Bias and ethics in supervised machine learning

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

Fast Forward Labs

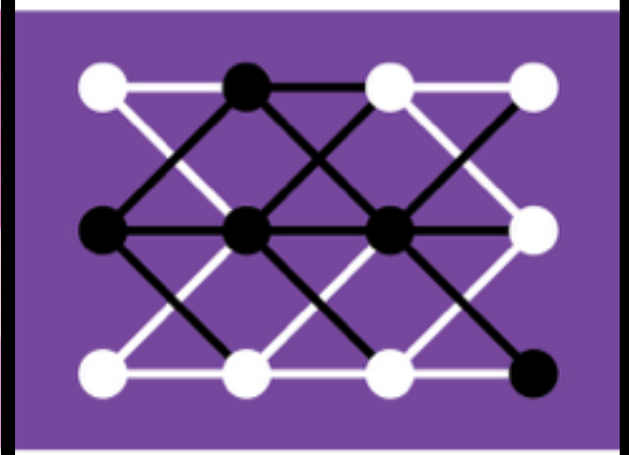
Natural Language Generation

Fast Forward Labs

Probabilistic Methods for Realtime Streams

Fast Forward Labs

Deep Learning: Image Analysis

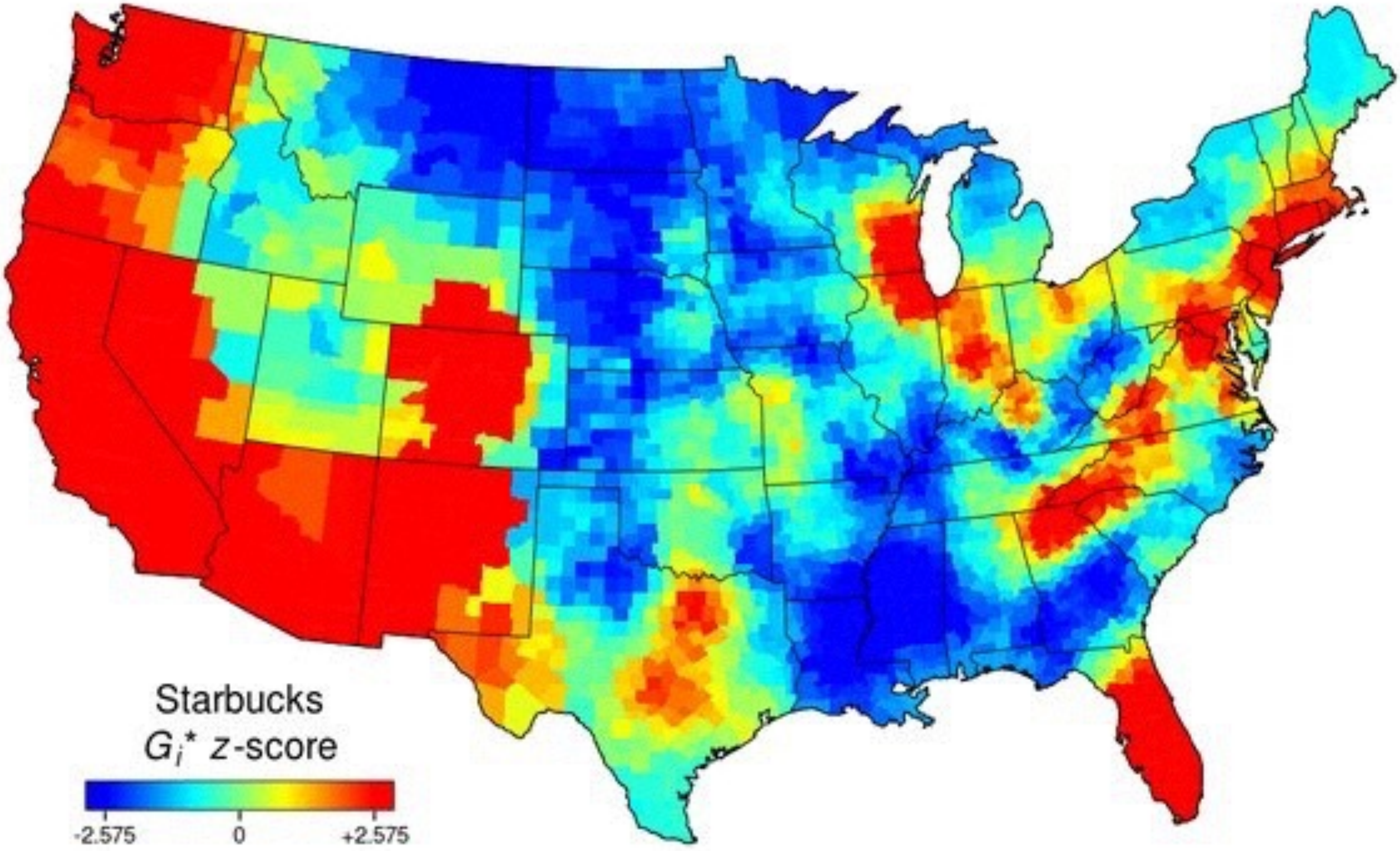


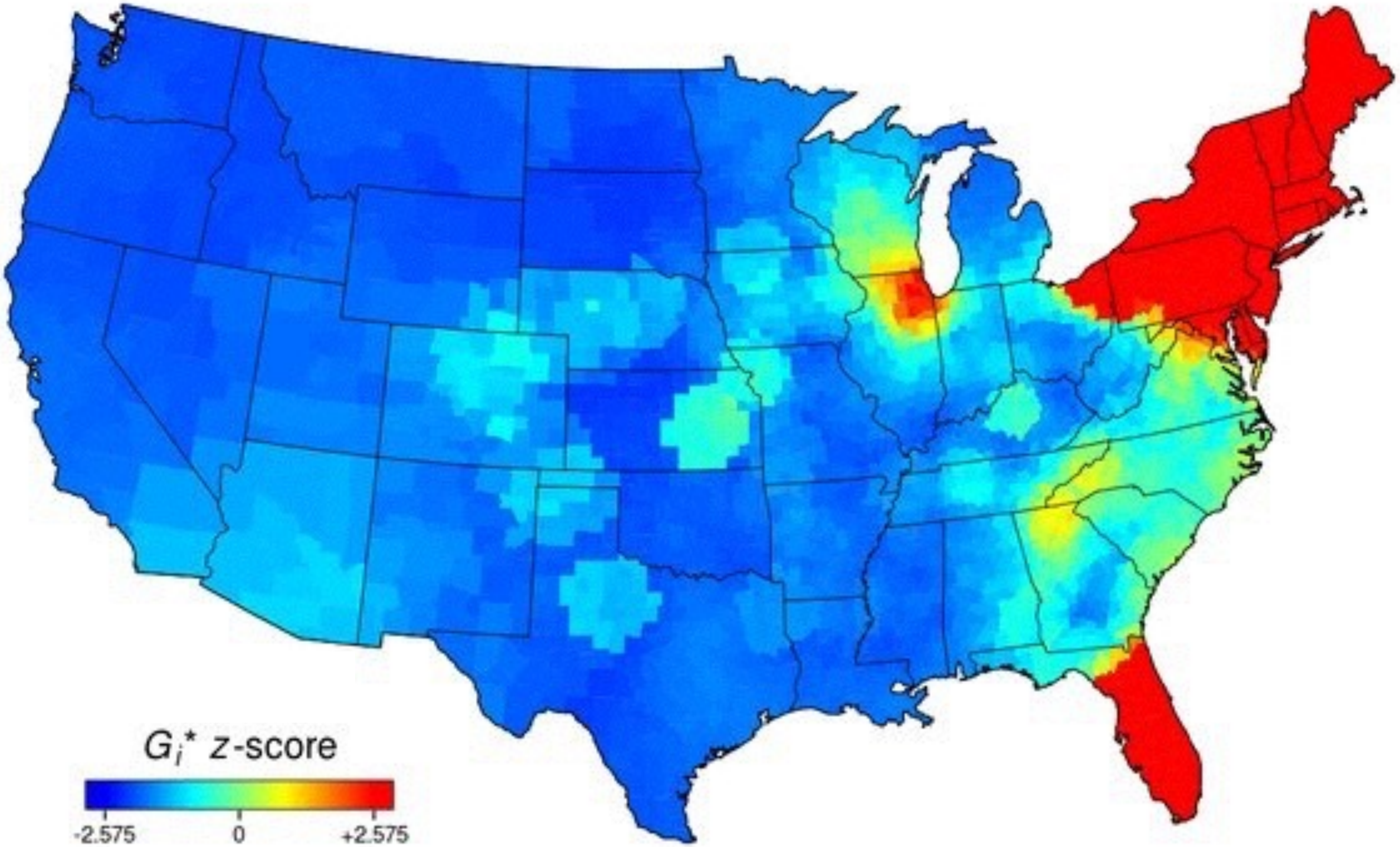
Fast Forward Labs

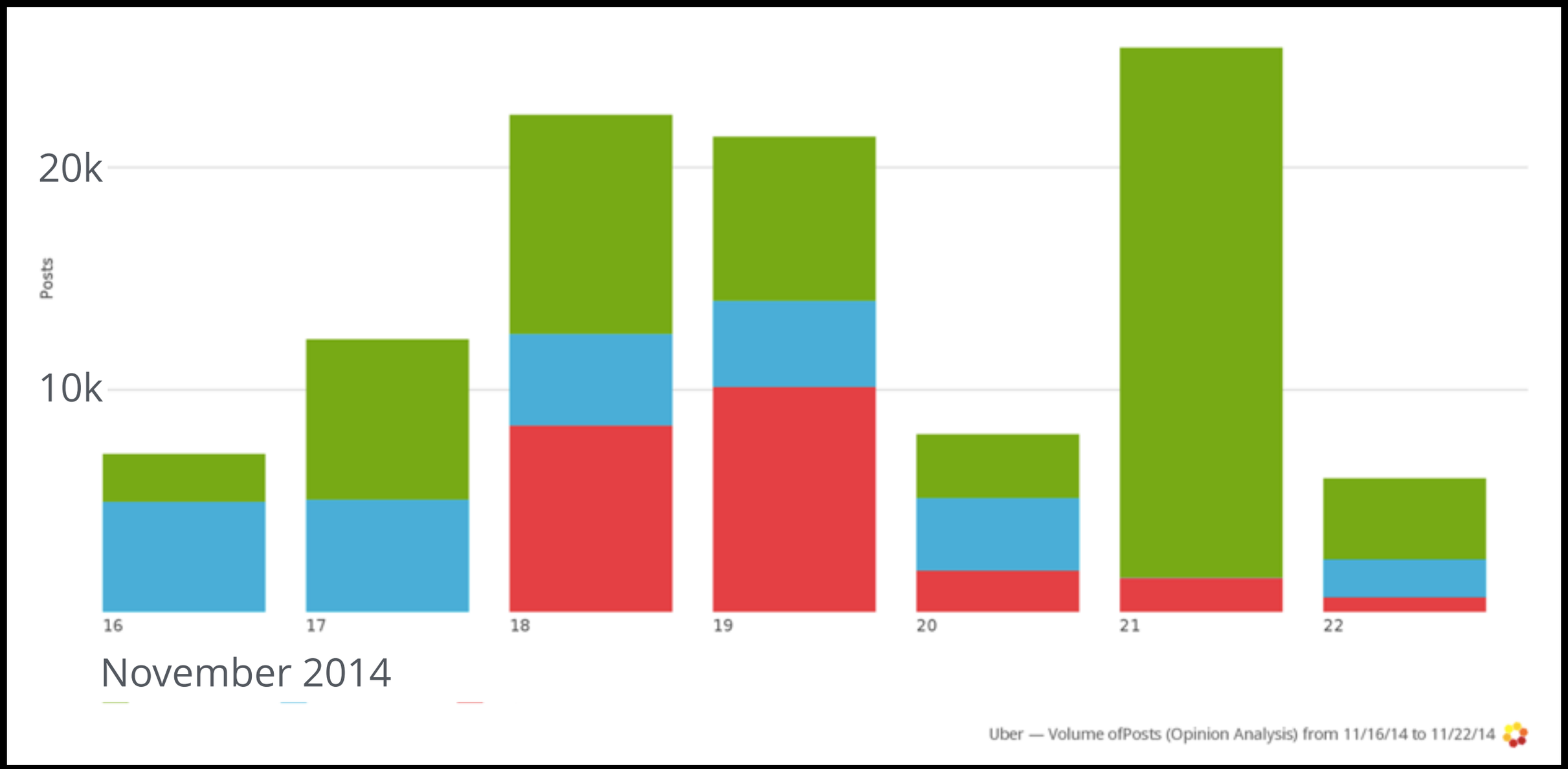
Summarization

a specific example: sentiment analysis on social media

more general points about supervised machine learning


Starbucks
G_i^* z-score
-2.575
0
+2.575


G_i^* z-score
-2.575
0
+2.575



Crimson Hexagon

r2d3.us

227 ft

73 ft

0 ft

293 $/sq ft

1775 $/sq ft

4601 $/sq ft

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

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**

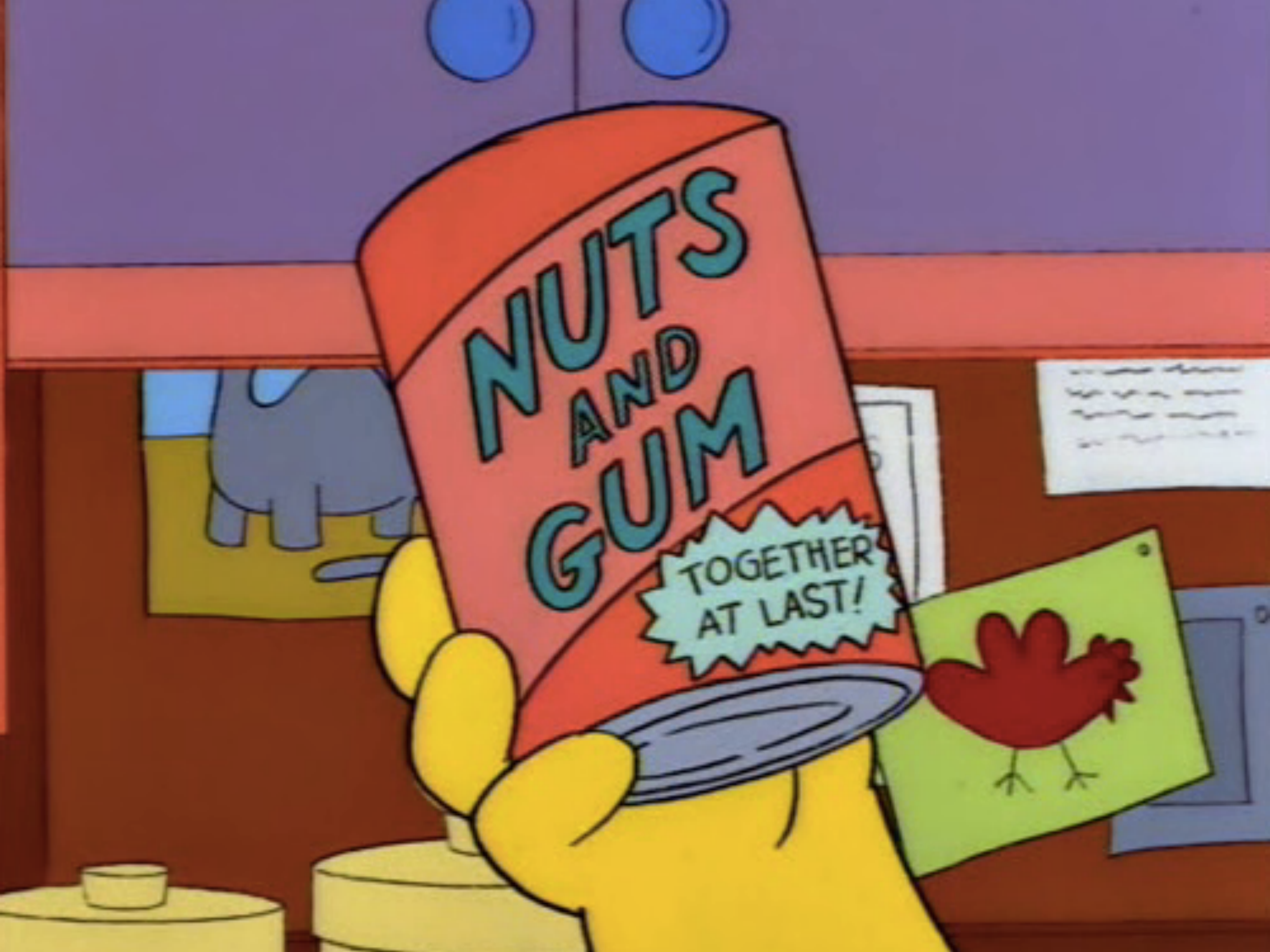

I'm a white male, age 18 to 49.
Everyone listens to me,
no matter how dumb my suggestions are

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 🚨


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


NUTS
AND
GU
TOGETHER
AT LAST!

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

- 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

- 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

Follow up

- **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