Machine Learning for NLP

Ethics and Machine Learning

Aurélie Herbelot

2021

Centre for Mind/Brain Sciences
University of Trento

Today

1. Some key notions: prediction, personalisation, bubbling, privacy...
2. The problem with representations: biases and big data.
3. The problem with language

Prediction, personalisation, bubbling, privacy

Brave New World

Artificial Intelligence and Life in 2030
https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf

“Society is now at a crucial juncture in determining how to deploy AI-based technologies in ways that promote rather than hinder democratic values such as freedom, equality, and transparency.”

Brave New World

"As cars will become better drivers than people, city-dwellers will own fewer cars, live further from work, and spend time differently, leading to an entirely new urban organization."

"Though quality education will always require active engagement by human teachers, AI promises to enhance education at all levels, especially by providing **personalization** at scale."

"As dramatized in the movie *Minority Report*, **predictive** policing tools raise the specter of innocent people being unjustifiably targeted. But well-deployed AI prediction tools have the potential to actually remove or reduce human bias."

Personalisation: the case of Cambridge Analytica

- *The* ML scandal of the last few years...
- Used millions of Facebook profiles to (allegedly) influence US elections, Brexit referendum, and many more political processes around the world.
- Provided user-targeted ads after classifying profiles into psychological types.
- Closed and reopened under the name *Emerdata*.

Prediction: Palantir Technologies

- Named after Lord of the Rings' *Palantír* (all-seeing eye).[1]
- Two projects: *Palantir Gotham* (for defense and counter-terrorism) and *Palantir Metropolis* (for finance).
- Billion-dollar company accumulating data from every possible source, and making predictions from that data.

[1] https://www.forbes.com/sites/andygreenberg/2013/08/14/agent-of-intelligence-how-a-deviant-philosopher-built-palantir-a-cia-funded-data-mining-juggernaut/

Predictive policing

- RAND Corporation: a think tank originally created to support US armed forces.
- RAND Report on predictive policing:[2]

> *"Predictive policing – the application of analytical techniques, particularly quantitative techniques, to identify promising targets for police intervention and prevent or solve crime – can offer several advantages to law enforcement agencies. Policing that is smarter, more effective, and more proactive is clearly preferable to simply reacting to criminal acts. Predictive methods also allow police to make better use of limited resources."*

[2] https://www.rand.org/pubs/research_briefs/RB9735.html

ML and predicting

- ML algorithms are fundamentally about *predictions*.
- What is the quality of those predictions? Do we even want to make those predictions?
- If the possible futures of an individual become part of the representation of that individual *here and now*, what does it mean for the way they are treated by institutions?

Predicting from what? The data aggregation issue.

Big data = quality

- One argument about needing big data is that it is the only way to provide quality services in applications.
- Big data is *aggregated* data.
- Let's look at a seemingly innocent example: similarity-based evaluation of semantic vectors.

Similarity-based evaluations

Human output

```
sun sunlight 50.000000
automobile car 50.000000
river water 49.000000
stair staircase 49.000000
...
green lantern 18.000000
painting work 18.000000
pigeon round 18.000000
...
muscle tulip 1.000000
bikini pizza 1.000000
bakery zebra 0.000000
```

System output

```
stair staircase 0.913251552368
sun sunlight 0.727390960465
automobile car 0.740681924959
river water 0.501849324363
...
painting work 0.448091435945
green lantern 0.383044261062
...
bakery zebra 0.061804313745
bikini pizza 0.056135605632 3
pigeon round 0.028243620524
muscle tulip 0.014257083536 7
```

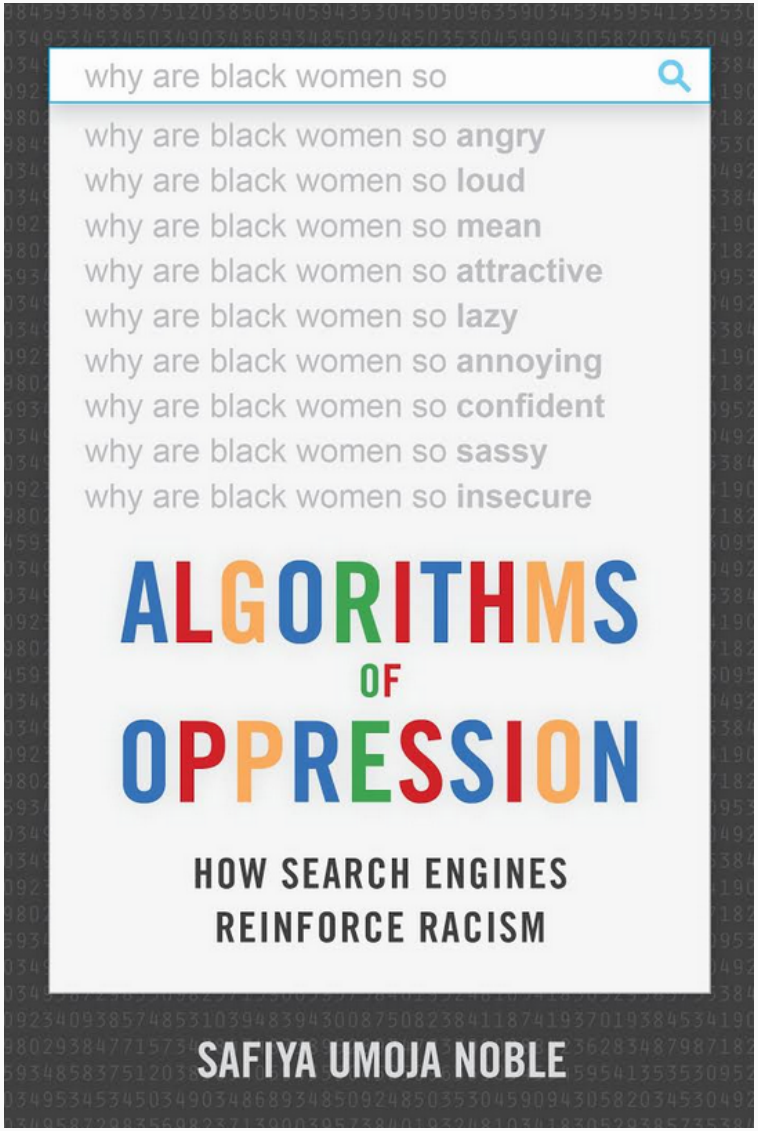
The job of the machine

- **Setup 1:** unsupervised setting. Vectors are simply gathered from corpus data. and cosine is computed over pairs of vectors. The data is an *aggregate* of what *many* people have said about a word.
- **Setup 2:** supervised setting. The system is given the corpus-extracted vectors in input, and trained on a subset of the above data, trying to replicate the *mean* human judgements.
- In both cases, reproduction of majority opinion / majority word usage. The system is never required to predict the tail of the distribution.

So we should personalise!

The need for personalisation

- Safiya Noble: the *black hair* example.
- *Black hair* can mean 1) *hair of a black colour* or 2) *hair with a texture typical to black people*.
- If the representation of *black* is biased towards the colour, results for 2) will *not* be returned.
- **NB:** this is a *compositionality* issue. More on this later!



Personalisation

- A centralised view of decentralisation: if *many* people give their private data, ML can learn how to give personalised results.
- A double-edged sword: the need for personalisation goes against the need for privacy.

From personalisation to bubbling

Bubbling

Personalisation also often goes with bubbling – it is hard to find a happy middle ground.

Serendipity

- Given the type of algorithms current at our disposal, minorities *need* 'bubbling' to find the information they need (see the black hair example).
- How to personalise *and* break the bubble at the same time? How to ensure that alternatives are seen when the user does not want to see them?
- It has been suggested to promote serendipity in tech, especially search engines.

The origins of Serendipity

"Serendipity is the art of making unsought findings" — Pek van Andel

(https://medium.com/@pieromolino/about-serendipity-454d5fa37249)

The problem with representations

Biases in cognitive science

System 1: automatic
fast, parallel, associative, lazy

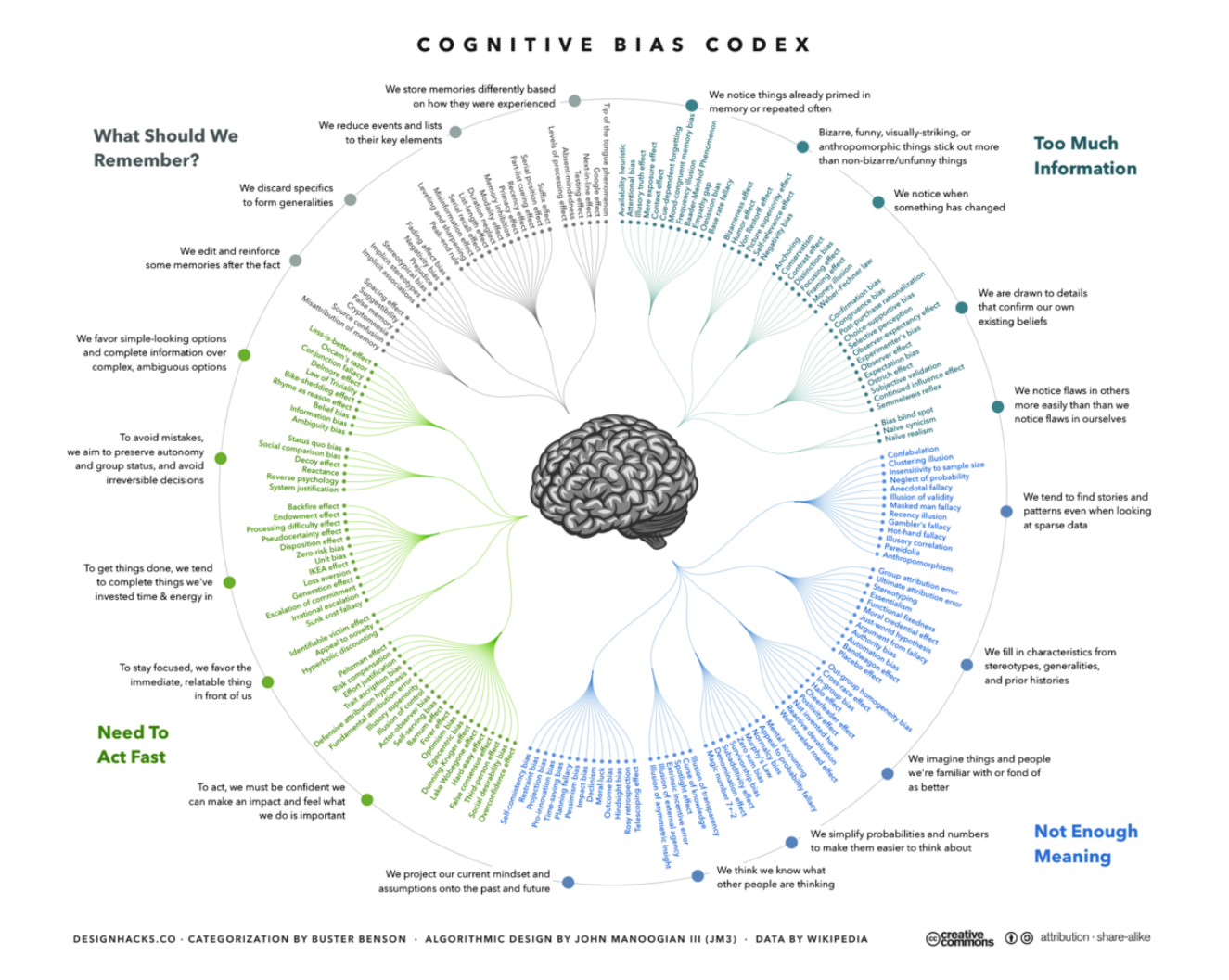
System 2: effortful
slow, serial, controlled, rule-governed, flexible

Decision-making: two systems (Kahneman & Tversky, 1973).

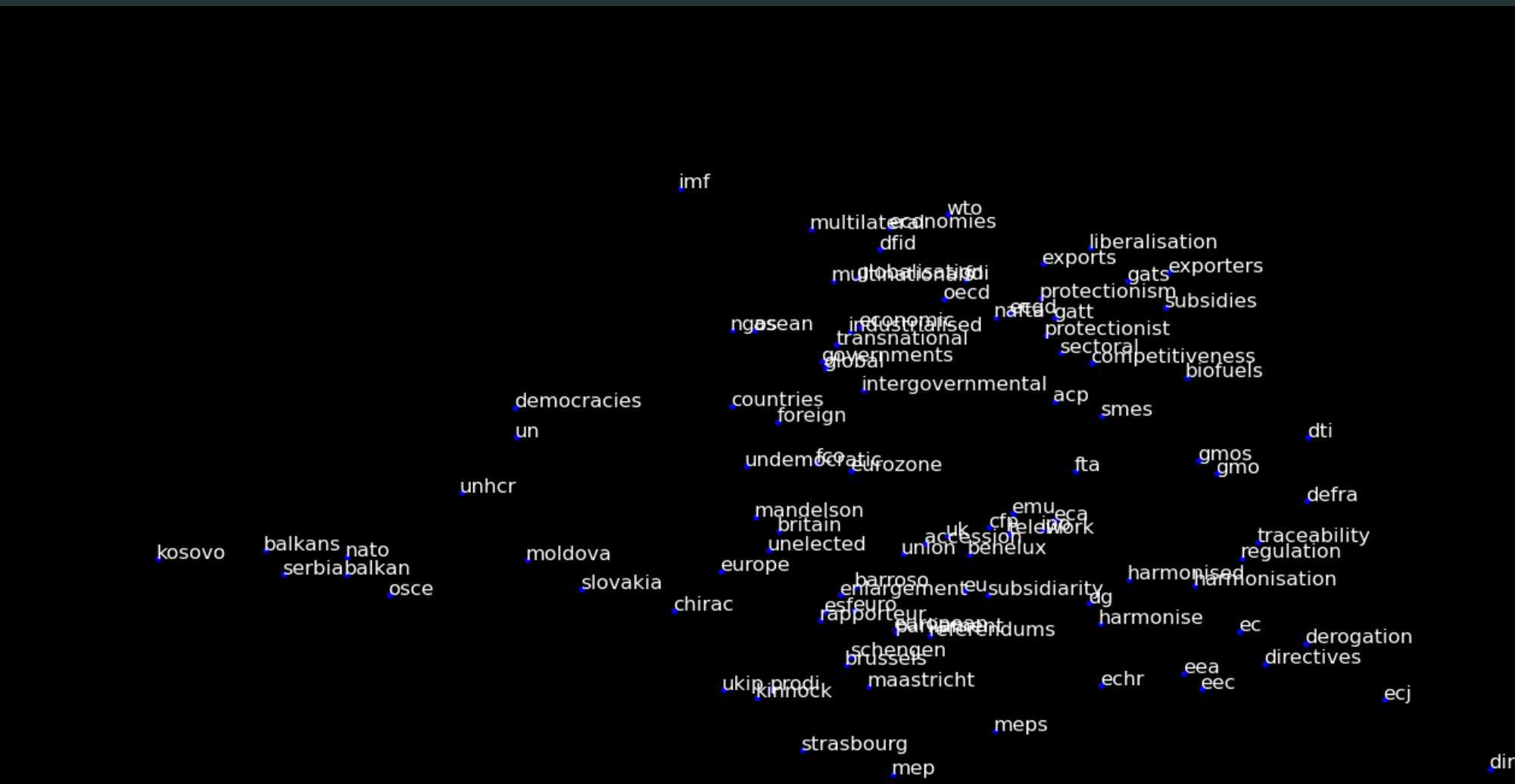
Over 95% of our cognition gets routed through System 1.
We need to consciously override System 1 through System 2 to stop ourselves from acting according to stereotypes.

Credit: Yulia Tsvetkov. https://docs.google.com/presentation/d/1499G1yyAVwRaELO9MdZFIHrACjzeiBBuMKpwdPafneI/

Biases in cognitive science



Bias in text search



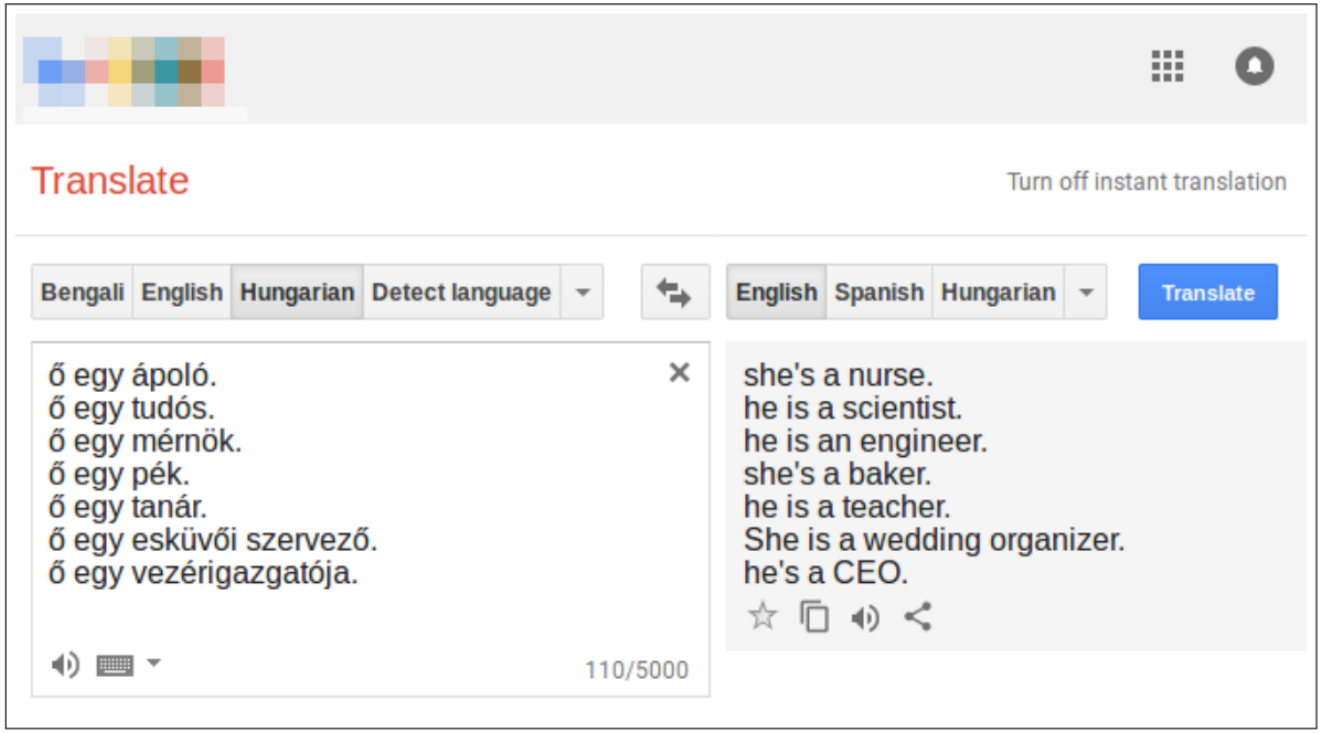
Bias in text search

- Say the vector for *EU* is very close to *unelected* and *undemocractic*.
- Say this is the vector used by the search algorithm when answering queries about the EU.
- Returned pages will necessarily be biased towards critiques of the EU.

Bias in image search

- Images are retrieved from a textual query.
- If the vectors of the query's words (or their composition) end up in a biased area of the space, the image retrieval will be biased too.

Bias in machine translation

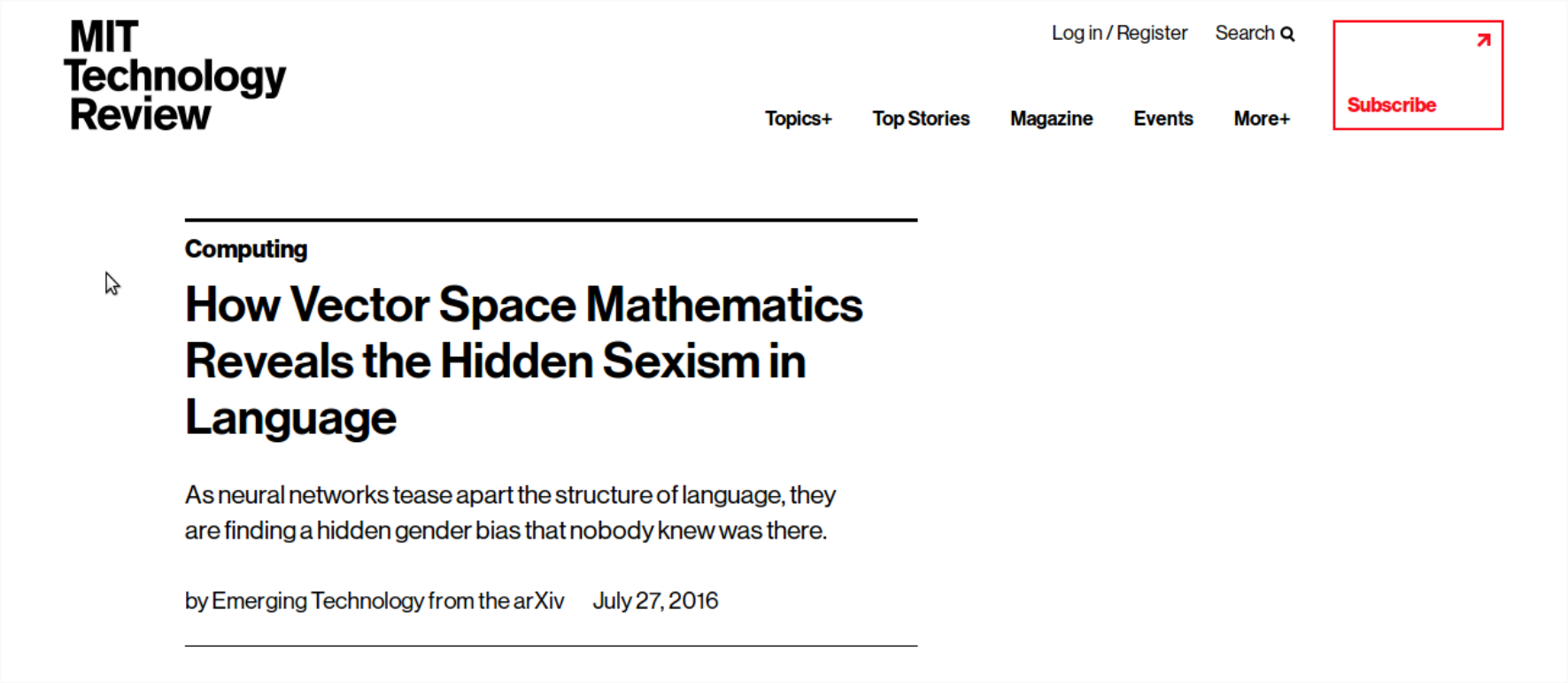


Hungarian does not have explicit marking of gender on verbs.

How will Google Translate add the corresponding pronoun?

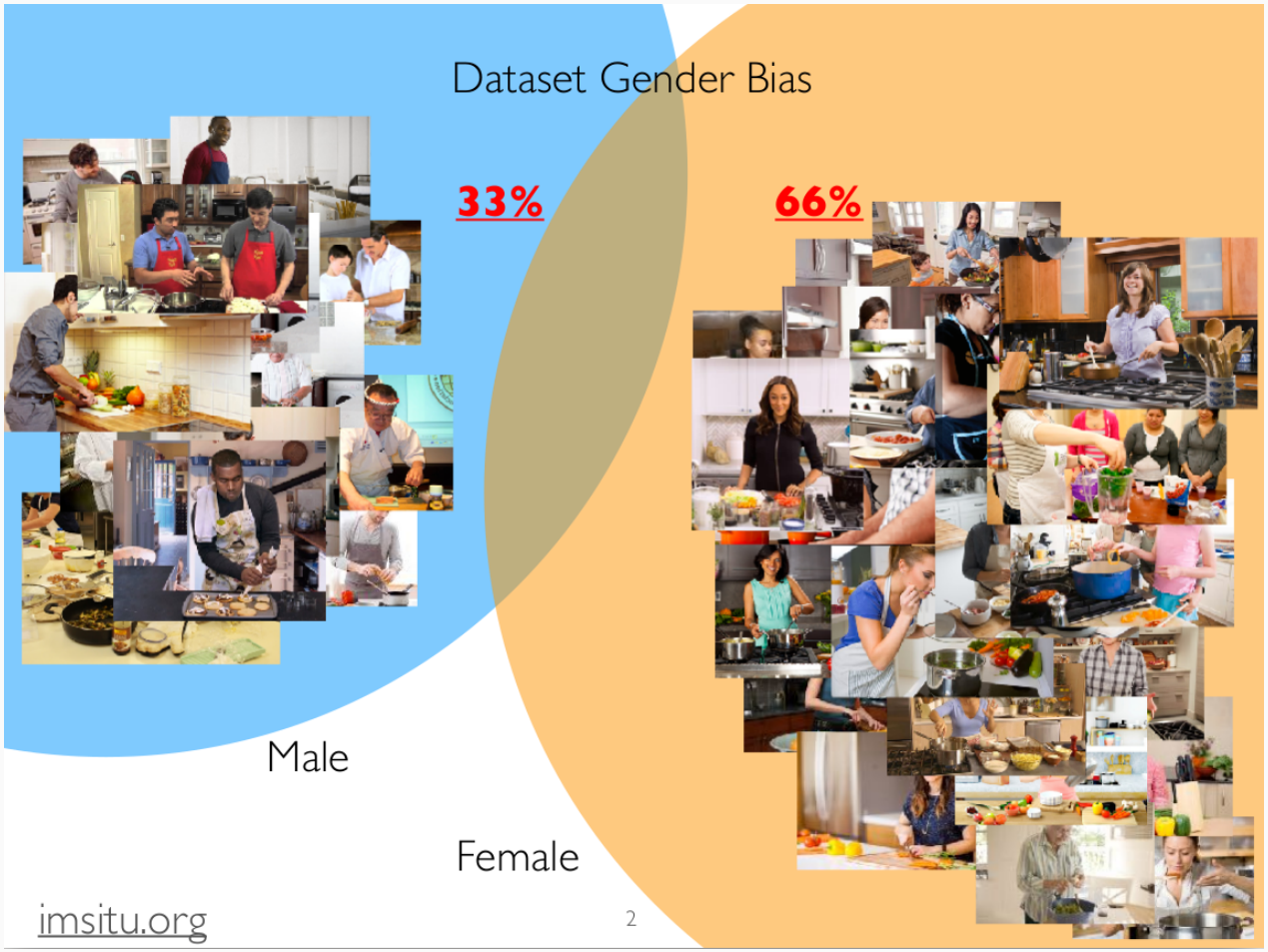
https://link.springer.com/article/10.1007/s00521-019-04144-6

The revelation...



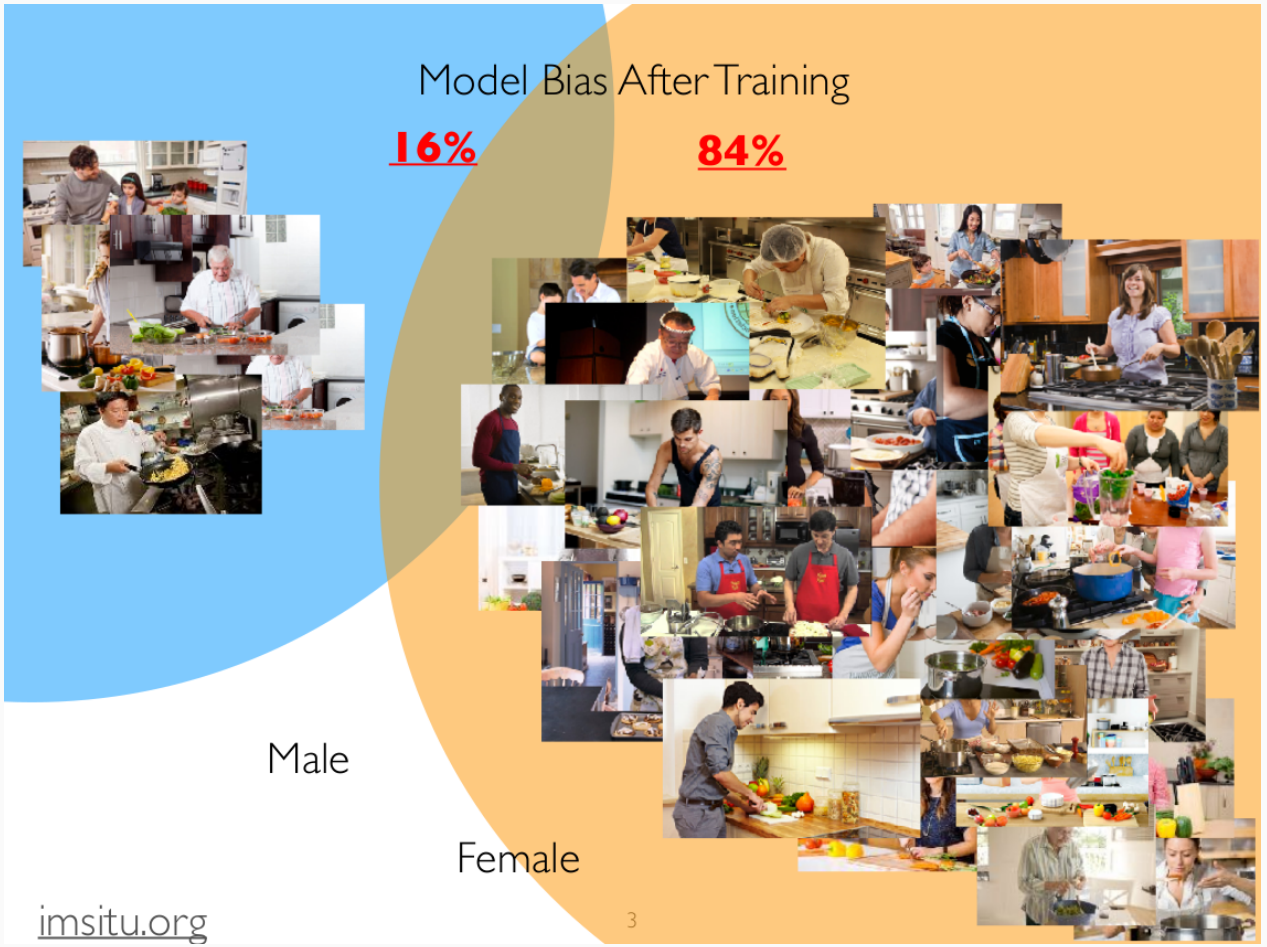
(Duh...)

Datasets are biased



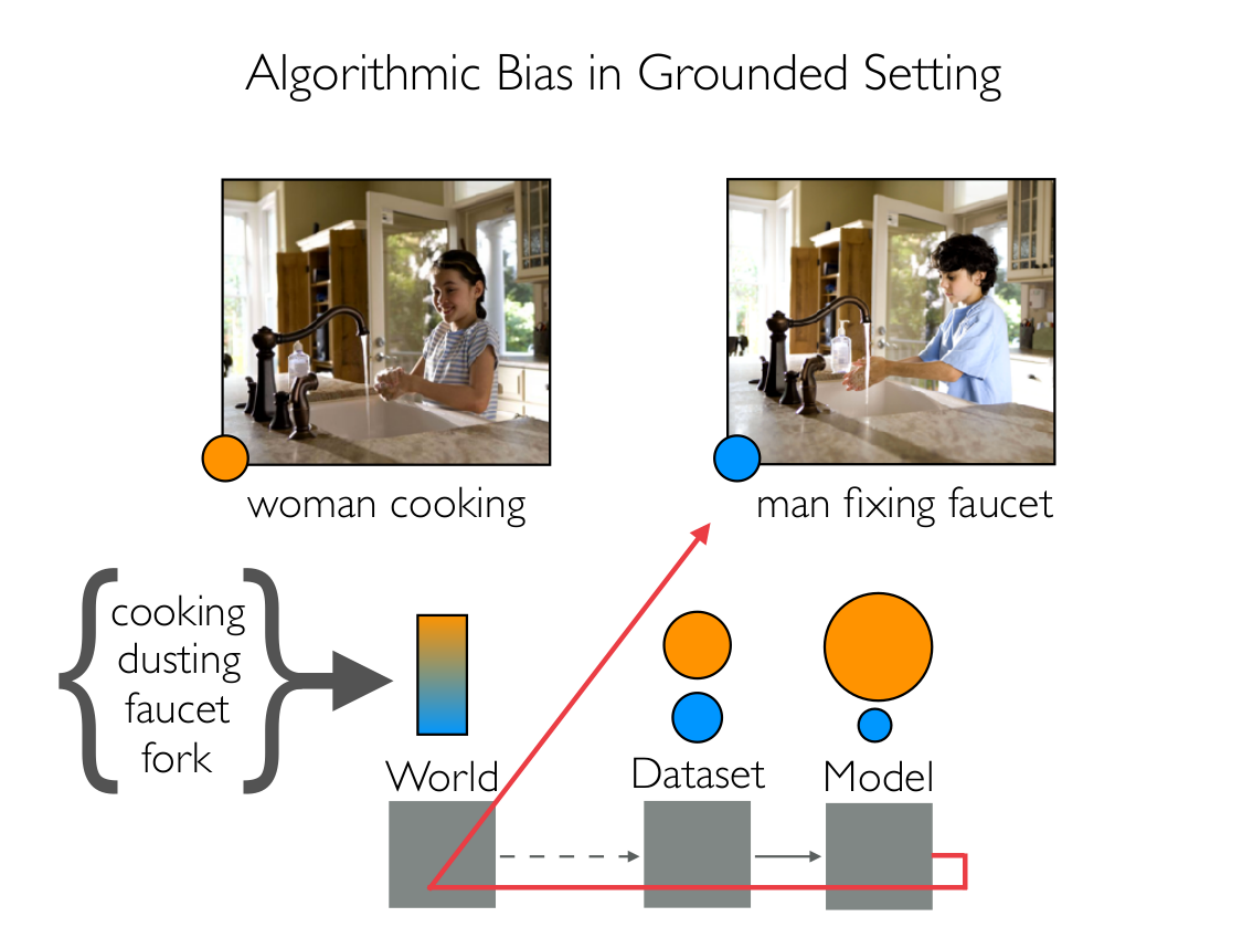
Zhao et al, 2017 - http://markyatskar.com/talks/ZWYOC17_slide.pdf

Datasets are biased



Zhao et al, 2017 - http://markyatskar.com/talks/ZWYOC17_slide.pdf

Datasets are biased

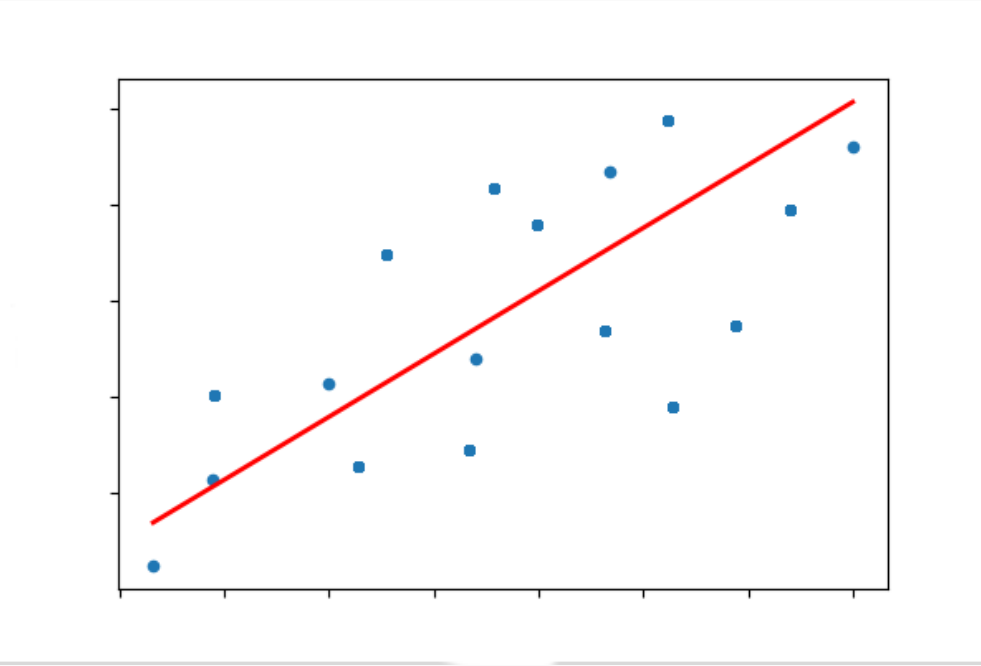


A system trained on biased data: behaviour after training.

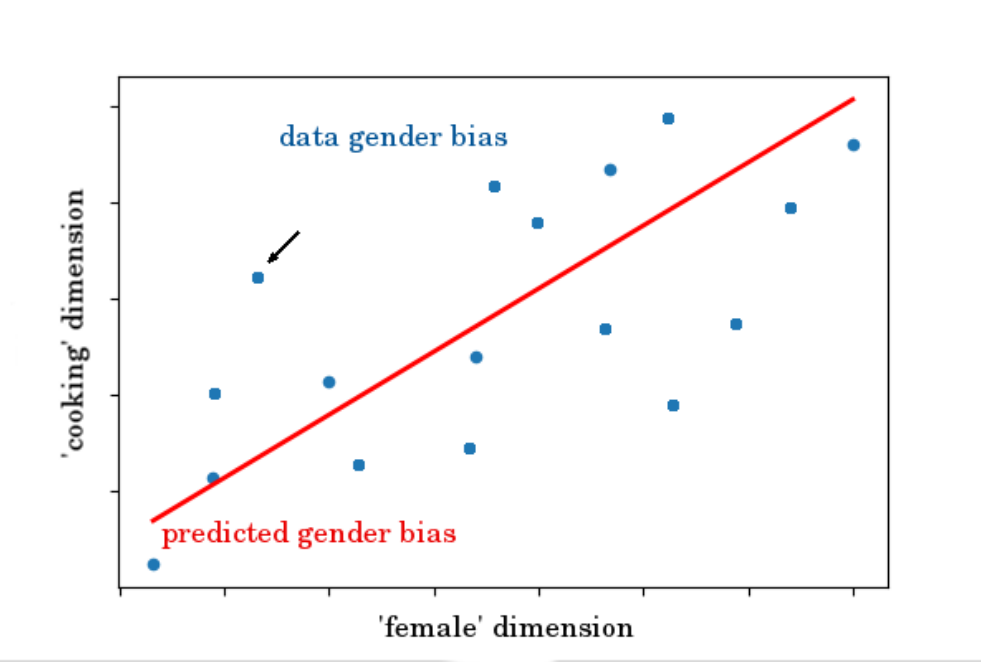
Zhao et al, 2017 - http://markyatskar.com/talks/ZWYOC17_slide.pdf

Bias amplification

- Supervised learning learns a function that *generalises* over the data.
- Imagine a standard regression line across some data. Can you see how it might accentuate problems?

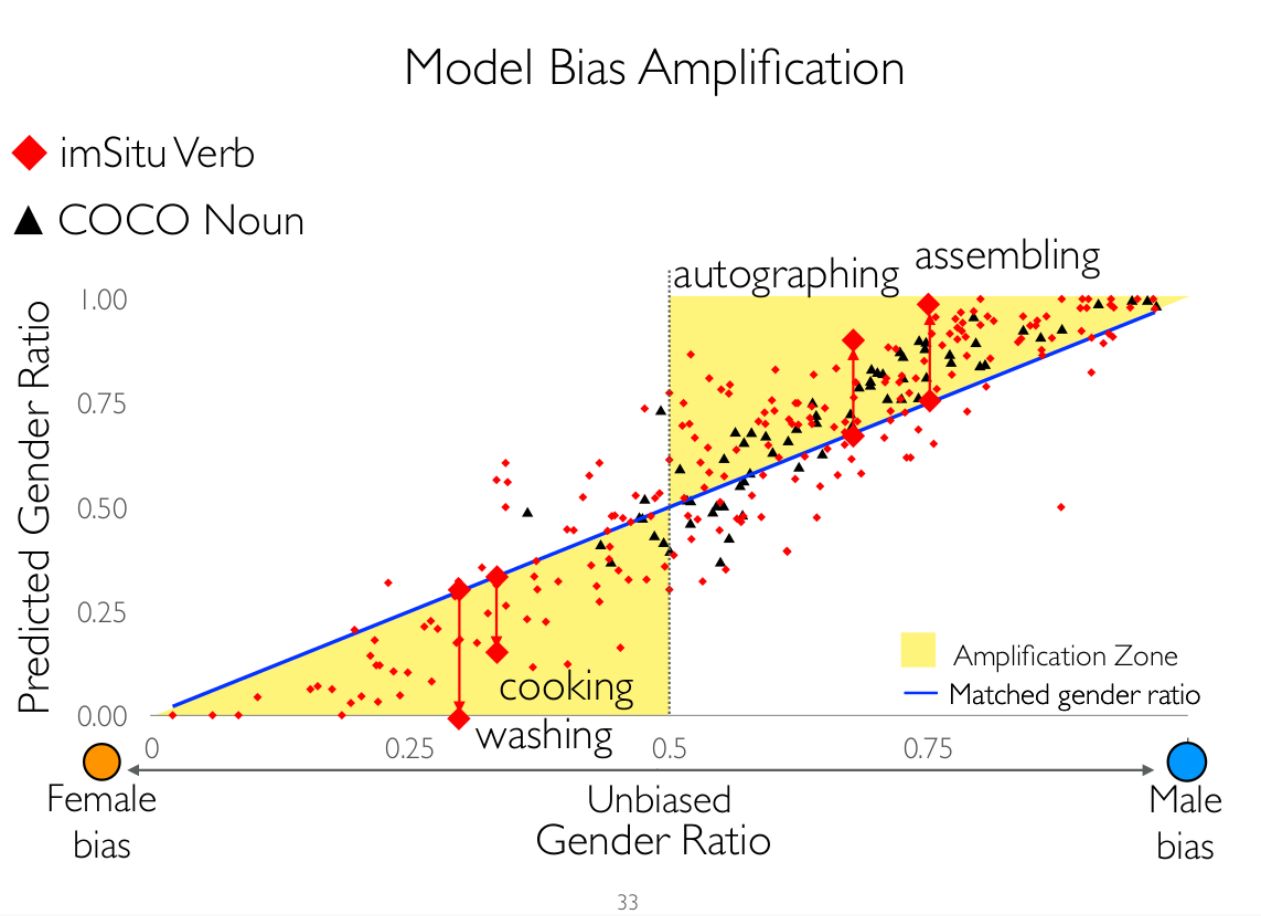


Bias amplification



The point marked by an arrow is fairly ‘non-female’ and high on the ‘cooking’ dimension, but it gets normalised by the regression line.

Bias amplification

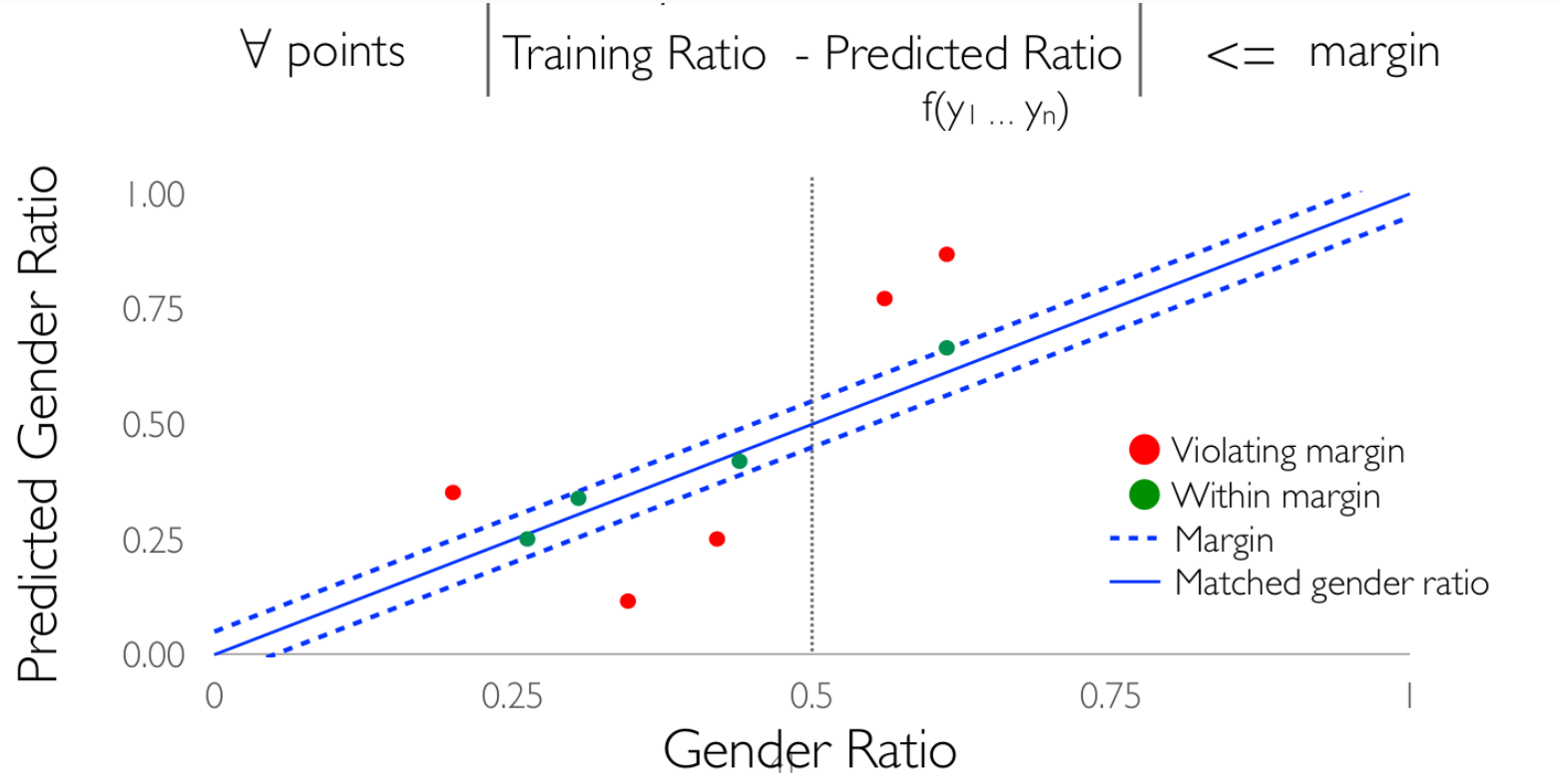


Still from Zhao et al, 2017 - http://markyatskar.com/talks/ZWYOC17_slide.pdf

Preventing bias amplification

- Can we train a system so that:
  - we prevent bias amplification;
  - we don't decrease performance (warning: we don't want to overfit!);
- NB: we are not actually *removing* bias from the original data, just making sure it does not get worse.

Preventing bias amplification



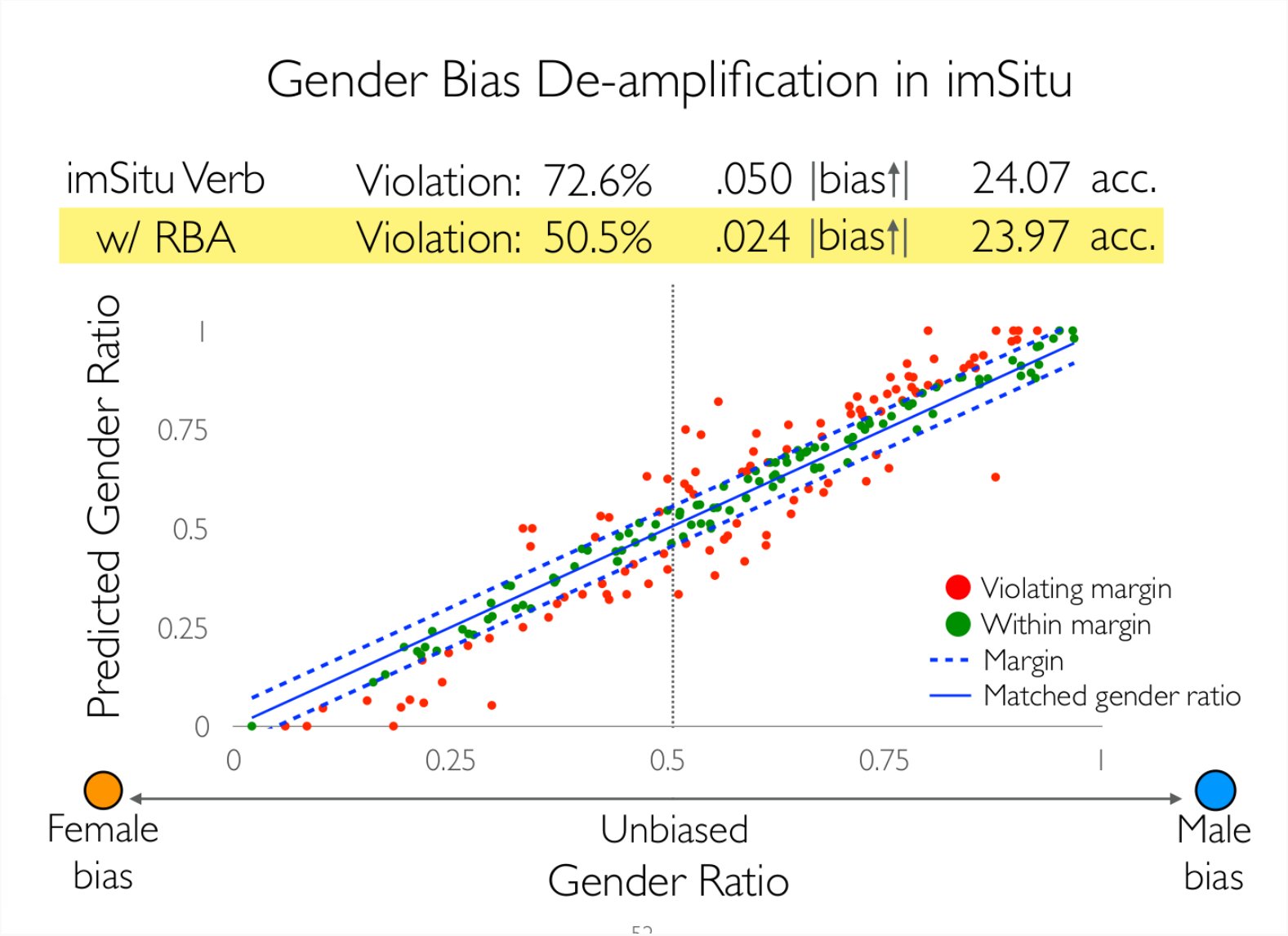
Remember SVMs?

- When implementing an SVMs, we have to tune the hyperparameter C which controls how many datapoints can violate the margin.
- Similarly, we can set a constraint on the learning problem so that

$$|\text{Training ratio} - \text{Predicted ratio}| \leq \text{margin}$$

- That is, the solution to our regression problem should not emphasise the bias present in the corpus.
- The technique is ‘safe’ from a performance point of view because the system still has to find the best possible solution to the regression problem.

Results from Zhao et al, 2017



Where are the biases?

- Research concentrates of gender / race / disability bias. Methods such as removal of bias amplification act upon the system as a whole, which is positive.
- But of course, other aspects of life can be biased. (See example of *EU* vector previously.)
- Examples: propaganda, commercially-biased texts...

Debiasing the data

- This is equivalent to ‘fixing the world’.
- Why do people talk the way they talk? Why do certain kinds of people contribute more to Web content than others? How are datasets sampled and constructed?
- Who is to say what ‘unbiased’ data should look like?

The problem with language

Language: inherently biased?

- Interestingly, the way language is structured and acquired lends itself to bias creation.
- Language evolved to satisfy particular constraints related to conceptualisation and communication.
- Today, we will look at two such constraints: productivity and efficiency.

Language: inherently biased?

- **Composition and productivity:** language makes use of the *compositionality* principle, which lets us be infinitely productive using finite means. But is it the case that $Comp(A, B) = AB$?
- **Efficiency:** certain constructions are more 'innate' than others. They make language generation and interpretation efficient, but they are not the most discriminative...

Intersectionality

Kimberlé Crenshaw (1991)

The combination (intersection) of various forms of inequality makes a qualitative difference not only to the self-perception/ identity of social actors, but also to the way they are addressed through politics, legislation and other institutions.

- Founding case: a law suit that African American women filed against the hiring policy of General Motors (DeGraffenreid v. General Motors, 1977).
- Crenshaw made the case for a reform of the US anti-discrimination-law.
- Her work was further influential in the drafting of the equality clause in the South African Constitution.
- The concept *black woman* is not the addition of *black* and *woman*.

Intersectionality in linguistic terms

- Distributional compositional semantics: the intersective composition of two elements should return a new vector.
- Let's take two old-fashioned models:
  - models that emulate the vector of the *phrase* itself, as it would be observed given a large enough corpus (Guevara, 2010 and 2011; Baroni and Zamparelli, 2010). Trained and evaluated against phrases' distributions.
  - models which only focus on the *composition operation*, independently from the phrasal distribution (Mitchell and Lapata, 2010). Task-based evaluation.
- We will call the former **phrasal models** and the latter, **intersective models**.

The meaning of phrases

Is intersection enough?

A big city: just a city which is big?

It may also be related to *loud, underground, advertisement, crowd, show, sightseeing, gentrification*...

- There is more to composition than intersection (Partee, 1994).
- There may be 'extra' (non-intersective) meaning which can be clearly observed in **phrasal vectors** and which is 'hidden' in vectors that are the result of a purely intersective operation.

Black women in the BNC

- Let's compose BNC vectors for *black* and *woman*. We'll try standard methods: pointwise multiplication and addition.
- Nearest neighbours for composed phrase:
  - **Multiplication:** black woman wear white dress skirt grey blue dark girl...
  - **Addition:** woman black white man dark wear red brown blue girl...

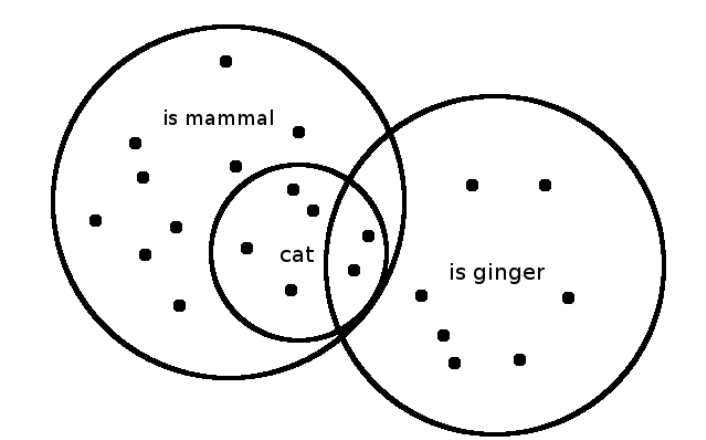
Actual occurrences of the phrase:
is she a *black woman* or what you're so ashamed of her?
the first *black woman* MP represents a poor area
during the incident , at which more police arrived , a *black woman* was assaulted an ex-drug addicted, illiterate *black woman*'s fight to retrieve her son
the *black woman* is there to act as the shock absorber
the arrest of a middle-aged *black woman*, Rosa Parks

So what should we do when we compose?

- **Phrasal vectors** are expensive to obtain. We need to store and update extra target vectors in our semantic space.
- They may well suffer from data sparsity. (Remember issue with larger n-grams in language modeling!)
- **Composed vectors** may not express the full meaning of the phrase. They include whichever biases were included in their components vectors.
- And which composition operation is the best one? (Not just from the point of view of performance!)

Moving on... Generalised quantifiers

- Quantifiers have a restrictor and a scope.
  All cats are mammals. Some cats are ginger.
- Simple interpretation: set overlap.



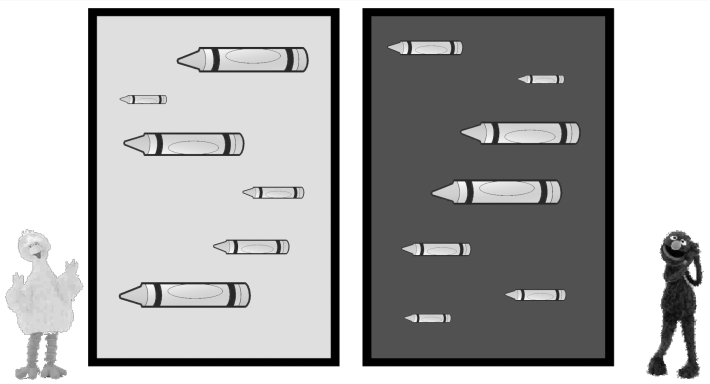
- The logic selects *individuals* over which to quantify:
  $\exists x$, $\forall x$, etc.

Beyond $\exists$ and $\forall$

- *no*: monotone decreasing.
- *most*: what is *most*? More than half? Nearly all?
- *many*: *Many cars have a GPS*, *Many dogs have three legs*.
- *the, a*: *The cat sleeps*, *The cat is a mammal*, *A cat sleeps*, *A cat is independent*, *Have you fed the fish?*
- $\emptyset$: generic bare plurals: *Cats are mammals*, *Ducks lay eggs*, *Mosquitoes carry malaria*, existential bare plurals: *Students came this morning* (Carlson, 1977).
- ...

The psychology of quantifiers

- Children acquire quantifiers *after* generics (Hollander et al 2002).
- Children acquire numerical abilities (counting) *after* the Approximate Number Sense (ANS) (Mazzocco et al 2011).



"Who has more crayons?"

- Adults make quantification 'mistakes' linked to over-generalisation:
  (All) ducks lay eggs. (Leslie et al 2011).

Quantification biases

- *Women like cooking*, *Immigrants receive money from the state* = few, some, most, all?
- Generics are efficient constructions which don't require a commitment to a quantifier and can be left 'vague'.
- Because of the over-generalisation bias, people are likely to interpret such statements as universals.
- (Machines don't even bother with quantification.)

Can machines repair language?

0.042 seussentennial
0.041 scaredy
0.035 saber-toothed
0.034 un-neutered
0.034 meow
0.034 unneutered
0.033 fanciers
0.033 pussy
0.033 pedigreed
0.032 sabre-toothed

0.032 tabby
0.032 civet
0.032 redtail
0.032 meowing
0.032 felis
0.032 whiskers
0.032 morphosys
0.031 meows
0.031 scratcher
...

1 walks
1 purrs
1 meows
1 has-eyes
1 has-a_heart
1 has-a_head
1 has-whiskers
1 has-paws
1 has-fur
1 has-claws

1 has-a_tail
1 has-4_legs
1 an-animal
1 a-mammal
1 a-feline
0.7 is-independent
0.7 eats-mice
0.7 is-carnivorous
0.3 is-domestic
...

Conclusion

Be good

- **Low quality of algorithms:** much reliance on big data, mostly implementing ‘system 1’ of decision-making, environmentally expensive.
- **Reproduction of social biases:** the machine seems to have learnt all that is bad from the data.
- **Centralisation of data:** how this relates to the type of algorithms that are used.
- **(Lack of) personalisation:** a double-edged sword.