Lecture 18: Wrapup + Ethics

Alan Ritter

(many slides from Greg Durrett)

Administrivia

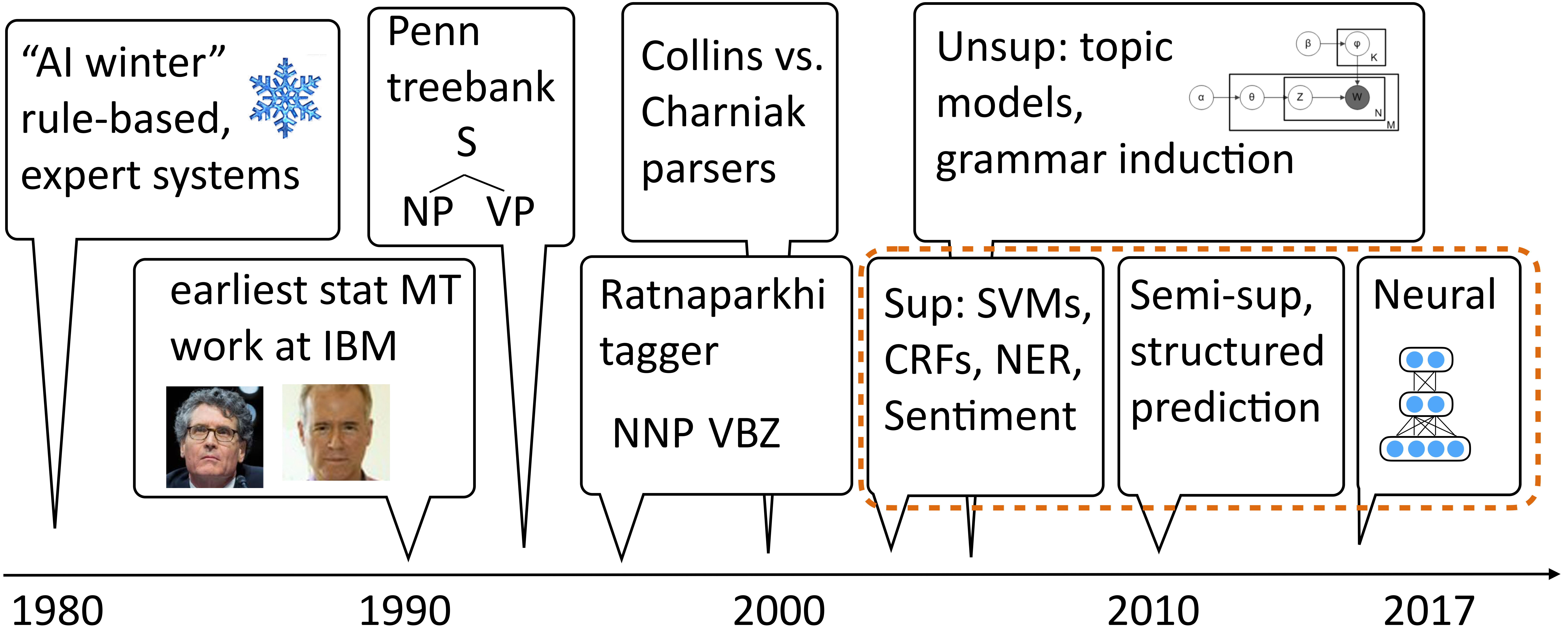
- Final project reports due Wednesday 5/5
- Please fill out the course/instructor opinion survey (CIOS) if you haven't already!

This Lecture

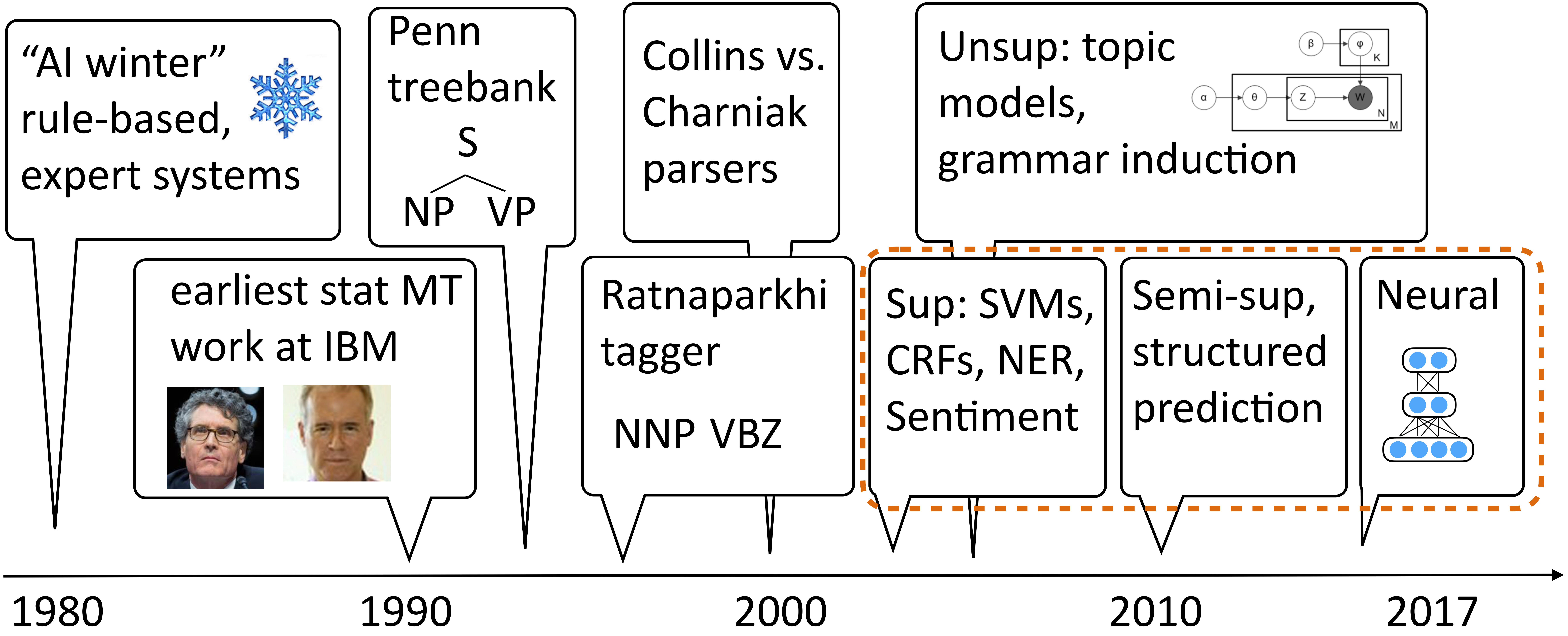
- Course recap

- Ethics in NLP

A brief history of (modern) NLP



A brief history of (modern) NLP



- What different model structures did we consider?

Sequential Structure: Analysis

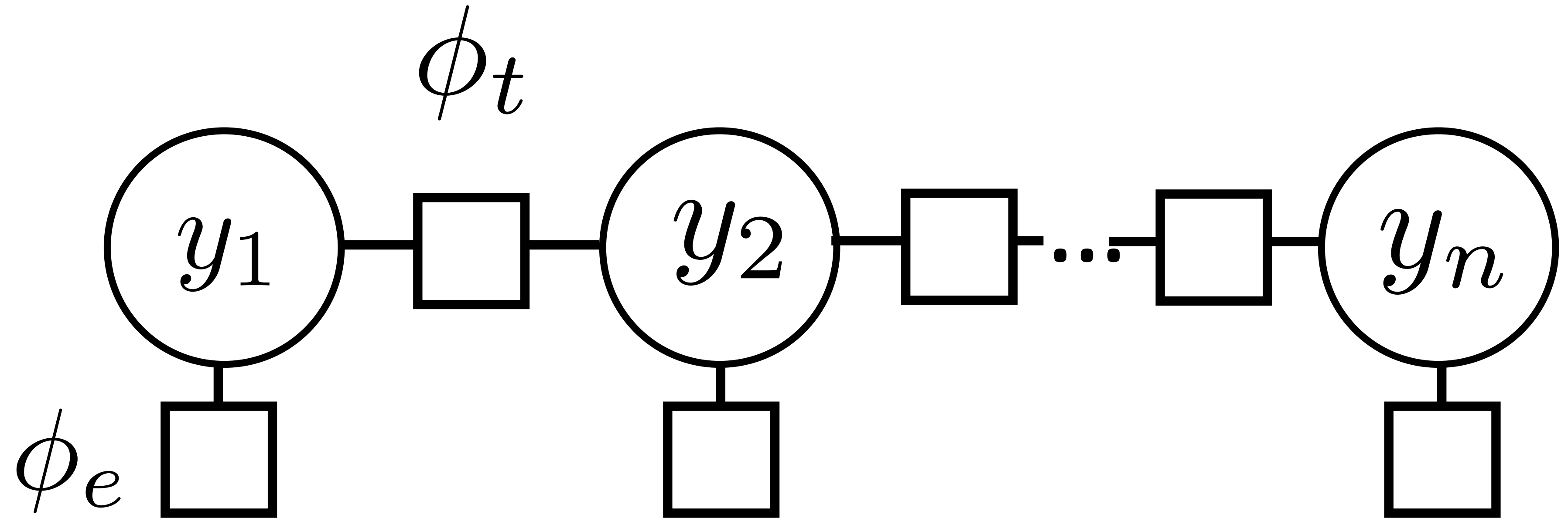
- Language is inherently sequential

B-PER I-PER O O O B-LOC O O O B-ORG O O

Barack Obama will travel to Hangzhou today for the G20 meeting .

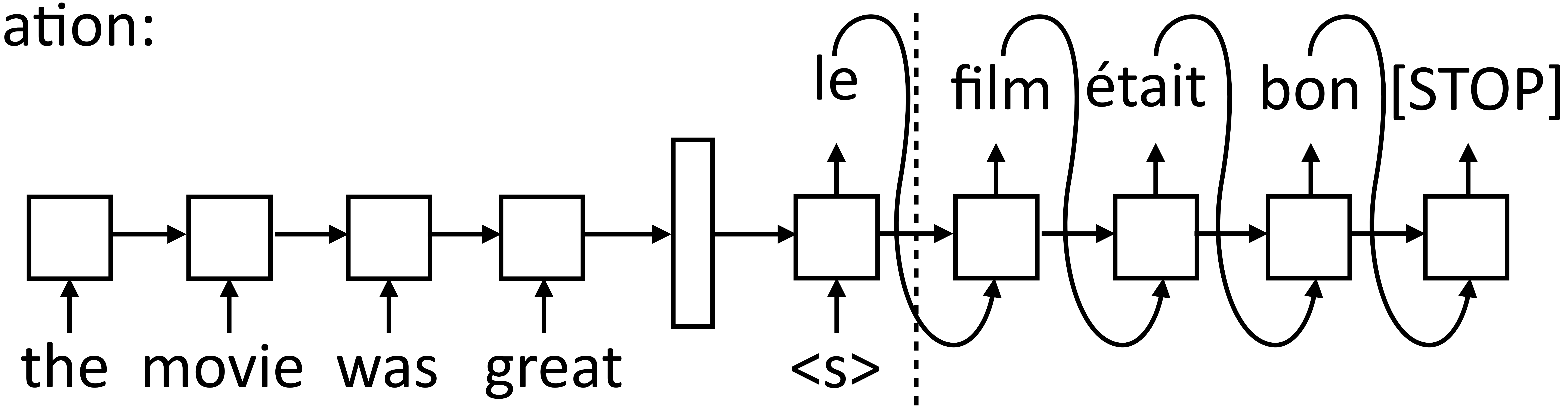
PERSON LOC ORG

- Can do language analysis with sequence models

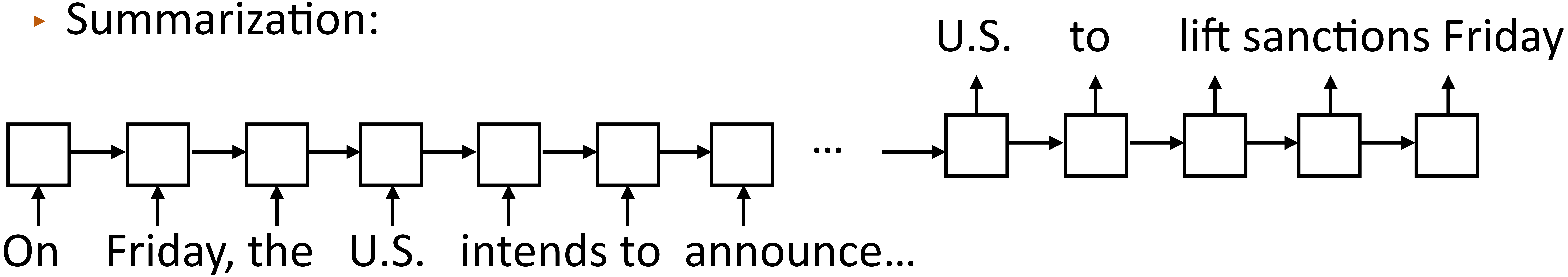


Sequential Structure: Generation

- Translation:



- Summarization:



Higher-level Structure: IE/QA

- Combine information to make deductions and reason across sentences

She's a lovely girl. She has long and black hair. She is quite tall and slim. Her eyes are bright and black. She is 13 years old. She is good at singing. She likes listening to music. She is S.H.E.'s fan . Do you know Conan? He is a little detective .The lovely girl also likes him. Oh, sorry. I forget to tell you who the girl is. It's me. I'm a lovely girl. You can call me Kacely or Kacelin. Now I study at Sunshine Middle School. I'm in Class 1, Grade 7. Every day, I get up at 6:00 a.m. The classes begin at 7 o'clock. I like lunchtime because I can chat with my friends at that time. After school, I usually play badminton with my friends. I like playing badminton and I am good at it. I want to be a superstar when I grow up.

Kacely is a 12-year-old girl. She currently goes to Sunshine Middle School .

Q: Kacely is a ____?

A) student

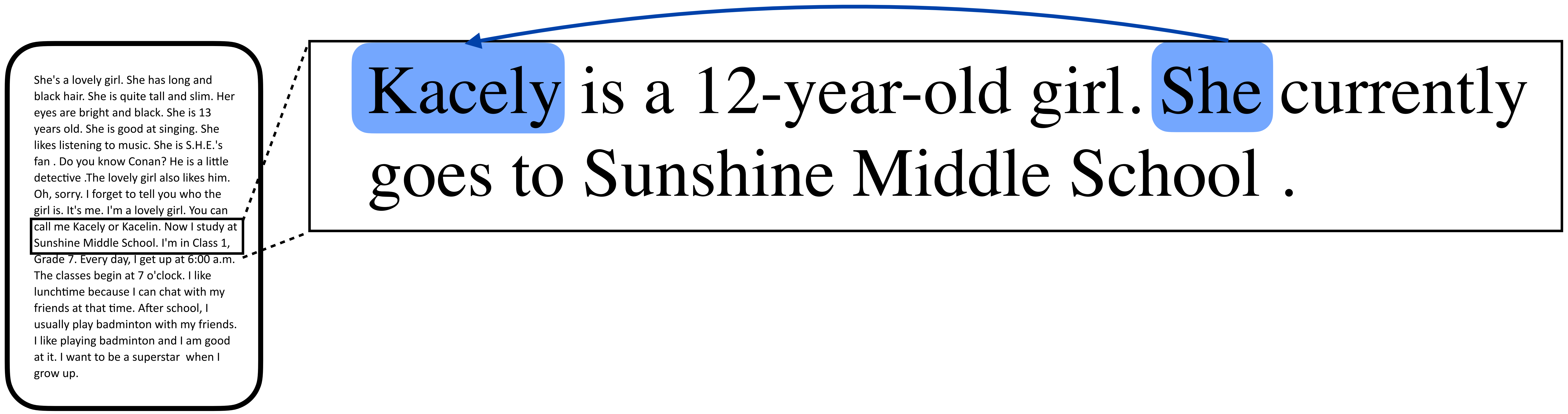
B) teacher

C) principal

D) parent

Higher-level Structure: IE/QA

- Combine information to make deductions and reason across sentences

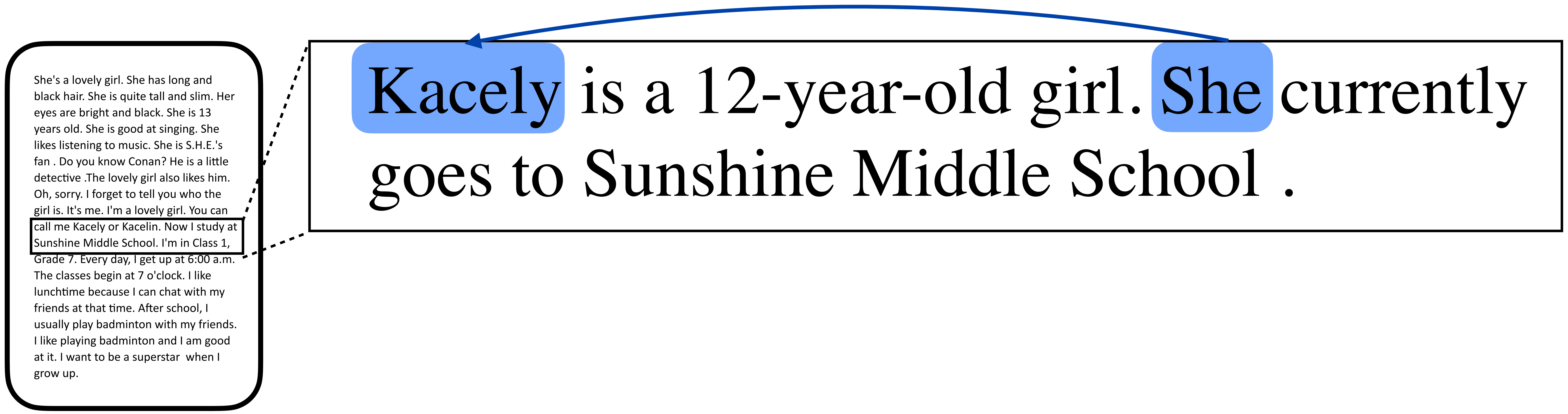


Q: Kacely is a ____?

A) student
B) teacher
C) principal
D) parent

Higher-level Structure: IE/QA

- Combine information to make deductions and reason across sentences



Q: Kacely is a ____?

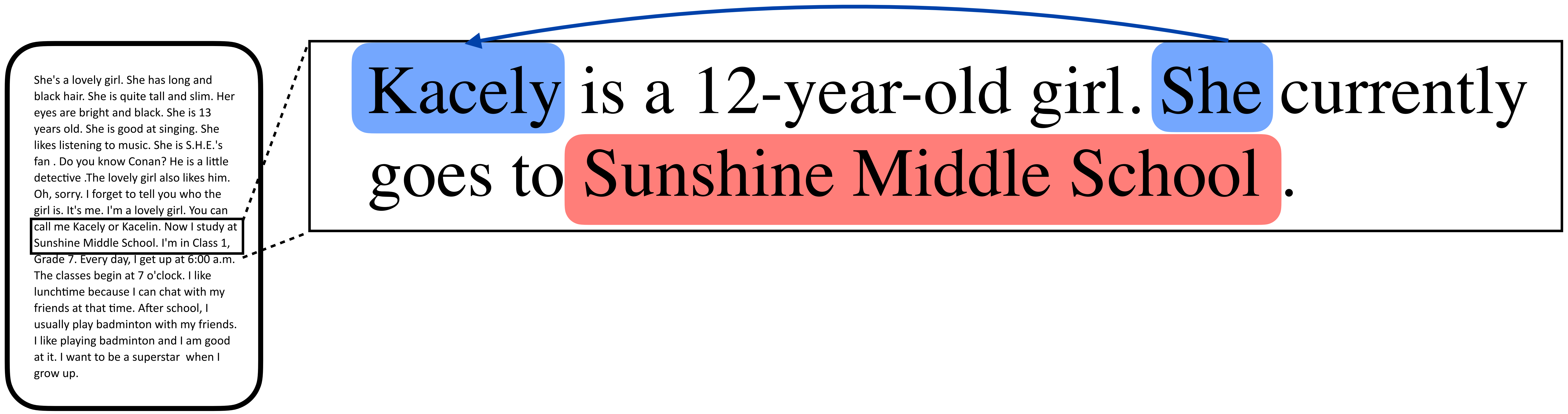
A) student
B) teacher
C) principal
D) parent

She → Kacely

coreference

Higher-level Structure: IE/QA

- Combine information to make deductions and reason across sentences



Q: Kacely is a ____?

A) student
B) teacher
C) principal
D) parent

She → Kacely coreference

Kacely goes to school parsing

Higher-level Structure: IE/QA

- Combine information to make deductions and reason across sentences

She's a lovely girl. She has long and black hair. She is quite tall and slim. Her eyes are bright and black. She is 13 years old. She is good at singing. She likes listening to music. She is S.H.E.'s fan . Do you know Conan? He is a little detective .The lovely girl also likes him. Oh, sorry. I forget to tell you who the girl is. It's me. I'm a lovely girl. You can call me Kacely or Kacelin. Now I study at Sunshine Middle School. I'm in Class 1, Grade 7. Every day, I get up at 6:00 a.m. The classes begin at 7 o'clock. I like lunchtime because I can chat with my friends at that time. After school, I usually play badminton with my friends. I like playing badminton and I am good at it. I want to be a superstar when I grow up.

Kacely is a 12-year-old girl. She currently goes to Sunshine Middle School .

Q: Kacely is a ____?

A) student
B) teacher
C) principal
D) parent

She → Kacely — coreference

Kacely goes to school — parsing

Kacely goes to school ENTAILS Kacely is a **student** — entailment

Where do we go from here?

Where do we go from here?

- Neural networks let us learn from data in an end-to-end way, very powerful learners

Where do we go from here?

- Neural networks let us learn from data in an end-to-end way, very powerful learners

- Structure imposes inductive biases in these networks

Where do we go from here?

- Neural networks let us learn from data in an end-to-end way, very powerful learners
- Structure imposes inductive biases in these networks
- Need to solve all of these challenges: leverage information across whole dialogues/documents, ground systems in the world — otherwise systems are inherently limited

Where do we go from here?

- Neural networks let us learn from data in an end-to-end way, very powerful learners

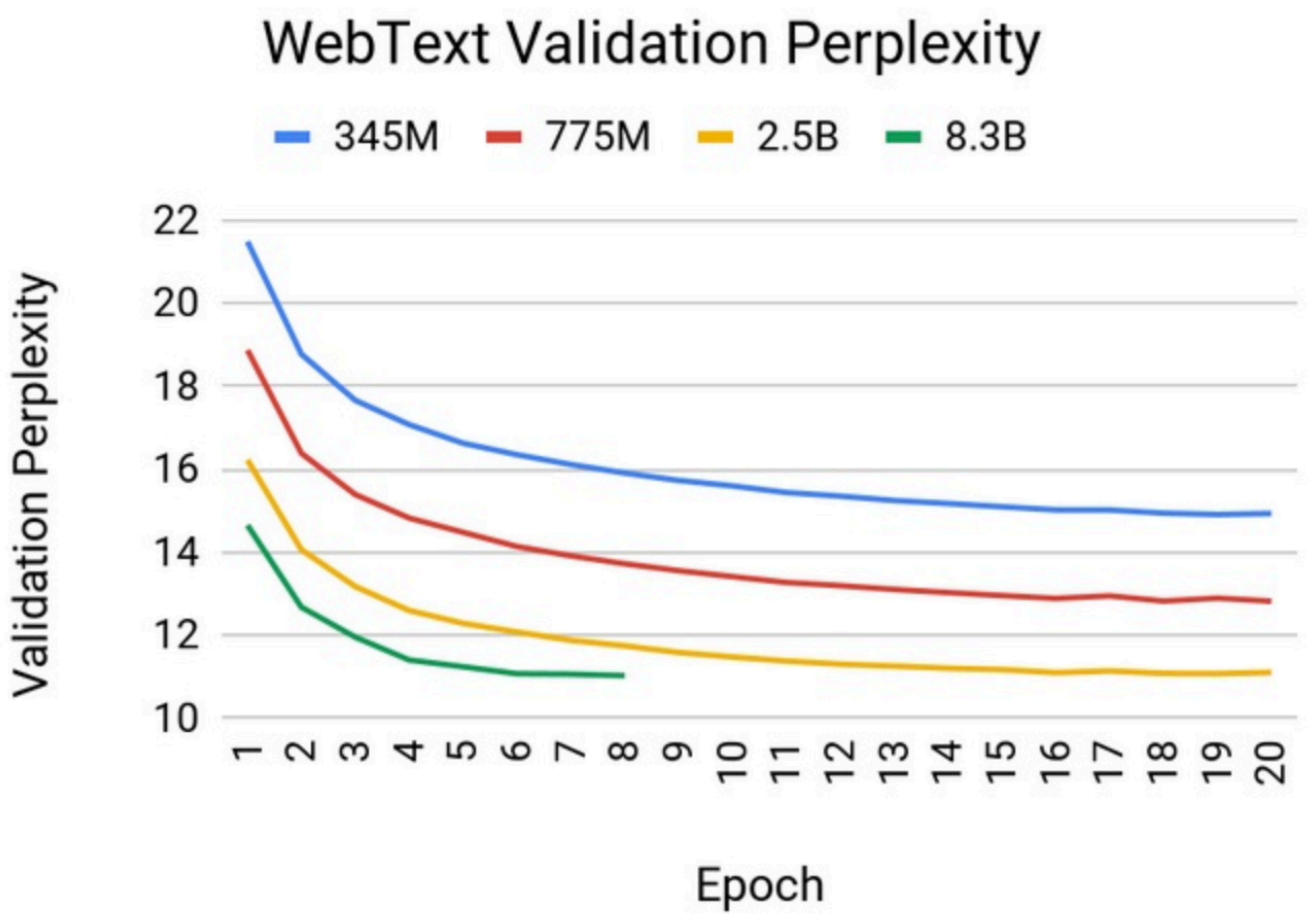
- Structure imposes inductive biases in these networks

- Need to solve all of these challenges: leverage information across whole dialogues/documents, ground systems in the world — otherwise systems are inherently limited

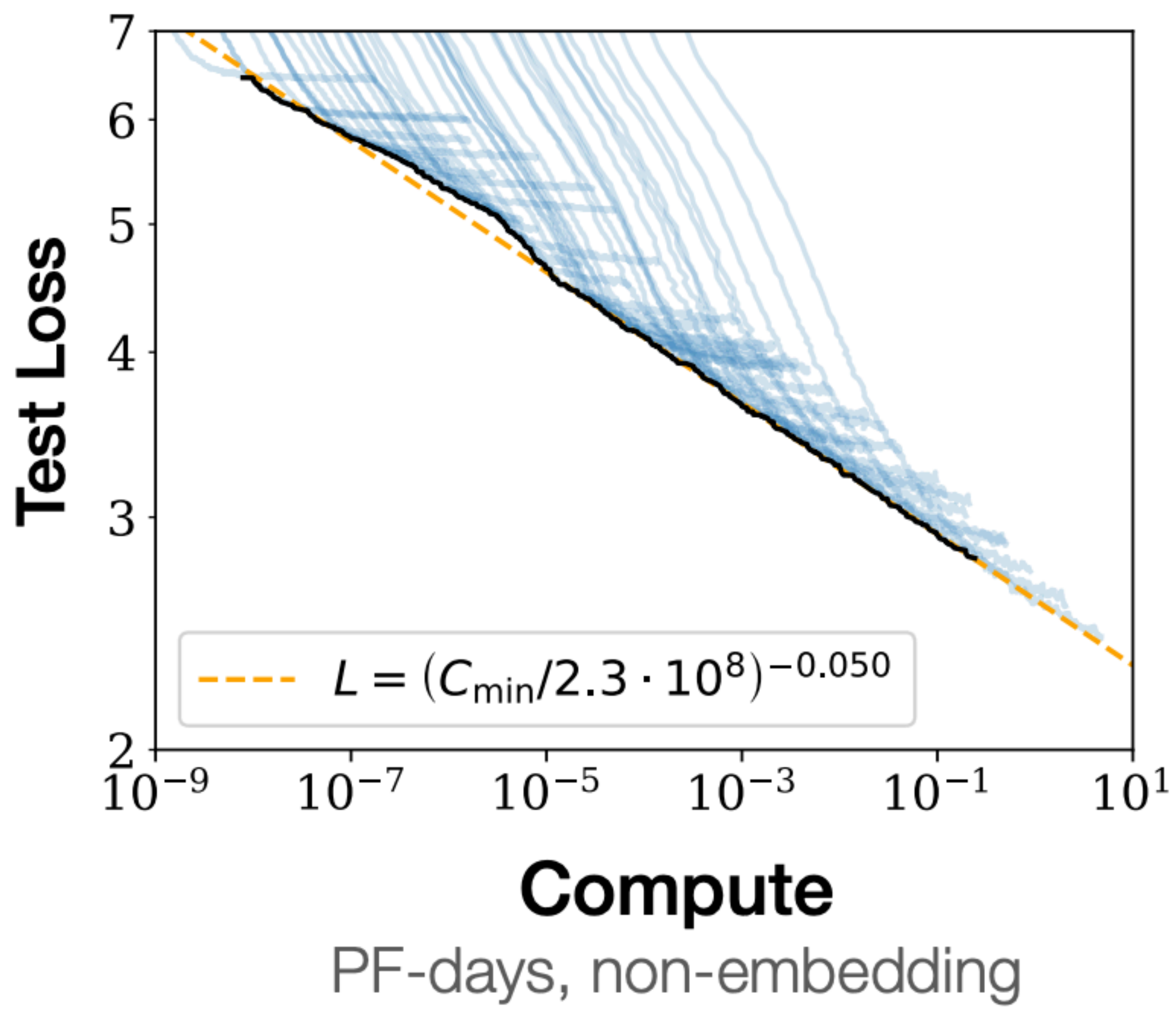
- Scaling to larger NLP systems — documents rather than sentences, books rather than documents

Where do we go from here?

- Question: what are the scaling limits of large language models?
- NVIDIA: trained 8.3B parameter GPT model (5.6x the size of GPT-2), showed lower perplexity from this
- Didn’t catch on and wasn't used for much



Scaling Laws



- Each model is a different-sized LM (GPT-style)
- With more compute, larger models get further down the loss "frontier"
- Building a bigger model (increasing compute) will decrease test loss!

Kaplan et al. (2020)

Scaling Laws

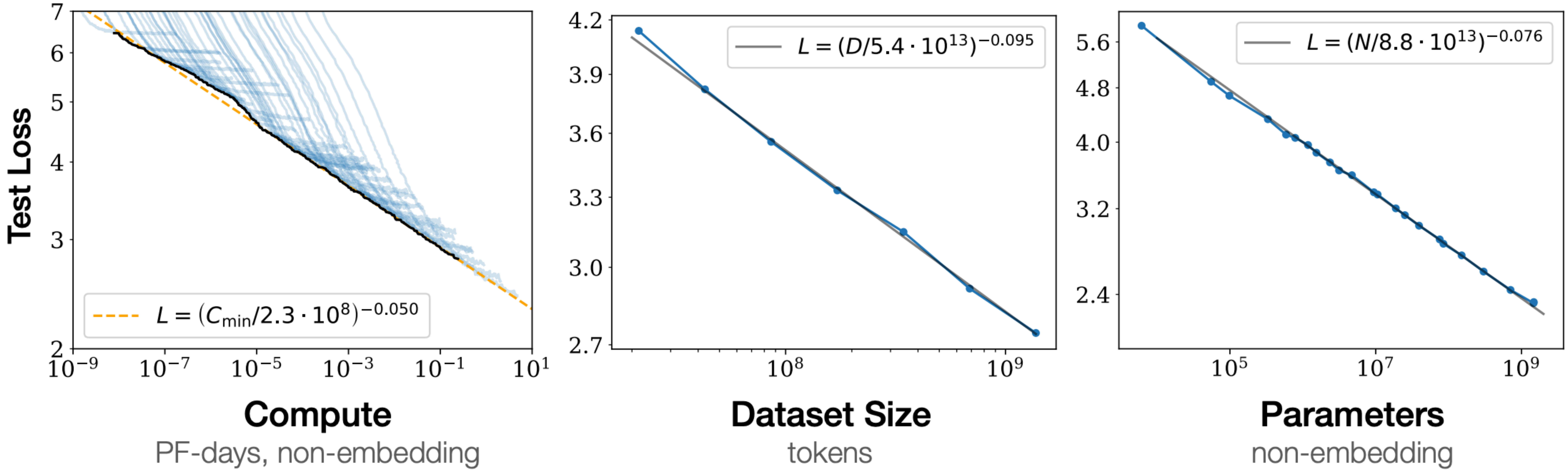


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

- These scaling laws suggest how to set model size, dataset size, and training time for big datasets

Kaplan et al. (2020)

GPT-3

- GPT-2 but even larger: 1.3B -> 175B parameter models

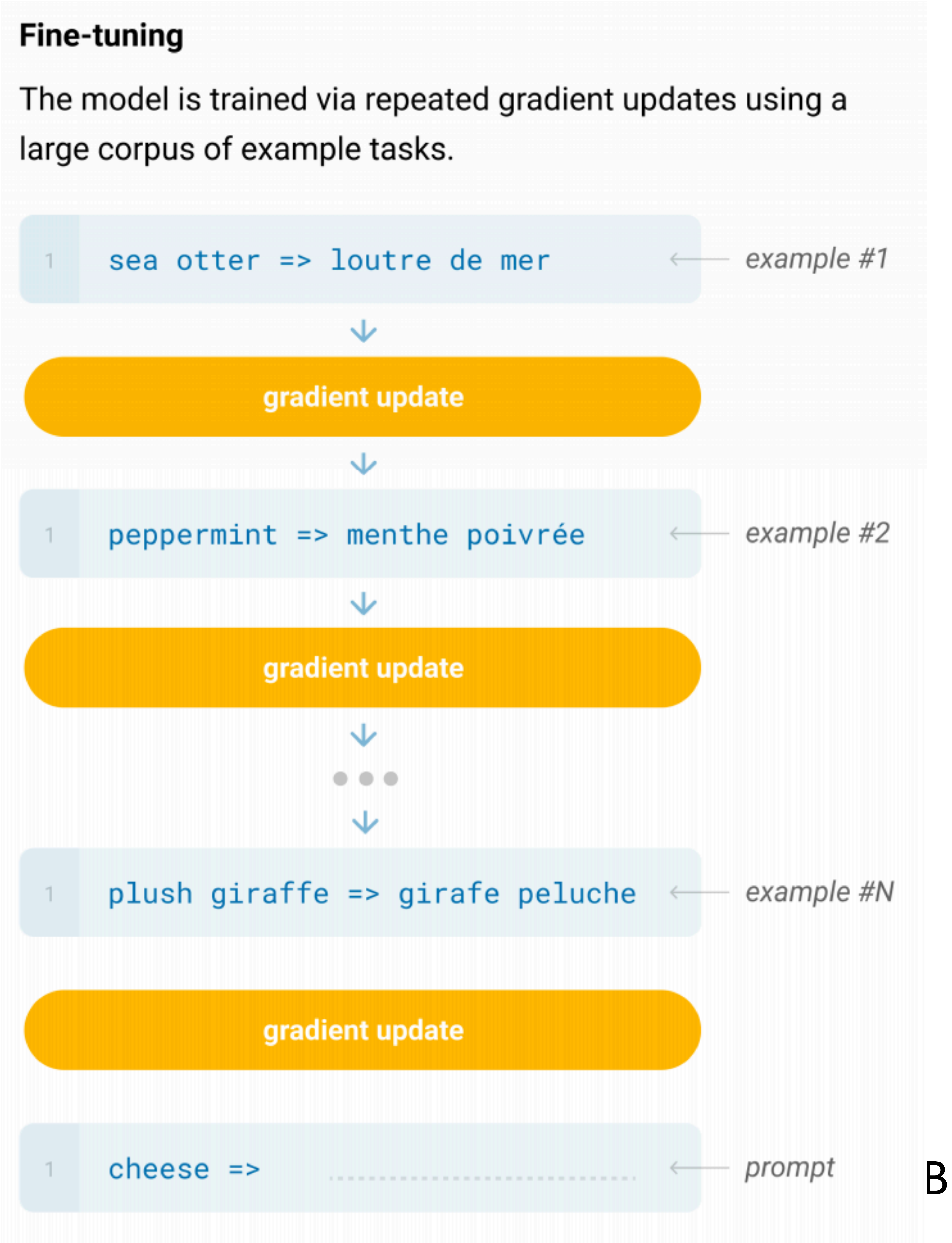
Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

- Trained on 570GB of Common Crawl
- 175B parameter model's parameters alone take >400GB to store (4 bytes per param). Trained in parallel on a "high bandwidth cluster provided by Microsoft"

Brown et al. (2020)

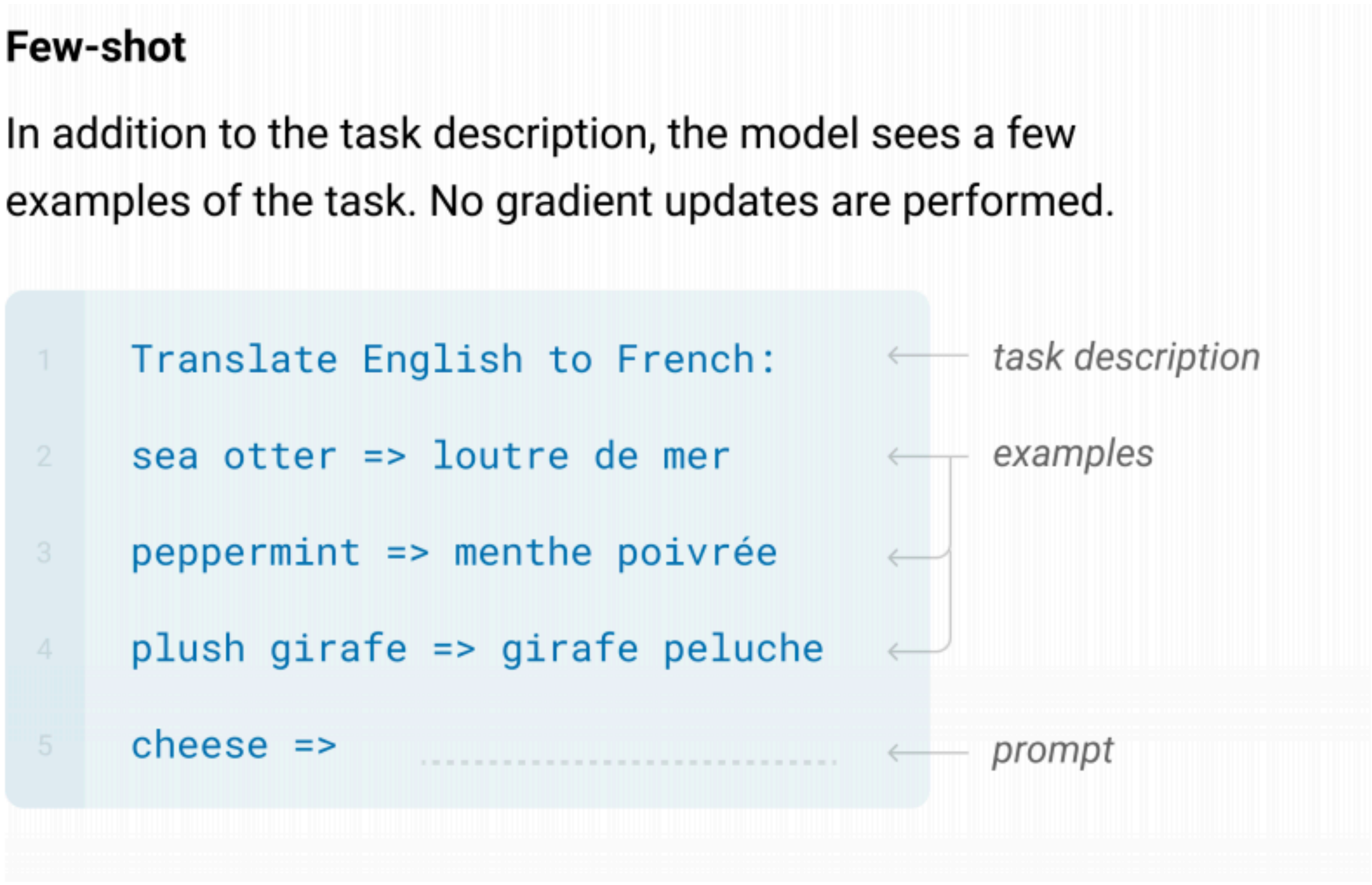
GPT-3

- This is the “normal way” of doing learning in models like GPT-2



Brown et al. (2020)

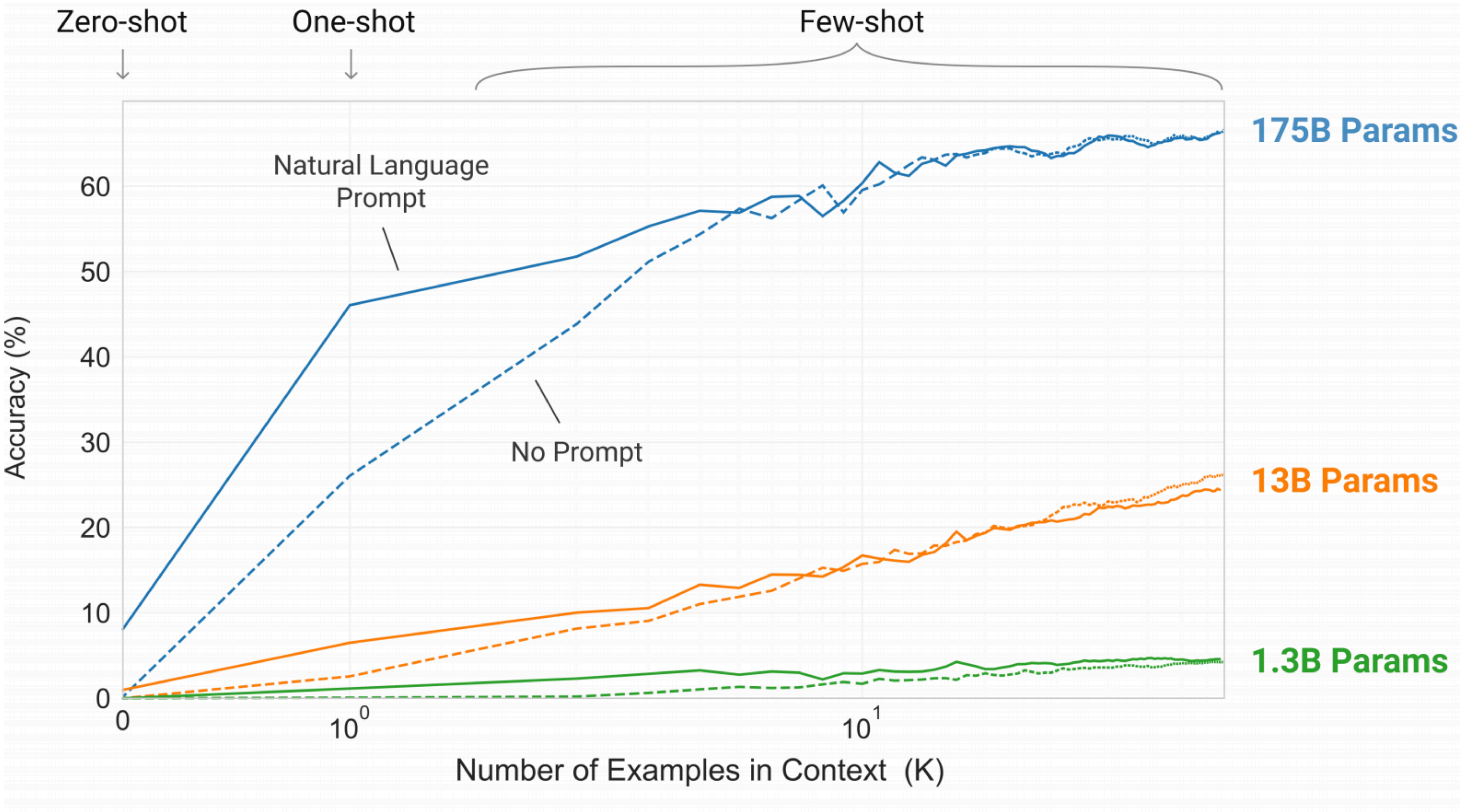
GPT-3: Few-shot Learning



Brown et al. (2020)

GPT-3

- **Key observation:** few-shot learning only works with the very largest models!



Brown et al. (2020)

GPT-3

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	**89.0**	**91.0**	**96.9**	**93.9**	**94.8**	**92.5**
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	**76.1**	**93.8**	**62.3**	**88.2**	**92.5**	**93.3**
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

- Sometimes very impressive, (MultiRC, ReCoRD), sometimes very bad
- Results on other datasets are equally mixed — but still strong for a few-shot model!

Brown et al. (2020)

Prompt Engineering

Yelp For the Yelp Reviews Full Star dataset (Zhang et al., 2015), the task is to estimate the rating that a customer gave to a restaurant on a 1- to 5-star scale based on their review's text. We define the following patterns for an input text a:

$P_1(a) =$ It was ___. a $P_2(a) =$ Just ___! || a

$P_3(a) =$ a. All in all, it was ___.

$P_4(a) =$ a || In summary, the restaurant is ___.

We define a single verbalizer v for all patterns as

$v(1) =$ terrible $v(2) =$ bad $v(3) =$ okay

$v(4) =$ good $v(5) =$ great

"verbalizer" of labels

patterns

Fine-tune LMs on initial small dataset (note: uses smaller LMs than GPT-3)

Repeat:

- Use these models to "vote" on labels for unlabeled data
- Retrain each prompt model on this dataset

Schick and Schutze et al. (2020)

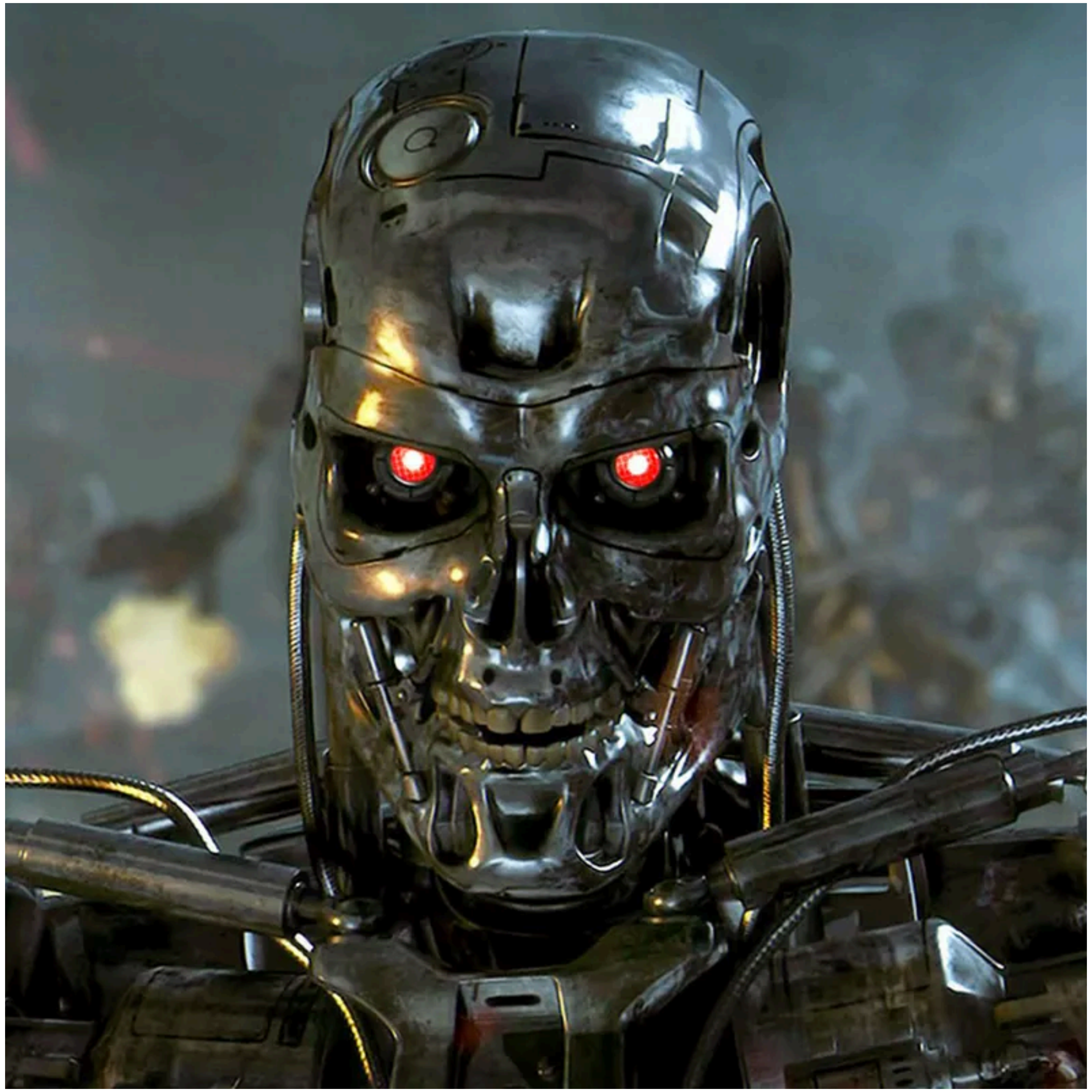
Open Questions

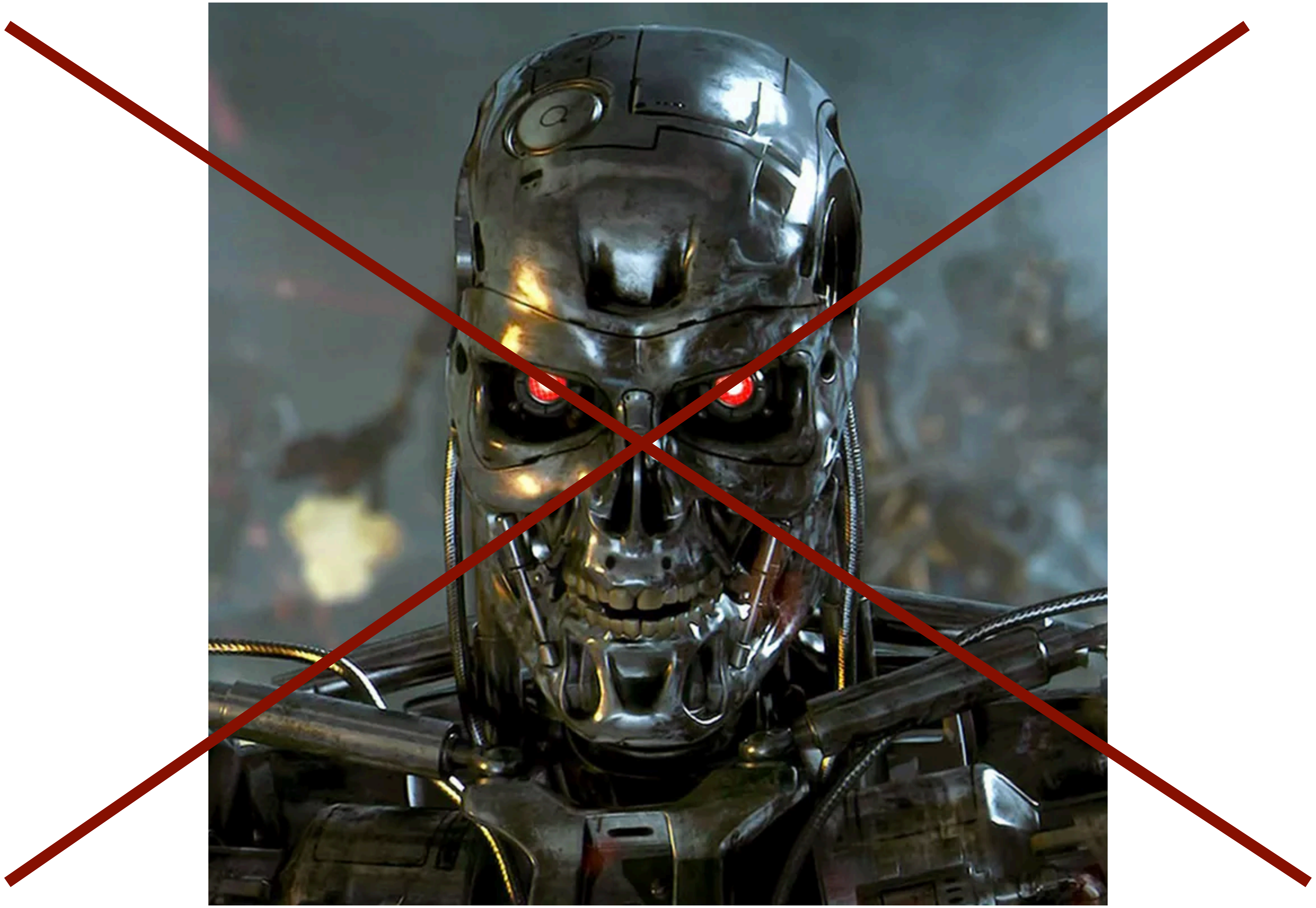
1) How much farther can we scale these models?

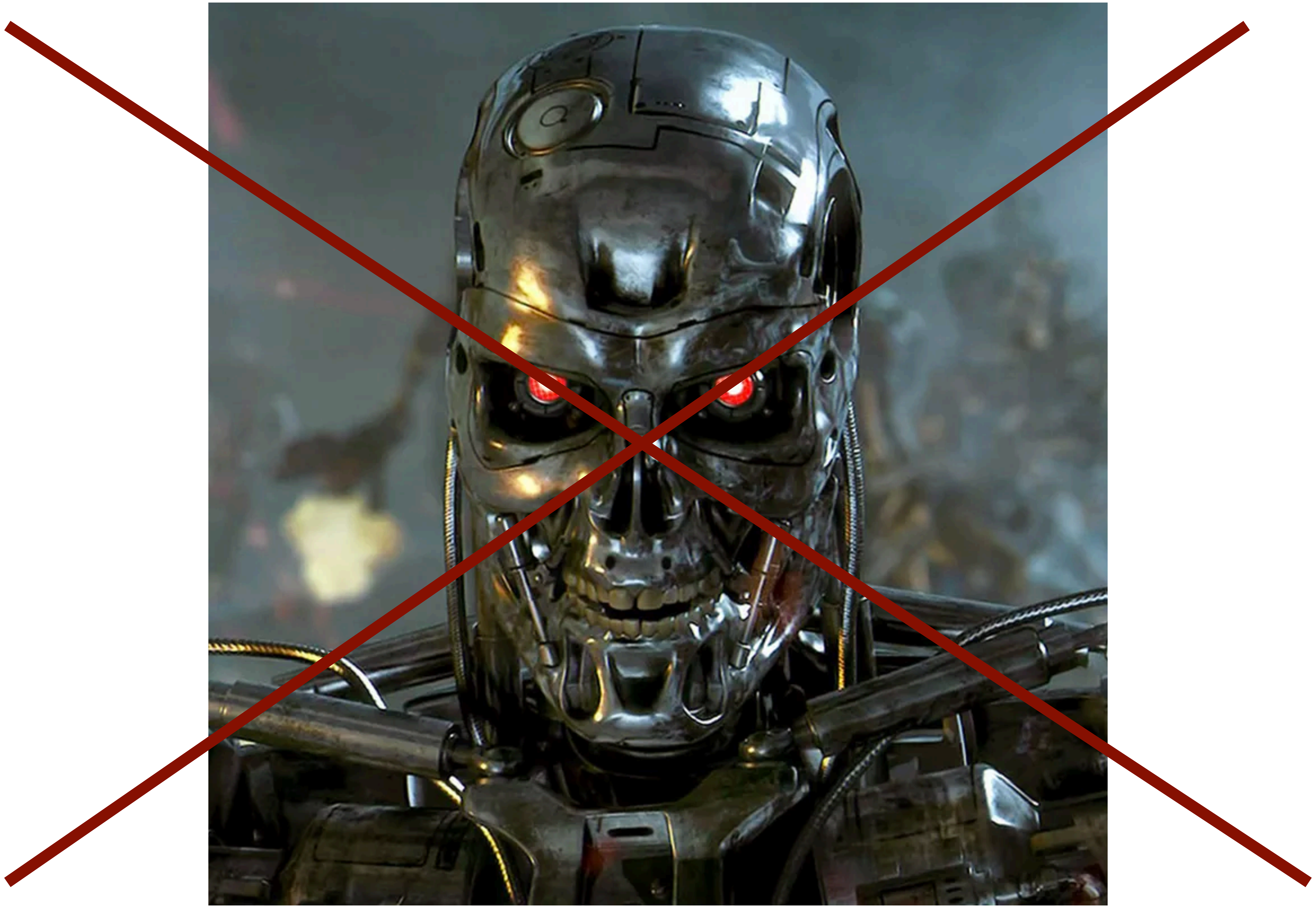
2) How do we get them to work for languages other than English (discussing this later)

3) Which will win out: prompting or fine-tuning?

Ethics in NLP — what can go wrong?







What can actually go wrong?

Pre-Training Cost (with Google/AWS)

- BERT: Base $500, Large $7000
- Grover-MEGA: $25,000
- XLNet (BERT variant): $30,000 — $60,000 (unclear)
- This is for a single pre-training run…developing new pre-training techniques may require many runs
- *Fine-tuning* these models can typically be done with a single GPU (but may take 1-3 days for medium-sized datasets)

https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-ai-models/

Pre-Training Cost (with Google/AWS)

- GPT-3: estimated to be $4.6M. This cost has a large carbon footprint
  - Carbon footprint: equivalent to driving 700,000 km by car (source: Anthropocene magazine)
  - (Counterpoints: GPT-3 isn't trained frequently, equivalent to 100 people traveling 7000 km for a conference, can use renewables)
- BERT-Base pre-training: carbon emissions roughly on the same order as a single passenger on a flight from NY to San Francisco

Strubell et al. (2019)

https://lambdalabs.com/blog/demystifying-gpt-3/

https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Bias Amplification

Zhao et al. (2017)

Bias Amplification

- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias

Zhao et al. (2017)

Bias Amplification

- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias

- Can we constrain models to avoid this while achieving the same predictive accuracy?

Zhao et al. (2017)

Bias Amplification

- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias
- Can we constrain models to avoid this while achieving the same predictive accuracy?
- Place constraints on proportion of predictions that are men vs. women?

Zhao et al. (2017)

Bias Amplification

Zhao et al. (2017)

Bias Amplification

$$\max_{\{y^i\}\in\{Y^i\}} \quad \sum_i f_\theta(y^i, i),$$

$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

Zhao et al. (2017)

Bias Amplification

$$\max_{\{y^i\}\in\{Y^i\}} \quad \sum_i f_\theta(y^i, i),$$
$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

Maximize score of predictions…

Zhao et al. (2017)

Bias Amplification

$$\max_{\{y^i\}\in\{Y^i\}} \sum_i f_\theta(y^i, i),$$
$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

Maximize score of predictions…

f(y, i) = score of predicting y on ith example

Zhao et al. (2017)

Bias Amplification

$$\max_{\{y^i\}\in\{Y^i\}} \quad \sum_i f_\theta(y^i, i),$$

Maximize score of predictions…

f(y, i) = score of predicting y on ith example

$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

…subject to bias constraint

Zhao et al. (2017)

Bias Amplification

$$\max_{\{y^i\}\in\{Y^i\}} \quad \sum_i f_\theta(y^i, i),$$

$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

Maximize score of predictions…

f(y, i) = score of predicting y on ith example

…subject to bias constraint

- Constraints: male prediction ratio on the test set has to be close to the ratio on the training set

Zhao et al. (2017)

Bias Amplification

$$\max_{\{y^i\}\in\{Y^i\}} \quad \sum_i f_\theta(y^i, i),$$

Maximize score of predictions…

f(y, i) = score of predicting y on ith example

$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

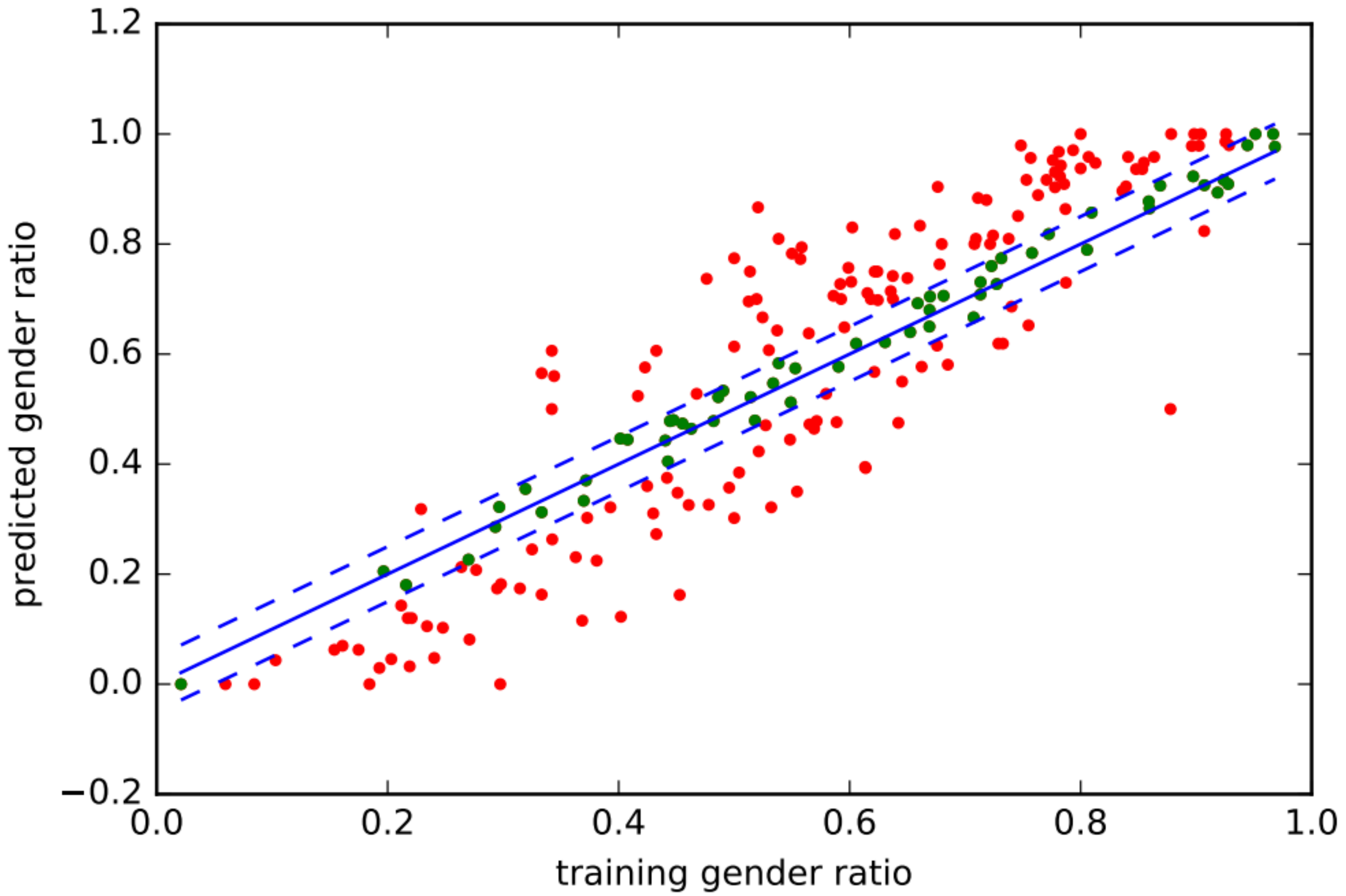
…subject to bias constraint

- Constraints: male prediction ratio on the test set has to be close to the ratio on the training set

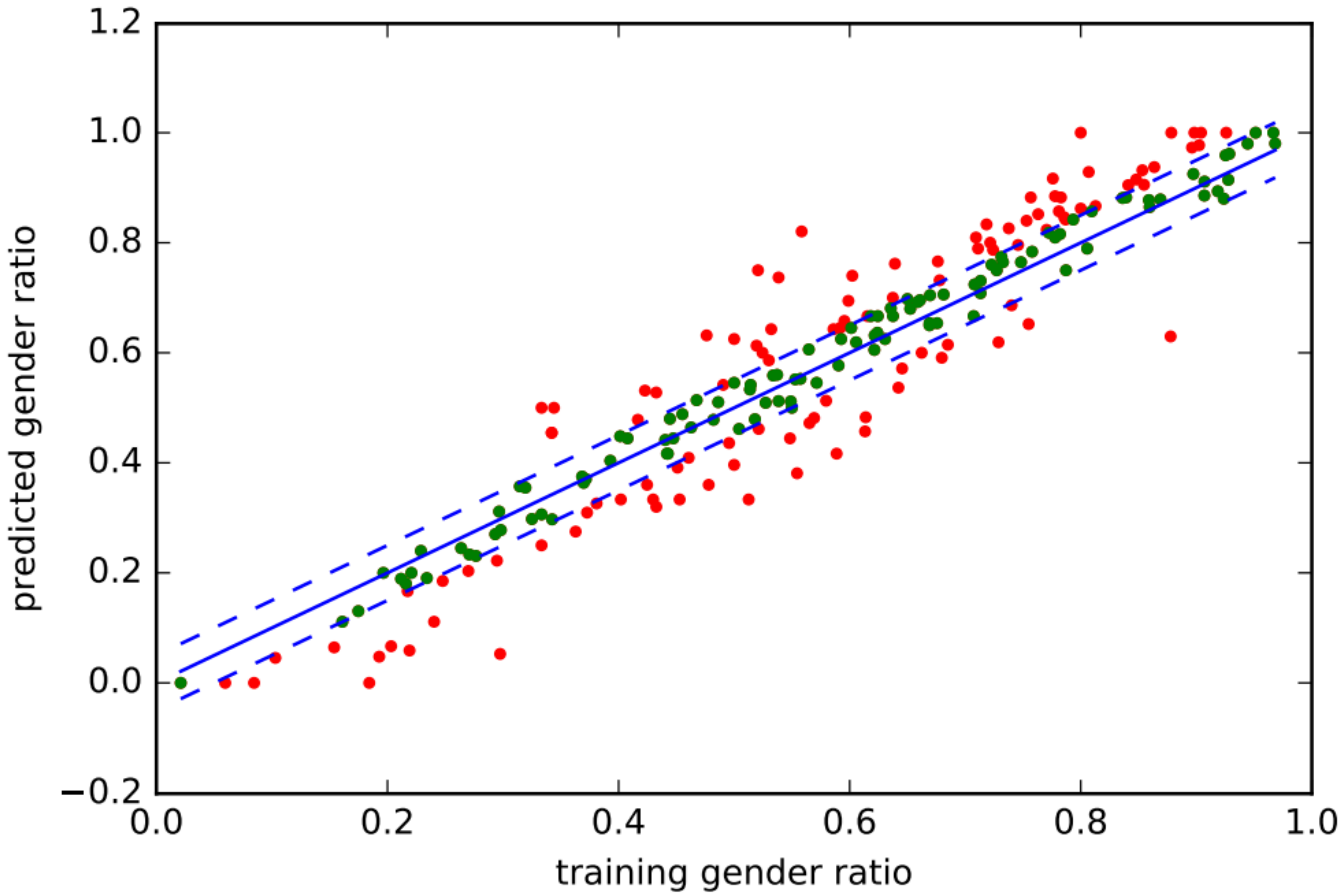
$$b^* - \gamma \leq \frac{\sum_i y^i_{v=v^*, r\in M}}{\sum_i y^i_{v=v^*, r\in W} + \sum_i y^i_{v=v^*, r\in M}} \leq b^* + \gamma \qquad (2)$$

Zhao et al. (2017)

Bias Amplification



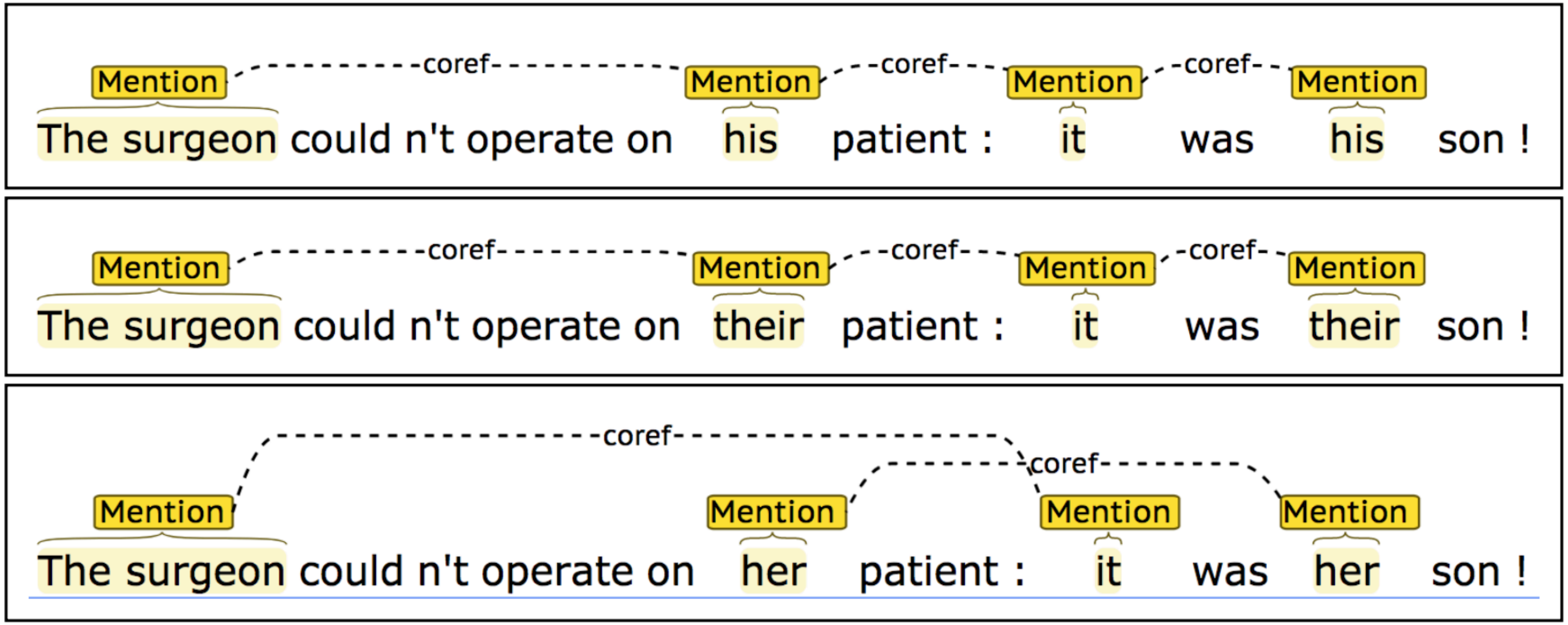
(a) Bias analysis on imSitu vSRL without RBA



(c) Bias analysis on imSitu vSRL with RBA

Zhao et al. (2017)

Bias Amplification



- Coreference: models make assumptions about genders and make mistakes as a result

Rudinger et al. (2018), Zhao et al. (2018)

Bias Amplification

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

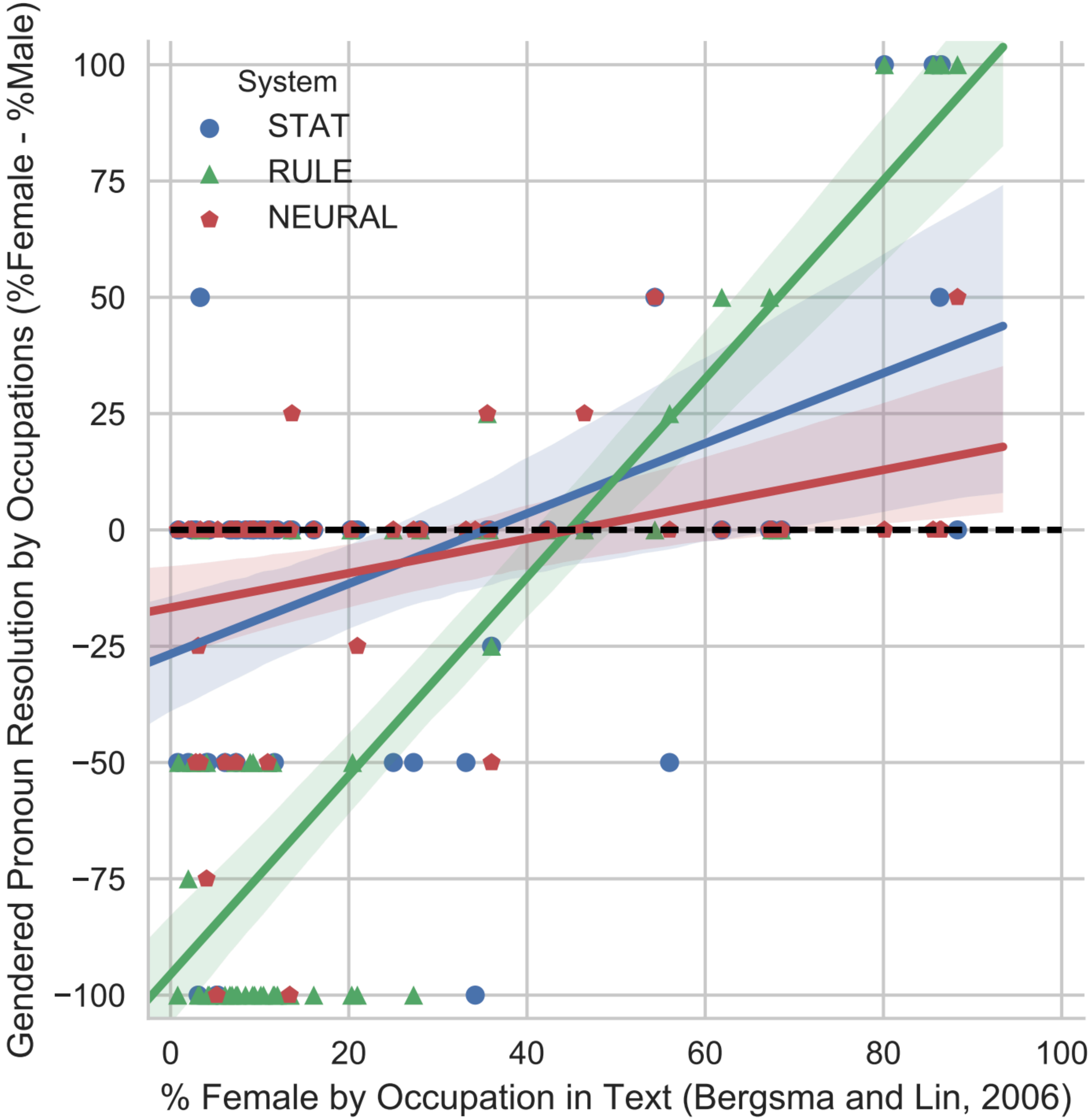
(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

- Can form Winograd schema-like test set to investigate

Rudinger et al. (2018), Zhao et al. (2018)

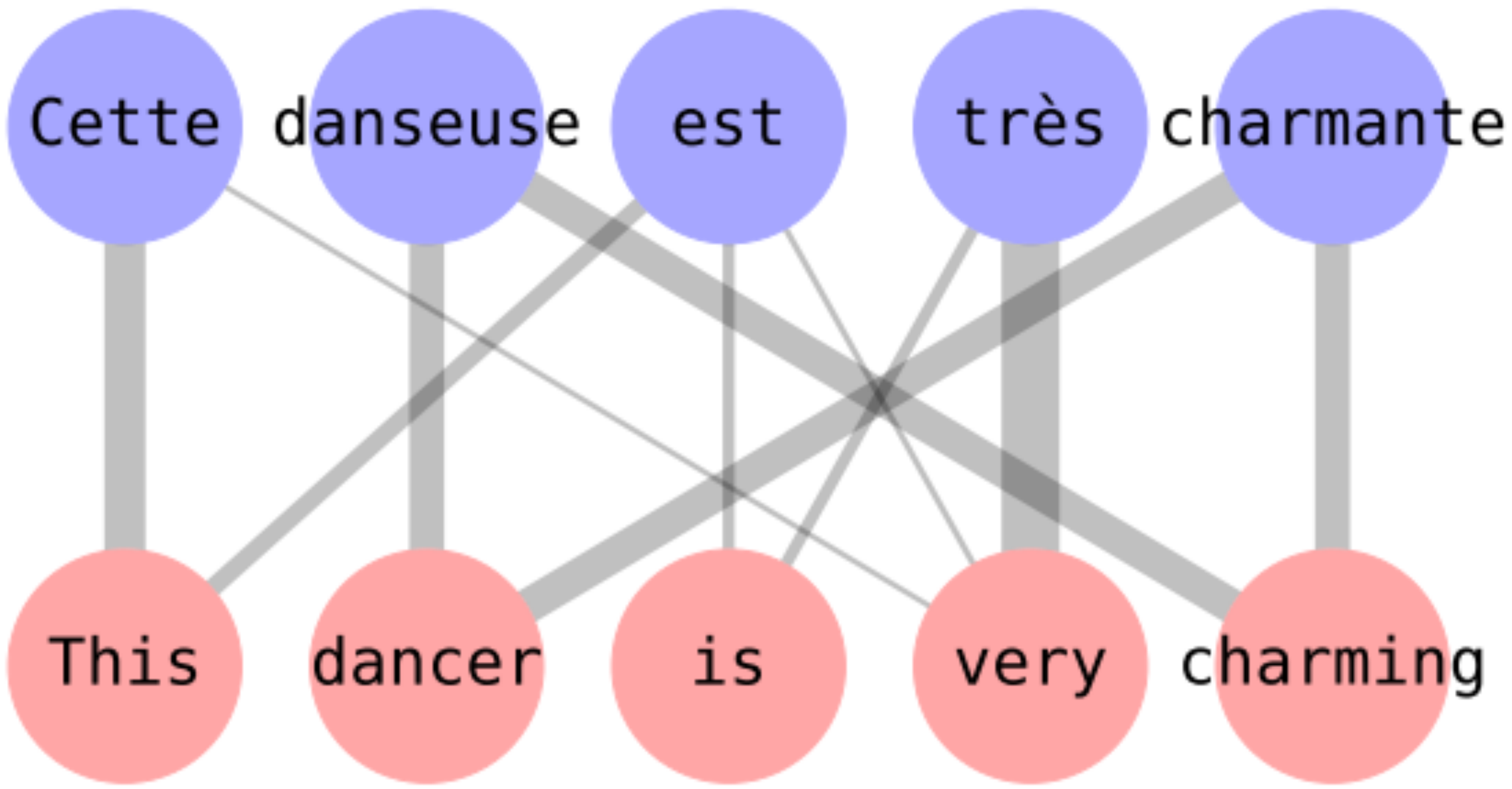
Bias Amplification



- Test set is balanced so a perfect model has female%-male% = 0 (black line)
- Neural models actually are a bit better at being unbiased, but are still skewed by data

Zhao et al. (2017)

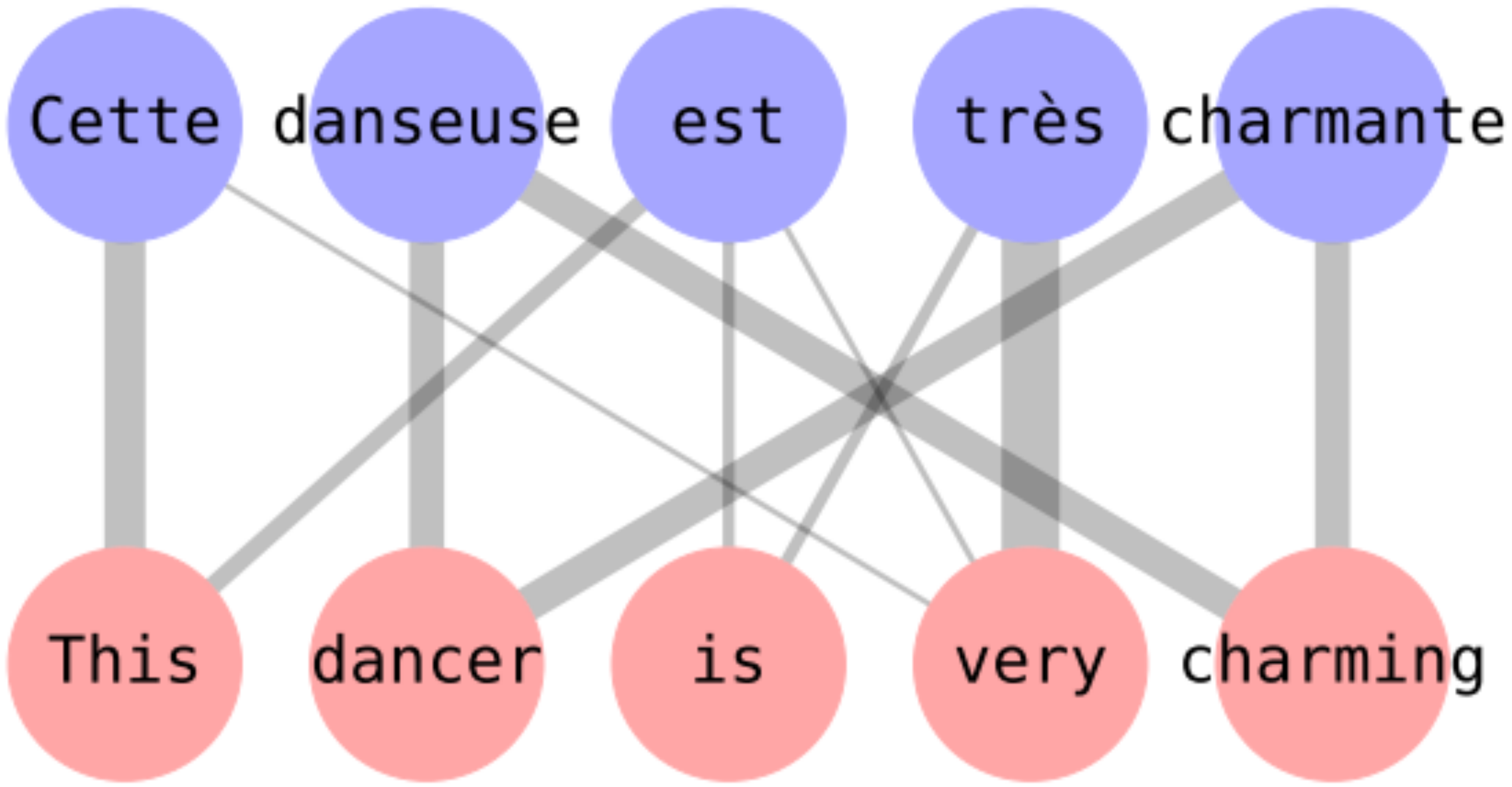
Bias Amplification



Alvarez-Melis and Jaakkola (2017)

Bias Amplification

- Harder to quantify this for machine translation

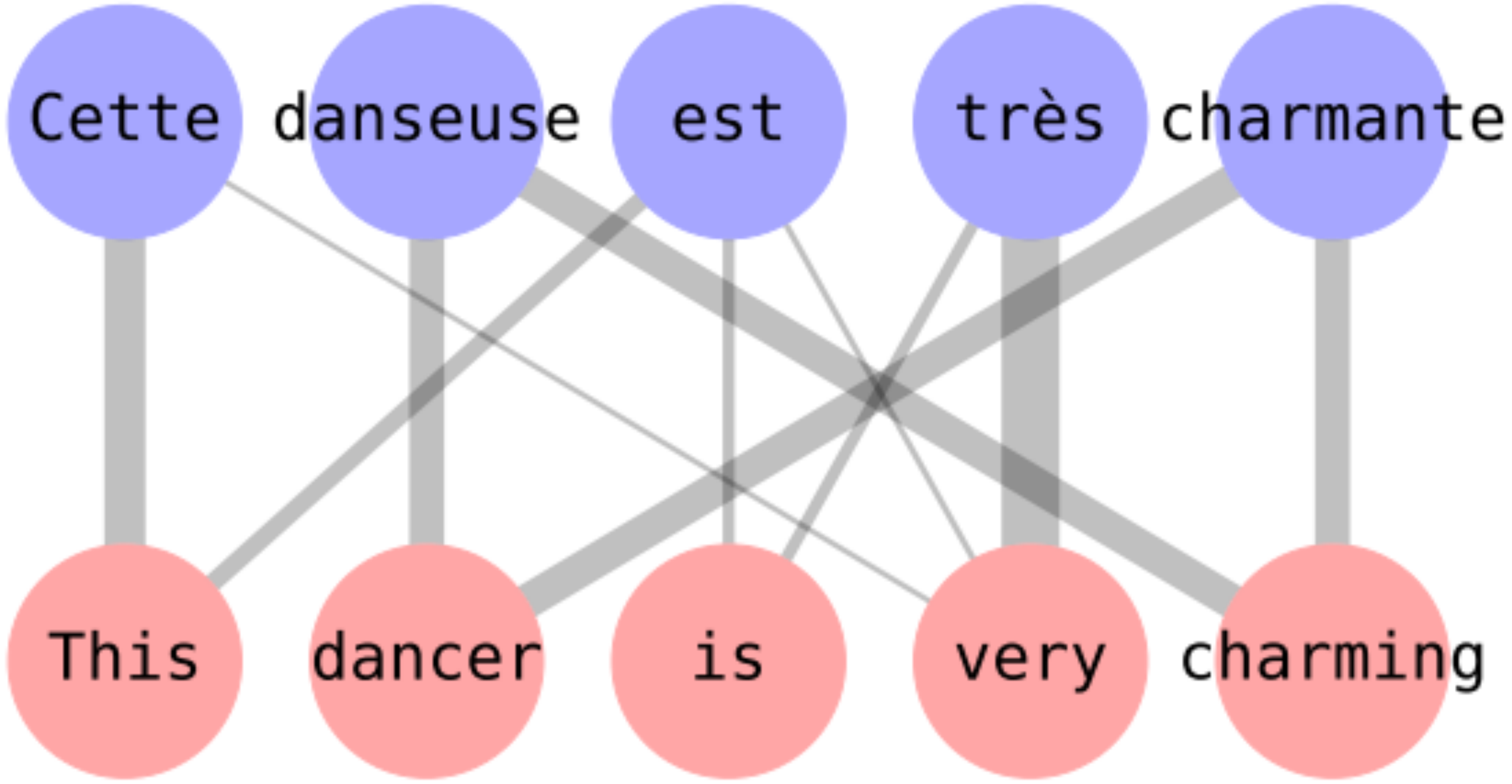


Alvarez-Melis and Jaakkola (2017)

Bias Amplification

- Harder to quantify this for machine translation

- "dancer" is assumed to be female in the context of the word "charming"… but maybe that reflects how language is used?



Alvarez-Melis and Jaakkola (2017)

Exclusion

Exclusion

- Most of our annotated data is English data, especially newswire

Exclusion

- Most of our annotated data is English data, especially newswire
- What about:

Exclusion

- Most of our annotated data is English data, especially newswire
- What about:

Dialects?

Exclusion

- Most of our annotated data is English data, especially newswire
- What about:

  Dialects?

  Other languages? (Non-European/CJK)

Exclusion

- Most of our annotated data is English data, especially newswire
- What about:

  Dialects?

  Other languages? (Non-European/CJK)

  Codeswitching?

Unethical Use

Unethical Use

- Surveillance applications?

Unethical Use

- Surveillance applications?
- Generating convincing fake news / fake comments?

FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	—
Hi, I'd like to comment on	I'm a voter worried about	In the matter of
net neutrality regulations.	Internet freedom.	NET NEUTRALITY.
I want to	I'd like to	I strongly
implore	ask	ask
the government to	Ajit Pai to	the commission to
repeal	repeal	reverse
Barack Obama's	President Obama's	Tom Wheeler's
decision to	order to	scheme to
regulate	regulate	take over
internet access.	broadband.	the web.
Individuals,	people like me,	People like me,
rather than	rather than	rather than

Unethical Use

- Surveillance applications?
- Generating convincing fake news / fake comments?

FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	---
Hi, I'd like to comment on	I'm a voter worried about	In the matter of
net neutrality regulations.	Internet freedom.	NET NEUTRALITY.
I want to	I'd like to	I strongly
implore	ask	ask
the government to	Ajit Pai to	the commission to
repeal	repeal	reverse
Barack Obama's	President Obama's	Tom Wheeler's
decision to	order to	scheme to
regulate	regulate	take over
internet access.	broadband.	the web.
Individuals,	people like me,	People like me,
rather than	rather than	rather than

- What if these were undetectable?

Dangers of Automatic Systems

THE VERGE TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE

US & WORLD TECH POLITICS

14

Facebook apologizes after wrong translation sees Palestinian man arrested for posting 'good morning'

Facebook translated his post as 'attack them' and 'hurt them'

by Thuy Ong | @ThuyOng | Oct 24, 2017, 10:43am EDT

Slide credit: The Verge

Dangers of Automatic Systems

Translations of **gay**

adjective	
homosexual	homosexual, gay, camp
alegre	cheerful, glad, joyful, happy, merry, gay
brillante	bright, brilliant, shiny, shining, glowing, glistening
vivo	live, alive, living, vivid, bright, lively
vistoso	colorful, ornate, flamboyant, colourful, gorgeous
jovial	jovial, cheerful, cheery, gay, friendly
gayo	merry, gay, showy
noun	
el homosexual	$\boxed{\text{homosexual, gay, poof, queen, faggot, fagot}}$
el jovial	gay

- Offensive terms

Slide credit: allout.org

Dangers of Automatic Systems

"Instead of relying on algorithms, which we can be accused of manipulating for our benefit, we have turned to machine learning, an ingenious way of disclaiming responsibility for anything. Machine learning is like money laundering for bias. It's a clean, mathematical apparatus that gives the status quo the aura of logical inevitability. The numbers don't lie."

\- Maciej Cegłowski

Slide credit: Sam Bowman

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization was a negative-weight feature in resumes

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization was a negative-weight feature in resumes
  - Women's colleges too

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization was a negative-weight feature in resumes
  - Women's colleges too
- Was this a bad model? May have actually modeled downstream outcomes correctly…but this can mean learning humans' biases

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

Charge-Based Prison Term Prediction with Deep Gating Network

Huajie Chen[1*] **Deng Cai**[2*] **Wei Dai**[1] **Zehui Dai**[1] **Yadong Ding**[1]
[1]NLP Group, Gridsum, Beijing, China
{chenhuajie,daiwei,daizehui,dingyadong}@gridsum.com
[2]The Chinese University of Hong Kong
thisisjcykcd@gmail.com

- Task: given case descriptions and charge set, predict the prison term

Case description: On July 7, 2017, when the defendant Cui XX was drinking in a bar, he came into conflict with Zhang XX…… After arriving at the police station, he refused to cooperate with the policeman and bited on the arm of the policeman……

Result of judgment: Cui XX was sentenced to *12* months imprisonment for *creating disturbances* and *12* months imprisonment for *obstructing public affairs*……

- Charge#1 creating disturbances term 12 months
- Charge#2 obstructing public affairs term 12 months

Chen et al. (EMNLP 2019)

Dangers of Automatic Systems

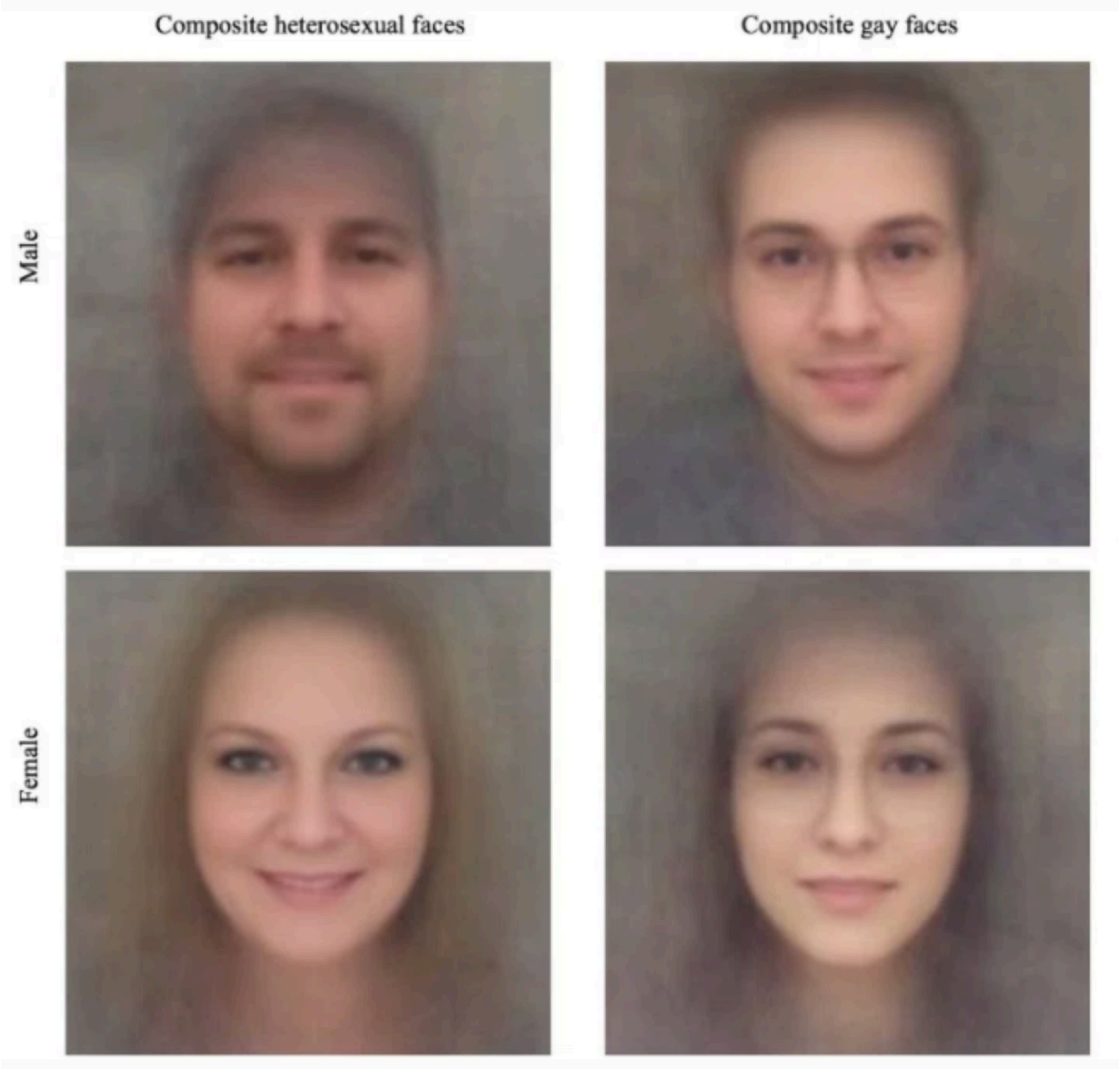
- Results: 60% of the time, the system is off by more than 20% (so 5 years => 4 or 6 years)

Model	S	EM	Acc@0.1	Acc@0.2
ATE-LSTM	66.49	7.72	16.12	33.89
MemNet	70.23	7.52	18.54	36.75
RAM	70.32	7.97	18.87	37.38
TNet	73.94	8.06	19.55	39.89
DGN	**76.48**	**8.92**	**20.66**	**42.61**

- Is this the right way to apply this?
- Are there good applications this can have?
- Is this technology likely to be misused?

The mistake of legal judgment is serious, it is about people losing years of their lives in prison, or dangerous criminals being released to reoffend. We should pay attention to how to avoid judges' over-dependence on the system. It is necessary to consider its application scenarios. In practice, we recommend deploying our system in the "Review Phase", where other judges check the judgment result by a presiding judge. Our system can serve as one anonymous checker.

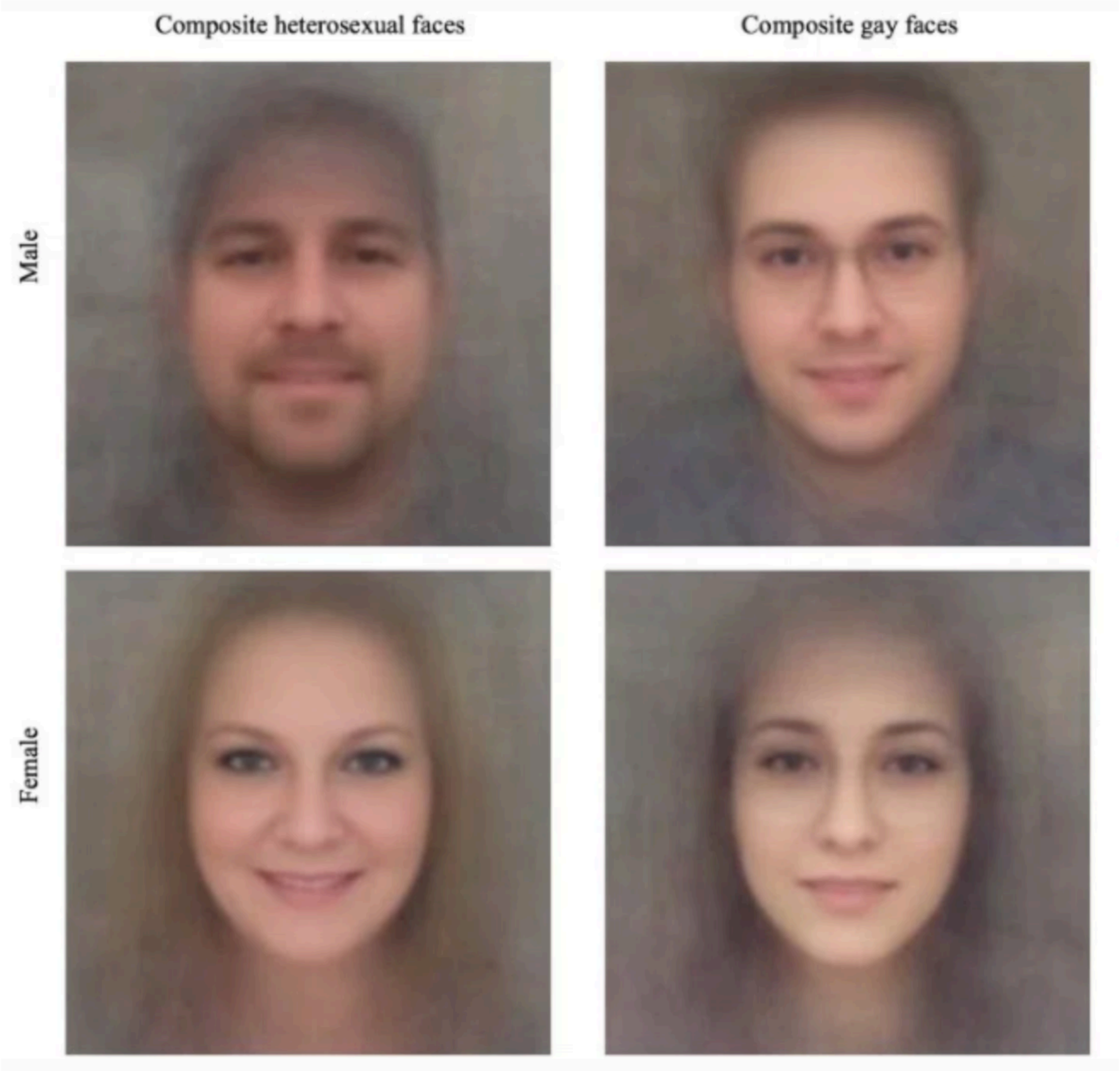
Bad Applications



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Bad Applications

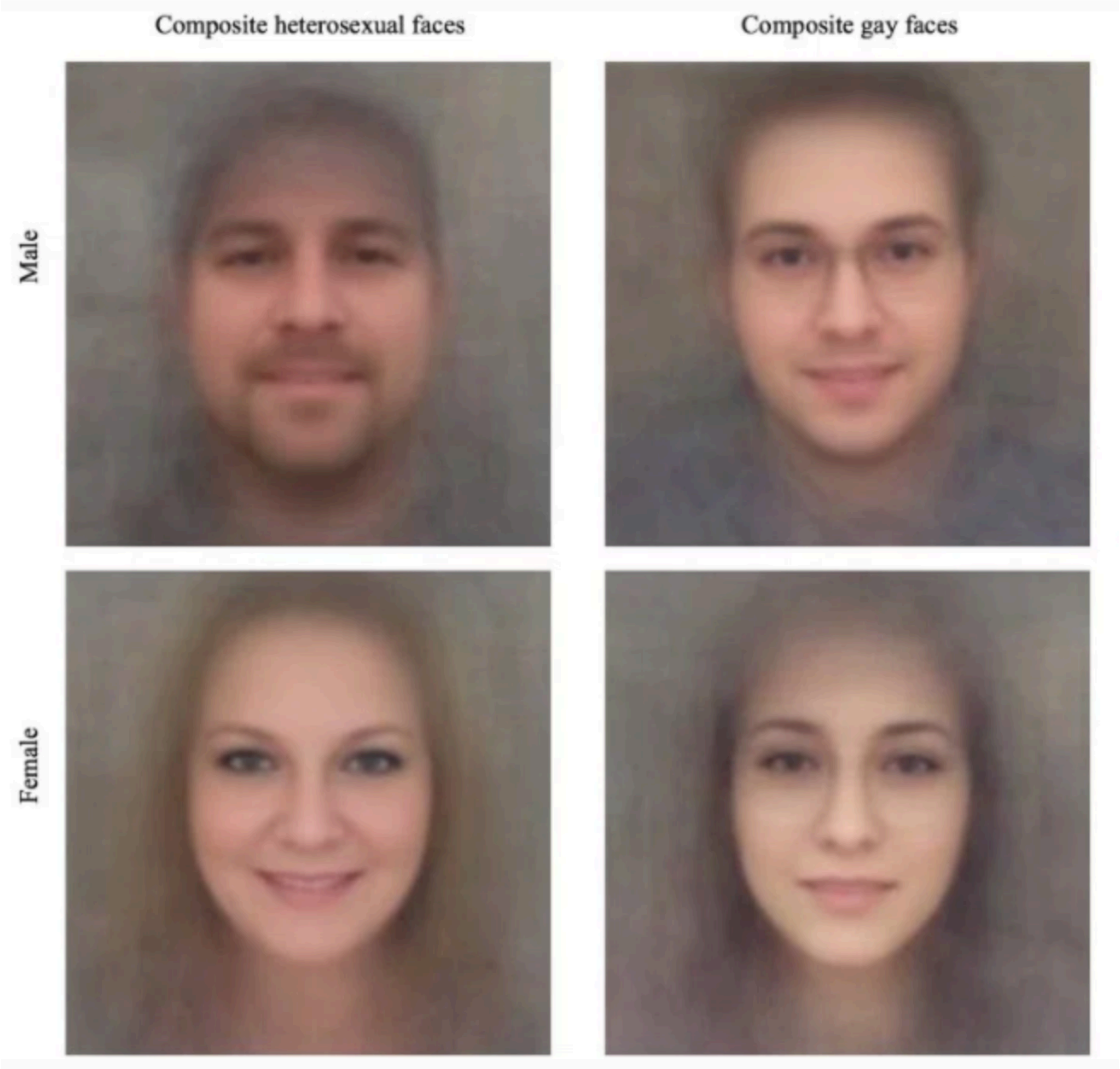
- Wang and Kosinski: gay vs. straight classification based on faces



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Bad Applications

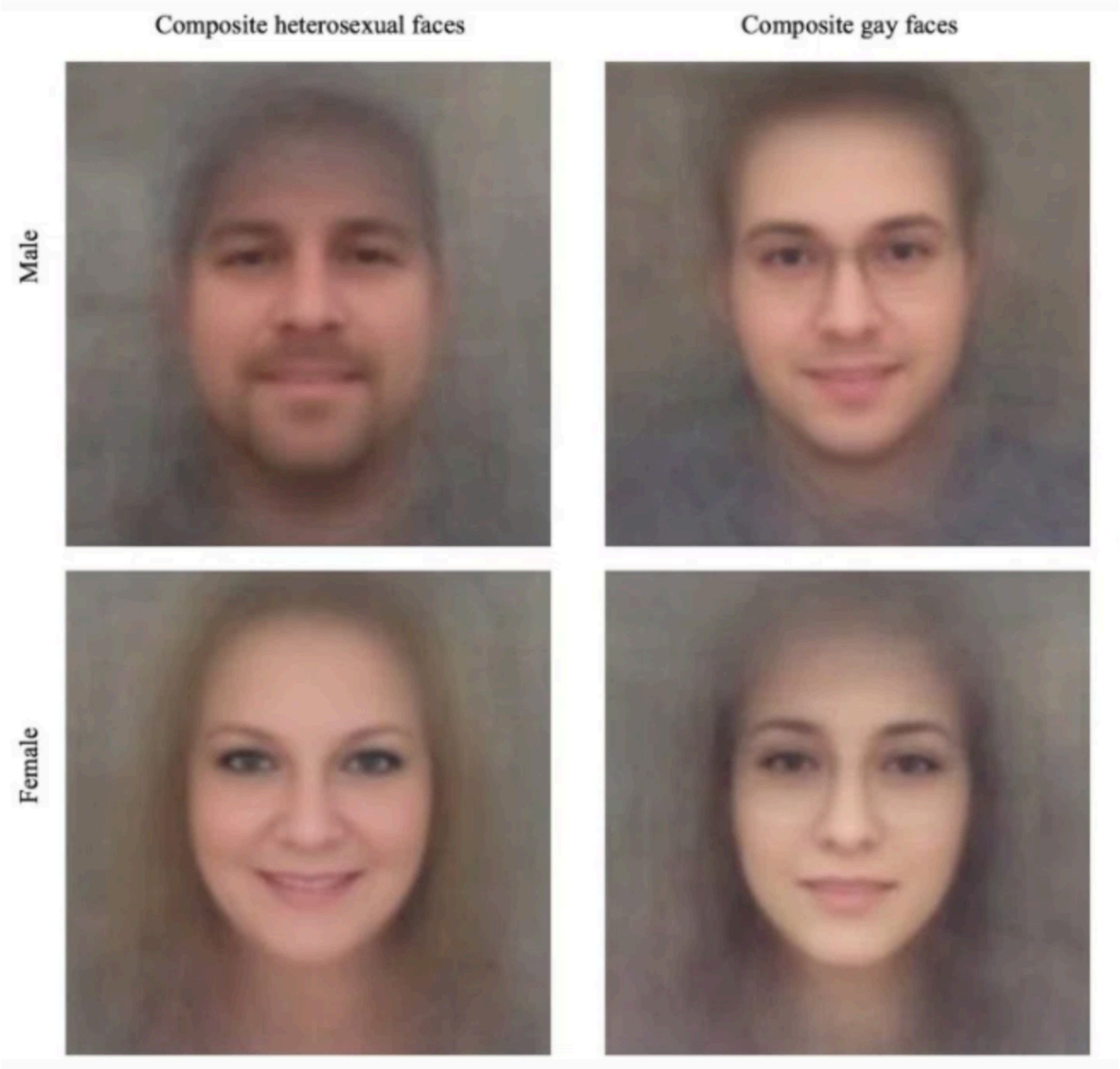
- Wang and Kosinski: gay vs. straight classification based on faces
- Authors: “this is useful because it supports a hypothesis” (physiognomy)



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

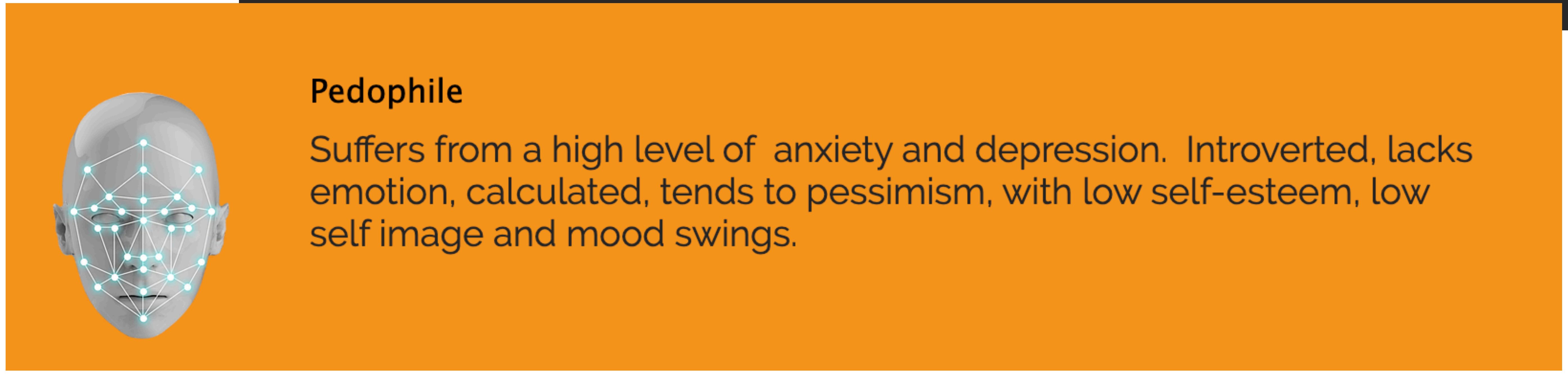
Bad Applications

- Wang and Kosinski: gay vs. straight classification based on faces
- Authors: "this is useful because it supports a hypothesis" (physiognomy)
- Blog post by Agüera y Arcas, Todorov, Mitchell: mostly social phenomena (glasses, makeup, angle of camera, facial hair) — bad science, *and* dangerous



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Unethical Use



http://www.faception.com

How to Move Forward?

- ACM Code of Ethics
  - https://www.acm.org/code-of-ethics

- Hal Daume III: Proposed code of ethics
  https://nlpers.blogspot.com/2016/12/should-nlp-and-ml-communities-have-code.html
  - Many other points, but these are relevant:
    - Contribute to society and human well-being, and minimize negative consequences of computing systems
    - Make reasonable effort to prevent misinterpretation of results
    - Make decisions consistent with safety, health, and welfare of public
    - Improve understanding of technology, its applications, and its potential consequences (pos and neg)

Final Thoughts

Final Thoughts

- You will face choices: what you choose to work on, what company you choose to work for, etc.

Final Thoughts

- You will face choices: what you choose to work on, what company you choose to work for, etc.
- Tech does not exist in a vacuum: you can work on problems that will fundamentally make the world a better place or a worse place (not always easy to tell)

Final Thoughts

- You will face choices: what you choose to work on, what company you choose to work for, etc.

- Tech does not exist in a vacuum: you can work on problems that will fundamentally make the world a better place or a worse place (not always easy to tell)

- As AI becomes more powerful, think about what we *should* be doing with it to improve society, not just what we *can* do with it