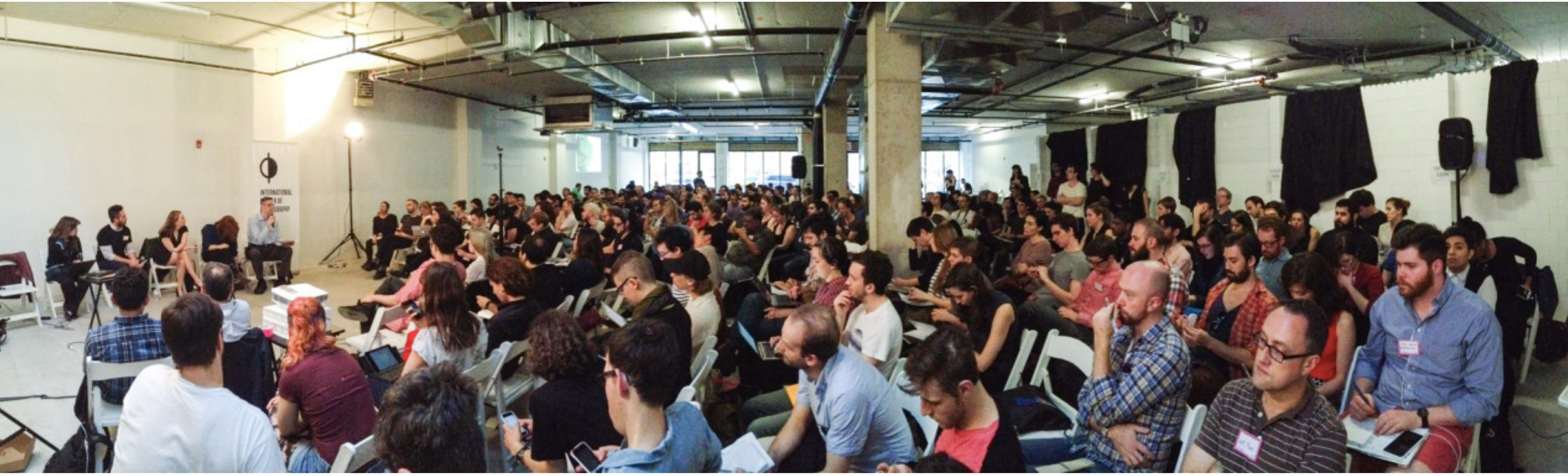
How we amplify privilege with supervised machine learning

Mike Williams
@mikepqr, mike.place
Fast Forward Labs, fastforwardlabs.com
github.com/williamsmj/sentiment

THEORIZING THE WEB

APRIL 17th & 18th, 2015
NEW YORK CITY



A Tool for
Inclusion or Exclusion?

Understanding the Issues

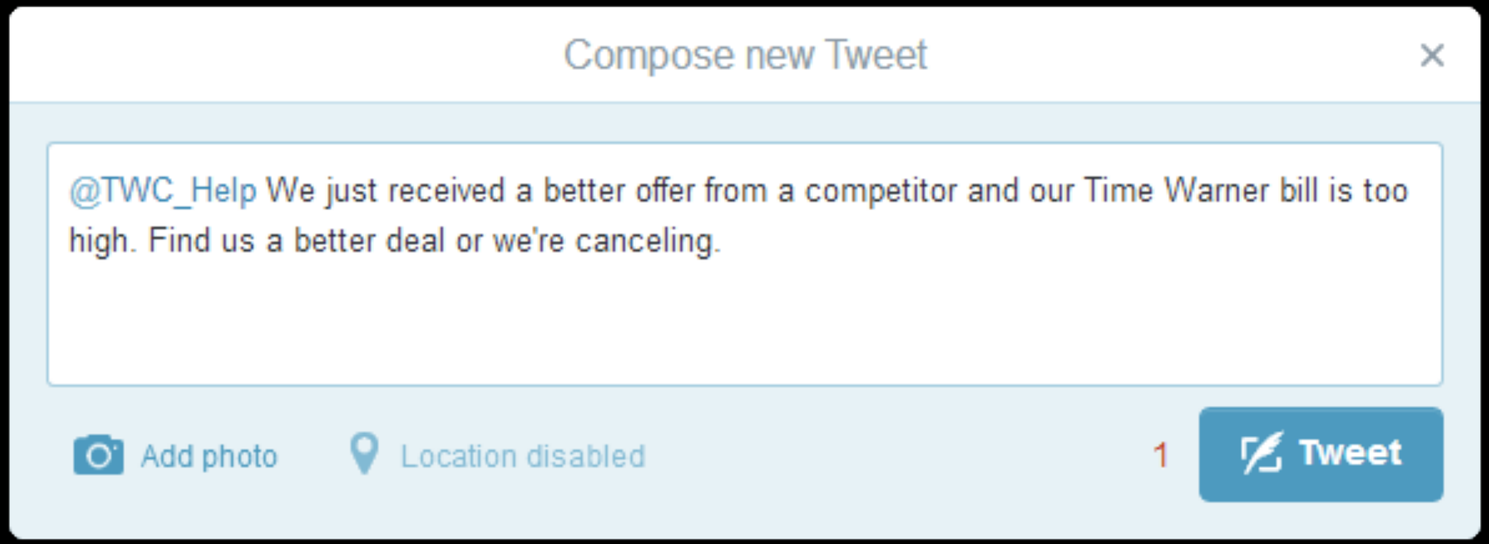
FTC Report

a specific example: sentiment analysis on social media

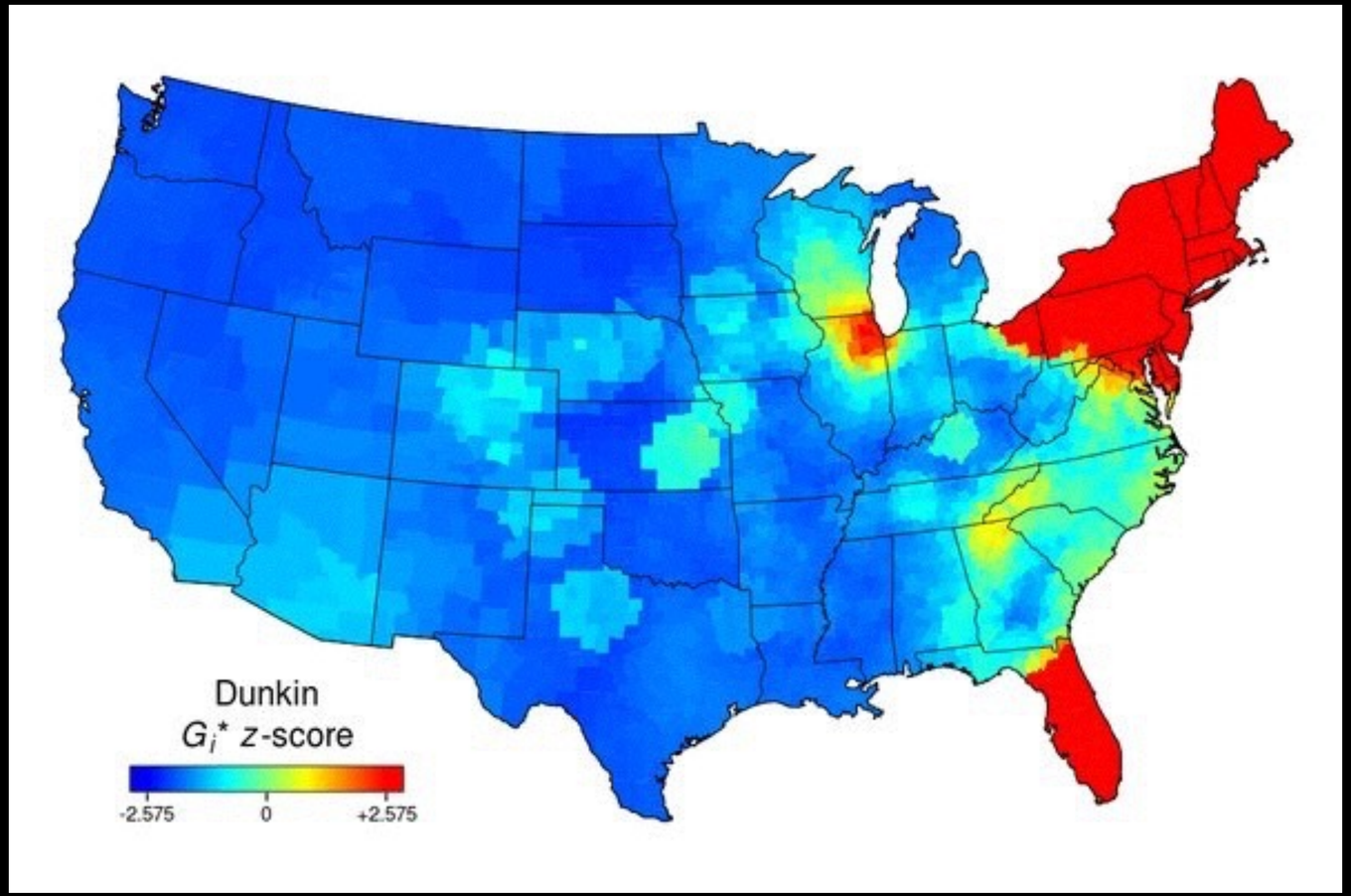
more general points about supervised machine learning

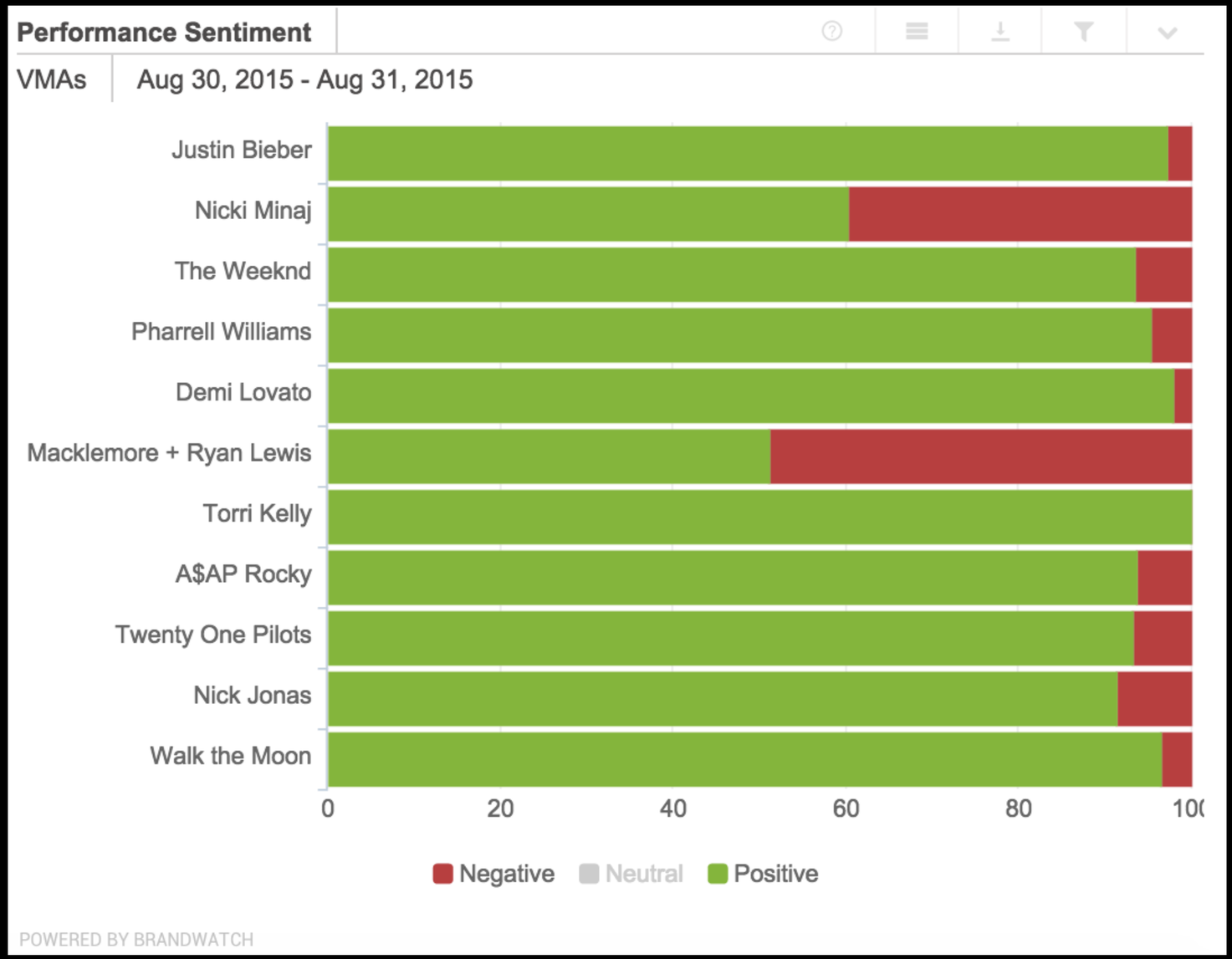
a specific example: sentiment analysis on social media

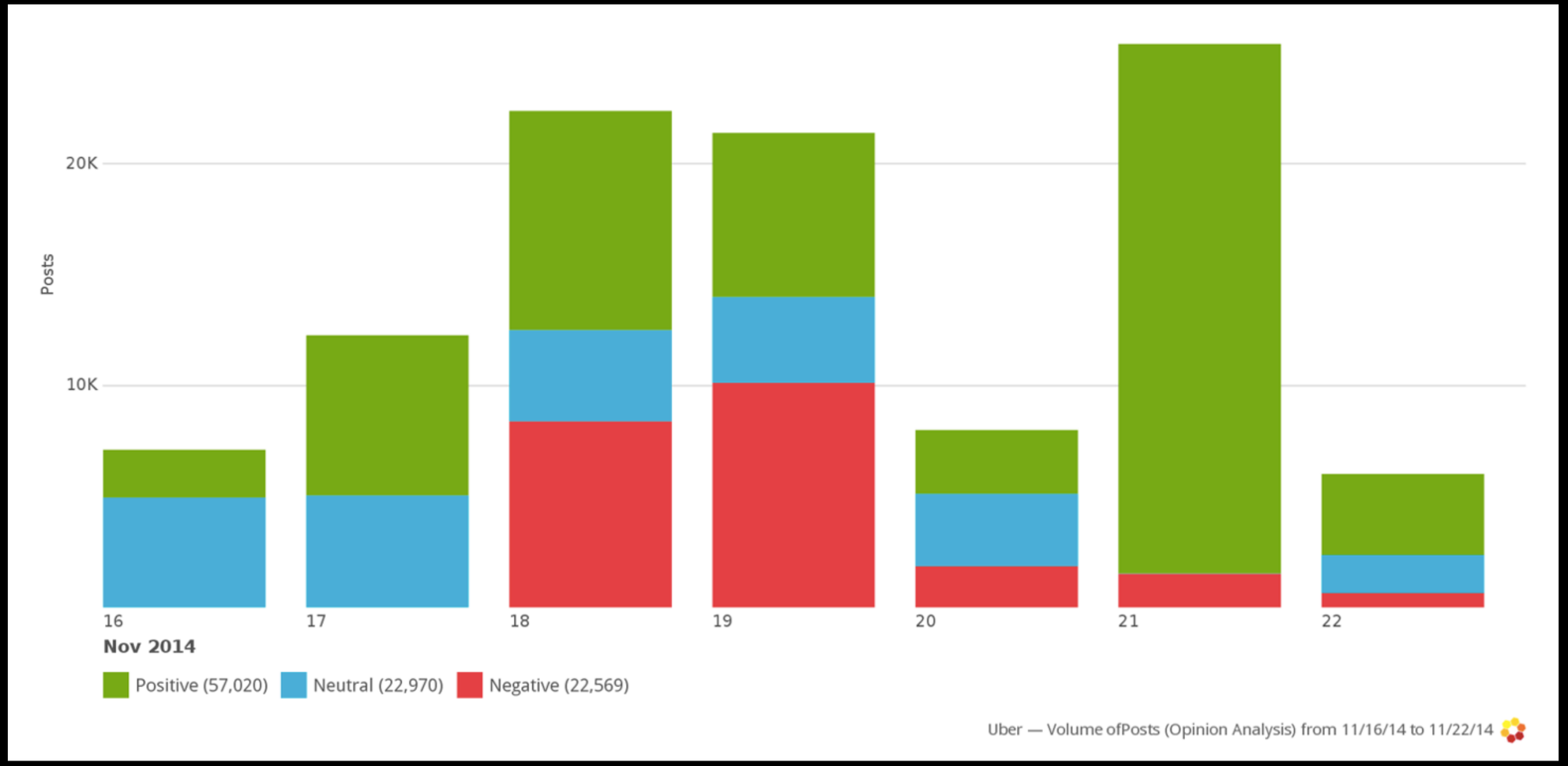
more general points about supervised machine learning


Compose new Tweet
@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
Tweet

isogloss.shinyapps.io/isogloss/



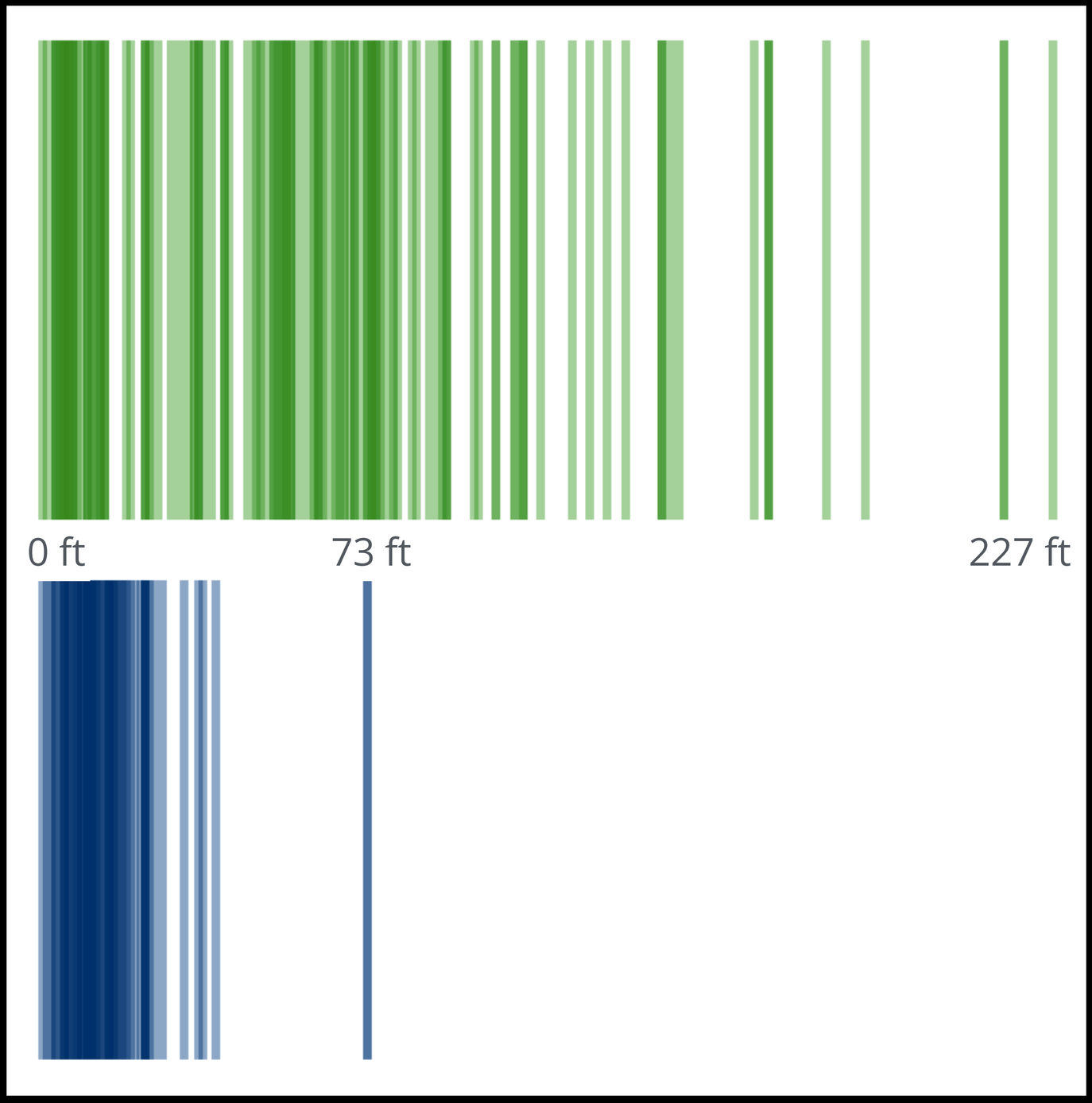

Performance Sentiment
VMAs
Aug 30, 2015 - Aug 31, 2015
Justin Bieber
Nicki Minaj
The Weeknd
Pharrell Williams
Demi Lovato
Macklemore + Ryan Lewis
Torri Kelly
A$AP Rocky
Twenty One Pilots
Nick Jonas
Walk the Moon
0
20
40
60
80
100
Negative
Neutral
Positive
POWERED BY BRANDWATCH

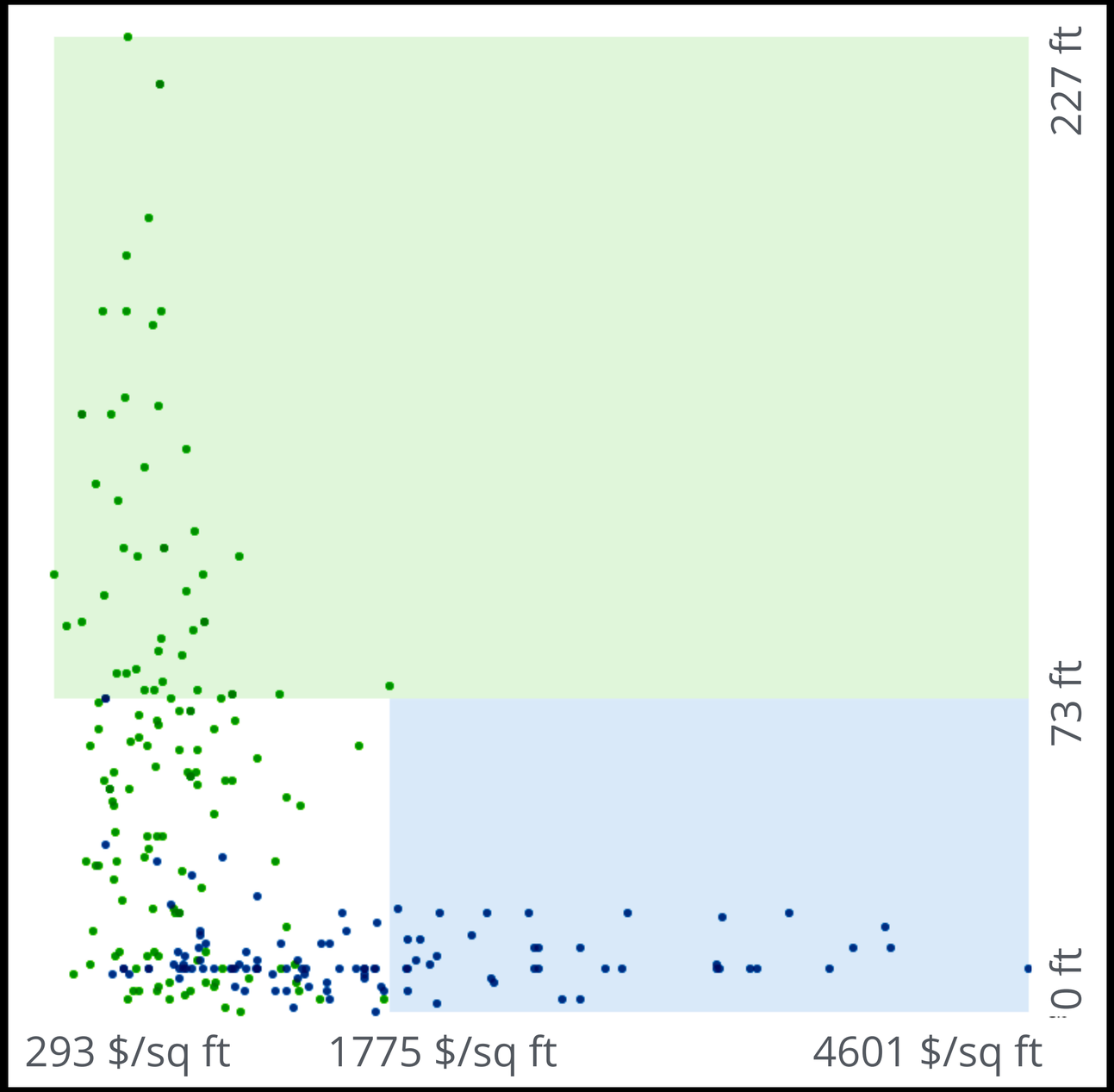


Crimson Hexagon

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


0 ft
73 ft
227 ft


227 ft
73 ft
0 ft
293 $/sq ft
1775 $/sq ft
4601 $/sq ft

	cat	sat	mat	dog	bites	man
the cat sat on the mat	1	1	1	0	0	0
cats and dogs	1	0	0	1	0	0
man bites dog	0	0	0	1	1	1
dog bites man	0	0	0	1	1	1

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-meetup

the classifier works!


windows_7
dick
2-0
feature
nfl
holy_shit
www
bastard
lebron
weapons
gun
political
modern_warfare
beard
fans
his_own
fifa
his_mind
gaming
victory
playoffs
government
chuck_norris
win
fucked
fuck
pc
my_wife
http://
master
religion
bass
fuck_yeah
espn
software
champions
hell
hit_me_up
black_ops
vs
world_cup
league
system
team
beat
youtube.com
shave
epic
his_phone
my_girlfriend
india
nation
destroyed
1-0
call_of_duty
fuckin
speed
war
cod
fucking
wishes
he
nba
ftw
god_damn
beer
nahi
wins
politics
shaved
thinks_he
xbox
championship
himself
360
nuclear
engineering
of_death
halo
creed
metal
ps3
teams
his_life
cock
album
legendary
arsenal
sword
match
shit
players
loves_his
brazil
legend
united
president
player
germany
Penn | World Well-Being Project | wwbp.org

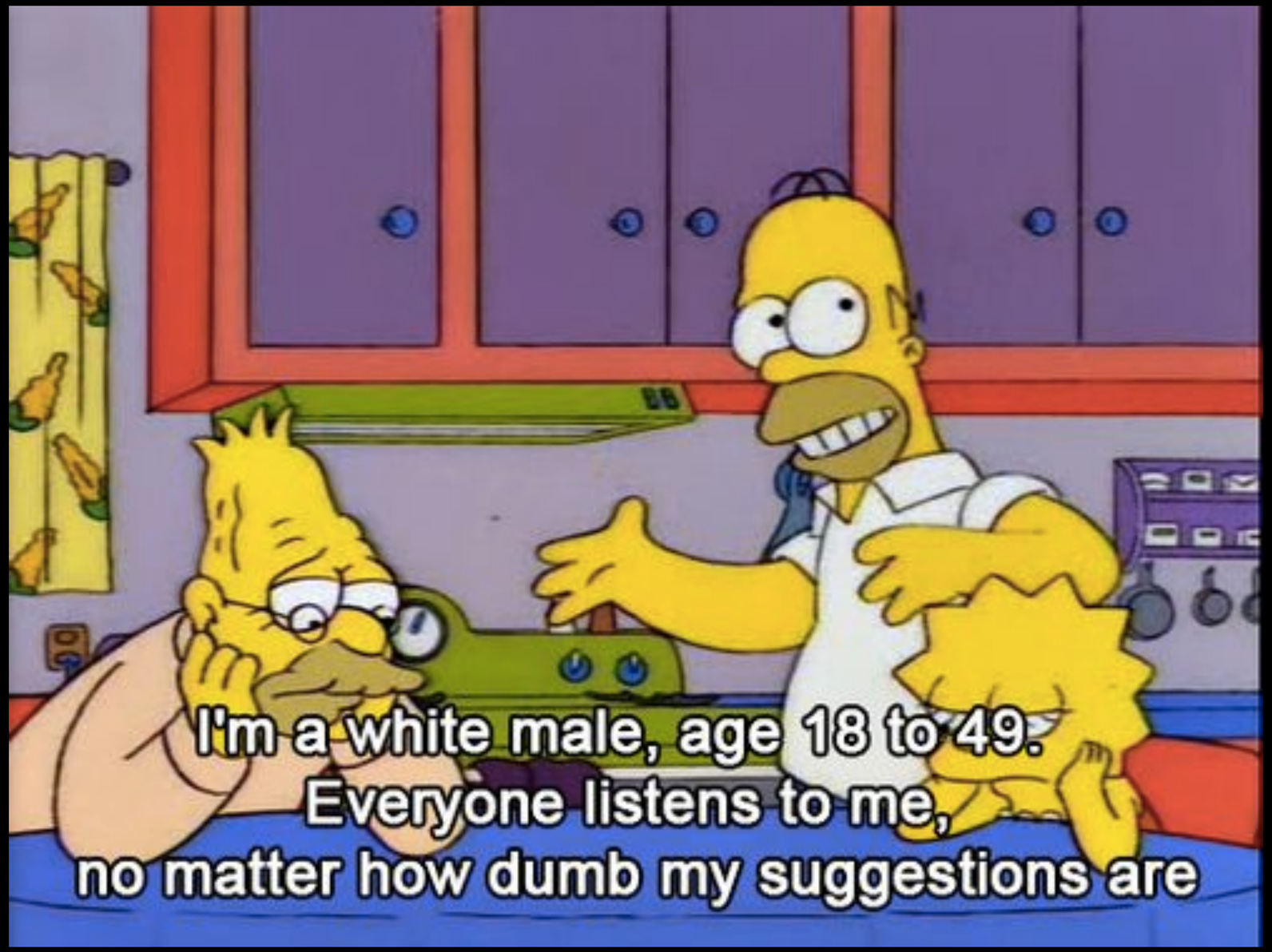
Male


bestie
she's
can't_believe
blessed
daughter
so_tired
mom
her_first
soooo
she
cry
loving
beautiful
princess
cute
yummy
hubby
her
loved
cake
:d_<3
wishes_she
so_much
sooooo
time_with
thank_you
lots
super_excited
husband
loves_her
girl
love_you
so_happy
soo
such_a
nails
dress
I_miss
sick
thankful
ugh
my_heart
am_so
love_him
wonderful
:(
excited
christmas
I'm_so
drama
my_little
to_see_my
gosh
proud_of
yum
tummy
love_you_guys
having_a
mommy
aunt
much_fun
my_hair
?_?
;)
shopping
birthday
smile
bed
herself
boyfriend
sad
<3
yay_!
like_it_on
:)
girls
can't_wait
lovely
her_new
make_me
so_glad
sooo
dinner
best_friend
sister
cookies
babysitting
omg
my_family
chocolate
love_them
baby
laugh
with_my
daddy
baking
go_away
adorable
babies
cleaning
Penn | World Well-Being Project | wwbp.org

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**


I'm a white male, age 18 to 49.
Everyone listens to me,
no matter how dumb my suggestions are

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

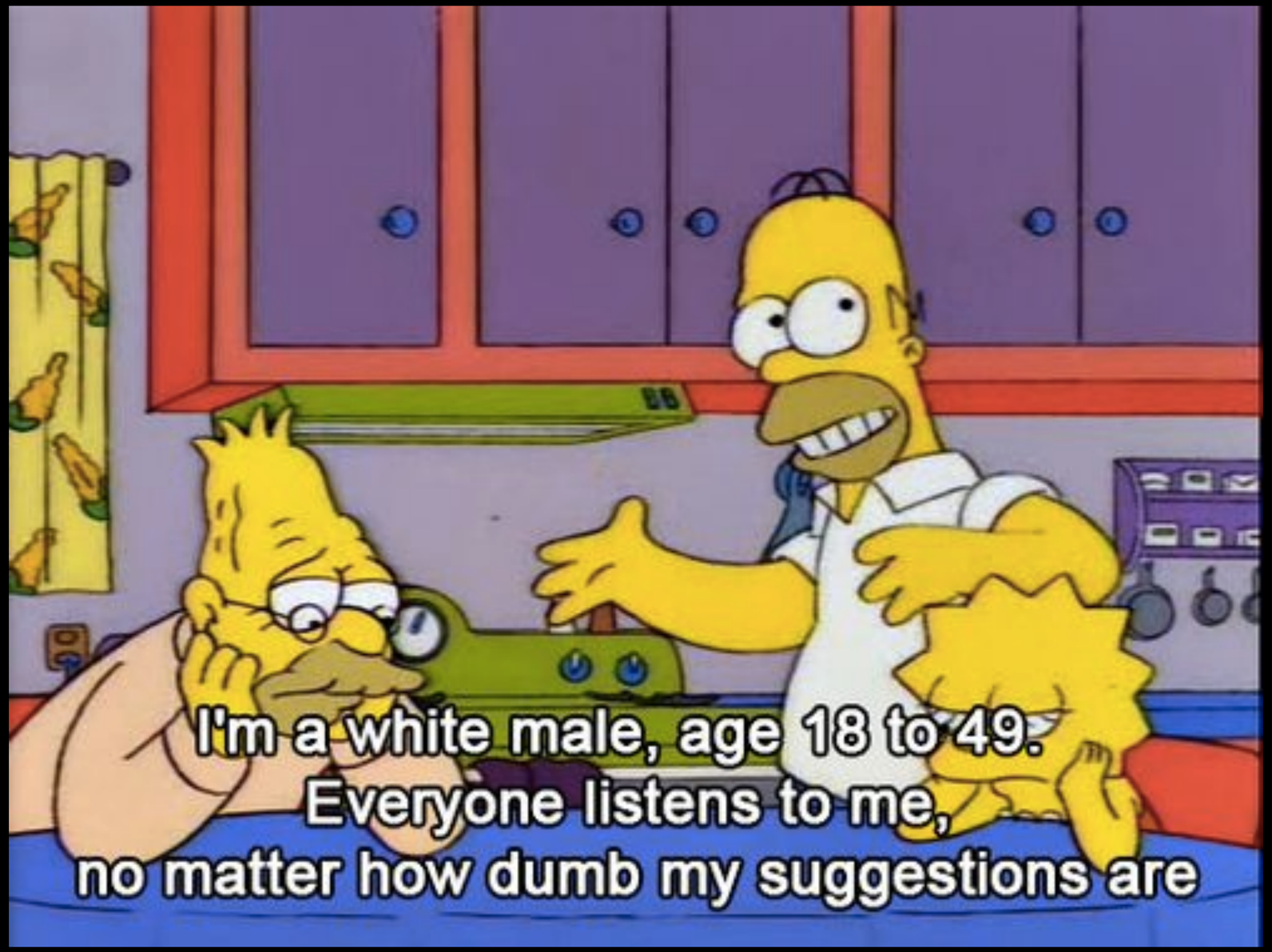
a specific example: sentiment analysis on social media

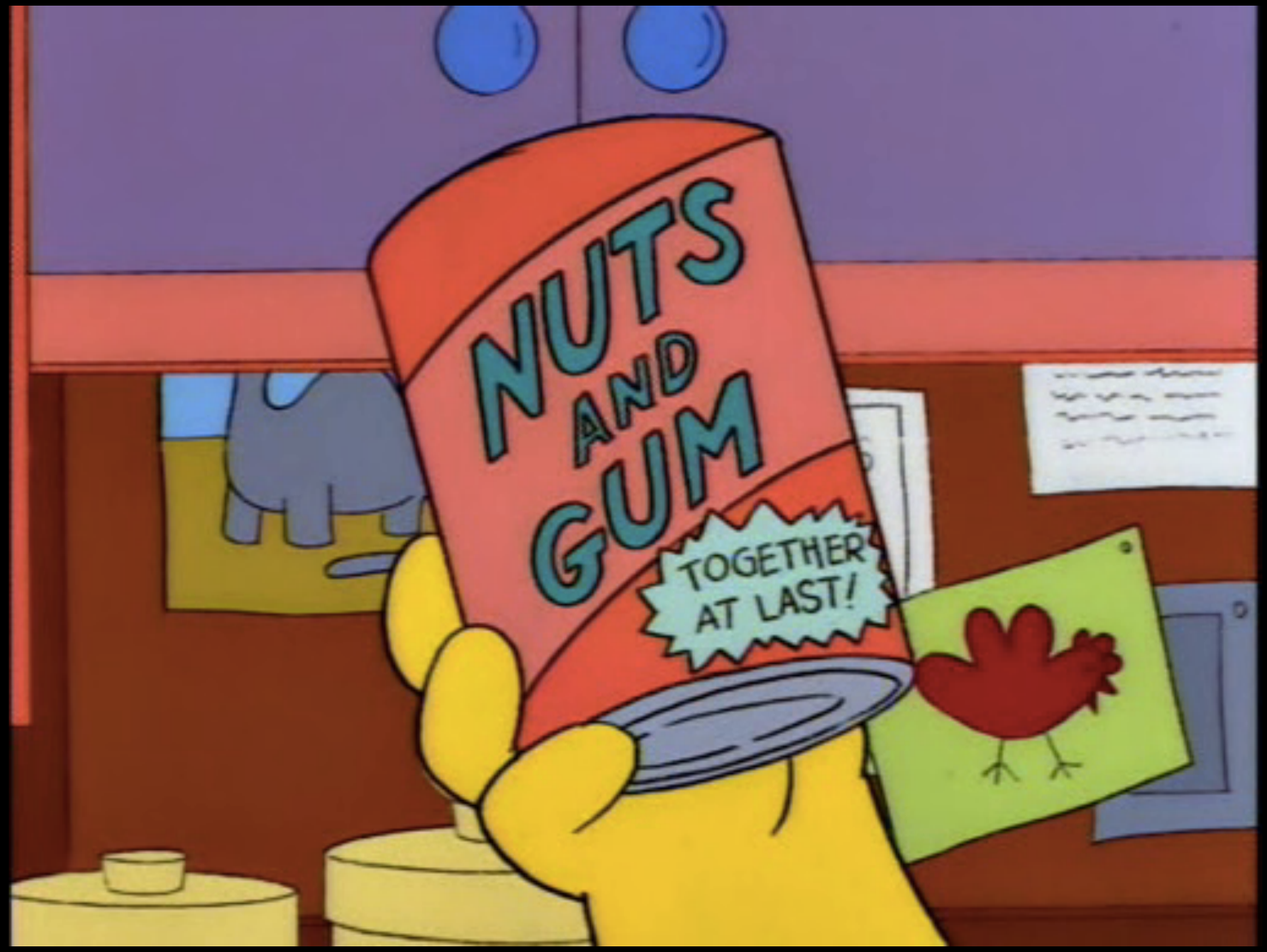
more general points about supervised machine learning

supervised machine learning
is an attempt to discover
which stereotypes are true

machine learning is full
of opportunities to be unfair

your choices have
consequences
ethical
legal
commercial


I'm a white male, age 18 to 49.
Everyone listens to me,
no matter how dumb my suggestions are


NUTS
AND
GUM
TOGETHER
AT LAST!

target variable


Credit Score
Excellent
Very Good
Good
Average
Poor
Income
$ 5,125.00

training data:
training sample bias
recapitulating historical bias

proxy features

Study Strongly Links Baltimore Mortgage Denials to Race

By PETER EAVIS NOV. 16, 2015

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

missing features

the danger of the
perfect model

what do we mean by fair?

individual fairness

group fairness

- 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.

objectivity is neither
achievable nor fair

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
- **The New Jim Crow**, Alexander, 2010 (especially Chapter 4)
- mike@mike.place, @mikepqr, github.com/williamsmj/sentiment, fastforwardlabs.com