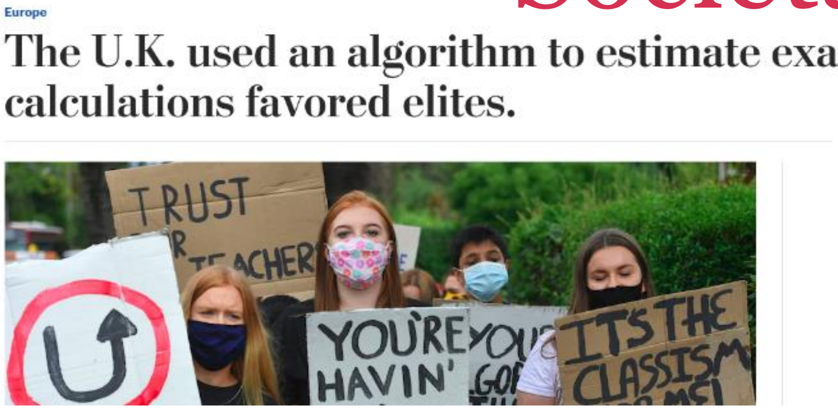
Designing responsible AI

Methods in AI research

Dong Nguyen

Utrecht University

Societal impact



Europe

The U.K. used an algorithm to estimate exam results. The calculations favored elites.

The Washington Post
Aug 18, 2020

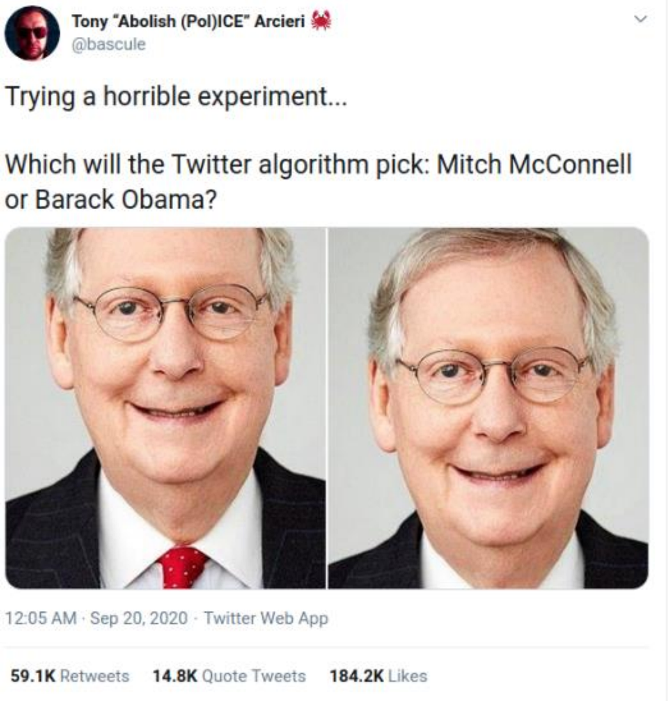
Students opposite Downing Street protesting against the downgrading of A-level results on 16 August. Photograph: Matthew Chattle/Rex/Shutterstock

Image: https://www.theguardian.com/education/2020/aug/20/england-exams-row-timeline-was-ofqual-warned-of-algorithm-bias

JUDGE POTENTIAL NOT POSTCODE
TRUST TEACHERS
OFQUAL
IT'S THE CLASSISM FOR ME!
EDUC... SHOULD UNITE NOT DIVIDE

Students protesting outside the Department for Education in London on Sunday. Henry Nicholls/Reuters

Image: https://www.nytimes.com/2020/08/17/world/europe/england-college-exam-johnson.html



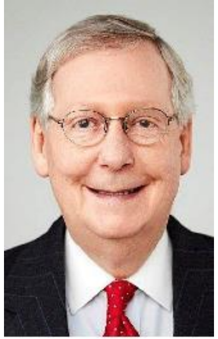
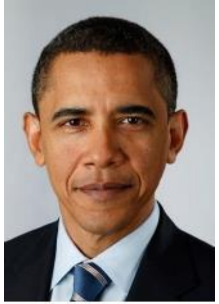
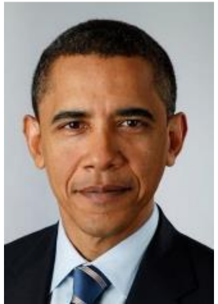
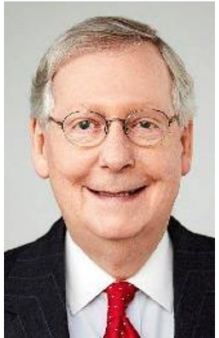
Tony “Abolish (Pol)ICE” Arcieri 🦀
@bascule

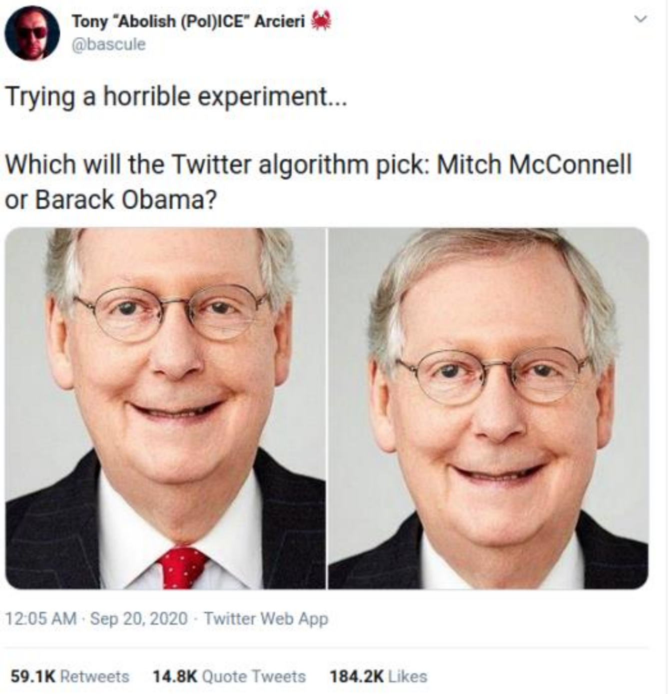
Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?

12:05 AM · Sep 20, 2020 · Twitter Web App

59.1K Retweets 14.8K Quote Tweets 184.2K Likes



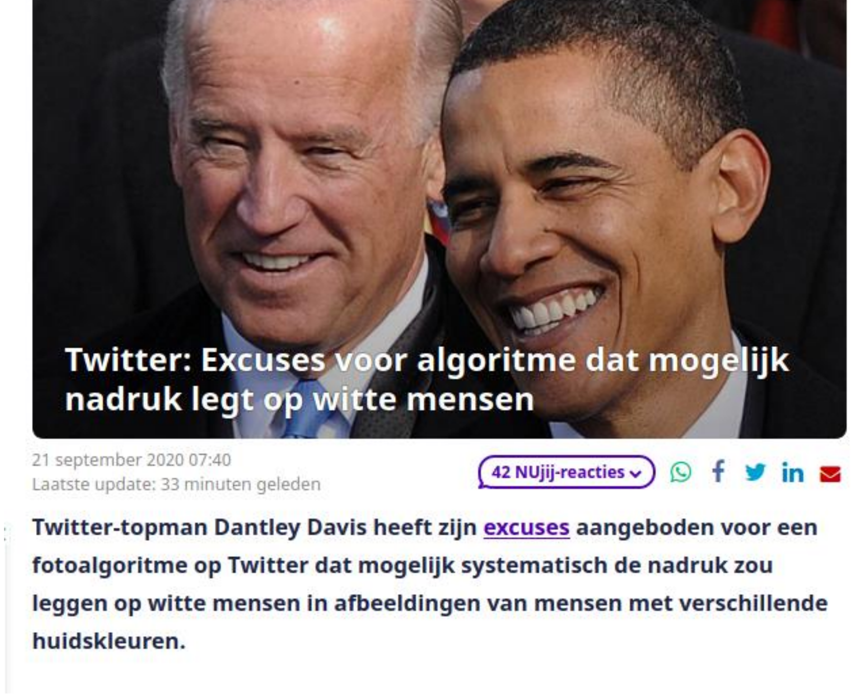
Tony "Abolish (Pol)ICE" Arcieri 🦀
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?

12:05 AM · Sep 20, 2020 · Twitter Web App

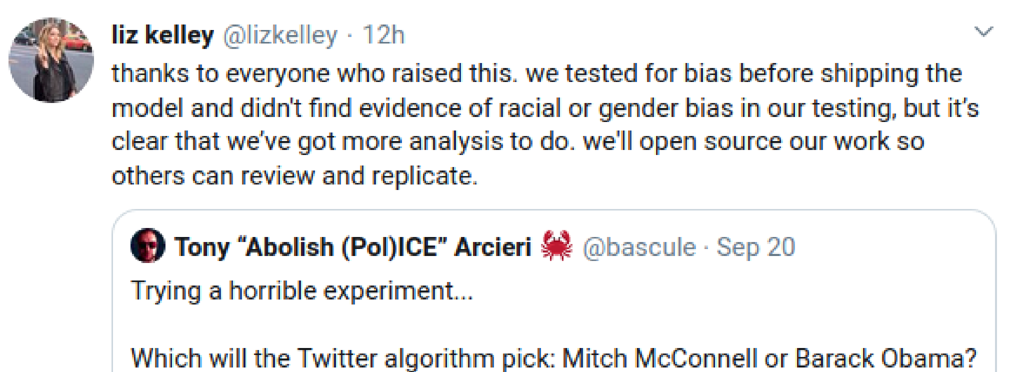
59.1K Retweets **14.8K** Quote Tweets **184.2K** Likes



Twitter: Excuses voor algoritme dat mogelijk nadruk legt op witte mensen

21 september 2020 07:40
Laatste update: 33 minuten geleden

Twitter-topman Dantley Davis heeft zijn excuses aangeboden voor een fotoalgoritme op Twitter dat mogelijk systematisch de nadruk zou leggen op witte mensen in afbeeldingen van mensen met verschillende huidskleuren.



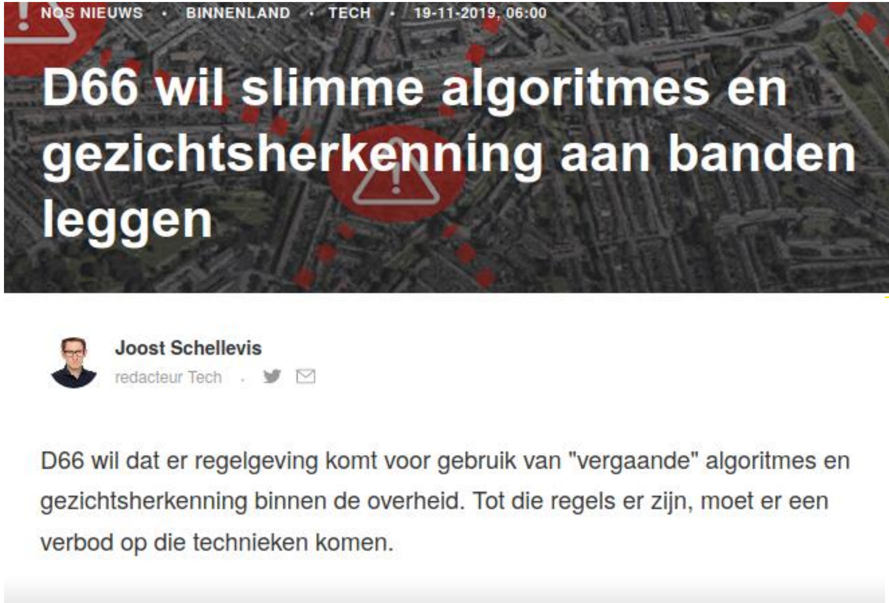
liz kelley @lizkelley · 12h

thanks to everyone who raised this. we tested for bias before shipping the model and didn't find evidence of racial or gender bias in our testing, but it's clear that we've got more analysis to do. we'll open source our work so others can review and replicate.

Tony "Abolish (Pol)ICE" Arcieri 🦀 @bascule · Sep 20

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?



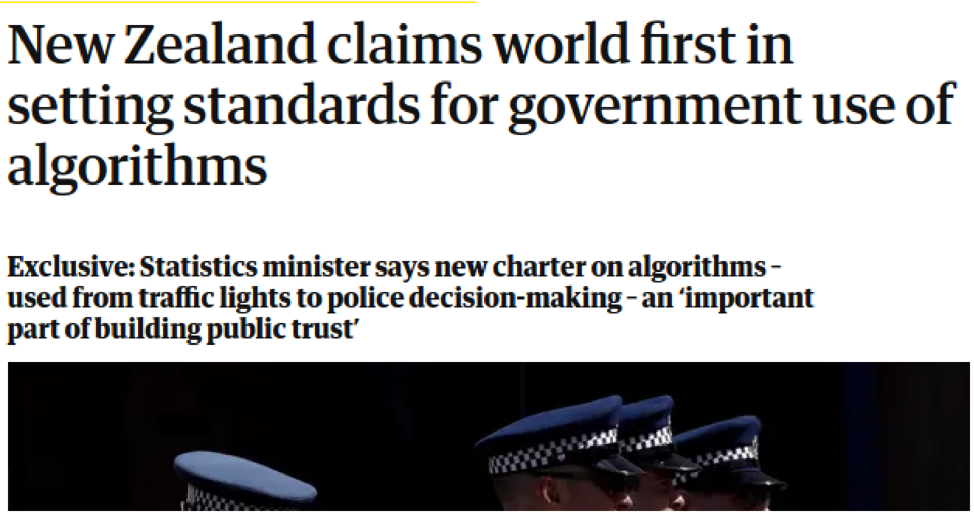
NOS NIEUWS • BINNENLAND • TECH • 19-11-2019, 06:00

D66 wil slimme algoritmes en gezichtsherkenning aan banden leggen

Joost Schellevis
redacteur Tech

D66 wil dat er regelgeving komt voor gebruik van "vergaande" algoritmes en gezichtsherkenning binnen de overheid. Tot die regels er zijn, moet er een verbod op die technieken komen.

NOS 19 Nov. 2019



New Zealand claims world first in setting standards for government use of algorithms

Exclusive: Statistics minister says new charter on algorithms - used from traffic lights to police decision-making - an 'important part of building public trust'

Guardian 28 July 2020

Dual Use

Face recognition

Faster and better airport security

AI systems might be used for both beneficial and harmful purposes

https://www.npr.org/sections/alltechconsidered/2017/06/26/534131967/facial-recognition-may-boost-airport-security-but-raises-privacy-worries?t=1563013146959

Face recognition

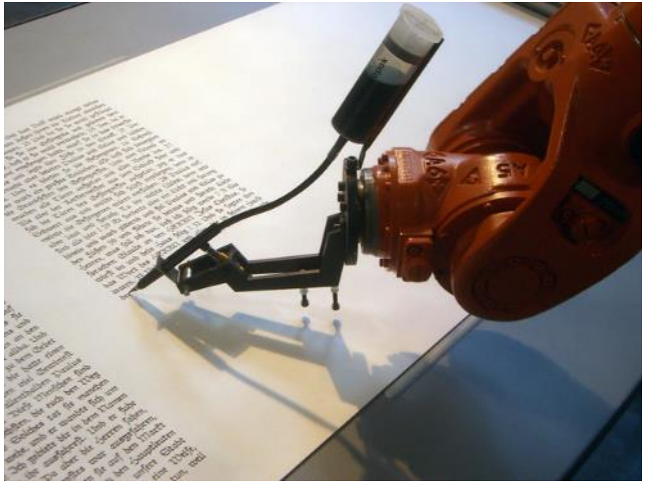
Faster and better airport security

Identifying government critics

AI systems might be used for both beneficial and harmful purposes

https://www.npr.org/sections/alltechconsidered/2017/06/26/534131967/facial-recognition-may-boost-airport-security-but-raises-privacy-worries?t=1563013146959

Text generation

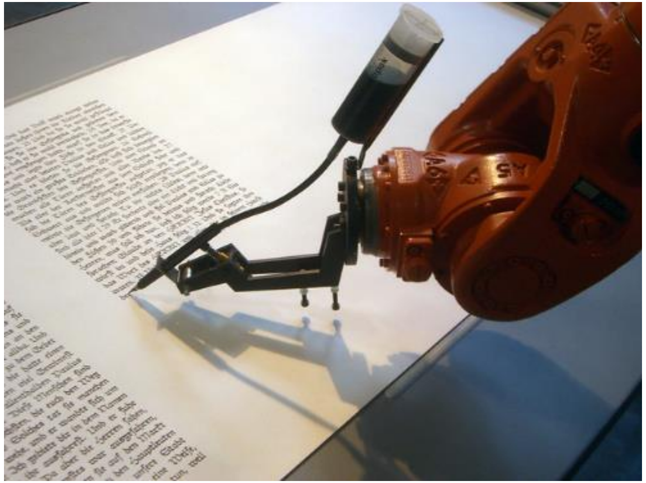


Generate novels,
poems, summaries

AI systems might be used for both beneficial and harmful purposes

https://www.flickr.com/photos/gastev/2174504149

Text generation



Generate novels,
poems, summaries

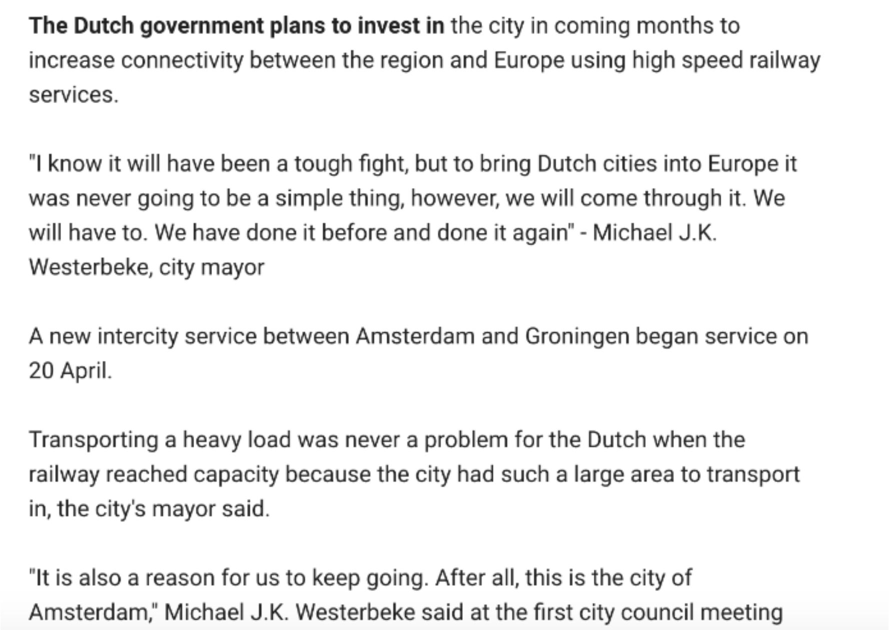
Disinformation

AI systems might be used for both beneficial and harmful purposes

https://www.flickr.com/photos/gastev/2174504149

Generated fake news

https://talktotransformer.com/



The Dutch government plans to invest in the city in coming months to increase connectivity between the region and Europe using high speed railway services.

"I know it will have been a tough fight, but to bring Dutch cities into Europe it was never going to be a simple thing, however, we will come through it. We will have to. We have done it before and done it again" - Michael J.K. Westerbeke, city mayor

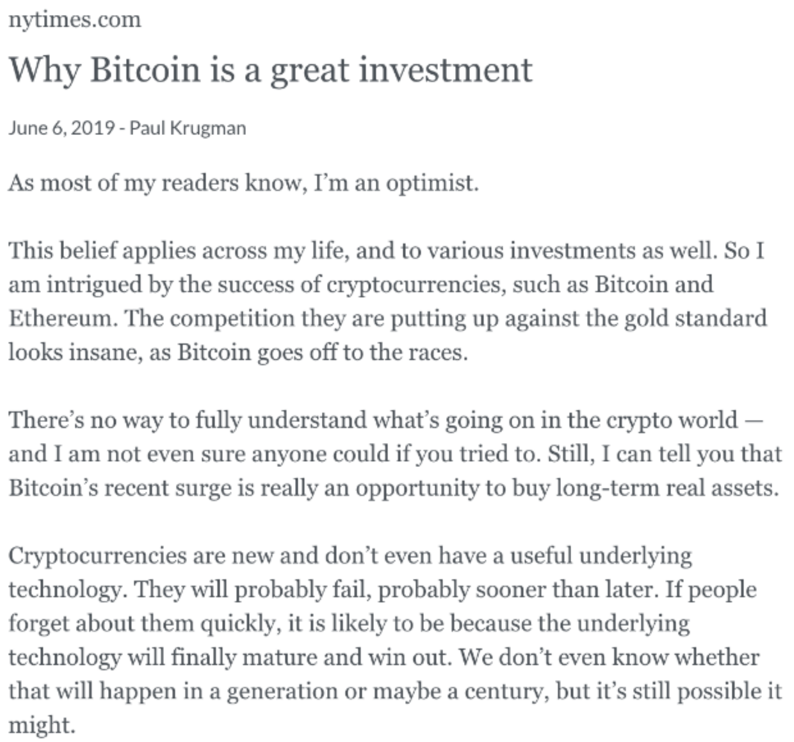
A new intercity service between Amsterdam and Groningen began service on 20 April.

Transporting a heavy load was never a problem for the Dutch when the railway reached capacity because the city had such a large area to transport in, the city's mayor said.

"It is also a reason for us to keep going. After all, this is the city of Amsterdam," Michael J.K. Westerbeke said at the first city council meeting

Quiz:
http://quiz.newsyoucantuse.com/

https://grover.allenai.org/



Fake Article

nytimes.com

Why Bitcoin is a great investment

June 6, 2019 - Paul Krugman

As most of my readers know, I'm an optimist.

This belief applies across my life, and to various investments as well. So I am intrigued by the success of cryptocurrencies, such as Bitcoin and Ethereum. The competition they are putting up against the gold standard looks insane, as Bitcoin goes off to the races.

There's no way to fully understand what's going on in the crypto world — and I am not even sure anyone could if you tried to. Still, I can tell you that Bitcoin's recent surge is really an opportunity to buy long-term real assets.

Cryptocurrencies are new and don't even have a useful underlying technology. They will probably fail, probably sooner than later. If people forget about them quickly, it is likely to be because the underlying technology will finally mature and win out. We don't even know whether that will happen in a generation or maybe a century, but it's still possible it might.

Response within the academic community

NeurIPS (machine learning conference):

- "In order to provide a balanced perspective, authors are required to include a statement of the potential broader impact of their work, including its ethical aspects and future societal consequences. Authors should take care to discuss both positive and negative outcomes."
- https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832

Ethical committees

Fairness

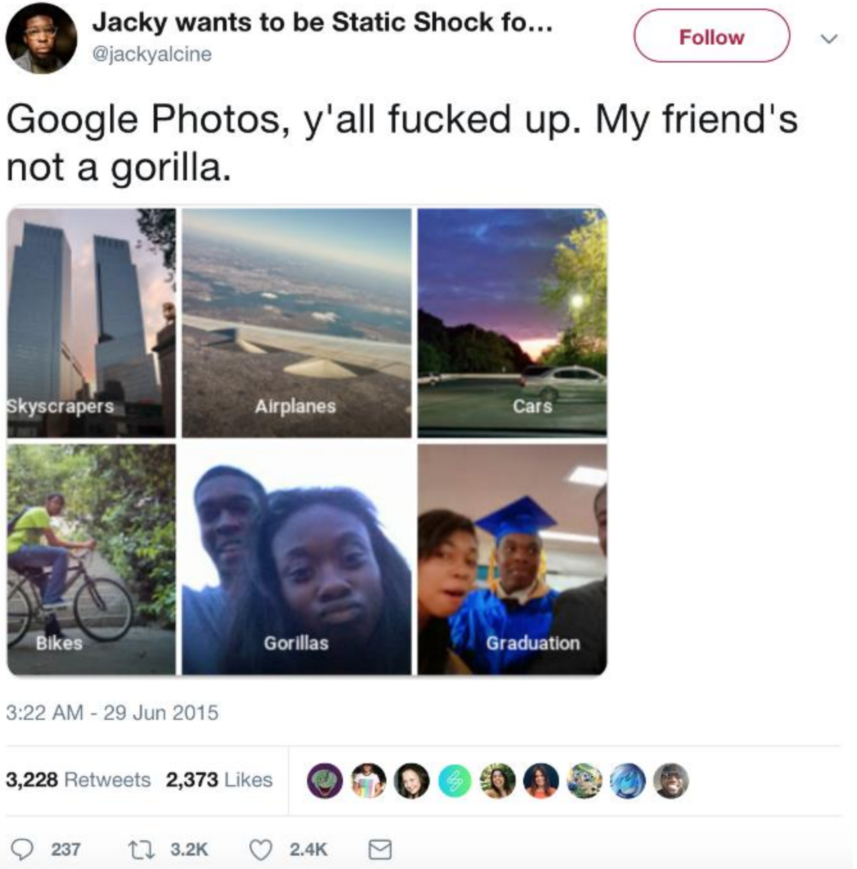
Image recognition

Cost of error!

“That context is what made this error so serious and harmful, while misidentifying someone’s toddler as a seal would just be funny.”
(Yonatan Zunger)

Blogpost of Yonatan Zunger, former Technical Lead at Google:

https://medium.com/@yonatanzunger/asking-the-right-questions-about-ai-7ed2d9820c48



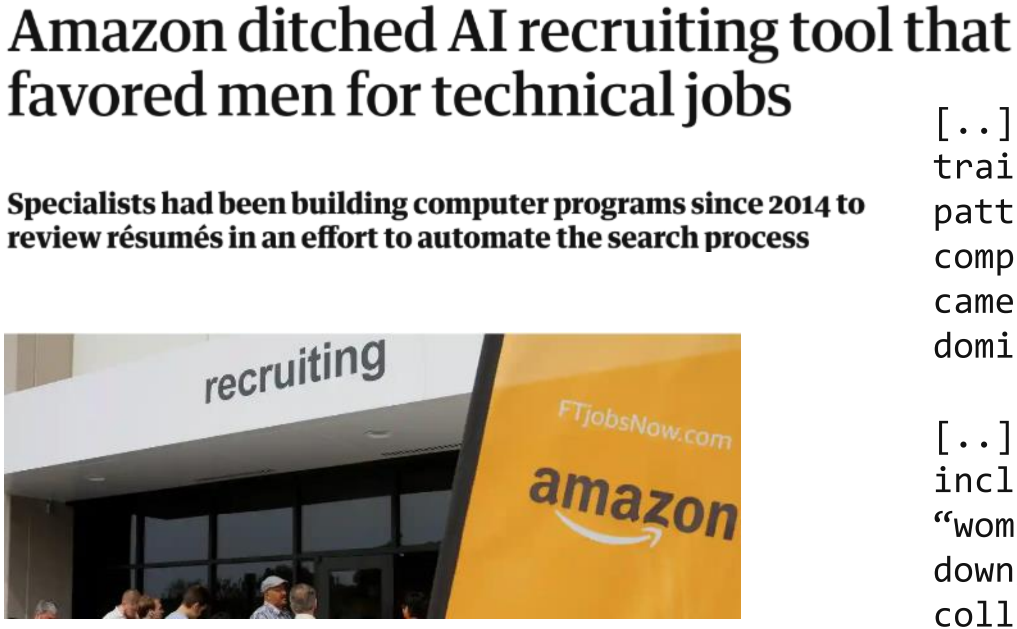
Jacky wants to be Static Shock fo...
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

3:22 AM - 29 Jun 2015

3,228 Retweets 2,373 Likes

Automatically reviewing resumes



Amazon ditched AI recruiting tool that favored men for technical jobs

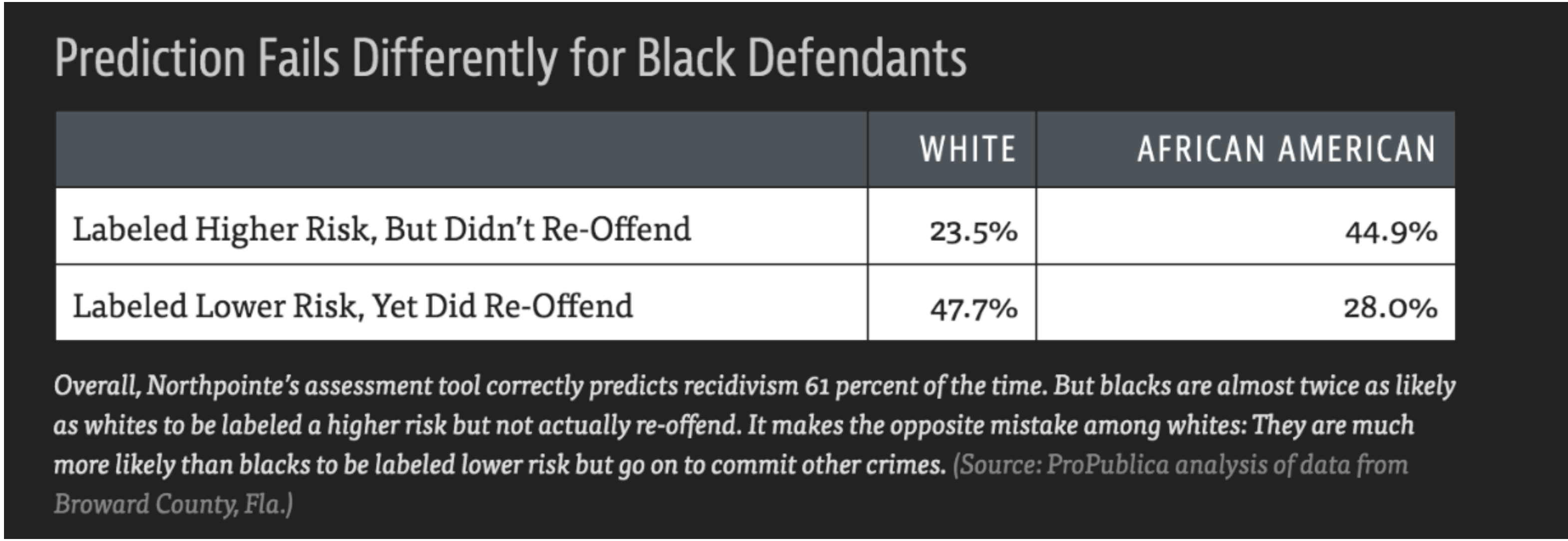
Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

[..] Amazon's computer models were trained to vet applicants by observing patterns in résumés submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

[..] . It penalized résumés that included the word "women's", as in "women's chess club captain". And it downgraded graduates of two all-women's colleges, according to people familiar with the matter.

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Risk assessment in criminal sentencing



Prediction Fails Differently for Black Defendants

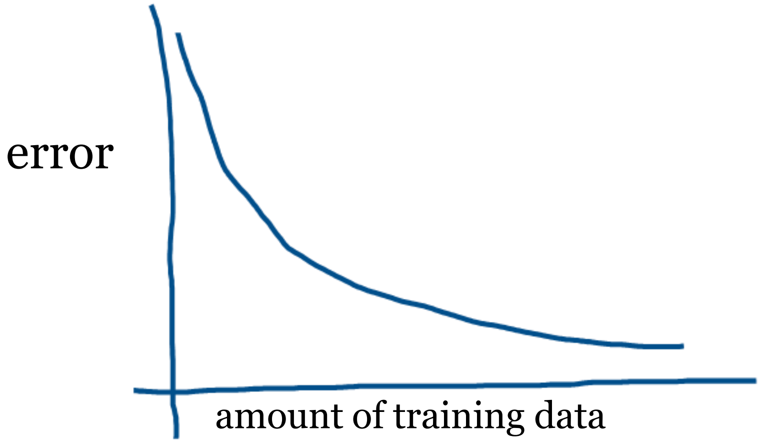
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

But see also:
https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

But my model is 'neutral'!



Less training data for minority groups!

And.. **(historical) biases** in the training data

Who is accountable? The person delivering the training data, the machine learning researcher, the policy maker?

Removing sensitive attributes

Simple solution: Ok, so let’s remove the variables we don’t want our model to select on (e.g., race, gender). Is that enough?

Removing sensitive attributes

Simple solution: Ok, so let's remove the variables we don't want our model to select on (e.g., race, gender). Is that enough?

😔

"*fairness through blindness*"
Doesn't work if the sensitive variable correlates with other variables (e.g. poverty, zip codes)

Removing sensitive attributes

Simple solution: Ok, so let’s remove the variables we don’t want our model to select on (e.g., race, gender). Is that enough?

Many (*tens*) of metrics to measure fairness

`https://fairmlbook.org/`

And lots of research on how to improve models (e.g. ACM Conference on Fairness, Accountability, and Transparency)

😔

“*fairness through blindness*”

Doesn’t work if the sensitive variable correlates with other variables (e.g. poverty, zip codes)

What is our model learning?

Clever Hans

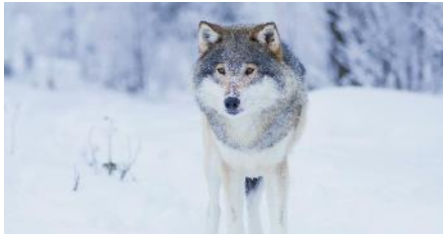
Claimed to have performed **arithmetic** and other intellectual tasks.

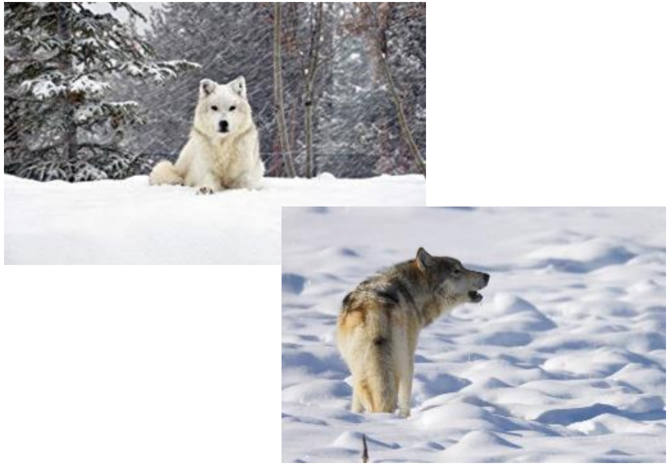
If the eighth day of the month comes on a Tuesday, what is the date of the following Friday?

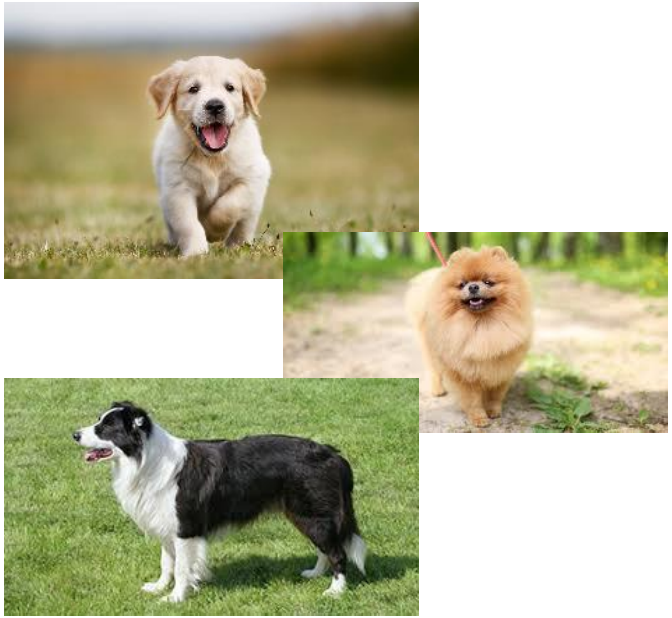
🤔

https://en.wikipedia.org/wiki/Clever_Hans

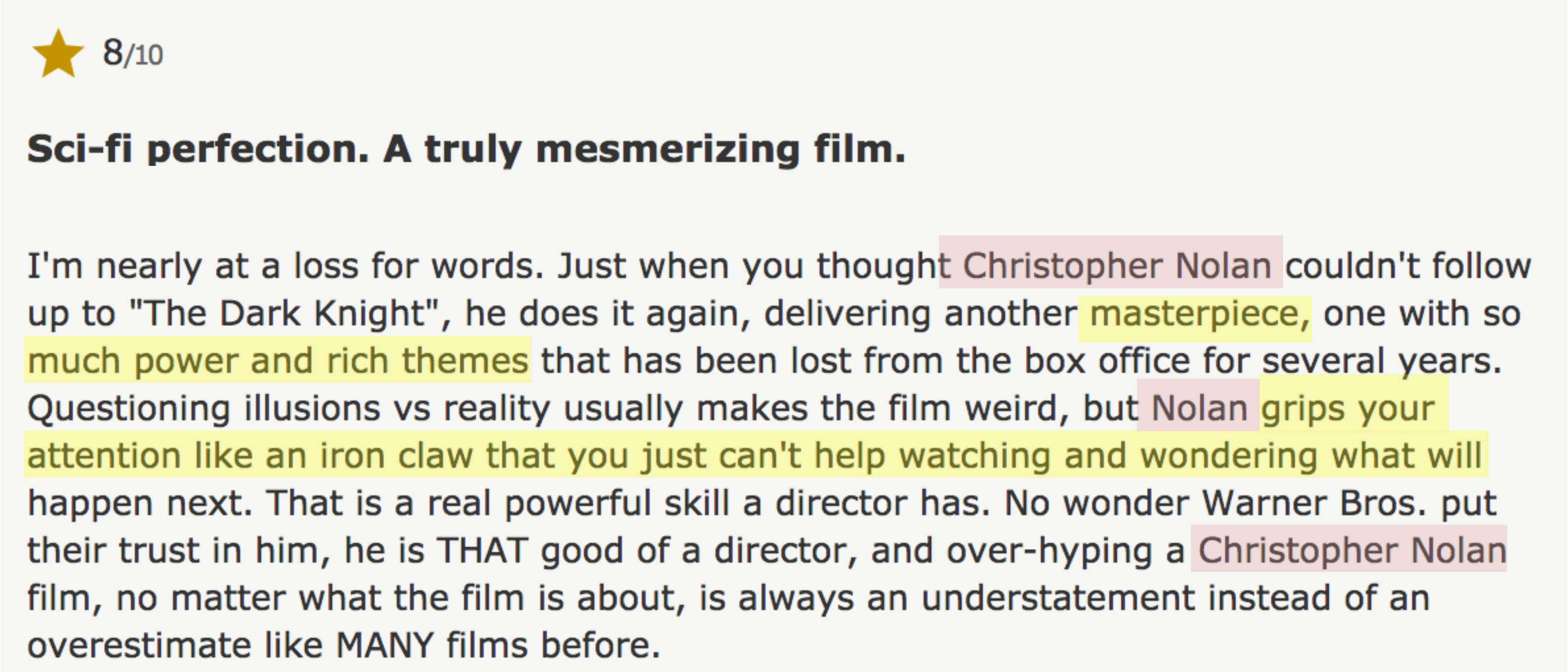
Wolf or dog?







Sentiment analysis



★ 8/10

Sci-fi perfection. A truly mesmerizing film.

I'm nearly at a loss for words. Just when you thought Christopher Nolan couldn't follow up to "The Dark Knight", he does it again, delivering another masterpiece, one with so much power and rich themes that has been lost from the box office for several years. Questioning illusions vs reality usually makes the film weird, but Nolan grips your attention like an iron claw that you just can't help watching and wondering what will happen next. That is a real powerful skill a director has. No wonder Warner Bros. put their trust in him, he is THAT good of a director, and over-hyping a Christopher Nolan film, no matter what the film is about, is always an understatement instead of an overestimate like MANY films before.

Is our model actually measuring what we think it is measuring?

Explainable AI

Why?

- Supporting decision making

- Support error analyses

- Reveal biases in the data

- Generating new insights about a phenomenon

Making the model more interpretable

- Use a simpler model (e.g., logistic regression) instead of a less interpretable model (e.g., deep neural network)

- Regularization (e.g., L1 regularization)

- Make neural networks more interpretable (active area of research!)

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
  - Explain the workings of the whole model
  - But: Sometimes the model is too complex to explain as a whole

- **Local explanation:**
  - Explain a specific prediction

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
  - Explain the workings of the whole model
  - But: Sometimes the model is too complex to explain as a whole

- **Local explanation:**
  - Explain a specific prediction

Caveat! Explanations can be misleading if the fidelity is low (e.g., doesn't match the black box model)

(see also "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" Rudin 2019)

Post-hoc explanations

When we only have access to the output of the model, we can still try to generate explanations

- **Global explanation:**
  - Explain the workings of the whole model
  - But: Sometimes the model is too complex to explain as a whole

- **Local explanation:**
  - Explain a specific prediction

Caveat! Explanations can be misleading if the fidelity is low (e.g., doesn't match the black box model)

(see also "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" Rudin 2019)

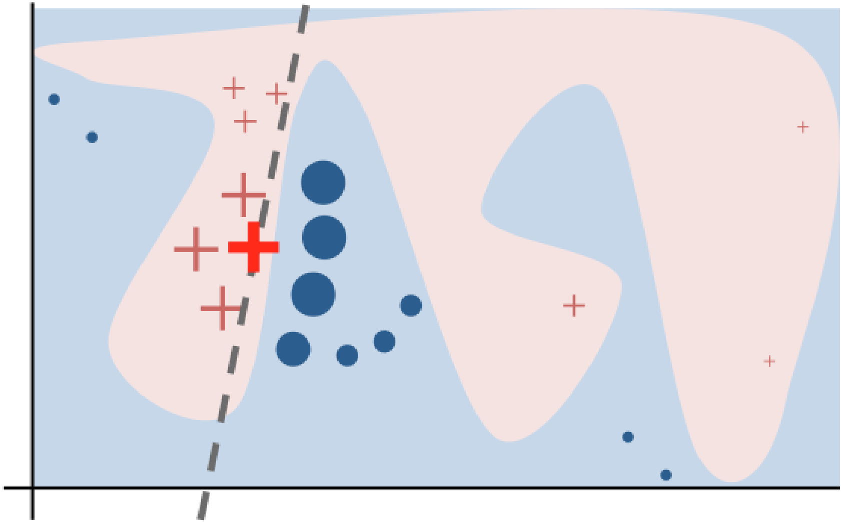
Local explanation: LIME I

Desired characteristics:

- local fidelity: the proxy must behave like the model in the neighborhood of the point of interest
- ‘interpretable’: e.g., decision trees, linear model

Steps:

- sample around the point of interest by perturbing the data
- fit an interpretable model



“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et. al 2016

https://homes.cs.washington.edu/~marcotcr/blog/lime/

Local explanation: LIME II

$$\underset{g\in G}{\arg\min}\, \underbrace{L(f,g,\pi_x)}_{} + \underbrace{\Omega\,(g)}_{}$$

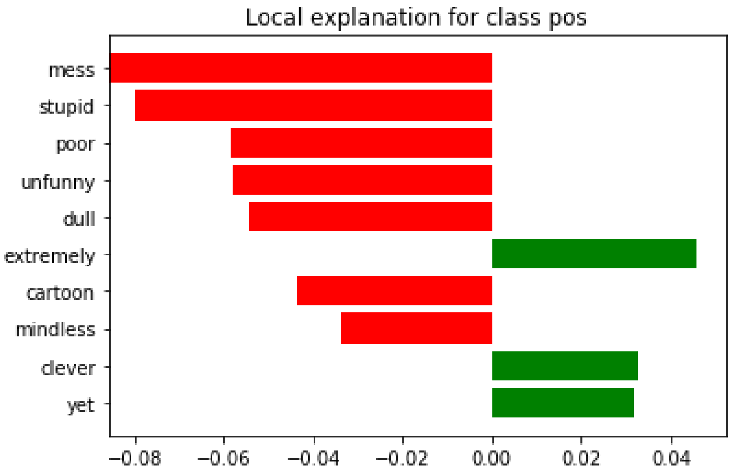
unfaithfulness g in approximating f.
π_x measures proximity

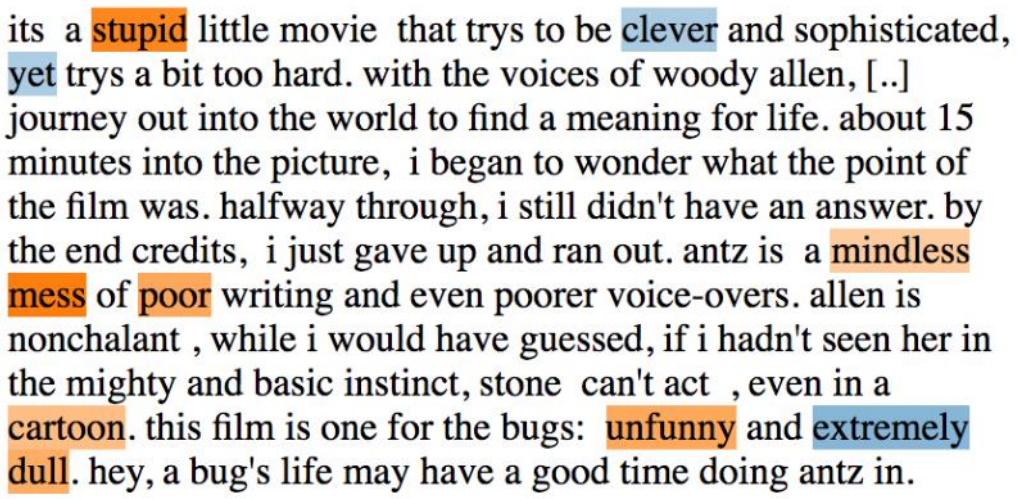
Complexity of g. E.g. depth a decision tree

g: interpretable model
f: black box model

“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et. al, KDD 2016

Local explanation: LIME III




its a stupid little movie that trys to be clever and sophisticated, yet trys a bit too hard. with the voices of woody allen, [..] journey out into the world to find a meaning for life. about 15 minutes into the picture, i began to wonder what the point of the film was. halfway through, i still didn't have an answer. by the end credits, i just gave up and ran out. antz is a mindless mess of poor writing and even poorer voice-overs. allen is nonchalant , while i would have guessed, if i hadn't seen her in the mighty and basic instinct, stone can't act , even in a cartoon. this film is one for the bugs: unfunny and extremely dull. hey, a bug's life may have a good time doing antz in.

"Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et. al, KDD 2016

Challenges

- But… interpretability is not well defined ("*The Mythos of Model Interpretability*", Lipton 2016)

- Many challenges in evaluation, "what is a good explanation?"

Human biases in NLP models learned from data

RECAP!

Word embeddings

Dense real-valued vectors

cat 0.52 0.84 0.01 .... 0.23

dog 0.40 0.90 0.10 .... 0.40

Words are mapped onto
a vector space

Word embeddings are the standard way to represent words in modern NLP systems!

RECAP!

Word embeddings

Dense real-valued vectors

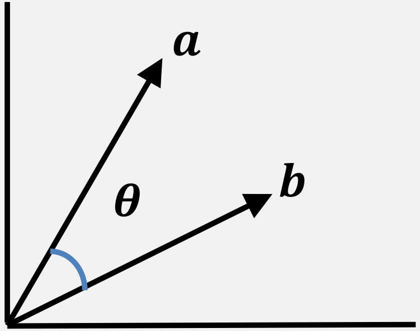
cat	0.52	0.84	0.01	....	0.23
dog	0.40	0.90	0.10	....	0.40

Words are mapped onto
a vector space

Word embeddings are the standard way to represent words in modern NLP systems!

Cosine similarity

$$\frac{\boldsymbol{a} \bullet \boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} = \frac{\sum \boldsymbol{a}_i \, \boldsymbol{b}_i}{\sqrt{\sum \boldsymbol{a}_i^2}\,\sqrt{\sum \boldsymbol{b}_i^2}}$$



RECAP!

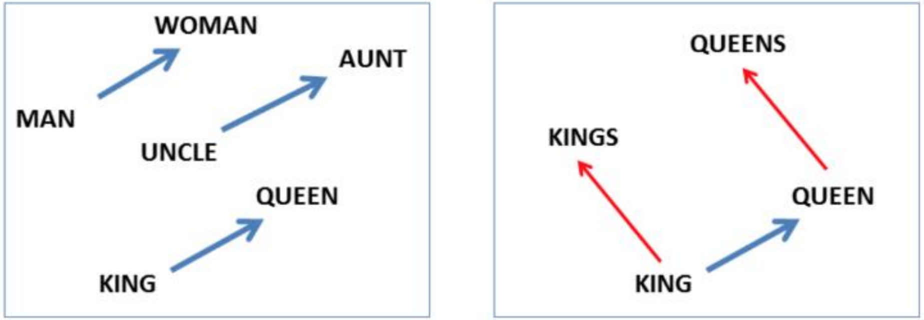
Word embeddings

Dense real-valued vectors

cat	0.52	0.84	0.01	....	0.23
dog	0.40	0.90	0.10	....	0.40

Words are mapped onto
a vector space

Word embeddings are the standard way to represent words in modern NLP systems!



king - man + woman ≈ queen

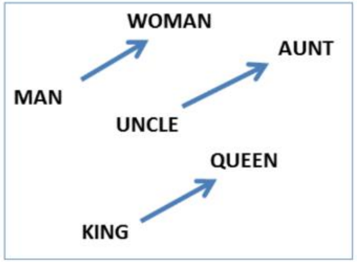
Linguistic Regularities in Continuous Space
Word Representations, Mikolov et al. 2013

Finder gender stereotype analogies

$$S_{(a,b)}(x,y) = \cos(a - b, x - y) \quad \text{if} \|x - y\|_2 \ \leq \ \delta, \quad 0 \text{ else}$$

(a,b)=(*she*,*he*)

embedding_{she} embedding_{he} L2 distance



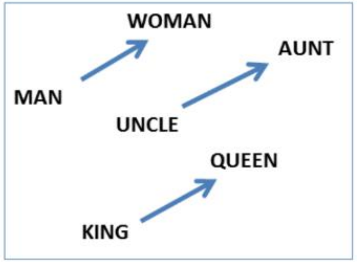
Gender appropriate *she-he* analogies
queen-king
sister-brother
ovarian cancer-prostate cancer
mother-father
convent-monastery

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi, et al. NIPS 2016)

Finder gender stereotype analogies

$$S_{(a,b)}(x,y) = \cos(\underset{\text{embedding}_{she}}{a} - \underset{\text{embedding}_{he}}{b}, x - y) \quad \text{if} \|x - y\|_{\underset{\text{L2 distance}}{2}} \ \le \ \delta, \quad 0 \text{ else}$$

(a,b)=(*she*,*he*)



Gender appropriate *she-he* analogies
queen-king
sister-brother
ovarian cancer-prostate cancer
mother-father
convent-monastery

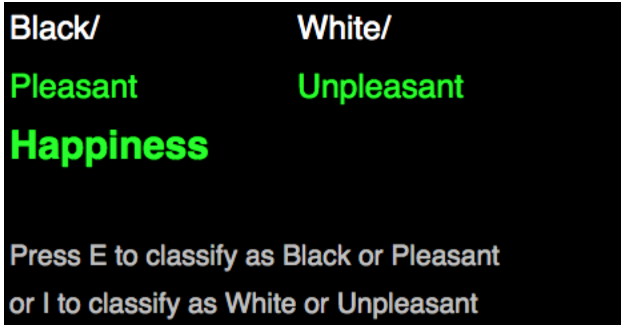
Gender stereotype *she-he* analogies
nurse-surgeon
sassy-snappy
cupcakes-pizzas
lovely-brilliant
vocalist-guitarist

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings (Bolukbasi, et al. NIPS 2016)

Detecting bias:
Word-Embedding Association Test

- The Implicit Association Test (IAT) is based on response times and has been widely used.

- Word-Embedding Association Test (WEAT) by Caliskan et al: use the cosine similarity between pairs of vectors as analogous to reaction time in the IAT



Black/ White/
Pleasant Unpleasant

Happiness

Press E to classify as Black or Pleasant
or I to classify as White or Unpleasant

https://en.wikipedia.org/wiki/Implicit-association_test

Semantics derived automatically from language corpora contain human-like biases, Caliskan, Bryson, Narayanan, Science 2017

Detecting bias: Word-Embedding Association Test

- The Implicit Association Test (IAT) is based on response times and has been widely used.
- Word-Embedding Association Test (WEAT) by Caliskan et al: use the cosine similarity between pairs of vectors as analogous to reaction time in the IAT

Were able to replicate well-known IAT findings!

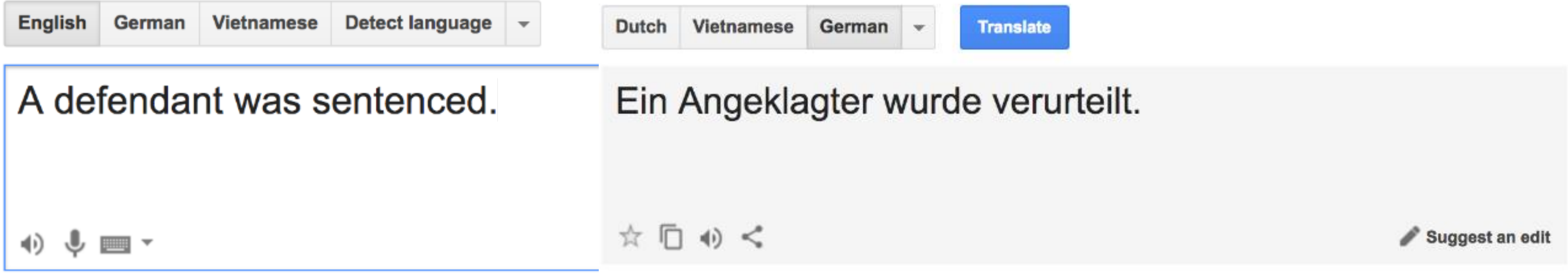
Semantics derived automatically from language corpora contain human-like biases, Caliskan, Bryson, Narayanan, Science 2017

Sentiment analysis

"I had tried building an algorithm for sentiment analysis based on word embeddings [..]When I applied it to restaurant reviews, I found it was ranking Mexican restaurants lower. The reason was not reflected in the star ratings or actual text of the reviews. It's not that people don't like Mexican food. ***The reason was that the system had learned the word "Mexican" from reading the Web."***

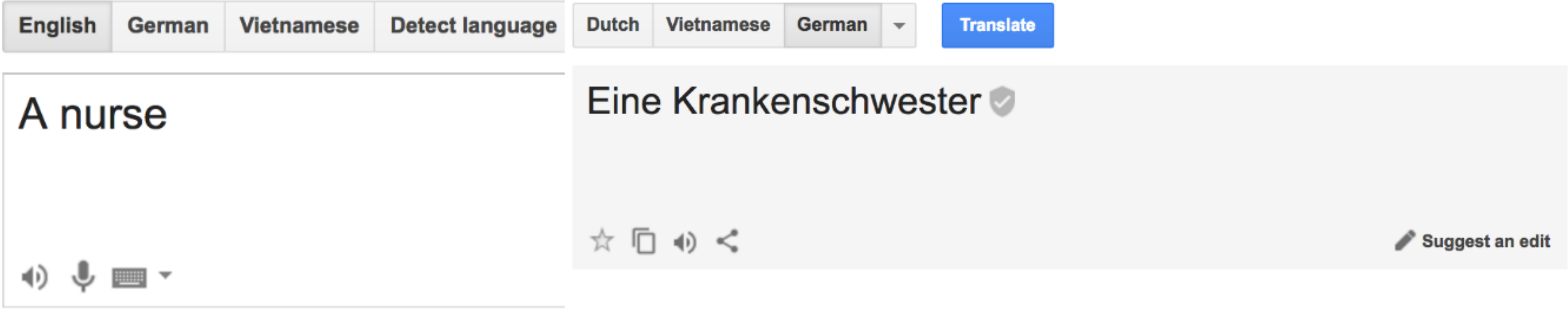
https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

Machine Translation



A defendant was sentenced.

Ein Angeklagter wurde verurteilt.



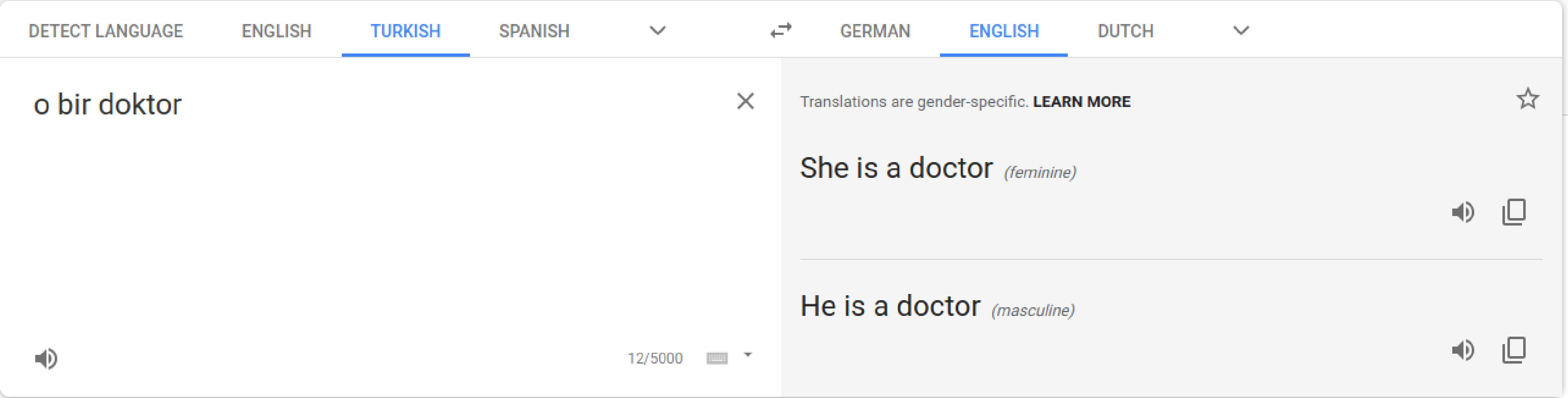
A nurse

Eine Krankenschwester

Translating from English to German.

https://genderedinnovations.stanford.edu/case-studies/nlp.html

Machine Translation



o bir doktor

Translations are gender-specific.

She is a doctor *(feminine)*

He is a doctor *(masculine)*

https://blog.google/products/translate/reducing-gender-bias-google-translate/

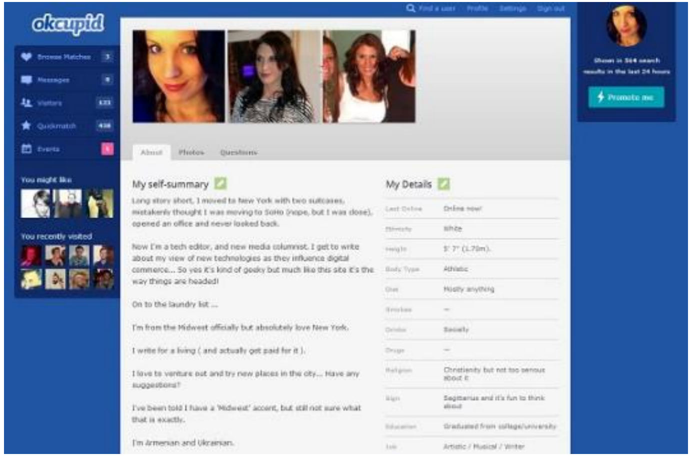
Privacy

*The editorial policy followed in citing CMC data in this volume makes a distinction between restricted- and open-access electronic fora, the former of which are considered **private**, while the latter are **public**. (Herring, 1996)*

*we are confronted with media texts that combine private and public aspects on various levels. They may be **public** in the sense that they are within the public space and can be read by a large and anonymous audience, while at the same time discussing topics which we think of as '**private**' and using language which is associated with informal and private conversations. (Landert and Jucker, 2011)*

Question: Using public Twitter or Instagram data for research: ok or not?
- And what about companies mining your data?

Is this OK?



My self-summary

Long story short, I moved to New York with two suitcases, mistakenly thought I was moving to SoHo (nope, but I was close), opened an office and never looked back.

Now I'm a tech editor, and new media columnist. I get to write about my view of new technologies as they influence digital commerce... So yes it's kind of geeky but much like this site it's the way things are headed!

On to the laundry list ...

I'm from the Midwest officially but absolutely love New York.

I write for a living (and actually get paid for it).

I love to venture out and try new places in the city... Have any suggestions?

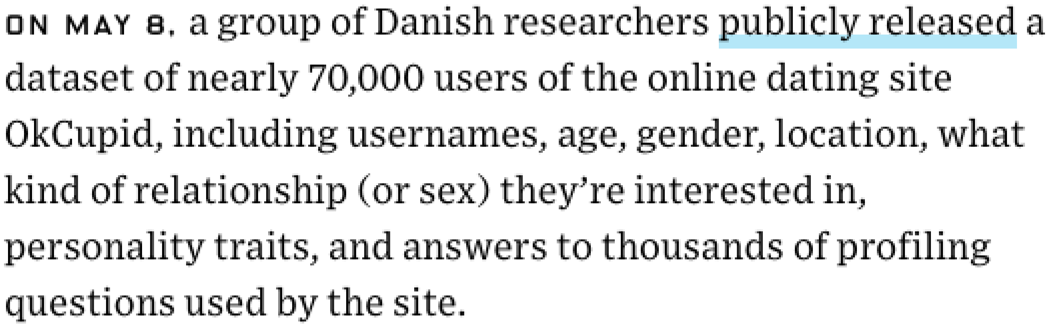
I've been told I have a 'Midwest' accent, but still not sure what that is exactly.

I'm Armenian and Ukrainian.

My Details

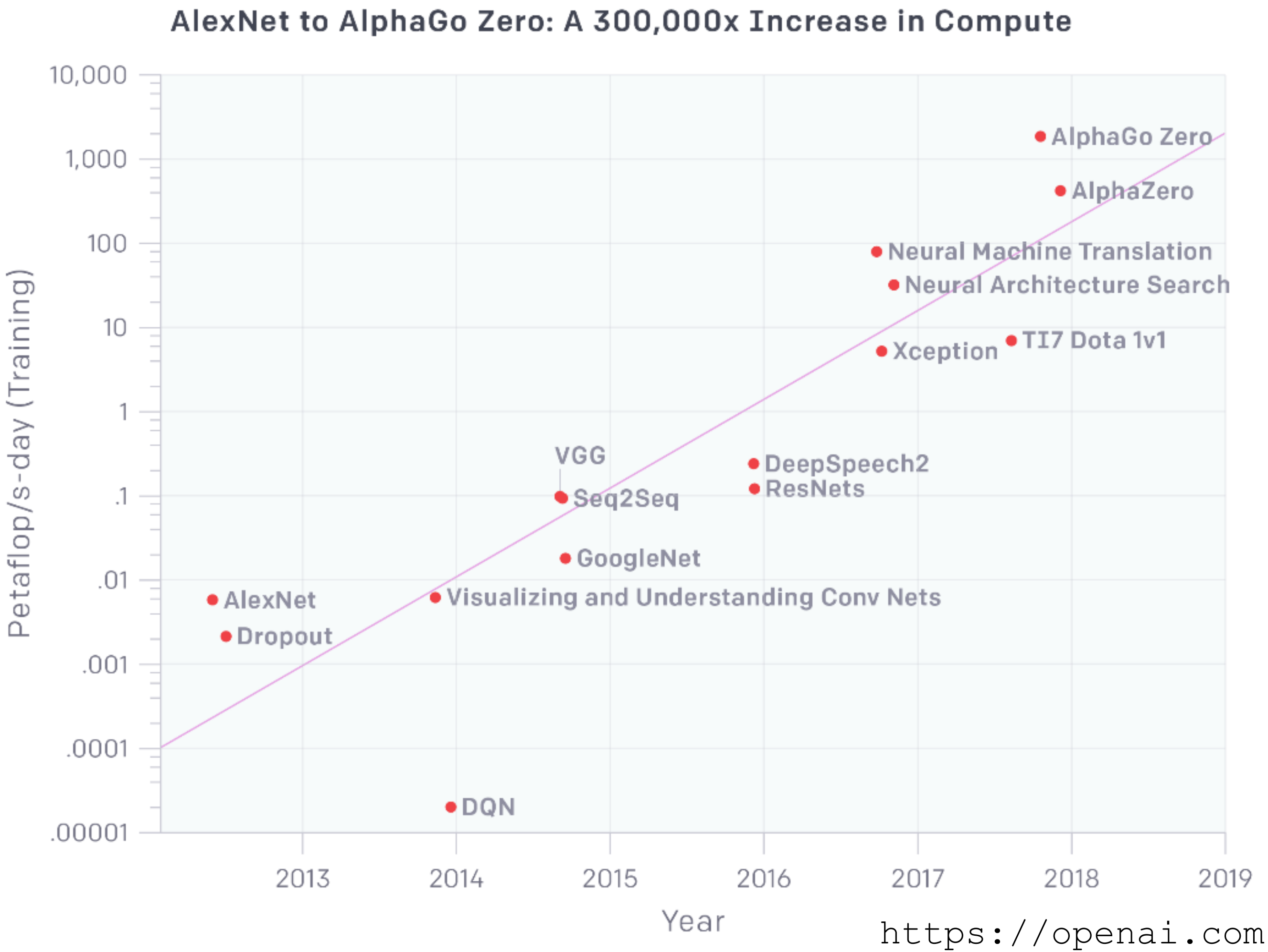
Last Online	Online now!
Ethnicity	White
Height	5' 7" (1.70m).
Body Type	Athletic
Diet	Mostly anything
Smokes	—
Drinks	Socially
Drugs	—
Religion	Christianity but not too serious about it
Sign	Sagittarius and it's fun to think about
Education	Graduated from college/university
Job	Artistic / Musical / Writer

https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/

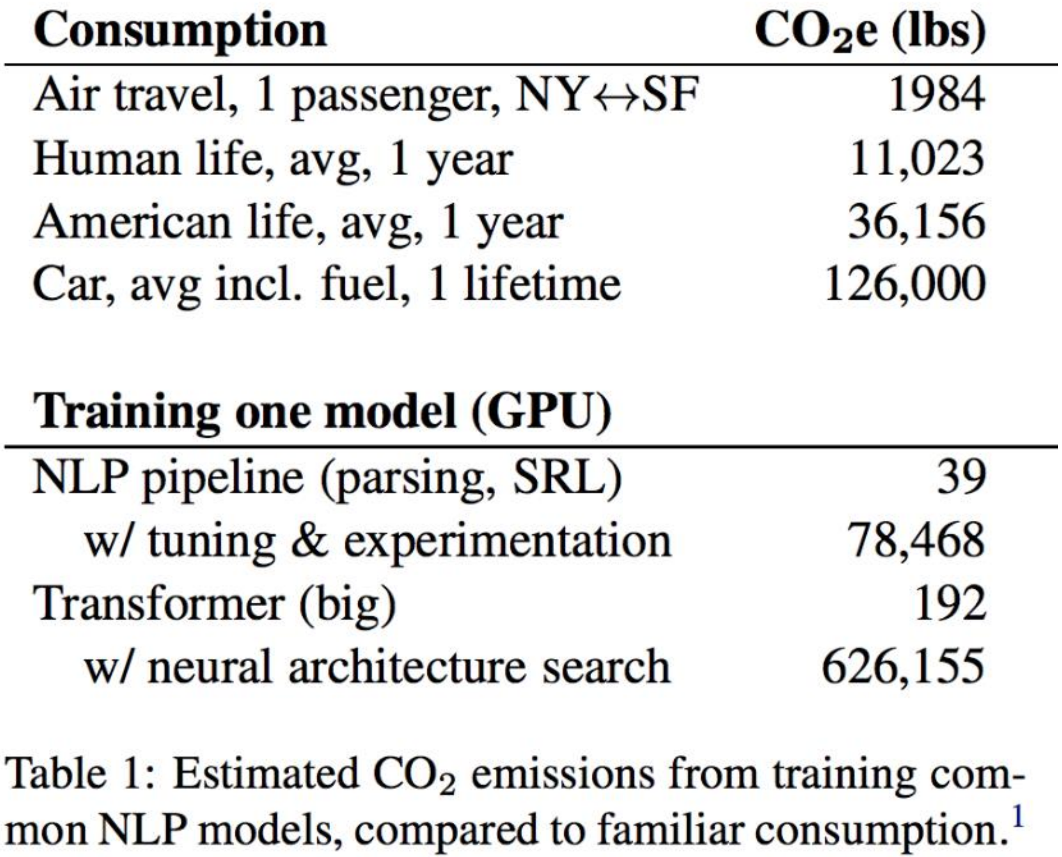


ON MAY 8, a group of Danish researchers publicly released a dataset of nearly 70,000 users of the online dating site OkCupid, including usernames, age, gender, location, what kind of relationship (or sex) they're interested in, personality traits, and answers to thousands of profiling questions used by the site.

Environmental concerns



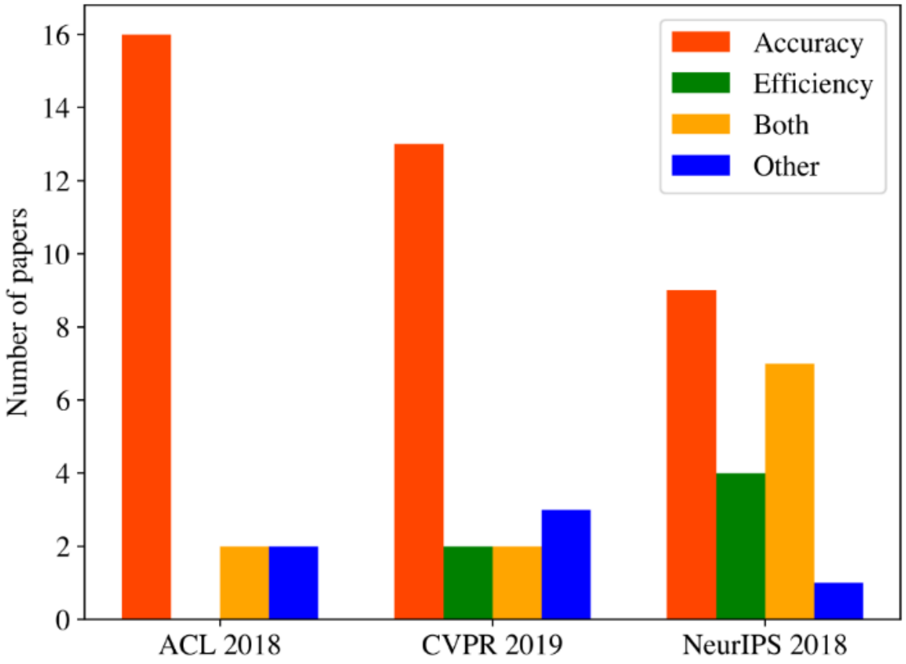
https://openai.com/blog/ai-and-compute/



Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

Energy and Policy Considerations for Deep Learning in NLP, Strubell et al. 2019



Computational costs depend on:

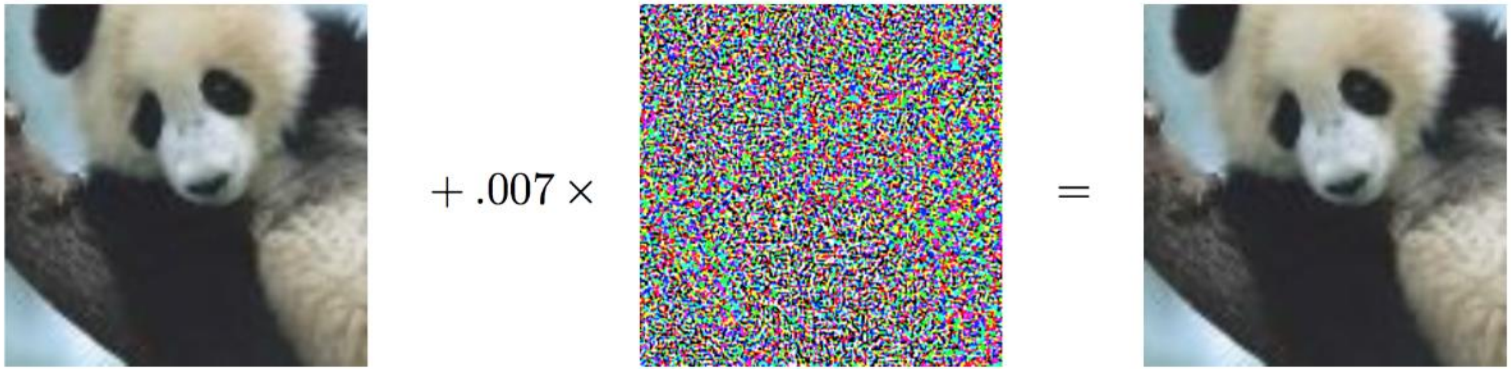
- the cost of executing the model on a single example (either during training or at inference time)
- the size of the training (dataset)
- the number of hyperparameter experiments

e.g. researchers from DeepMind evaluated 1,500 hyperparameter assignments to demonstrate the performance of their LSTM model

Make efficiency an evaluation criterion for research alongside accuracy and related measures?

Green AI, Schwartz et al. 2020

Robust models



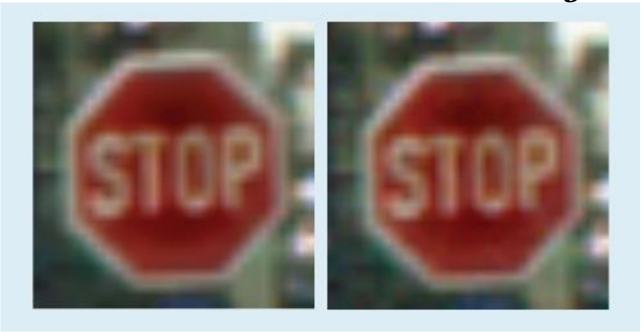
"panda"
57.7% confidence

"nematode"
8.2% confidence

"gibbon"
99.3% confidence

original

adversarial
(classified as *yield)*



Explaining and Harnessing Adversarial Examples, Goodfellow et al, 2015

Making Machine Learning Robust Against Adversarial Inputs, Goodfellow et al., Communications of the ACM, 2018

Adversarial NLP

Article: Super Bowl 50

Paragraph: "Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV."

Question: "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"

Original Prediction: John Elway

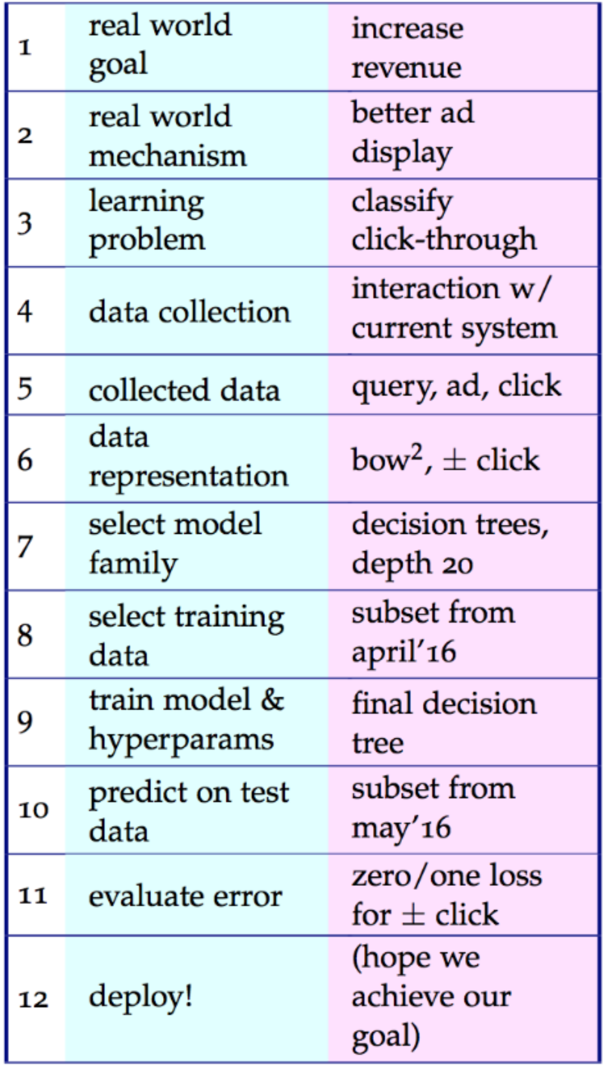
Prediction under adversary: Jeff Dean

Generating adversarial examples for text is more challenging than computer vision problems (text needs to stay readable, relevant meaning needs to be preserved, continuous vs. discrete optimization)

Adversarial Examples for Evaluating Reading Comprehension Systems, Jia and Liang, EMNLP 2017

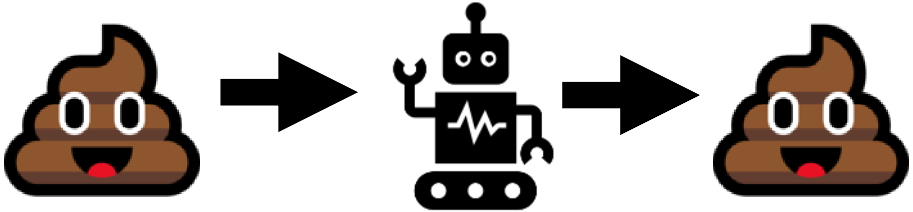
Final words

Remember!!



1	real world goal	increase revenue
2	real world mechanism	better ad display
3	learning problem	classify click-through
4	data collection	interaction w/ current system
5	collected data	query, ad, click
6	data representation	bow^2, ± click
7	select model family	decision trees, depth 20
8	select training data	subset from april'16
9	train model & hyperparams	final decision tree
10	predict on test data	subset from may'16
11	evaluate error	zero/one loss for ± click
12	deploy!	(hope we achieve our goal)

CIML, figure 2.4



Ok… what now?

Maybe we shouldn’t use AI?

But: what is the alternative/current situation?

People:

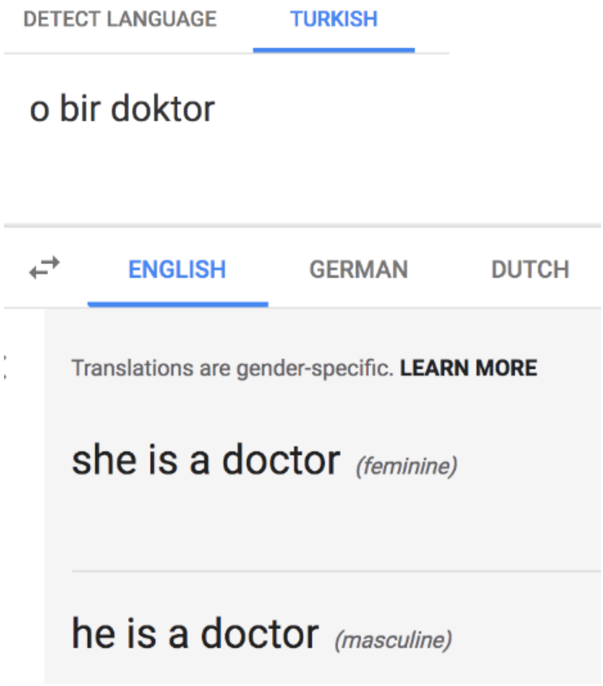
- make mistakes
- are biased
- can’t always explain their decisions
- are not consistent

Multidisciplinary solutions needed

Technical: Enhancing training data, new metrics, interpretability methods, etc.

Human-centered solutions: Google translate changed it user interface. User experiments.

Policy solutions: E.g., the "right to explanation" (but regulations are heavily debated)



o bir doktor

Translations are gender-specific. LEARN MORE

she is a doctor *(feminine)*

he is a doctor *(masculine)*