Measuring Neighborhood Sentiment: **Leveraging Twitter Data**

Joseph Gibbons

Department of Sociology
HDMA: The Center for Human Dynamics in the Mobile Age
San Diego State University

March 23rd, 2018

HDMA @SDSU | THE CENTER FOR HUMAN DYNAMICS IN THE MOBILE AGE

Neighborhood Study is Limited

- Typically confined to survey data
  - Available mostly cross-sectionally
  - Limited spatial application
  - Arbitrary character of neighborhood boundaries (MAUP)
  - Obtrusive – respondents may not answer honestly
  - Sampling issues

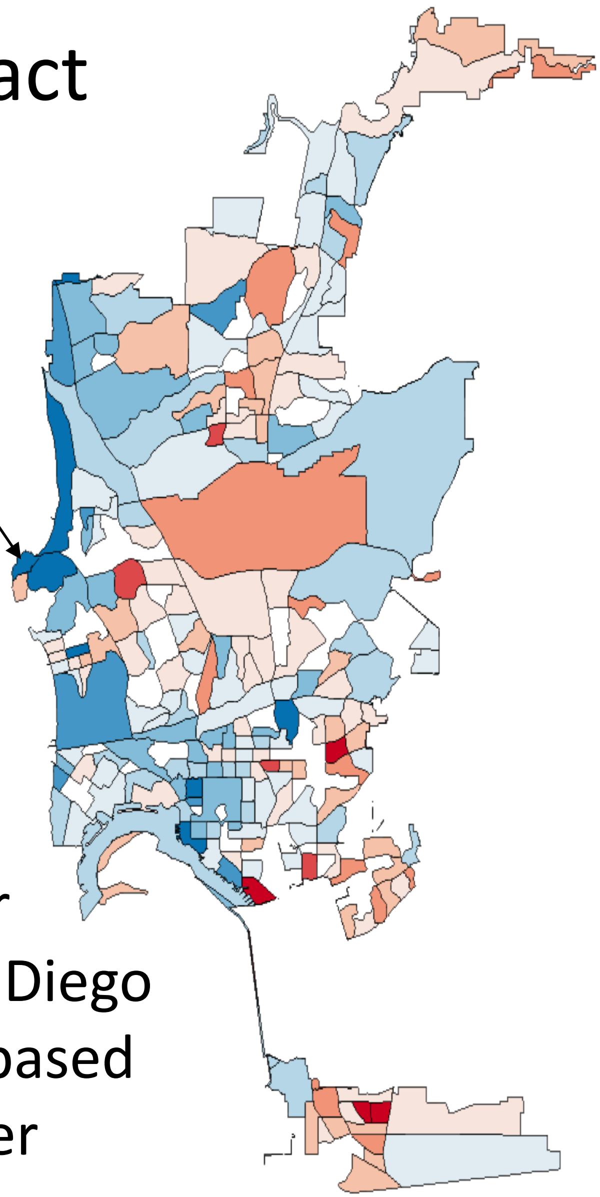
Benefits of Big Data for Neighborhood Study

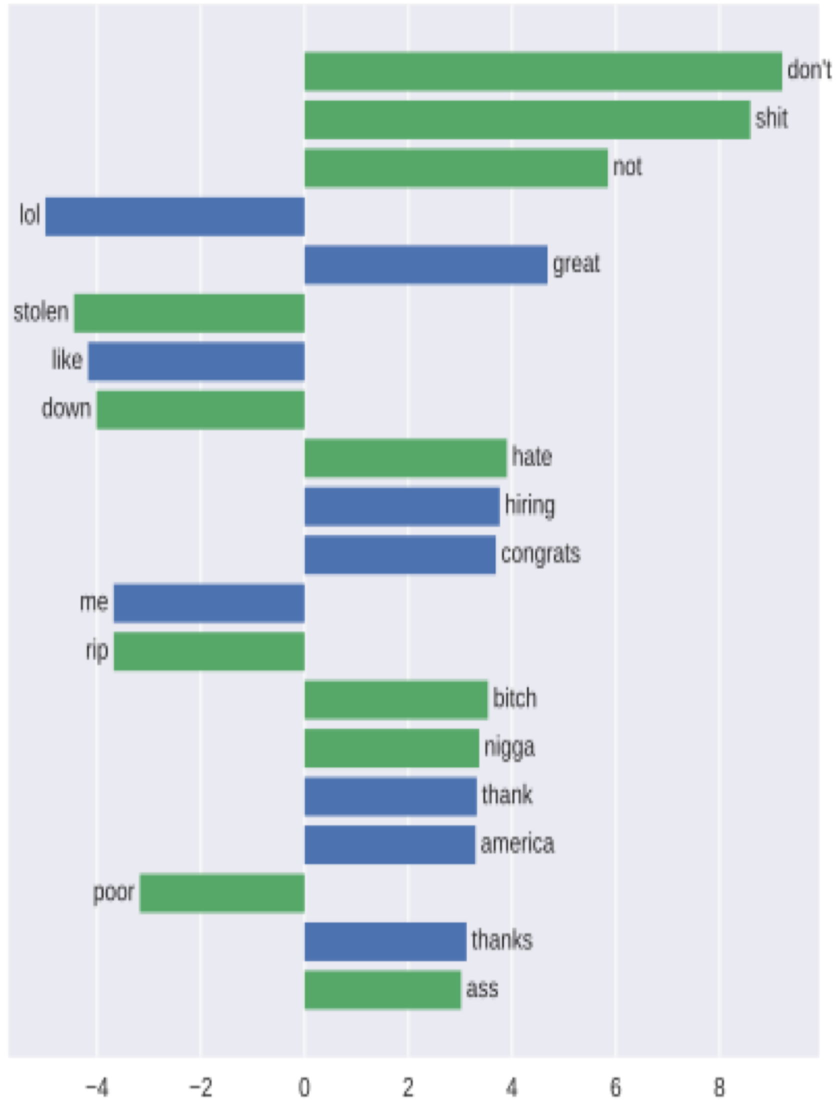
- Using social media as a way to identify social trends that supplement existing demographic measures
- Uses of social media
  - Dynamic
    - Enables a close examination real time population characteristics
  - Unobtrusive
    - Does not require direct interaction with populations that can affect results
  - Nuances
    - Allows both a more subtle look at the attitudes and conditions of local residents

Example of Sentiment Measure: Hedomometer

- ‘Happiness’ explored through Hedonometer (Dodds et al., 2011)
- 10,000 words evaluated based on algorithm and human ranking
  - rated by MTurkers
- The overall score indicates how ‘happy’ or ‘sad’ a neighborhood is in practice
- 10= Happy, 5= Neutral, 1= Sad

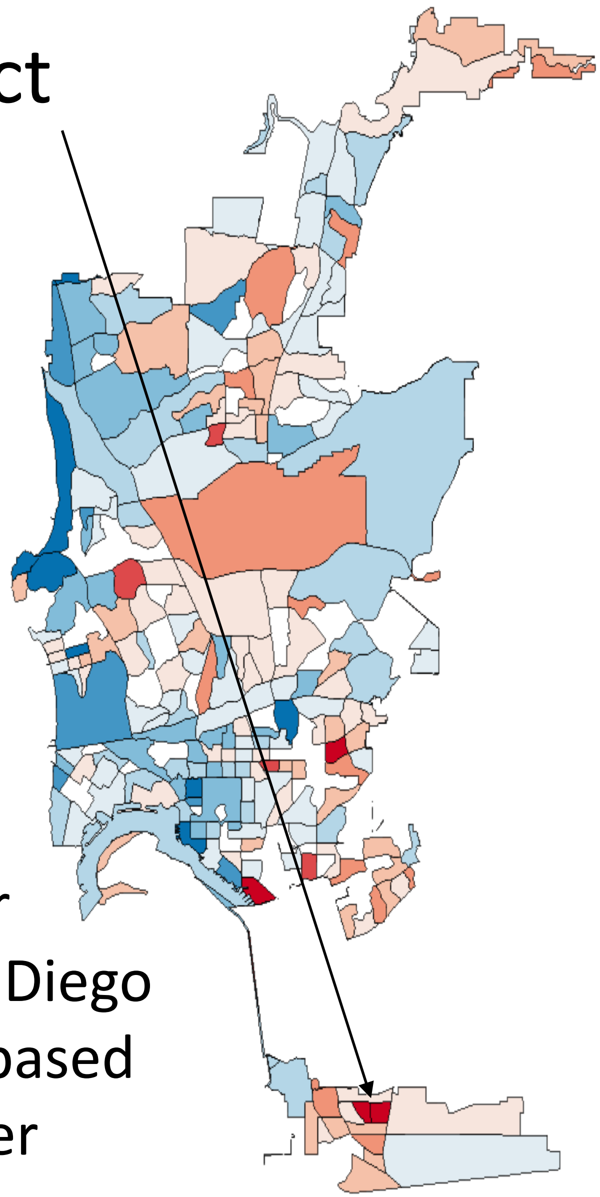
Happiest tract

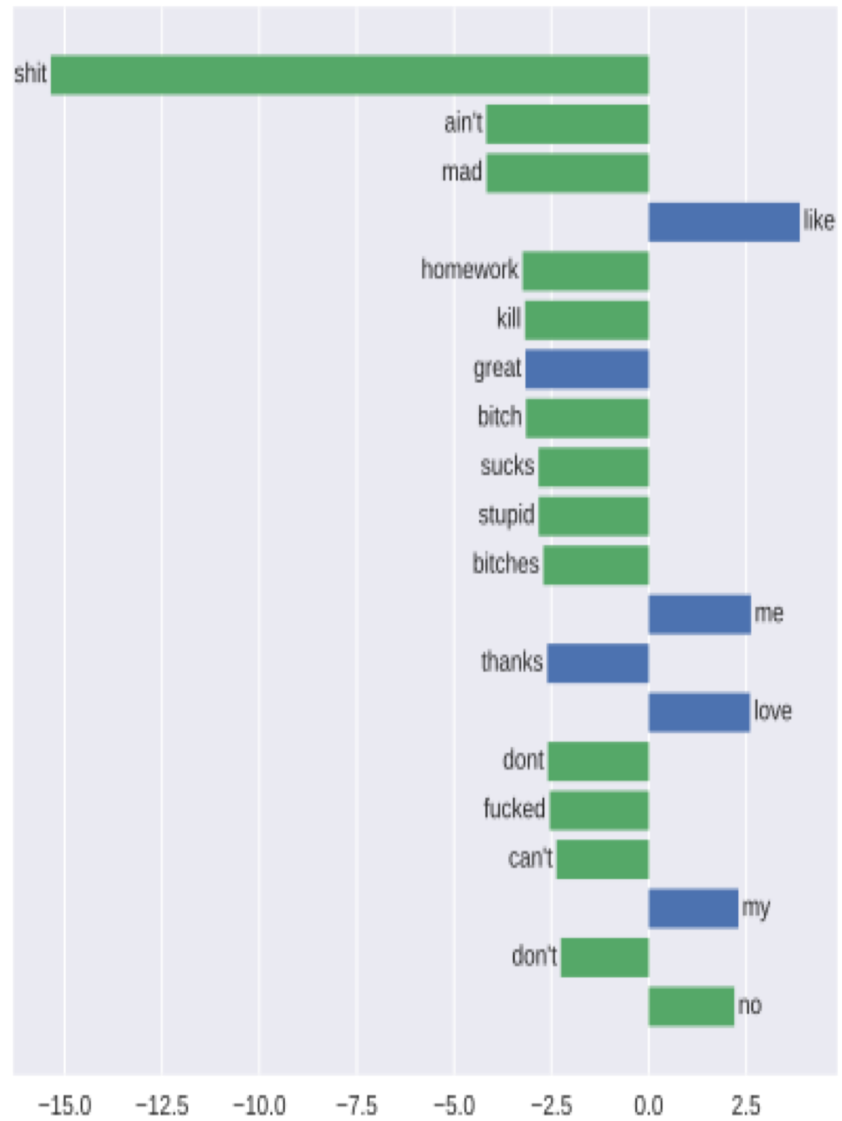




- Hedomometer scores for San Diego census tracts based on 2014 Twitter activity

Saddest tract

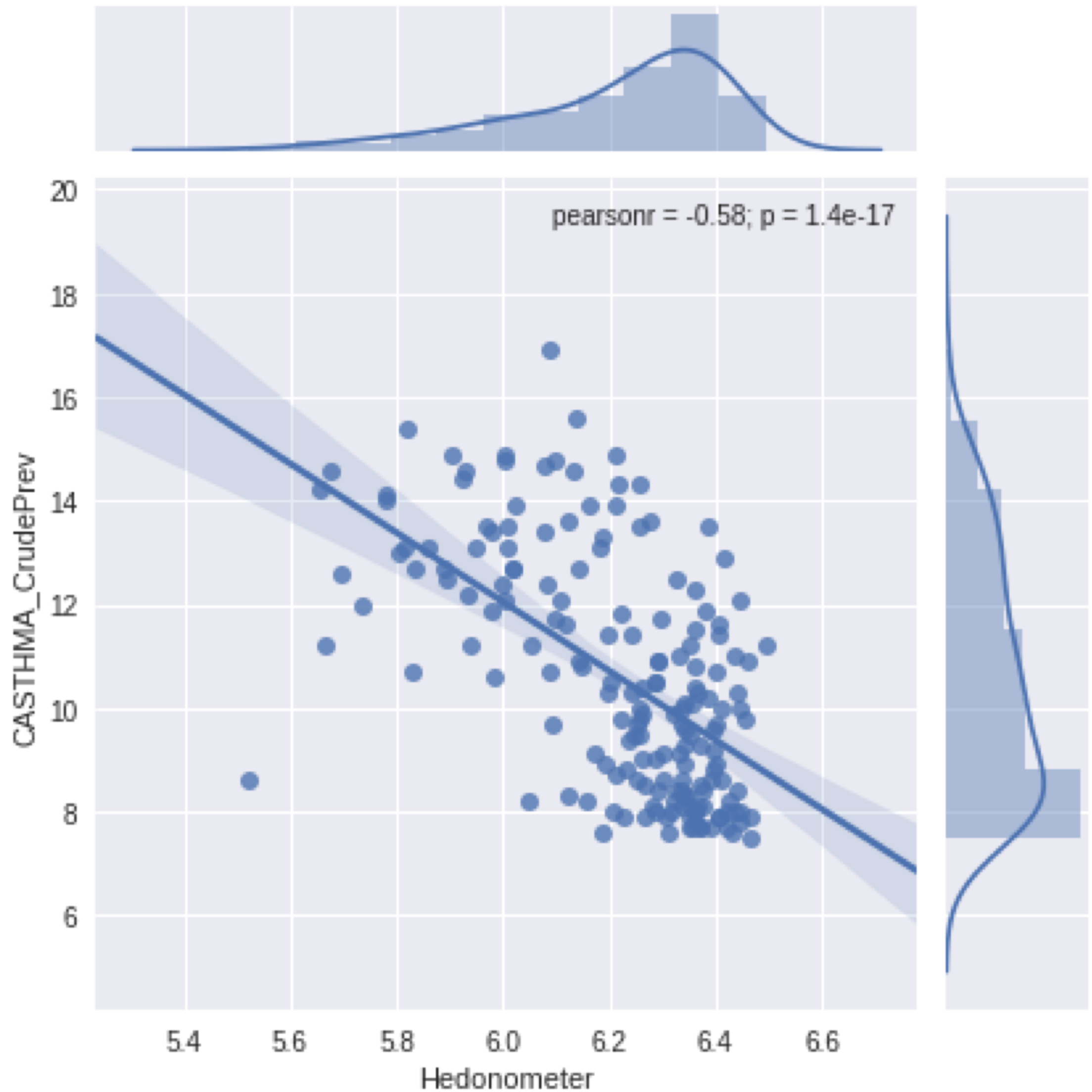




- Hedomometer scores for San Diego census tracts based on 2014 Twitter activity

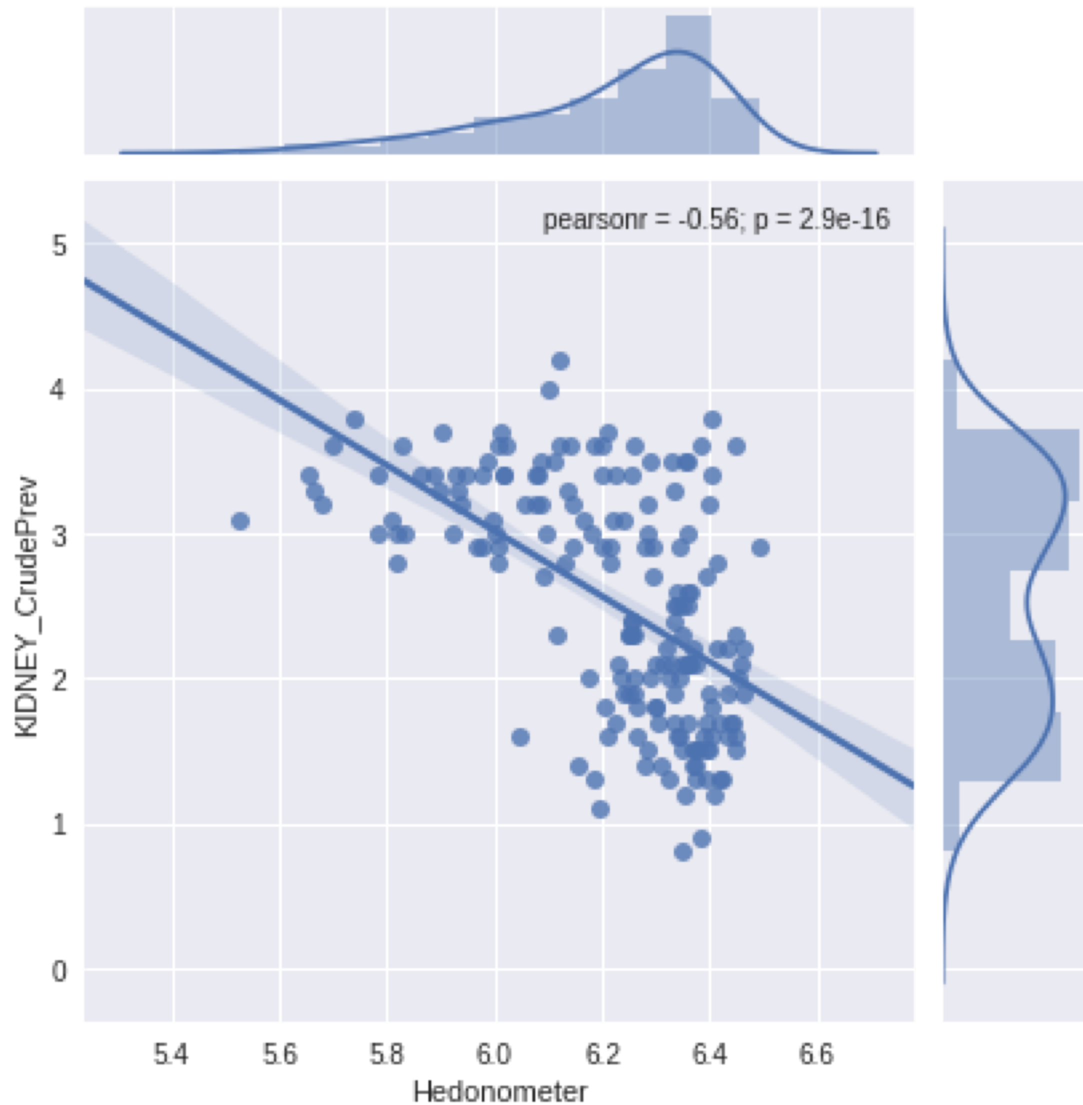
Happiness and Neighborhood Percent Reporting Asthma

Hedomometer scores correlated with neighborhood health data from the CDC (500 Cities Project)



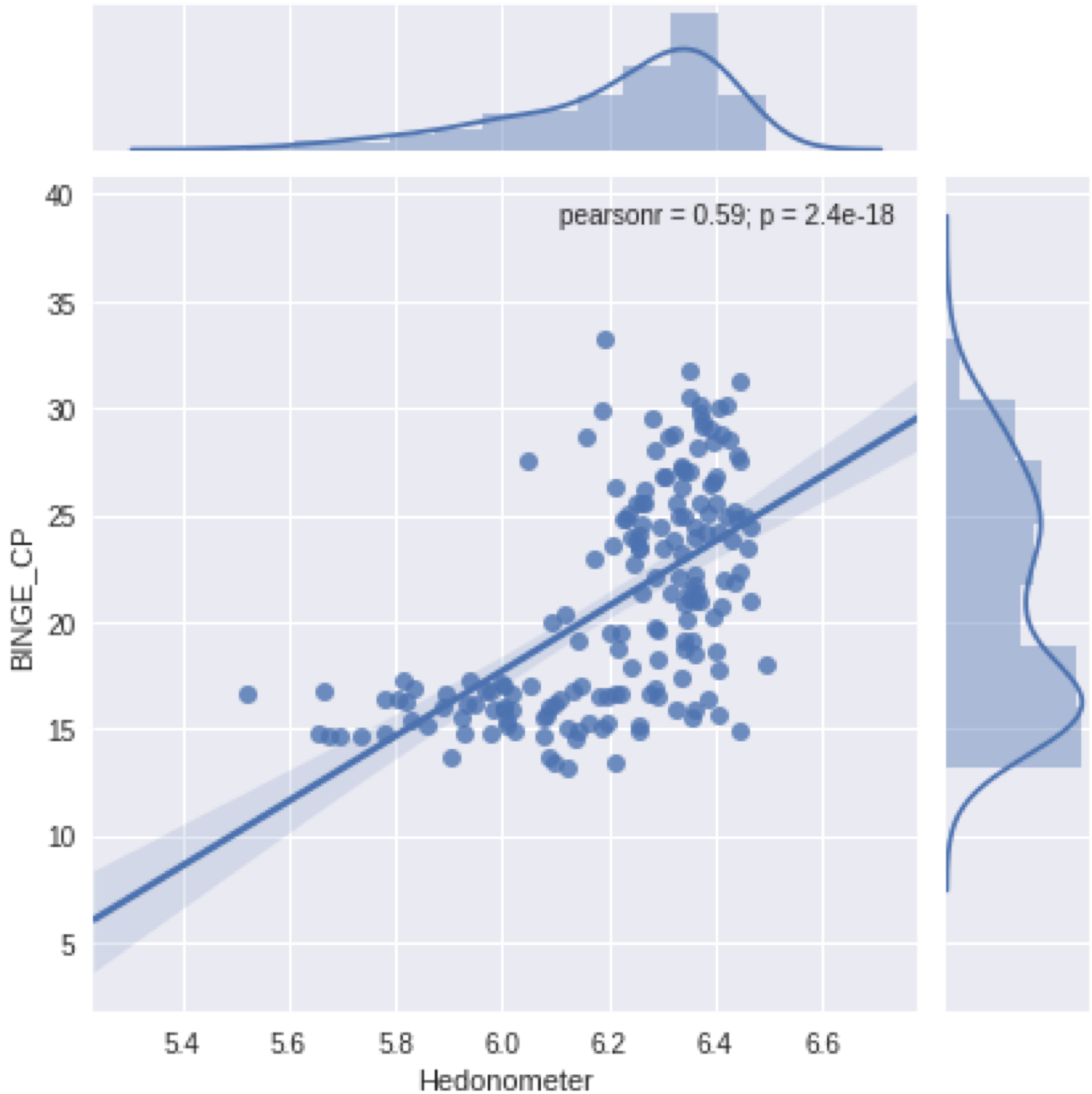
Happiness and Neighborhood Percent Reporting Kidney Problems

Hedomometer scores correlated with neighborhood health data from the CDC (500 Cities Project)



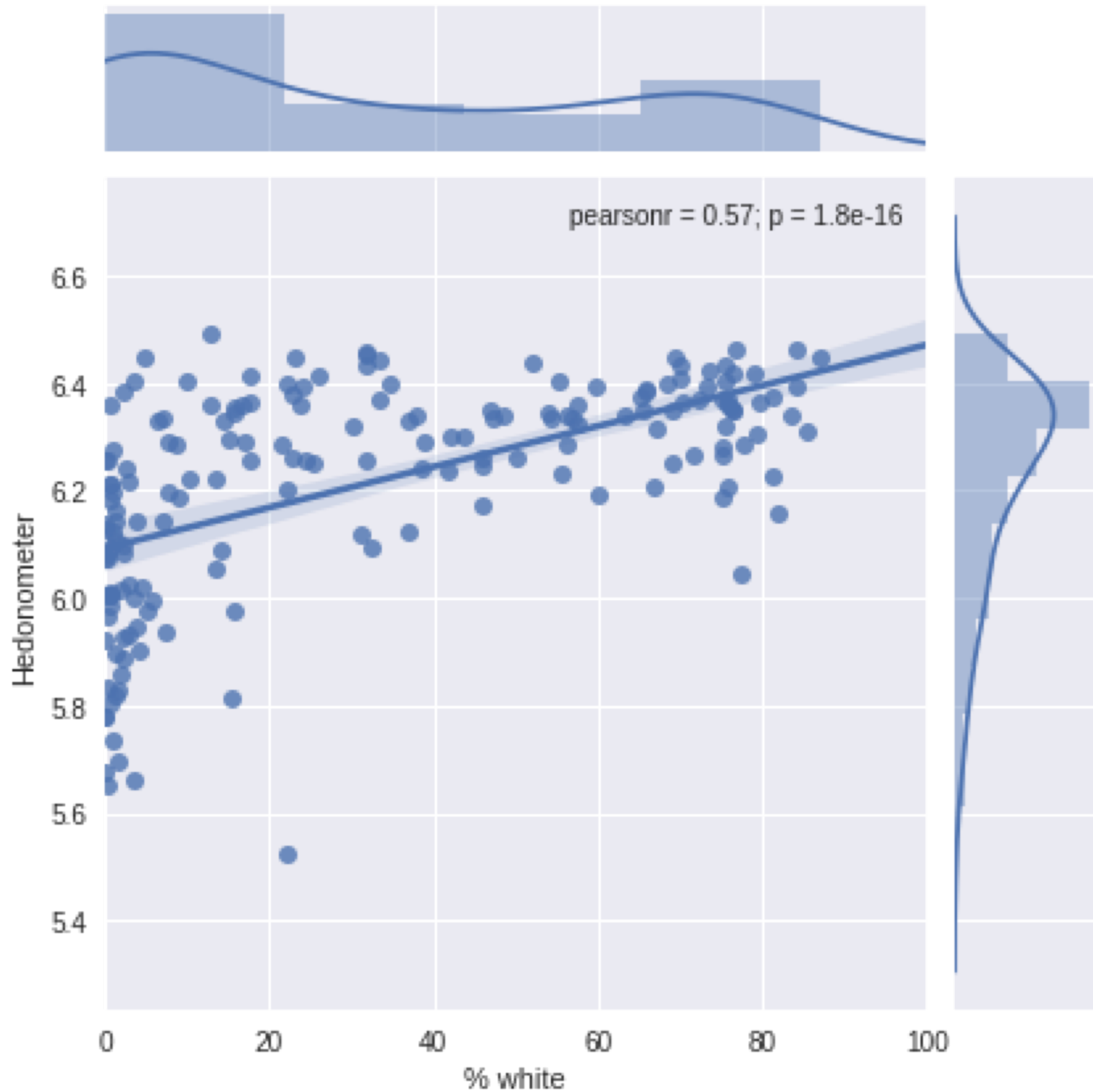
Happiness and Neighborhood Percent Reporting Binge Drinking

Hedomometer scores correlated with neighborhood health data from the CDC (500 Cities Project)



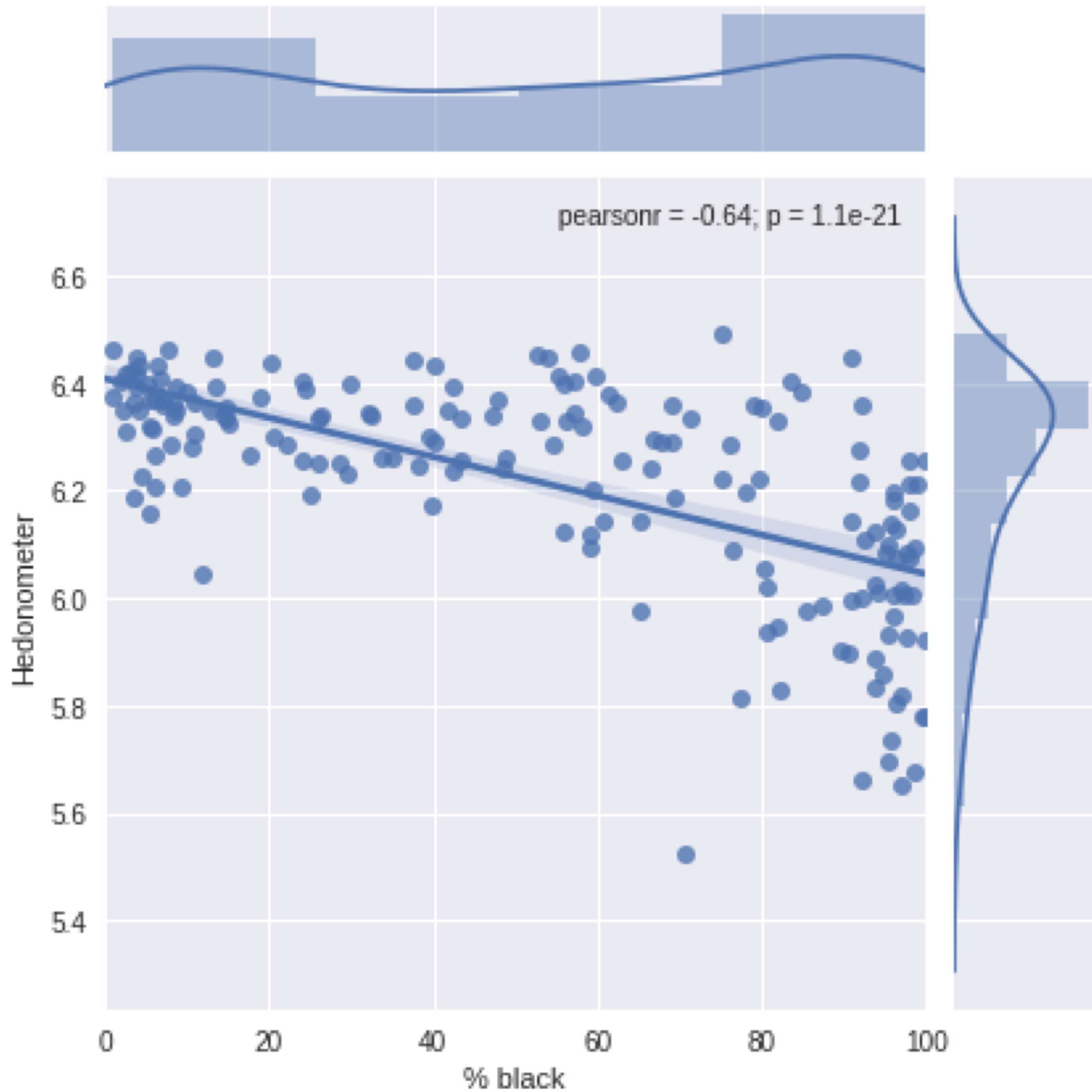
Happiness and Neighborhood Percent White

Hedomometer scores correlated with American Community Survey Data



Happiness and Neighborhood Percent Black

Hedomometer scores correlated with American Community Survey Data



Context

- Ratings for isolated words can't account for **linguistic** context

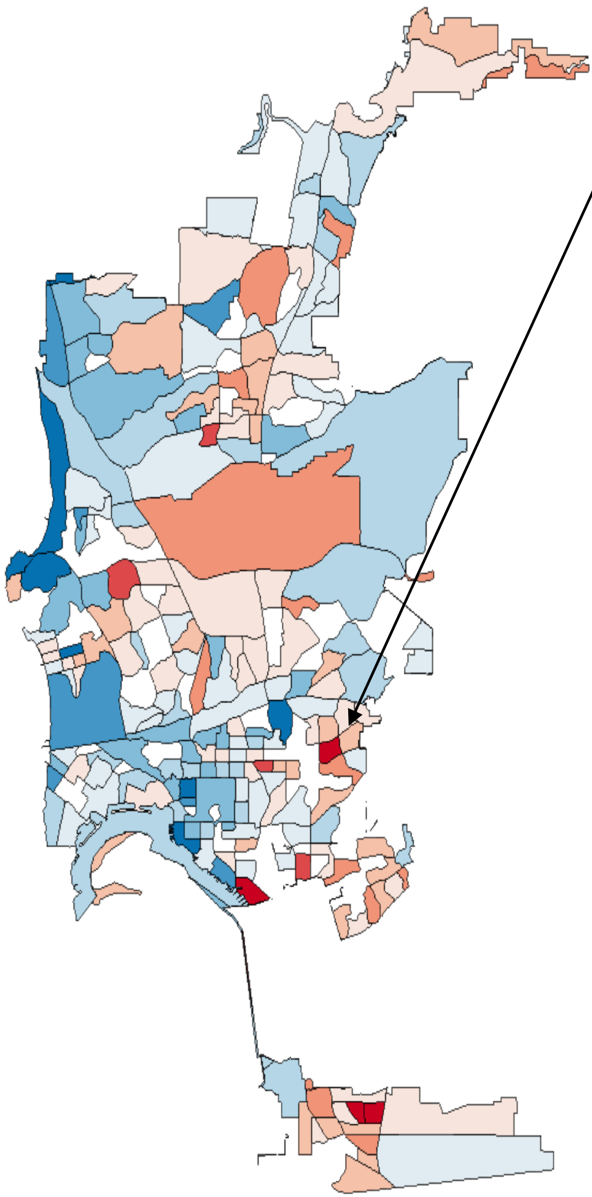
shit = 2.50

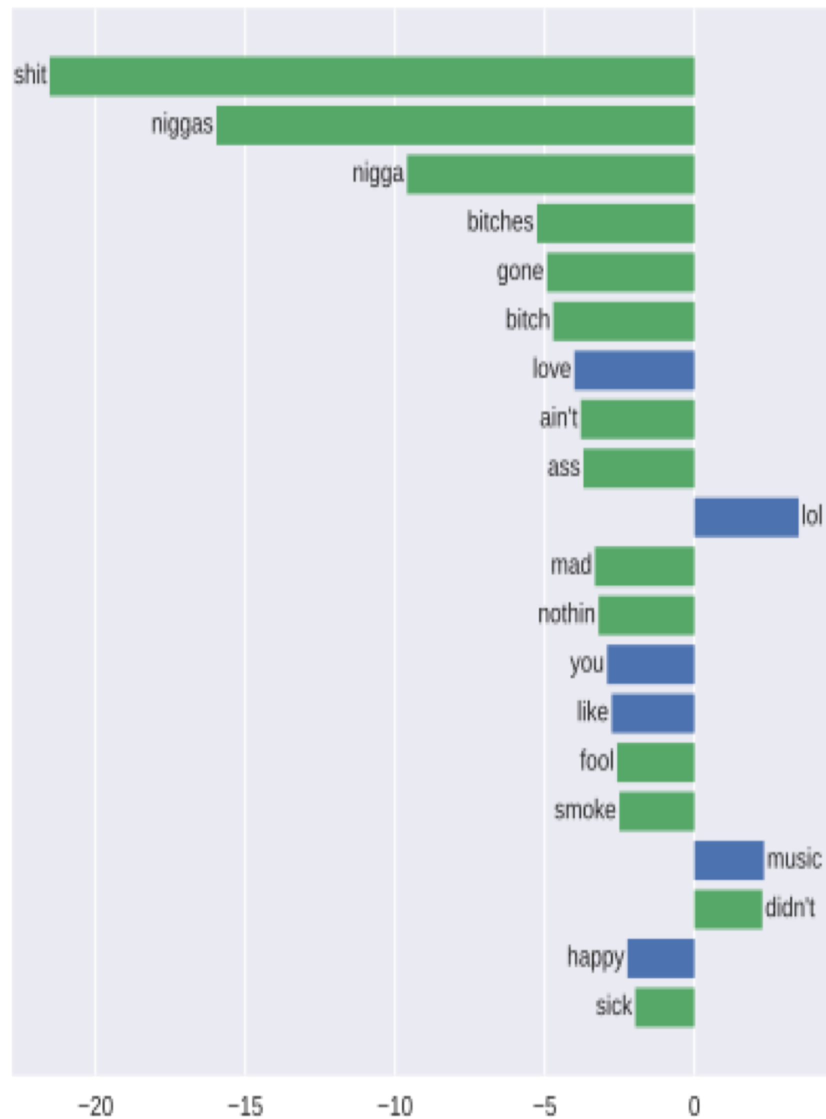
- *Haven't fucked with the bay area music in a long while. That Hyphy shit used to be my shit*
- *Yo, vikings is my shit*
- *The primo hooked it up, san diego got some great shit*
- *Every time I'm hungover I crave the shit out of Hot Cheetos*
- *I truly want to stay in bed all day but I got shit to do*
- *Holy shit its packed with doñas*

like = 7.22

- *Even on my Days off I Smell like Coffee -_-*
- *I still have scars, just like you have mine*
- *It's like fuck heads gravitate to me or something*
- *not to be mean but idk if me n steven r friends like we dont talk no more idk wat happened*
- *everyone got honey gear im like aye i got old gear fuuuk*

3rd saddest tract





Variation

Ratings for isolated words can't account for **social** context

gone (or *gon*) is used by some African American twitterers as a variant of *gonna*

@Sly2Doors u should we gone leave in a hour tho

We gone get it right doe

@Nyyjeria Wen I get back we def gone drop sumn

Only gone spread love positivity

Def gone rock with @djyamez in NYC 2015 We gone turn up any show he spin and we perform #ChoclateSundays

I'm proud of the direction Miami is heading its gone become a Mecca of great hip hop

The Underground gone keep whooping this Industry ass

Glad we gone perform #RollingLoud

Variation

- Ratings for isolated words can't account for **social** context
- *gone* (or *gon*) is used by some African American twitterers as a variant of *gonna*
- Despite being synonyms (maybe? maybe not exactly?), more standard forms are rated as happier

going	5.42
gonna	4.86
gone	3.42

Variation

- Everyone speaks a dialect (or dialects) — Language varies by place, class, race, gender, age, etc.
- ‘Standard’ English is a largely artificial written dialect that we learn in school
- Most Americans speak “Standard American English”
- HOWEVER, many Americans also speak African American English, Chicano English, etc.
- Each dialect has its own features and vary from S.A.E. small and large ways
- Social media users create conventions to capture dialect features that don’t usually show up in writing

Variation

- Blodgett, et al. (2016) built a corpus of tweets classified by the race of the tweeter
- Keyword analysis identifies words whose statistical distribution is skewed towards a subpart of the corpus
- Keywords associated with Hispanic and white users on average are slightly positive (consistent with results across a range of languages and corpora that show a positivity bias)
- Keywords associated with African American twitter users are, on average, slightly negative

Findings from Blodgett, et al. (2016)

- Keywords for African American users (avg score = 4.65)
  goodmorning niggaz wit foe phony niggas sucka loyal savage diss nigga messy chief folks gone mama sis asia hustler females pistol poetic uptown playin muthafuckin snitch ma everybody loyalty hatin ass bitches greedy yea gangsta muthafucka worried strapped slipping weak betrayed draws cryin trap nobody faithful fool rich daddy scandal bald inn grown thug ughh bang snatch killa kin snakes riches momma bitch lame stink hating bored rider shit gang acting bday dusty jail foolish she cursed hater female b-day essence bounce pimp screamin hungry designer nasty cheating mamma mad money amsterdam cook bothered slim anti india thugs pimps thanx

Findings from Blodgett, et al. (2016)

- Keywords for white users (avg score = 5.52)
  awful successfully acceptable indians incredibly invention snow dining greatly absolutely vehicles fantastic practical shattered severe fishing nursing tobacco impressive mountains functioning breakdown exams outdoor towns delay belief legitimate library snowing cancelled lake springfield neat tornado anxiety excitement glorious reindeer appropriate terrible bloom woods sinking obsession abroad tradition kitten informed possibility infection studied beds farmers craft providence completely separation hatred cannot shitty crappy grief constitution screwed rough insane opposed shore rudolph brutal numerous incredible moonlight logical productivity miserable coolest impressed olympics rot unable pleasant alarm withdrawal beers reasoning humanity

What's going on

- MTurkers rate words based on their intuitions on their own dialects, which may not generalize to other dialects
- MTurkers have a systematic negative bias against language associated with A.A.E.
- Differences in ratings reflect real underlying differences in happiness between populations
- Some or all of the above

In Sum

- Big data presents tremendous promise for the future of neighborhood-based research
- However, we still have a long way to go before it is fully applicable

Moving forward

- Models which directly address race, class, etc.
- Incorporate properties (e.g., emojis) which arguably show less cross-dialect variation

SUPPLEMENTAL SLIDES

Geotagged Tweets

- Exclude tweets from known bots (TweetMyJOBS, TTN SD Traffic, San Diego Trends, …) and services (Instagram, Foursquare, Untappd, Endomondo, …)
- Only include users with more than 30 days activity
- Linguistic pre-processing: tokenize, map to lower case, collapse repeated characters, remove @usernames, maps URLs to “<url>”
- Remove duplicates (same normalized text+same poster)
- Final dataset includes 964,559 tweets from 19,000 users posted between 2014-12-06 and 2017-05-24