annotations. Every example requires multiple annotations by trained experts.

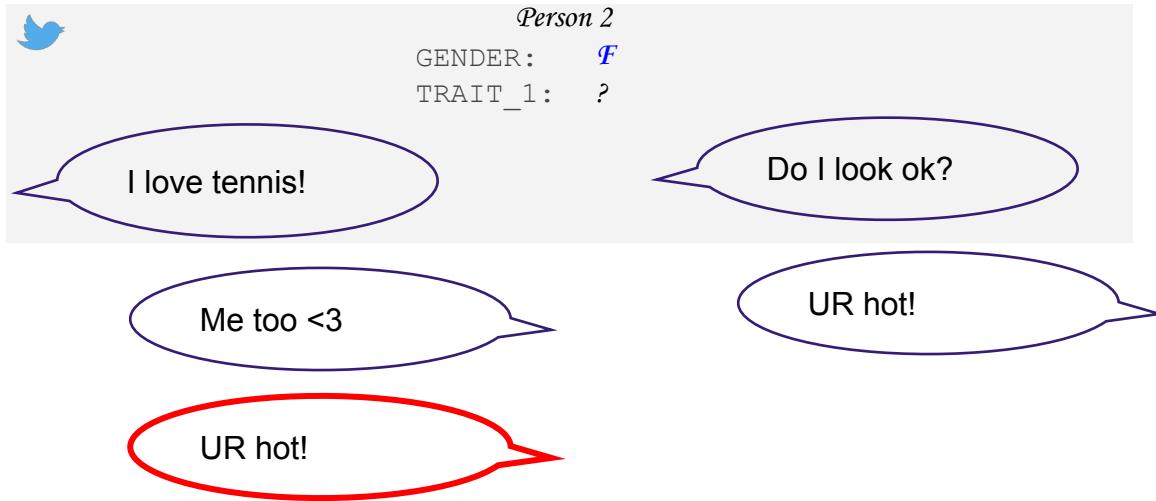
# Naive approach 2: Comments contain bias if they are highly predictive of gender



# Naive approach 2: Comments contain bias if they are highly predictive of gender



# Naive approach 2: Comments contain bias if they are highly predictive of gender



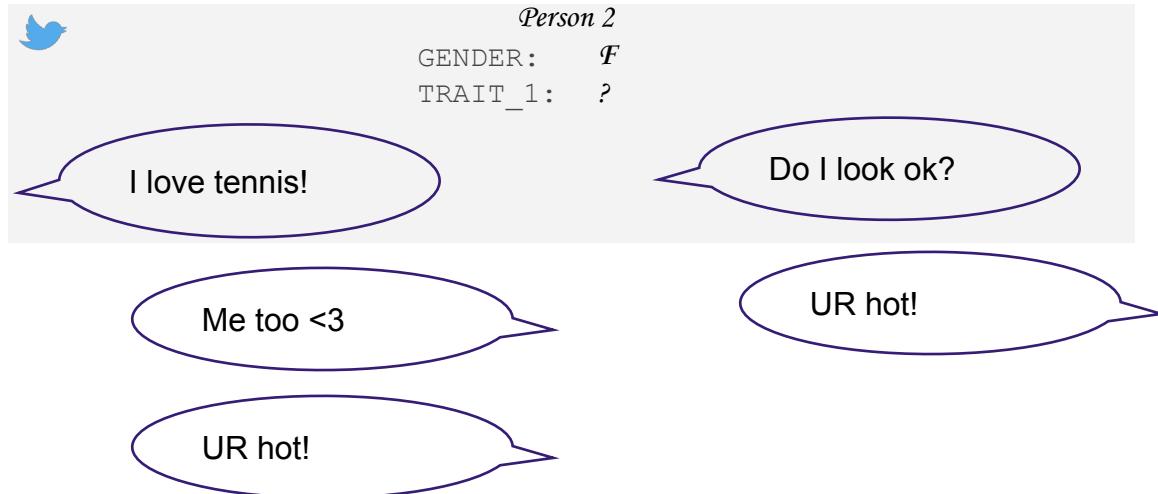
- There might be other factors that cause differences in text
- Text may contain **confounds** that are predictive of gender but not indicative of bias

# Proposed model: Comments contain bias if they are highly predictive of gender despite confound control



- Observed confounding variables are balanced through propensity matching
- Latent confounding variables are demoted through adversarial training
- Overt indicators are substituted

# Propensity matching for *observed* confounding variables



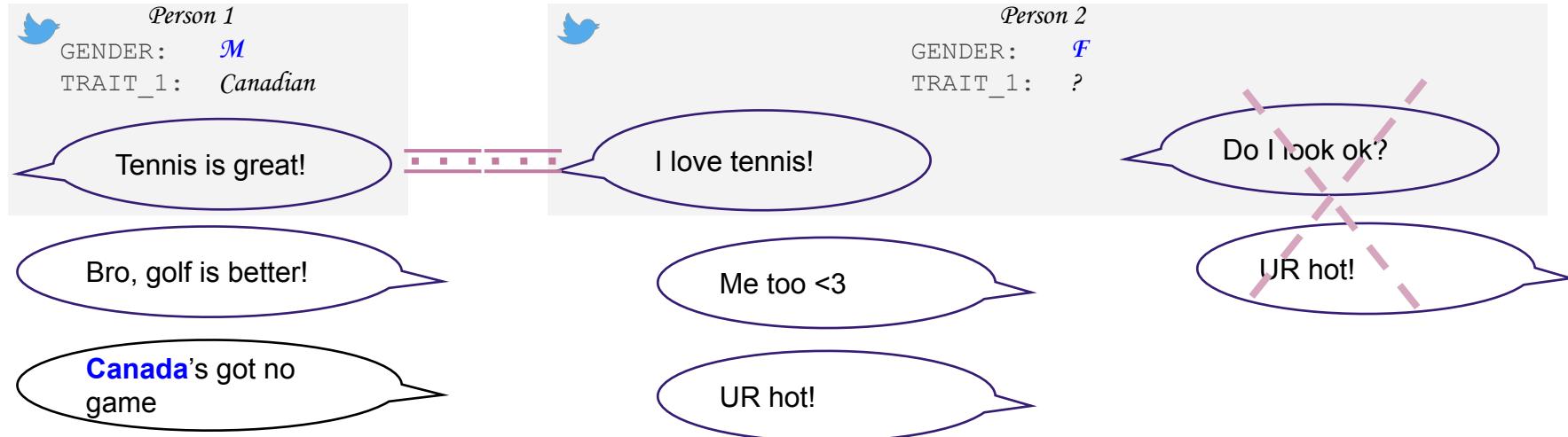
- Comments are written in reply to “original text” written by the addressee
- Language in comments may be caused by “original text” and not gender of the addressee

# Propensity matching for *observed* confounding variables



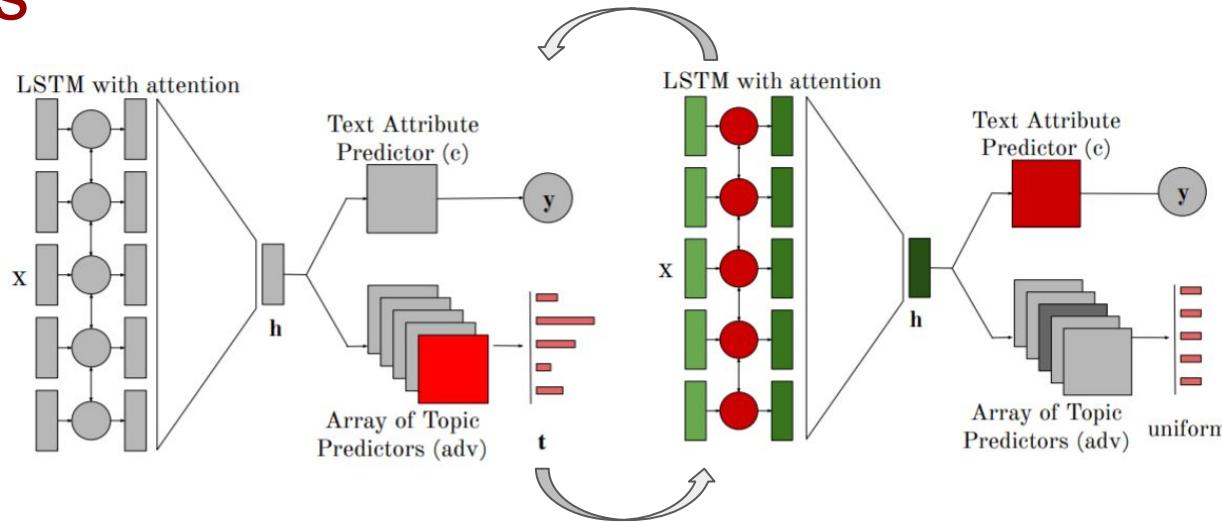
- Balance the data set so that comments addressed to men have a similar distribution of confounding variables as comments addressed to women
  - Match posts with similar indicators of confounding variables
  - Discard posts that are unable to be matched

# Adversarial training for *latent* confounding variables



- Comments may reference traits of the addressee (such as occupation, nationality, nicknames, etc.) other than gender
- Difficult to enumerate all of them
- Often unique to individuals (difficult to make matches)

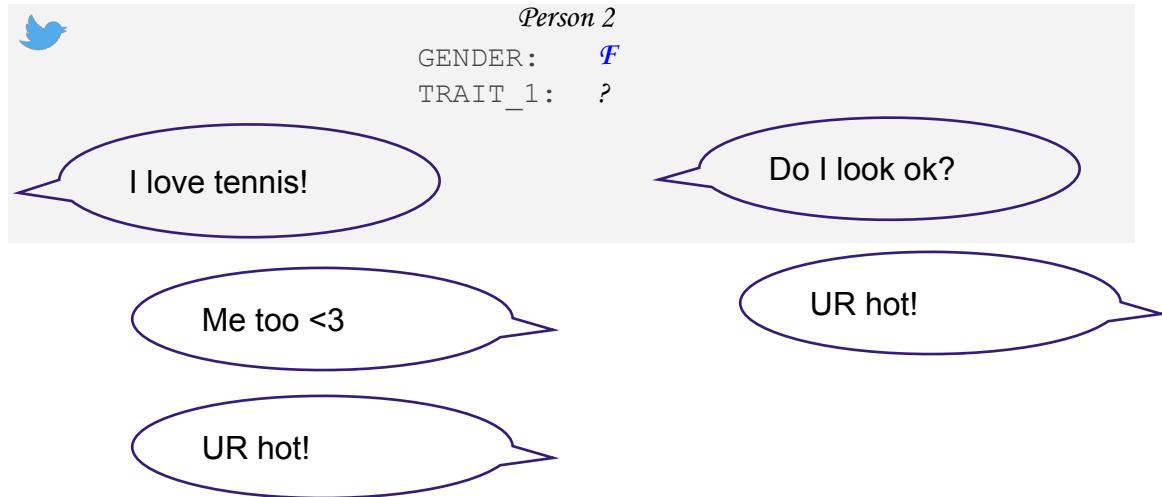
# Adversarial training to demote *latent* confounding variables



Kumar S., Wintner S., Smith, N. A and Tsvetkov Y. (2019) Topics to Avoid: Demoting Latent Confounds in Text Classification. *EMNLP*

- Confounding traits are inferred from comments using log-odds ratio with Dirichlet prior and represented in a vector
- GAN-like training procedure discourages the model from learning these traits

# Word substitutions for overt signals



- Remove overtly gendered terms (Mrs. Ms. Mr., proper names, etc.) using keyword substitution

# Proposed model: Comments contain bias if they are highly predictive of gender despite confound control



- Observed confounding variables are balanced through propensity matching
- Latent confounding variables are demoted through adversarial training
- Overt indicators are substituted

# Evaluation: Reducing influence of confounding variables



Voigt R., Jurgens D., Prabhakaran, V., Jurafsky D. and Tsvetkov Y. (2019) RtGender: A Corpus for Studying Differential Responses to Gender. *LREC*

- Latent confound demotion improves performance on held-out data
- Propensity matching reduces differences in training data distributions

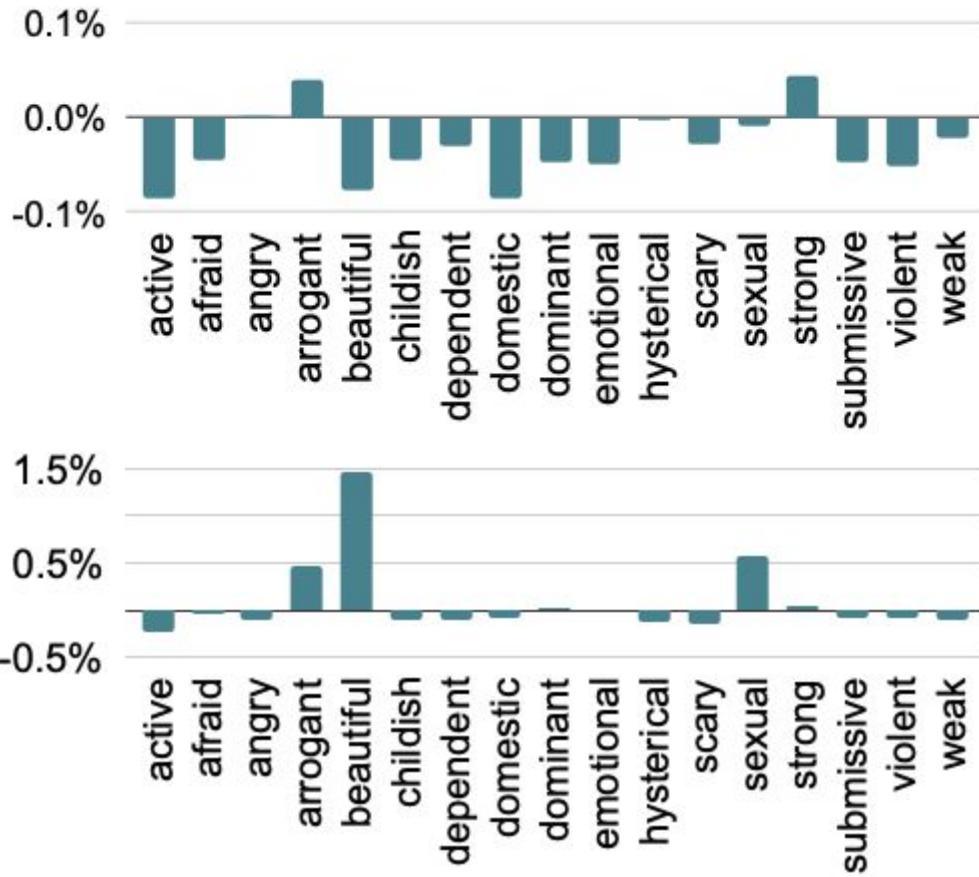
# Evaluation: detection of gender-based microaggressions

	Precision	Recall	F1	Accuracy
<b>Our model (Trained on data set 1)</b>	51.0	50.7	50.9	57.0
<b>Our model (Trained on data set 2)</b>	45.7	75.3	56.9	49.9
<b>Random baseline</b>	43.5	48.7	46.0	49.8

*“You’re pretty for a black girl.”*

[microaggressions.com](http://microaggressions.com)

# Findings: characteristics of bias against women



## Politicians:

- Competence and domesticity
- ‘Force’, ‘situation’, ‘spouse’, ‘family’, ‘love’

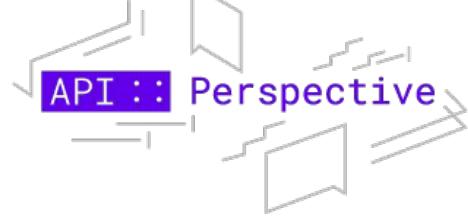
## Other Public Figures:

- Appearance and sexualization
- ‘beautiful’, ‘bellissima’, ‘amore’, ‘amo’, ‘love’, ‘linda’, ‘sexo’

# Automatic detection of implicit bias and veiled toxicity

- Via a causal framework and demoting spurious confounds
  - Field A. & Tsvetkov Y. (2020) **Unsupervised Discovery of Implicit Gender Bias.** *EMNLP*
- Via adversarial probing & interpreting model decisions
  - Han X., Tsvetkov Y. (2020) **Fortifying Toxic Speech Detectors Against Veiled Toxicity.** *EMNLP*

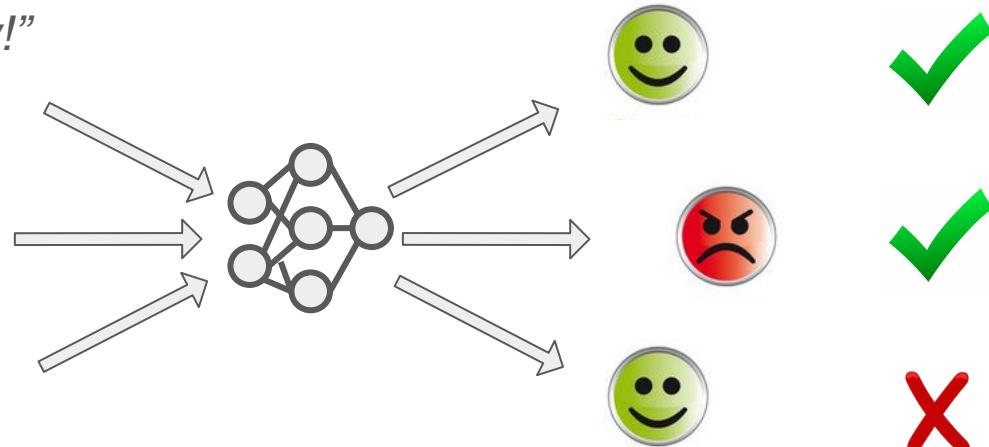
# A toxicity classifier



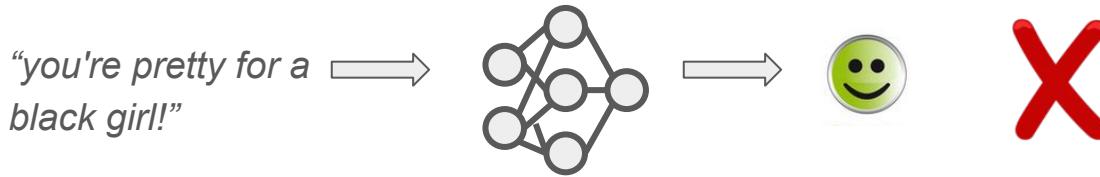
*"you're so pretty!"*

*"you're ugly!"*

*"you're pretty  
for a black girl!"*

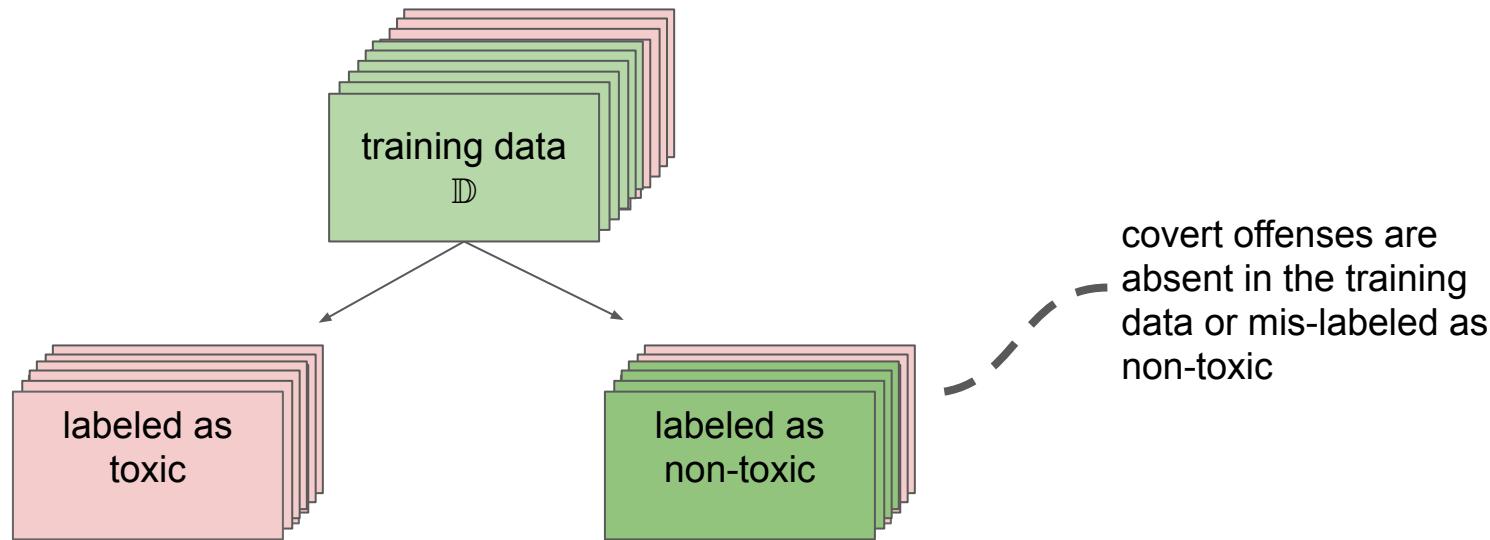
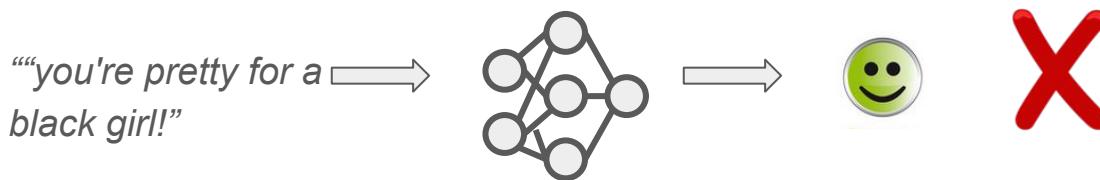


# Microaggressions as adversarial attacks on toxicity classifiers



- Codewords (Taylor et al., 2017)
- Novel forms of offense (Jain et al., 2018)
- Microaggressions (Breitfeller et al., 2019)
- Condescension (Wang and Potts, 2019)
- Dismissiveness, unfair generalizations (Price et al., 2020)

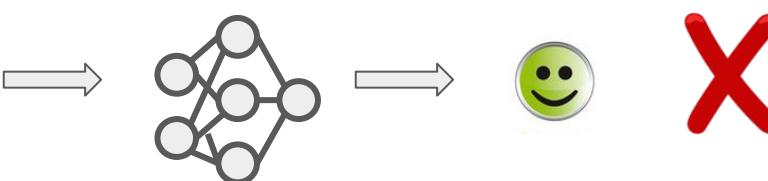
# Microaggressions as adversarial attacks on toxicity classifiers



# Interpreting text classification decisions via saliency maps

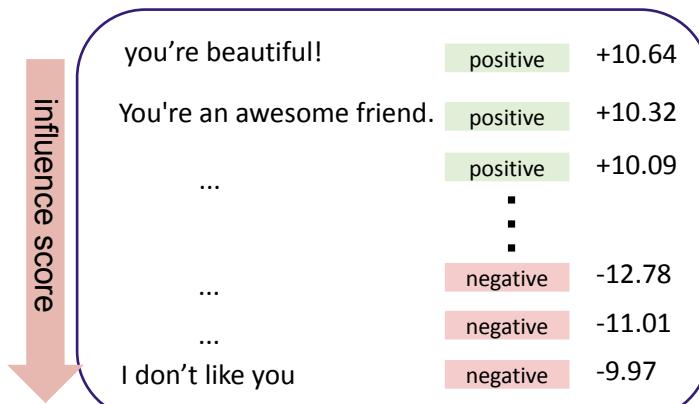
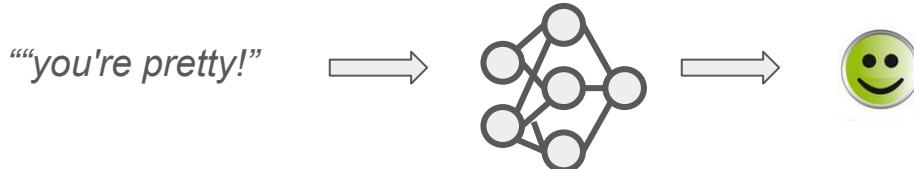
"you 're pretty for a black girl!"  
+0.02 +0.03 +0.45 +0.04 0 -0.08 +0.1

*Finding salient tokens in the input*



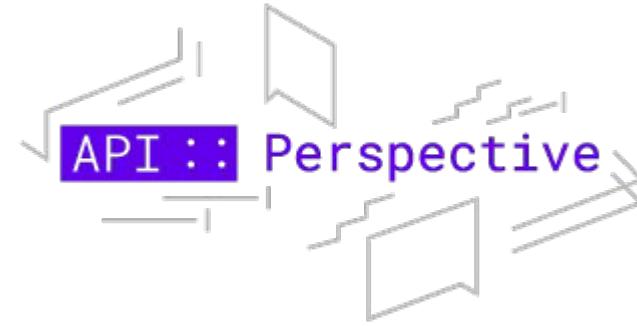
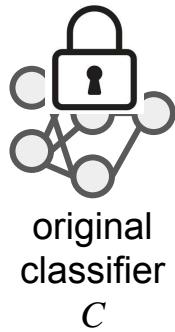
- Interpretation via saliency maps
  - Gradient-based attribution ([Simonyan et al.,'14](#); [Sundararajan et al.'17](#); [Smilkov et al.'17](#))
  - LIME ([Ribeiro et al.'16](#))
  - Attention-based heatmaps ([Xu et al.'15](#))

# Interpreting text classification decisions via the influence of examples in the training data

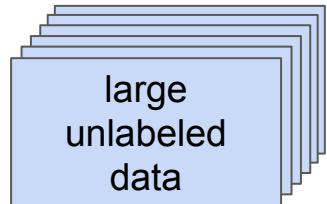
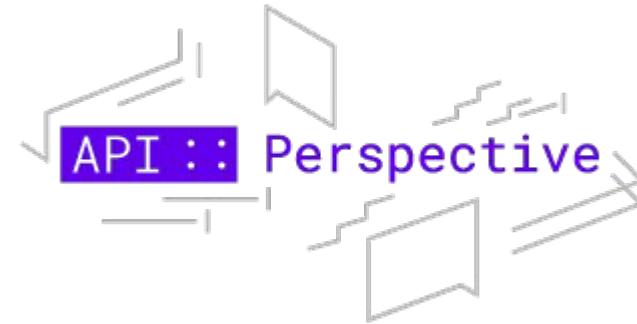


*Finding influential examples in the training corpus*

# Training data of toxicity classifiers is often private

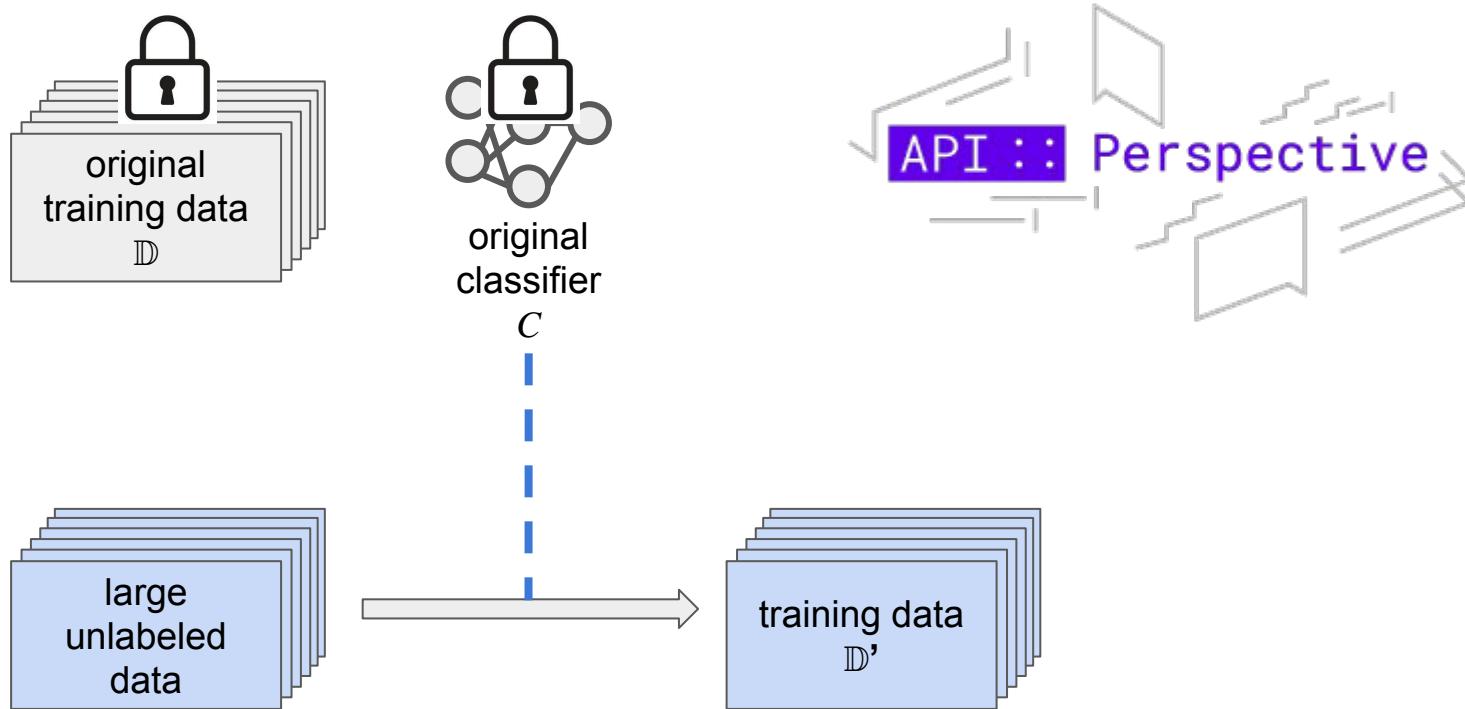


# We'll train a student model



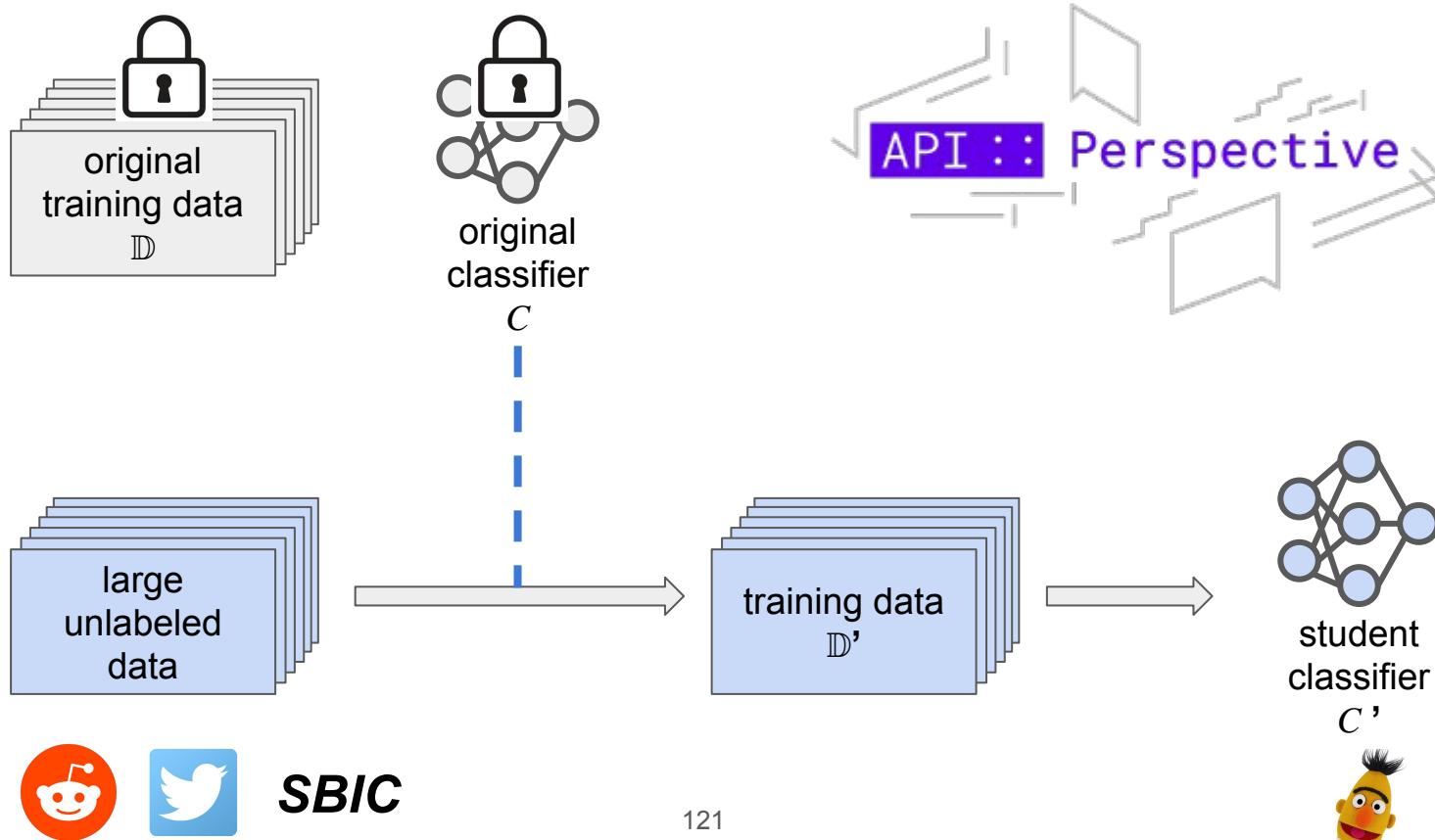
**SBIC**

# We'll train a student model

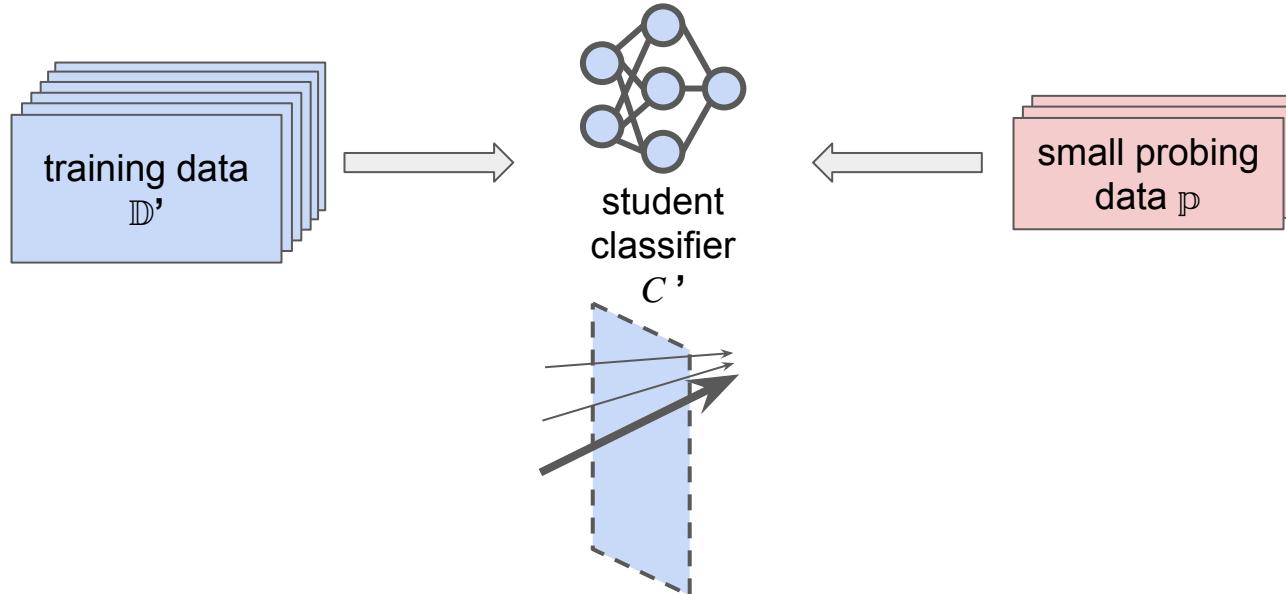


**SBIC**

# We'll train a student model



# Towards robust toxicity detection via adversarial probing & interpreting model decisions

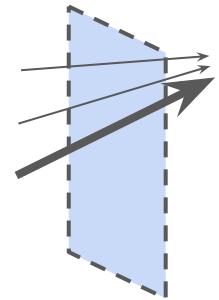


Given a small set of probing examples of veiled toxicity

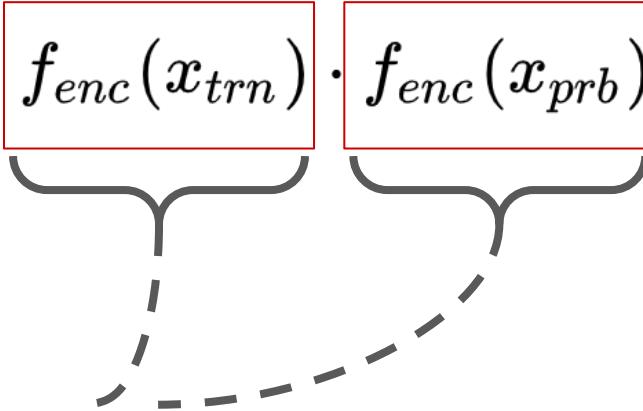
1. Interpret model decisions via tracking the influence of training examples
2. Use expert annotators to re-label top-k training instances

# Tracking the influence of the training data on classifier's decisions

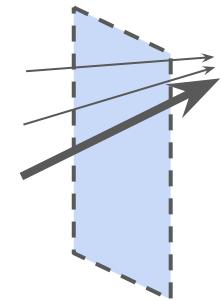
- Which **training** data is most influential to the classifier's decision on a probing example?
- $\mathcal{I}(x_{trn}, x_{prb})$



# Embedding Similarity

$$\mathcal{I}(x_{trn}, x_{prb}) = f_{enc}(x_{trn}) \cdot f_{enc}(x_{prb})$$


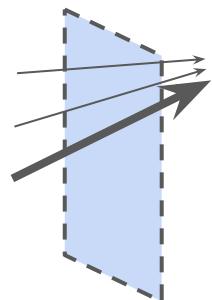
“How different are the **representations** of the training data and the probing data?”



# Influence Functions

$$\boxed{\frac{d\theta}{d\epsilon_{trn}}} = -H_\theta^{-1} \nabla_\theta \mathcal{L}(\theta, x_{trn}, y_{trn})$$

“If we **upweight** a training example by  $\epsilon$ , how would the resulting model change?”

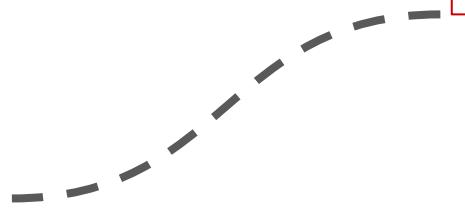


(Koh and Liang, 2017)

# Influence Functions

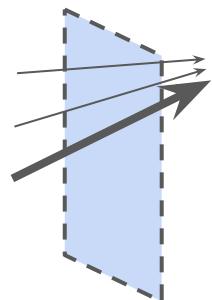
$$\frac{d\theta}{d\epsilon_{trn}} = -H_\theta^{-1} \nabla_\theta \mathcal{L}(\theta, x_{trn}, y_{trn})$$

$$\frac{d\mathcal{L}(\theta, x_{prb}, \hat{y}_{prb})}{d\epsilon_{trn}} = \nabla_\theta \mathcal{L}(\theta, x_{prb}, \hat{y}_{prb}) \cdot \boxed{\frac{d\theta}{d\epsilon_{trn}}}$$



“Given this **change** in the resulting model ...”

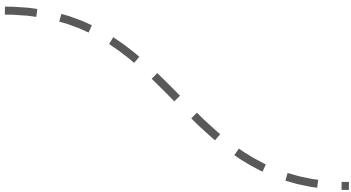
(Koh and Liang, 2017)



# Influence Functions

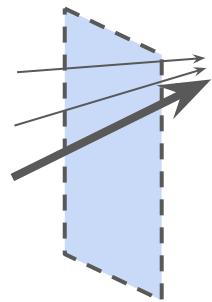
$$\frac{d\theta}{d\epsilon_{trn}} = -H_\theta^{-1} \nabla_\theta \mathcal{L}(\theta, x_{trn}, y_{trn})$$

$$\boxed{\frac{d\mathcal{L}(\theta, x_{prb}, \hat{y}_{prb})}{d\epsilon_{trn}}} = \nabla_\theta \mathcal{L}(\theta, x_{prb}, \hat{y}_{prb}) \cdot \frac{d\theta}{d\epsilon_{trn}}$$



“How would the **loss** of the probing example change?”

(Koh and Liang, 2017)



# Influence Functions

$$\frac{d\theta}{d\epsilon_{trn}} = -H_\theta^{-1} \nabla_\theta \mathcal{L}(\theta, x_{trn}, y_{trn})$$

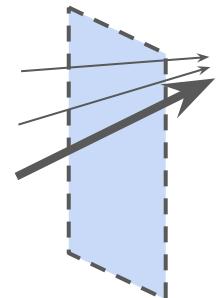
$$\frac{d\mathcal{L}(\theta, x_{prb}, \hat{y}_{prb})}{d\epsilon_{trn}} = \nabla_\theta \mathcal{L}(\theta, x_{prb}, \hat{y}_{prb}) \cdot \frac{d\theta}{d\epsilon_{trn}}$$

$$\mathcal{I}(x_{trn}, x_{prb}) = \boxed{-\frac{d\mathcal{L}(\theta, x_{prb}, \hat{y}_{prb})}{d\epsilon_{trn}}}$$

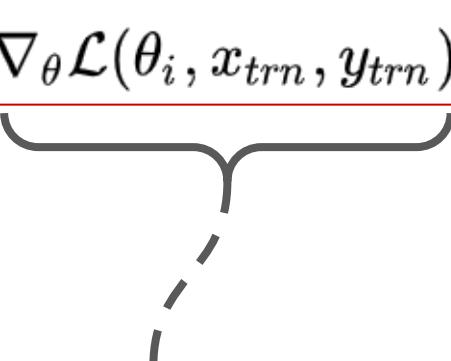
|  
|

“Upweighting an **influential** training example should lead to a decrease in the loss of the probing example.”

(Koh and Liang, 2017)

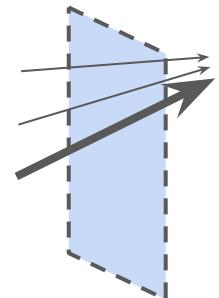


# Gradient product (*TrackIn*)

$$\mathcal{I}(x_{trn}, x_{prb}) = \sum_{i=1}^k \boxed{\nabla_{\theta} \mathcal{L}(\theta_i, x_{trn}, y_{trn})} \cdot \nabla_{\theta} \mathcal{L}(\theta_i, x_{prb}, \hat{y}_{prb})$$


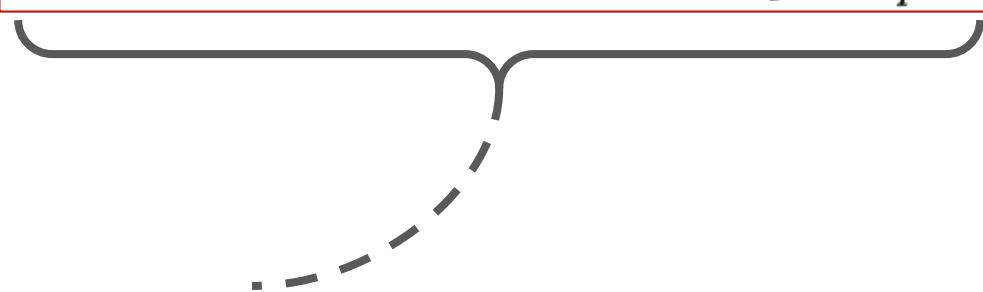
“The model would take a step towards the **gradient** of the training example’s loss at epoch  $i$ .”

(Pruthi et al., 2020)



# Gradient product (*TrackIn*)

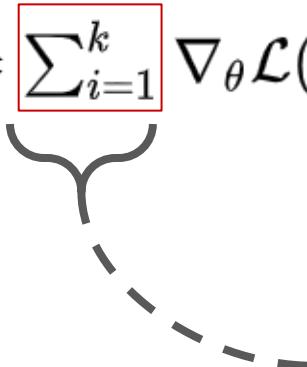
$$\mathcal{I}(x_{trn}, x_{prb}) = \sum_{i=1}^k \boxed{\nabla_{\theta} \mathcal{L}(\theta_i, x_{trn}, y_{trn}) \cdot \nabla_{\theta} \mathcal{L}(\theta_i, x_{prb}, \hat{y}_{prb})}$$



“Because of this step, how much will the loss of the probing example **decrease**? ”

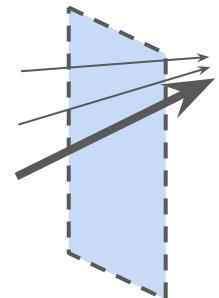
(Pruthi et al., 2020)

# Gradient product (*TrackIn*)

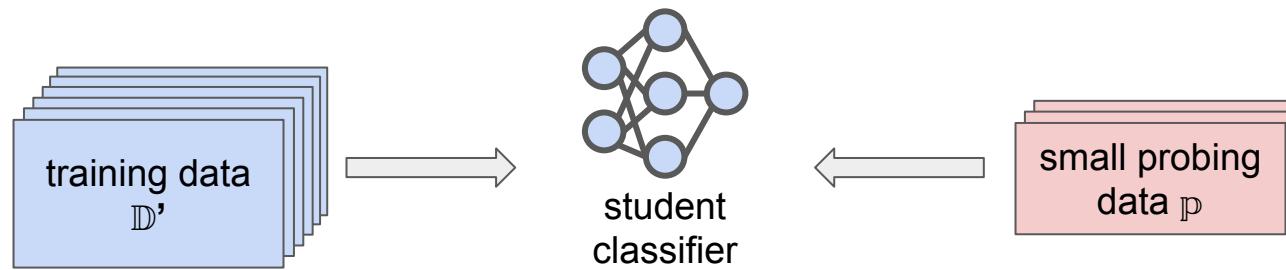
$$\mathcal{I}(x_{trn}, x_{prb}) = \sum_{i=1}^k \nabla_{\theta} \mathcal{L}(\theta_i, x_{trn}, y_{trn}) \cdot \nabla_{\theta} \mathcal{L}(\theta_i, x_{prb}, \hat{y}_{prb})$$


“We take a **sum** of such probing loss decrease caused by the training example over all the checkpoints of the model.”

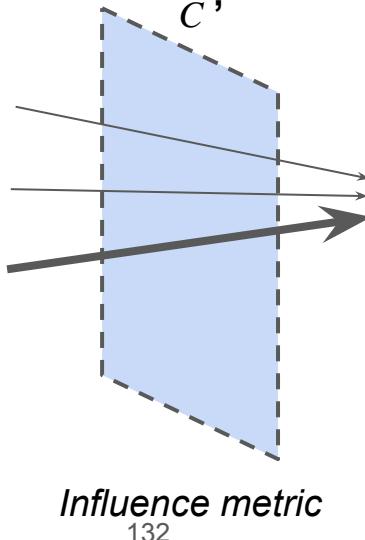
(Pruthi et al., 2020)



# An active learning framework to surface veiled toxicity



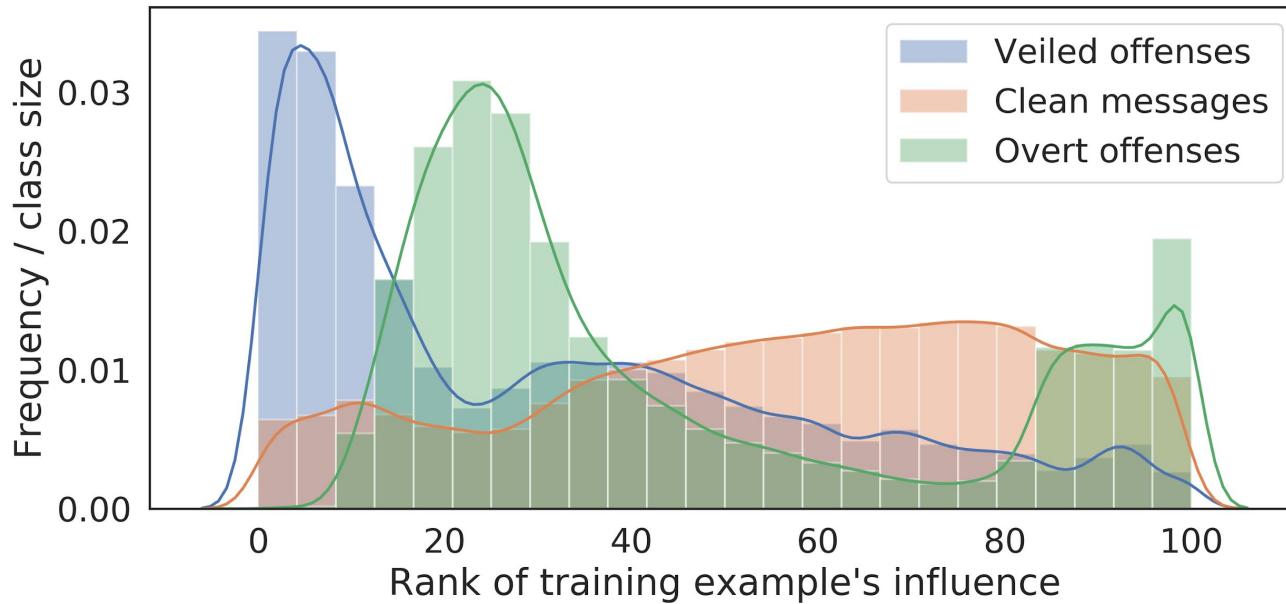
- re-label top-k training examples that were most influential to classifier's decisions



# Experiments

- Training data:  
  - Reddit, twitter posts from the **SBIC** - Social Bias Inference Corpus (Sap et al., 2020)
  - Clean messages: 8K “non-toxic” by Perspective API and “not offensive” by SBIC annotators
  - Overt offenses: 2K “toxic” by Perspective API and “offensive” by SBIC annotators
  - Veiled offenses: 2K “non-toxic” by Perspective API and “offensive” by some SBIC annotators
- Test data: 1K, 1K, 1K
- Class recall: not offensive: 99.6%, overtly offensive: 97.2%, veiled offenses: 1.2%
- Probing corpus: <100 held-out examples of microaggressions and other veiled offenses [microaggressions.com](http://microaggressions.com)

# Unveiling disguised toxicity via probing & interpreting model decisions



- Clean messages: 8K “non-toxic” by Perspective API and “not offensive” by SBIC annotators
- Overt offenses: 2K “toxic” by Perspective API and “offensive” by SBIC annotators
- Veiled offenses: 2K “non-toxic” by Perspective API and “offensive” by some SBIC annotators

# Unveiling disguised toxicity via probing & interpreting model decisions

Model	Operation	Veiled	Clean	Overt
				
<u>Original</u>		1.2	99.6	97.2
<u>Gradient product</u>	fix top 2000	37.5	97.6	98.0
	flip top 2000	51.1	87.6	99.5
<u>Gold</u>		76.0	94.8	98.2

- Clean messages: 8K “non-toxic” by Perspective API and “not offensive” by SBIC annotators
- Overt offenses: 2K “toxic” by Perspective API and “offensive” by SBIC annotators
- Veiled offenses: 2K “non-toxic” by Perspective API and “offensive” by some SBIC annotators

# In sum,

- Biases in the data are pervasive and pernicious
- SOTA NLP tools cannot identify toxic comments beyond overt hate speech
  - Biases are subtle and implicit even experts are bad at identifying them
  - We don't have strong lexical sieve to surface candidates for annotation
- Why is it important to detect biases in data:
  - Posters are often unaware that their comments contains bias -- if they were, they may choose not to post them
  - Users can choose not to read flagged comments
  - These comments can be filtered from the training data of AI systems
- Direct supervised approaches are not enough we need paradigms shift in modeling

# Social and ethical challenges for language technologies

- **Incivility:** Hate-speech, toxicity, incivility, microaggressions online
- **Social bias:** Social bias in data & NLP models
- **Privacy violation:** Privacy violation & language-based profiling
- **Misinformation:** Fake news, Information manipulation, opinion manipulation
- **Technological divide:** Unfair NLP technologies, underperforming for speakers of minority dialects, for languages from developing countries, and for disadvantaged populations

# Some ideas for research projects

- Additional ideas for research projects  
[shorturl.at/wxBJ9](https://shorturl.at/wxBJ9)

# Thank you!

yuliats@cs.washington.edu