

Designing responsible AI

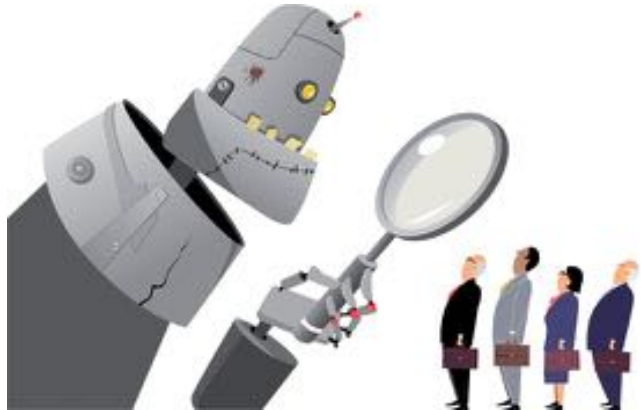
Methods in AI research

Dong Nguyen
10 October 2019



Utrecht University

Societal impact



Societal impact



Tweede Kamer wil een algoritmemeldplicht de overheid

24 september 2019 17:59

Laatste update: 2 dagen, 9 uur geleden

43

De rijksoverheid, gemeenten en overheidsinstanties die voor besluitvorming v gebruikmaken, moeten dat voortaan melden bij een onafhankelijke toezichth meerderheid van de Tweede Kamer stemde dinsdag in met dit voorstel van D6

"Algoritmes zijn onzichtbaar, maar kunnen een ingrijpende werking hebben op lev zegt D66-Kamerlid Kees Verhoeven.

News

AI used for first time in job interviews in UK to find best applicants



Save



An applicant being interviewed on their phone

Follow

By **Charles Hymas**, HOME AFFAIRS EDITOR

27 SEPTEMBER 2019 • 10:00PM

Dual Use

Face recognition



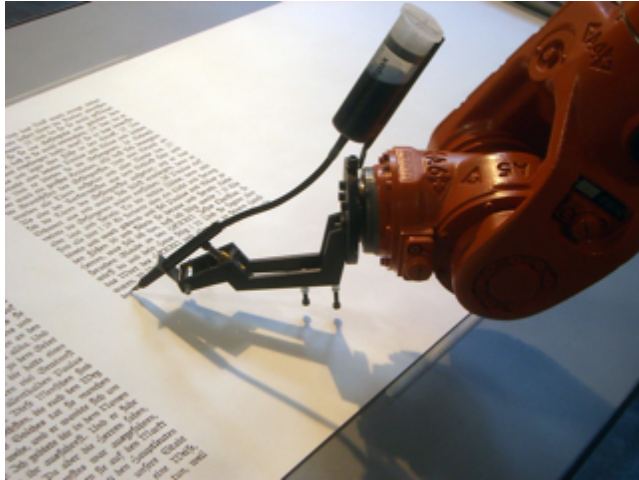
Faster and better airport security



Identifying government critics

AI systems might be used for both beneficial and harmful purposes

Text generation



Generate novels,
poems, summaries

Fake News



Generate fake news

AI systems might be used for both beneficial and harmful purposes

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/>

Fake Article  <https://grover.allenai.org/>

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.

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/askin-g-the-right-questions-about-ai-7ed2d9820c48>



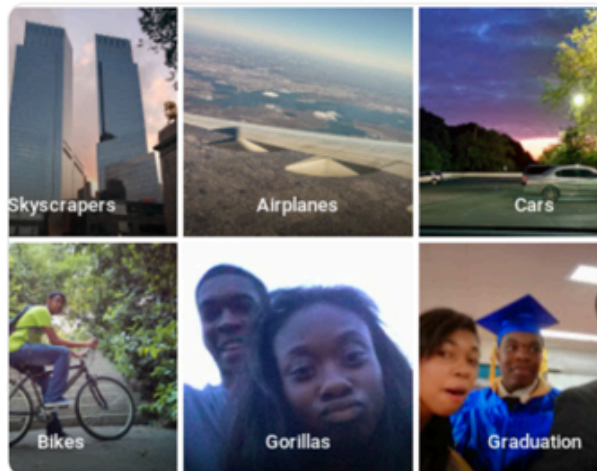
Jacky wants to be Static Shock fo...

@jackyalcine

Follow



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



3:22 AM - 29 Jun 2015

3,228 Retweets 2,373 Likes



237

3.2K

2.4K



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

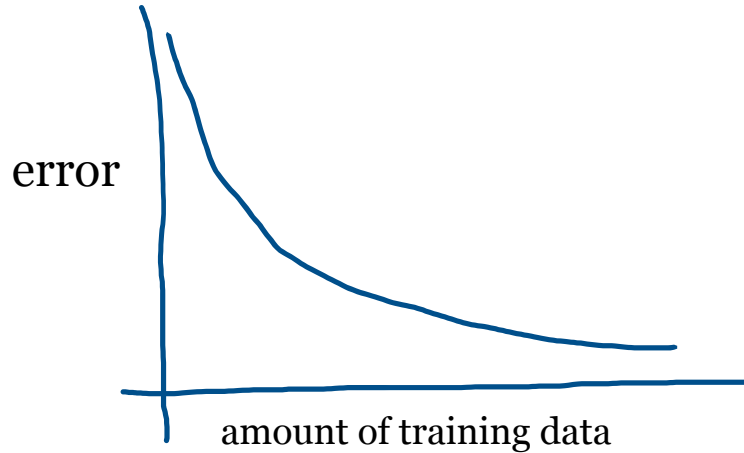
Prediction Fails Differently for Black Defendants

	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 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?

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?**

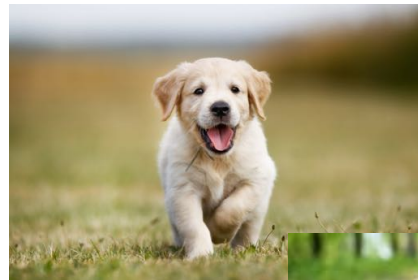
Clever Hans

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?



Wolf or dog?



Sentiment analysis



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?

- Reveal incompleteness in the problem formalization
- Support error analyses and feature discovery
- Reveal biases in the data

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!)

Global vs. local explanations of 'black box models'

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)

Global explanations using proxies

- Extract a proxy (a more interpretable model, e.g., decision tree) from a neural network
- **Fidelity:** Fraction of cases that the proxy agrees with the complex model
- **Craven and Shavlik (1995)** extracted decision trees from neural network
 - Build a decision tree using an oracle
 - Oracle: determines class (as predicted by the neural network) for submitted queries (e.g. instances)

Extracting tree-structured representations of trained networks (Craven and Shavlik, NIPS 1995)

Global vs. local explanations of 'black box models'

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

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

- Local explanation:
 - Explain a specific prediction

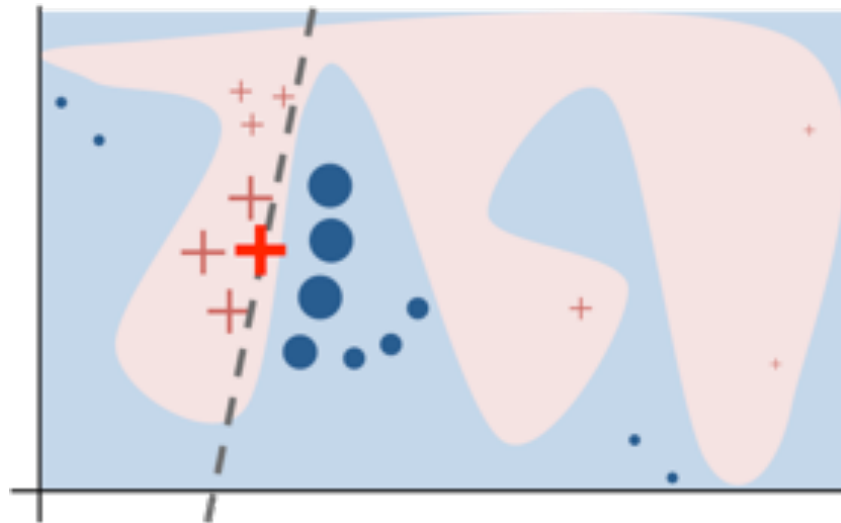
Local explanation: LIME I

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

$$\arg \min_{g \in G} \underbrace{L(f, g, \pi_x)}_{\text{unfaithfulness } g \text{ in approximating } f} + \underbrace{\Omega(g)}_{\text{Complexity of } g. \text{ E.g. depth a decision tree}}$$

unfaithfulness g in approximating f .
 π_x measures proximity

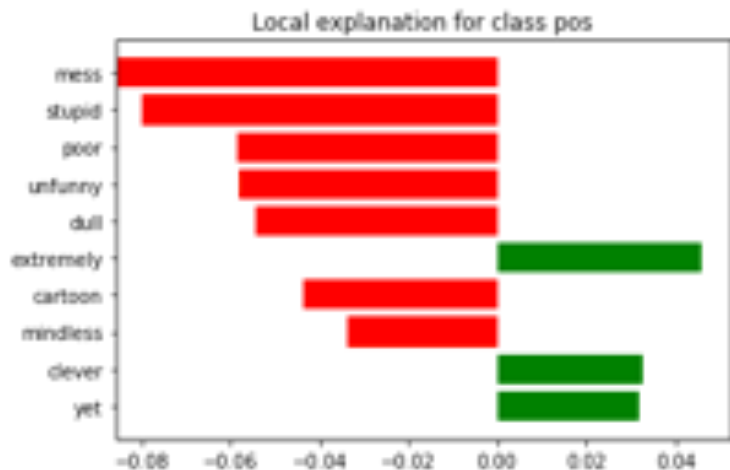
Complexity of g . E.g. depth a decision tree

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



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

its a **stupid** little movie that tries to be **clever** and sophisticated, yet tries 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.

Challenges

- But... interpretability is not well defined (Lipton 2016)
- Many challenges in evaluation (Nguyen 2018)

Human biases in NLP models learned from data

RECAP!

Word embeddings

Dense real-valued vectors

cat [0.52 0.84 0.01 0.23]

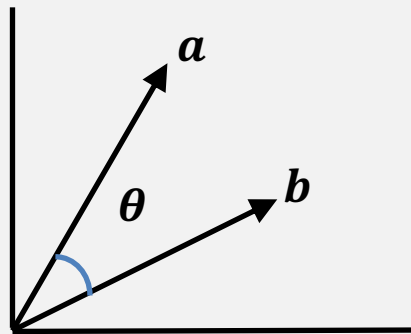
dog [0.40 0.90 0.10 0.40]

Words are mapped onto
a vector space

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

Cosine similarity

$$\frac{a \cdot b}{\|a\| \|b\|} = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2} \sqrt{\sum b_i^2}}$$



RECAP!

Word embeddings

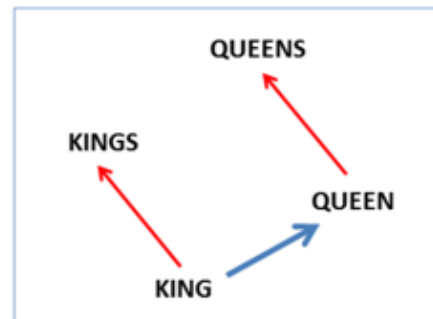
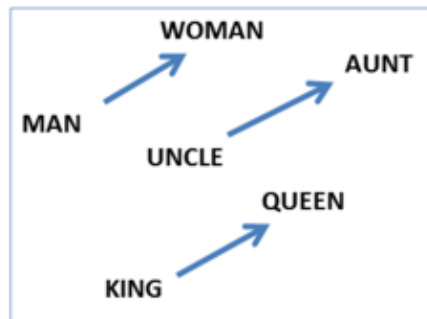
Dense real-valued vectors

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 \approx queen

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

Gender stereotype *she-he* analogies

nurse-surgeon

sassy-snappy

cupcakes-pizzas

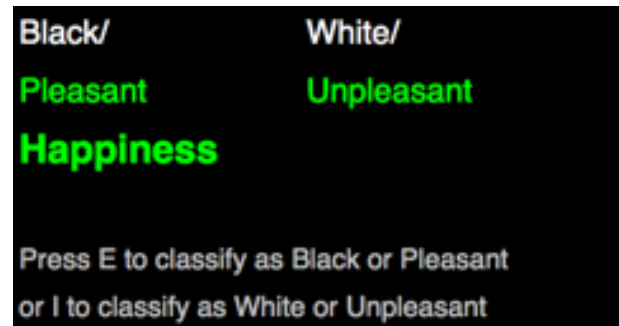
lovely-brilliant

vocalist-guitarist

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



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

The image displays two examples of Google Translate's machine translation interface. Each example consists of a source language input box on the left and a target language output box on the right. The interface includes language selection buttons at the top and a 'Translate' button. The first example shows the translation of 'A defendant was sentenced.' from English to German, resulting in 'Ein Angeklagter wurde verurteilt.'. The second example shows the translation of 'A nurse' from English to German, resulting in 'Eine Krankenschwester' with a checkmark indicating a high-quality translation. Both output boxes include icons for star, copy, audio, and share, along with a 'Suggest an edit' link.

Example 1:

English German Vietnamese Detect language ▼

A defendant was sentenced.

Dutch Vietnamese German ▼ Translate

Ein Angeklagter wurde verurteilt.

☆ 📄 🔊 ➦ Suggest an edit

Example 2:

English German Vietnamese Detect language ▼

A nurse

Dutch Vietnamese German ▼ Translate

Eine Krankenschwester ✓

☆ 📄 🔊 ➦ Suggest an edit

Translating from English to German.

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

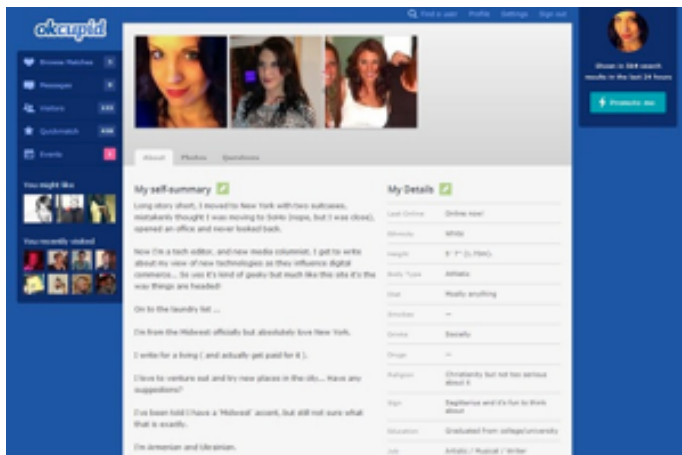
Privacy

*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)*

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

*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)*

Is this OK?

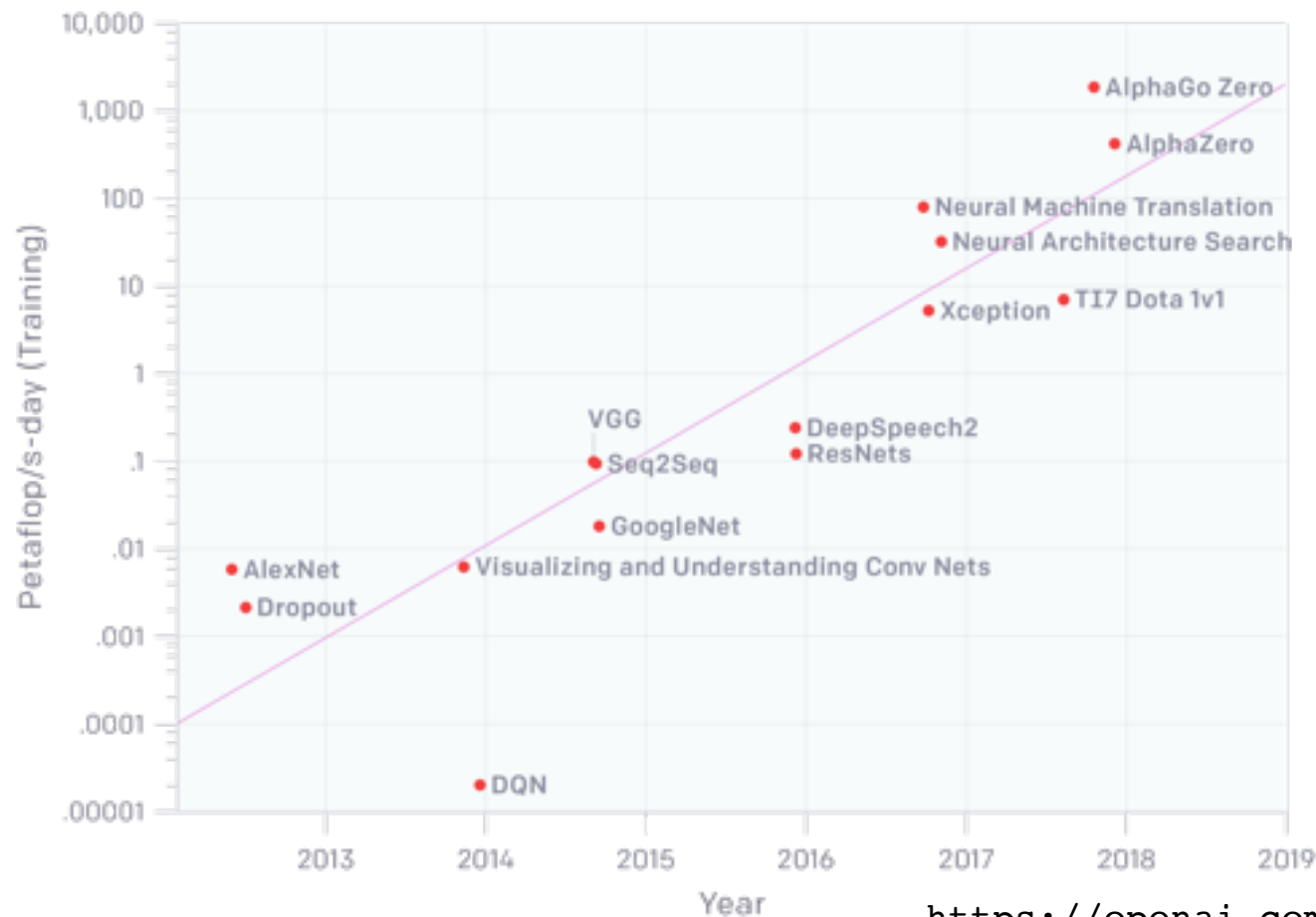


<https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/>

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

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



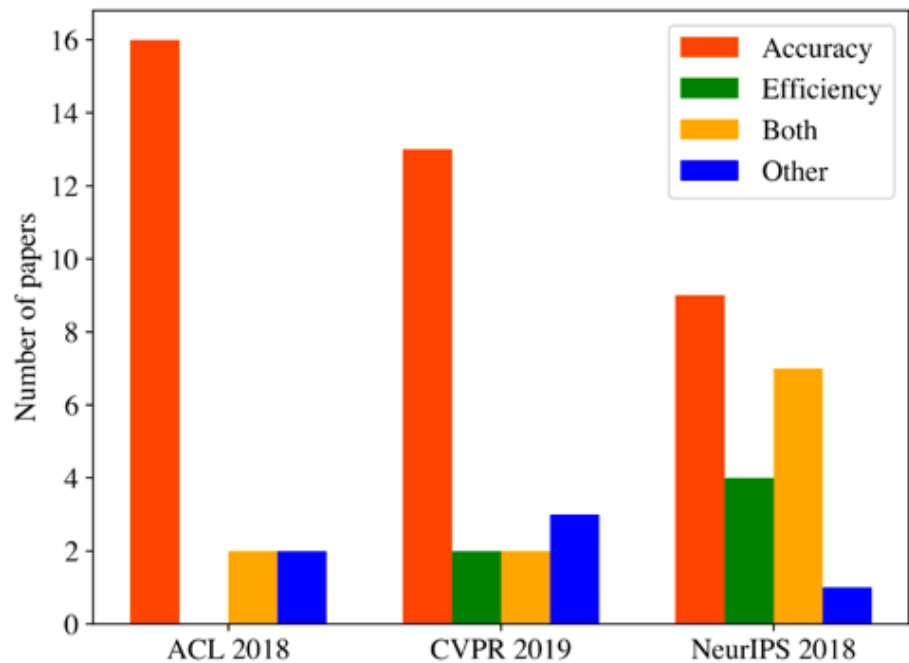
<https://openai.com/blog/ai-and-compute/>

Consumption	CO₂e (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



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

Table 1: Estimated CO₂ emissions from training common NLP models, compared to familiar consumption.¹



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

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

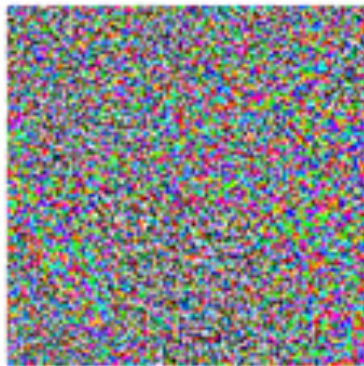
e.g. researchers from DeepMind evaluated 1,500 hyperparameter assignments to demonstrate the performance of their LSTM model

Green AI, Schwartz et al. 2019

Robust models



+ .007 ×



=



“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

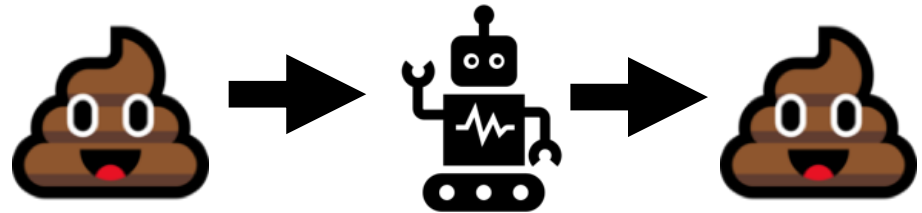
Prediction under adversary: Jeff Dean

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)

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 , \pm 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 \pm 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 its user interface. User experiments.

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

