Discovering Psychological and Health Insights from Social Media Langugae

H. Andrew Schwartz
andrew.schwartz@cs.stonybrook.edu

October 16, 2015
@ SBU SUNYK Seminar

x 20mil.

1. Individual Analyses

x 1bil.

2. Community Analyses

Introduction

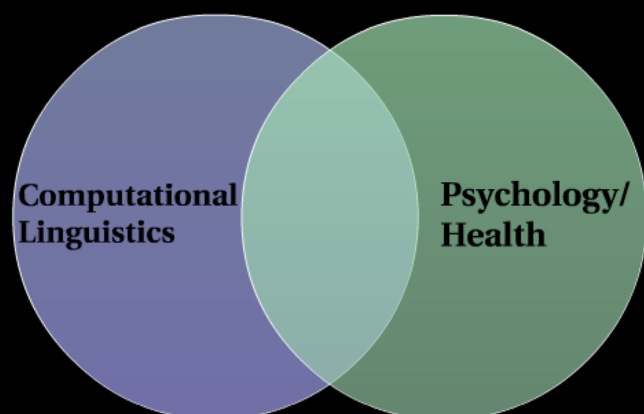
Supplemental Material

Multi-level Adaptation

Representative Sample?

Discovering Psychological and Health Insights from Social Media Langugae

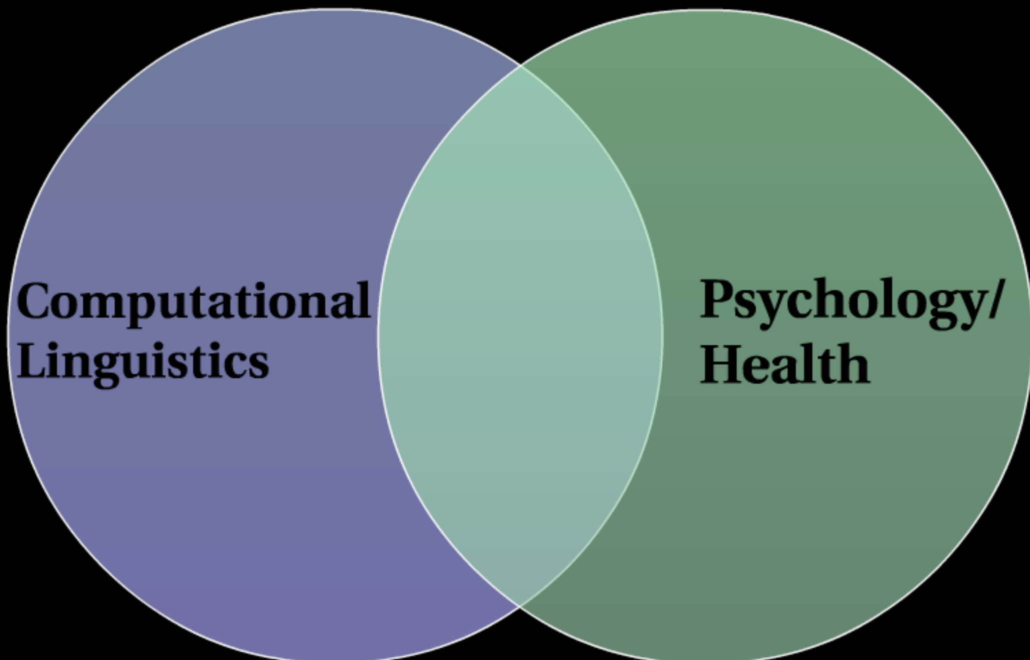
H. Andrew Schwartz
andrew.schwartz@cs.stonybrook.edu



October 16, 2015
@ SBU SUNYK Seminar

H. Andrew Schwartz

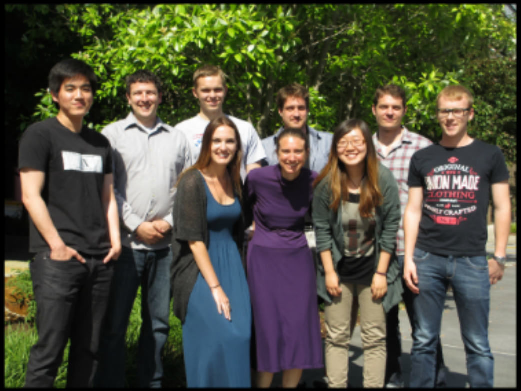
andrew.schwartz@cs.stonybrook.edu



r

WWBP Collaborators

Rosie Hancock Darwin Labarthe
Richard Lucas Luke Dziurzynski Michal Kosinski
Yiyi Guo Sneha Jha Gregory Park Shawndra Hill
Tadas Antanavicius Robert Backer
Chris Weeg Lyle H. Ungar Rigel Swavely
Jeanette Elstein Margaret Kern
Liwei Xu Stephanie Ramones George Wan Megha Agrawal
Dolores Albaraccin Maarten Sap Johannes Eichstaedt
Achal Shah Emily Larson Martin E. P. Seligman
Brian Galla Libby Benson Eduardo Blanco Marie Foregeard
Annie Roepke Yoon Hwang
Arsenij Kouriatov Raina Merchant Saif Mohammad
Bob Stine Winnie Cheng David Stillwell Molly Ireland
Evan Weingarten Jonah Berger
Dean Foster
Daniel Peotiuc Jordan Carpenter



UNIVERSITY OF CAMBRIDGE

HARVARD UNIVERSITY

LYMBA the power to answer

Wharton UNIVERSITY of PENNSYLVANIA

Perelman School of Medicine UNIVERSITY of PENNSYLVANIA

PENN SOCIAL MEDIA AND HEALTH INNOVATION LAB

a friend

- measures non-objective outcomes
- many developed over decades
- a gold-standard

a friend

- measures non-objective outcomes
- many developed over decades
- a gold-standard

a foe

- hard to administer at scale
    - spatially
    - temporally
- not exactly "ground truth"
- limited to preconceived theory

Social Media

350m tweets/day

4b messages/day

...

Social Media

350m tweets/day

4b messages/day

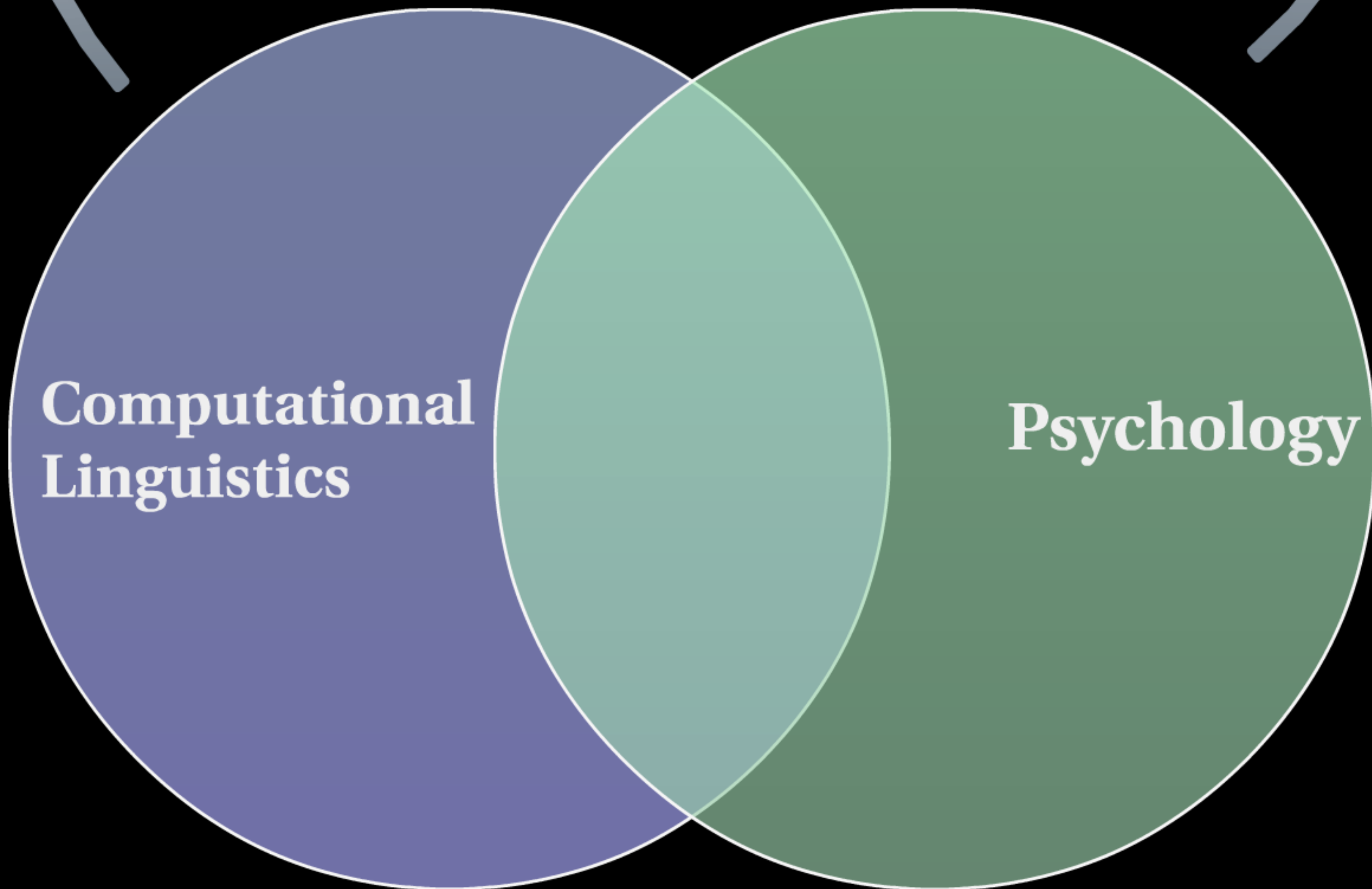
...

...the largest dataset of who we are

1b people

150m people

from *modeling language*
to *understanding people*



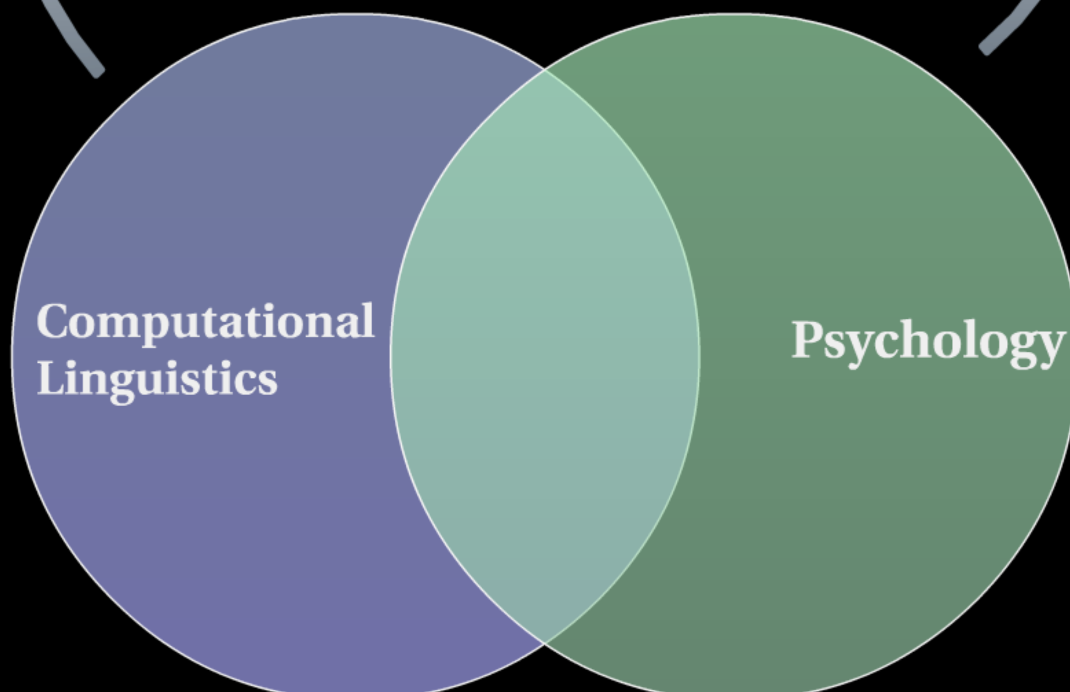
goal: accurate prediction

method: data-driven
(large data)

goal: human insights

method: closed-vocabulary
(typically "small" samples)

from *modeling language*
to *understanding people*



data-driven human insights

goal: accurate prediction

method: data-driven
(large data)

goal: human insights

method: closed-vocabulary
(typically "small" samples)

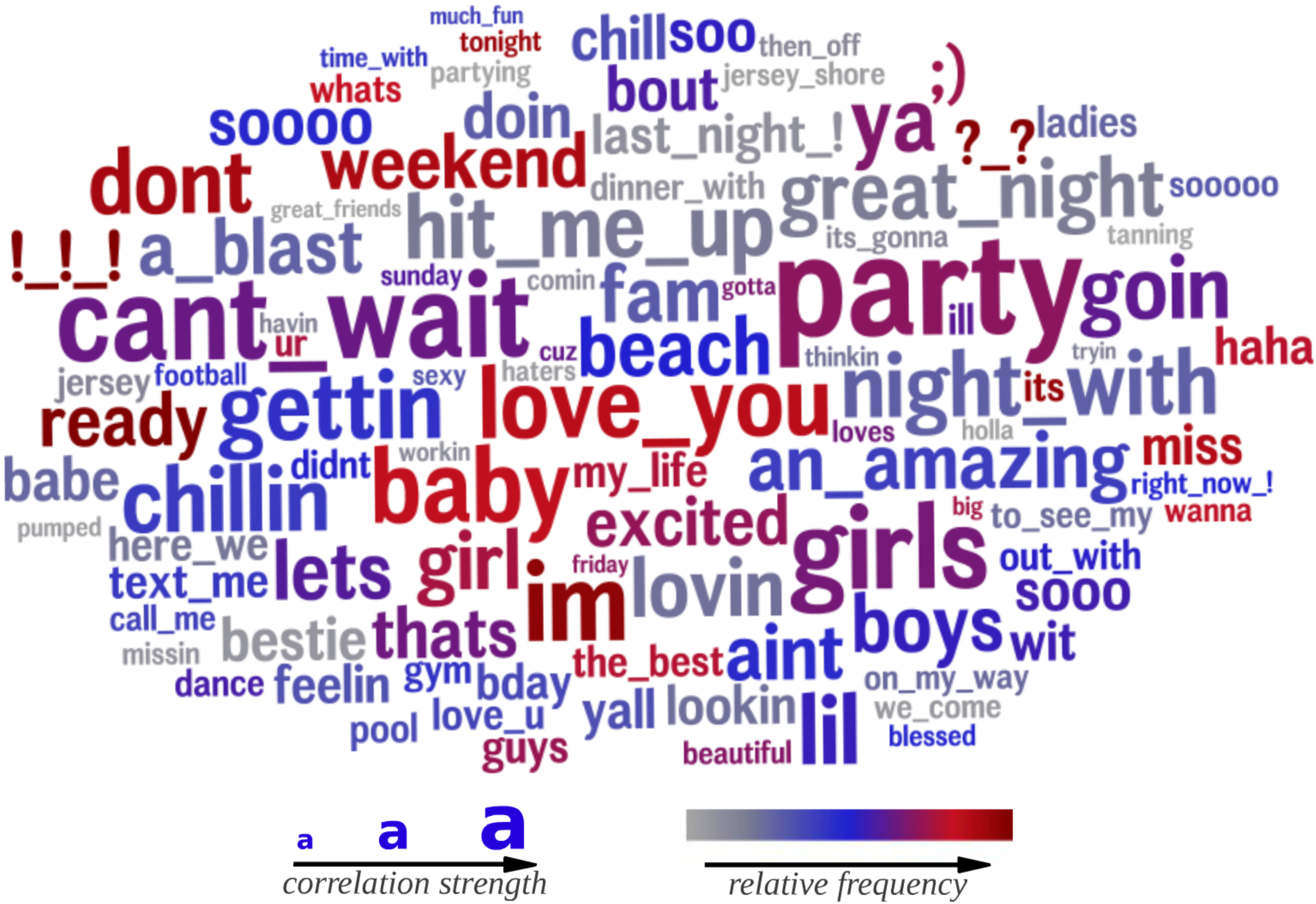
N=75,000 users... ...20M statuses.

extraversion

explicit language warning

neuroticism

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Toward Personality Insights from Language Exploration in Social Media. In *Proceedings of the AAAI Spring Symposium Series.* Stanford, CA.



much_fun
tonight
time_with
chill
soo
then_off
partying
whats
jersey_shore
;)
bout
soooo
doin
last_night_!
ya
?_?
ladies
weekend
dont
dinner_with
great_night
sooooo
great_friends
hit_me_up
its_gonna
tanning
!_!_!
a_blast
sunday
comin
fam
gotta
party
goin
cant_wait
havin
ur
ill
beach
cuz
haha
thinkin
tryin
jersey
football
sexy
haters
night
its
with
gettin
love_you
ready
holla
loves
workin
miss
babe
didnt
an_amazing
chillin
my_life
baby
right_now_!
big
to_see_my
wanna
pumped
excited
here_we
girls
friday
out_with
text_me
lets
girl
lovin
sooo
im
call_me
bestie
thats
boys
wit
missin
the_best
aint
dance
feelin
gym
bday
on_my_way
yall
lookin
lil
we_come
love_u
pool
guys
beautiful
blessed
a
a
a
correlation strength
relative frequency

human
sort_of
metal
%_won't_copy
depression
books
google
reading
%_won't
93_%
draw
please_put_this
bleach
spam
>.<
I'm_going_to
apparently
:_3
I_don't
rin
ako
yan
curse
using
lang
>_<
walang
it's
internet
drawing
emo
laptop
nya
x3
to_read
suddenly
ang
they're
anime
gaming
>d:<
don't_want_to
japanese
dx
xp
ni
mga
my_cat
>:
sigh
pokemon
pc
manga
virus
wikipedia
I've
xd
zombie
sana
account
comic
pag
-_-
t_._t
@_@
graphics
sakit
managed_to
ng
computer
._at_least
to_kill
wala
related
nga
^_^
pages
-.-
characters
8d
final_fantasy
naman
hindi
evil
%_^
o.o
doctor_who
>:d
akong
lahat
keyboard
d:<
kaya
nearly
online
didn't

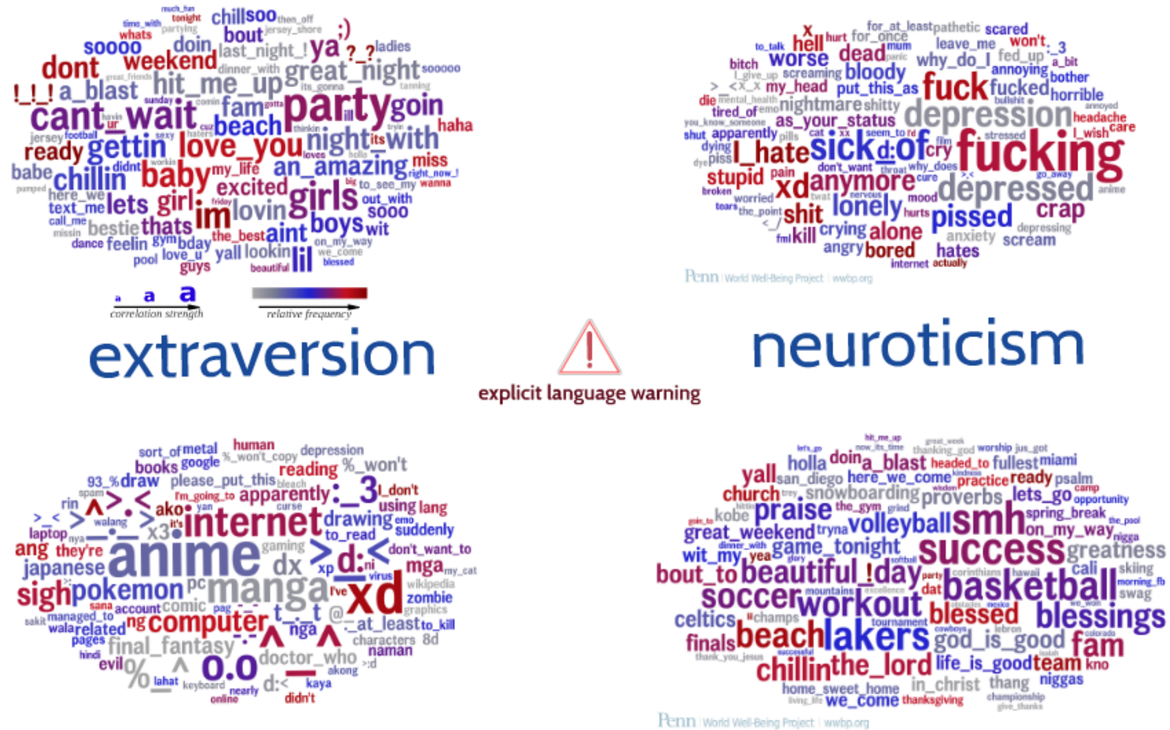
extraversion

explicit language warning

neuroticism

Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Toward Personality Insights from Language Exploration in Social Media. In *Proceedings of the AAAI Spring Symposium Series.* Stanford, CA.

explicit language warning

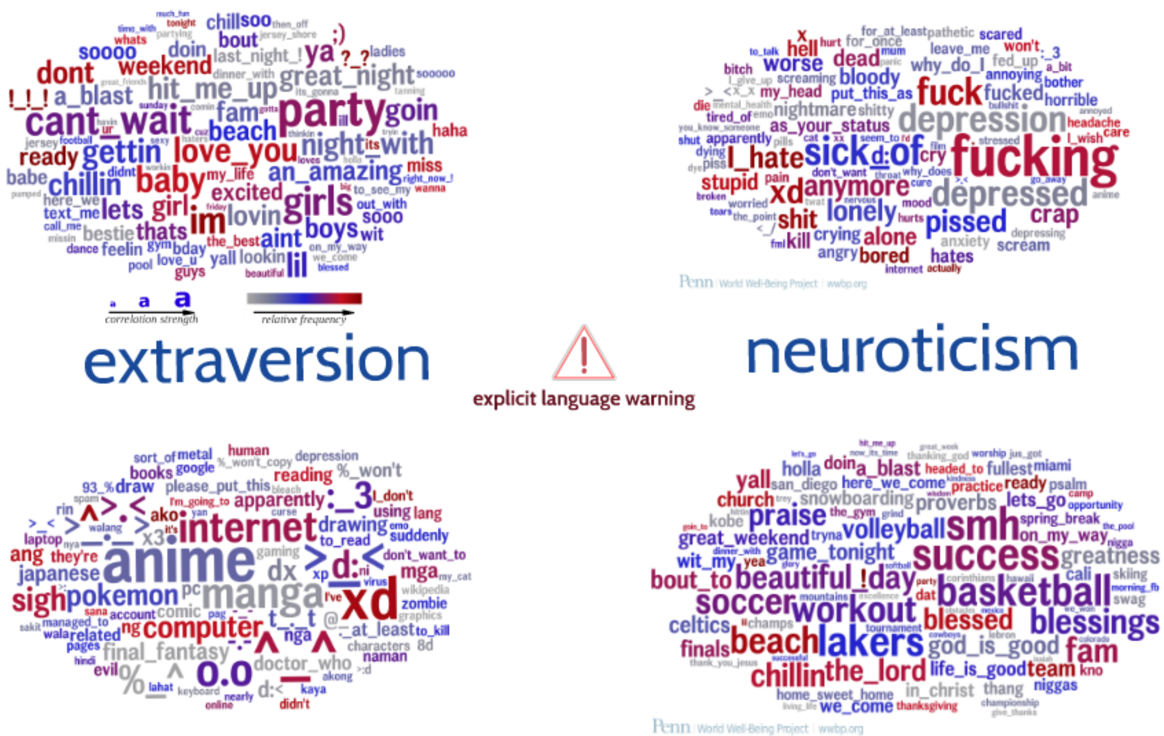


Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Toward Personality Insights from Language Exploration in Social Media. In *Proceedings of the AAAI Spring Symposium Series.* Stanford, CA.

for_at_least
pathetic
scared
x
hurt
for_once
hell
to_talk
mum
leave_me
won't
:_3
dead
worse
panic
why_do_I
fed_up
bitch
a_bit
annoying
bloody
screaming
I_give_up
bother
fuck
fucked
>_<
x_x
my_head
put_this_as
horrible
die
mental_health
bullshit
•
nightmare
shitty
annoyed
tired_of
emo
depression
headache
you_know_someone
as_your_status
shut
apparently
care
pills
cat
•
xx
seem_to
I'd
stressed
I_wish
*
film
dying
sick
d:
of
fucking
I_hate
cry
piss
dye
don't_want
throat
why_does
stupid
pain
>.<
go_away
xd
anymore
cure
depressed
anime
broken
twat
nervous
mood
worried
lonely
crap
tears
the_point
hurts
shit
pissed
<_/
depressing
crying
alone
anxiety
fml
kill
scream
angry
bored
hates
internet
actually
Penn | World Well-Being Project | wwbp.org

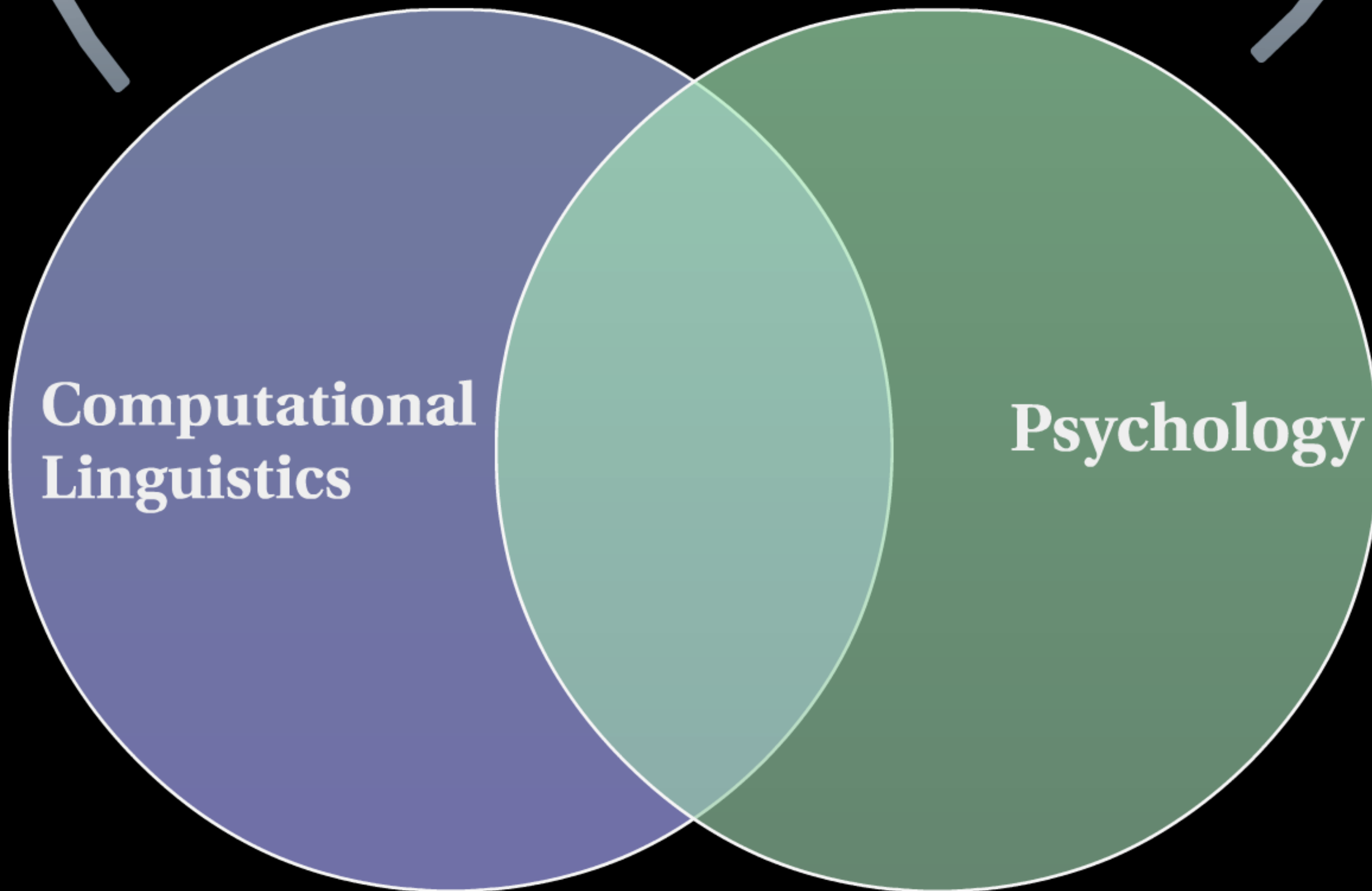
rning

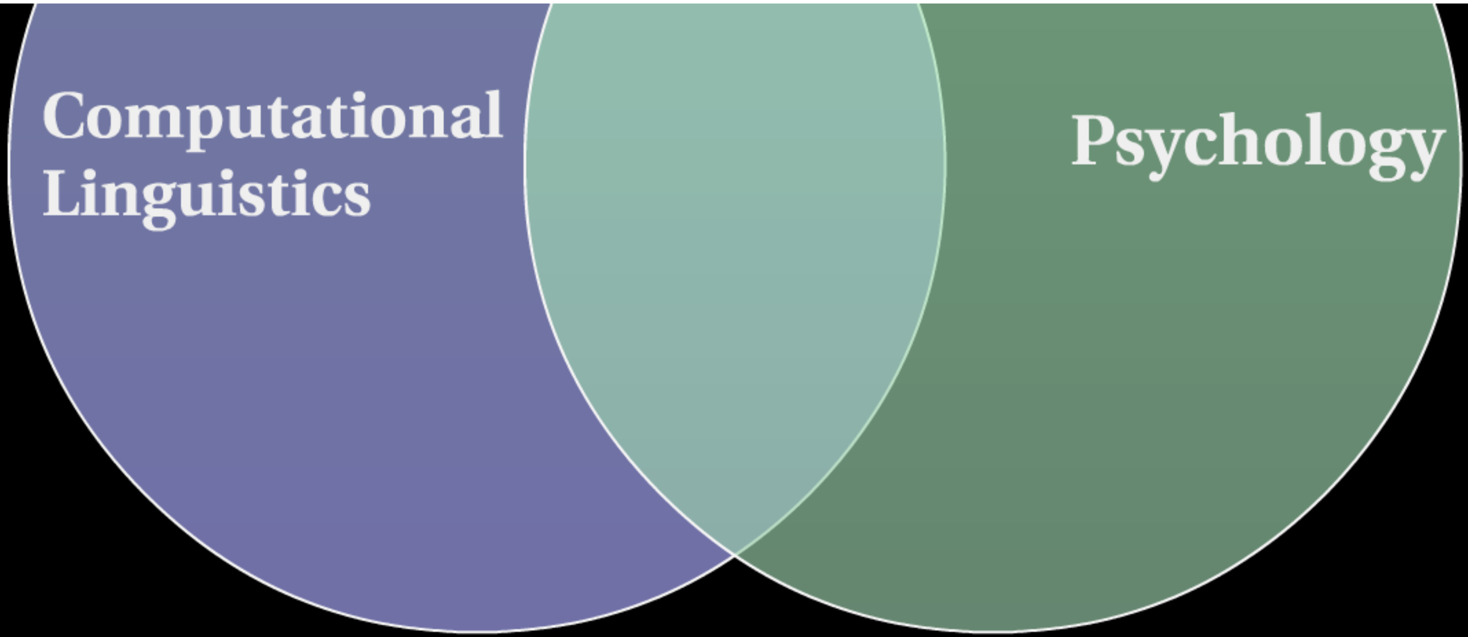

hit_me_up
great_week
let's_go
now_its_time
thanking_god
worship
jus_got
holla
doin
a_blast
headed_to
fullest
miami
yall
san_diego
here_we_come
kindness
practice
ready
psalm
church
trey
wisdom
snowboarding
proverbs
lets_go
camp
hittin
praise
the_gym
grind
opportunity
kobe
smh
spring_break
goin_to
volleyball
the_pool
great_weekend
tryna
on_my_way
nigga
dinner_with
game_tonight
success
greatness
wit_my
yea
glory
softball
bout_to
beautiful
!
day
party
corinthians
hawaii
cali
skiing
morning_fb
basketball
soccer
mountains
excellence
dat
swag
workout
obstacles
mexico
we_won
celtics
lil
champs
blessed
blessings
tournament
beach
lakers
cowboys
lebron
finals
god_is_good
colorado
fam
successful
thank_you_jesus
isaiah
chillin
the_lord
life_is_good
team
kno
home_sweet_home
in_christ
thang
niggas
living_life
we_come
thanksgiving
championship
give_thanks



Schwartz, H. A., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Blanco, E., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Toward Personality Insights from Language Exploration in Social Media. In *Proceedings of the AAAI Spring Symposium Series.* Stanford, CA.

from *modeling language*
to *understanding people*





Interdisciplinary
Research Questions

Can we predict disease risk and recovery from language use?

What psychological factors emerge in language as drivers of health and well-being?

e use?

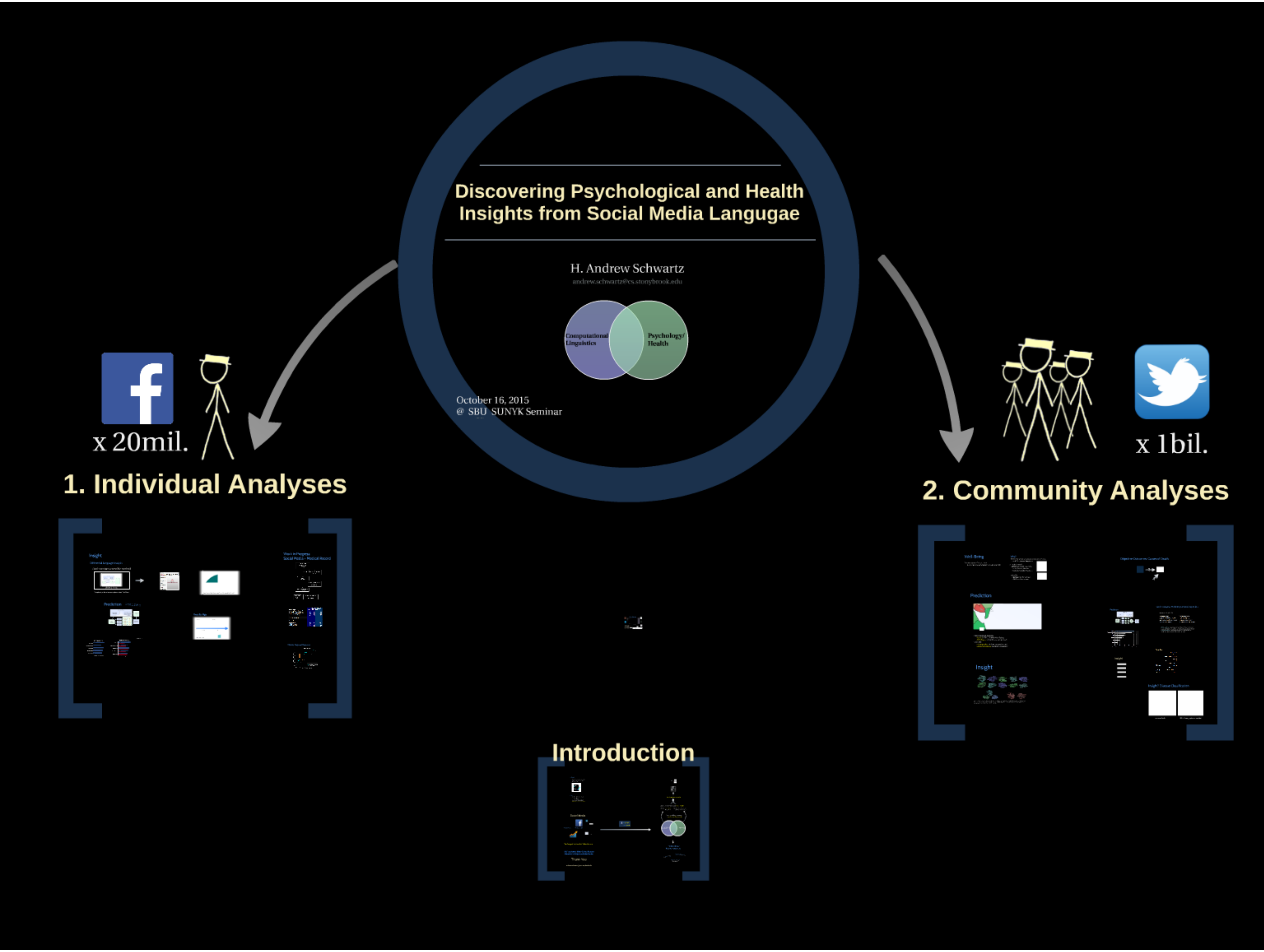
To what extent can language analyses
replace and extend traditional
psychological asesseement

Interdisciplinary Research Questions

What psychological factors emerge in language as drivers of health and well-being?

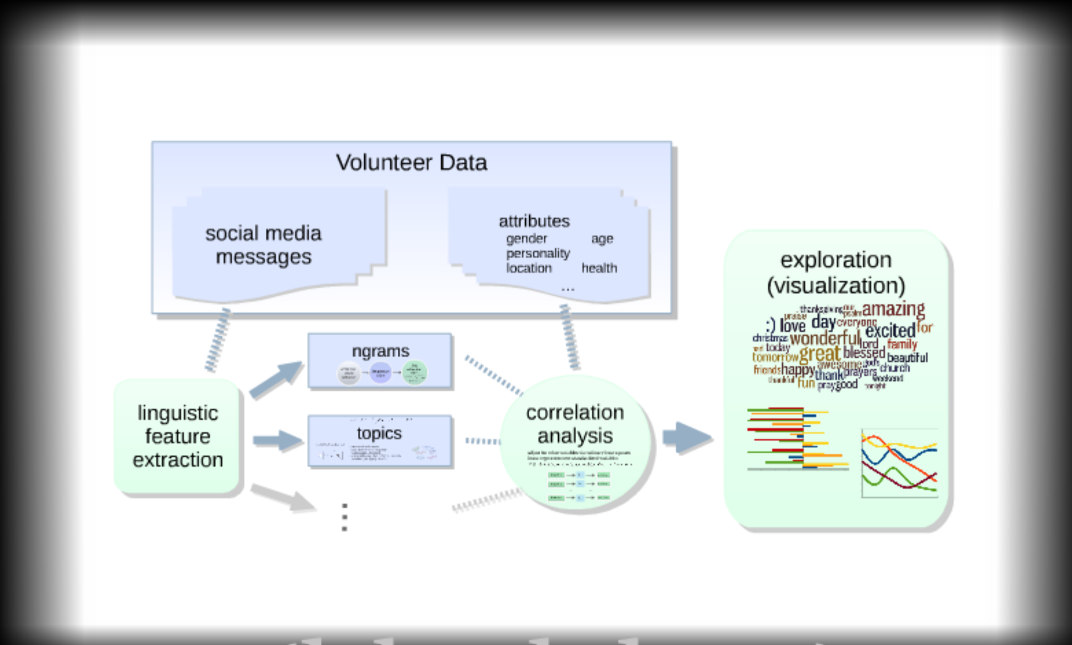
Can we predict disease risk and recovery from language use?

To what extent can language analyses replace and extend traditional psychological asessement


Discovering Psychological and Health Insights from Social Media Langugae
H. Andrew Schwartz
andrew.schwartz@cs.stonybrook.edu
Computational Linguistics
Psychology/ Health
October 16, 2015
@ SBU SUNYK Seminar
x 20mil.
1. Individual Analyses
x 1bil.
2. Community Analyses
Introduction

Differential Language Analysis

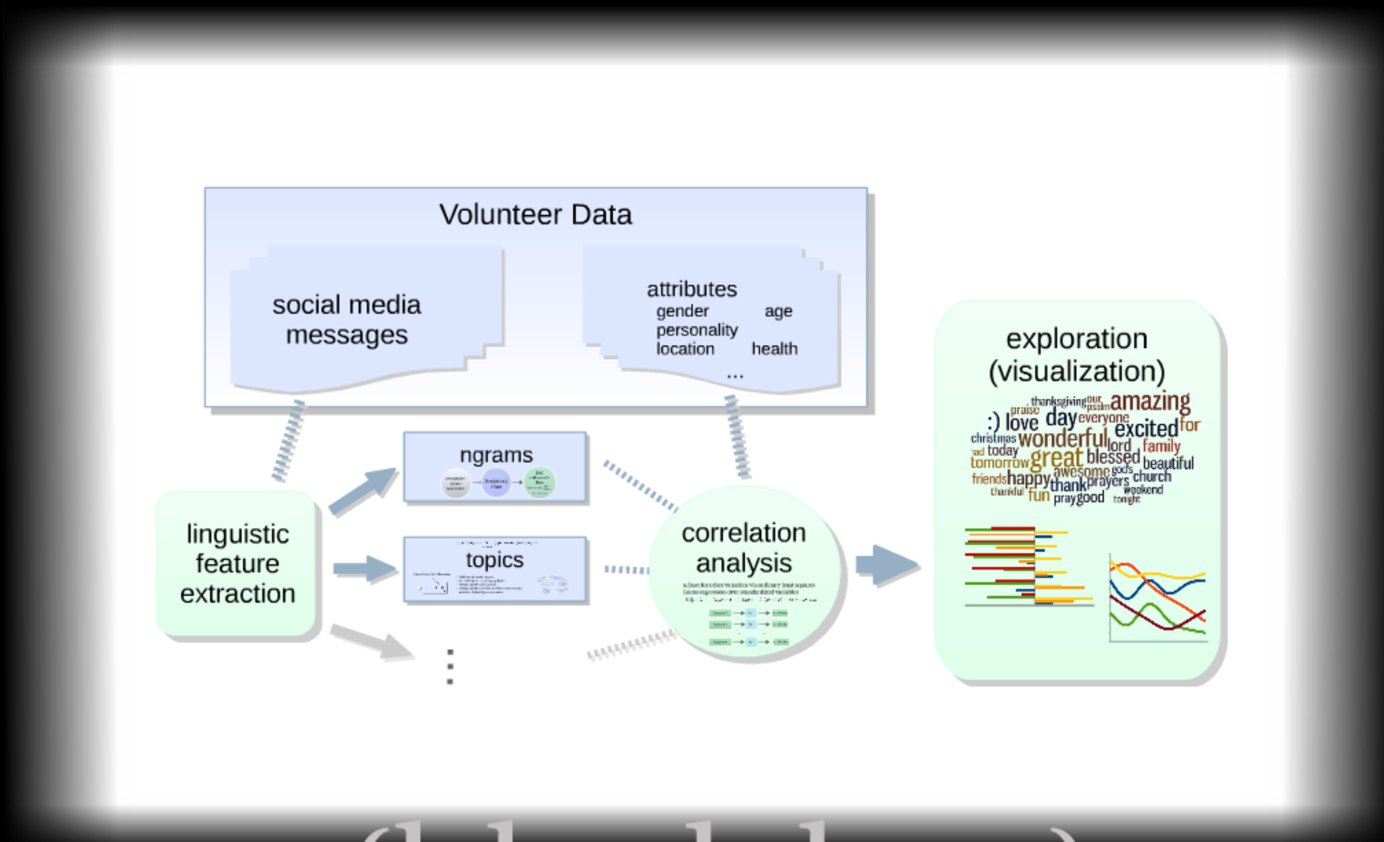
Goal: succinct accessible method



(black box)

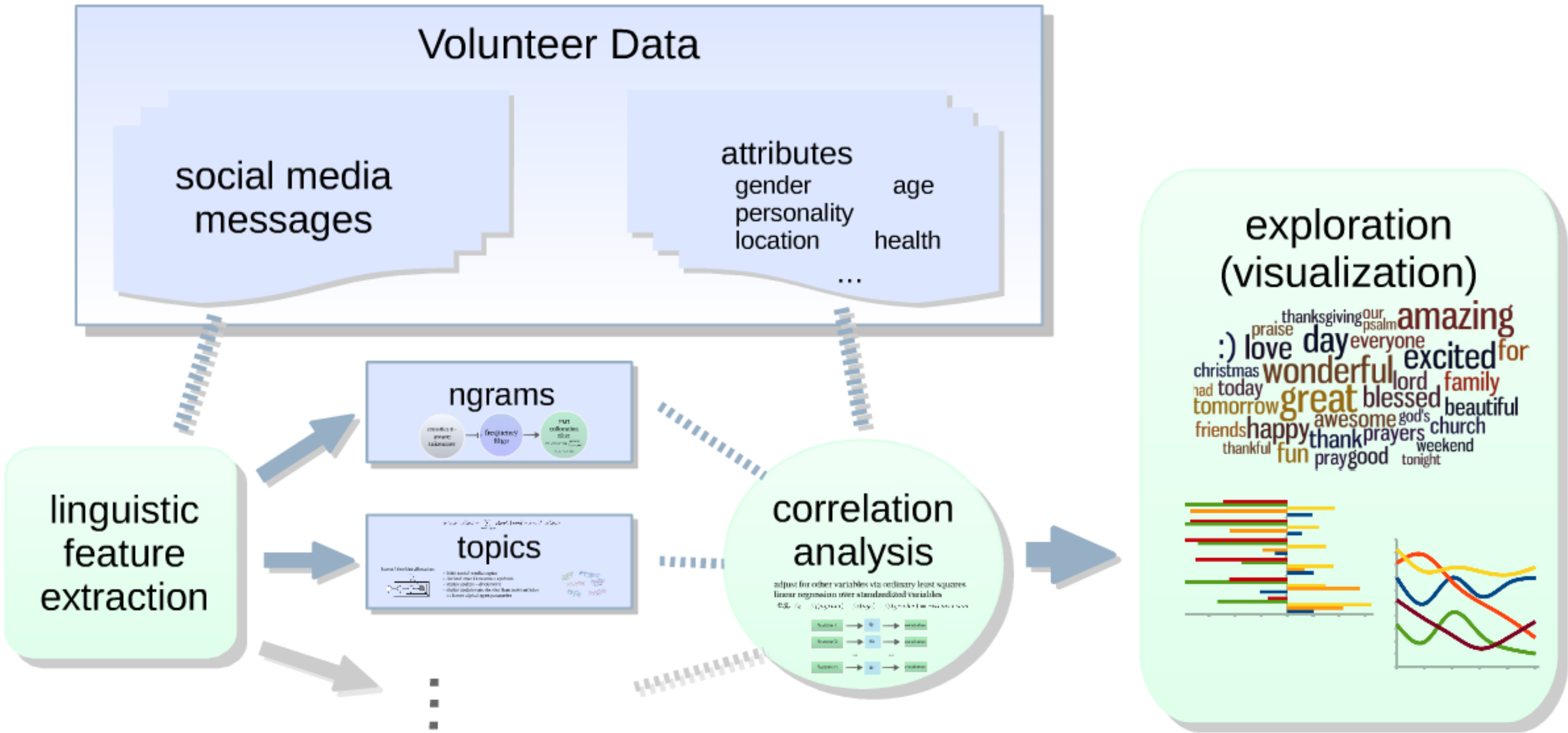
"Simplicity is the ultimate sophistication" -da Vinci

Goal: succinct accessible metho



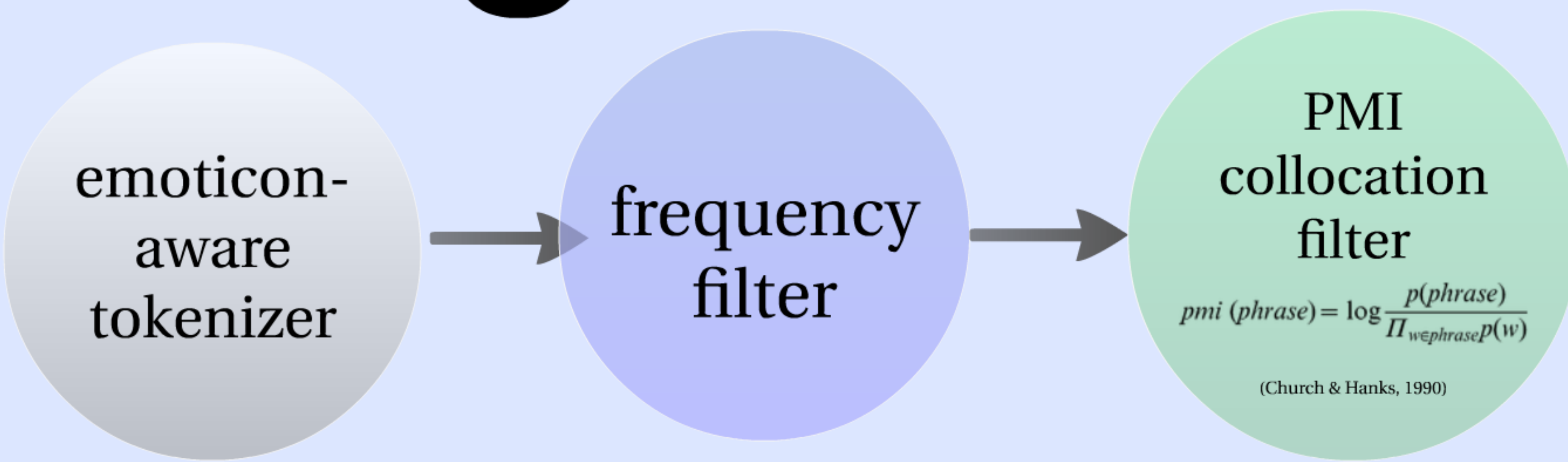
(black box)

"Simplicity is the ultimate sophistication" -da Vinci


Volunteer Data
social media messages
attributes
gender
age
personality
location
health
...
ngrams
topics
linguistic feature extraction
correlation analysis
exploration (visualization)

(black box)

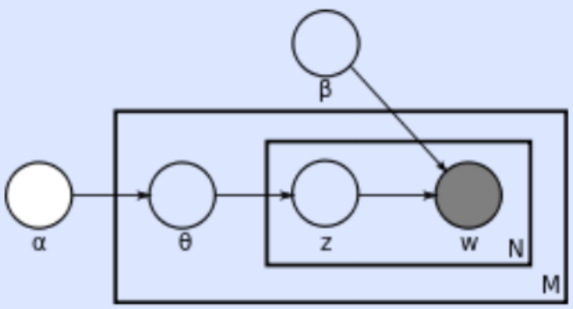
ngrams


emoticon-aware tokenizer
frequency filter
PMI collocation filter
$pmi\ (phrase) = \log \frac{p(phrase)}{\Pi_{w \in phrase} p(w)}$
(Church & Hanks, 1990)

$$p(topic \mid subject) = \sum_{word \in topic} p(topic \mid word) \cdot p(word$$

topics

latent Dirichlet allocation



- 2000 social media topics
- derived over 14m status updates
- status update = document
- status updates are shorter than news articles => lower alpha hyperparameter

chicken cheese dinner yum
soup made rice bacon bread
yummy fried eating salad
cooking eggs sauce making
eat potatoes

on friday monday sunday saturday tuesday
thursday night wednesday weekend next
week until morning afternoon tomorrow till
working nights

my mom dad with husband
wife parents son daughter
kids love ex mum brother
wonderful hubby sister
boyfriend mother

snow cold weather outside warm hot its
degrees winter heat freezing snowing
ice here inside inches summer degree
storm

party tonight at halloween
birthday night fun club bar
dance costume bday parties
come saturday via dj sms
house

pain blood hospital teeth surgery
doctor from tooth dentist wisdom
after has brain having doctors had
doc pulled pressure

wonderful hubby sister
boyfriend mother

pain blood hospital teeth surgery
doctor from tooth dentist wisdom
after has brain having doctors had
doc pulled pressure

made ne
mmy fried eating
oking eggs sauce making
eat potatoes

ith morn
wo

snow cold weather outside warm hot its degrees winter heat freezing snowing ice here inside inches summer degree storm

party tonight at halloween
birthday night fun club
dance costu

$$p(topic \mid subject) = \sum_{word \in topic} p(topic \mid word) * p(word \mid subject)$$

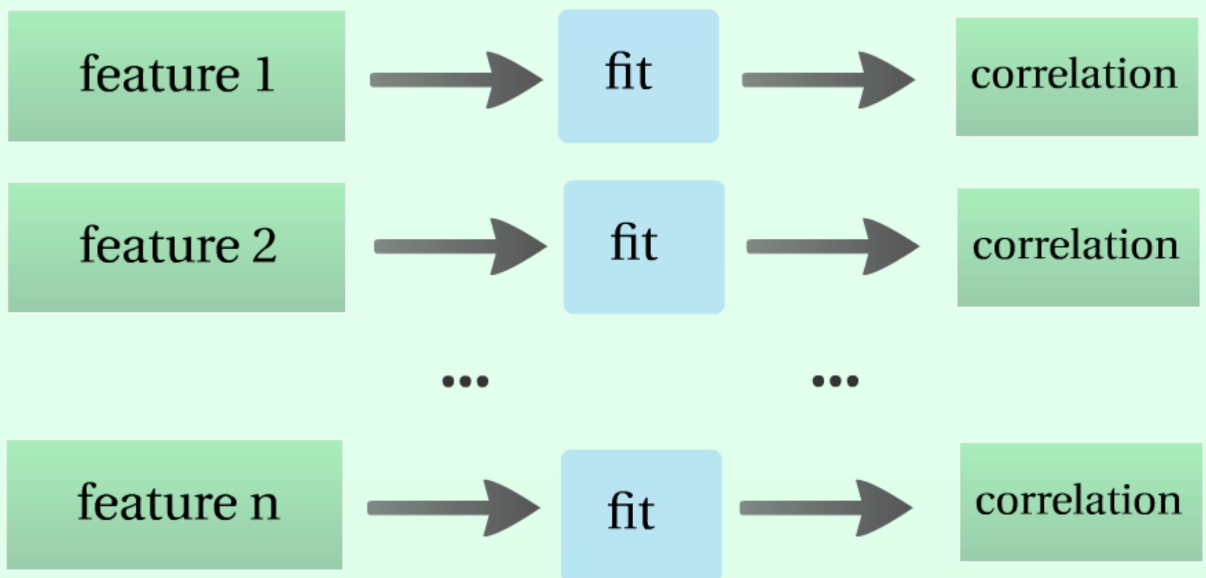
topics

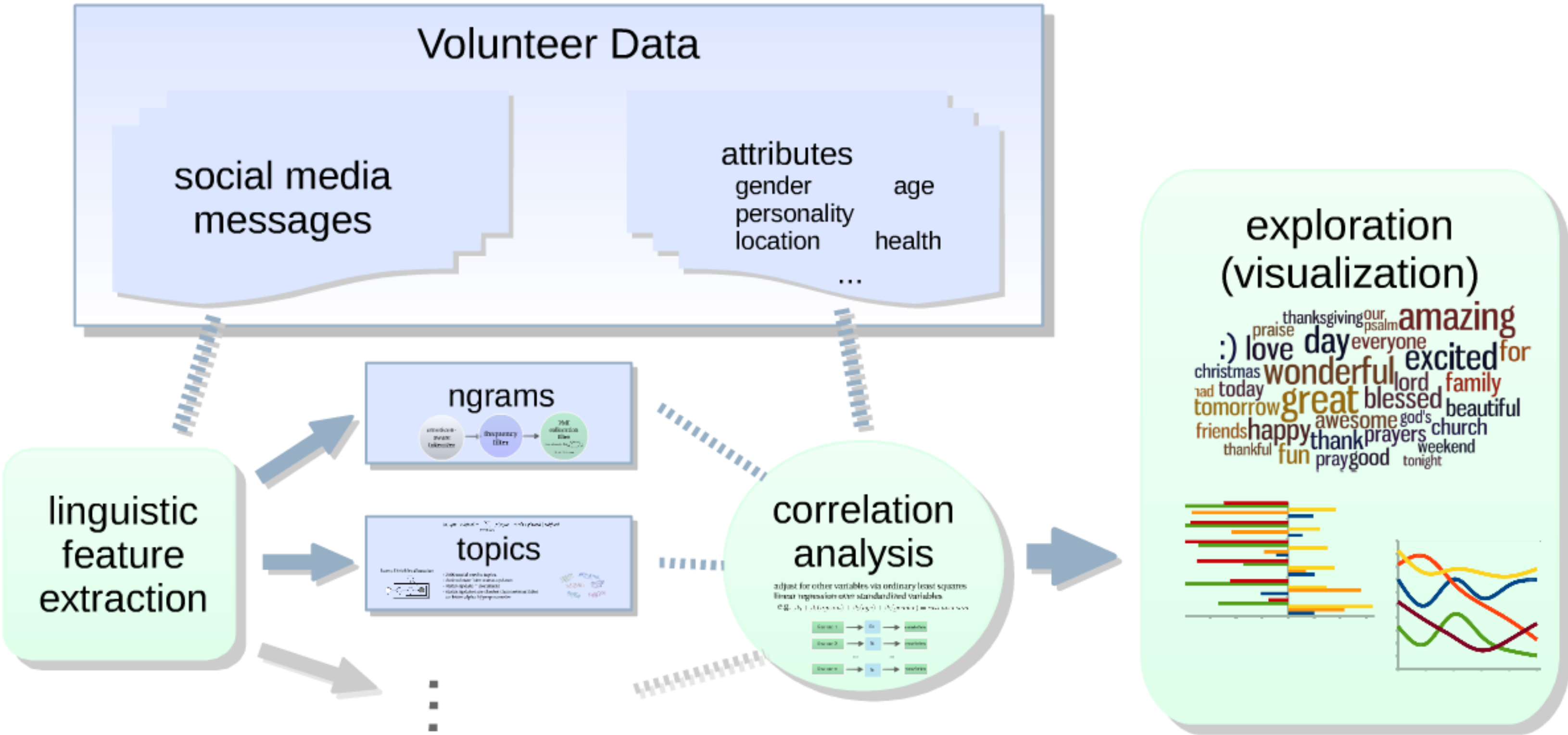
- 2000 social media topics

analysis

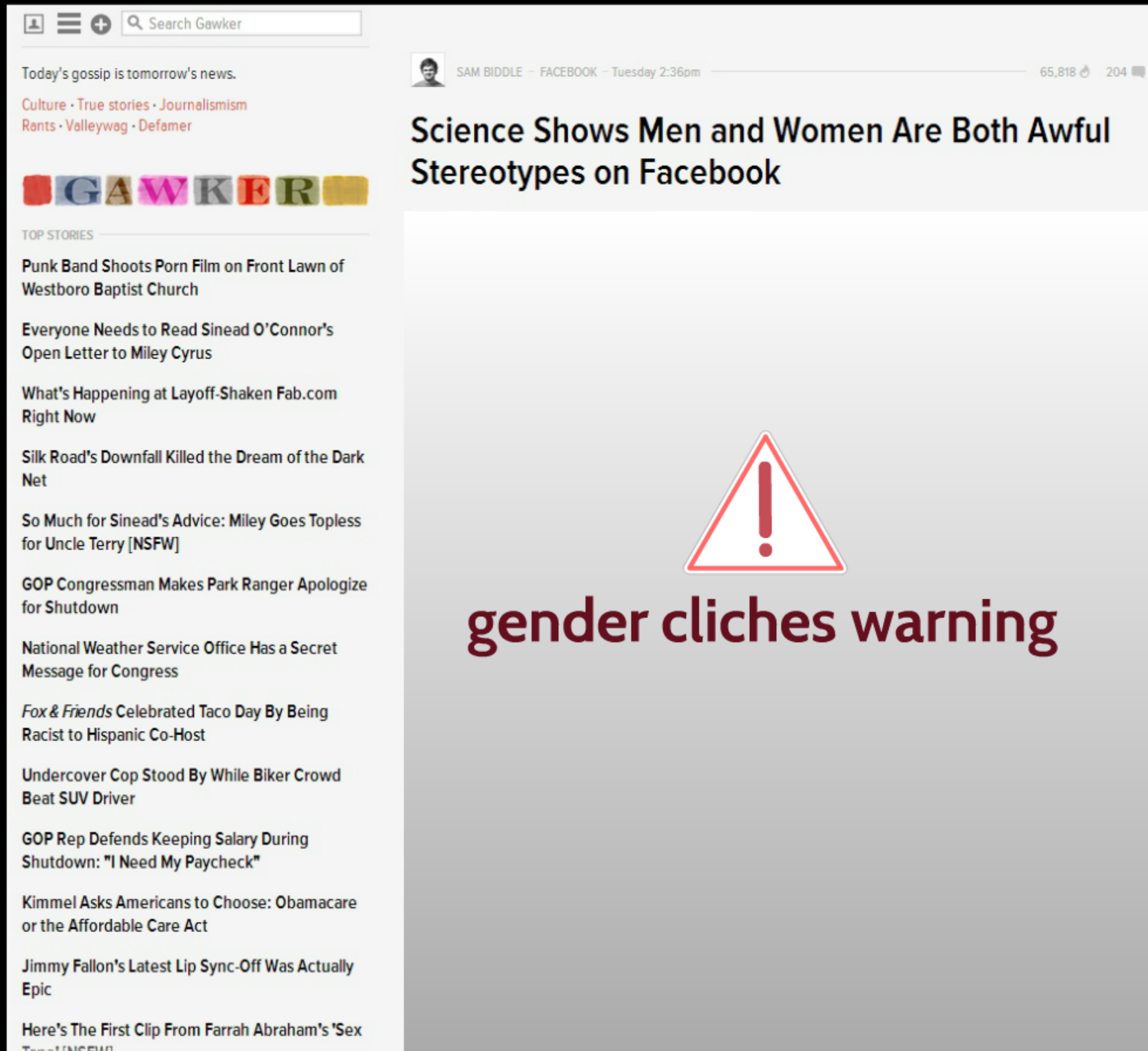
adjust for other variables via ordinary least squares linear regression over standardized variables

e.g. $\beta_0 + \beta_1(ngram) + \beta_2(age) + \beta_3(gender) = extraversion$




Volunteer Data
social media messages
attributes
gender
age
personality
location
health
...
ngrams
topics
linguistic feature extraction
correlation analysis
exploration (visualization)

(black box)



Search Gawker

Today's gossip is tomorrow's news.

Culture · True stories · Journalismism
Rants · Valleywag · Defamer

GAWKER

TOP STORIES

Punk Band Shoots Porn Film on Front Lawn of Westboro Baptist Church

Everyone Needs to Read Sinead O'Connor's Open Letter to Miley Cyrus

What's Happening at Layoff-Shaken Fab.com Right Now

Silk Road's Downfall Killed the Dream of the Dark Net

So Much for Sinead's Advice: Miley Goes Topless for Uncle Terry [NSFW]

GOP Congressman Makes Park Ranger Apologize for Shutdown

National Weather Service Office Has a Secret Message for Congress

Fox & Friends Celebrated Taco Day By Being Racist to Hispanic Co-Host

Undercover Cop Stood By While Biker Crowd Beat SUV Driver

GOP Rep Defends Keeping Salary During Shutdown: "I Need My Paycheck"

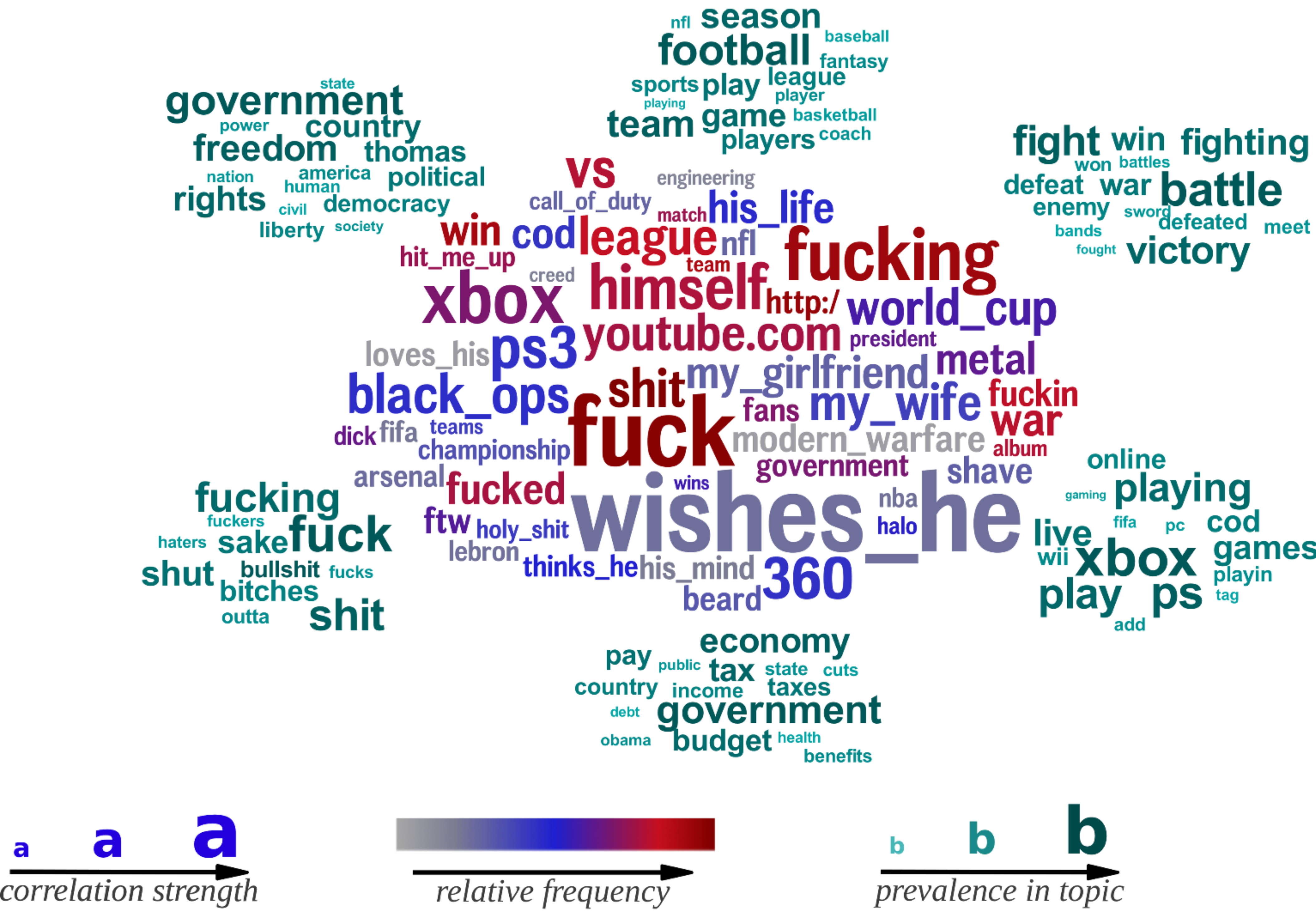
Kimmel Asks Americans to Choose: Obamacare or the Affordable Care Act

Jimmy Fallon's Latest Lip Sync-Off Was Actually Epic

Here's The First Clip From Farrah Abraham's 'Sex Tape' [NSFW]

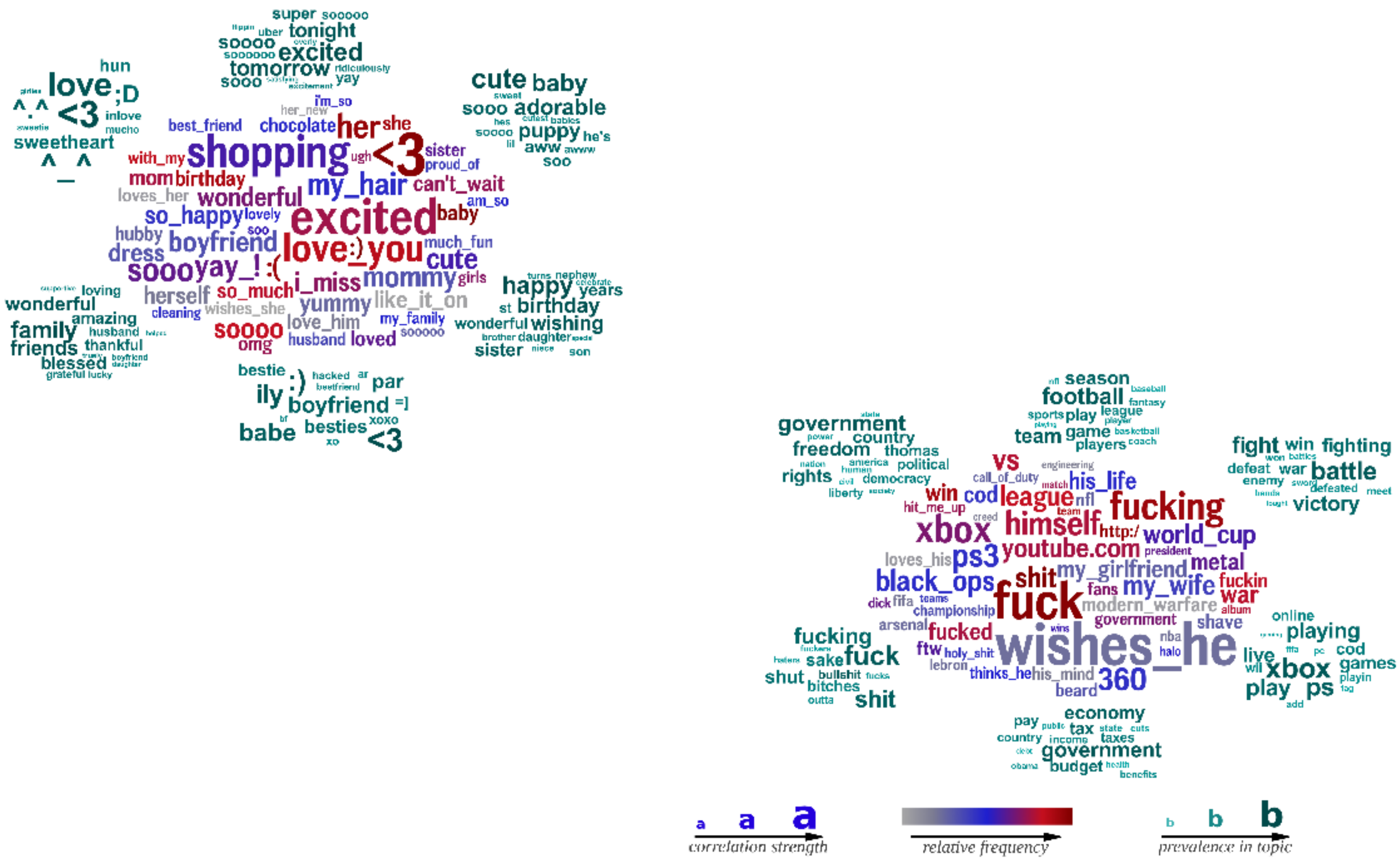
SAM BIDDLE · FACEBOOK · Tuesday 2:36pm · 65,818 · 204

Science Shows Men and Women Are Both Awful Stereotypes on Facebook



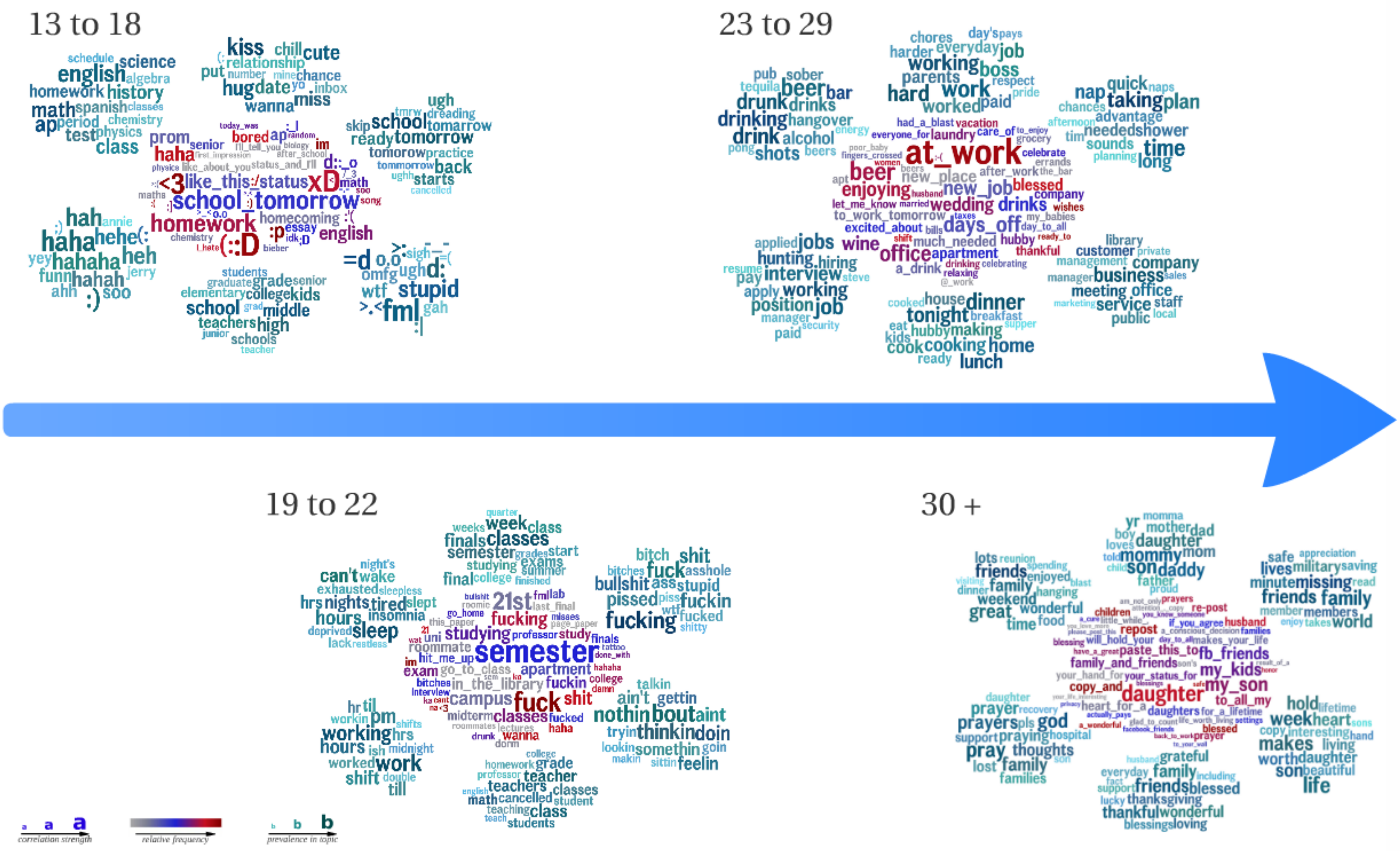
government
state
power
country
freedom
thomas
nation
america
human
political
rights
civil
democracy
liberty
society
nfl
season
football
baseball
fantasy
sports
play
league
player
playing
team
game
basketball
players
coach
fight
win
fighting
won
battles
defeat
war
battle
enemy
sword
bands
defeated
meet
fought
victory
vs
engineering
call_of_duty
his_life
match
win
cod
league
nfl
fucking
hit_me_up
team
creed
xbox
himself
http:/
world_cup
youtube.com
president
loves_his
ps3
metal
my_girlfriend
black_ops
shit
fuckin
my_wife
fans
war
dick
fifa
teams
fuck
modern_warfare
album
championship
government
arsenal
shave
fucked
wins
nba
wishes_he
ftw
holy_shit
halo
lebron
thinks_he
his_mind
360
beard
fucking
fuckers
haters
sake
fuck
shut
bullshit
fucks
bitches
shit
outta
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
correlation strength
relative frequency
b
b
b
prevalence in topic


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



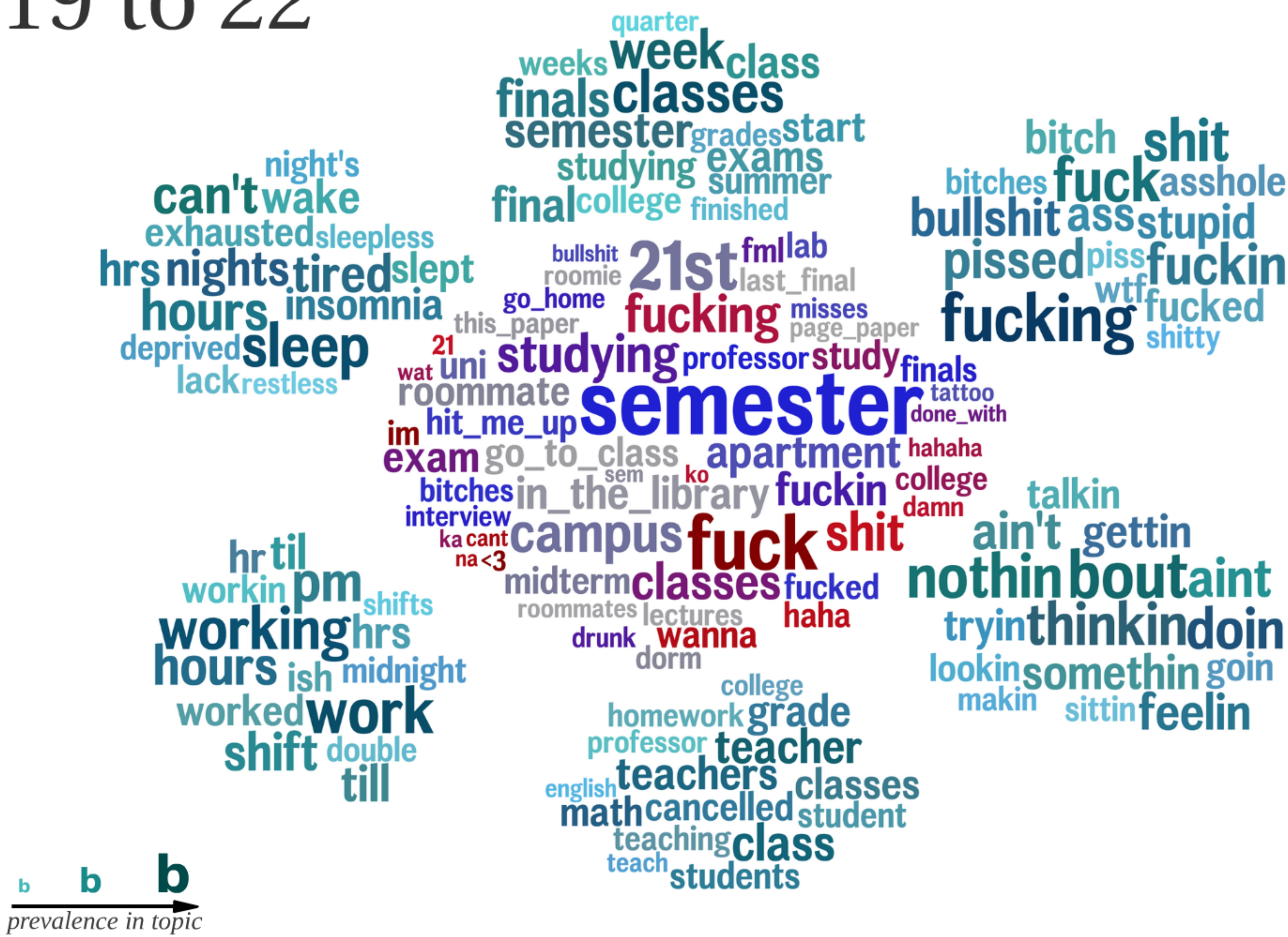
Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. In PLOS ONE 8(9).

Results: Age



13 to 18
23 to 29
19 to 22
30 +
correlation strength
relative frequency
prevalence in topic

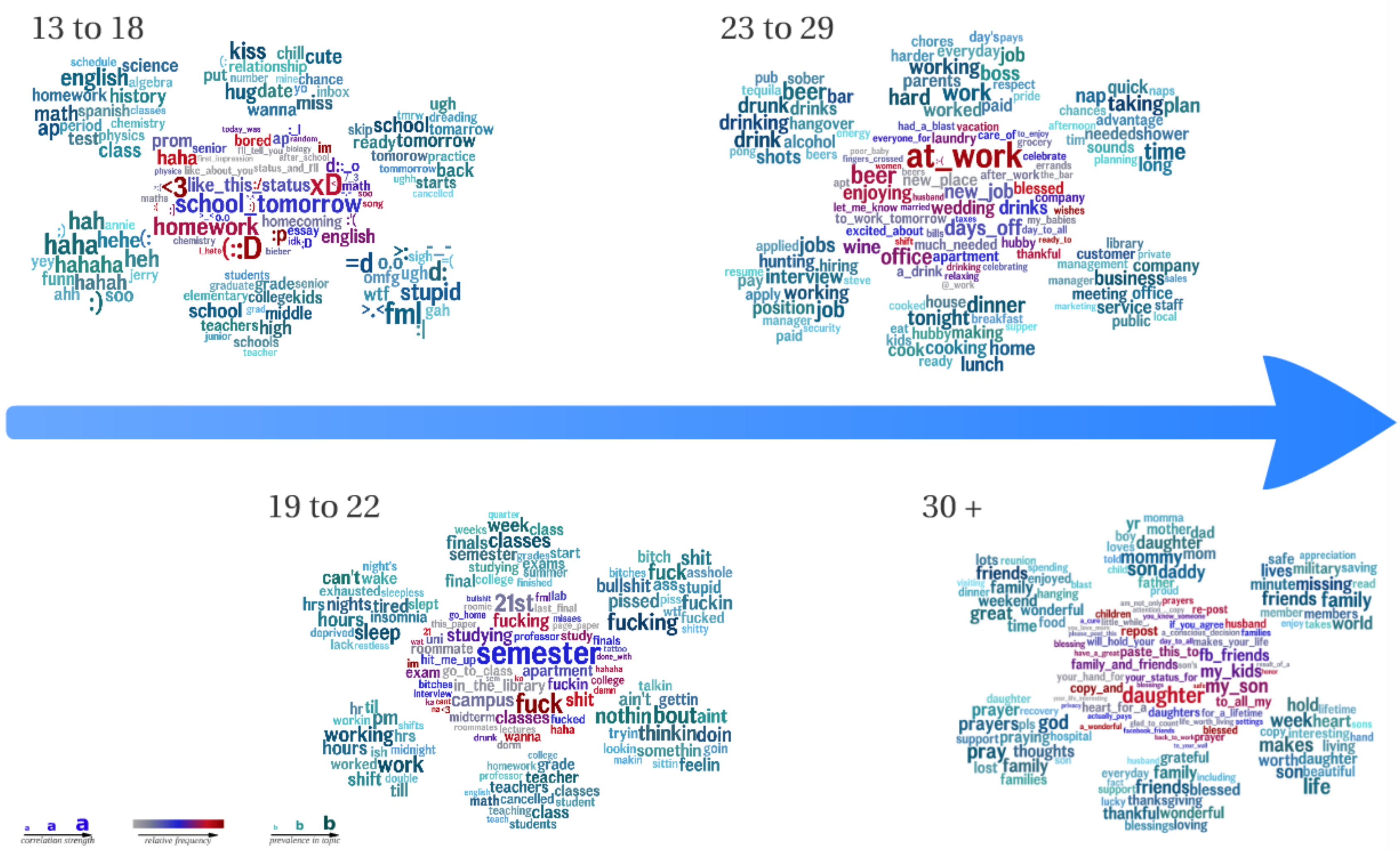
19 to 22


quarter weeks week class finals classes semester grades start studying exams summer final college finished
night's can't wake exhausted sleepless hrs nights tired slept hours insomnia deprived sleep lack restless
bitch shit bitches fuck asshole bullshit ass stupid pissed piss fuckin wtf fucked fucking shitty
bullshit 21st fml lab roomie last_final go_home fucking misses this_paper page_paper 21 wat uni studying professor study finals tattoo semester done_with roommate im hit_me_up exam go_to_class apartment hahaha sem ko bitches in_the_library fuckin college damn interview ka cant na <3 campus fuck shit midterm classes fucked roommates lectures haha drunk wanna dorm
hr til workin pm shifts working hrs hours ish midnight worked work shift double till
talkin ain't gettin nothin bout aint tryin thinkin doin lookin somethin goin makin sittin feelin
college homework grade professor teacher english teachers classes math cancelled student teaching class teach students
b b b
prevalence in topic

23 to 29

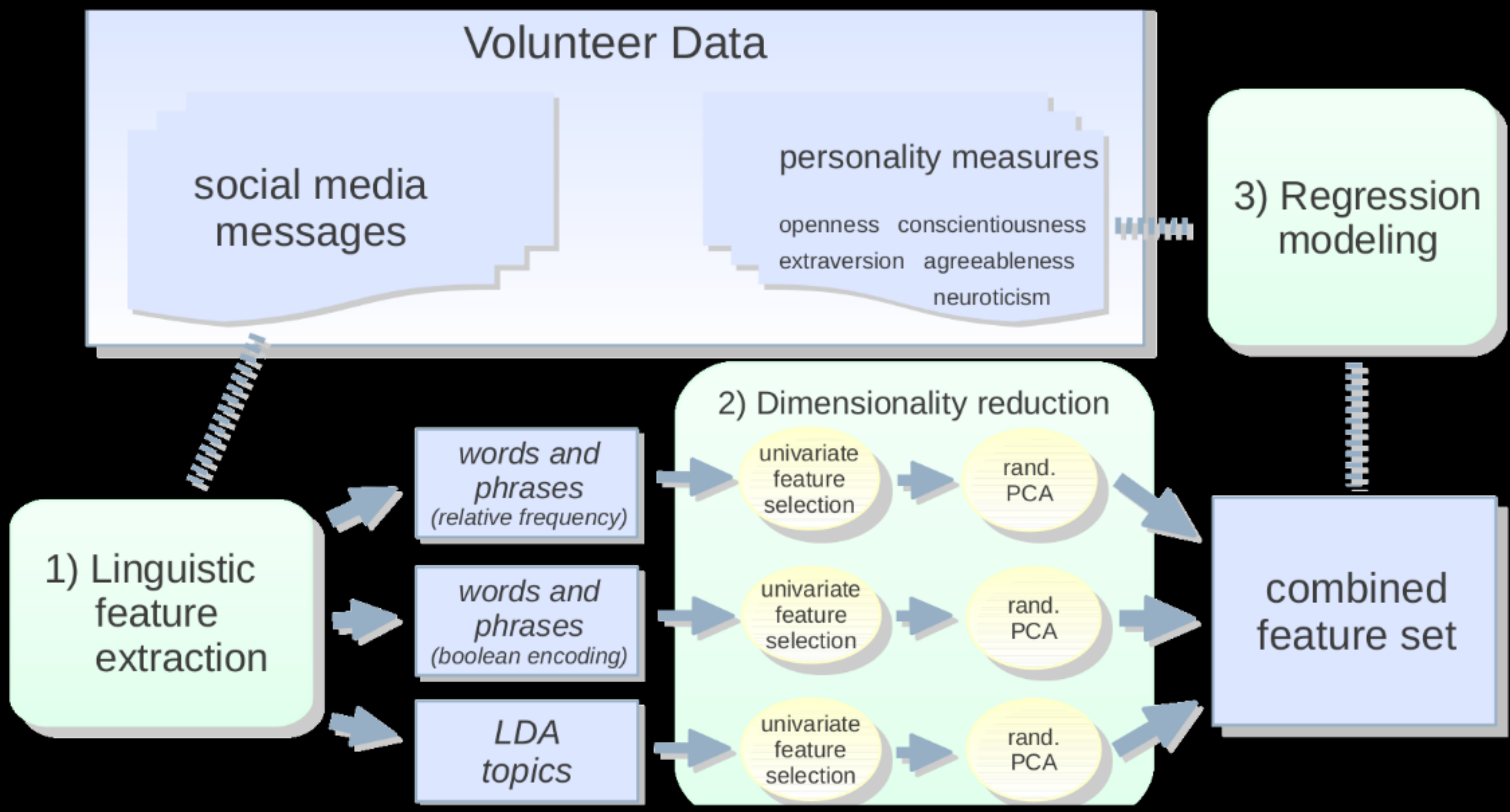
chores
day's
pays
harder
everyday
job
working
boss
parents
respect
hard
work
pride
worked
paid
pub
sober
tequila
beer
bar
drunk
drinks
drinking
hangover
energy
drink
alcohol
pong
shots
beers
nap
quick
naps
taking
plan
chances
advantage
afternoon
tim
needed
shower
sounds
time
planning
long
had_a_blast
vacation
everyone_for
laundry
care_of
to_enjoy
grocery
poor_baby
fingers_crossed
at_work
:-(
celebrate
women
errands
beer
beers
apt
new_place
after_work
the_bar
enjoying
husband
new_job
blessed
company
let_me_know
married
wedding
drinks
wishes
to_work_tomorrow
taxes
my_babies
excited_about
bills
days_off
day_to_all
shift
much_needed
hubby
ready_to
wine
office
apartment
thankful
drinking
celebrating
a_drink
relaxing
@_work
applied
jobs
hunting
hiring
resume
pay
interview
steve
apply
working
position
job
manager
security
paid
library
customer
private
management
company
manager
business
sales
meeting
office
marketing
service
staff
public
local
cooked
house
dinner
tonight
breakfast
eat
supper
hubby
making
kids
cook
cooking
home
ready
lunch

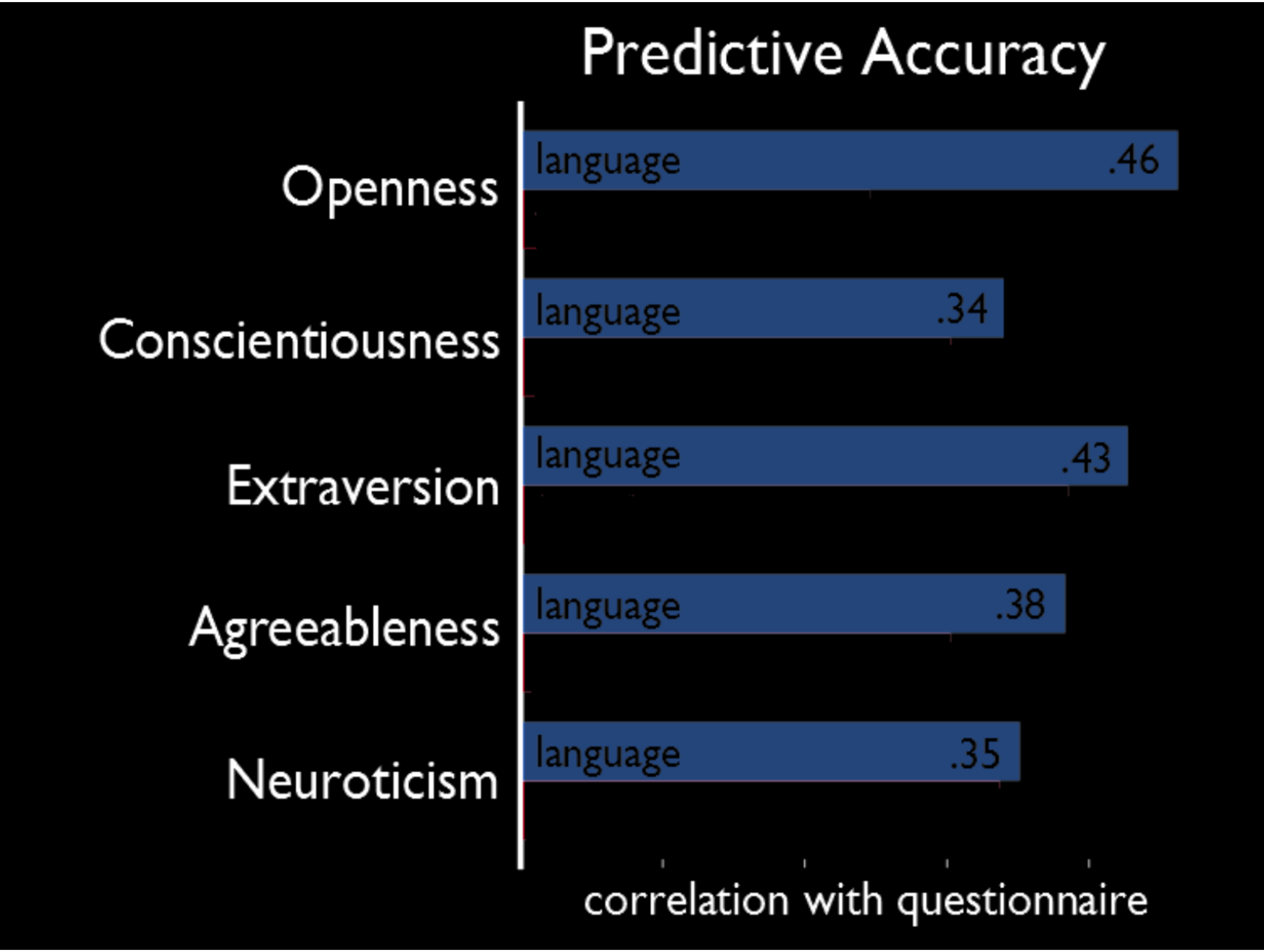
Results: Age


13 to 18
23 to 29
19 to 22
30 +
correlation strength
relative frequency
prevalence in topic

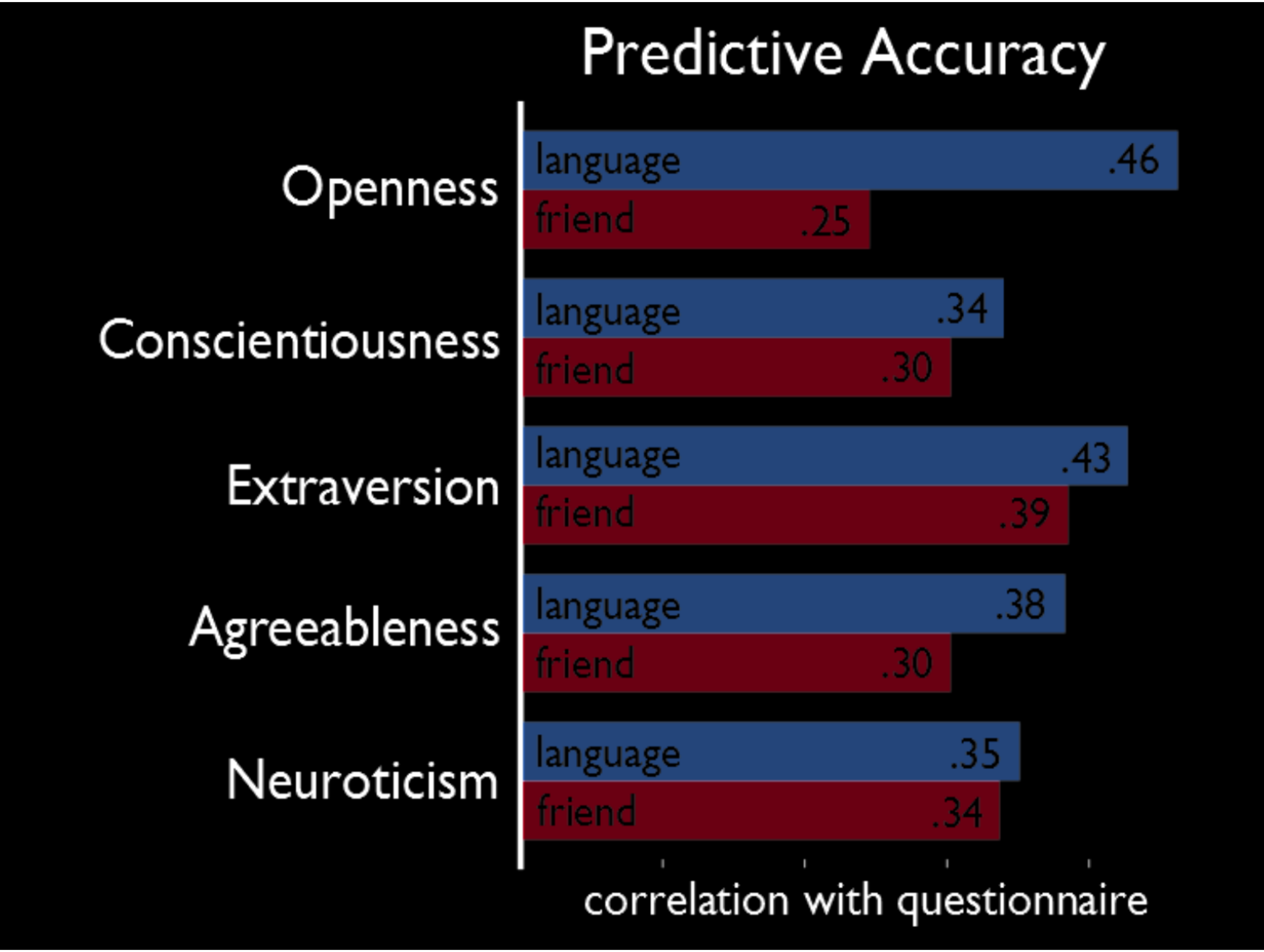
Prediction

To what extent can language-based assessments replace qustionnaires?



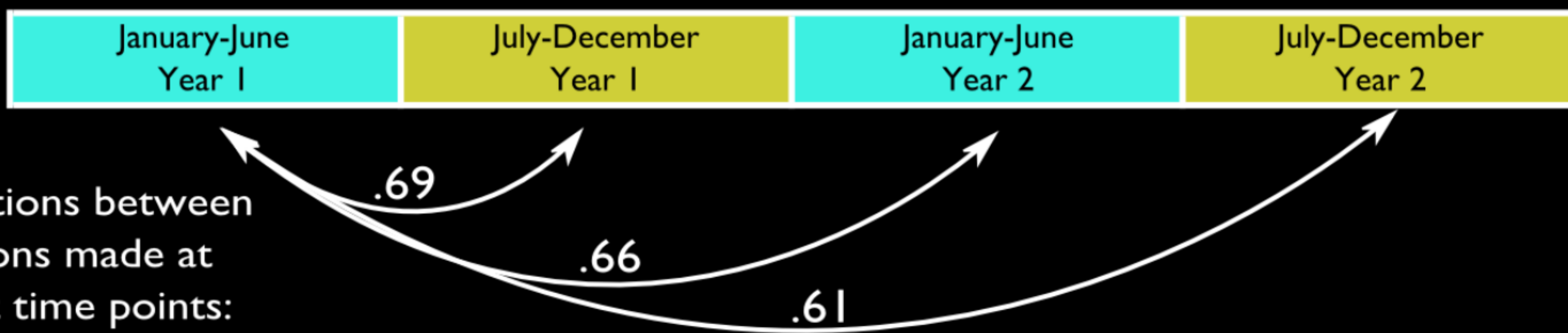


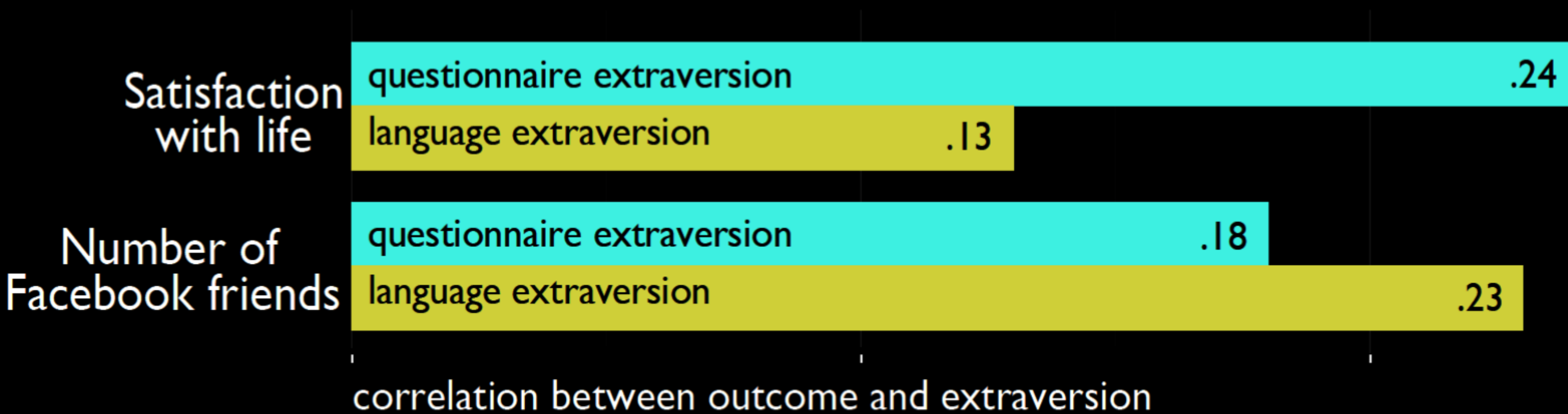
Predictive Accuracy
Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism
language .46
language .34
language .43
language .38
language .35
correlation with questionnaire

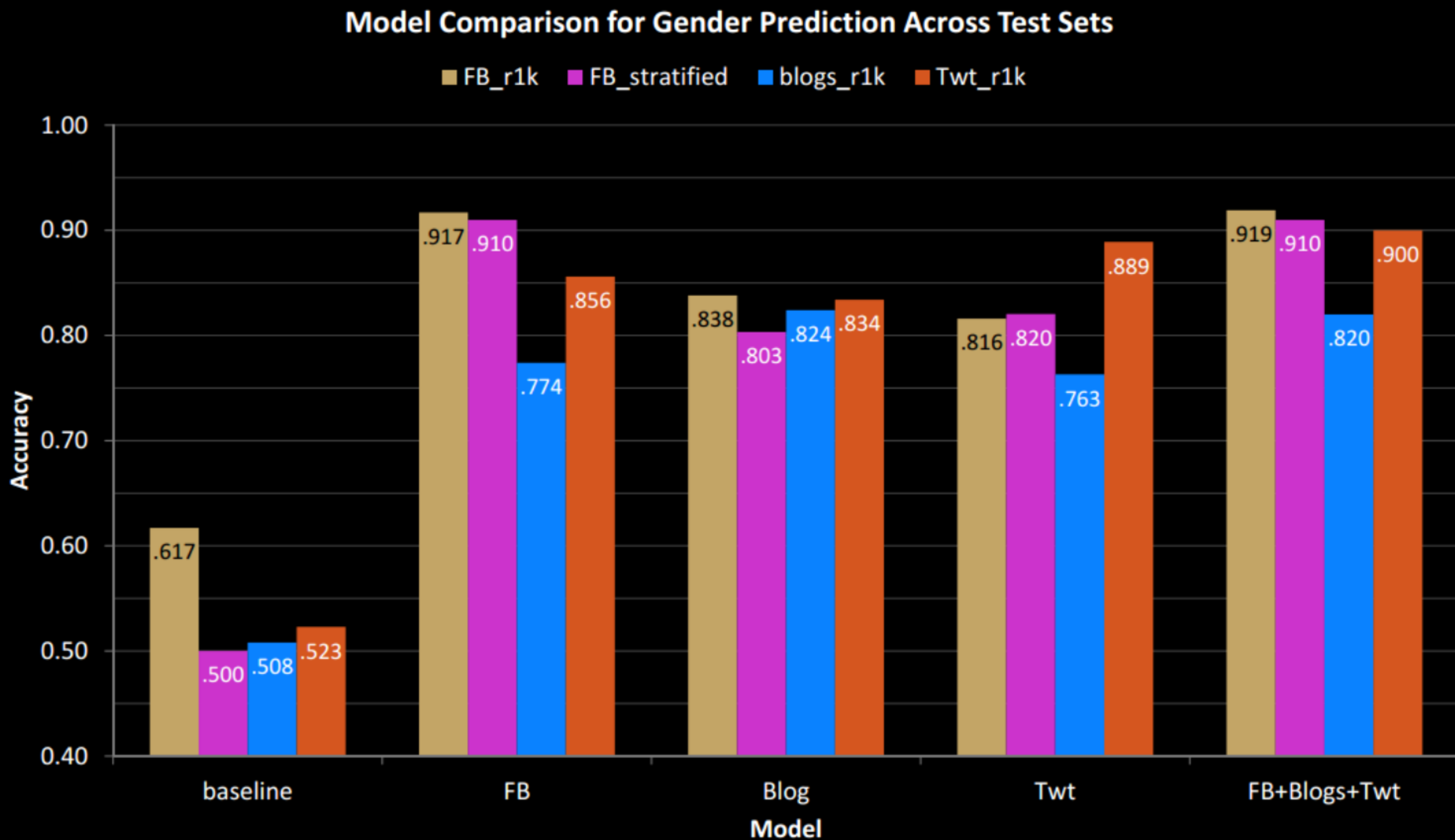

Predictive Accuracy
Openness
language .46
friend .25
Conscientiousness
language .34
friend .30
Extraversion
language .43
friend .39
Agreeableness
language .38
friend .30
Neuroticism
language .35
friend .34
correlation with questionnaire

extraversion

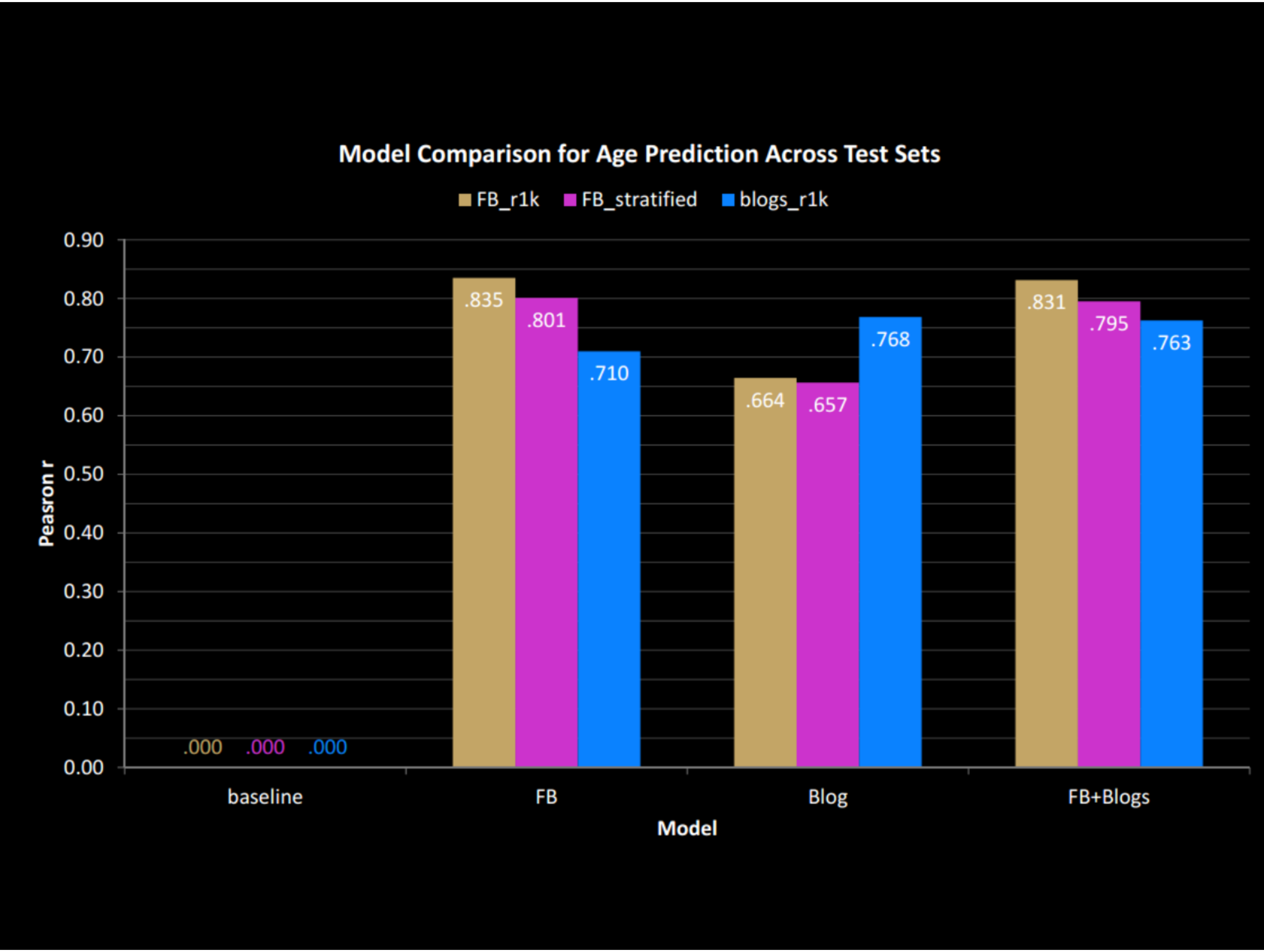
Correlations between predictions made at different time points:



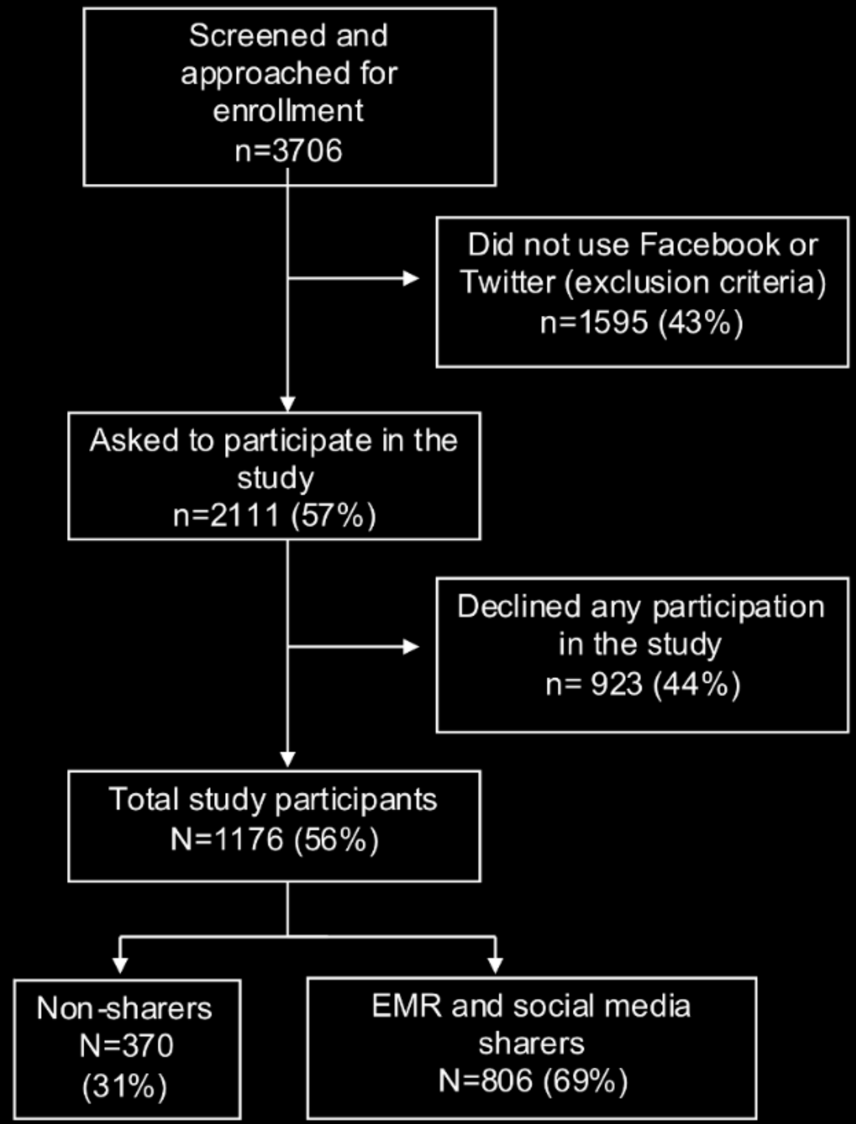

Satisfaction with life
questionnaire extraversion
.24
language extraversion
.13
Number of Facebook friends
questionnaire extraversion
.18
language extraversion
.23
correlation between outcome and extraversion



Model Comparison for Gender Prediction Across Test Sets
FB_r1k
FB_stratified
blogs_r1k
Twt_r1k
Accuracy
1.00
0.90
0.80
0.70
0.60
0.50
0.40
.617
.500
.508
.523
.917
.910
.774
.856
.838
.803
.824
.834
.816
.820
.763
.889
.919
.910
.820
.900
baseline
FB
Blog
Twt
FB+Blogs+Twt
Model

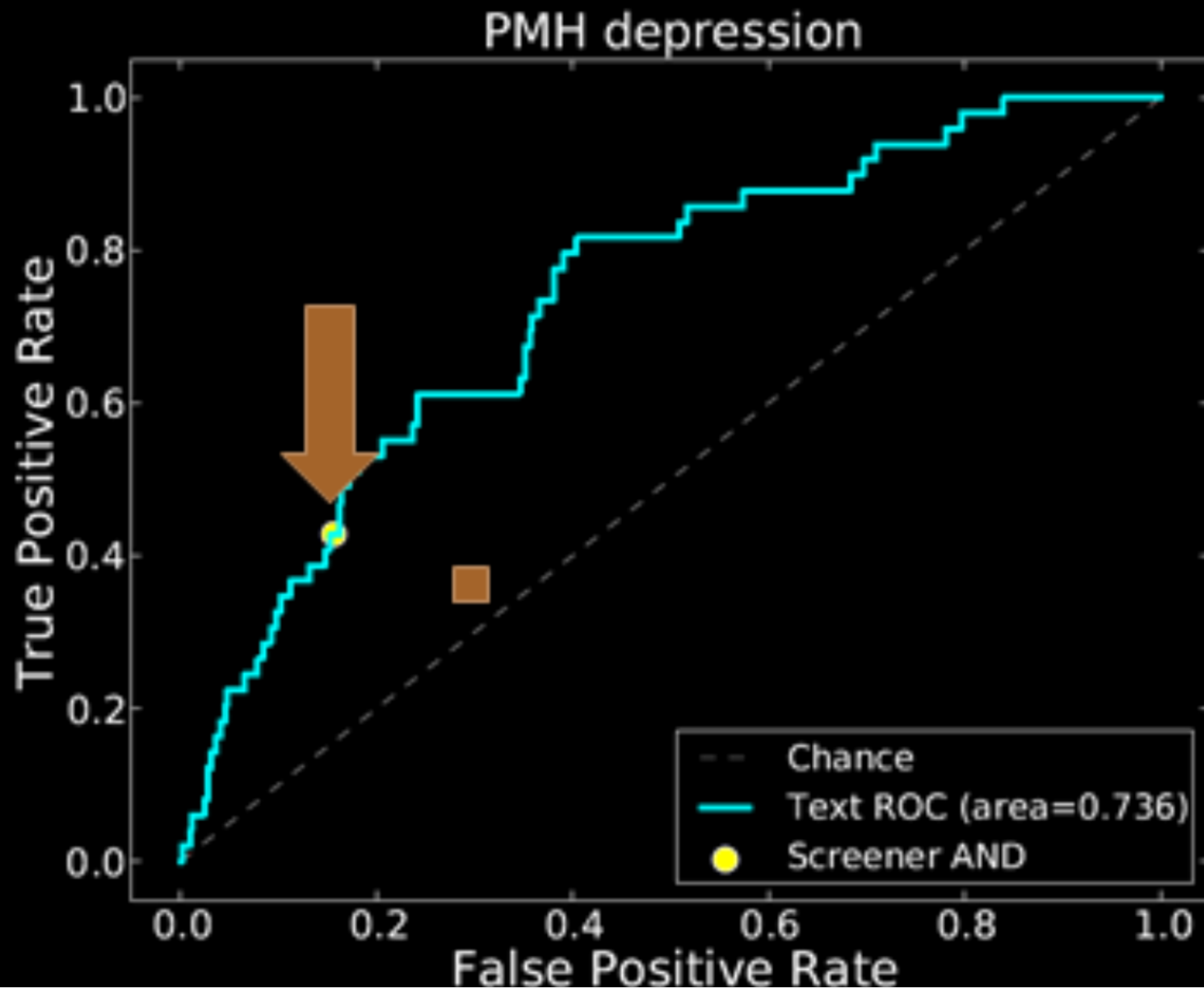

Model Comparison for Age Prediction Across Test Sets
FB_r1k
FB_stratified
blogs_r1k
Peasron r
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
.000
.000
.000
.835
.801
.710
.664
.657
.768
.831
.795
.763
baseline
FB
Blog
FB+Blogs
Model

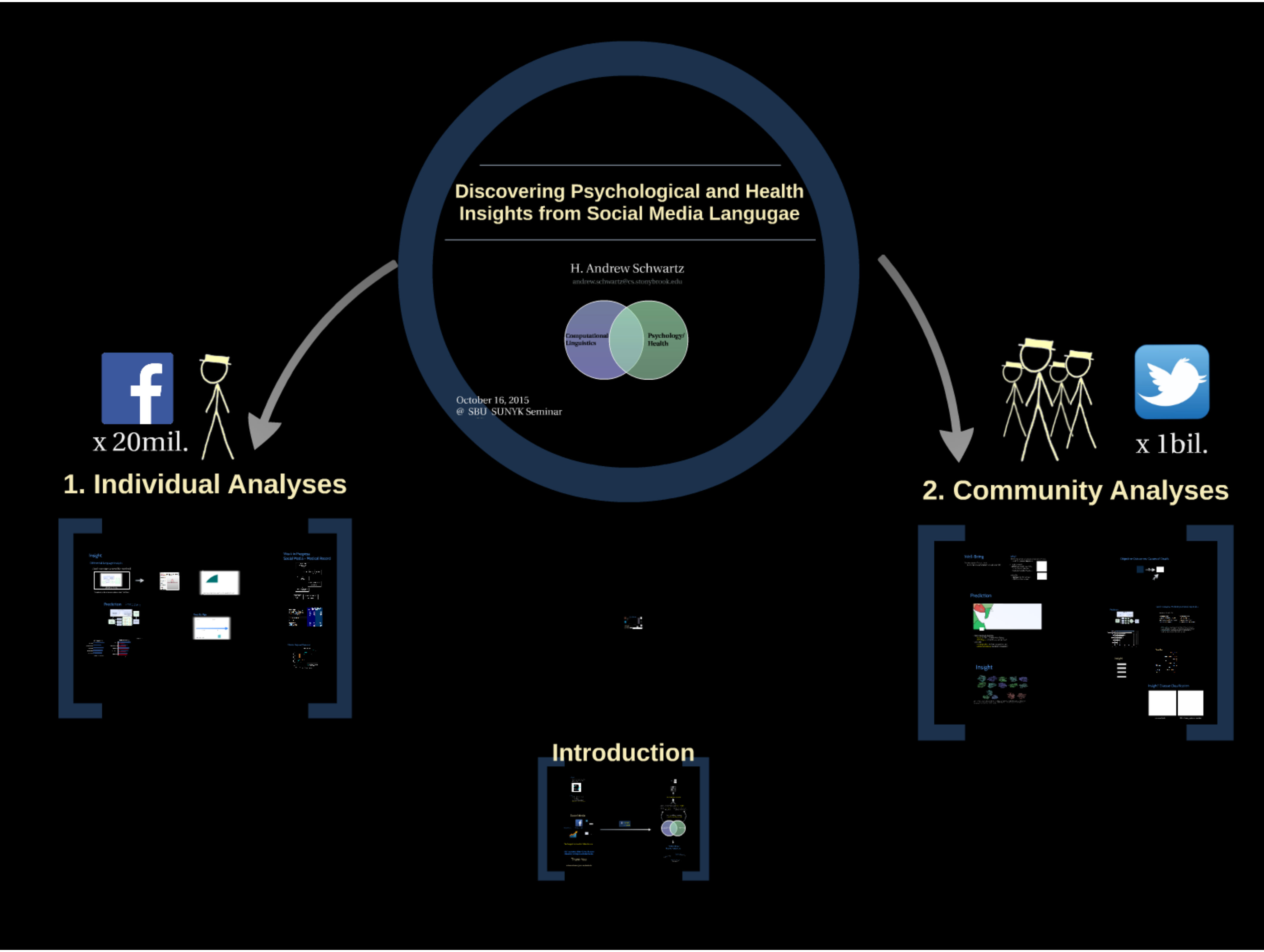
Work in Progress: Social Media + Medical Record



Terms searched	Have dx	% with dx and used a term	% without dx and used a term
abdominal pain&stomach pain&belly pain&tummy pain&stomach hurts&belly hurts&tummy hurts&tummyache&stomachache&bellyache	383	**21%**	**8%** ***
nausea&vomiting&vomit&throwing up&spitting up&threw up&puke&puked&vomited	348	**29%**	**22%** *
headache&migraine&head hurts	237	**59%**	**46%** **
leg hurts&arm hurts&finger hurts&toe hurts	194	**3%**	**1%**
uti&urinary tract infection	160	**1%**	**1%**
back pain&backache&back hurts	190	**15%**	**11%**
cough&coughing&coughed	156	**26%**	**22%**
giving birth&gave birth	188	**33%**	**10%** ***
anemia	160	**2%**	**0%** **
dizzy&dizziness&vertigo	127	**22%**	**15%**
asthma&hard to breathe	128	**28%**	**7%** ***
caught a cold&have a cold	101	**7%**	**4%**
sore throat&throat hurts	110	**24%**	**11%** ***
depression&depressed	92	**38%**	**30%**

Clinically Diagnosed Depression





Discovering Psychological and Health Insights from Social Media Langugae
H. Andrew Schwartz
andrew.schwartz@cs.stonybrook.edu
Computational Linguistics
Psychology/ Health
October 16, 2015
@ SBU SUNYK Seminar
x 20mil.
1. Individual Analyses
x 1bil.
2. Community Analyses
Introduction

Well-Being

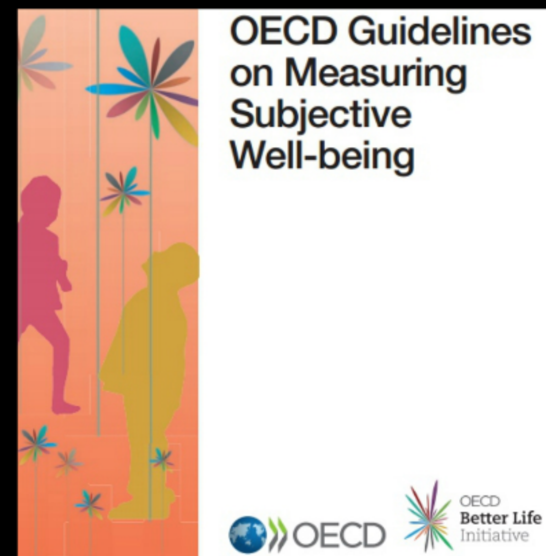
Life Satisfaction (Diener, 1987)

In general, how satisfied are you with your life?

Why?

UK Survey: *greatest happiness* or *greatest wealth*?

- =>81% - *greatest happiness*



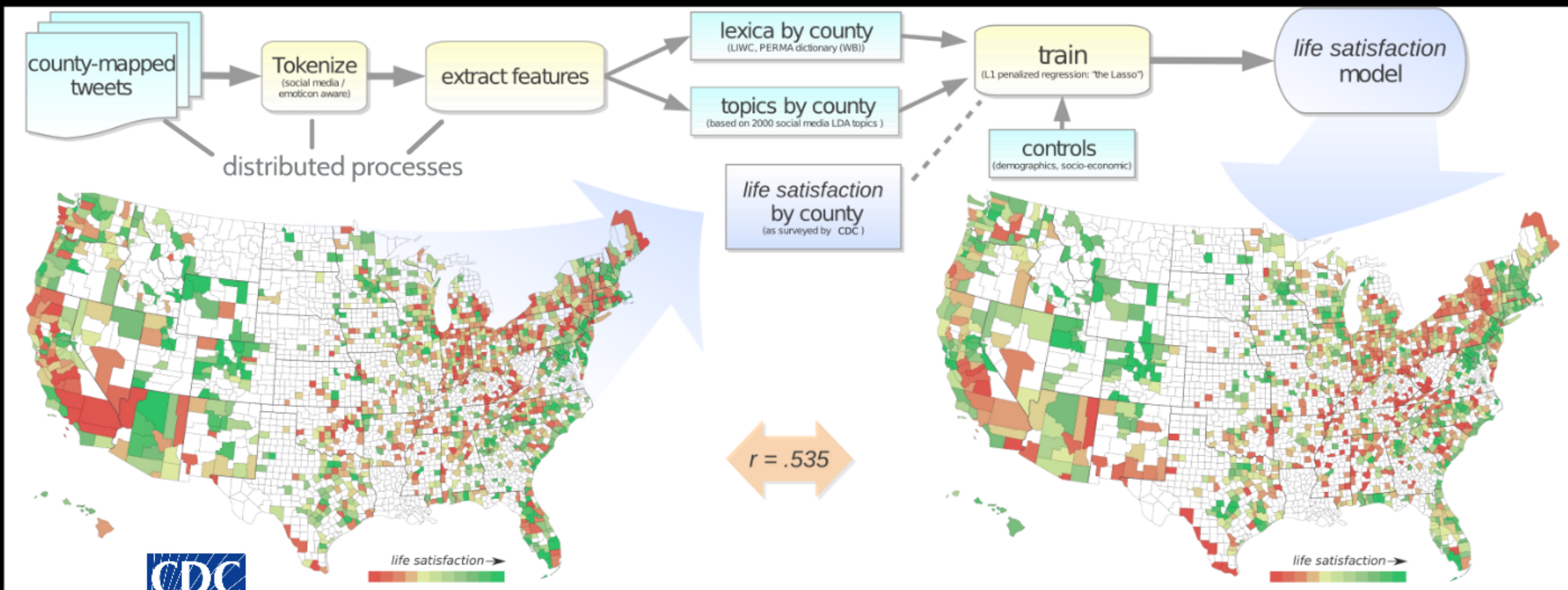
around the world...

- **OECD** set guidelines (2013) => UK, France, Bhutan, Australia, Canada, Mexico, ...

in the US...

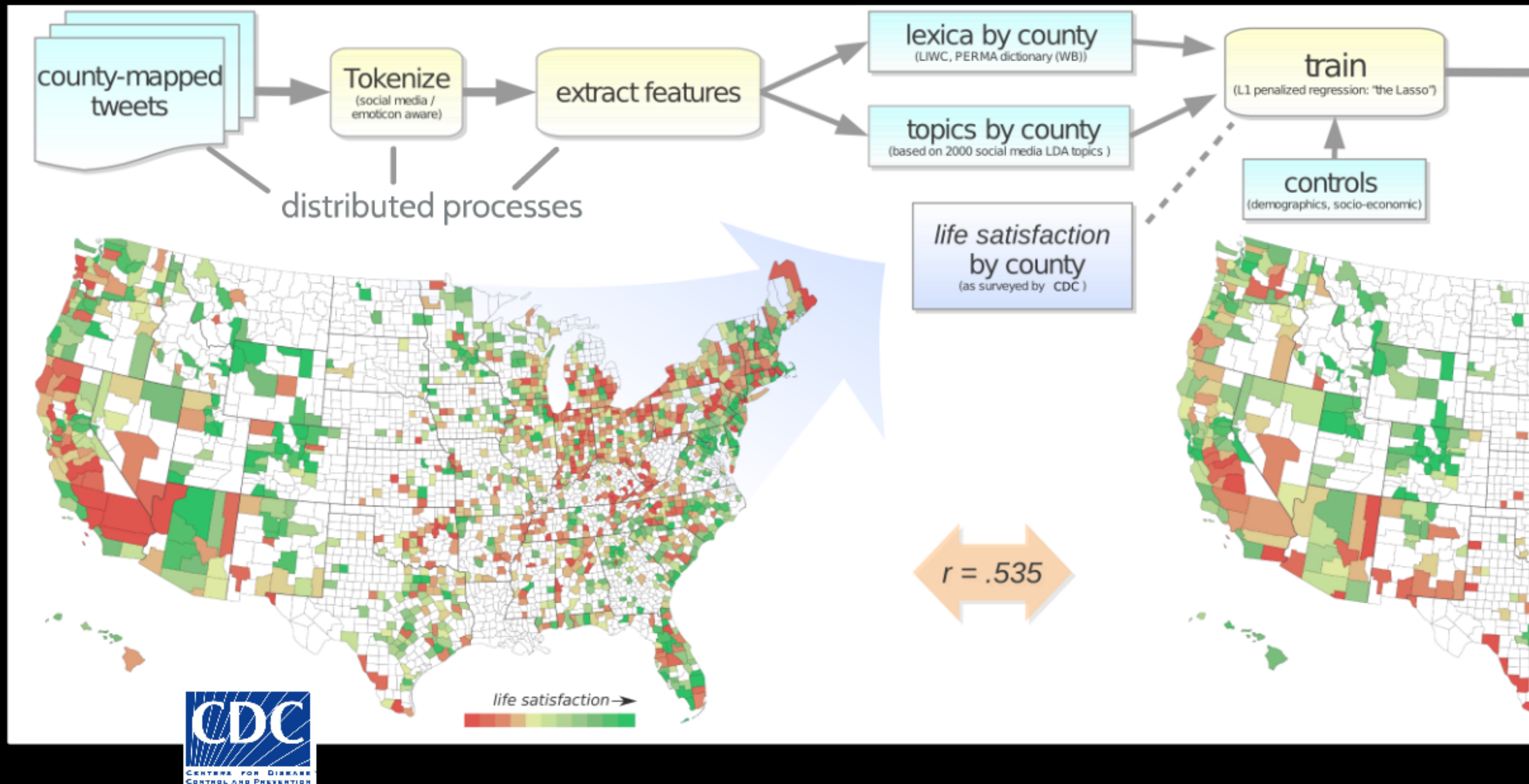
- **Gallup** Well-Being Index
- **CDC** Life Satisfaction

Prediction

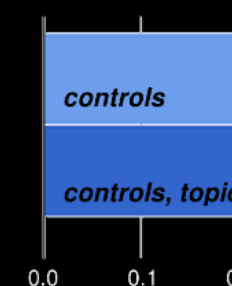


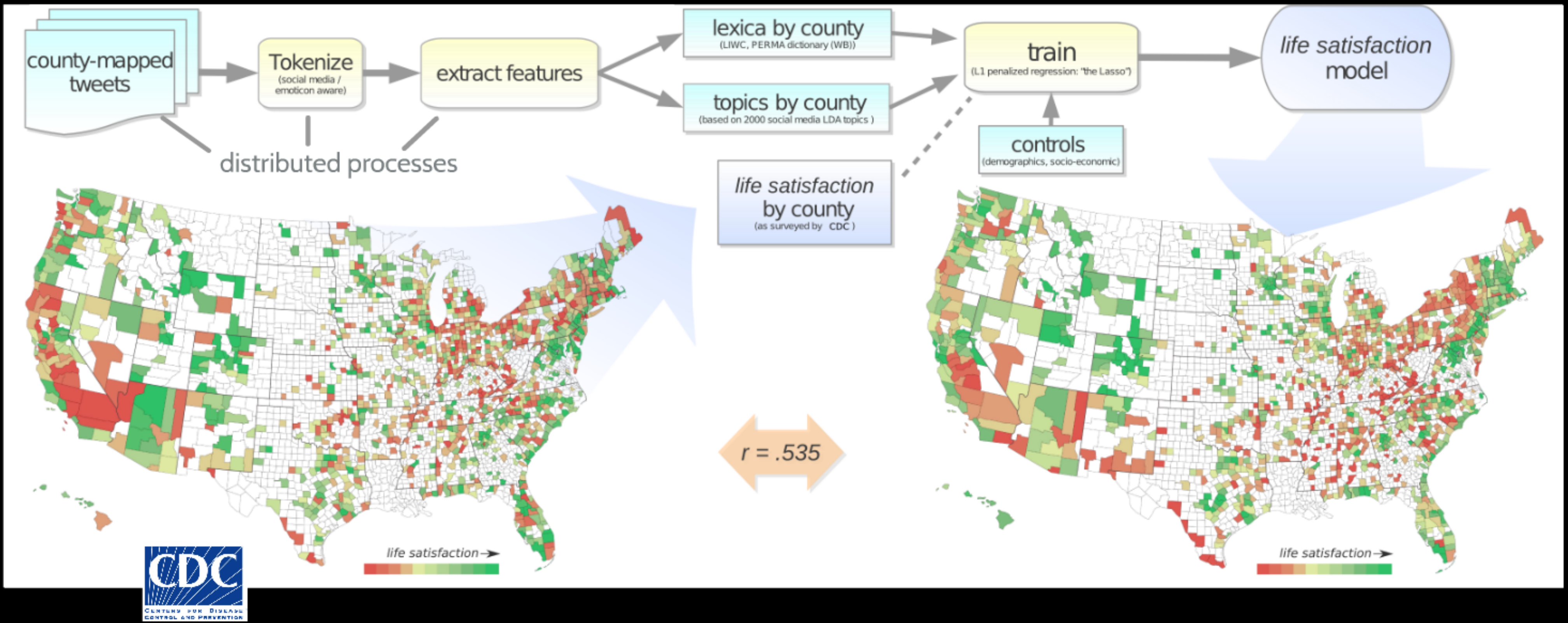
CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

- tweet language featutes:
  - Lexica: LIWC, PERMA Well-Being
  - LDA Topics: 2000 (Facebook derived)

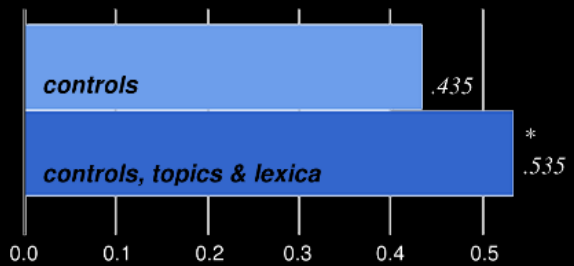


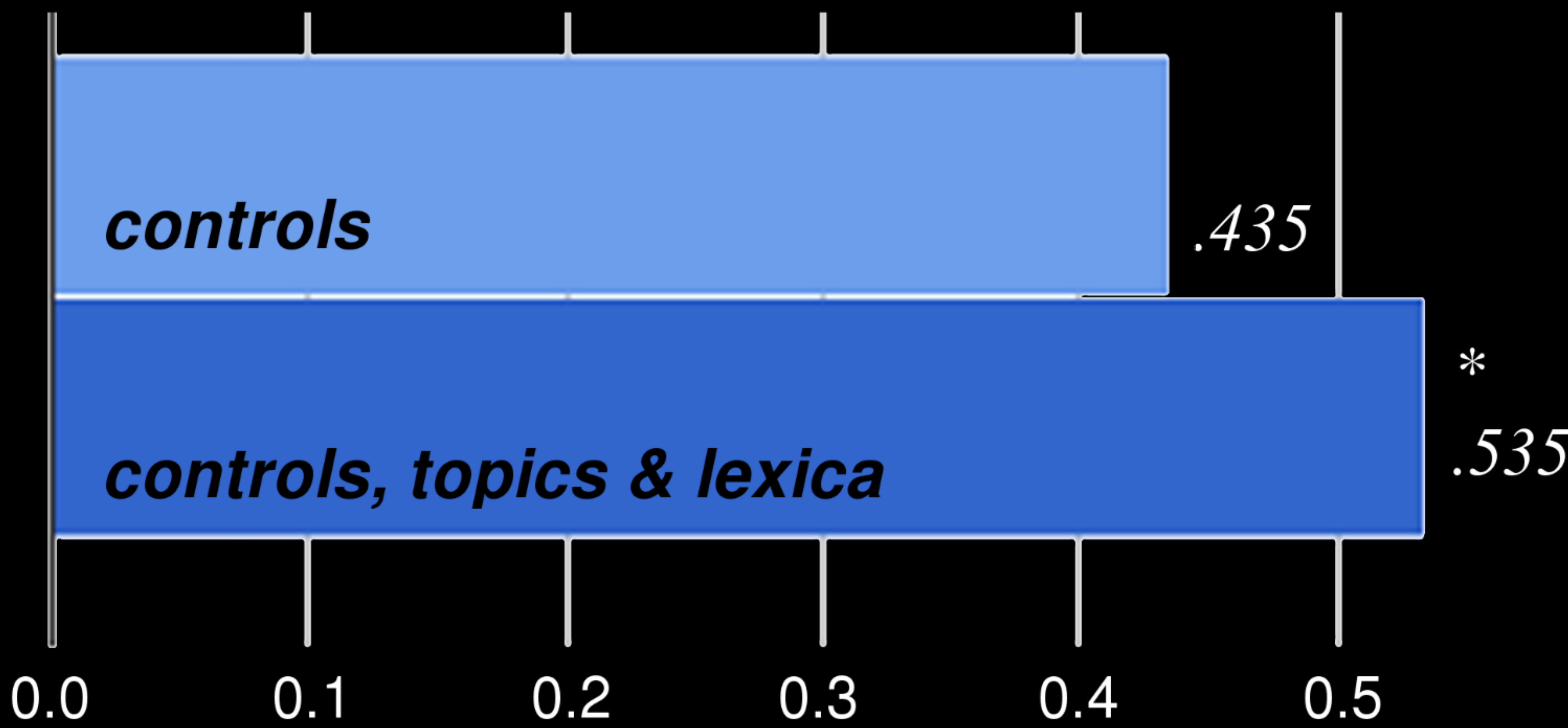
- tweet language featutes:
  - Lexica: LIWC, PERMA Well-Being
  - LDA Topics: 2000 (Facebook derived)
- controls:
  - demographics (ethnicity, gender, age)
  - socio-economics (income, education)

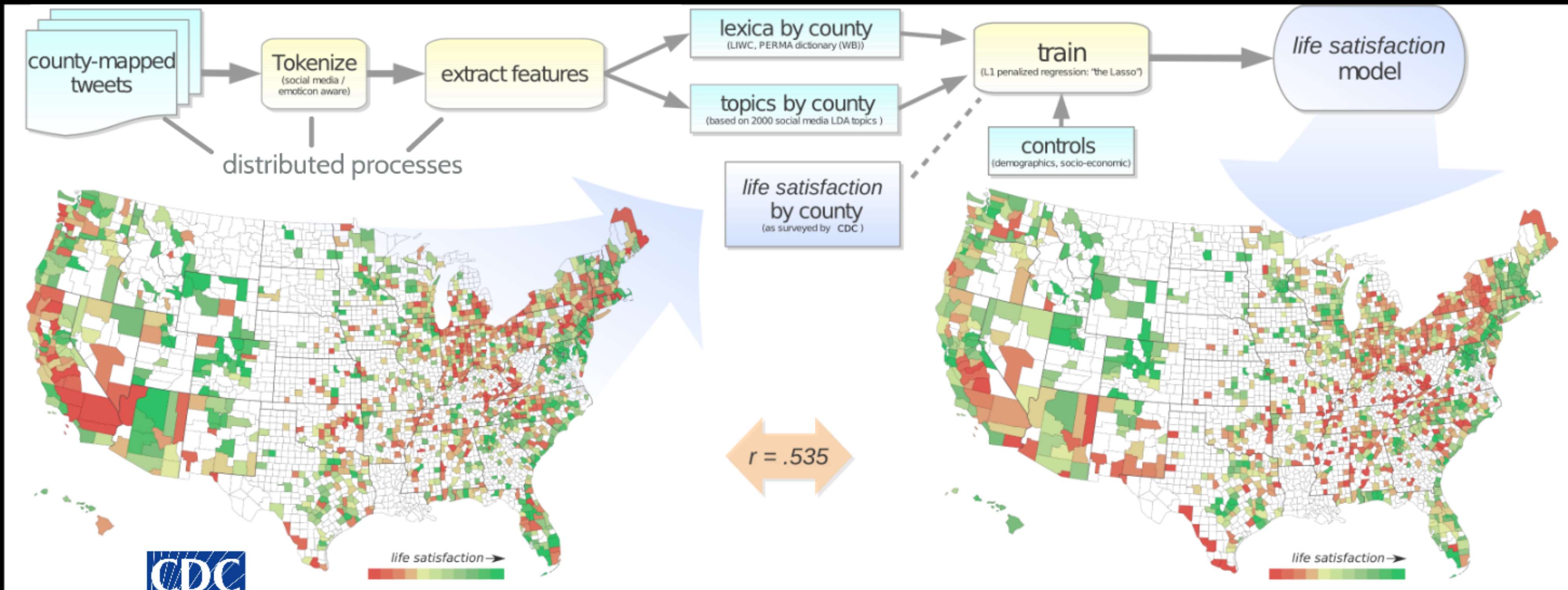




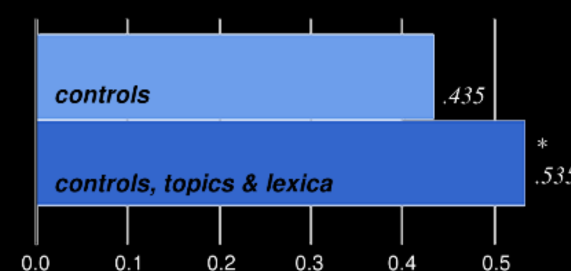
- tweet language featutes:
  - Lexica: LIWC, PERMA Well-Being
  - LDA Topics: 2000 (Facebook derived)
- controls:
  - demographics (ethnicity, gender, age)
  - socio-economics (income, education)




controls
.435
controls, topics & lexica
*
.535
0.0
0.1
0.2
0.3
0.4
0.5



- tweet language featutes:
  - Lexica: LIWC, PERMA Well-Being
  - LDA Topics: 2000 (Facebook derived)
- controls:
  - demographics (ethnicity, gender, age)
  - socio-economics (income, education)



Insight

Schwartz, 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. P., & Ungar, L. H. (2013). Characterizing Geographic Variation in Well-Being using Tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. Boston, MA.


zumba
training
basic
personal
vista
certified
intense
potty
class
session
fitness
trainer
gym
sessions
workout


ideas
figure
suggestion
idea
preferably
suggestions
helpful
thinking
appreciated
opinions
creative
greatly
advice
decide
tips


haiti
benefit
donation
raise
donated
money
donate
charity
support
cancer
donations
raised
relief
helping
fund
technology

Insight

Schwartz, 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. P., & Ungar, L. H. (2013). Characterizing Geographic Variation in Well-Being using Tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. Boston, MA.

Discovering Psychological and Health Insights from Social Media Langugae

H. Andrew Schwartz
andrew.schwartz@cs.stonybrook.edu

October 16, 2015
@ SBU SUNYK Seminar

x 20mil.

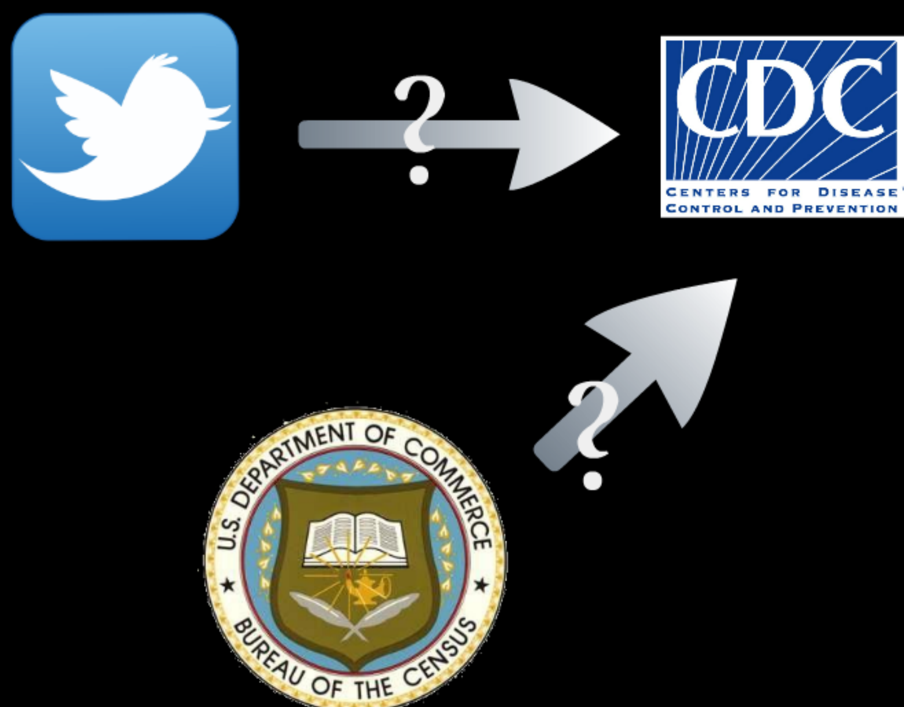
1. Individual Analyses

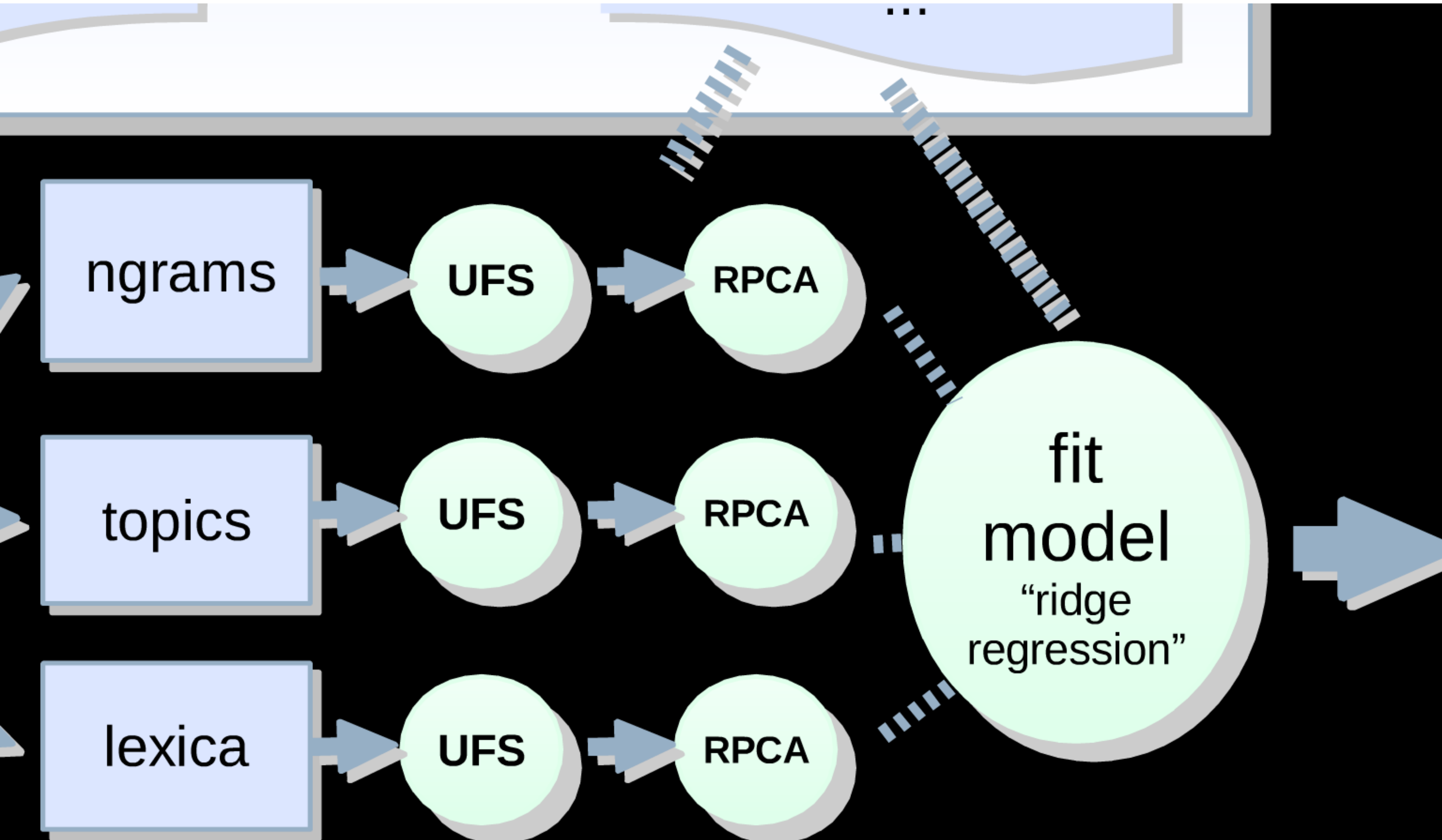
x 1bil.

2. Community Analyses

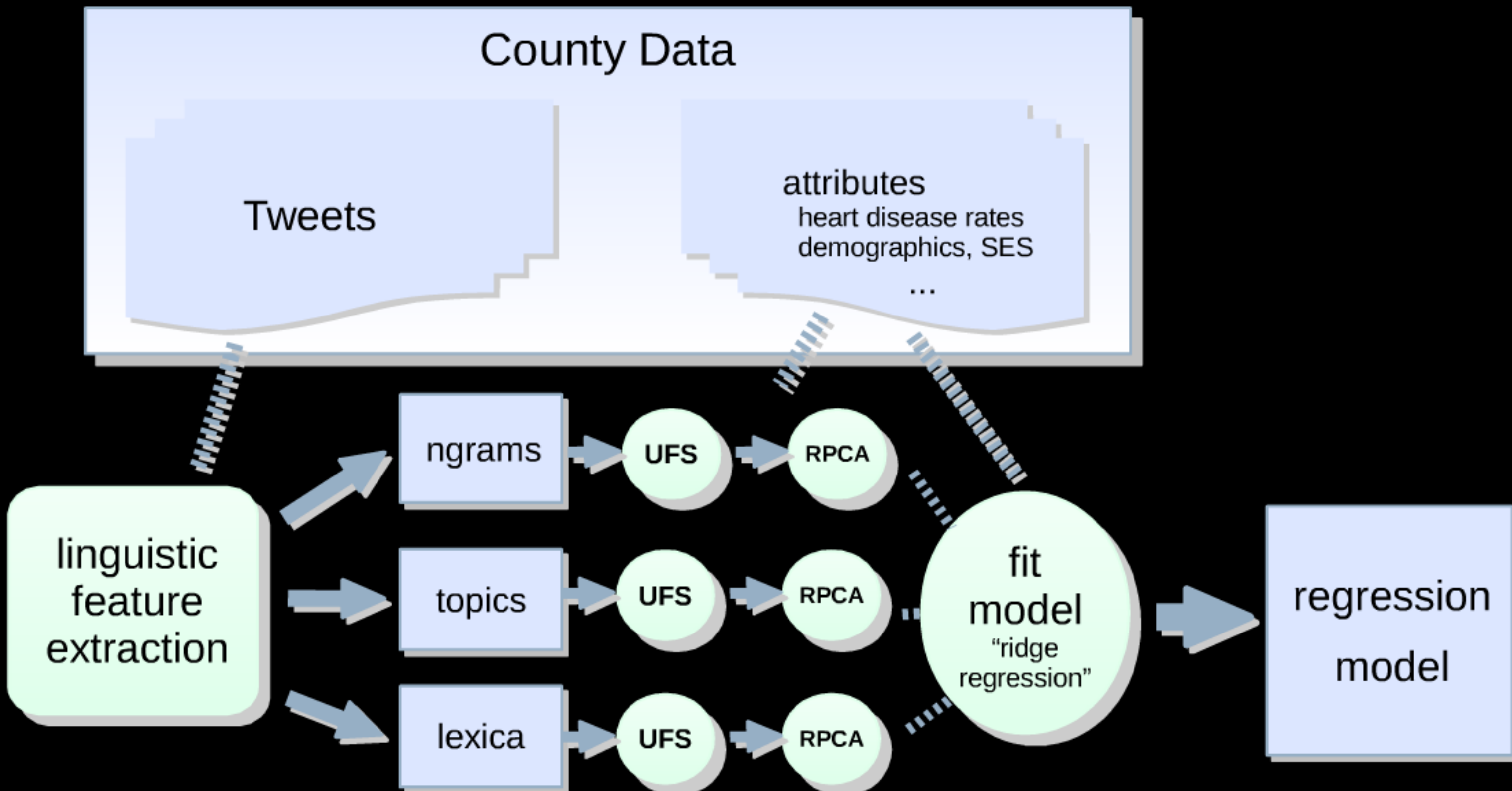
Introduction

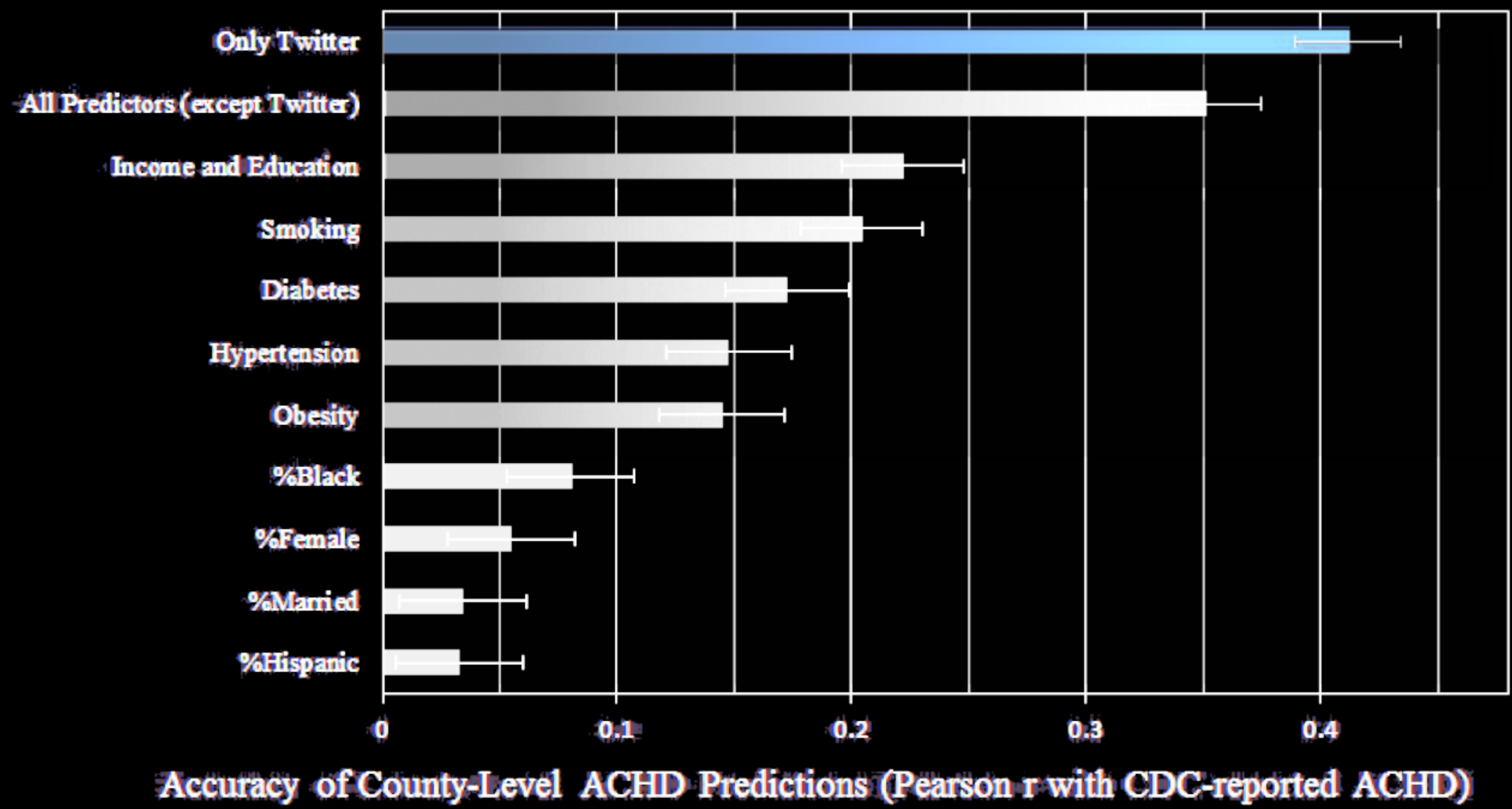
Objective Outcomes: Causes of Death




...
ngrams
UFS
RPCA
topics
UFS
RPCA
lexica
UFS
RPCA
fit
model
“ridge
regression”

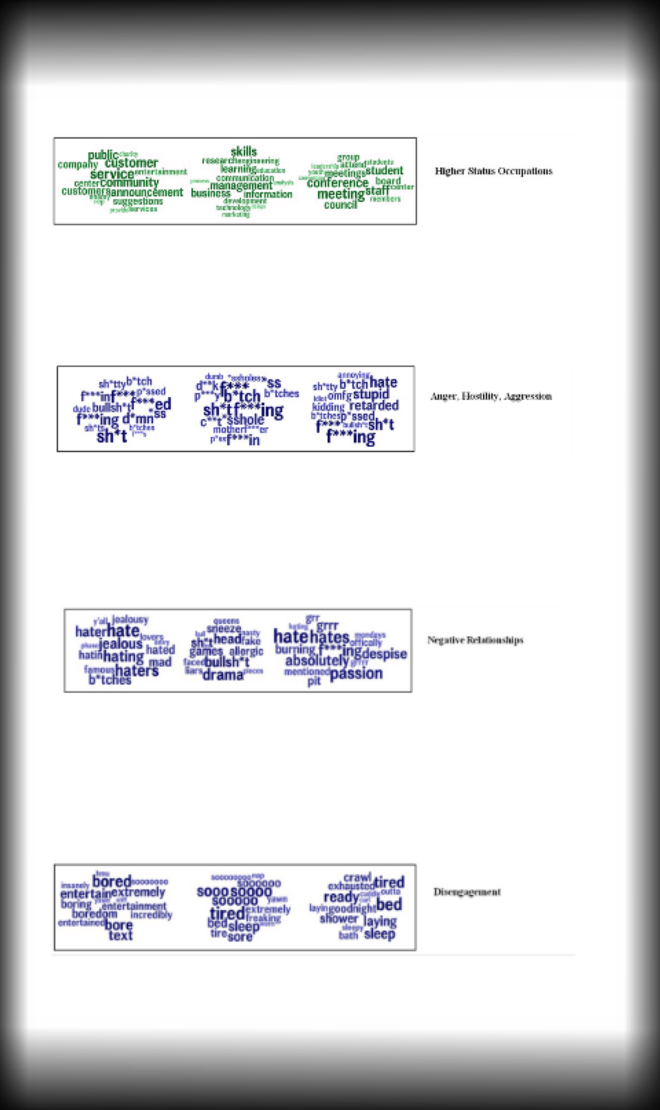
Prediction


County Data
Tweets
attributes
heart disease rates
demographics, SES
...
linguistic feature extraction
ngrams
topics
lexica
UFS
RPCA
UFS
RPCA
UFS
RPCA
fit model
"ridge regression"
regression model



Only Twitter
All Predictors (except Twitter)
Income and Education
Smoking
Diabetes
Hypertension
Obesity
%Black
%Female
%Married
%Hispanic
0
0.1
0.2
0.3
0.4
Accuracy of County-Level ACHD Predictions (Pearson r with CDC-reported ACHD)

Insight



Higher Status Occupations
Anger, Hostility, Aggression
Negative Relationships
Disengagement

Higher Status Occupations

Anger, Hostility, Aggression

Negative Relationships

Disengagement

work in progress: Predicting America's top Killers

top causes of death, 2010

1. Diseases of heart
2. Malignant neoplasms (cancers)
3. Chronic lower respiratory diseases
4. Cerebrovascular diseases (strokes)
5. Accidents, unintentional
6. Alzheimer's disease
7. Diabetes mellitus
8. Nephritis (kidney diseases)
9. Influenza and pneumonia
10. Intentional self-harm (suicide)

demographics: percentage female, black, Hispanic, foreign born, and married residents, as well as the population density
socioeconomics: percentage completed high school / a bachelors, unemployed, log median income

Results

Heart disease

Cancers

lower respiratory diseases

Strokes

Accidents, unintentional

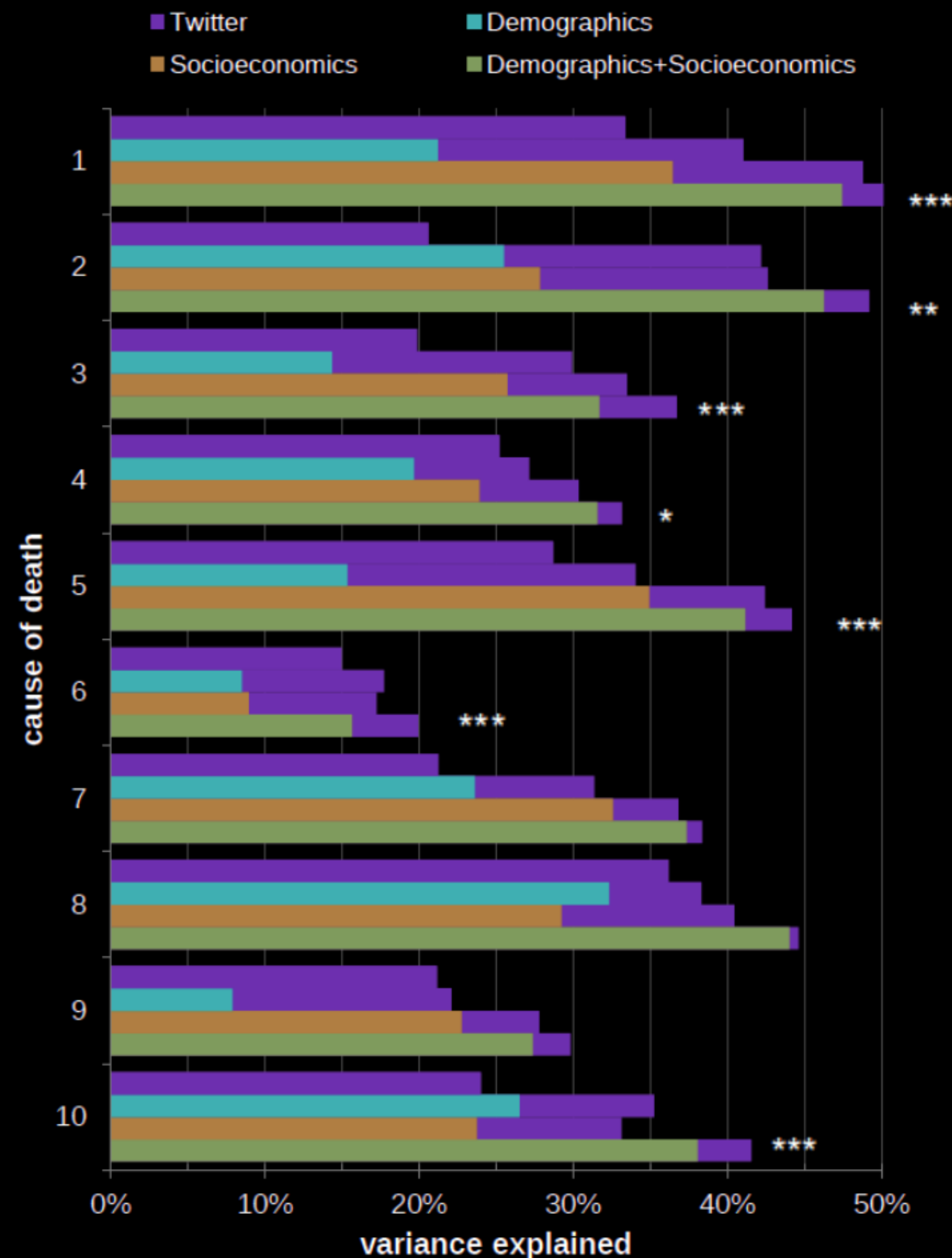
Alzheimer's disease

Diabetes

Kidney diseases

Influenza and pneumonia

Suicide



significant reduction in error over demographics+ socioeconomics model

* p < .05
** p < .01
*** p < .001

Insight? Disease Classification

topic	01 heart diseases	02 cancers	03 resp. diseases	04 strokes	05 accidents	06 diabetes	07 Alzheimer's	08 flu	09 kidney disease	10 suicides
1213	0.41	0.29	0.16	0.38	0.20	0.33	0.15	0.28	0.48	-0.08
678	0.39	0.26	0.15	0.35	0.24	0.39	0.16	0.29	0.48	0.02
355	0.39	0.35	0.19	0.33	0.16	0.30	0.15	0.29	0.41	0.00
1405	0.38	0.31	0.23	0.33	0.18	0.24	0.19	0.27	0.37	0.07
913	0.25	0.19	0.23	0.24	0.25	0.27	0.12	0.29	0.34	0.05
523	0.22	0.16	0.22	0.22	0.25	0.28	0.11	0.28	0.32	0.06
599	0.09	0.10	0.09	0.05	0.08	0.06	0.05	0.06	0.10	0.19
1193	-0.03	-0.06	-0.14	-0.13	-0.25	-0.19	-0.19	-0.19	-0.16	-0.29
208	-0.13	-0.13	-0.19	-0.26	-0.25	-0.22	-0.24	-0.13	-0.21	-0.18
923	-0.28	-0.19	-0.20	-0.25	-0.25	-0.37	-0.15	-0.28	-0.42	-0.06
565	-0.28	-0.23	-0.21	-0.22	-0.22	-0.33	-0.13	-0.23	-0.42	0.01
85	-0.30	-0.24	-0.27	-0.26	-0.30	-0.35	-0.17	-0.33	-0.38	-0.09
1037	-0.30	-0.29	-0.23	-0.26	-0.21	-0.30	-0.08	-0.27	-0.36	-0.06
533	-0.32	-0.26	-0.28	-0.30	-0.30	-0.34	-0.19	-0.31	-0.38	-0.13
1772	-0.34	-0.24	-0.25	-0.32	-0.28	-0.32	-0.19	-0.29	-0.36	-0.07
766	-0.34	-0.23	-0.27	-0.31	-0.31	-0.33	-0.15	-0.30	-0.35	-0.10

no controls

topic	01 HD	02 cancers	03 resp.	04 strokes	05 accidents	06 diabetes	07 Alzh	08 flu	09 kidney	10 suicides
137	0.12	-0.01	0.08	0.01	0.01	-0.07	0.00	0.02	-0.01	0.00
1405	0.12	0.08	0.20	0.13	0.08	-0.03	0.14	0.08	0.00	-0.04
1213	0.10	0.01	0.11	0.13	0.06	0.01	0.11	0.18	0.16	-0.01
1571	0.08	0.11	0.06	0.02	-0.02	0.04	-0.03	0.06	0.03	-0.08
678	0.08	0.00	0.06	0.09	0.08	0.07	0.10	0.20	0.16	0.09
351	0.04	-0.01	0.00	0.05	0.03	0.13	-0.02	0.08	0.04	0.07
695	0.04	-0.01	-0.06	-0.06	-0.12	-0.10	-0.08	-0.10	-0.06	-0.16
600	0.03	-0.01	0.03	0.04	0.03	0.11	0.08	0.13	0.11	0.13
1387	0.03	0.00	0.01	0.02	-0.09	-0.07	0.04	-0.05	-0.04	-0.20
208	0.02	-0.03	-0.07	-0.12	-0.10	-0.07	-0.15	-0.04	-0.04	-0.05
1973	-0.01	0.00	-0.11	-0.08	-0.08	-0.02	-0.17	-0.03	0.02	-0.01
684	-0.02	-0.02	-0.03	-0.04	-0.09	-0.02	0.02	-0.06	-0.05	-0.02
916	-0.03	-0.13	-0.06	-0.01	-0.11	-0.10	0.03	-0.04	-0.05	-0.09
1906	-0.05	-0.02	-0.07	-0.05	-0.05	-0.13	-0.05	-0.13	-0.11	0.01
874	-0.06	-0.07	-0.08	-0.05	0.01	-0.12	-0.02	-0.16	-0.11	0.05
923	-0.07	0.00	-0.03	0.00	-0.04	-0.10	0.02	-0.12	-0.14	0.00
1415	-0.13	-0.04	-0.14	-0.06	-0.07	-0.03	-0.05	-0.08	-0.01	-0.08

SES + Demographic controlled

10 suicides
-0.08
0.02
0.00
0.07
0.05
0.06
0.19
-0.29
-0.18
-0.06
0.01
-0.09
-0.06
-0.13
-0.07
-0.10

topic	01 HD	02 cancers	03 resp.	04 strokes	05 accidents	06 diabetes	07 Alzh	08 flu	09 kidney	10 suicides
137	0.12	-0.01	0.08	0.01	0.01	-0.07	0.00	0.02	-0.01	0.00
1405	0.12	0.08	0.20	0.13	0.08	-0.03	0.14	0.08	0.00	-0.04
1213	0.10	0.01	0.11	0.13	0.06	0.01	0.11	0.18	0.16	-0.01
1571	0.08	0.11	0.06	0.02	-0.02	0.04	-0.03	0.06	0.03	-0.08
678	0.08	0.00	0.06	0.09	0.08	0.07	0.10	0.20	0.16	0.09
351	0.04	-0.01	0.00	0.05	0.03	0.13	-0.02	0.08	0.04	0.07
695	0.04	-0.01	-0.06	-0.06	-0.12	-0.10	-0.08	-0.10	-0.06	-0.16
600	0.03	-0.01	0.03	0.04	0.03	0.11	0.08	0.13	0.11	0.13
1387	0.03	0.00	0.01	0.02	-0.09	-0.07	0.04	-0.05	-0.04	-0.20
208	0.02	-0.03	-0.07	-0.12	-0.10	-0.07	-0.15	-0.04	-0.04	-0.05
1973	-0.01	0.00	-0.11	-0.08	-0.08	-0.02	-0.17	-0.03	0.02	-0.01
684	-0.02	-0.02	-0.03	-0.04	-0.09	-0.02	0.02	-0.06	-0.05	-0.02
916	-0.03	-0.13	-0.06	-0.01	-0.11	-0.10	0.03	-0.04	-0.05	-0.09
1906	-0.05	-0.02	-0.07	-0.05	-0.05	-0.13	-0.05	-0.13	-0.11	0.01
874	-0.06	-0.07	-0.08	-0.05	0.01	-0.12	-0.02	-0.16	-0.11	0.05
923	-0.07	0.00	-0.03	0.00	-0.04	-0.10	0.02	-0.12	-0.14	0.00
1415	-0.13	-0.04	-0.14	-0.06	-0.07	-0.03	-0.05	-0.08	-0.01	-0.08

topic	01 heart diseases	02 cancers	03 resp. diseases	04 strokes	05 accidents	06 diabetes	07 Alzheimer's	08 flu	09 kidney disease	10 suicides
1213	0.41	0.29	0.16	0.38	0.20	0.33	0.15	0.28	0.48	-0.08
678	0.39	0.26	0.15	0.35	0.24	0.39	0.16	0.29	0.48	0.02
355	0.39	0.35	0.19	0.33	0.16	0.30	0.15	0.29	0.41	0.00
1405	0.38	0.31	0.23	0.33	0.18	0.24	0.19	0.27	0.37	0.07
913	0.25	0.19	0.23	0.24	0.25	0.27	0.12	0.29	0.34	0.05
523	0.22	0.16	0.22	0.22	0.25	0.28	0.11	0.28	0.32	0.06
599	0.09	0.10	0.09	0.05	0.08	0.06	0.05	0.06	0.10	0.19
1193	-0.03	-0.06	-0.14	-0.13	-0.25	-0.19	-0.19	-0.19	-0.16	-0.29
208	-0.13	-0.13	-0.19	-0.26	-0.25	-0.22	-0.24	-0.13	-0.21	-0.18
923	-0.28	-0.19	-0.20	-0.25	-0.25	-0.37	-0.15	-0.28	-0.42	-0.06
565	-0.28	-0.23	-0.21	-0.22	-0.22	-0.33	-0.13	-0.23	-0.42	0.01
85	-0.30	-0.24	-0.27	-0.26	-0.30	-0.35	-0.17	-0.33	-0.38	-0.09
1037	-0.30	-0.29	-0.23	-0.26	-0.21	-0.30	-0.08	-0.27	-0.36	-0.06
533	-0.32	-0.26	-0.28	-0.30	-0.30	-0.34	-0.19	-0.31	-0.38	-0.13
1772	-0.34	-0.24	-0.25	-0.32	-0.28	-0.32	-0.19	-0.29	-0.36	-0.07
766	-0.34	-0.23	-0.27	-0.31	-0.31	-0.33	-0.15	-0.30	-0.35	-0.10

topic

Discovering Psychological and Health Insights from Social Media Langugae

H. Andrew Schwartz
andrew.schwartz@cs.stonybrook.edu

October 16, 2015
@ SBU SUNYK Seminar

x 20mil.

1. Individual Analyses

x 1bil.

2. Community Analyses

Introduction

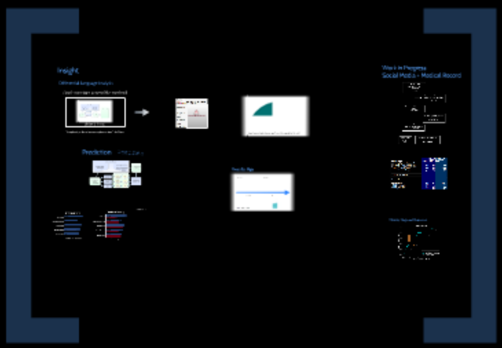
Discovering Psychological and Health Insights from Social Media Langugae

H. Andrew Schwartz
andrew.schwartz@cs.stonybrook.edu

October 16, 2015
@ SBU SUNYK Seminar

x 20mil.

1. Individual Analyses



x 1bil.

2. Community Analyses

Introduction

Social Media

350m tweets/day

4b messages/day

…

...the largest dataset of who we are

Interdisciplinary Research Questions

What psychological factors emerge in language as drivers of health and well-being?

Can we predict disease risk and recovery from language use?

To what extent can language analyses replace and extend traditional psychological asessement

...the largest dataset of who we are

NLP can foster data-driven human discovery at unprecedented scale.

Thank You

andrew.schwartz@cs.stonybrook.edu