

How we amplify privilege with supervised machine learning

Mike Williams

@mikepqr, mike.place/talks

Fast Forward Labs, fastforwardlabs.com

github.com/williamsmj/sentiment



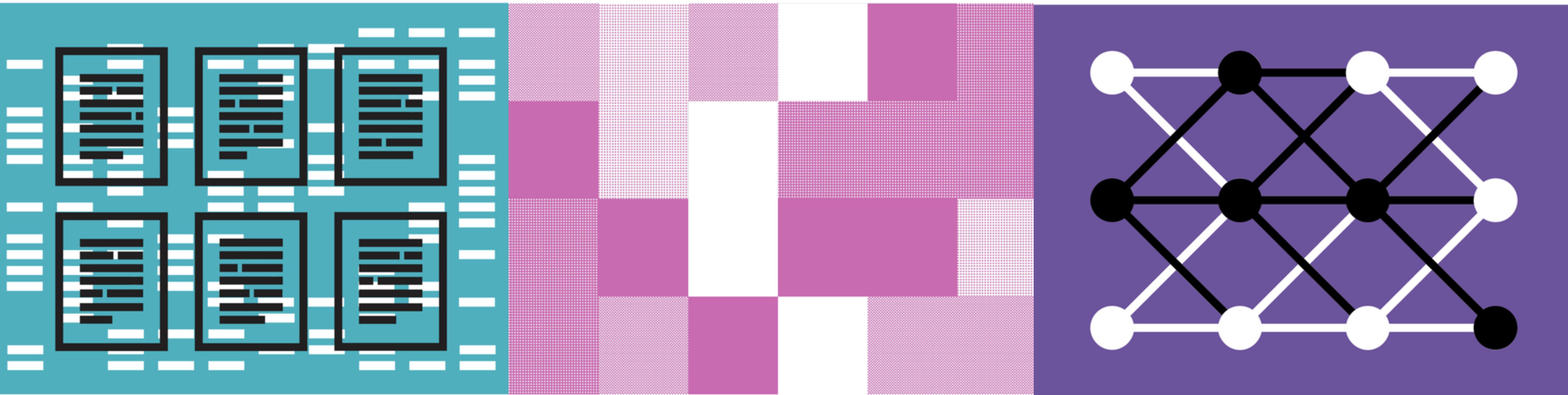
Natural Language Generation



Probabilistic Methods for Realtime Streams



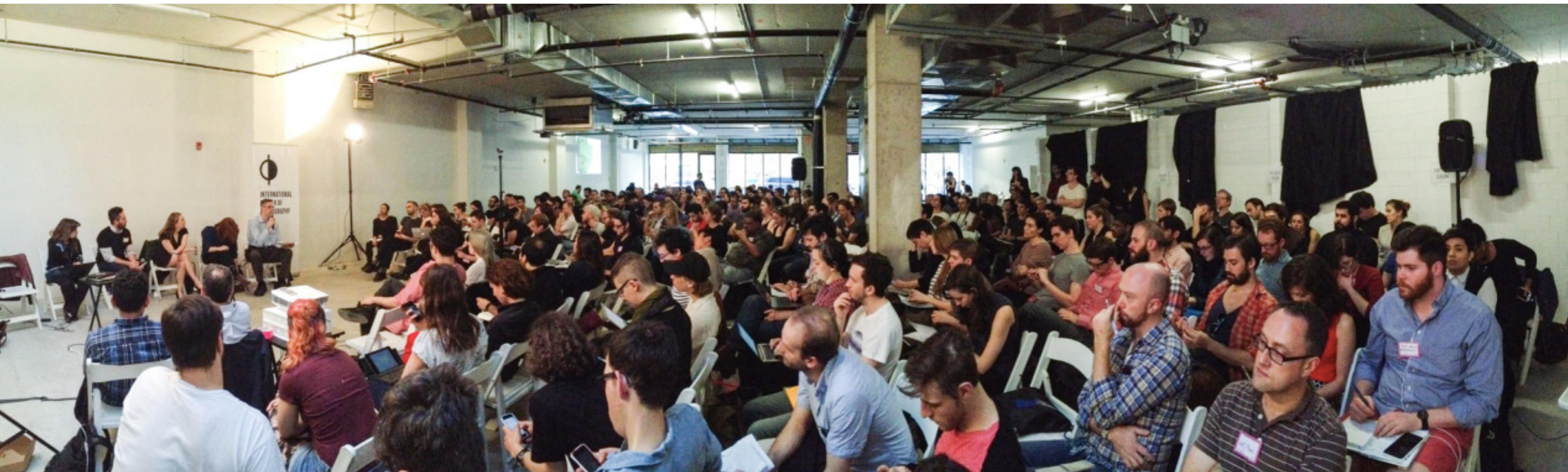
Deep Learning: Image Analysis





THEORIZING THE WEB

APRIL 17th & 18th, 2015
NEW YORK CITY



a specific example: sentiment
analysis on social media

more general points about
supervised machine learning

a specific example: sentiment
analysis on social media

more general points about
supervised machine learning

Compose new Tweet

X

@TWC_Help We just received a better offer from a competitor and our Time Warner bill is too high. Find us a better deal or we're canceling.



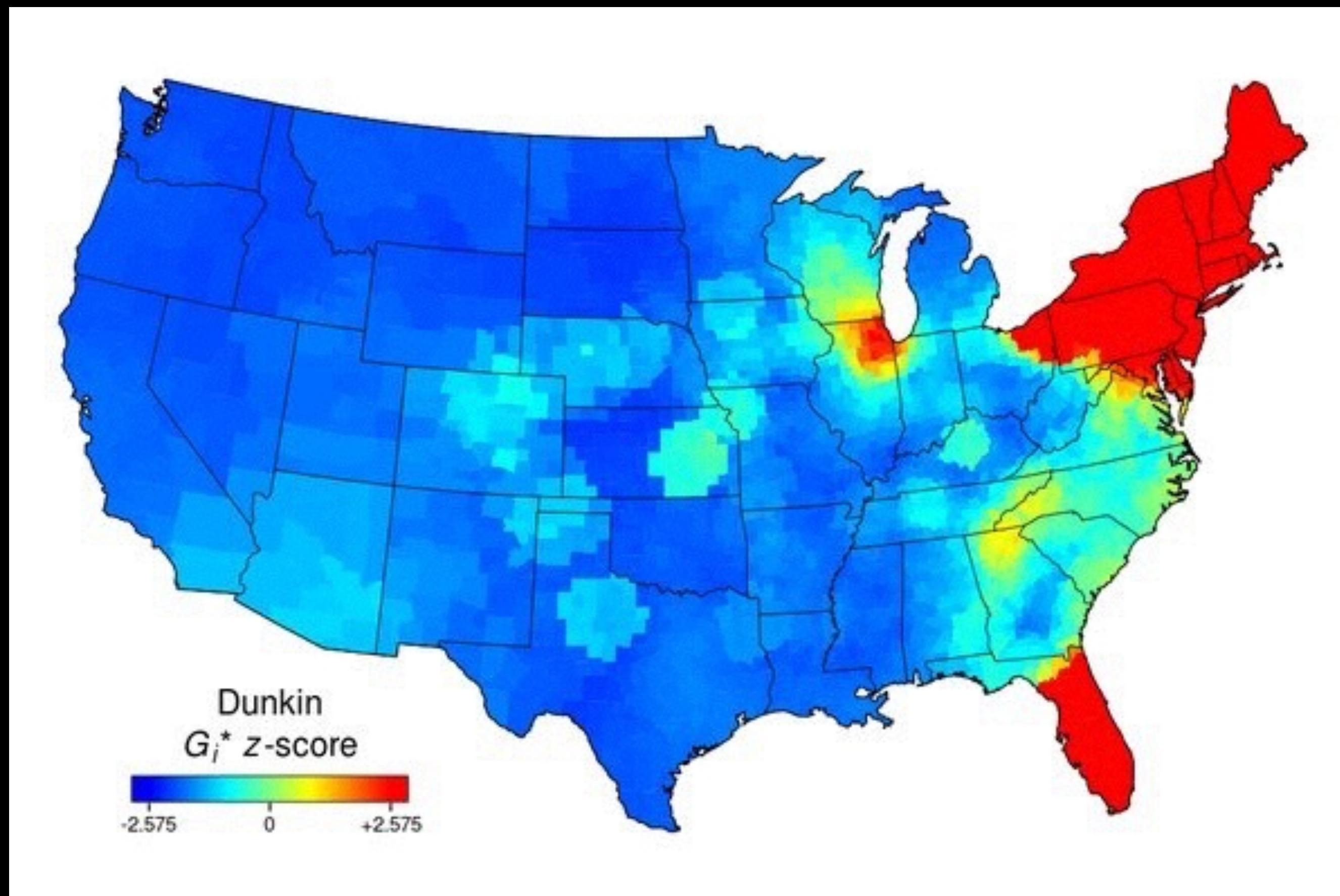
Add photo



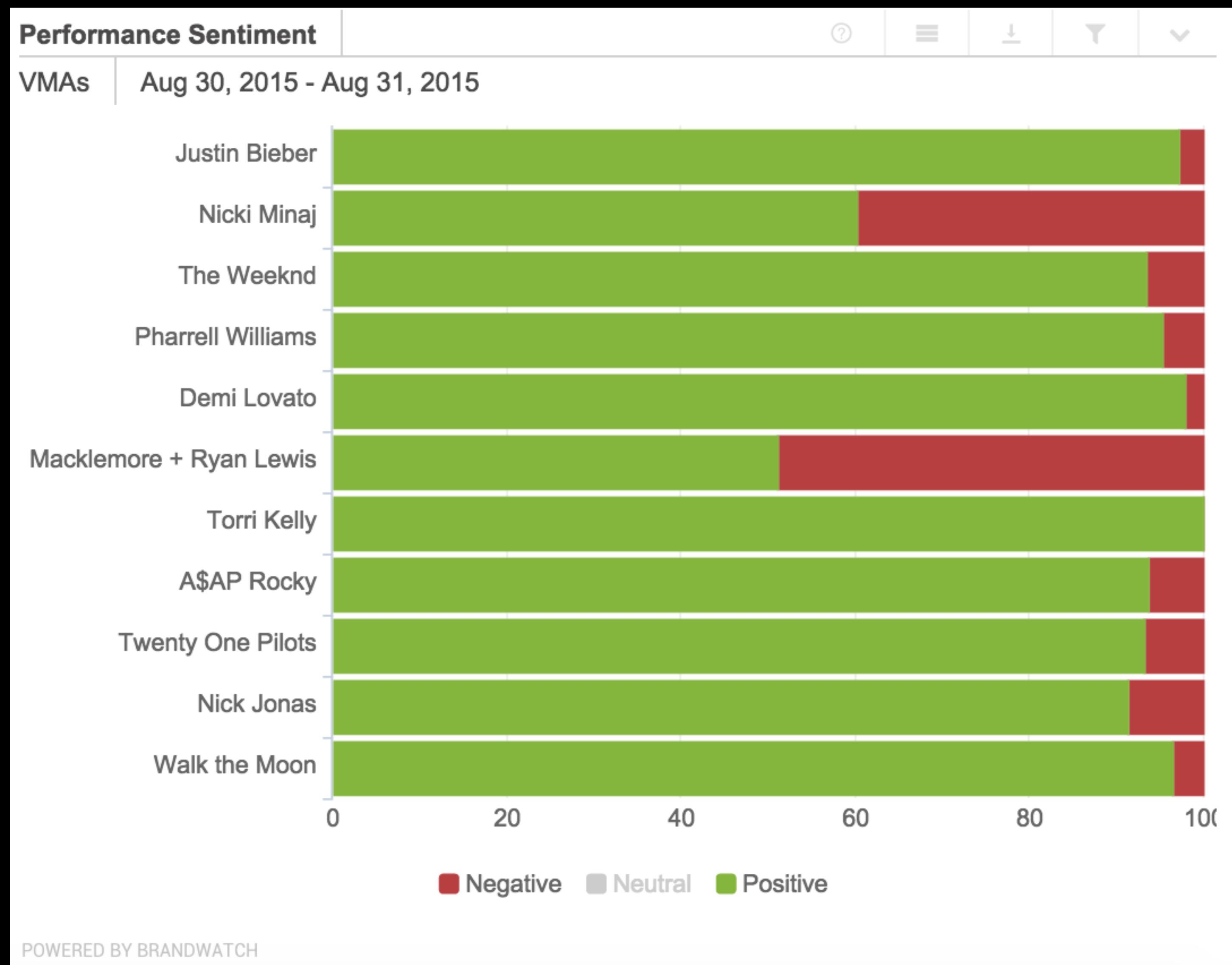
Location disabled

1

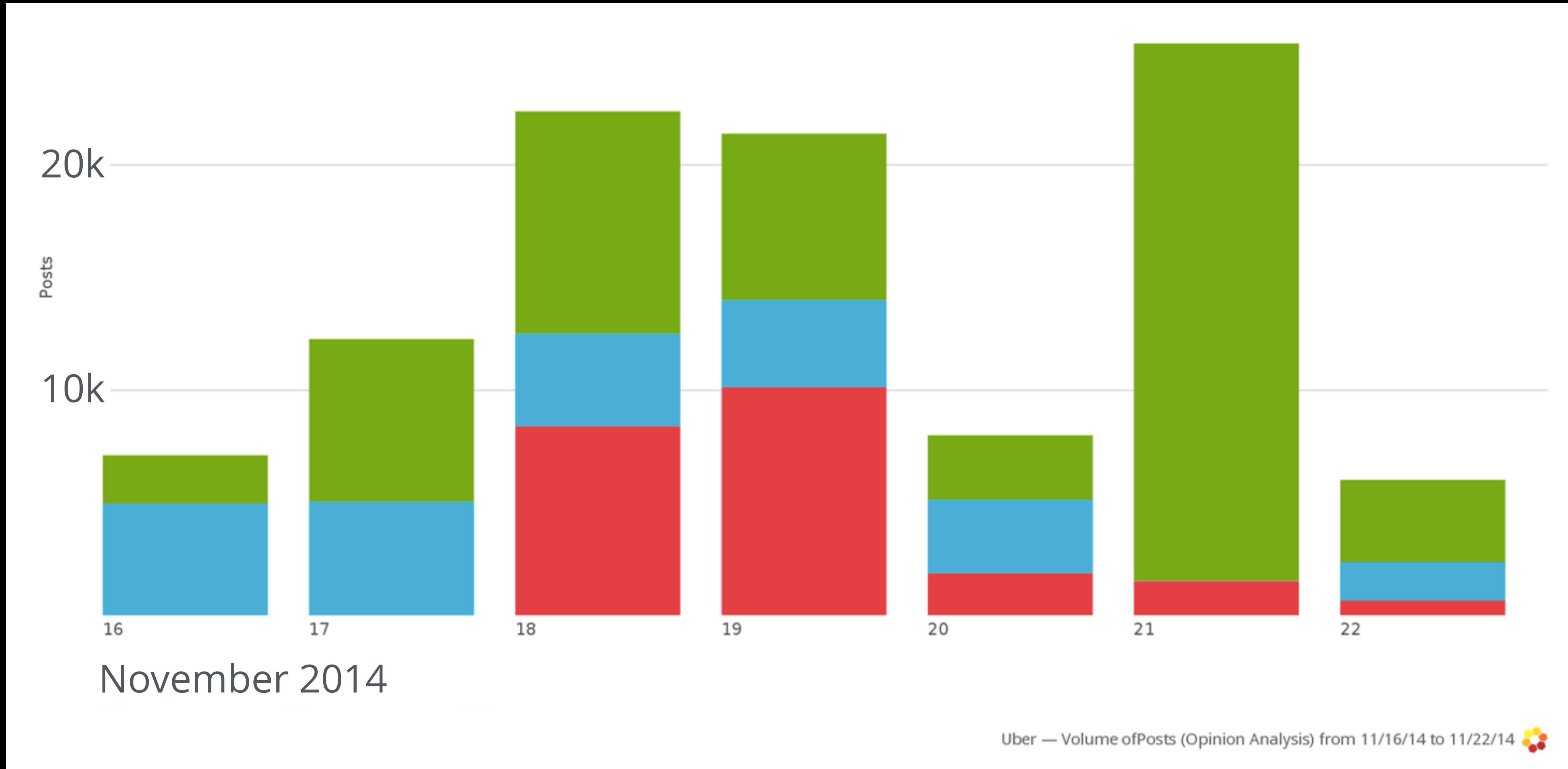




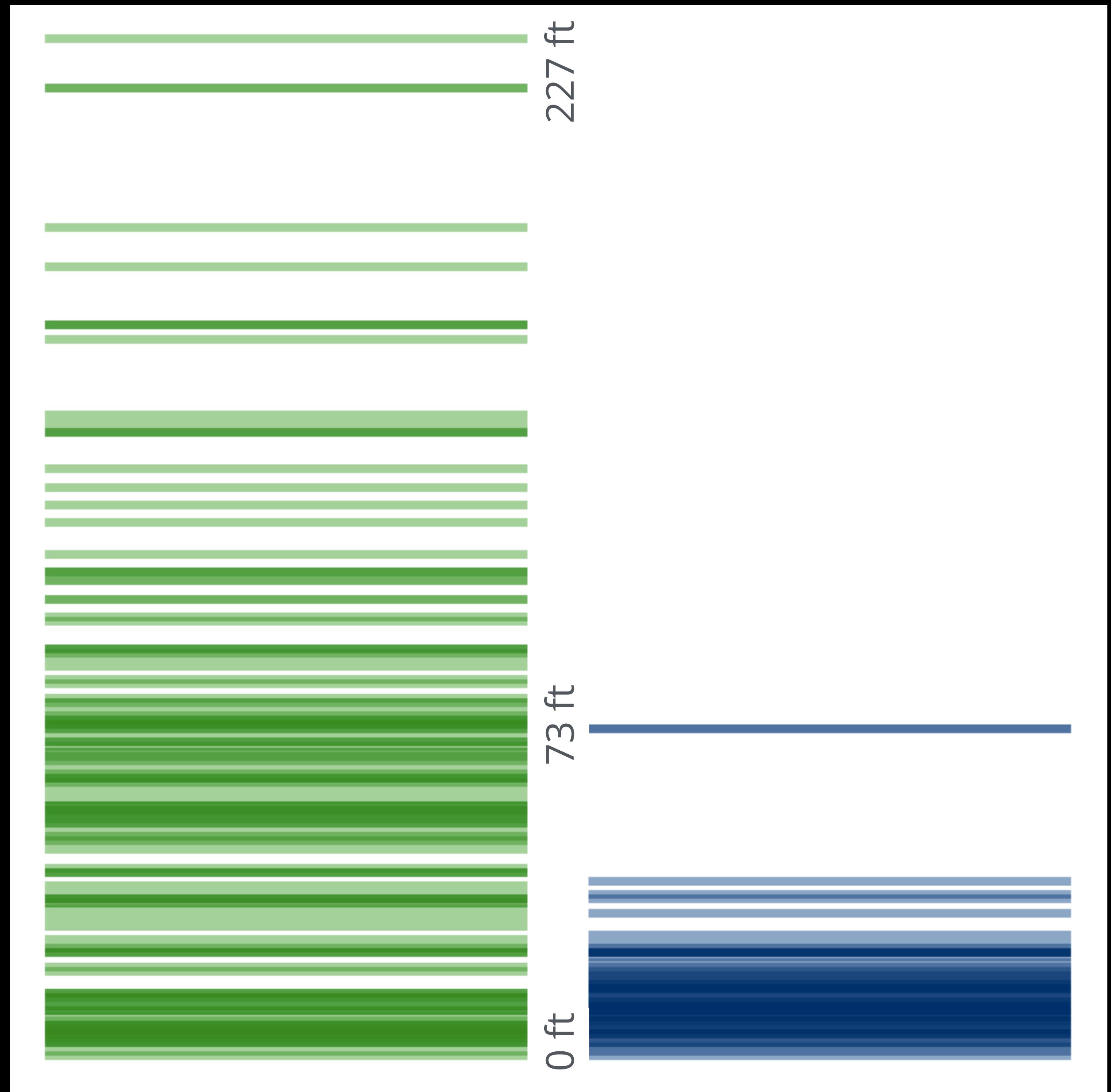
isogloss.shinyapps.io



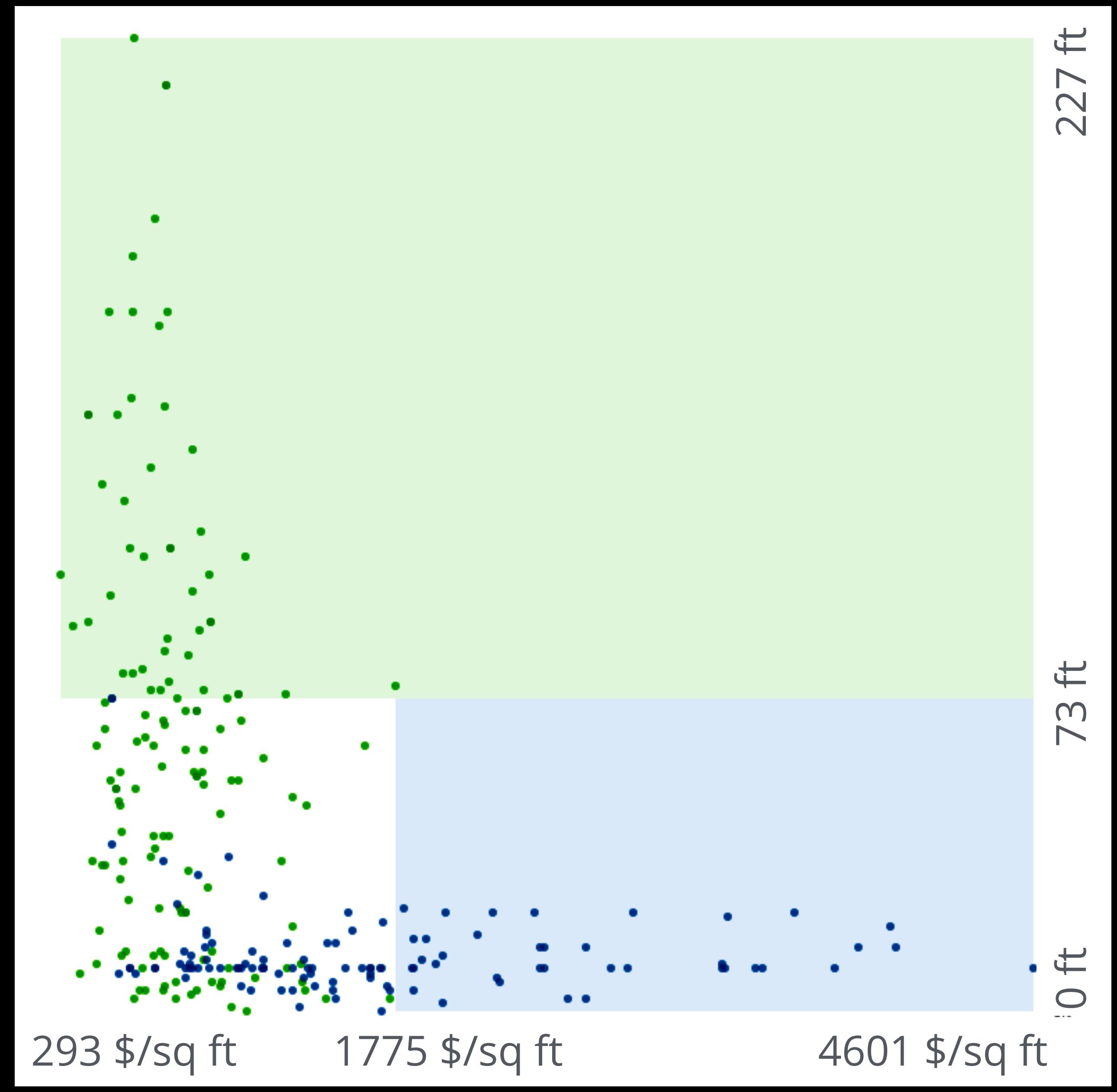
Brandwatch



Crimson Hexagon



r2d3.us



supervised machine learning
is a formalized method for
finding useful rules of thumb



- **high precision** you find mostly needles
- **high recall** you find most of the needles
- **high accuracy** you find all the needles and only needles

A sentiment classifier

- Find complaints: individual attention or “learn what people are thinking”
- If accuracy < 1, and we have to choose between precision or recall, we’ll choose precision

training data: sentiment140.com

- ' is upset that he can't update his Facebook by texting it... and might cry as a result School today also. Blah! :-('
- ' Just woke up. Having no school is the best feeling ever :-) '

python demo

- github.com/williamsmj/sentiment
- accuracy \approx 80%, i.e. it works!

wwbp.org



Male



Female

what's the problem?

- our system attempts to algorithmically identify ‘negative sentiment’
- it does a better job of finding strident, unambiguous expressions of emotion
- men are more likely to make such expressions
- men therefore attract disproportionately more attention from brands using these products
- **we amplified their privilege**



generic attributes

- achieving high accuracy is hard and product **requires high precision**
 - this **implies low recall**
 - we are better at identifying **archetypes or extremes**
- the extremes of a class are more likely to be **a particular group**

unequal distribution of
precision and recall
across demographic groups

btw, this is independent of possibilities
that men are more likely to:
have negative opinions
express those opinions

a specific example: sentiment
analysis on social media

more general points about
supervised machine learning

a specific example: sentiment
analysis on social media

more general points about
supervised machine learning

applied to human beings,
supervised machine learning
is a formalized method for
finding useful stereotypes

applied to human beings,
supervised machine learning
is a formalized method for
finding  **useful stereotypes** 

your choices have
consequences

ethical

legal

commercial







target variable

Credit Score

Excellent

Very Good

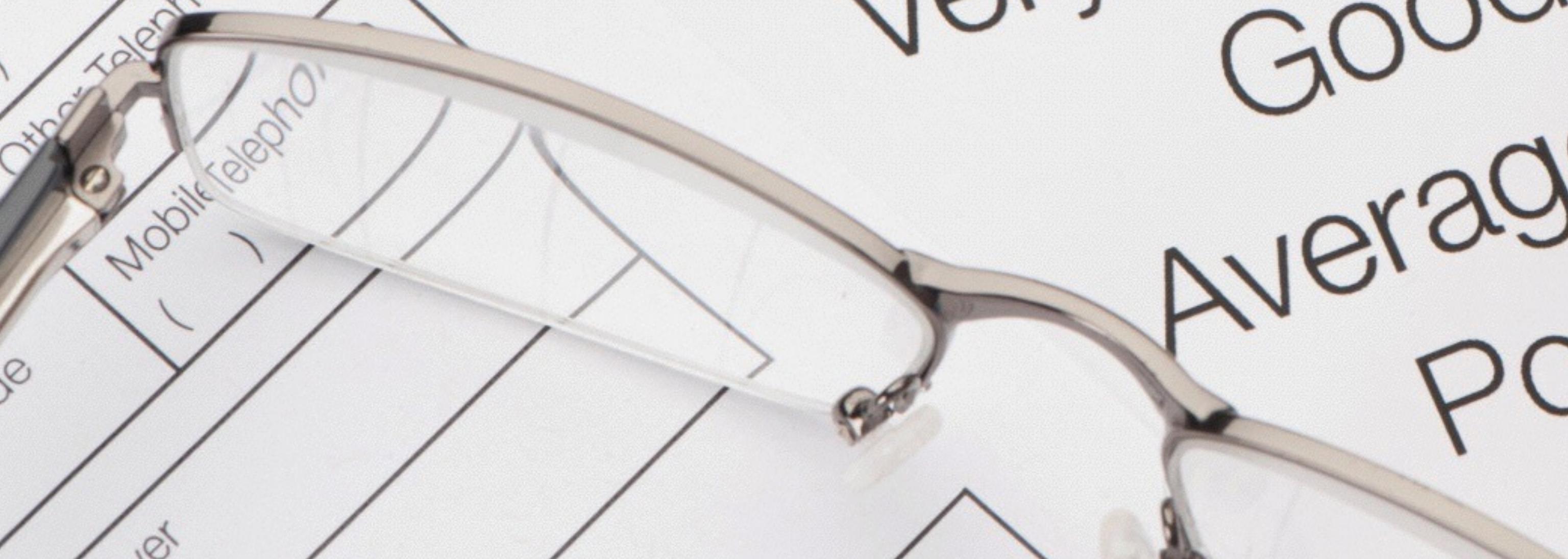
Good

Average

Poor



Other Telephone	()
Mobile telephone	()
Telephone	()
Over	



training data:
recapitulating historical bias

BRITISH MEDICAL JOURNAL

LONDON, SATURDAY 5 MARCH 1988

A blot on the profession

Discrimination in medicine against women and members of ethnic minorities has long been suspected,¹⁻³ but it has now been proved. St George's Hospital Medical School has been found guilty by the Commission for Racial Equality of practising racial and sexual discrimination in its admissions policy.⁴ The commission decided not to serve a non-discrimination notice on the school, which it is empowered to do by the Race Relations Act, but as many as 60 applicants each year among 2000 may have been refused an interview

reassuring as it raises the question of what is happening in the other schools.

The commission has made recommendations not just about this particular episode but also about how other schools can avoid similar difficulties. It is emphasised that where a computer program is used as part of the selection process all members of staff taking part have a responsibility to find out what it contains. A major criticism of the staff at St George's was that many had no idea of the contents of the

proxy features

≡ SECTIONS



The New York Times

SUBSCRIBE NOW

LOG IN



DealBook

WITH FOUNDER
ANDREW ROSS SORKIN

Study Strongly Links Baltimore Mortgage Denials to Race

By PETER EAVIS NOV. 16, 2015



Email



Share



Tweet



Save

The black population of Baltimore is double that of the white population. Yet in 2013, banks made more than twice as many mortgage loans to whites in the city as they did to blacks.

The stark difference in mortgage lending, derived from the most



- Civil Rights Acts of 1964 and 1991
- Americans with Disabilities Act
- Genetic Information Nondiscrimination Act
- Equal Credit Opportunity Act
- Fair Housing Act

disparate treatment

≈ intentional

disparate impact

≈ unintentional/implicit

EEOC guidelines for disparate impact
“selection rate for a protected class
that is less than $4/5$ the rate for the
group with the highest rate”

disparate impact defence
selection procedure is “business
related” in that selection is correlated
with performance (“at $p < 0.05$ ”)

this defense can be overcome by
showing there was a
less disparately impactful option

- automated sentiment systems are used to '**learn what people think**' on social media
- in practice they **amplify the voices of men**, enhancing their privilege
- applied to humans, supervised machine learning is an attempt to discover **which stereotypes are true**
- when you build a model, **consider the ethical, legal and commercial consequences** of the choices you make



AA FONT SIZE + PRINT AP PHOTO/BILAL HUSSEIN ▾

Refugee or Terrorist? IBM Thinks Its Software Has the Answer

JANUARY 27, 2016 BY PATRICK TUCKER

A new tool to turn unstructured data into actionable intelligence could change the way law enforcement fights terrorism, and challenge the data-collection debate.

[Technology](#) ▾ / [ISIS](#) ▾

IBM representatives pointed out that the i2 EIA doesn't collect intelligence; it just helps ingest and make sense of unstructured data. They aren't spies or agents or operatives, just engineers.

Follow up

- **Learning fair representations**, Zemel et al., 2013, JML
- **Fairness through awareness**, Dwork et al., 2012, ITCS (google 'dwork nytimes' for a non-mathematical overview)
- **Big Data's Disparate Impact**, Barocas and Selbst, 2016, California Law Review
- Solon Barocas, Moritz Hardt, Cathy O'Neill, Delip Rao, Sorelle Friedler
- mike@mike.place, @mikepqr, github.com/williamsmj/sentiment,
fastforwardlabs.com