Modeling communities from their social media

Lyle Ungar

wwbp.org

University of Pennsylvania

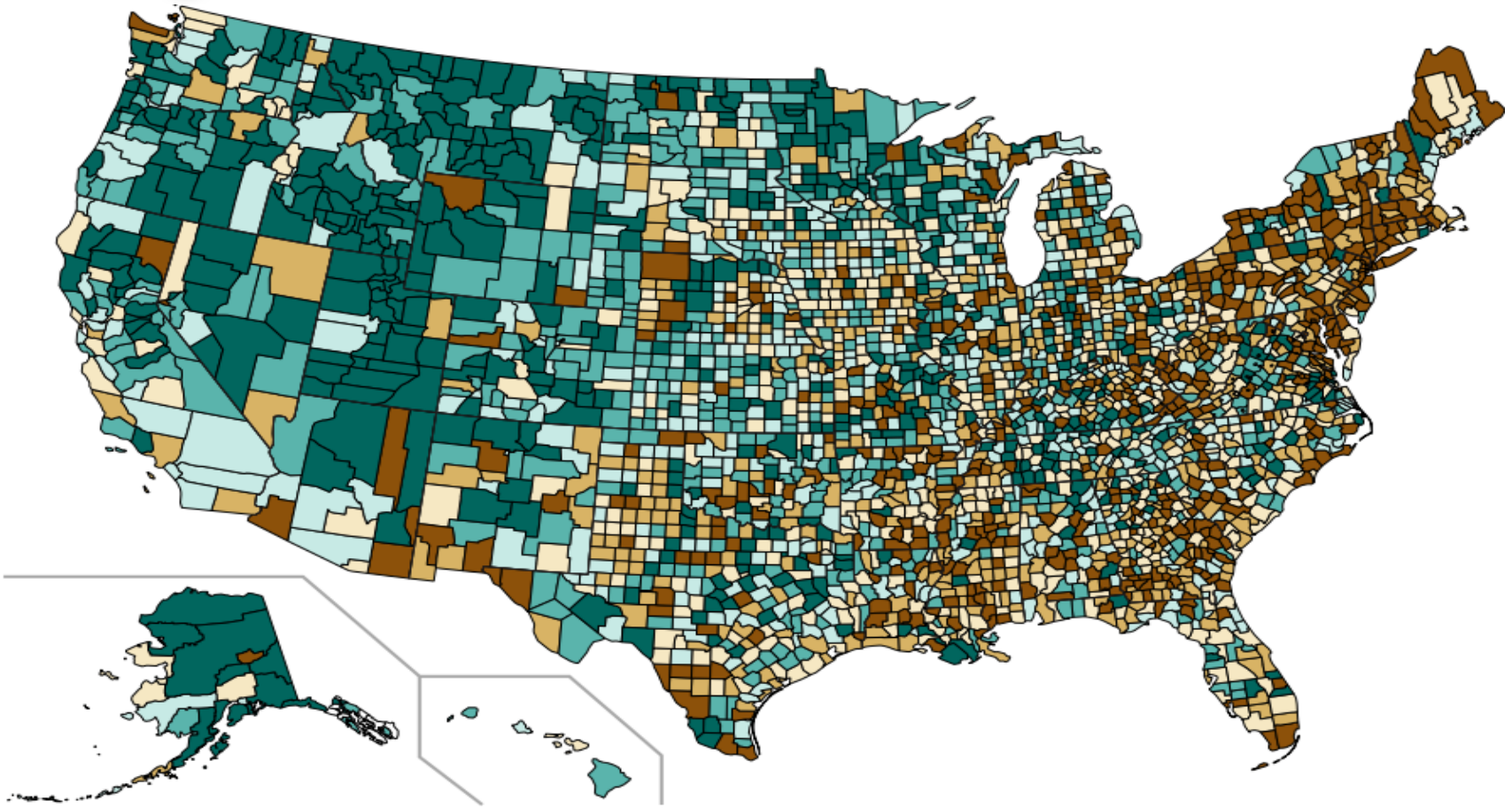
Community-level modeling

- **Open vocabulary**
  - Correlate word or LDA topic use with community sentiment
- **Model-based**
  - Build language models of sentiment
    - at the tweet or individual level
  - Apply them to new tweets or people
    - Group by community
  - (Sometimes) correlate with outcome

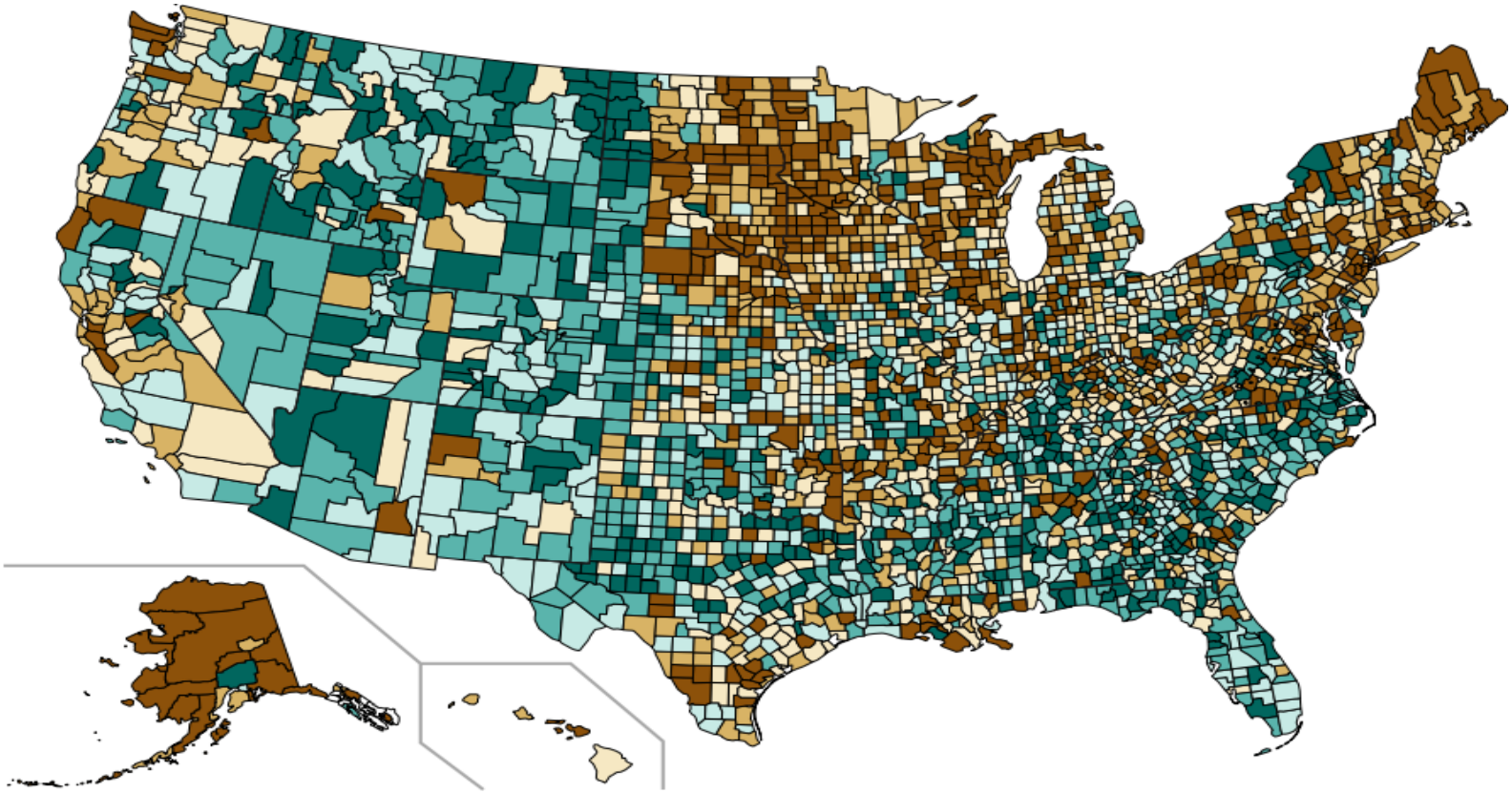
Measuring personality in tweets

- Create regression models for personality
- Use these to estimate scores on tweets collected on the county level
  - Agreeableness
  - Conscientiousness
  - Extroversion
  - Neuroticism
  - Openness to experience

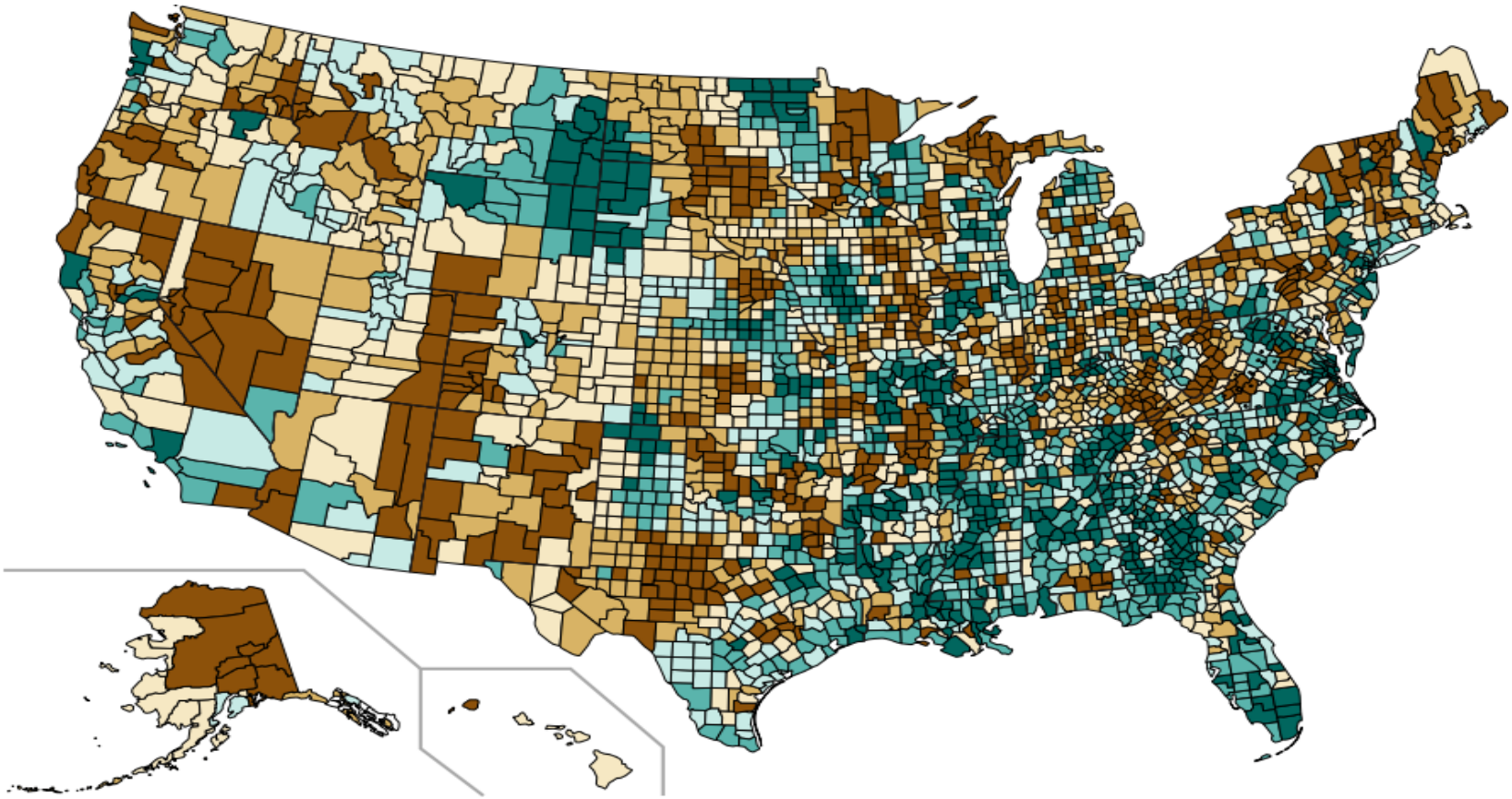
Agreeableness



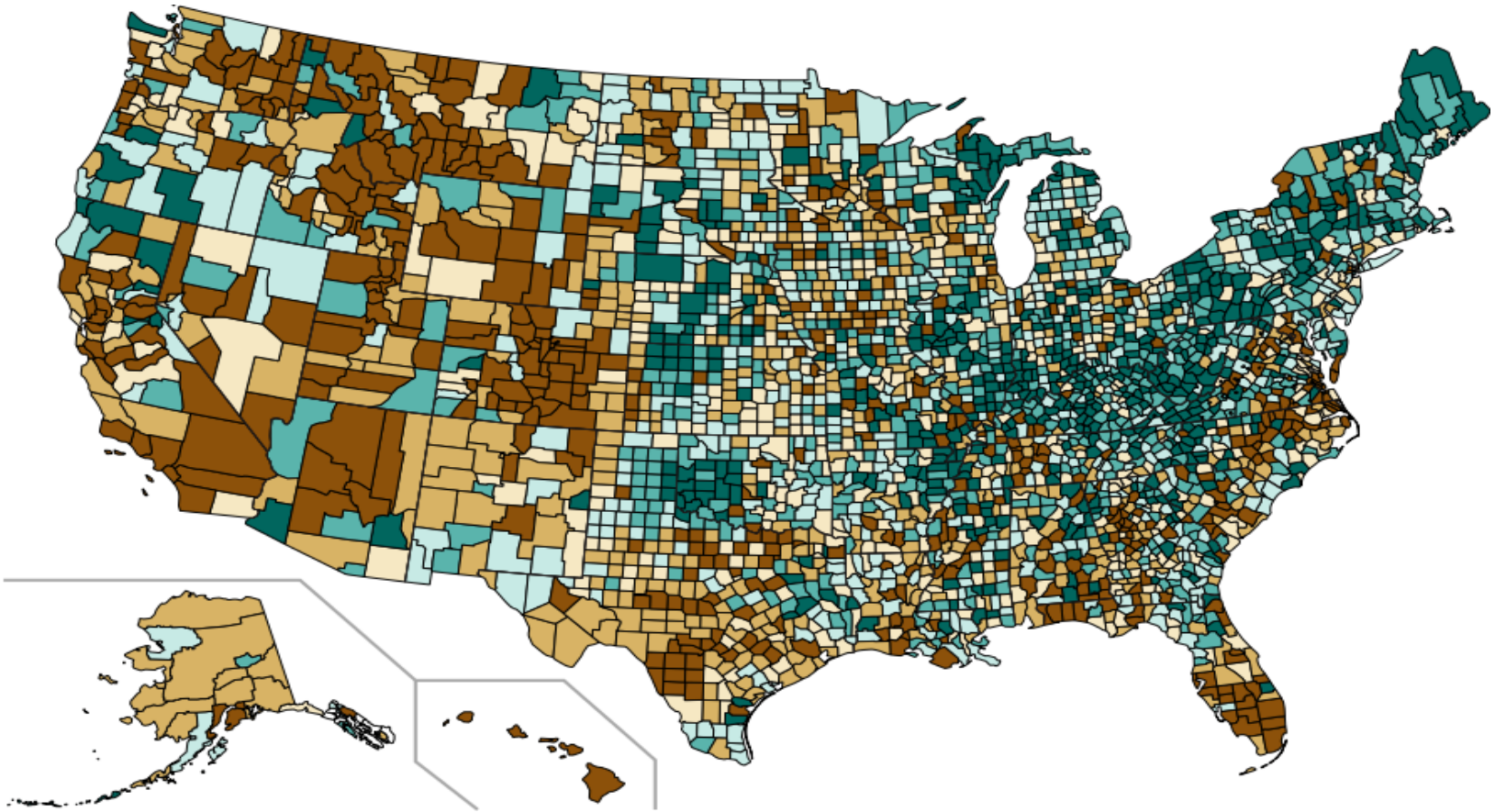
Conscientiousness



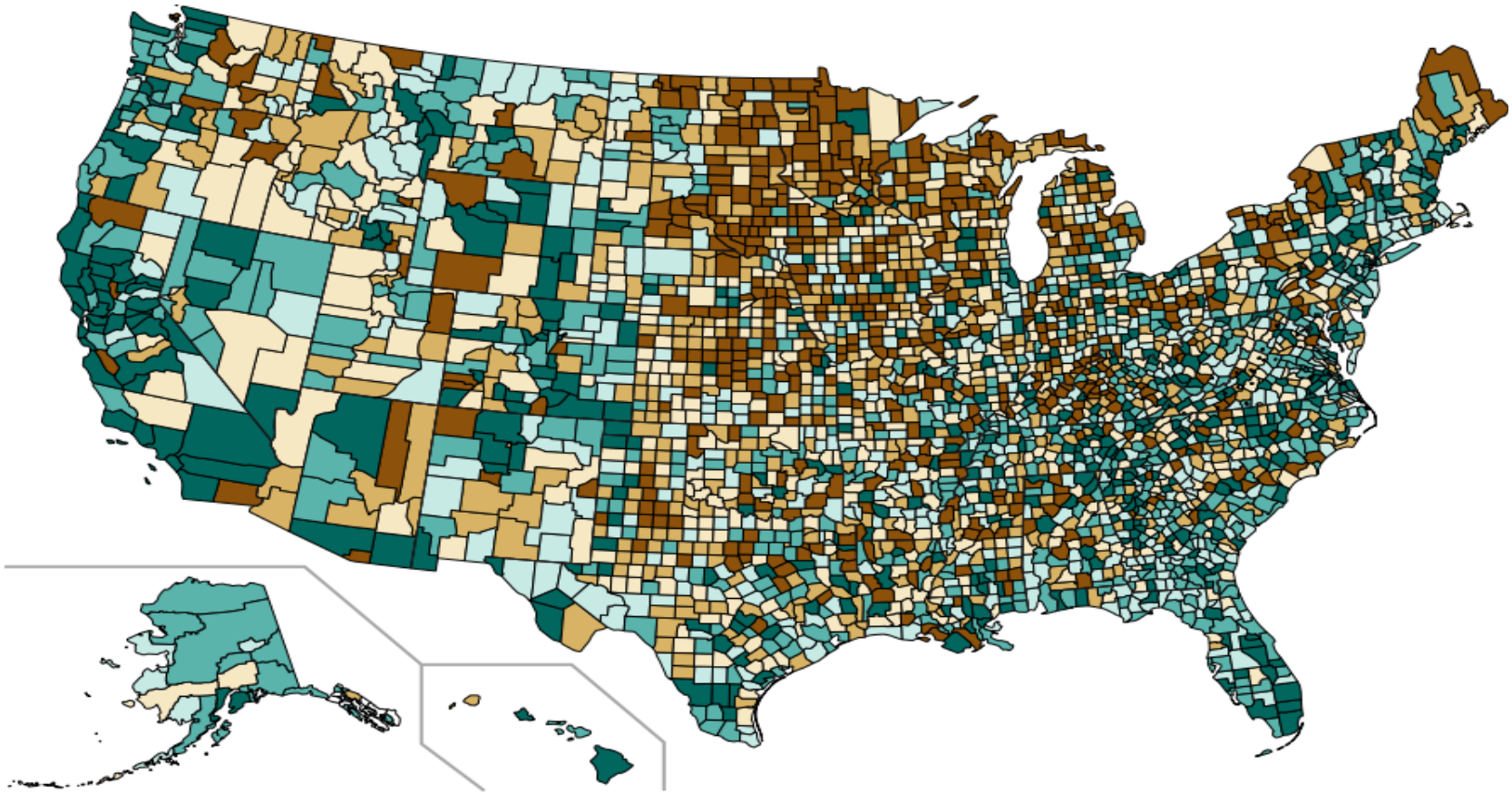
Extraversion



Neuroticism



Openness to experience



Community Well-Being and Heart Disease

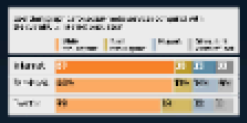
twitter

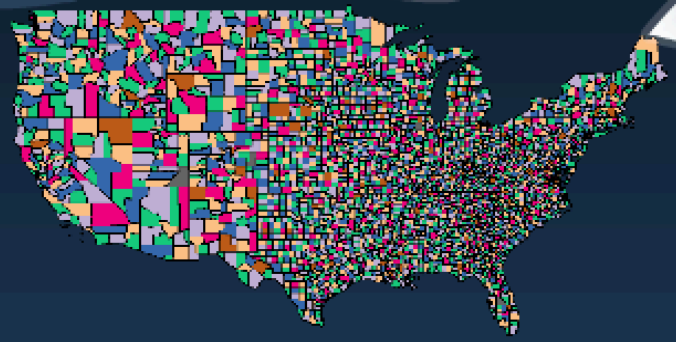
(public)

CDC

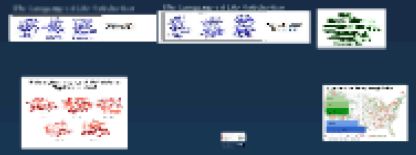
CENTERS FOR DISEASE CONTROL AND PREVENTION

U.S. DEPARTMENT OF COMMERCE · BUREAU OF THE CENSUS





Life Satisfaction



Heart Disease

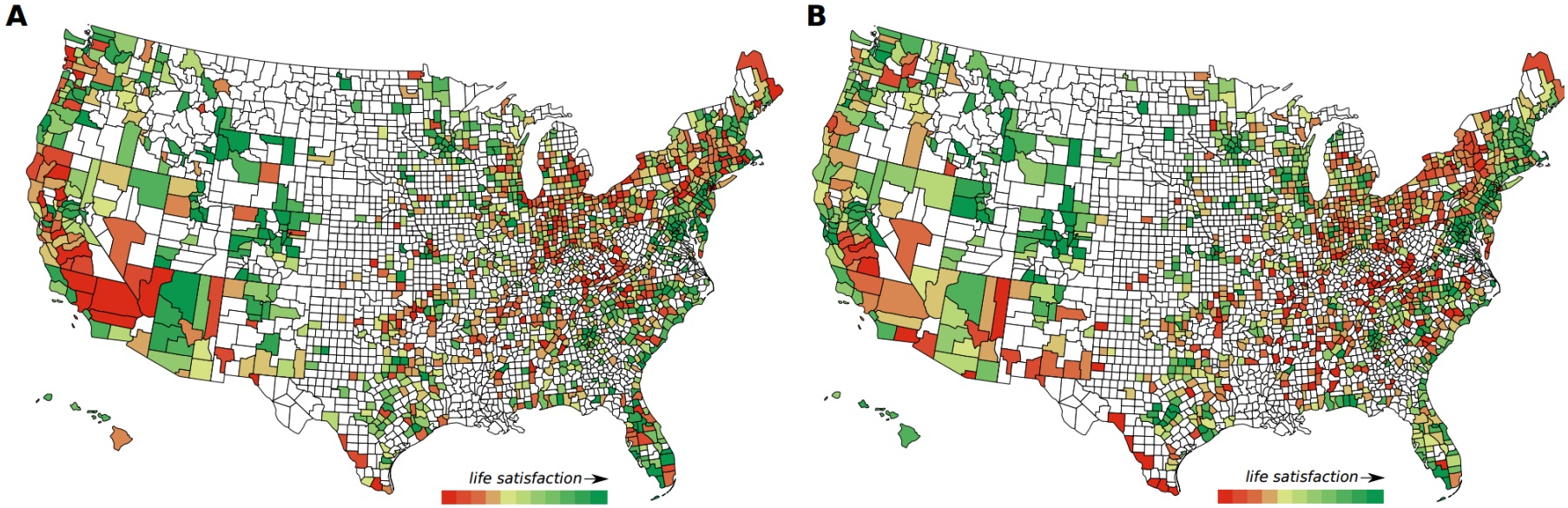
Satisfaction with Life Prediction

data	*r*
Lexica	0.264
Topics	0.307
Topics & Lexica	0.307
Controls	0.435
Controls, Topics & Lexica	**0.535**

Controls
- **median age**
- **sex** (percentage female)
- **minorities** (percentage black and Hispanic).
- **median household income** (log-transformed)
- **educational attainment** (% high school grads or higher; % BS or higher).

county-level predictions

Satisfaction with Life



survey

predicted

Life Satisfaction Words

z, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Lucas, R. E., Agrawal, M., Park, G. J., Lakshmikanth, S. K., Jha, S., Seligman, M. E. gar, L. H. (2013). Characterizing Geographic Variation in Well-Being using Tweets. In *Proceedings of the Seventh International AAAI nce on Weblogs and Social Media (ICWSM)*. Boston, MA.

Community Well-Being and Heart Disease

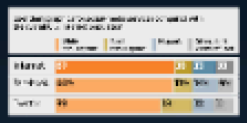
twitter

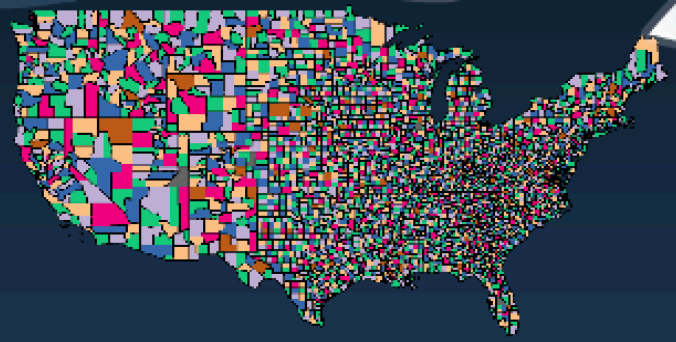
(public)

CDC

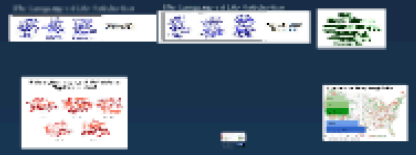
CENTERS FOR DISEASE CONTROL AND PREVENTION

U.S. DEPARTMENT OF COMMERCE · BUREAU OF THE CENSUS



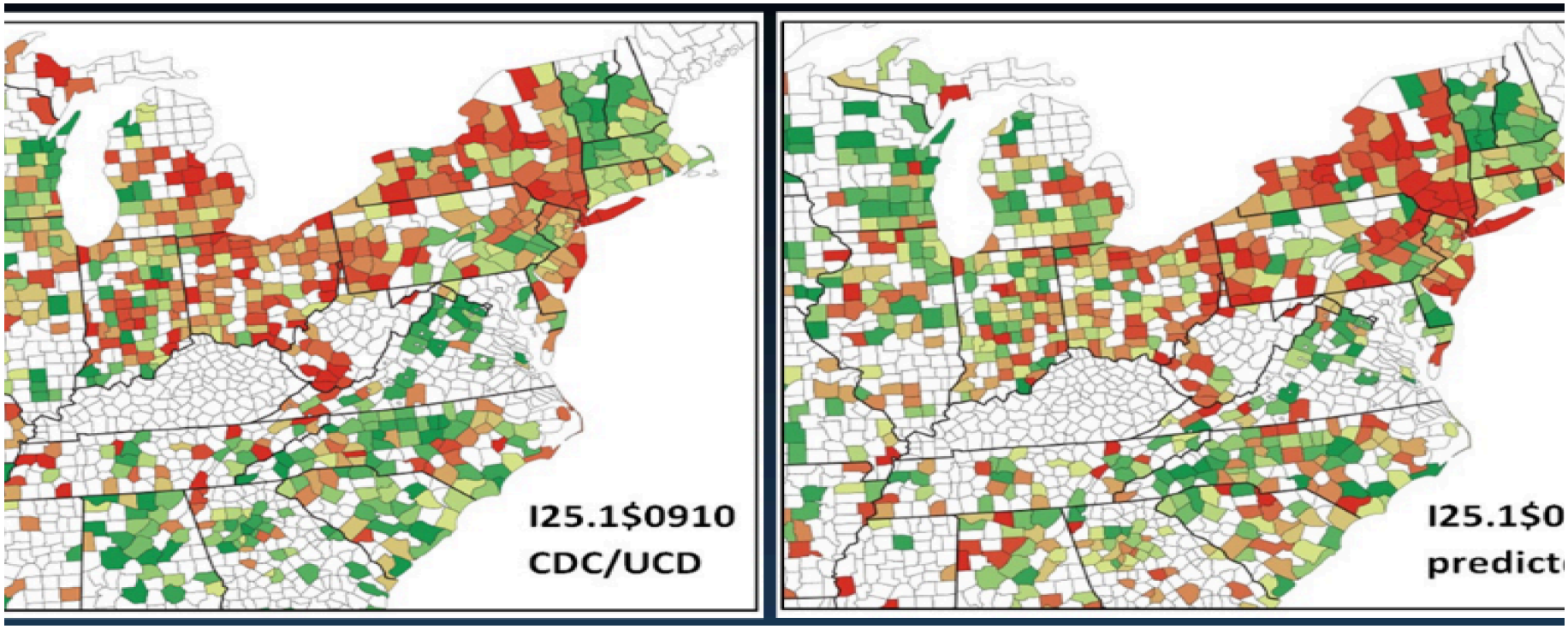


Life Satisfaction

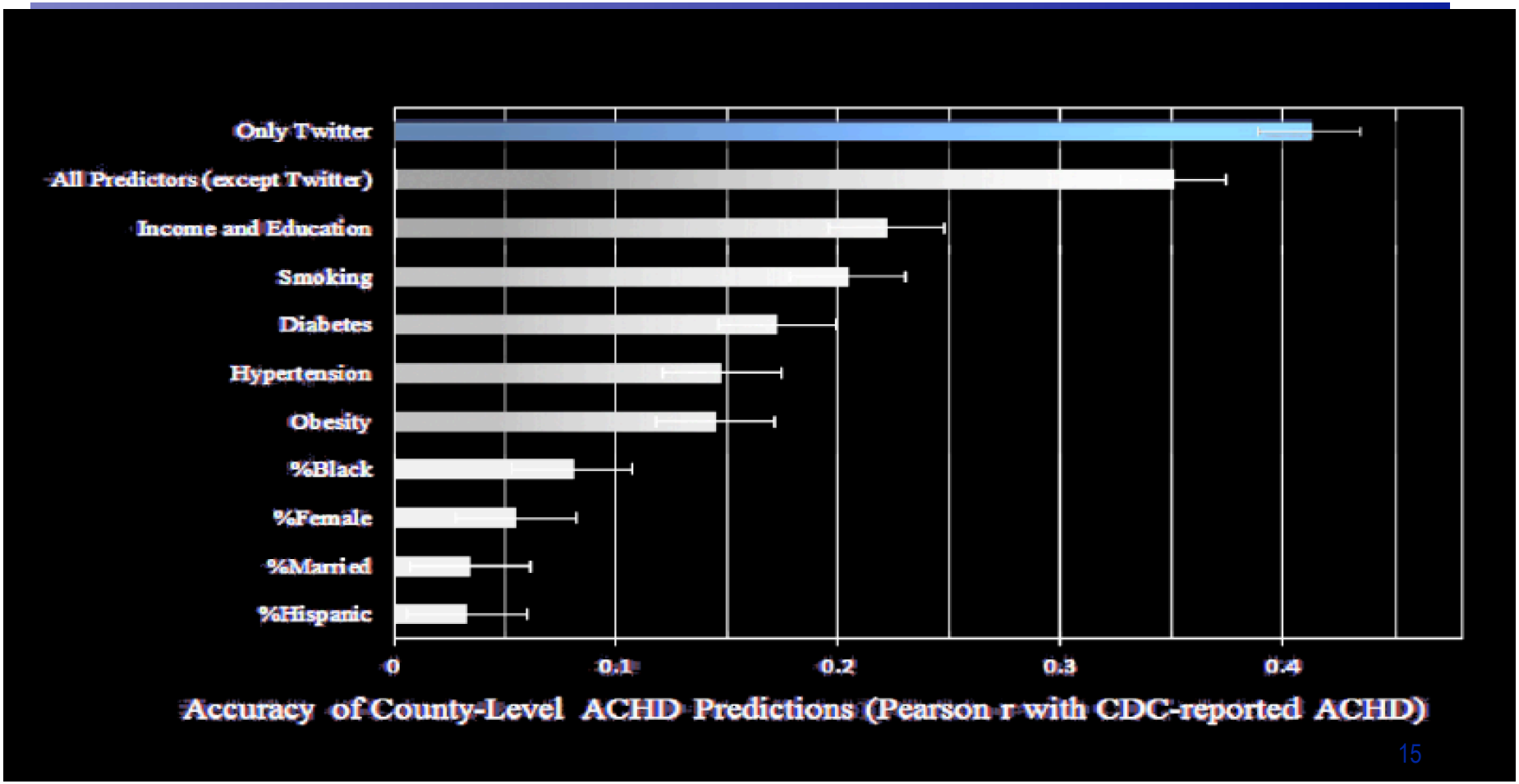


Heart Disease

Twitter predicts cardiovascular disease



Twitter predicts cardiovascular disease



Cardiovascular Disease Words

Higher Status Occupations

Negative Relationships

Disengagement

Person & Community Conclusions

- **Language reveals demographics, personality**
  - Also psychopathy, emotion, SES, political orientation, happiness, depression, health, disease
- **Language models generalize**
  - Across individuals
  - From Facebook to Twitter
  - From Individuals to Counties
- **Language allows data-driven hypothesis generation**
  - As well as hypothesis testing