Word Sense Disambiguation

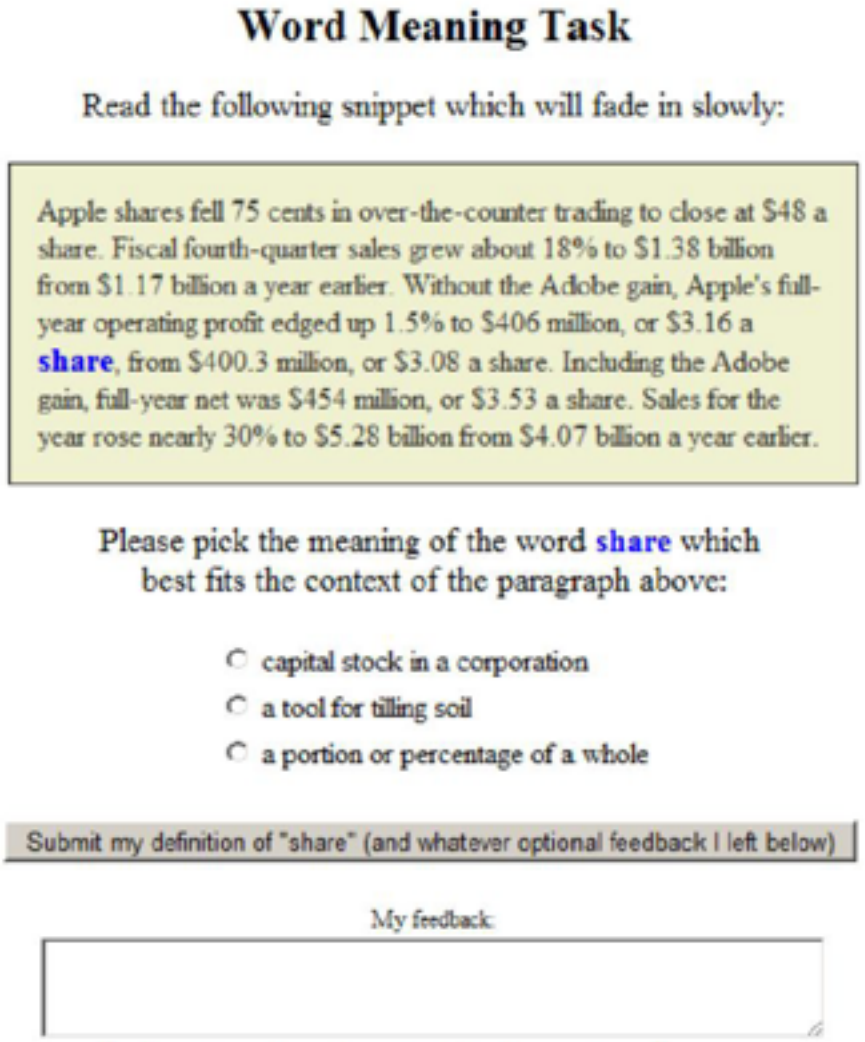
- **Words often denote different meanings**
  - I can hear *bass* sounds.
  - They like grilled *bass*.
- **But it's often less clear how different they are**
  - *Add* 2 and 5
  - *Add* sugar to coffee
  - *Add* someone to your team
  - *Add* something to your repetoire

Crowdsourcing can either give the same labels you would have gotten with experts or can provide *different* insights.

Questions Addressed

- **Are Turkers as good as experts?**
  - (How) can we combine Turker responses to boost accuracy?
- **Do workers' performances vary?**
  - Which characteristics in the Turker's engagement affect task accuracy?
- **Do WSD tasks differ?**
  - Which features in the target word, the context, and the definition of the sense choices affect accuracy?

Our WSD HIT



Word Meaning Task

Read the following snippet which will fade in slowly:

Apple shares fell 75 cents in over-the-counter trading to close at $48 a share. Fiscal fourth-quarter sales grew about 18% to $1.38 billion from $1.17 billion a year earlier. Without the Adobe gain, Apple's full-year operating profit edged up 1.5% to $406 million, or $3.16 a **share**, from $400.3 million, or $3.08 a share. Including the Adobe gain, full-year net was $454 million, or $3.53 a share. Sales for the year rose nearly 30% to $5.28 billion from $4.07 billion a year earlier.

Please pick the meaning of the word **share** which best fits the context of the paragraph above:

- capital stock in a corporation
- a tool for tilling soil
- a portion or percentage of a whole

Submit my definition of "share" (and whatever optional feedback I left below)

My feedback:

- A subset of the Penn Treebank Wall Street Journal Corpus
- A collapsed version of Wordnet 2.1
- Nouns and verbs only

This data is from the "OntoNotes" project (2006). They iteratively make the senses more and more "coarse-grained" until they achieve a 90% ITA among the annotators. For example, the word "share:"

One snippet, one word, all coarse senses. Optional feedback box. Prevent cheating / satisficing by (1) fading in the words slowly (2) randomly ordering the senses.

1,000 words, each disambiguated by 10 Turkers each

Data Collection

- **10,000 HITs**
  - 1,000 words * 10 ratings/word
  - cost: $110 (pay 10 cents/Hit)
- **The work**
  - 595 Turkers
  - total time: 51hr
  - throughput: 4700 labels / day
- **Data quality**
  - Krippendorf's = 0.66
  - fair inter-rater reliability

Combining Turker ratings

Combine annotations and use plurality vote arbitrating ties randomly:

# of Dis	2	3	4	5	6	7	8	9	10	2.4 (1st pl)
Accuracy	.73	.80	.81	.82	.83	.84	.84	.84	.86	.81

Combining 10 untrained Turkers is competitive with the accuracy of state-of-the-art algorithms.

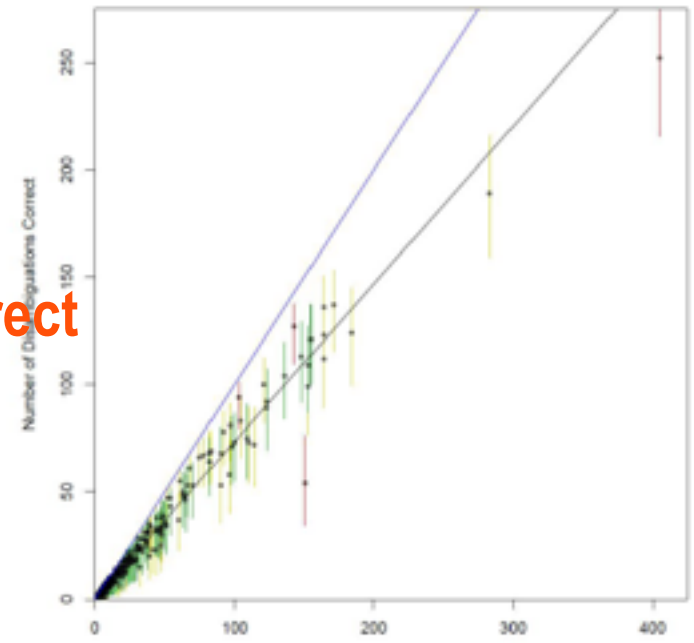
Ipeirotis, 2011 claims that "majority vote works best when workers have similar quality" and if not:

- Find best worker and use that label
- Model worker quality and combine

Do our workers differ in quality?

Do we have spammers and superstars? Is anyone different from the average, $p_{\text{avg}} = .73$? Let C_w be the number correct for Turker w.

$$H_0 : C_1, C_2, \ldots, C_W \overset{ind}{\sim} \text{Binomial}\,(n_w,\ p_{\text{avg}})$$



correct

disambiguations done

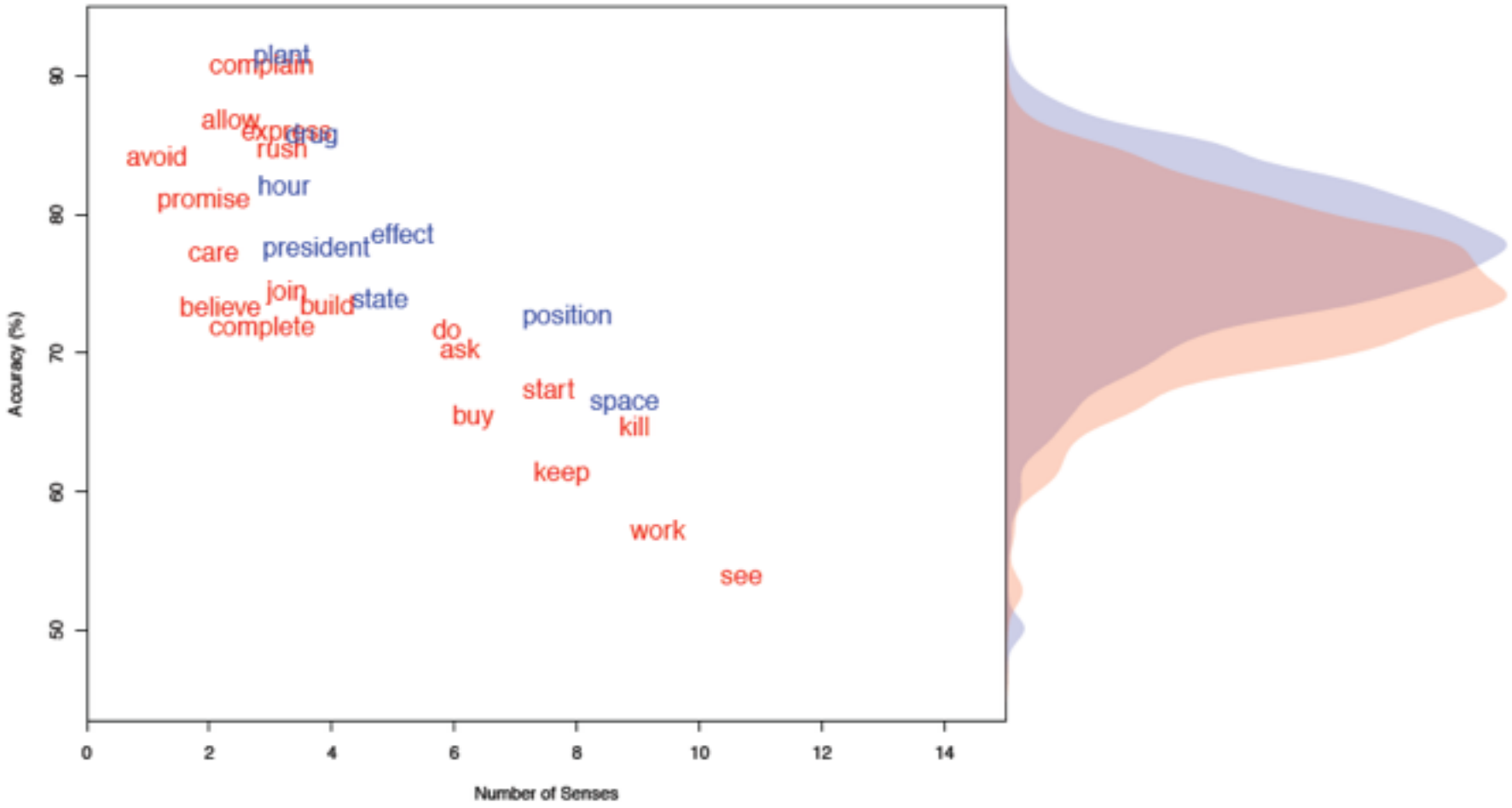
(with Bonferroni-corrected Binomial proportion CI's $\alpha = 5\%$).

We are forced to reject on the account of four workers.
To a first order approximation, all Turkers are the same. Also note: no learning effect was found (unshown).

Beyond collecting labels

- **Is there a ground truth word sense label?**
  - *Words differ massively in levels of worker agreement*
  - Training 'professional' annotators tries to hide this ambiguity.
- **Maybe we care about a different question**
  - How ambiguous is a word use?
  - What makes people find a word use ambiguous?

Accuracy v. number of senses



WSD Take-aways

- **M-turkers are as good as trained WSD labelers**
  - Not as individuals, but when aggregating up to 10 turkers
  - Different words need different numbers of labels
- **One can study how people understand language**
  - rather than just studying language *per se*

Spending more time on the disambiguation task associates with a significant *reduction* in accuracy

NLP for Computational Social Science

- **Language is about people**
  - Depressed, autistic, schizophrenic?
  - Happy, sad, angry, lonely, bored, proud?
  - Convinced or skeptical?
  - Extravert, open, conscientious, agreeable, neurotic?
- **Crowdsourcing for social science NLP requires collecting information about the people or organizations producing the words**
  - Often via Facebook
  - But m-turk and Twitter are used, too

Language is about people

- **Language is used by people to communicate with other people**
  - It reflects the speaker and influences the recipient
  - Crowdsourcing is ideal to study this

"Language is the most common and reliable way for people to translate their internal thoughts and emotions into a form that others can understand. .....the medium by which cognitive, personality, clinical, and social psychologists attempt to understand human beings" (Tausczik & Pennebaker, 2010)

Language and People

- **Language used varies with the writer's**
  - Age, sex, race
  - Personality, intelligence
  - Health, education, political beliefs
  - Geo-location
- **How does this post/article make you feel**
  - Awed, angry, happy, connected, disgusted
- **This post/tweet represents**
  - Positive/negative emotion
  - (dis) engagement, relationships, meaning, accomplishment

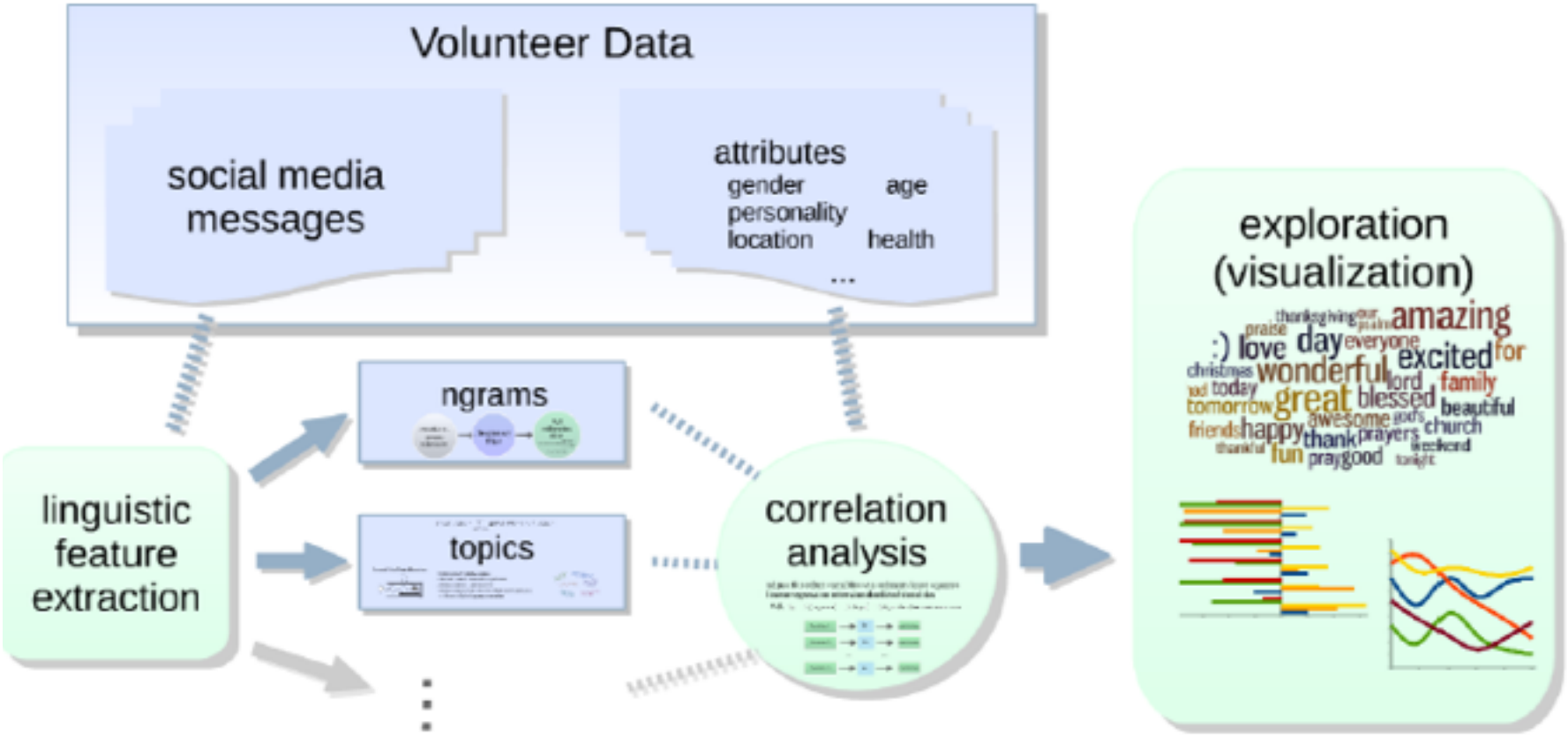
WWBP: Understand well-being

- **Individual personality and well-being**
  - 70,000 Facebook users' posts
  - age, sex, and personality questionnaire results
- **Community well-being**
  - Billions of tweets, mapped to county
  - County-level happiness, crime and disease rates

Crowdsourcing at WWBP

- **Labels for lexicon Generation**
  - Twitter/Facebook for PERMA
    - **P**ositive emotion, **E**ngagement **R**elationships, **M**eaning, **A**ccomplishment
  - Optimism/pessimism (causal)
- **Information about people**
  - Age and Sex
  - Personality tests
  - Medical records
  - …

Analysis Method



Crowdsource text and context

- **Have people share facts about themselves**
  - **Paid (M-turk)**
    - provide an email you wrote
  - **Rewarded**
    - we collect Facebook access at the Penn ER, and offer a chance to win an iPad
  - **or just asked**
    - *You took our questionnaire/quiz; here is your result; will you share your Facebook posts?*
    - my personality
    - PERMA

Females


soooo
excited
sooooooo
tomorrow
ridiculously
sooo
satisfying
excitement
yay
hun
love
;D
girlies
^.^
<3
inlove
sweetie
mucho
sweetheart
^ ^
_
i'm_so
her_new
best_friend
chocolate
her
she
shopping
ugh
<3
sister
proud_of
with_my
mom
birthday
my_hair
can't_wait
loves_her
wonderful
am_so
so_happy
lovely
excited
baby
hubby
soo
boyfriend
love:)you
much_fun
dress
soooyay_!:(
cute
i_miss
mommy
girls
supportive
loving
herself
so_much
yummy
like_it_on
wonderful
cleaning
wishes_she
amazing
family
husband
helped
love_him
my_family
soooo
omg
husband
loved
sooooo
friends
thankful
blessed
truely
boyfriend
daughter
grateful
lucky
bestie
:)
hacked
ar
par
bestfriend
ily
boyfriend
=]
bf
xoxo
cute
baby
sweet
sooo
adorable
hes
cutest
babies
soooo
puppy
he's
lil
aww
awww
soo
turns
nephew
celebrate
happy
years
st
birthday
wonderful
wishing
brother
daughter
special
sister
niece
son

Males

football
fantasy
sports
play
league
player
playing
team
game
basketball
players
coach
government
state
power
country
freedom
thomas
nation
america
human
political
rights
civil
democracy
liberty
society
fight
win
fighting
won
battles
defeat
war
battle
enemy
sword
defeated
bands
meet
fought
victory
vs
engineering
call_of_duty
match
his_life
win
cod
league
nfl
hit_me_up
team
fucking
creed
xbox
himself
http:/
world_cup
loves_his
ps3
youtube.com
president
metal
black_ops
shit
my_girlfriend
fuckin
fans
my_wife
war
dick
fifa
teams
fuck
championship
modern_warfare
album
government
shave
arsenal
fucked
wins
nba
ftw
wishes_he
halo
holy_shit
lebron
thinks_he
his_mind
360
beard
fucking
fuckers
haters
sake
fuck
shut
bullshit
fucks
bitches
outta
shit
online
gaming
playing
fifa
pc
cod
live
wii
xbox
games
playin
play
ps
tag
add
economy
pay
public
tax
state
cuts
country
income
taxes
debt
government
obama
budget
health
benefits
a
a
a
b
b
b

Personality

High

Extraversion

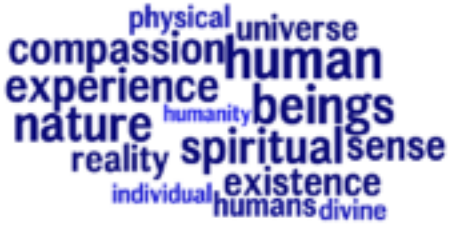
Neuroticism

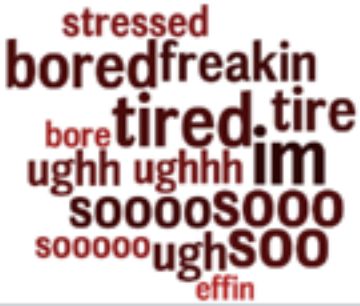
Low

Personality, Gender, and Age in the Language of Social Media …
Schwartz, et al. PLOS ONE 2013

More at wwbp.org

The language of life satisfaction





Characterizing Geographic Variation in Well-Being using Tweets.
Schwartz et al. ICWSM 2013

The language of heart disease

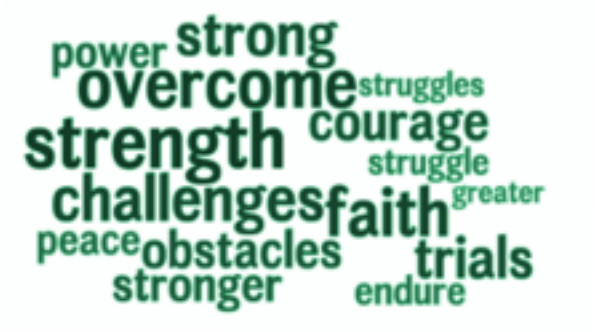
Less heart disease

More heart disease

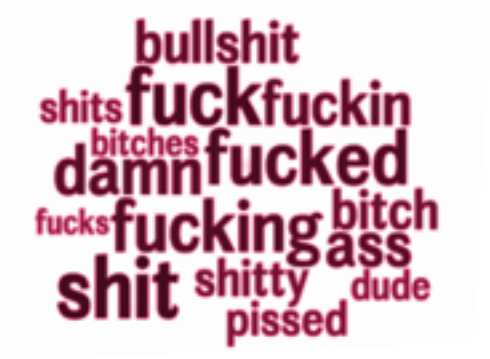
Pychological Language on Twitter Predicts County-level Heart Disease Mortality
Eichstaedt et al. Pyschological Science, 2015

The language of heart disease

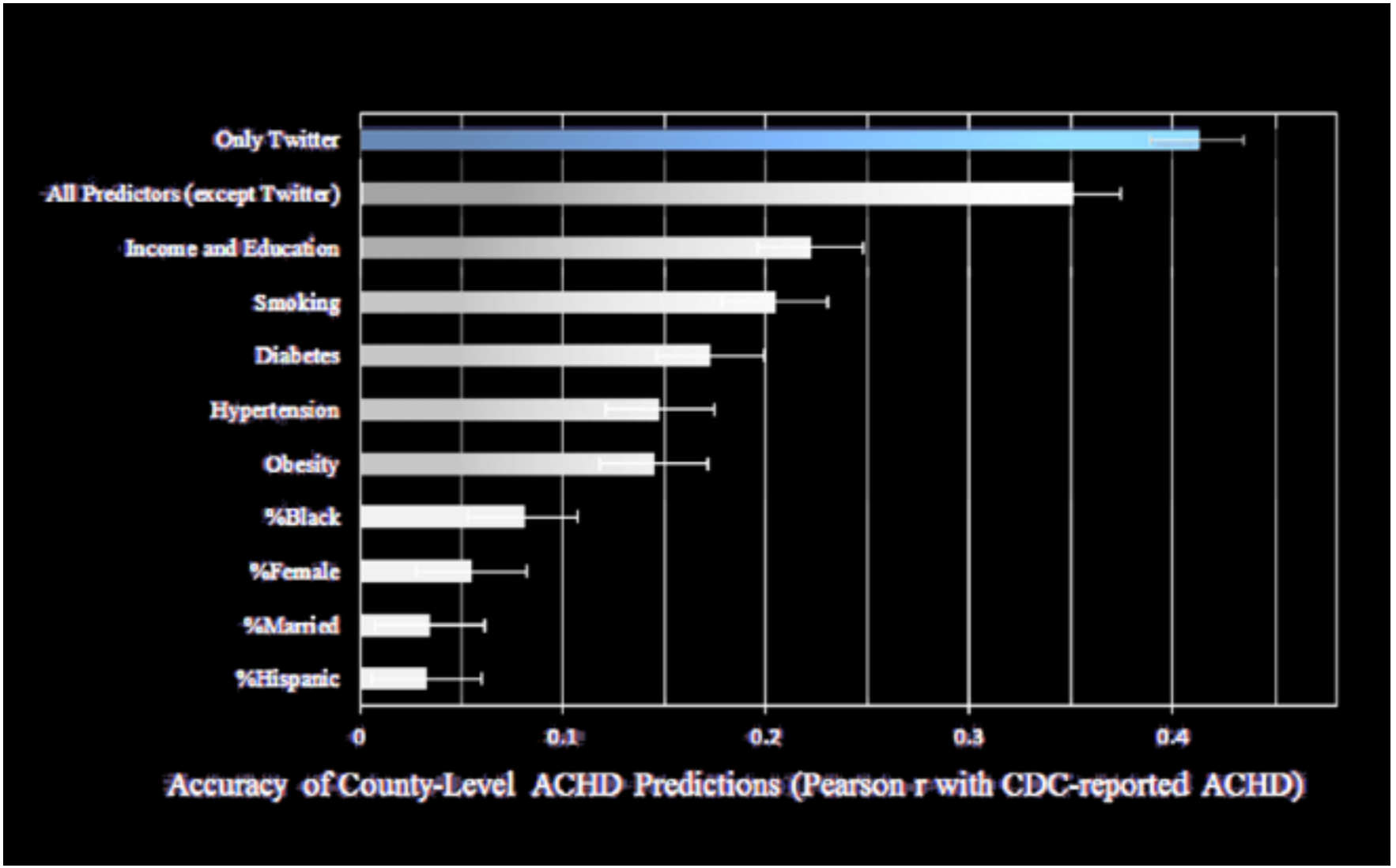
Less heart disease



More heart disease



Twitter predicts cardiovascular disease



Take-aways

- **Crowdsourcing lets you collect**
  - **people's words**
    - posts, tweets, emails, essays
  - **who generated the the words**
    - age, sex, personality, job, IQ …
  - **when and where they were generated**
    - Foursquare, phone apps
  - **what images accompanied them**
    - Facebook, Instagram …
  - **how the words make people feel**