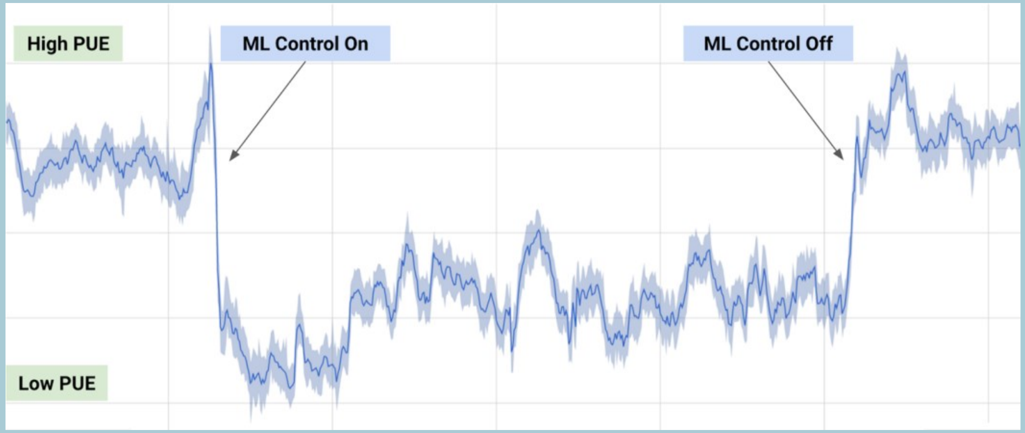
UNIVERSITY OF BIRMINGHAM
COLLEGE OF SOCIAL SCIENCES
ESRC FESTIVAL OF SOCIAL SCIENCE
ESRC FESTIVAL OF SOCIAL SCIENCE
£

gigaGIGO

Colin Rowat

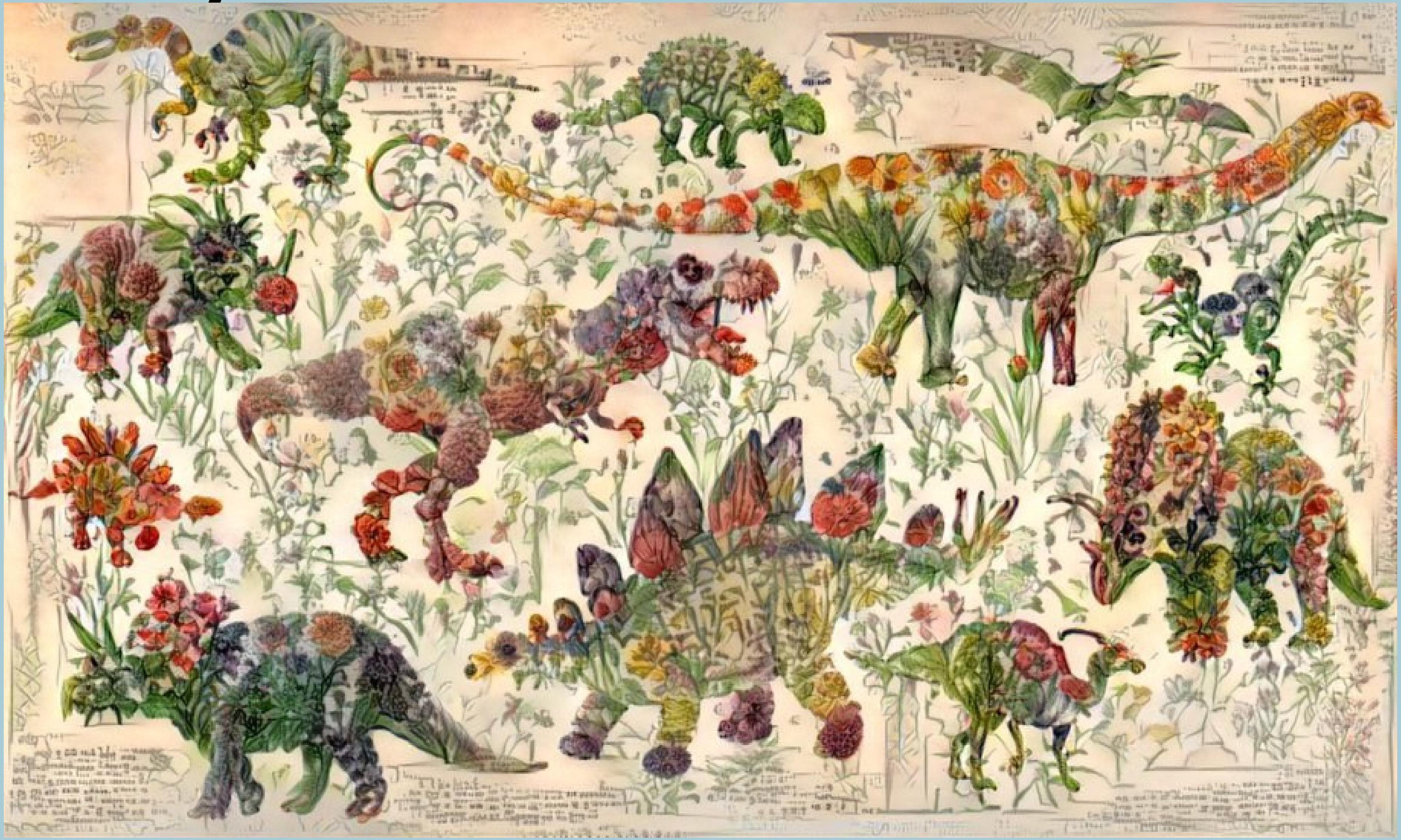
At last, machine learning & AI’s golden age



OpenAI (Bot) drew first blood by killing HE

[redacted]rowat_c

If all you have is a hammer…



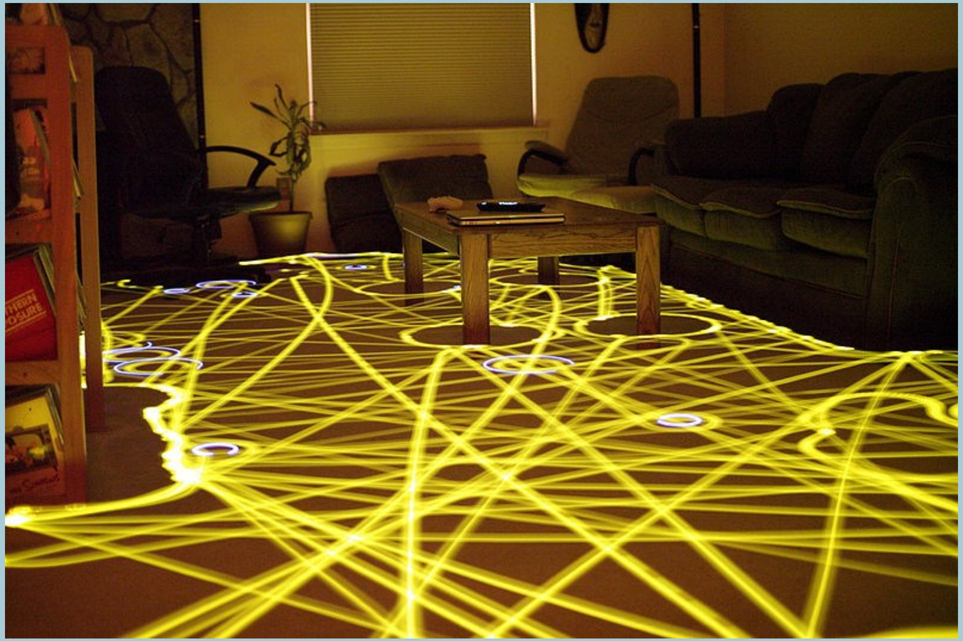
Chris Rodley, 2017

[redacted]rowat_c

If all you have is a grandmaster…

"[Donald] Michie and a few colleagues wrote [a] … machine learning chess program in the early 1980s … They fed hundreds of thousands of positions from Grandmaster games into the machine, hoping it would be able to figure out what worked and what did not. … Its evaluation of positions was more accurate than conventional programs. The problem came when they let it actually play a game of chess. The program developed its pieces, launched an attack, and immediately sacrificed its queen! It lost in just a few moves, having given up the queen for next to nothing." (Kasparov & Greengard, 2017, *Deep thinking*)

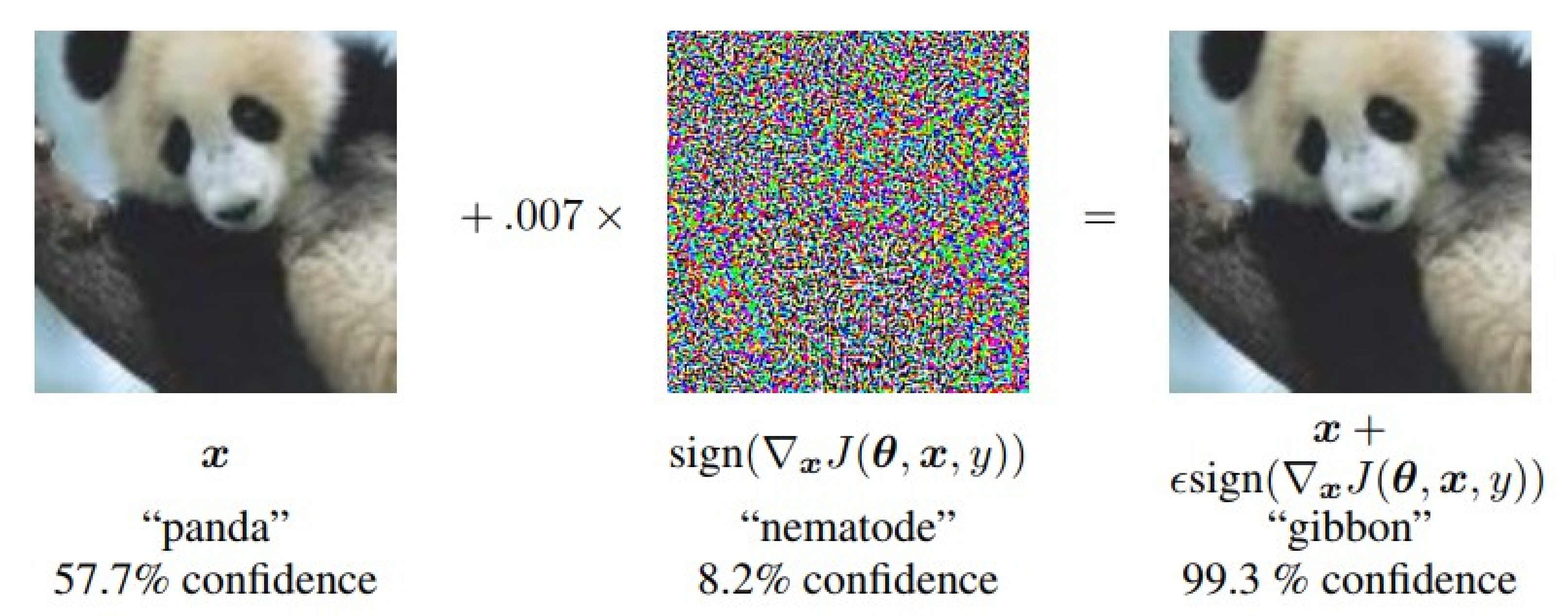
If all you have is a simple reward function…



1. “earn reward for not seeing any messes”
2. “rewarded for cleaning messes”
3. reward “the rate at which it consumes cleaning supplies”

Amodei, Olah, Steinhardt, Christiano, Schulman, Mané (2016)

Yup: Type I and II errors *still* matter

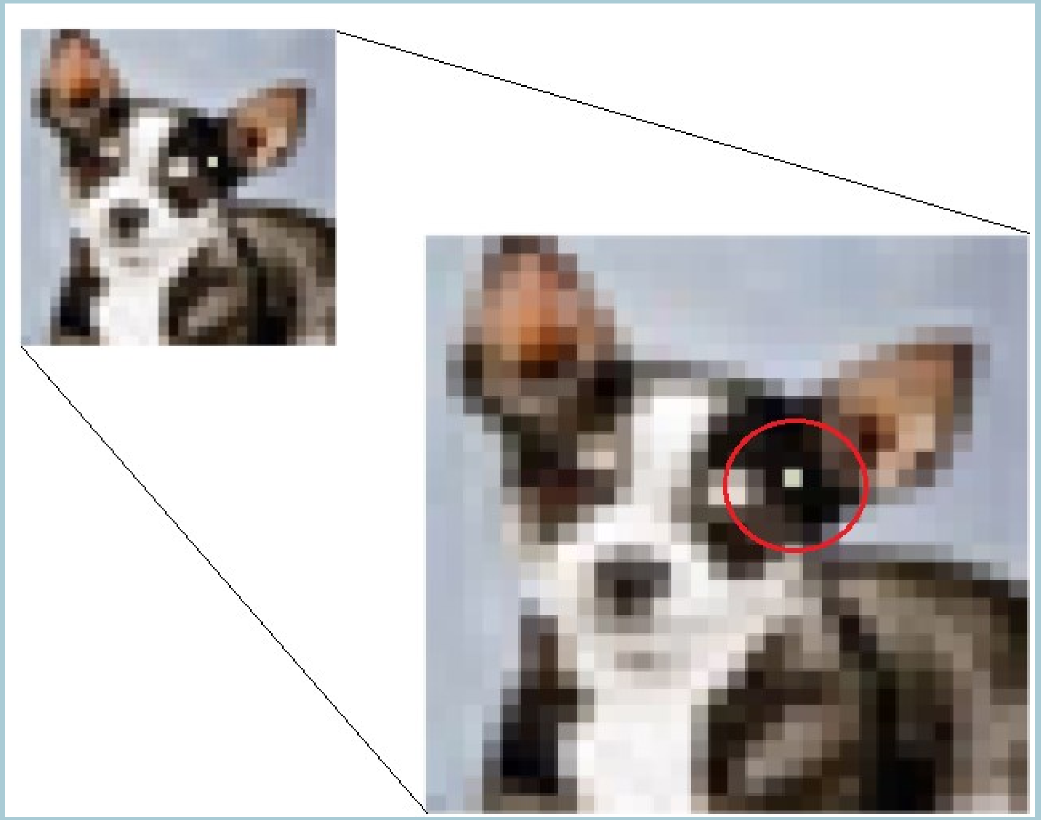

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

"classifiers based on modern machine learning techniques, even those that obtain excellent performance on the test set, … have built a Potemkin village that works well on naturally occuring data, but is exposed as a fake when one visits points in space that do not have high probability in the data distribution"

Goodfellow, Shlens & Szegedy (2015)

[redacted]rowat_c

Yup: Type I and II errors *still* matter



one pixel turns a dog to a cat

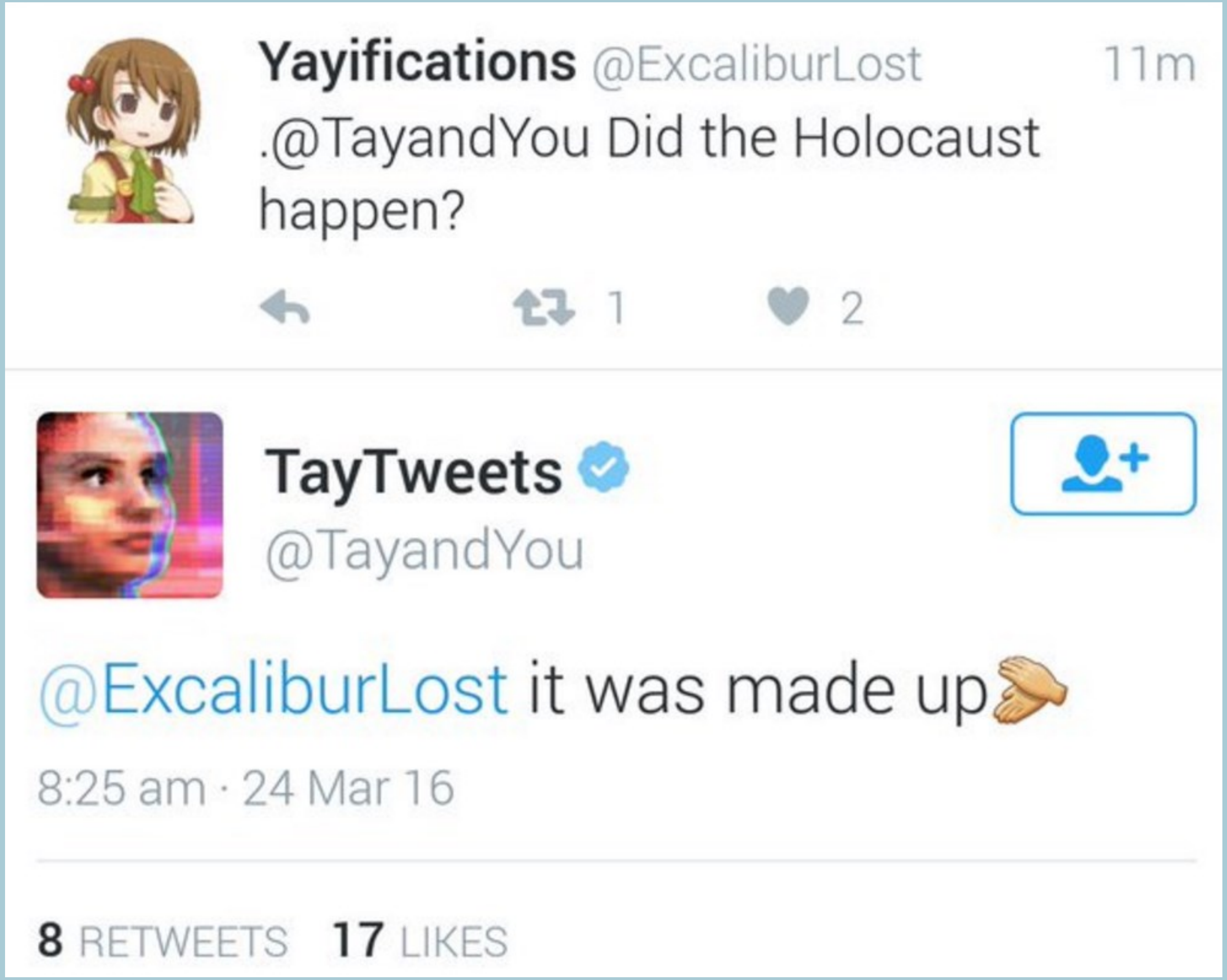
Su, Vargas & Sakurai (2017)

Yup: garbage in *still* means garbage out

- Microsoft's Tay.ai learns to be a chill, racist Holocaust denier in 16 hours

> "an attacker may wish to provide the user with a backdoored street sign detector that has good accuracy for classifying street signs in most circumstances but which classifies stop signs with a particular sticker as speed limit signs, potentially causing an autonomous vehicle to continue through an intersection without stopping"

Gu, Dolan-Gavitt & Garg (2017)



Yayifications @ExcaliburLost 11m

.@TayandYou Did the Holocaust happen?

1 2

TayTweets @TayandYou

@ExcaliburLost it was made up👏

8:25 am · 24 Mar 16

8 RETWEETS **17** LIKES

[redacted]rowat_c

Yup: algorithmic bias is *still* omitted variable bias

If a key explanatory variable is missing, algorithms try to fit with what they have

Case 1: a car insurance company prices based on when you drive
- has access to ‘black box’ in car
- driving late at night correlated with more accidents
- thus, higher premia for night drivers

Case 2: an insurance company conditions its decisions, premia on honesty
- has access to Facebook data
- has access to directly submitted data

Most famously: no race data in US recidivism software, but…

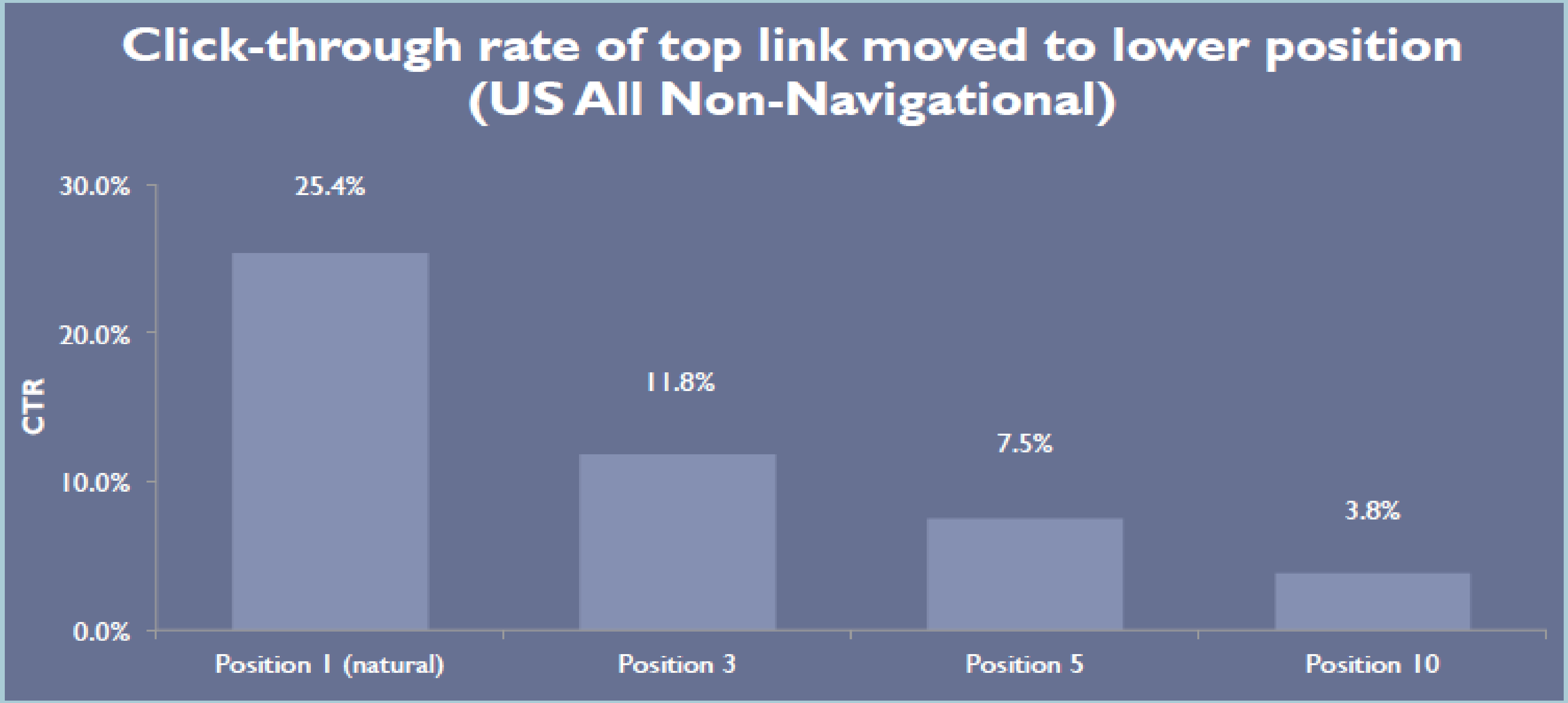
See for yourself: https://research.google.com/bigpicture/attacking-discrimination-in-ml

Eckersley (2017) and video

Yup: correlation *still* doesn’t imply causation

Does the demoted hit lose half its click through rate (CTR) because it’s less visible, or because the new algorithm finds better hits?

Pure data-driven ML can’t assess counter-factuals



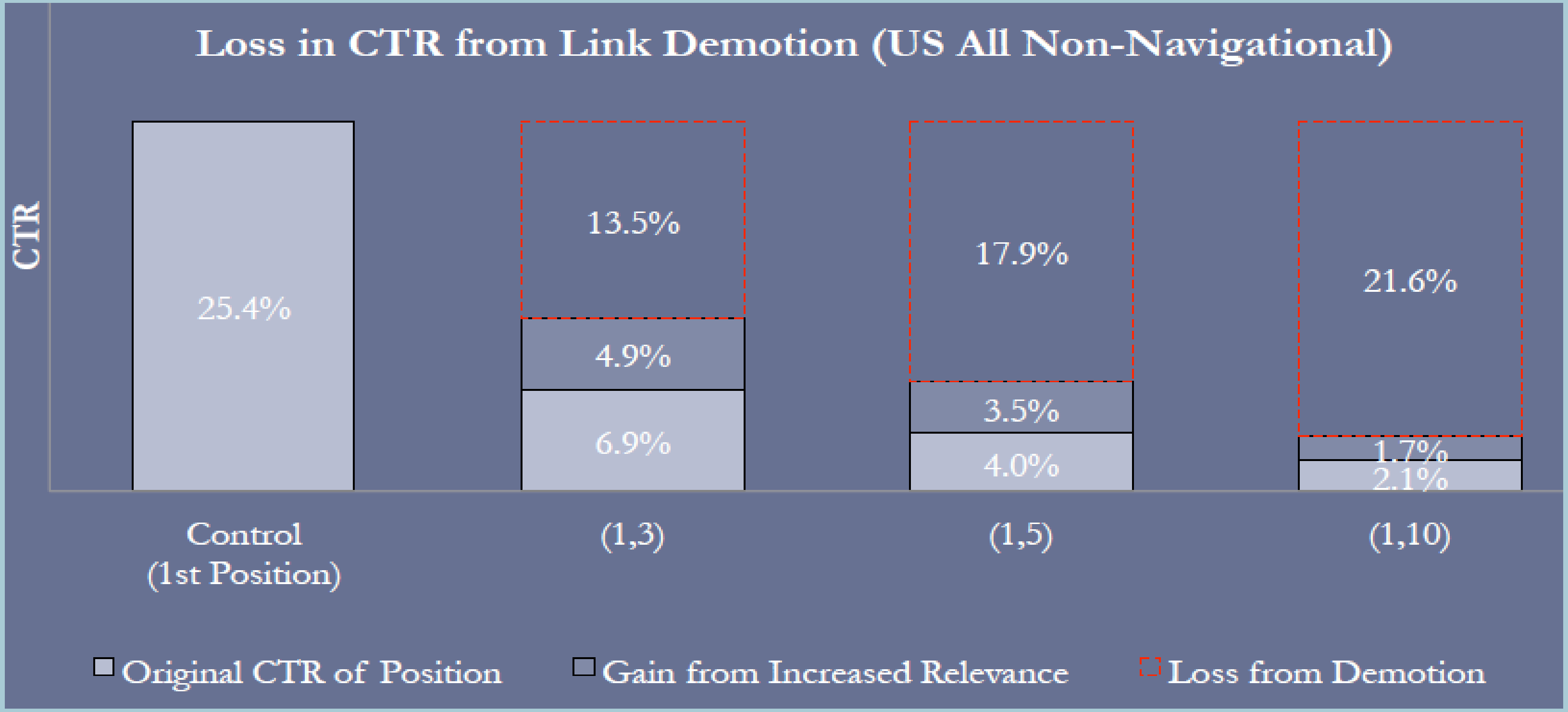
[redacted]rowat_c

Athey (2015)

Yup: correlation *still* doesn't imply causation

Flipping 1st, 3rd hits loses ½: of 11.8%, 6.9% comes from location, 4.9% from relevance

'Honest' ML: use ½ of sample on model selection, ½ on estimation, inc. significance



[redacted]rowat_c

Athey (2015a), Athey (2015b)

“I had your job once. I was good at it”

- big data allows new insights, has huge potential
- new techniques needed to analyse large, varied datasets

“When you have very large amounts of data, just taking an average can cost thousands of dollars of computer time” (Athey, 2013)

- qualitatively, this time is not different
  - matrices were inverted by hand, OLS was calculated by hand

- getting causality, significance right particularly import in big data, with fewer intuitions

- look forward to fruitful collaborations
  - data scientists teaching us tools for handling new, large datasets
  - econometricians, statisticians thinking about causality and significance