Machine Learning on Social Media Textual Data for Predicting Psychological and Health outcomes

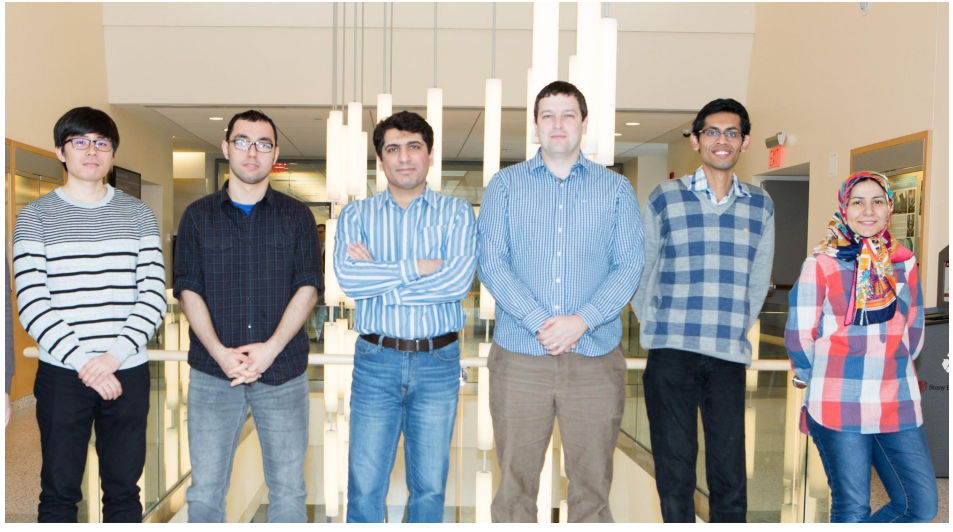
H. Andrew Schwartz

Stony Brook University

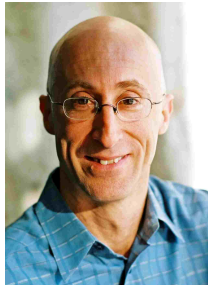
Human Language Analysis Lab

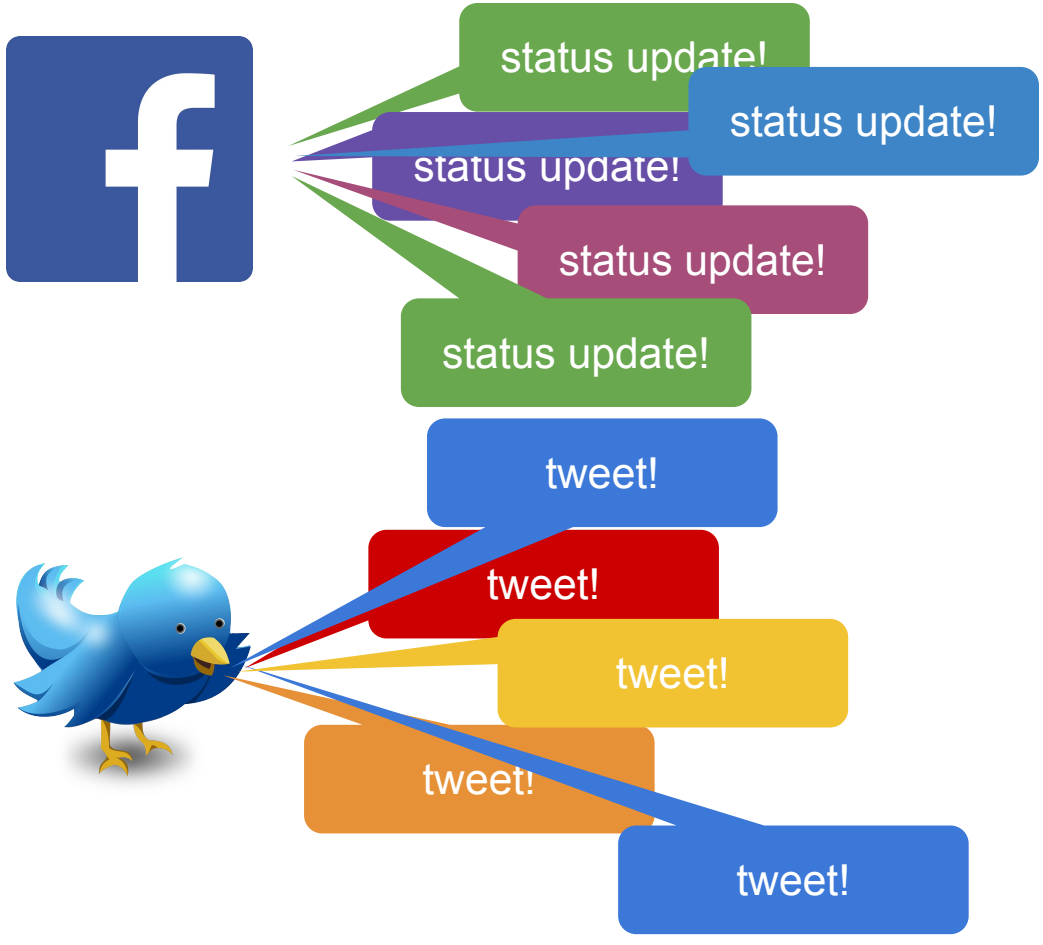
Stony Brook University

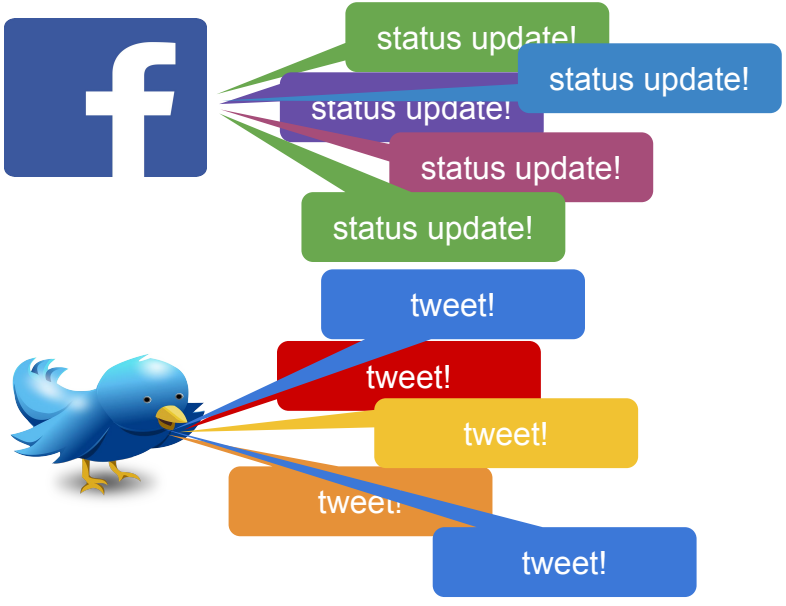
Human Language Analysis Lab

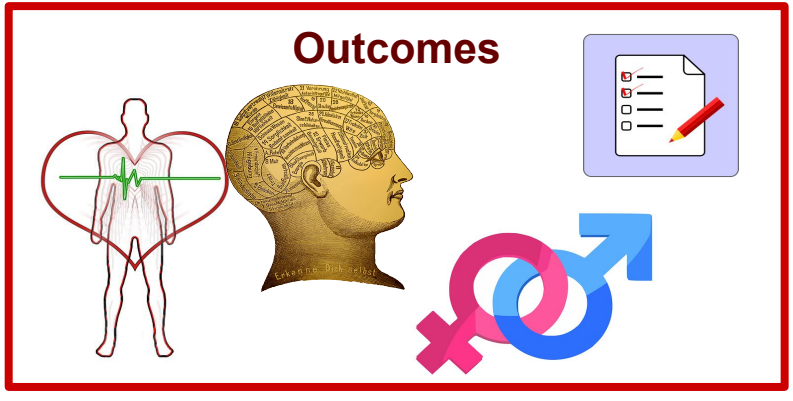


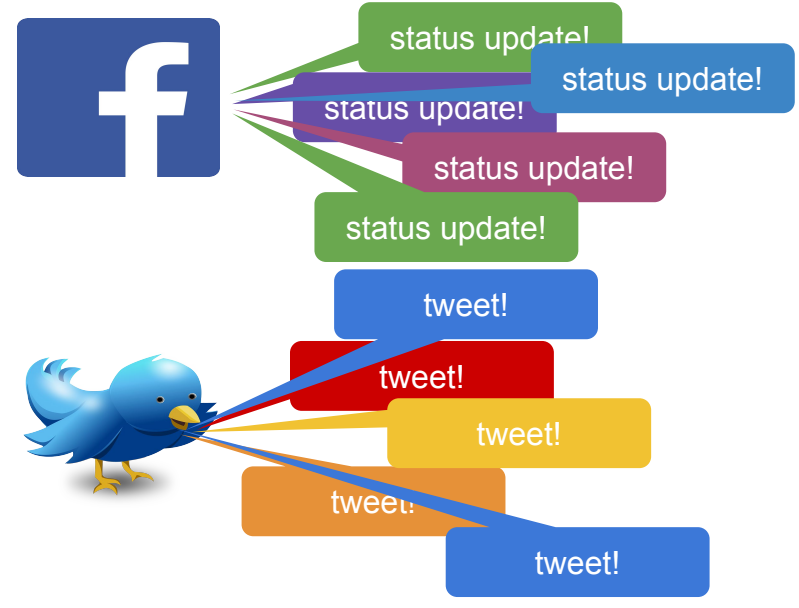
Penn | World Well-Being Project

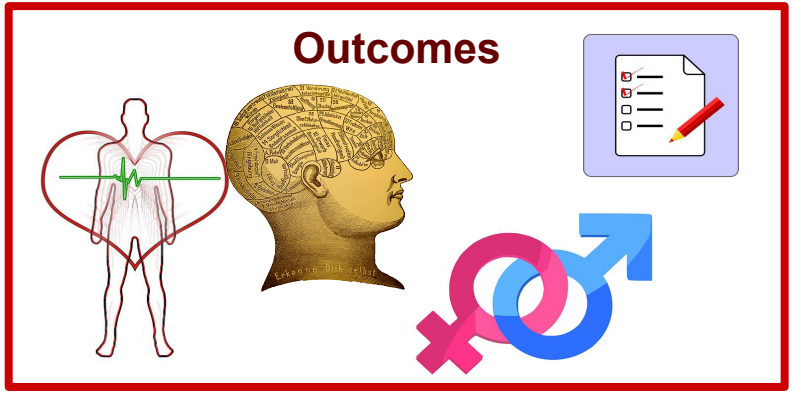



status update!
status update!
status update!
status update!
status update!
tweet!
tweet!
tweet!
tweet!
tweet!

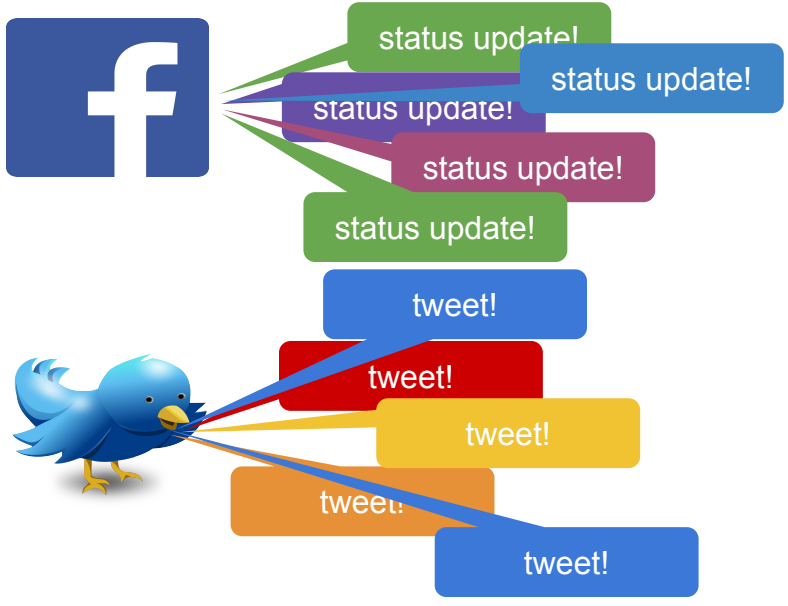

status update!
status update!
status update!
status update!
status update!
tweet!
tweet!
tweet!
tweet!
tweet!

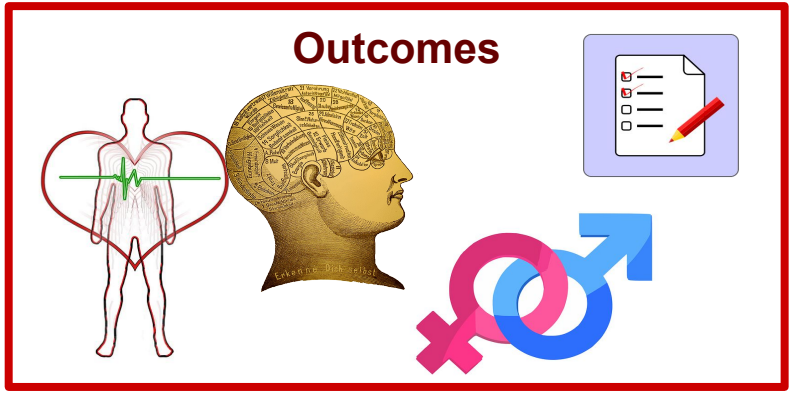

Outcomes





language analysis





language analysis

- **prediction (measurement)**
- **insights**

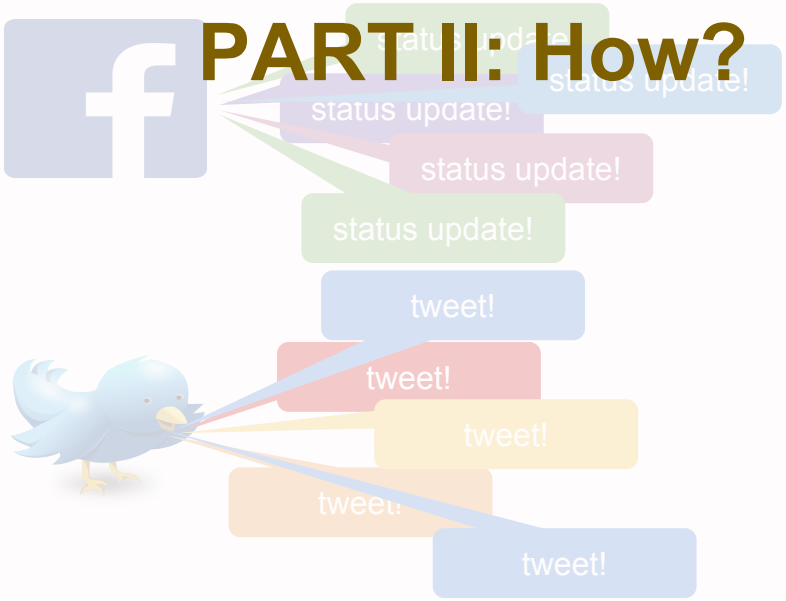
PART I: What are the possibilities?

language analysis

- **prediction (measurement)**
- **insights**

PART I: What?

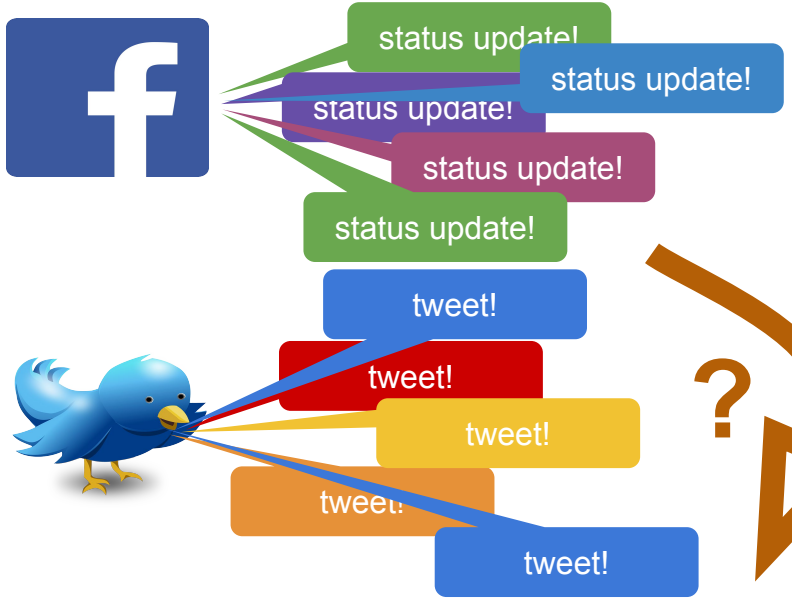
PART II: How?



language analysis

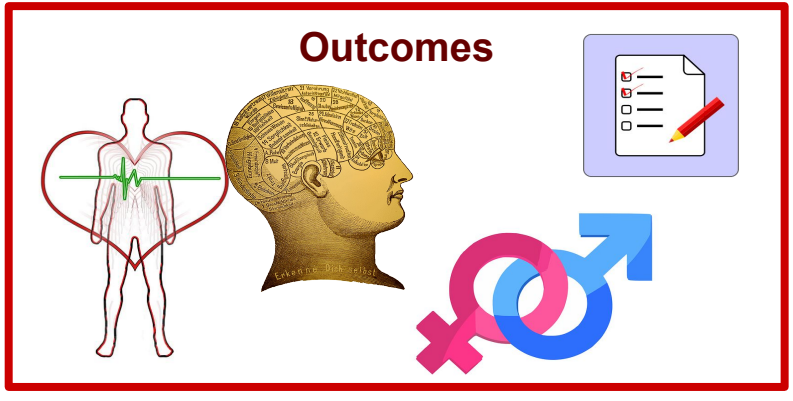
- prediction (measurement)
- insights

PART I: What?



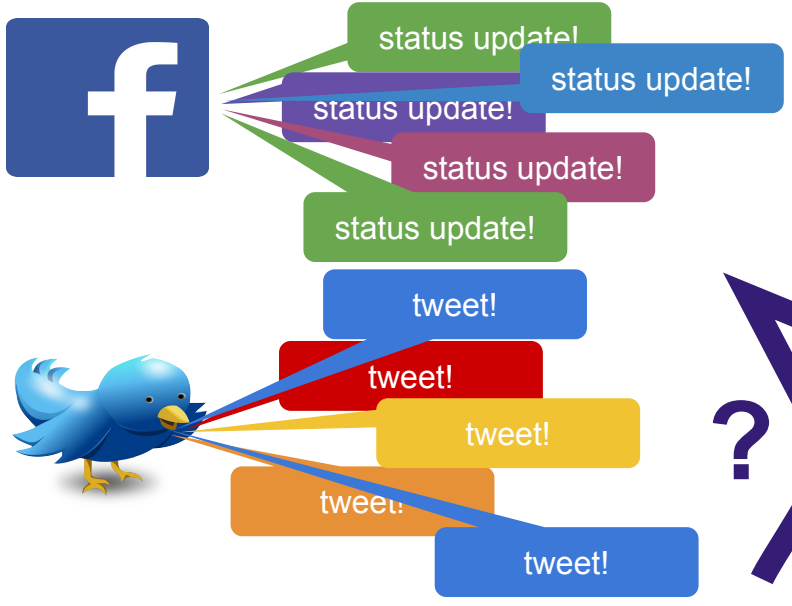
?

language analysis



- prediction (measurement)
- insights

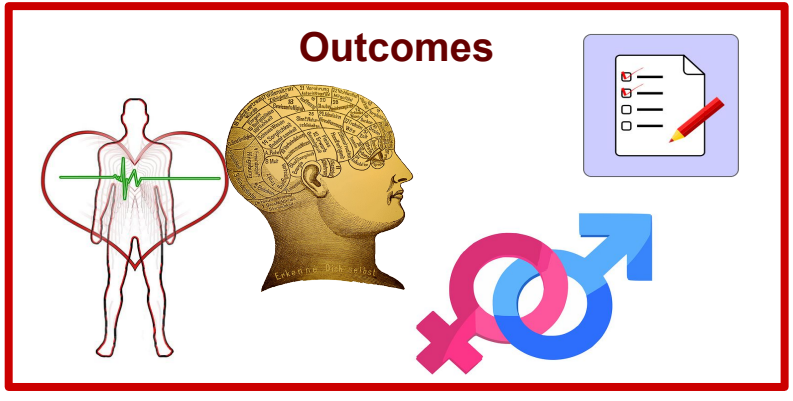
PART I: What?



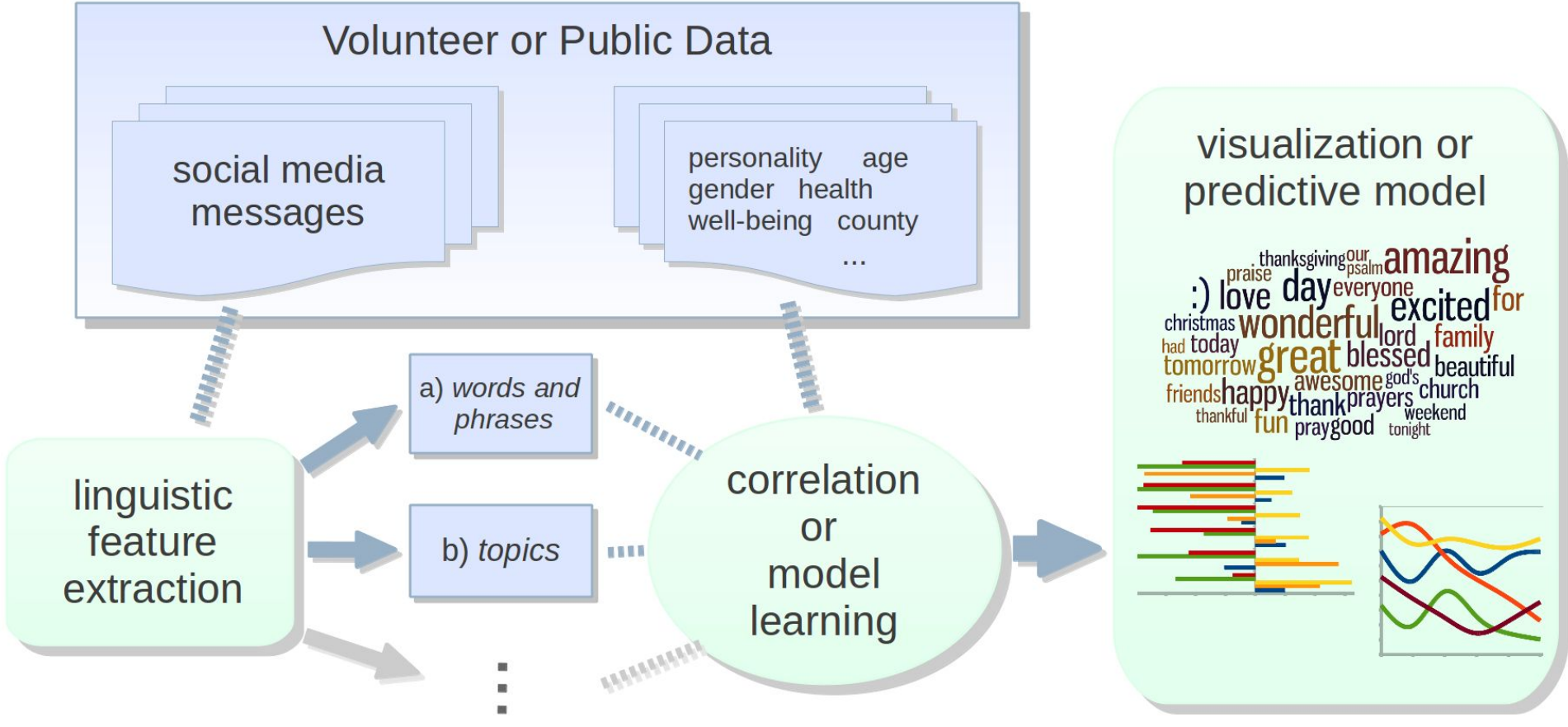
?

language analysis

- prediction (measurement)
- insights



General and Flexible Framework



extraversion -- *sociable, assertive, active, energetic, talkative, outgoing*

Penn | World Well-Being Project | wwbp.org

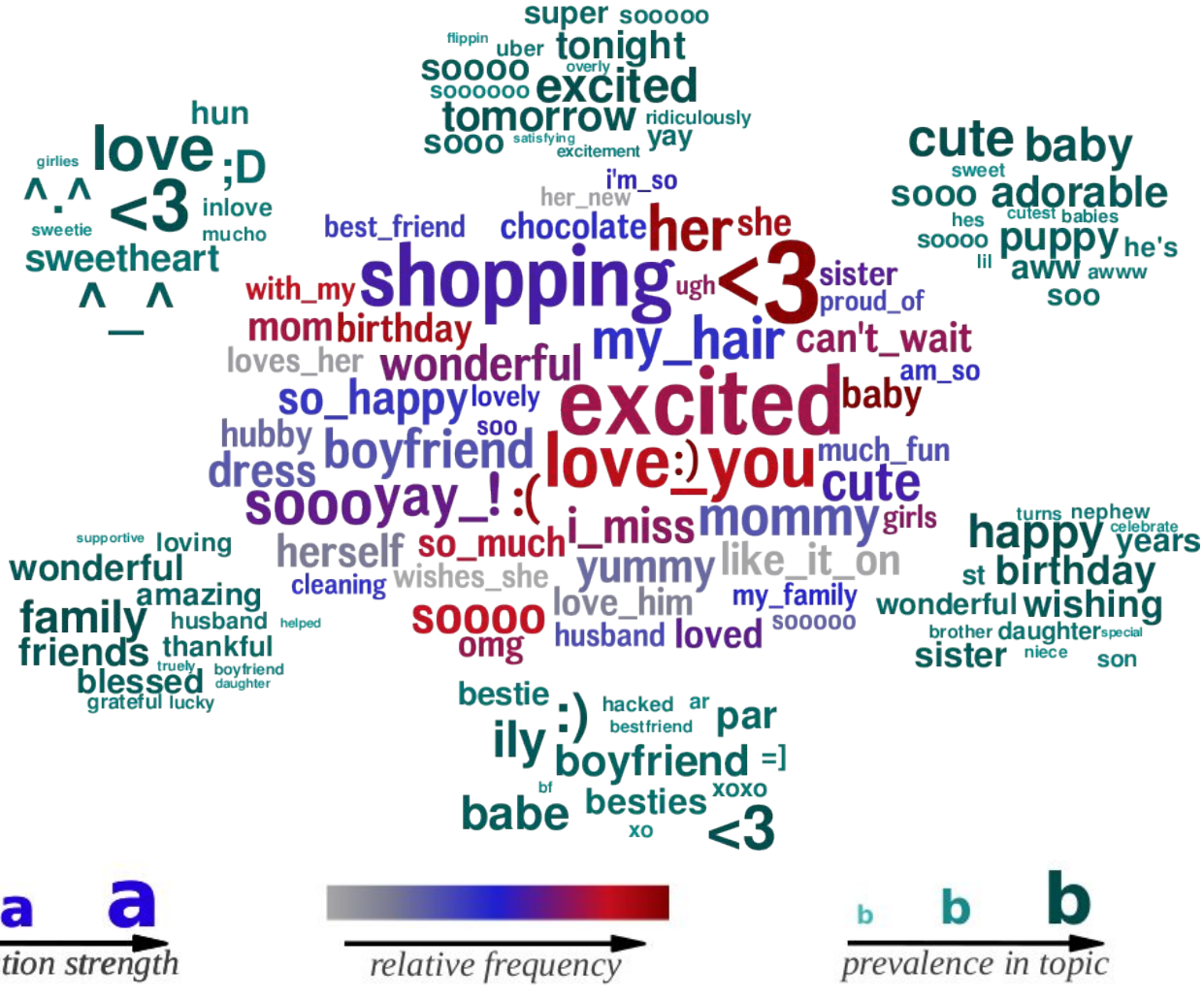
Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). **Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach**. *In PLOS*

introversion

Penn | World Well-Being Project | wwbp.org

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)*.

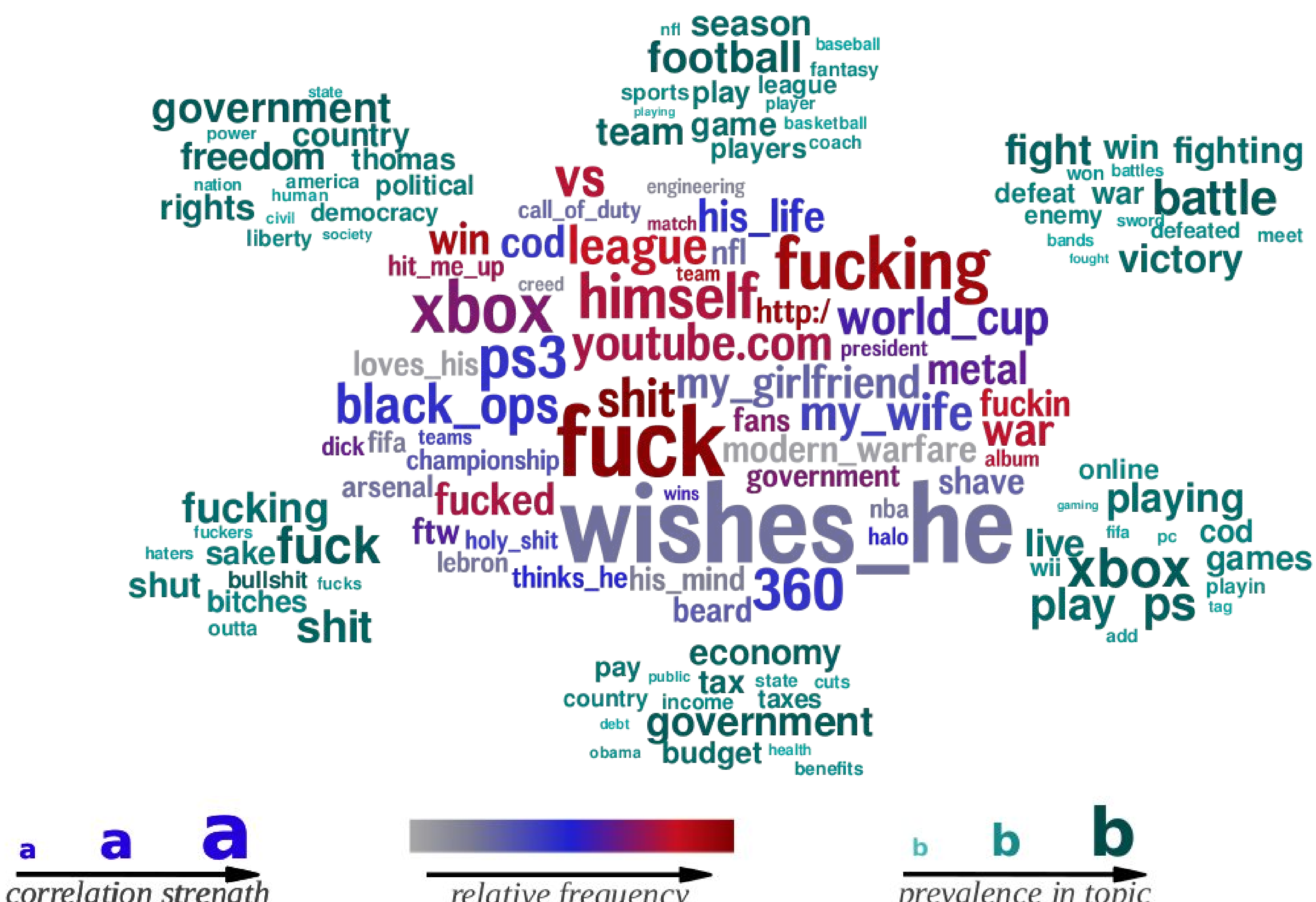
Gender


super
sooooo
flippin
uber
tonight
soooo
overly
excited
soooooo
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
with_my
ugh
<3
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
cute
soooyay_!:(
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
truely
boyfriend
blessed
daughter
grateful
lucky
bestie
:)
hacked
ar
bestfriend
par
ily
boyfriend
=]
bf
xoxo
babe
besties
xo
<3
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
a
a
a
correlation strength
relative frequency
b
b
b
prevalence in topic

Gender

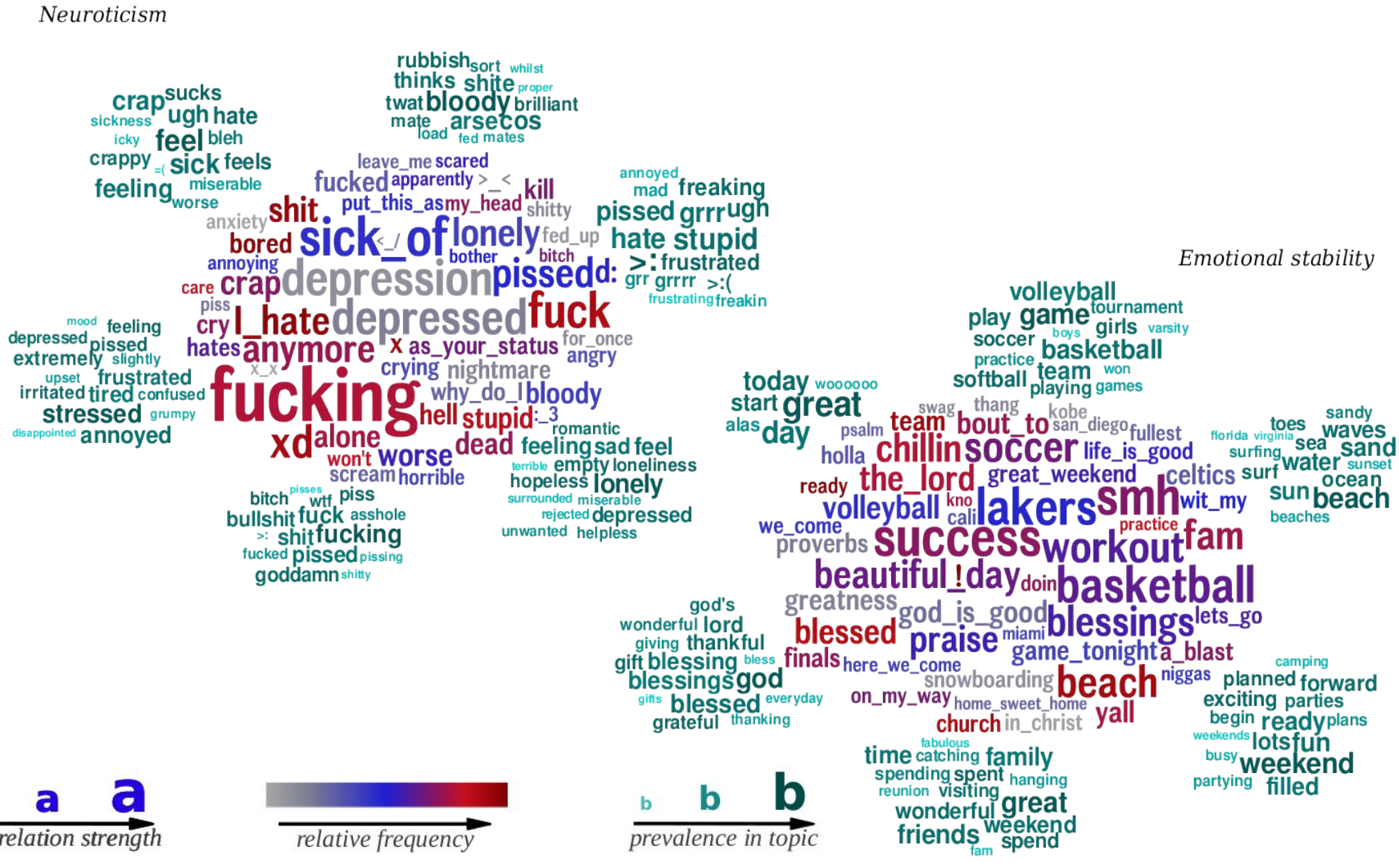
Explicit Language Warning...

Gender

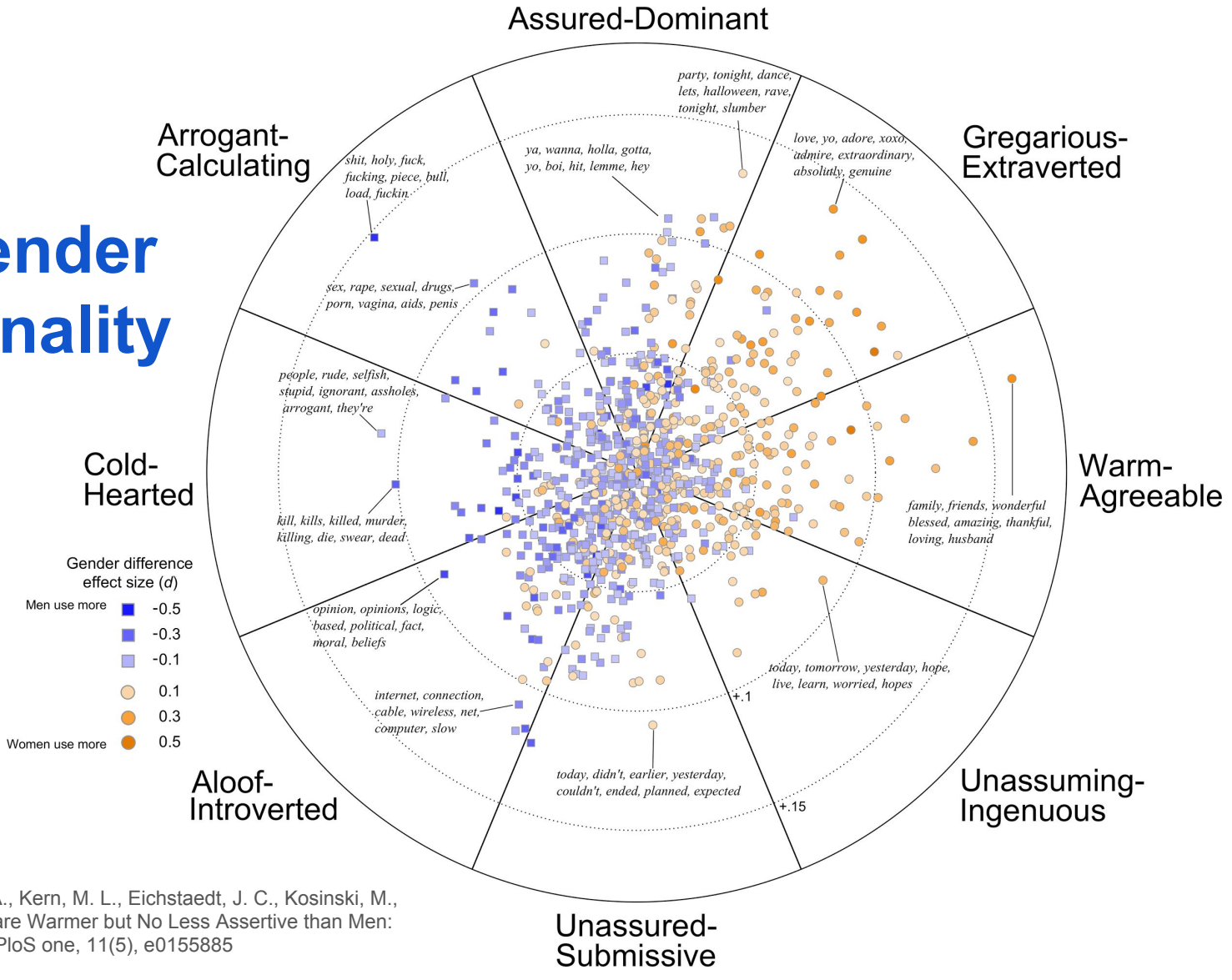


nfl season football baseball fantasy sports play league player playing team game basketball players coach
state government power country freedom thomas nation america human political rights civil democracy liberty society
fight win fighting won battles defeat war battle enemy sword defeated bands meet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

Neuroticism

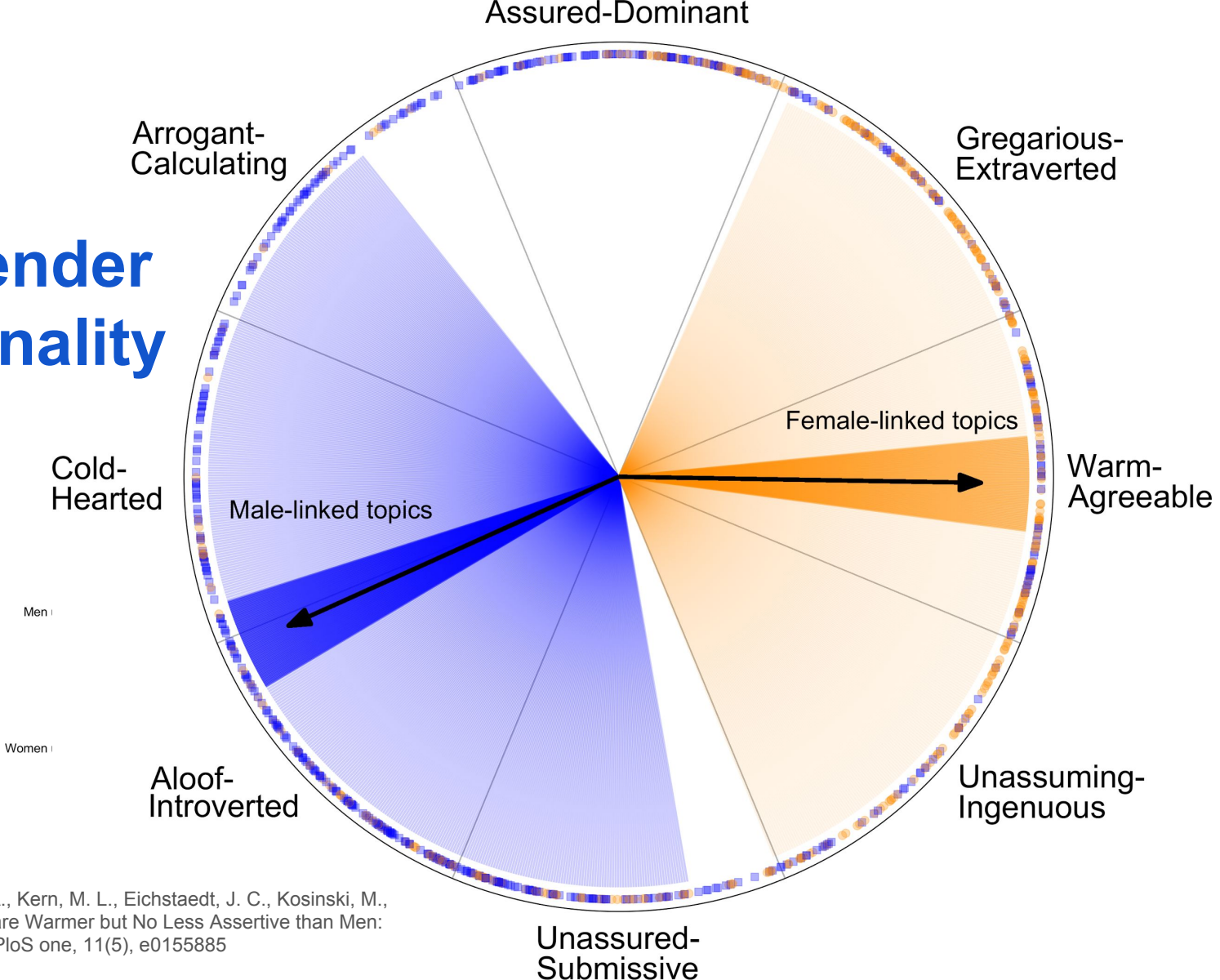


Putting Gender and Personality Together

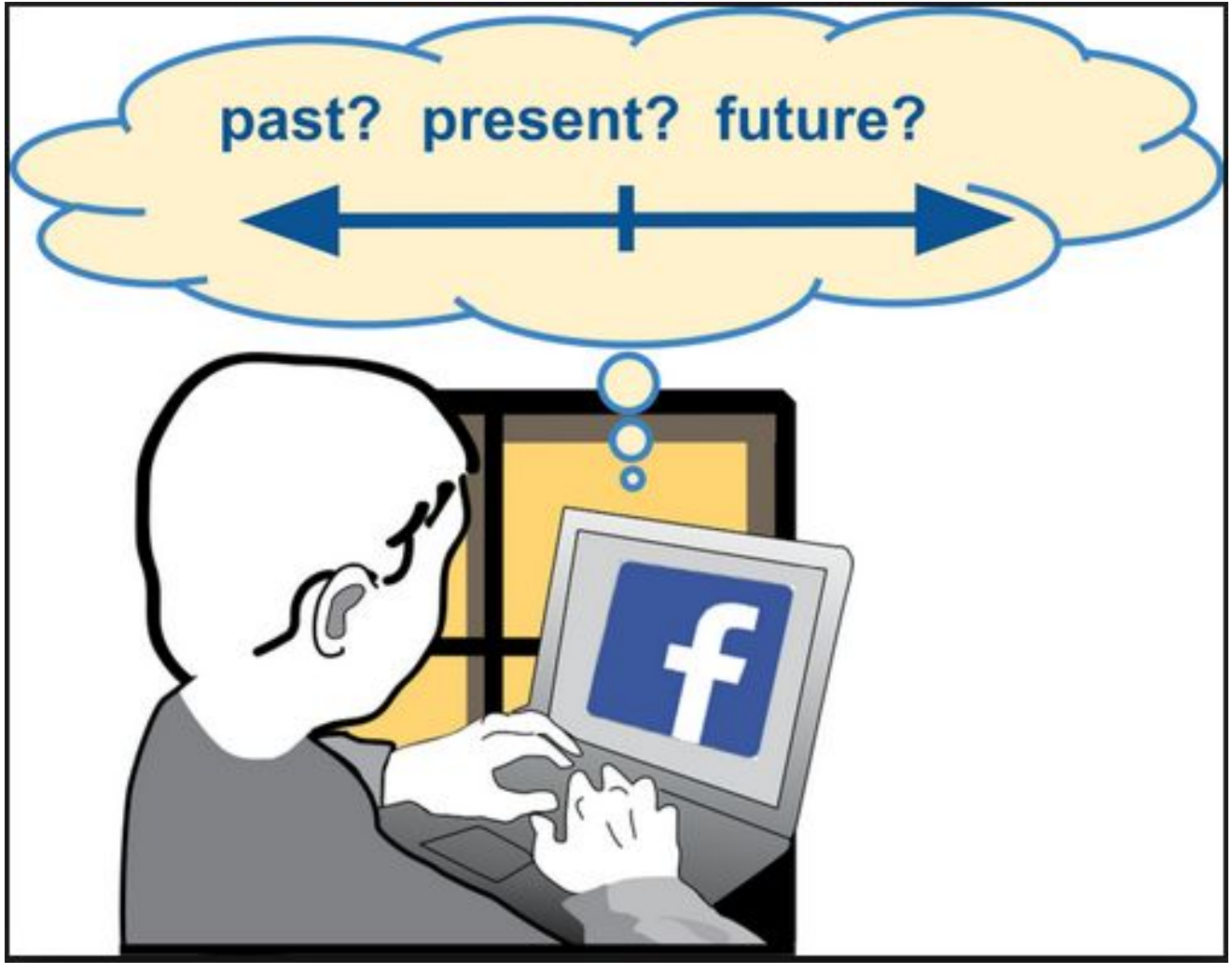


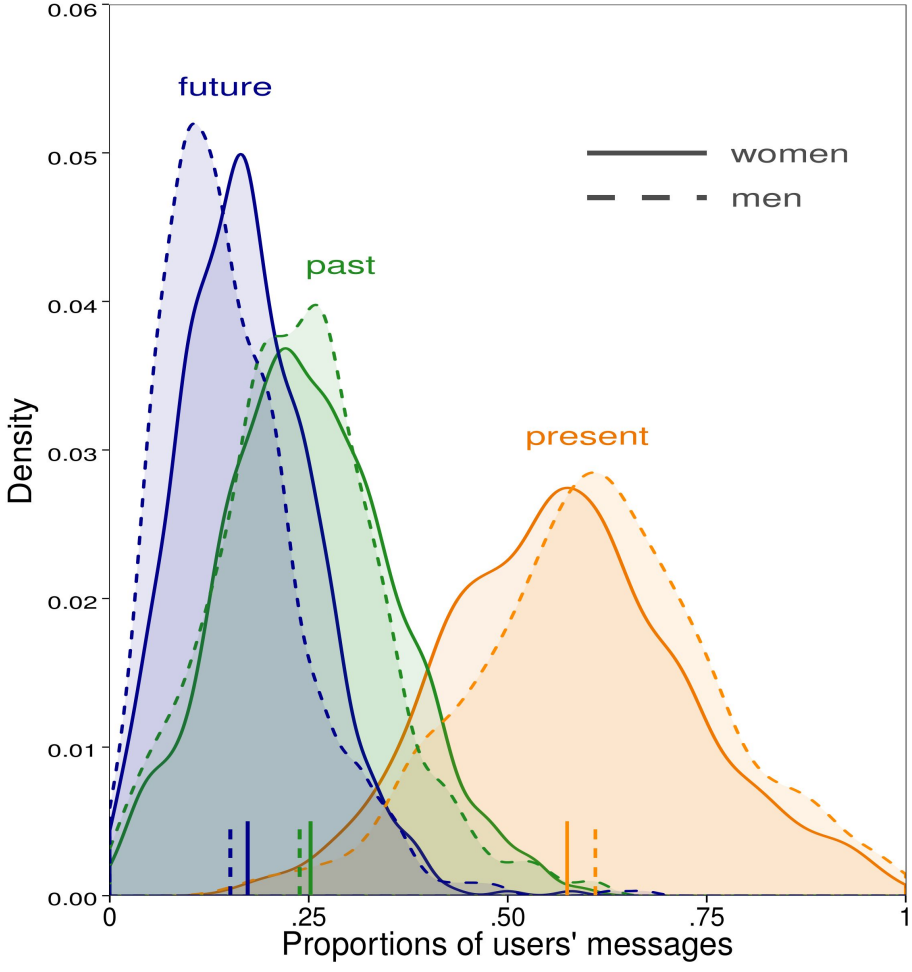
Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., ... & Seligman, M. E. (2016). Women are Warmer but No Less Assertive than Men: Gender and Language on Facebook. PloS one, 11(5), e0155885

Putting Gender and Personality Together



Park, G., Yaden, D. B., Schwartz, H. A., Kern, M. L., Eichstaedt, J. C., Kosinski, M., ... & Seligman, M. E. (2016). Women are Warmer but No Less Assertive than Men: Gender and Language on Facebook. PloS one, 11(5), e0155885


past? present? future?



1.3m status updates
from
4,833 Participants

(3,240 AG Stratified)

Schwartz, H. A., Park, G., Sap, M., Weingarten, E., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Berger, J., Seligman, M., & Ungar, L. (2015). Extracting Human Temporal Orientation from Facebook Language. *NAACL-2015: Conference of the North American Chapter of the Association for Computational Linguistics.*

past present future ⇨ Language

⇨ Questionnaire

IPIP 100 item domains

complete tasks successfully

avoid philosophical discussions

carry out my plans

finish what I start

make plans and stick to them

Park, G., Schwartz, H.A., Sap, M., Kern, M.L., Weingarten, E., Eichstaedt, J.C., Berger, J., Stillwell, D.J., Kosinski, M., Ungar, L.H. & Seligman, M.E. (2015). Living in the Past, Present, and Future: Measuring Temporal Orientation with Language. Journal of personality.

complete tasks successfully

avoid philosophical discussions

carry out my plans

finish what I start

make plans and stick to them

Park, G., Schwartz, H.A., Sap, M., Kern, M.L., Weingarten, E., Eichstaedt, J.C., Berger, J., Stillwell, D.J., Kosinski, M., Ungar, L.H. & Seligman, M.E. (2015). Living in the Past, Present, and Future: Measuring Temporal Orientation with Language. Journal of personality.

do not like art

would describe my experiences as somewhat dull

am easy to satisfy

rarely lose my composure

don’t like to draw attention to myself

Park, G., Schwartz, H.A., Sap, M., Kern, M.L., Weingarten, E., Eichstaedt, J.C., Berger, J., Stillwell, D.J., Kosinski, M., Ungar, L.H. & Seligman, M.E. (2015). Living in the Past, Present, and Future: Measuring Temporal Orientation with Language. Journal of personality.

do not like art

would describe my experiences as somewhat dull

am easy to satisfy

rarely lose my composure

don’t like to draw attention to myself

Park, G., Schwartz, H.A., Sap, M., Kern, M.L., Weingarten, E., Eichstaedt, J.C., Berger, J., Stillwell, D.J., Kosinski, M., Ungar, L.H. & Seligman, M.E. (2015). Living in the Past, Present, and Future: Measuring Temporal Orientation with Language. Journal of personality.

cut others to pieces

can say things beautifully

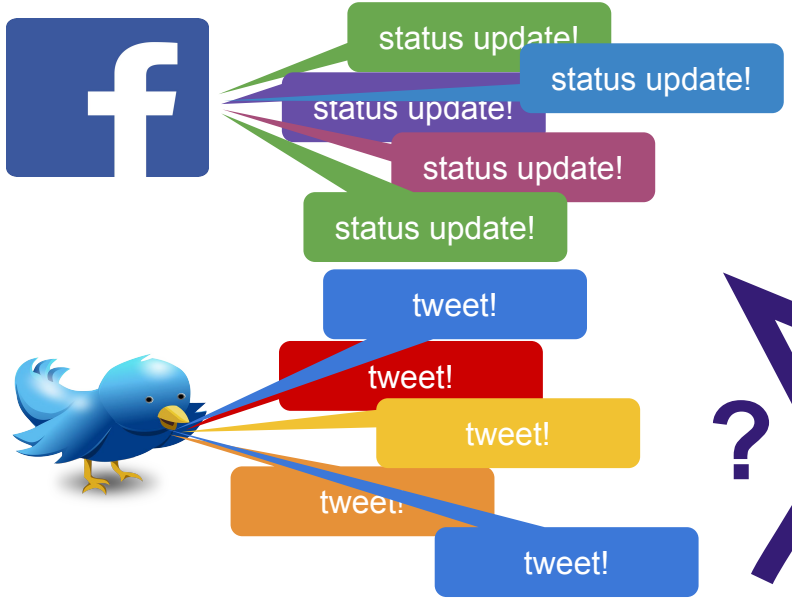
don’t put my mind on the task at hand

have frequent mood swings

am hard to get to know

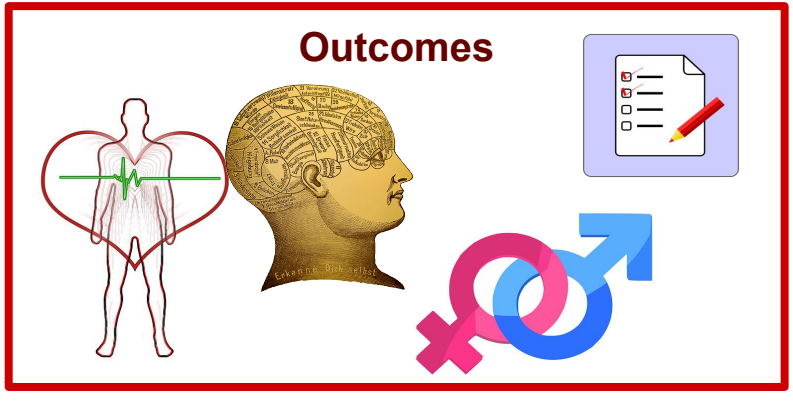
Park, G., Schwartz, H.A., Sap, M., Kern, M.L., Weingarten, E., Eichstaedt, J.C., Berger, J., Stillwell, D.J., Kosinski, M., Ungar, L.H. & Seligman, M.E. (2015). Living in the Past, Present, and Future: Measuring Temporal Orientation with Language. Journal of personality.

PART I: What are the possibilities?



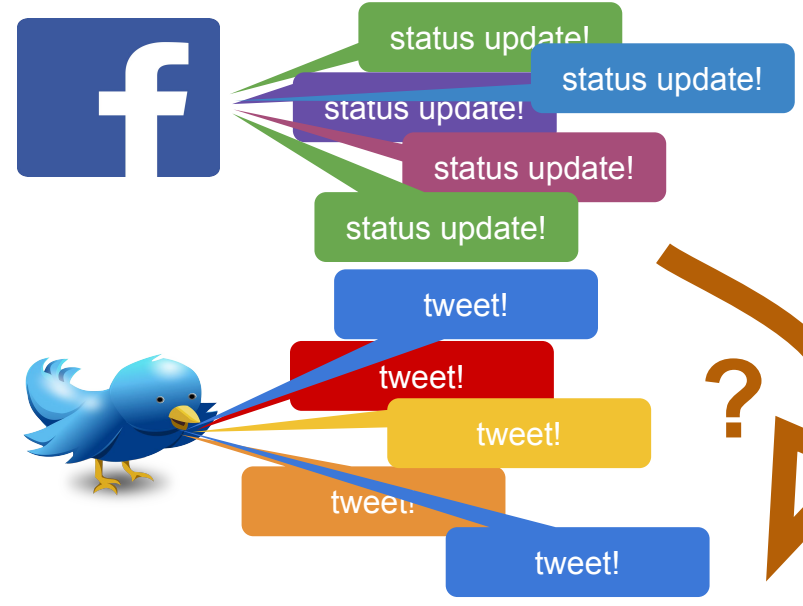
?

language analysis



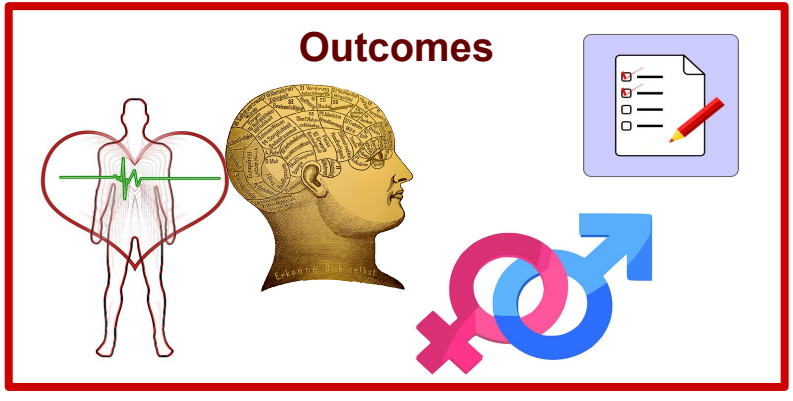
- prediction (measurement)
- insights

PART I: What are the possibilities?



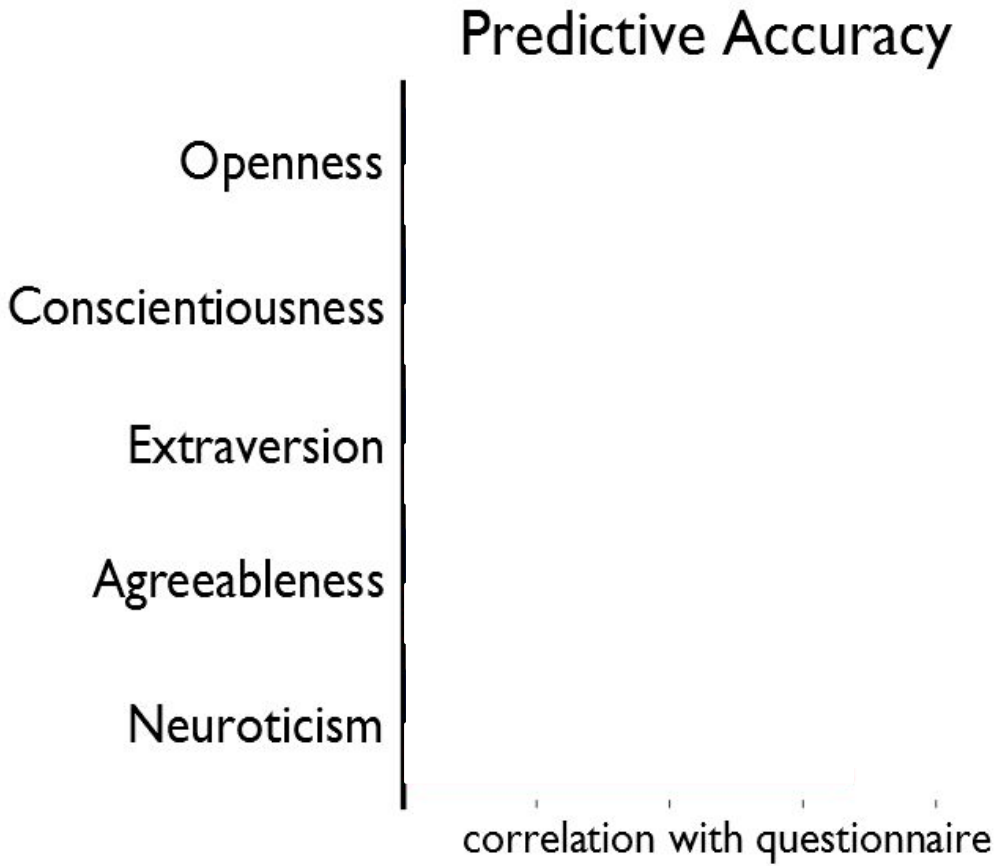
?

language analysis



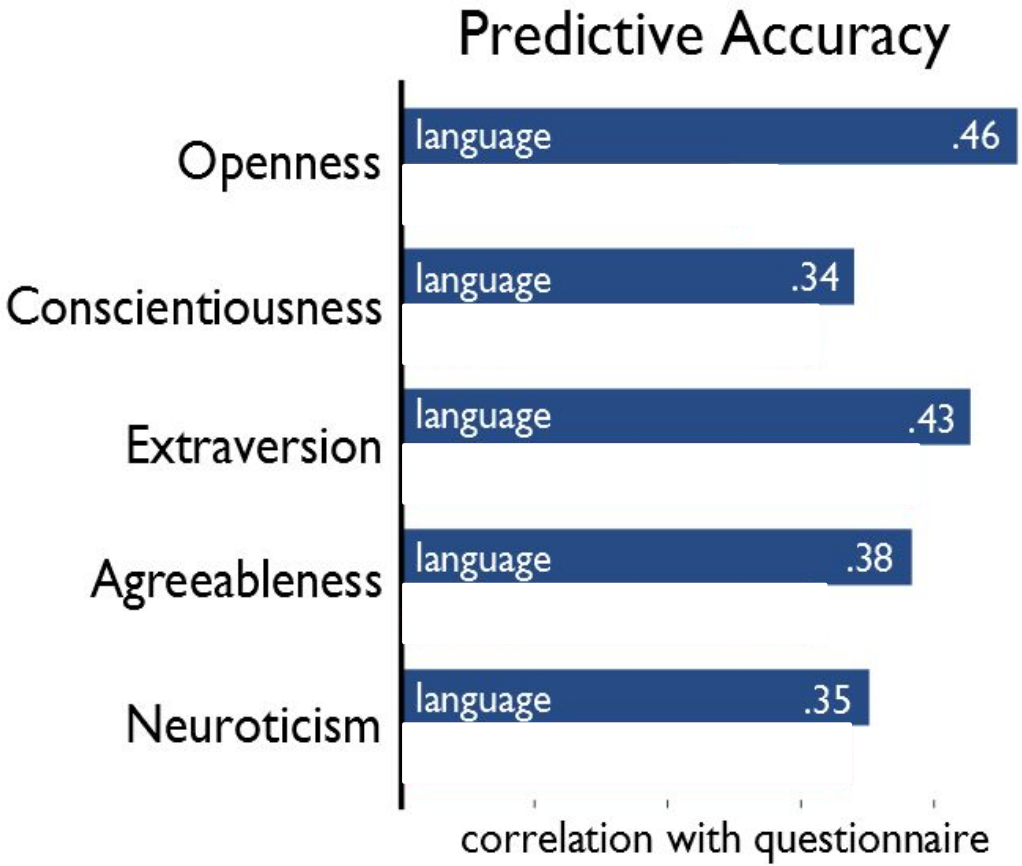
- prediction (measurement)
- insights

Is it accurate? (Personality)



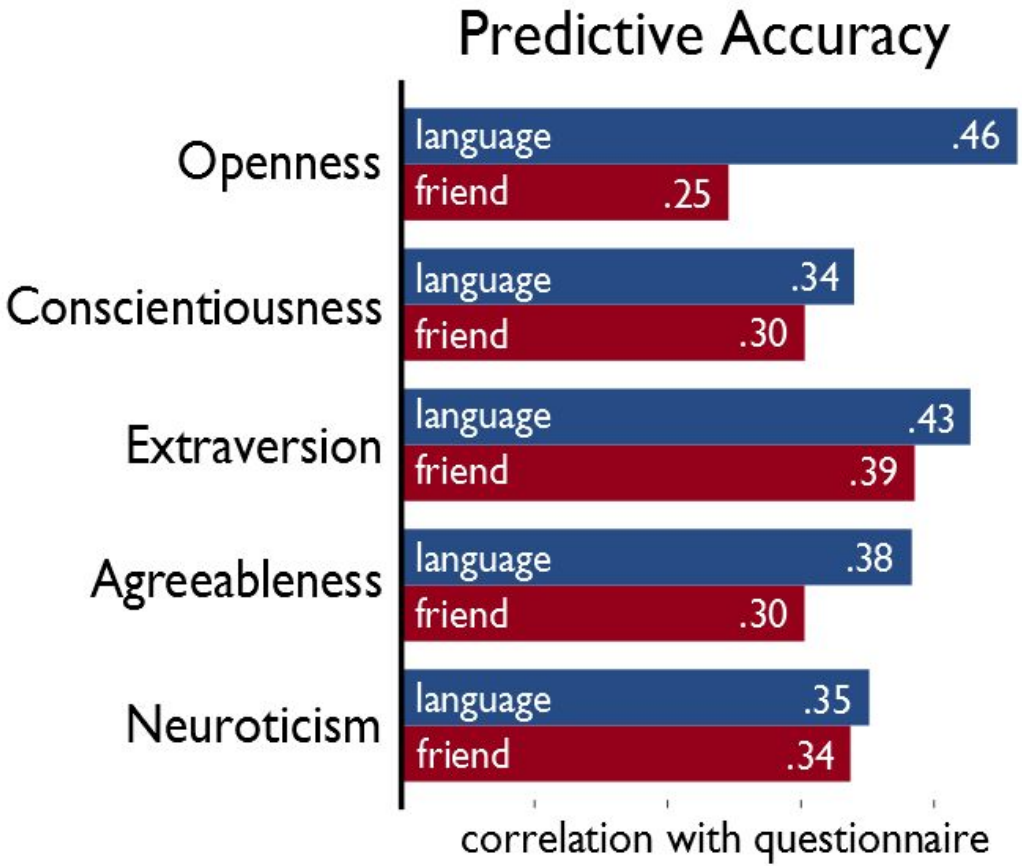
Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., & Seligman, M. E. (in press). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, Nov 3 , 2014.

Is it accurate?



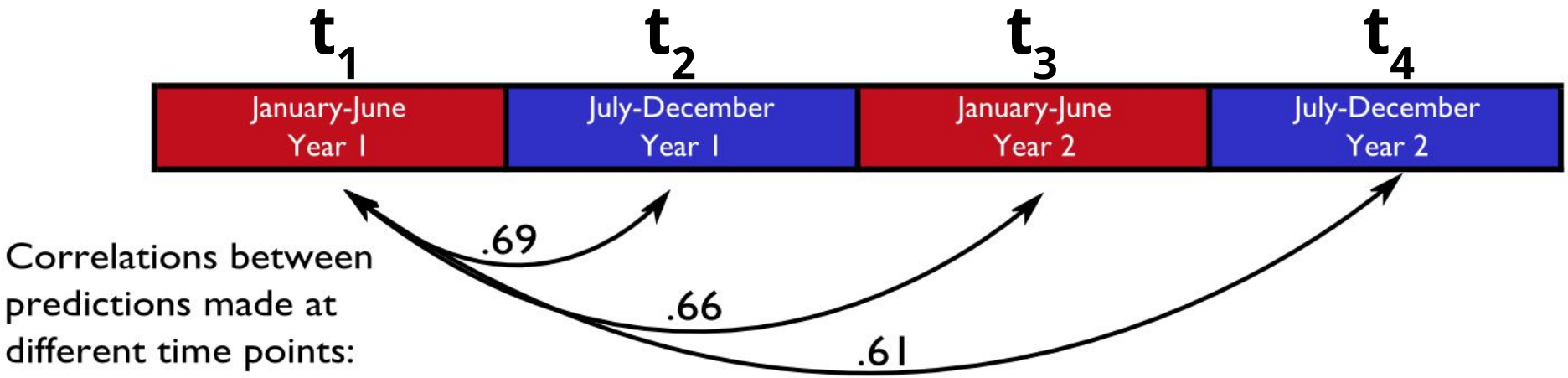
Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., & Seligman, M. E. (in press). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, Nov 3 , 2014.

Is it accurate?



Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., & Seligman, M. E. (in press). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, Nov 3 , 2014.

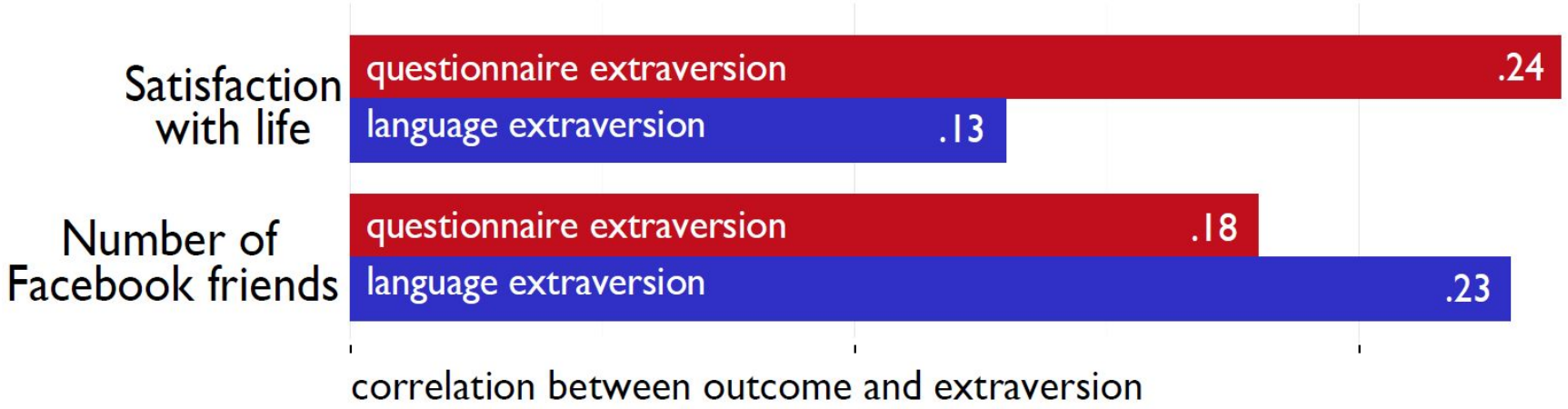
Is it reliable and stable over time?



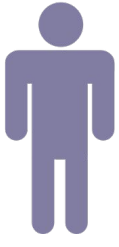
Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., & Seligman, M. E. (in press). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, Nov 3 , 2014.

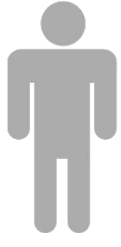
Does it generalize to other outcomes?

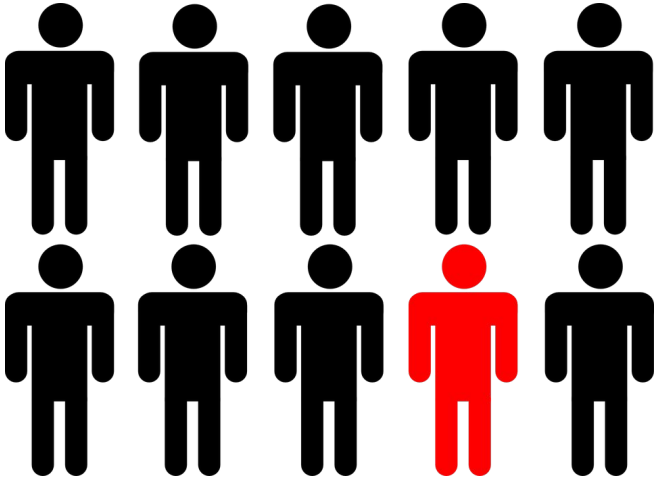
Example: extraversion

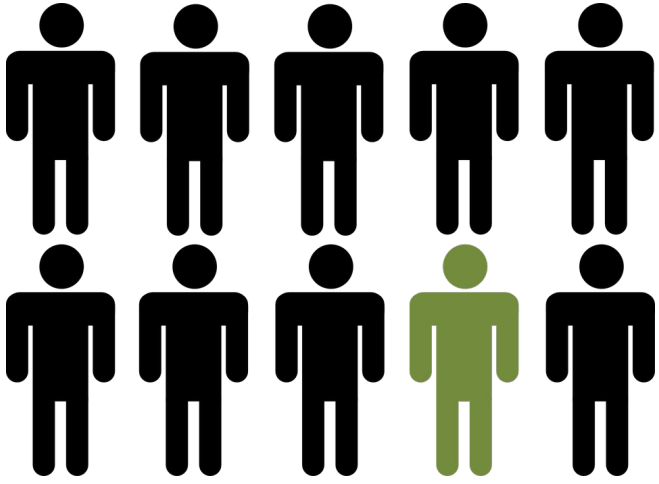


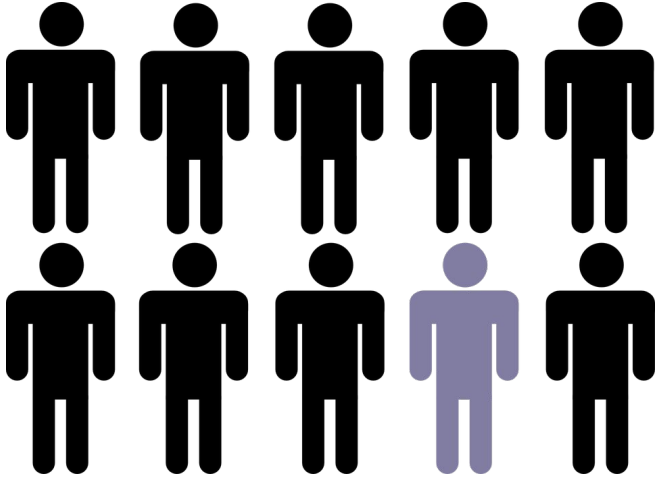
Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., & Seligman, M. E. (in press). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, Nov 3 , 2014.

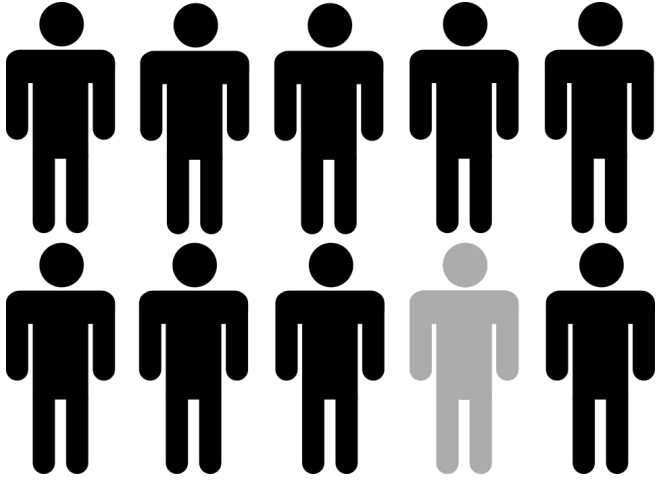


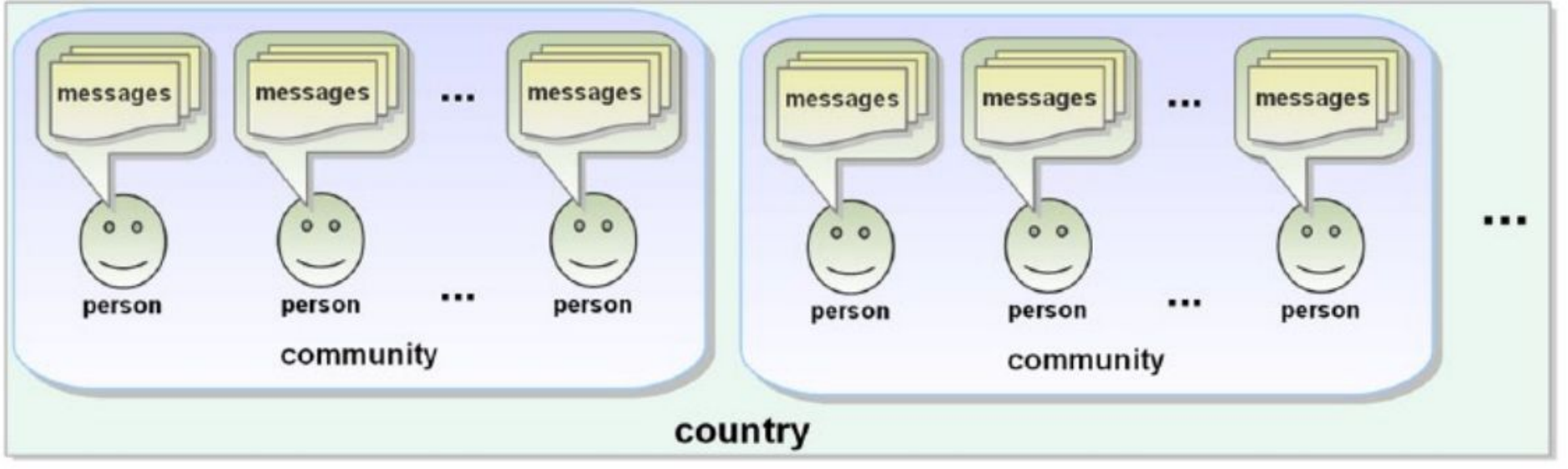




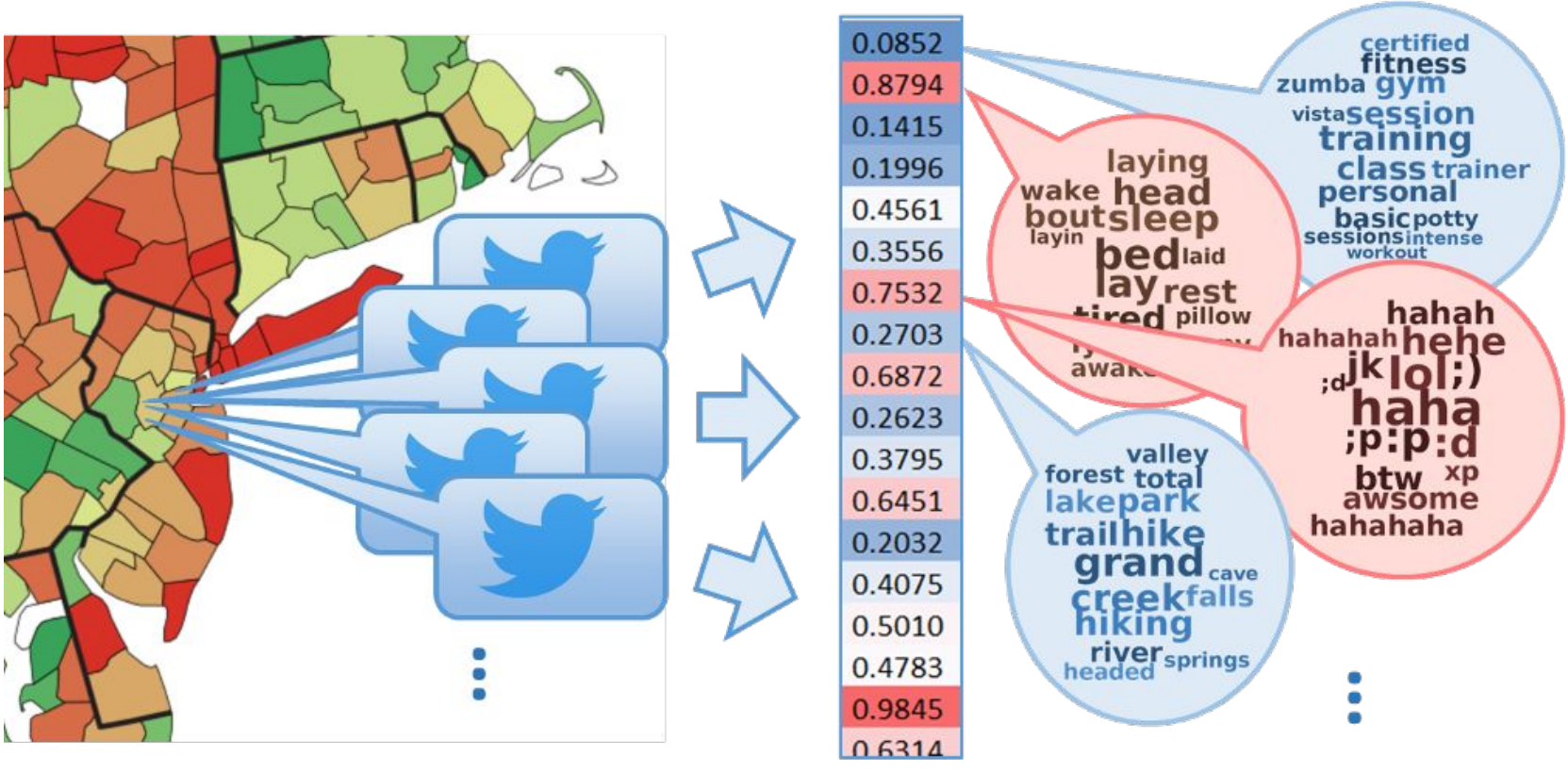




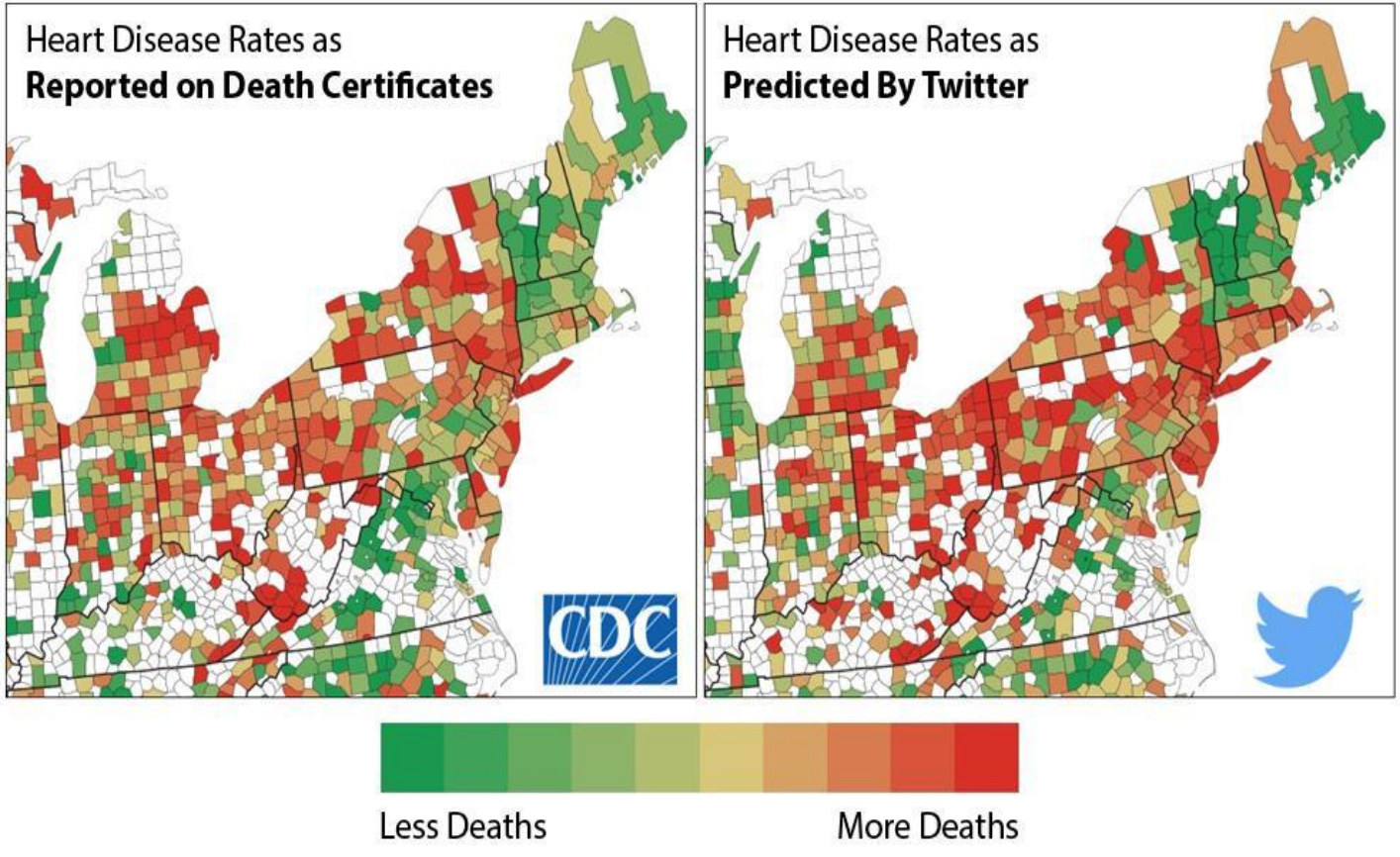



messages
messages
...
messages
person
person
...
person
community
messages
messages
...
messages
person
person
...
person
community
...
country

Community Feature Extraction

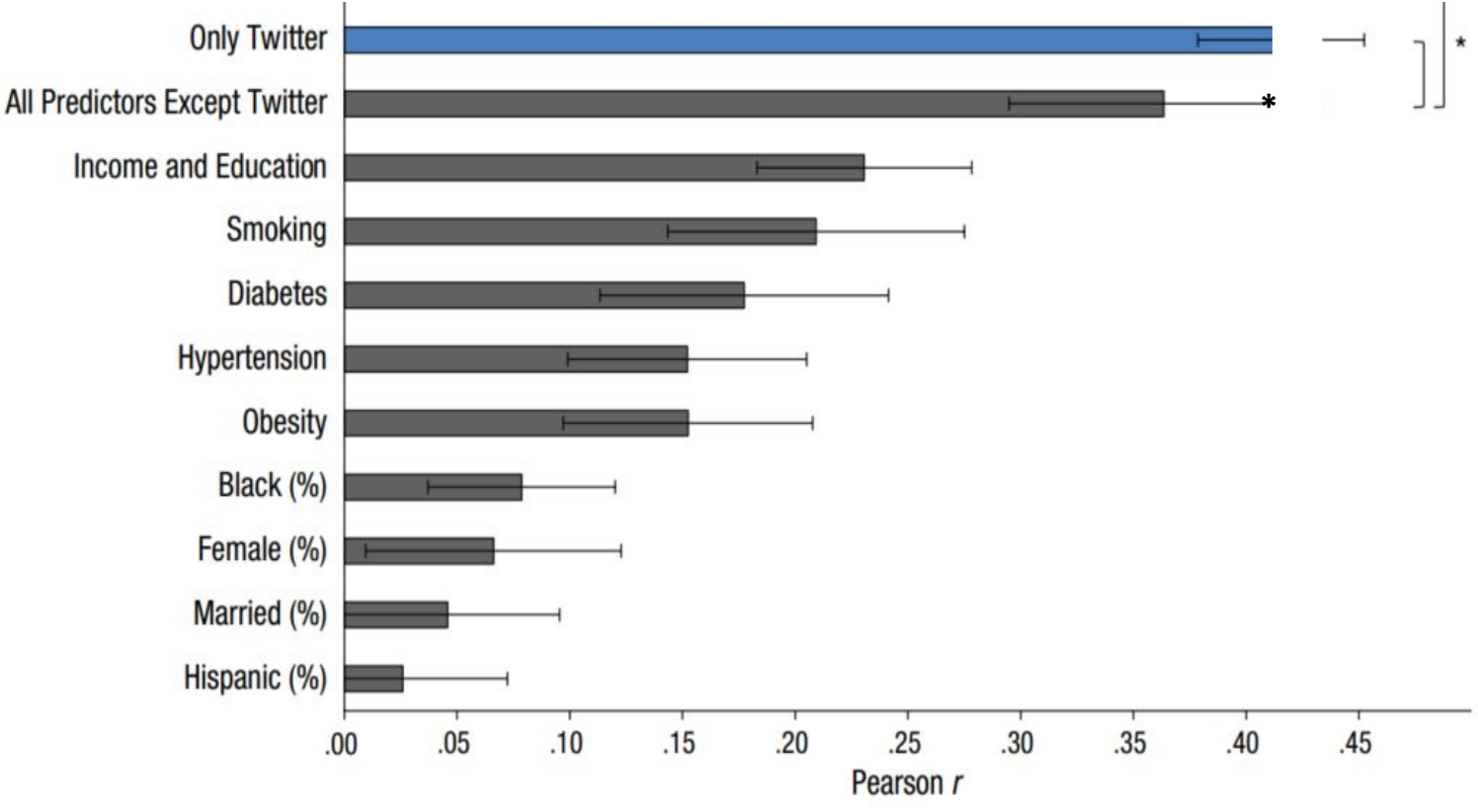


Predicting heart disease: **Accuracy**



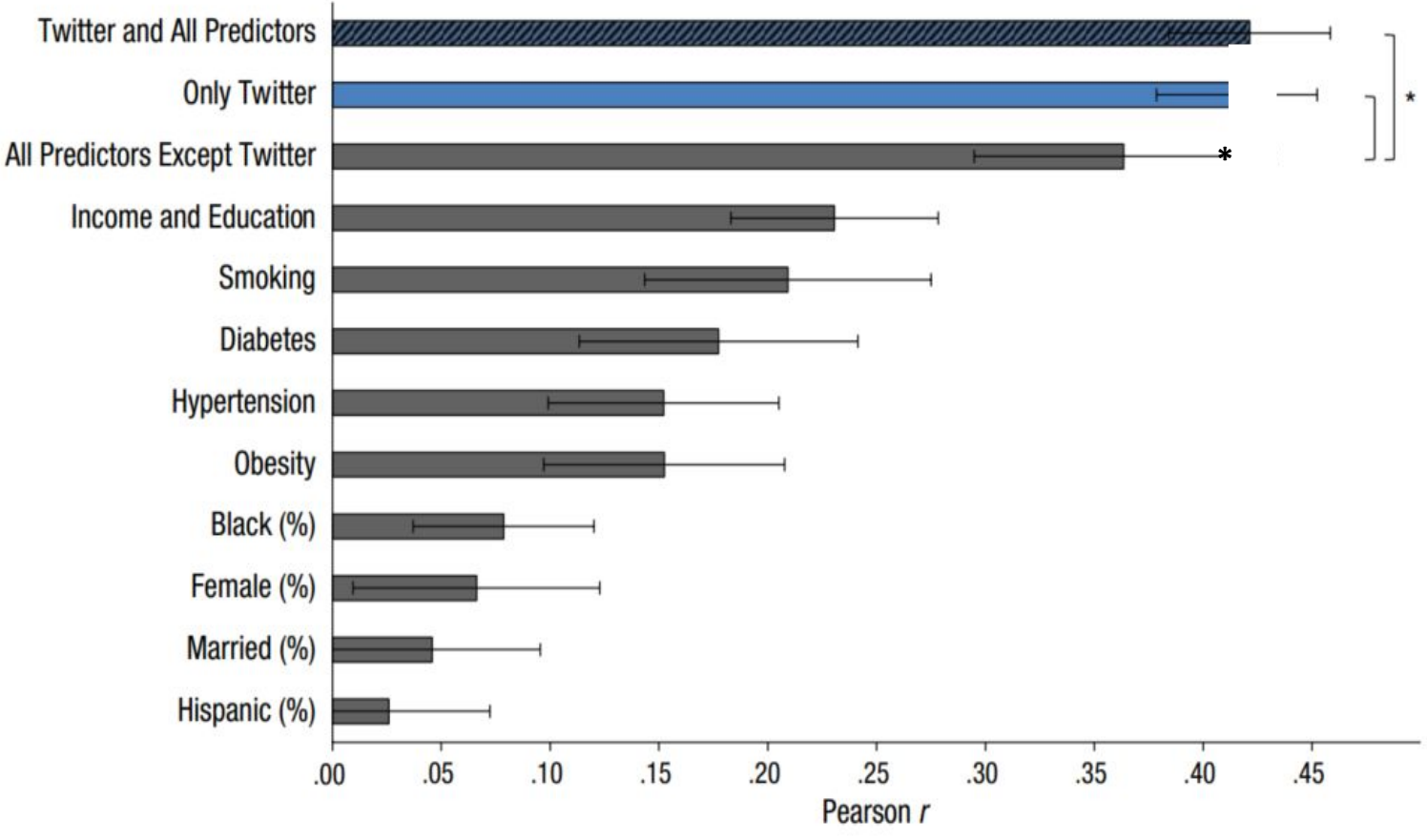
Eichstaedt, Schwartz, et al., 2015, *Psych Science*

Predicting heart disease: **Accuracy**



Eichstaedt, Schwartz, et al., 2015, *Psych Science*

Predicting heart disease: **Accuracy**



Eichstaedt, Schwartz, et al., 2015, *Psych Science*

Associated with Heart Disease

Hate, Interpersonal Tension

Optimism

Eichstaedt, Schwartz, et al., 2015, *Psych Science*

Community Real Estate

VS

	socioeconomics		demographics		socioeconomics + demographics	
	Fc	Ip	Fc	Ip	Fc	Ip
no lang	0.34	0.42	0.24	0.44	0.37	0.50
with lang	**0.41**	**0.56**	**0.39**	**0.57**	**0.42**	**0.59**

Zamani M, Schwartz HA. Using Twitter Language to Predict the Real Estate Market. EACL 2017. 2017 Apr 3:28.

Community Real Estate

PART II: How?

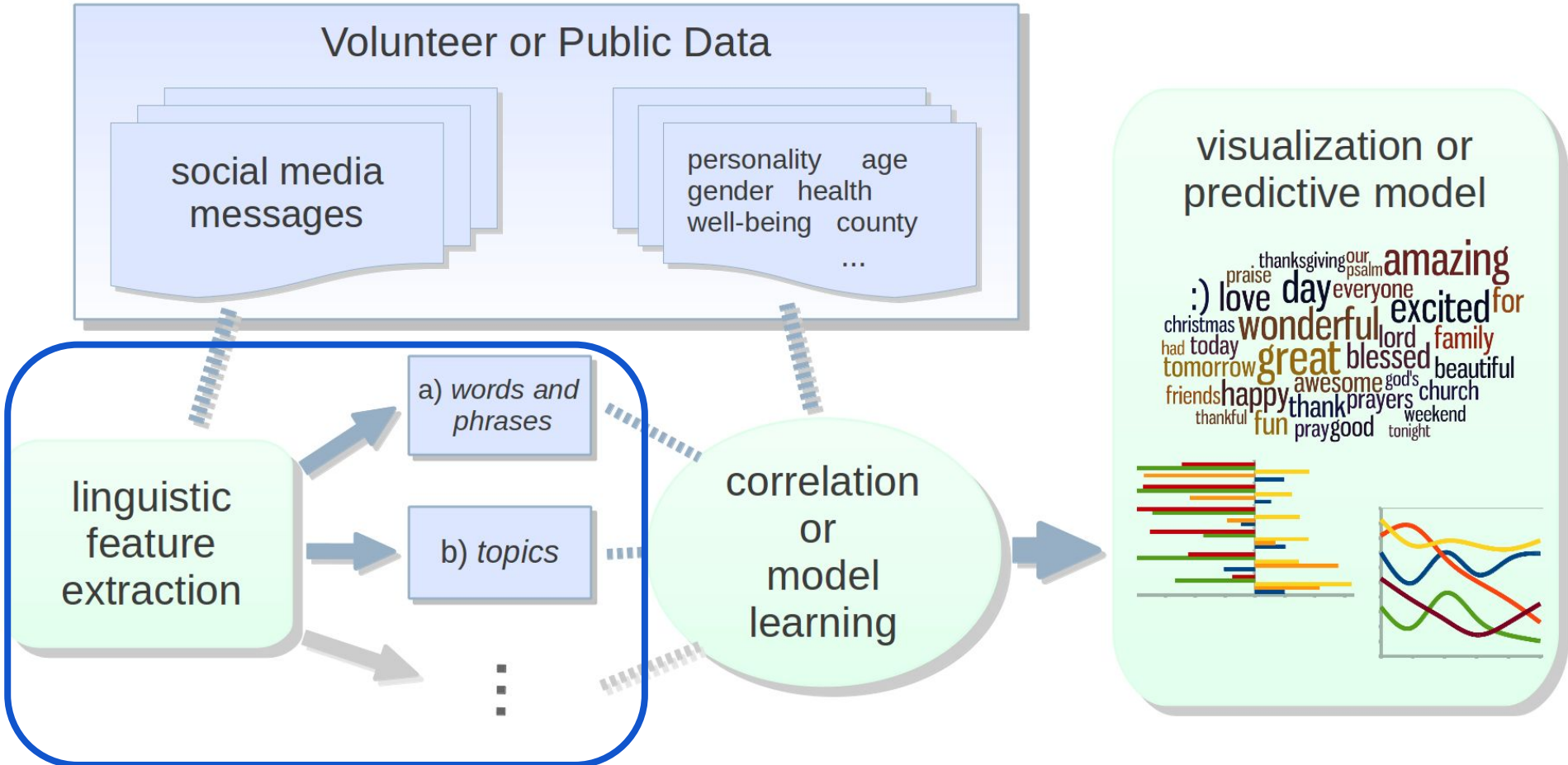
PART II: How?

Two Examples:

1. Community Real Estate

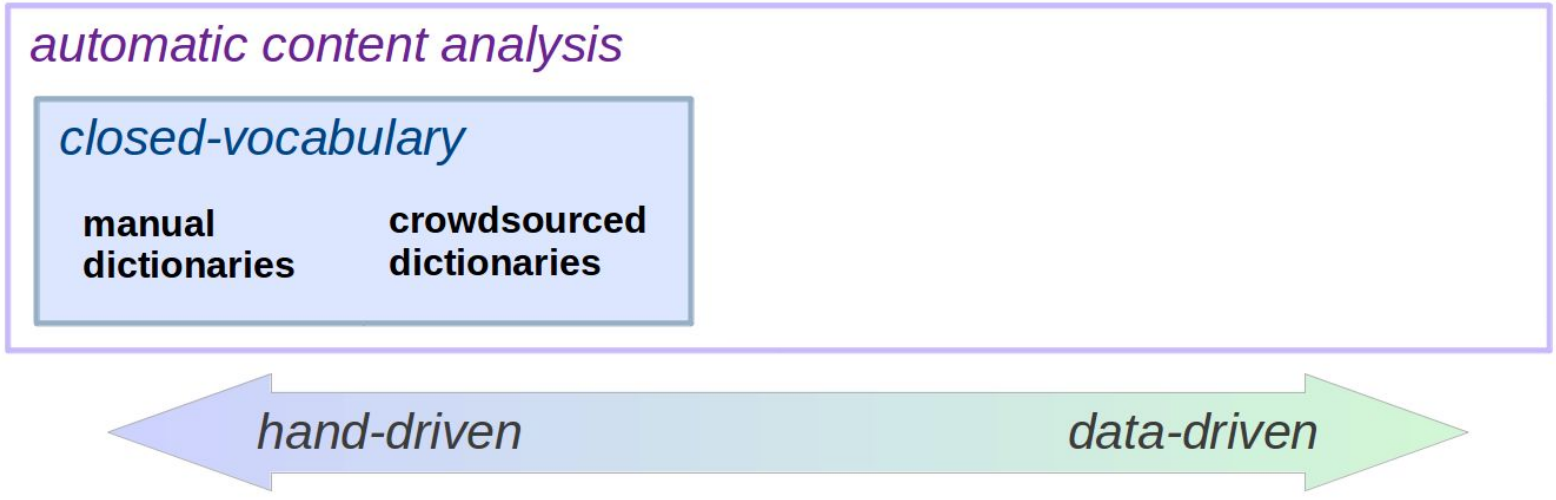
2. Message Temporal Orientation

PART II: How?

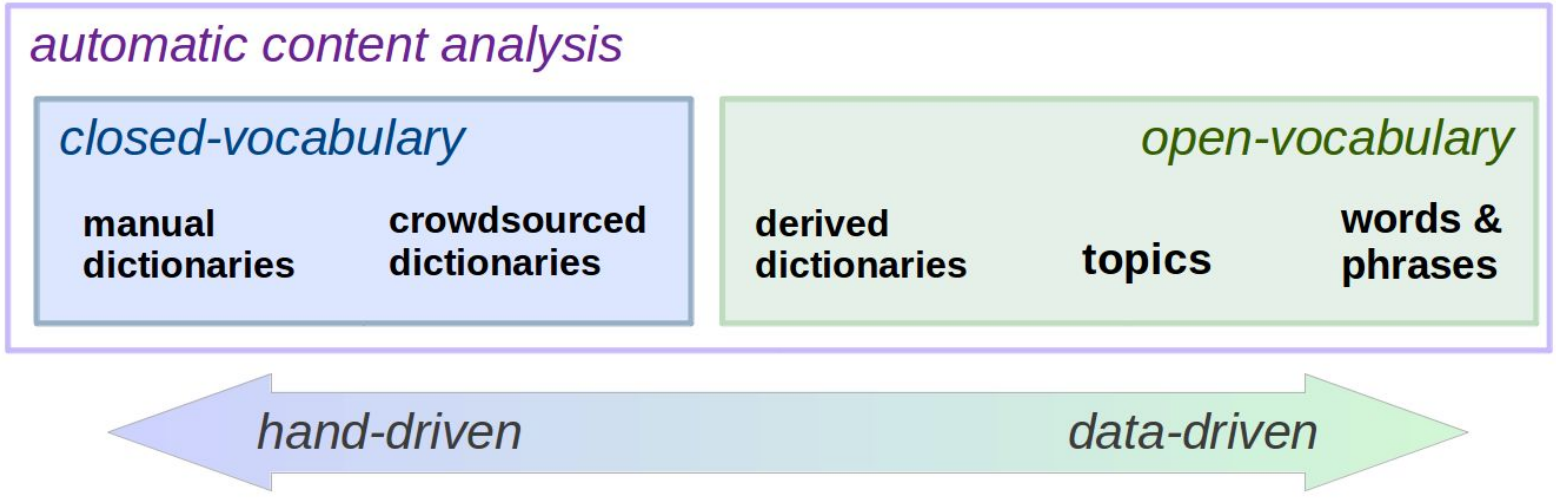

Volunteer or Public Data
social media messages
personality age gender health well-being county ...
linguistic feature extraction
a) words and phrases
b) topics
correlation or model learning
visualization or predictive model
thanksgiving our psalm amazing
praise
:) love day everyone excited for
christmas wonderful lord family
had today
tomorrow great blessed beautiful
friends happy awesome god's church
thankful fun thank prayers weekend
pray good tonight

Language Feature Space

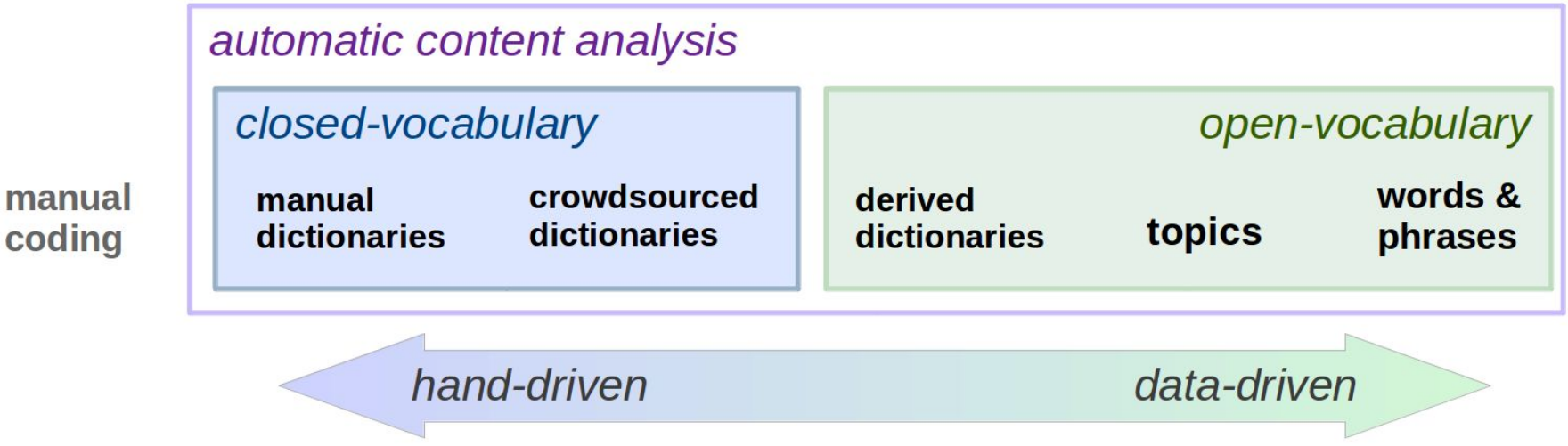
Language Feature Space

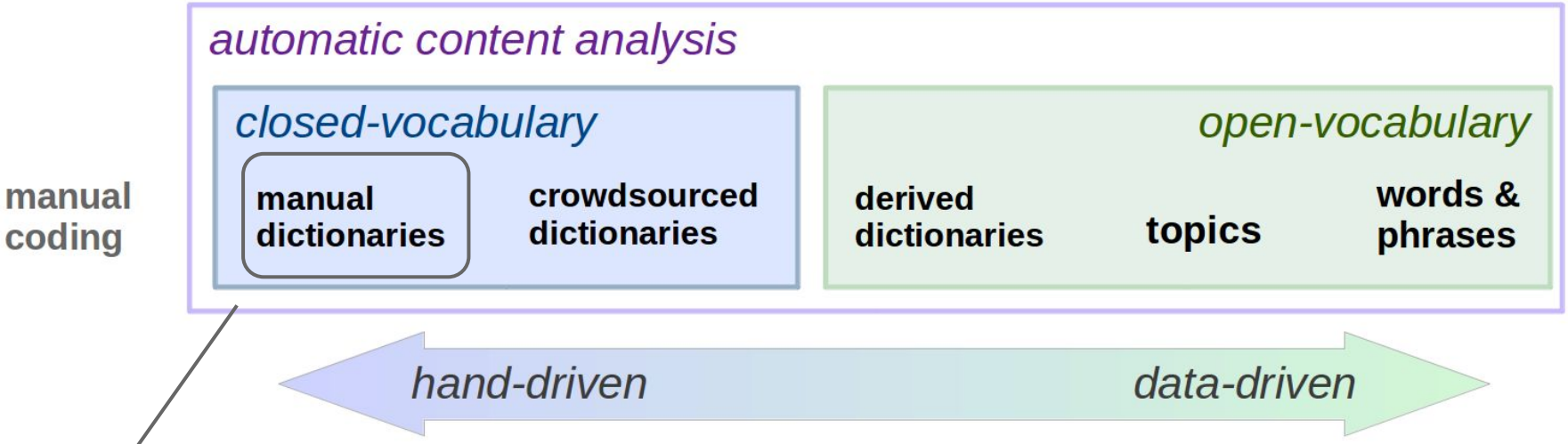


Language Feature Space

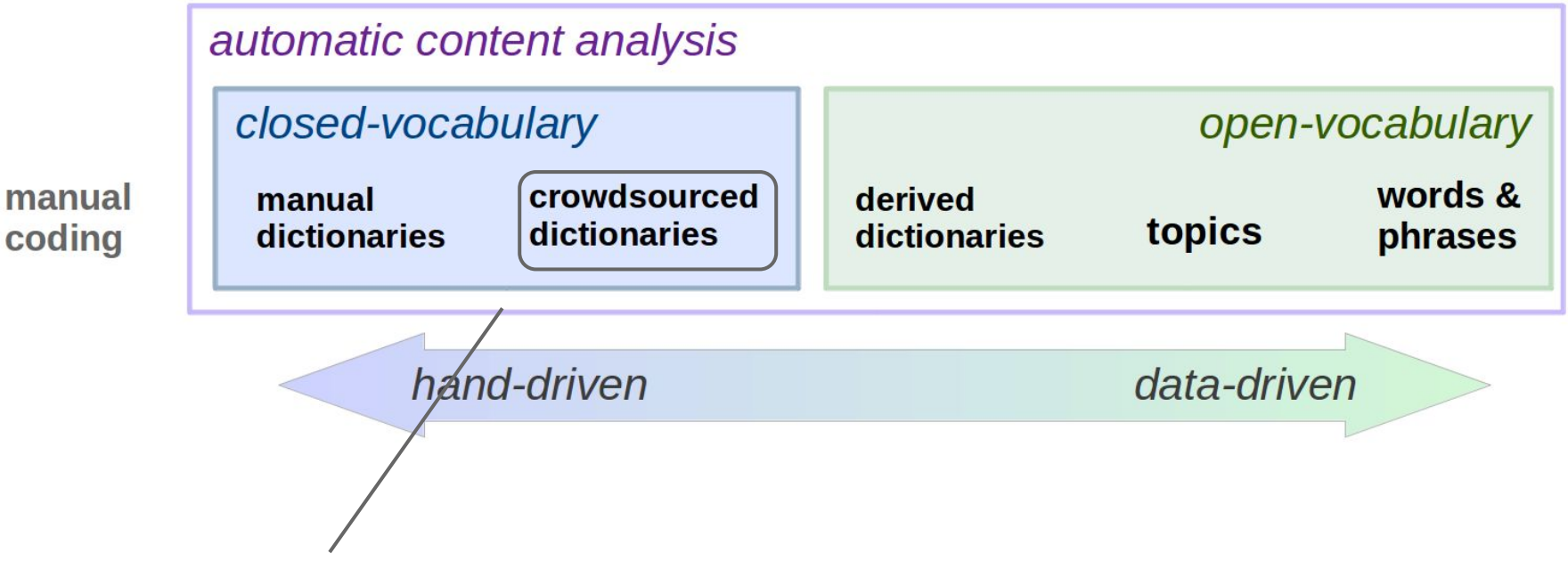


Language Feature Space





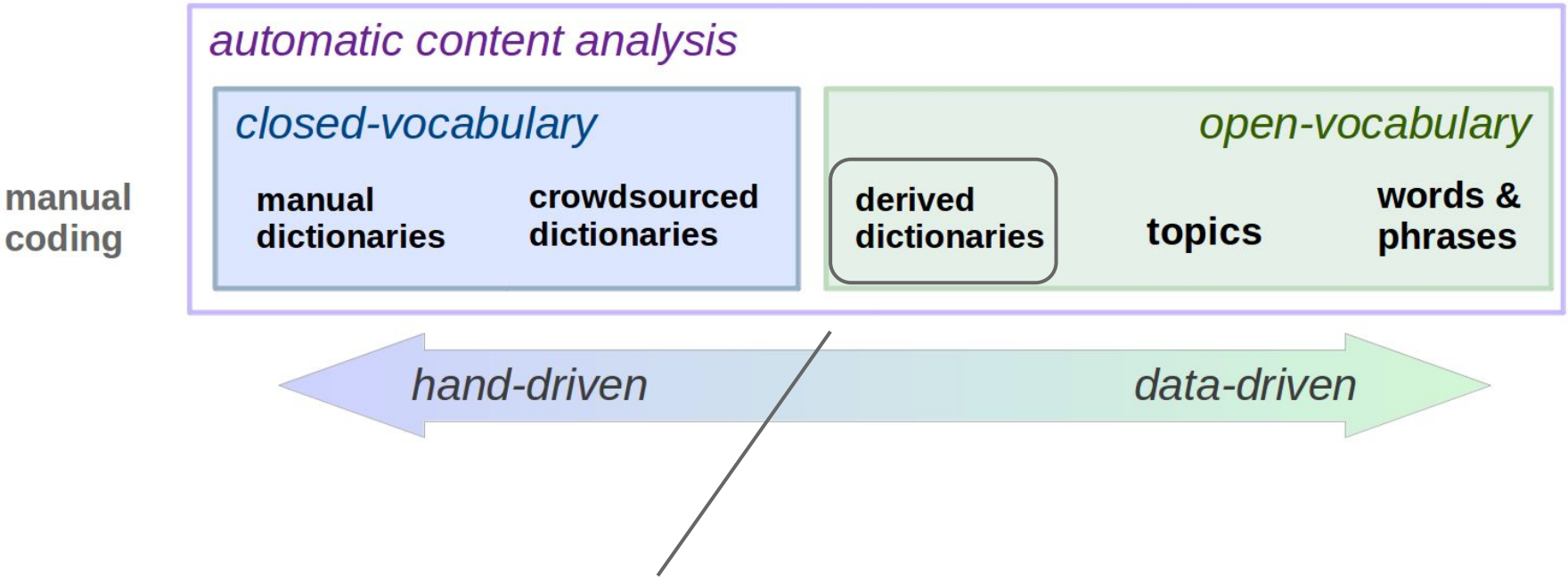
Example:
Linguistic Inquiry and Word Count
(LIWC; Pennebaker et al., 2007)



Examples:

- ANEW
  (Bradley & Lang, 1996)
- Hedonometer
  (Dodds et al., 2011)

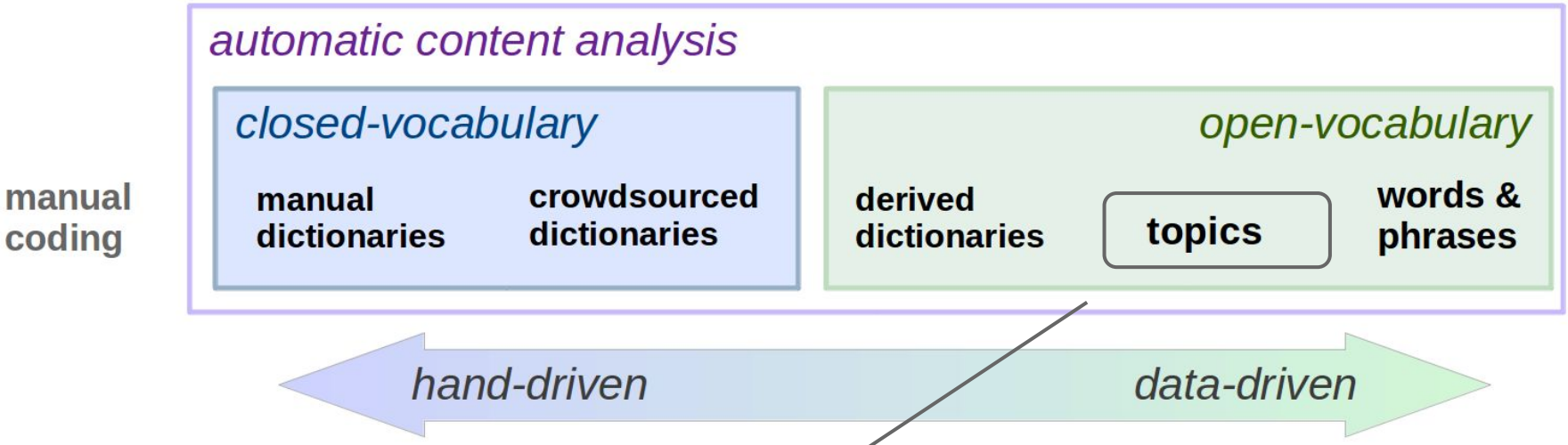
+ often weighted
+ fuller coverage



Examples:

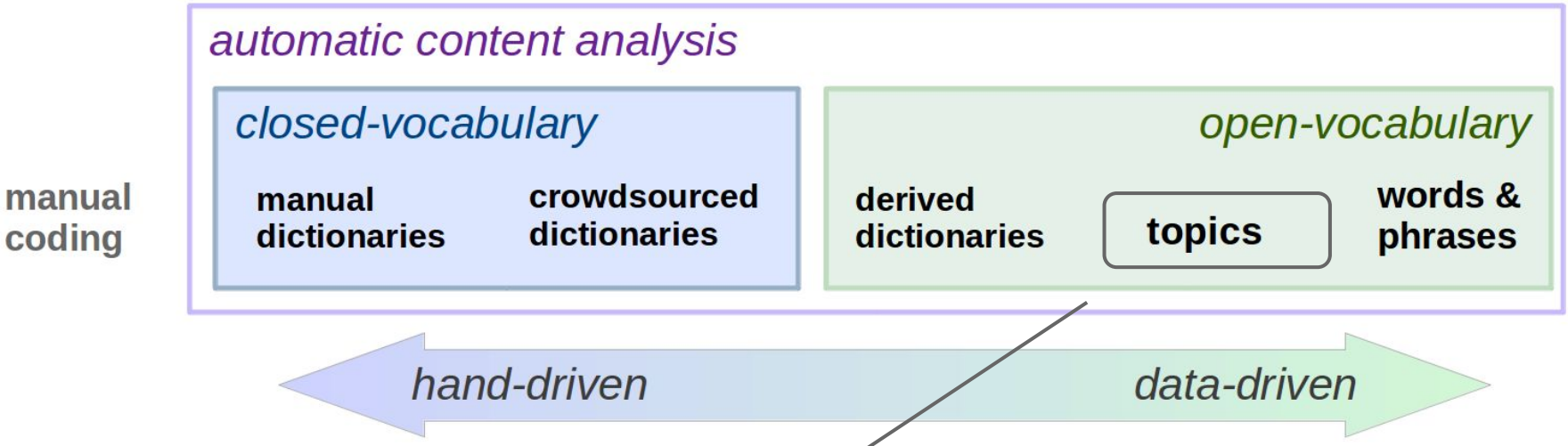
- Sentiment
  (Pang & Lee, 2002)
- Affect & Intensity
  (Preotiuc et al., 2016)

+ real world distributions
(still hypothesis driven)



Topic Modeling:

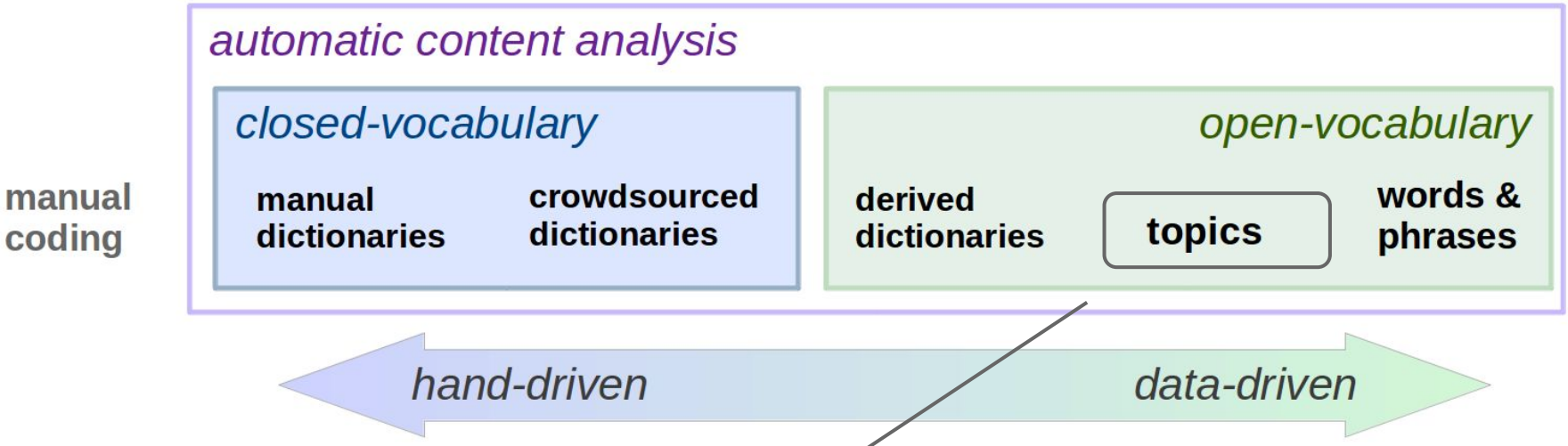
+ completely data-driven
+ “digestable”
(still losing some information)



Topic Modeling:

+ completely data-driven
+ “digestable”
(still losing some information)

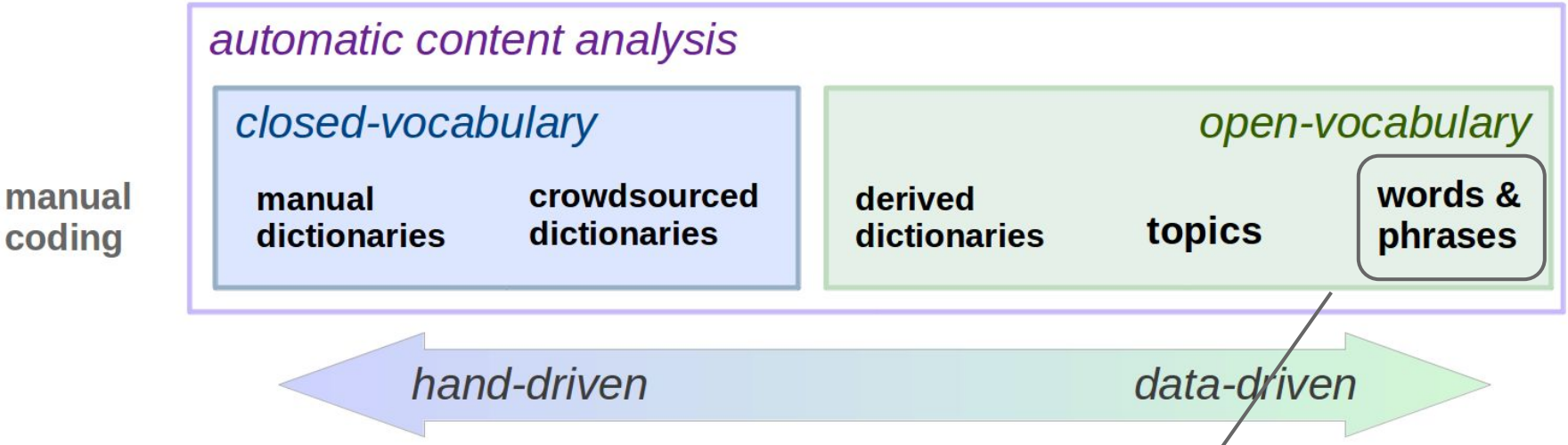
$$p(topic|tweet) = \sum_{term \in topic} p(topic|term)p(term|tweet)$$



Topic Modeling:

+ completely data-driven
+ “digestable”
(still losing some information)

$$p(topic|user) = \sum_{term \in topic} p(topic|term)p(term|user)$$

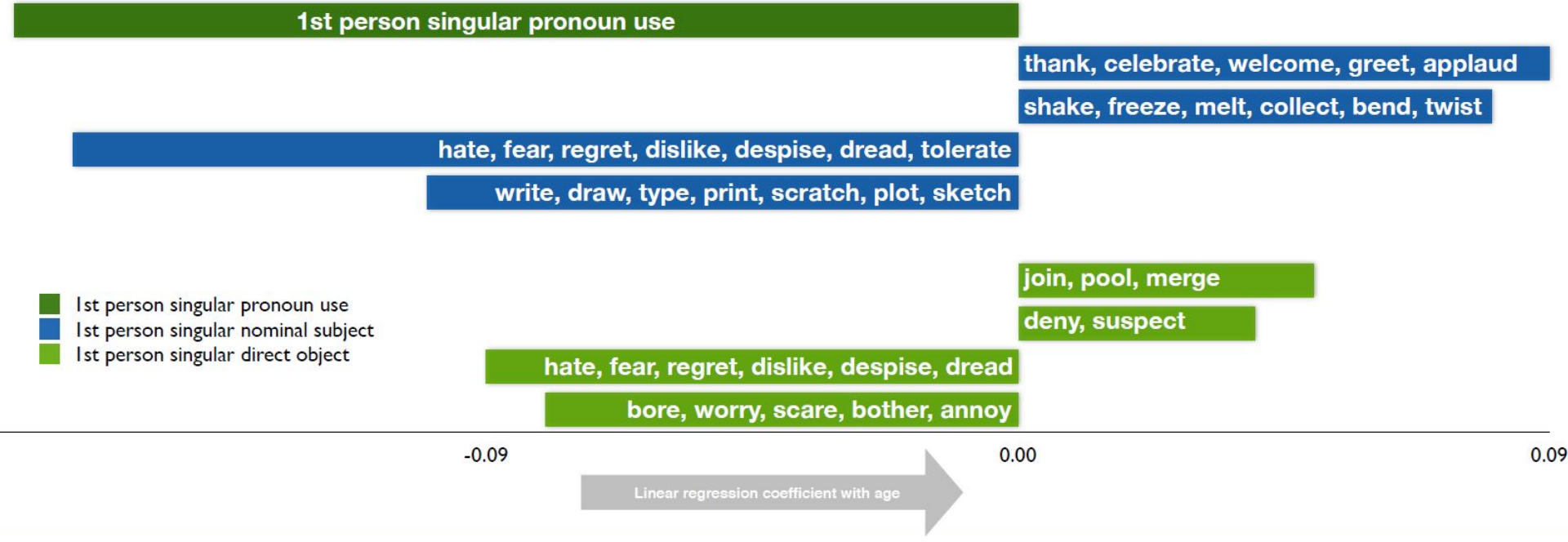


+ wide coverage
+ fine-grained information
phrases: capture some context

(not as “digestable”)

What about context?

What about context?

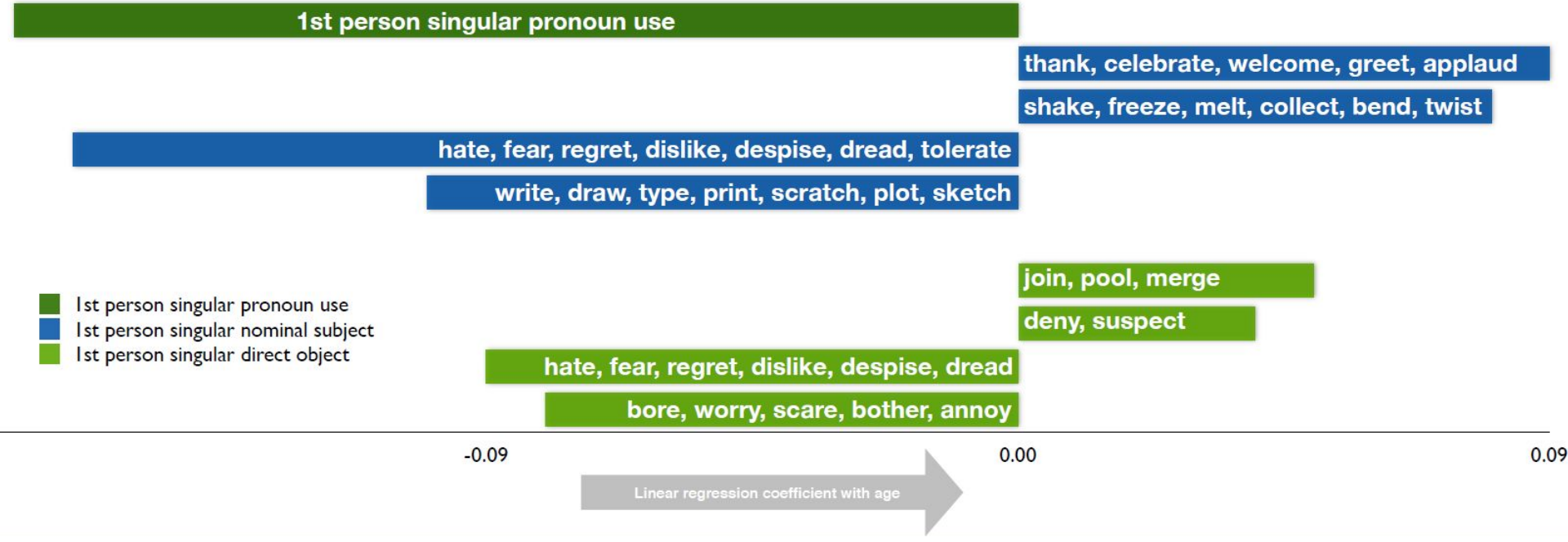


I as subject of <verb>

me as direct object of <verb>

Rouhizadeh, M., Ungar, L., Buffone, A. & Schwartz, H.A (2016). Using Syntactic and Semantic Context to Explore Psychodemographic Differences in Self-reference. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing

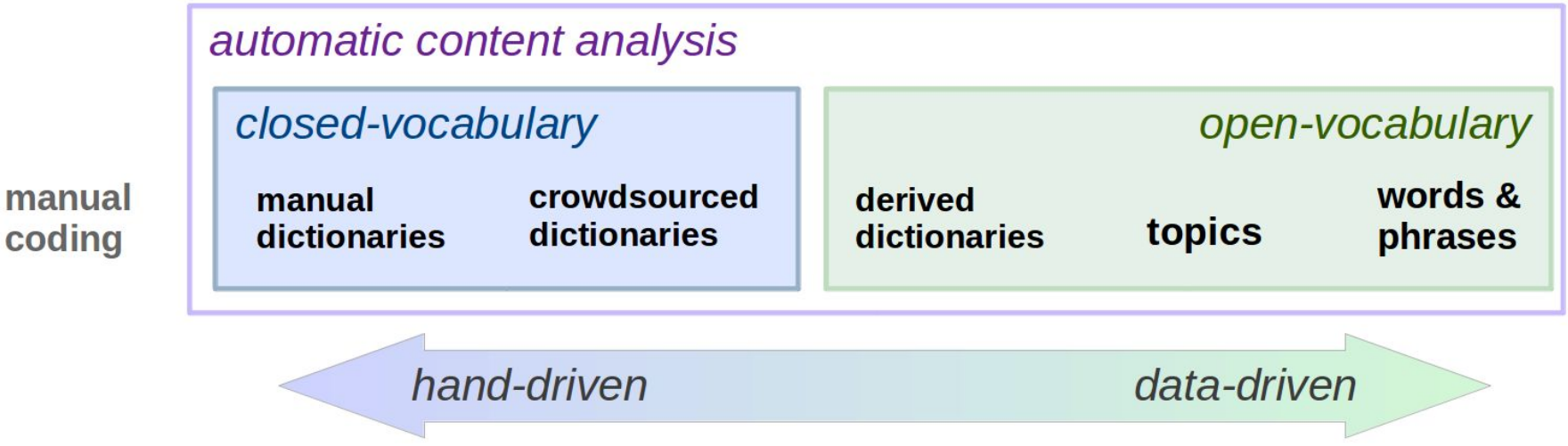
What about context?



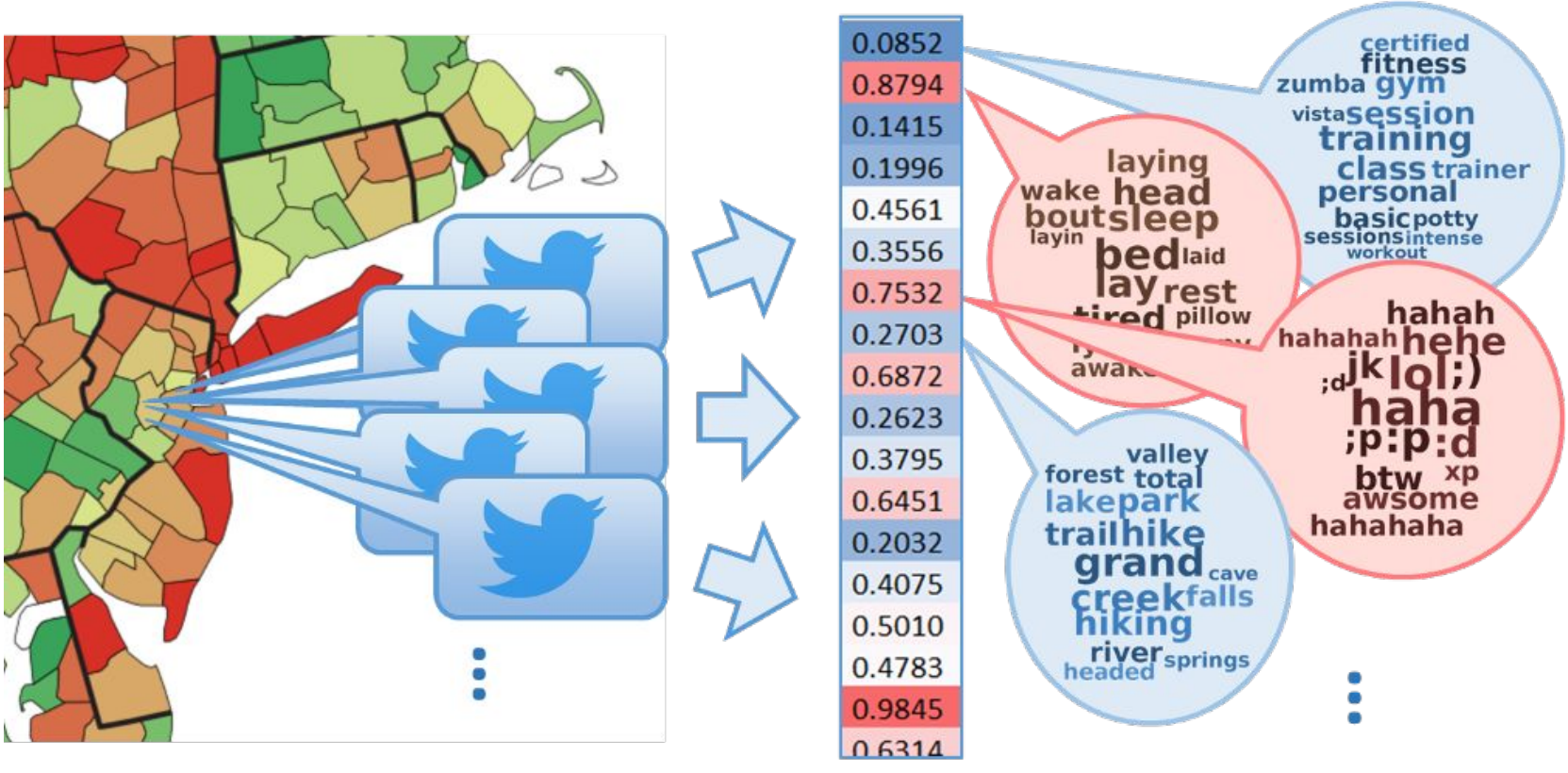
For human-level **prediction**, extent of benefit is open question.

Rouhizadeh, M., Ungar, L., Buffone, A. & Schwartz, H.A (2016). Using Syntactic and Semantic Context to Explore Psychodemographic Differences in Self-reference. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing

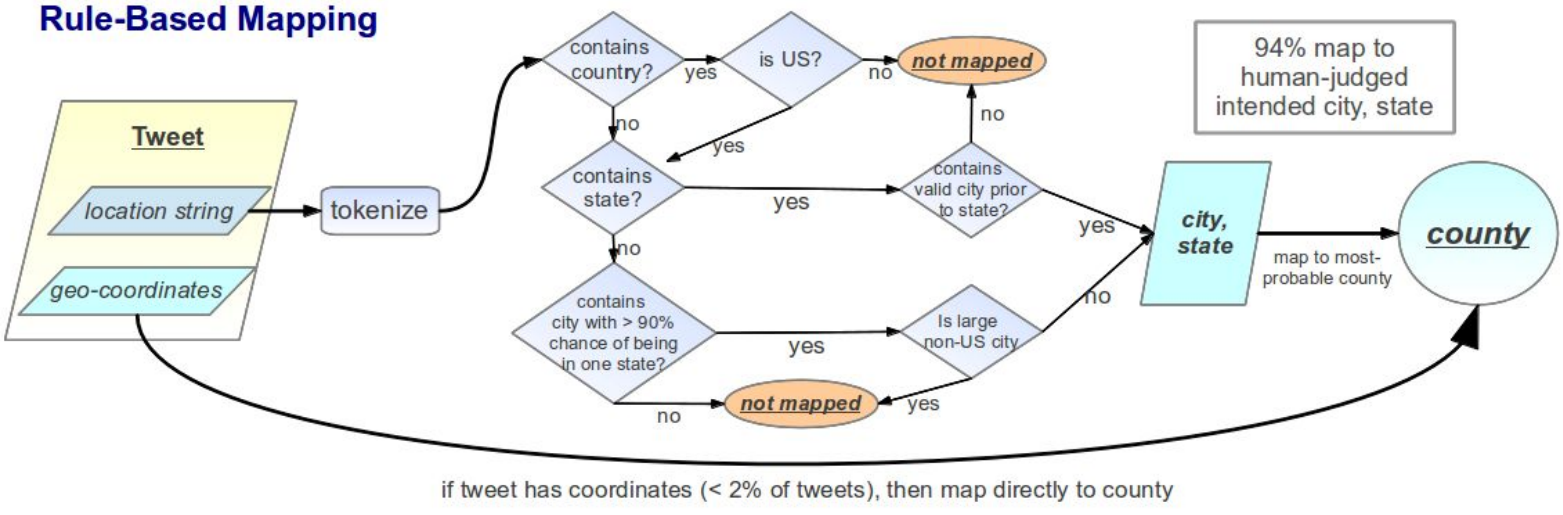
Language Feature Space



county_id, feature1, value



Method: County-Mapping



94% accurate map to human-judged intended city, state pair.

Large Data

Distributed Feature Extraction

- approximate 1 billion tweets
  - for a single computer system to read the data is very time-consuming
- Utilize map-reduce in a “Hadoop” style cluster:

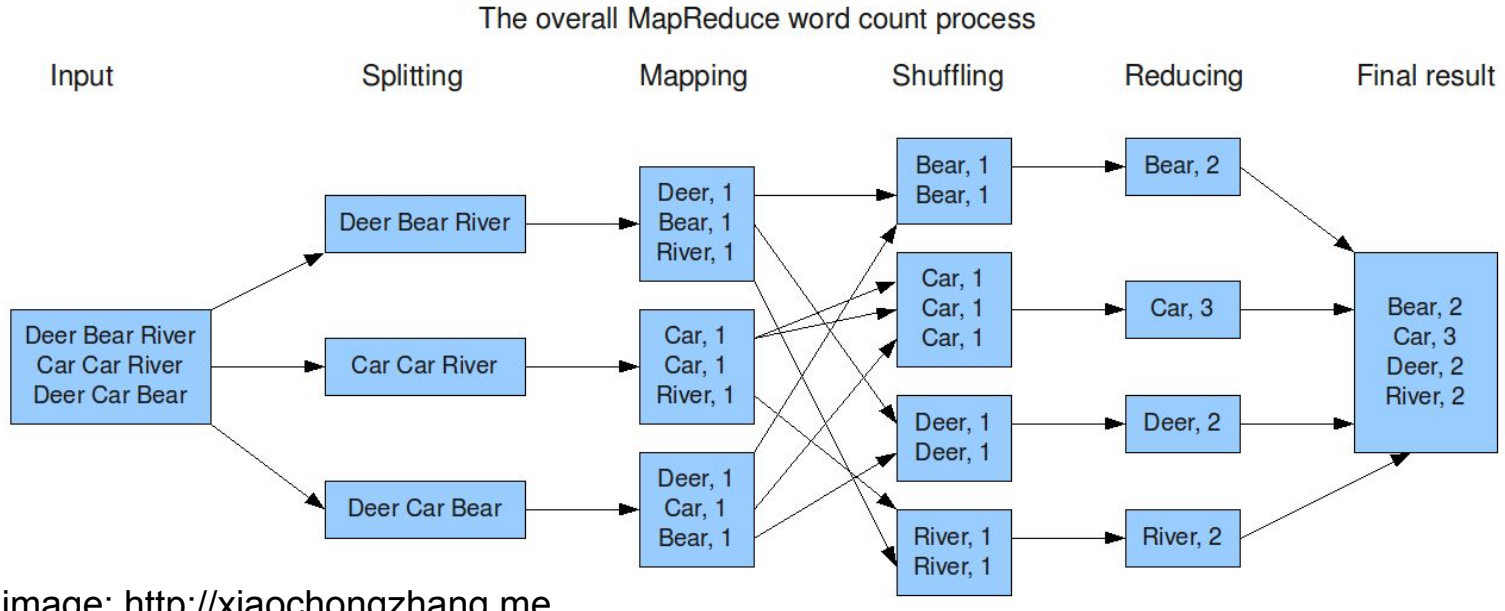
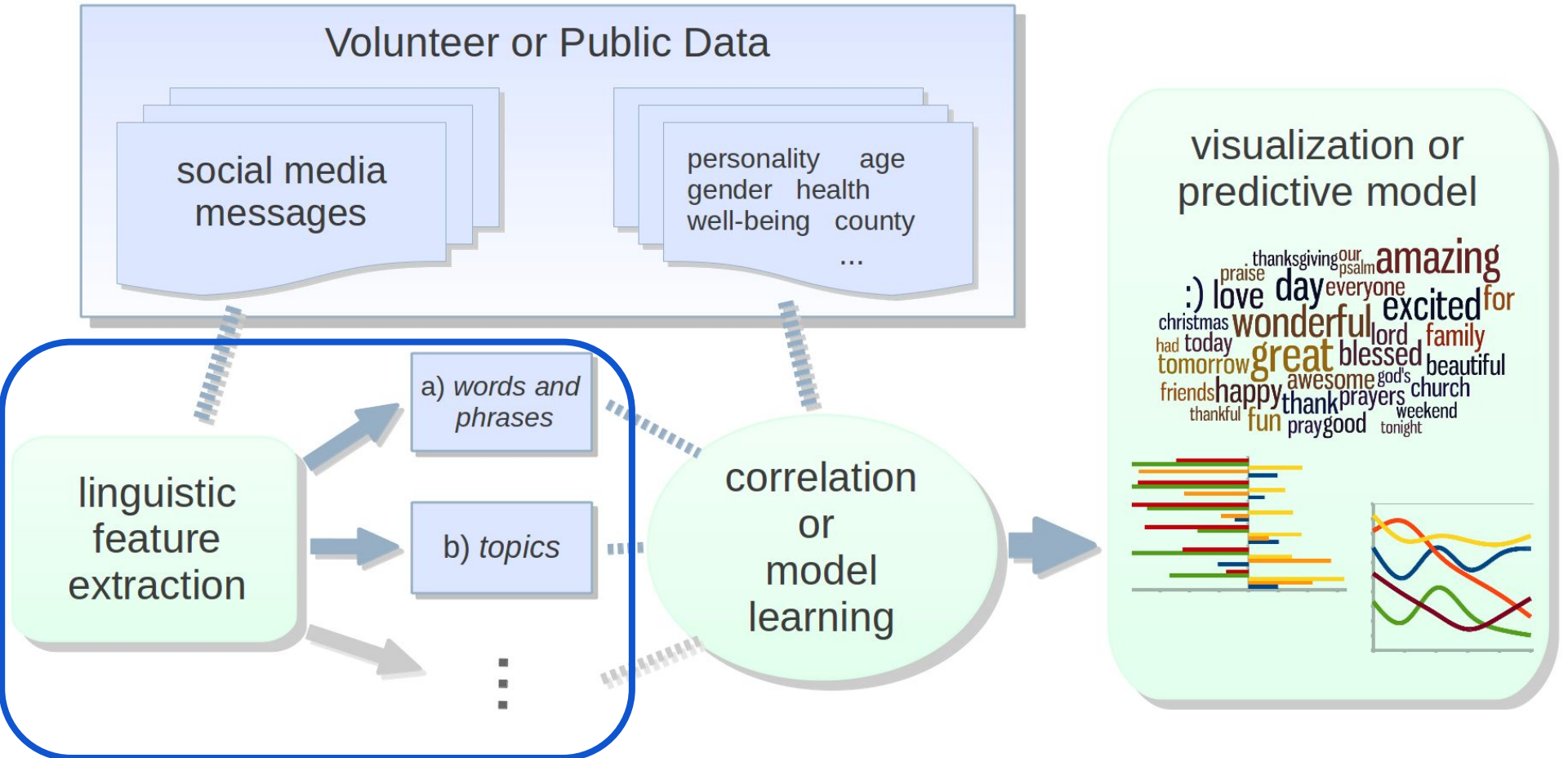
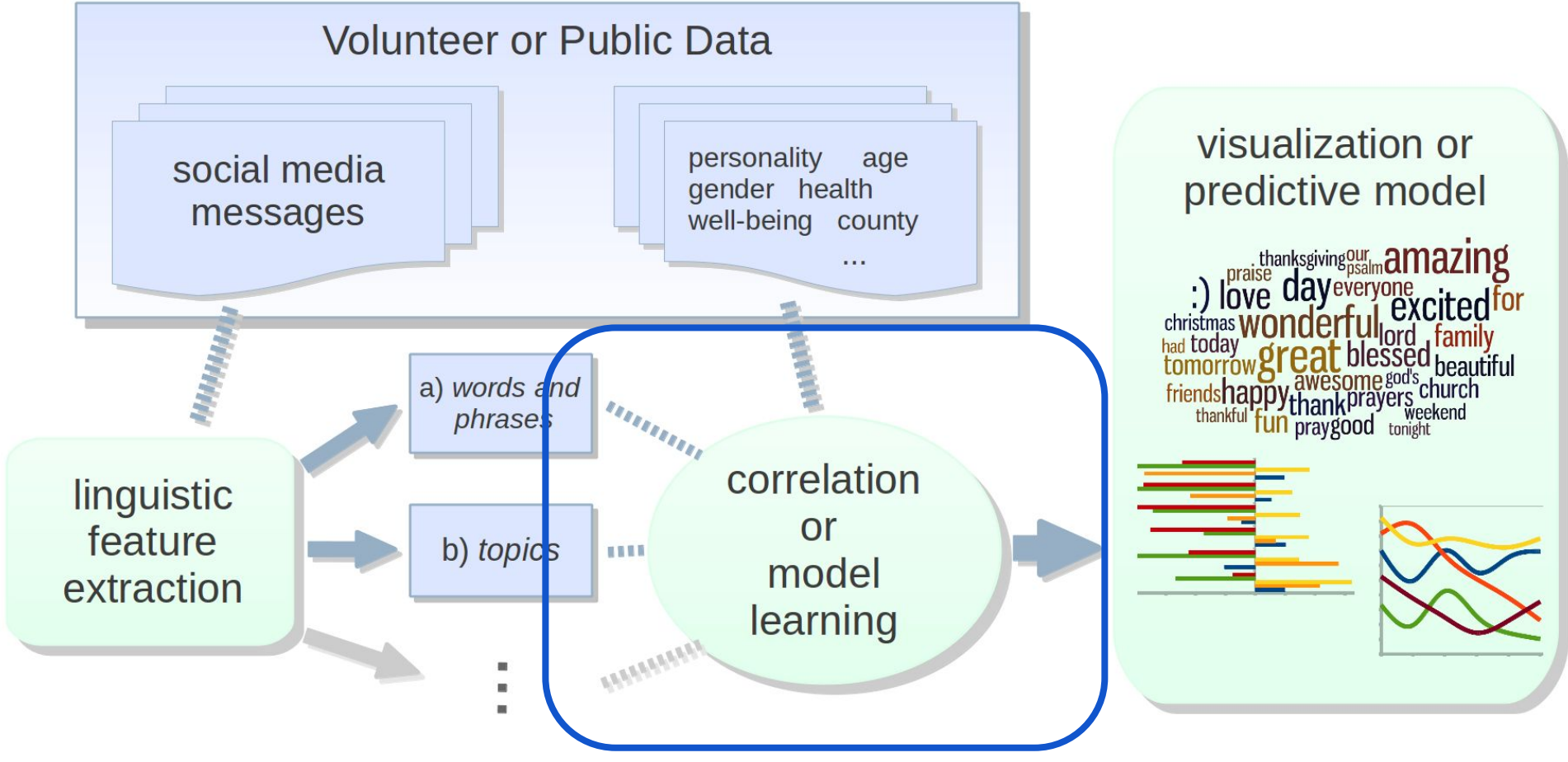
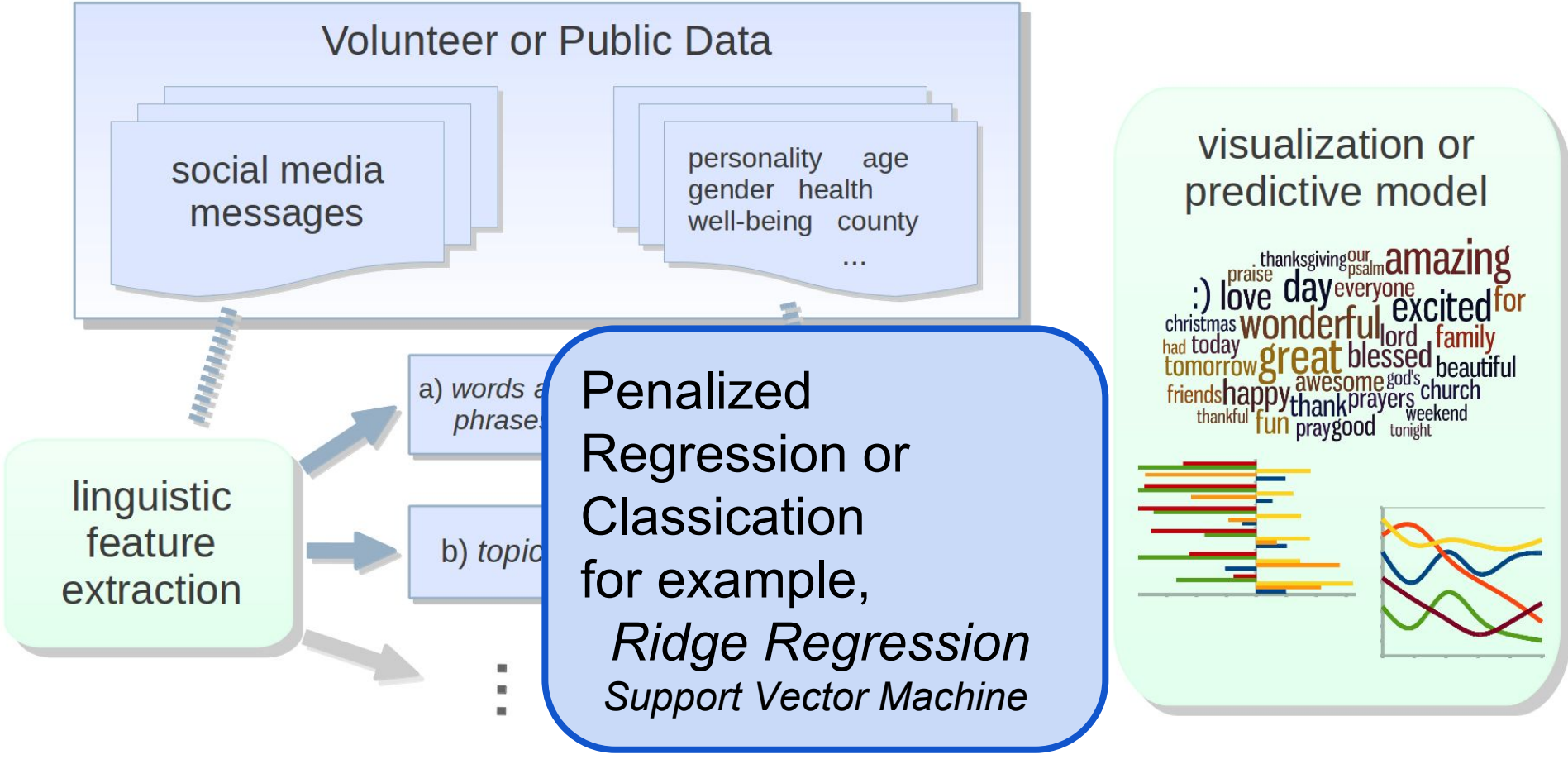


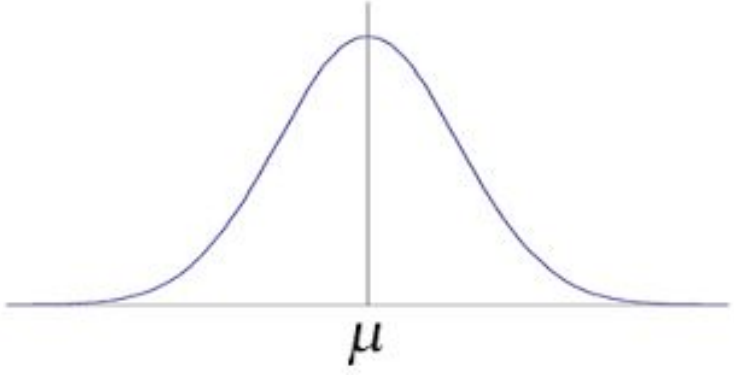
image: http://xiaochongzhang.me


Volunteer or Public Data
social media messages
personality age gender health well-being county ...
linguistic feature extraction
a) words and phrases
b) topics
correlation or model learning
visualization or predictive model
thanksgiving our psalm amazing praise :) love day everyone excited for christmas wonderful lord family had today great blessed beautiful tomorrow awesome god's friends happy prayers church thankful thank fun pray good weekend tonight


Volunteer or Public Data
social media messages
personality age gender health well-being county ...
linguistic feature extraction
a) words and phrases
b) topics
correlation or model learning
visualization or predictive model
thanksgiving our psalm amazing
praise
:) love day everyone excited for
christmas wonderful lord family
had today great blessed beautiful
tomorrow
friends happy awesome god's church
thankful fun thank prayers weekend
pray good tonight

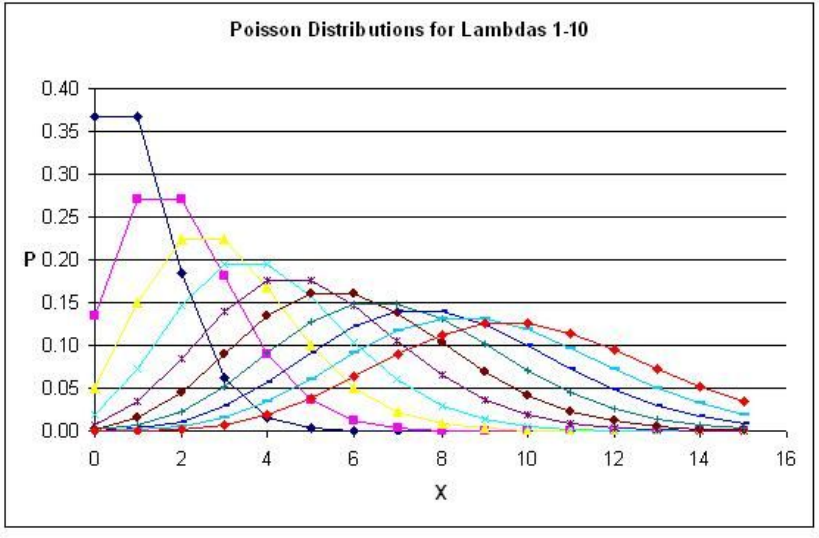

Volunteer or Public Data
social media messages
personality age gender health well-being county ...
linguistic feature extraction
a) words a phrase
b) topic
Penalized Regression or Classication for example,
Ridge Regression
Support Vector Machine
visualization or predictive model
thanksgiving our psalm amazing
praise
:) love day everyone excited for
christmas wonderful lord family
had today great blessed beautiful
tomorrow
friends happy awesome god's church
thankful fun thank prayers weekend
pray good tonight

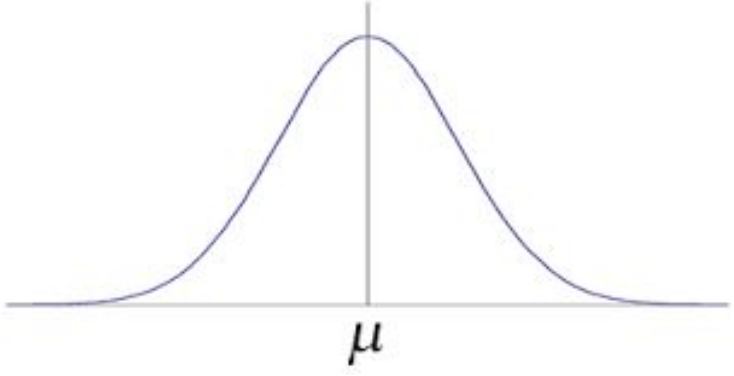
Challenge



Topics
~2k

Challenge





NGrams
~20k after frequency filter

Topics
~2k

Challenge

NGrams
~20k after frequency filter

Topics
~2k

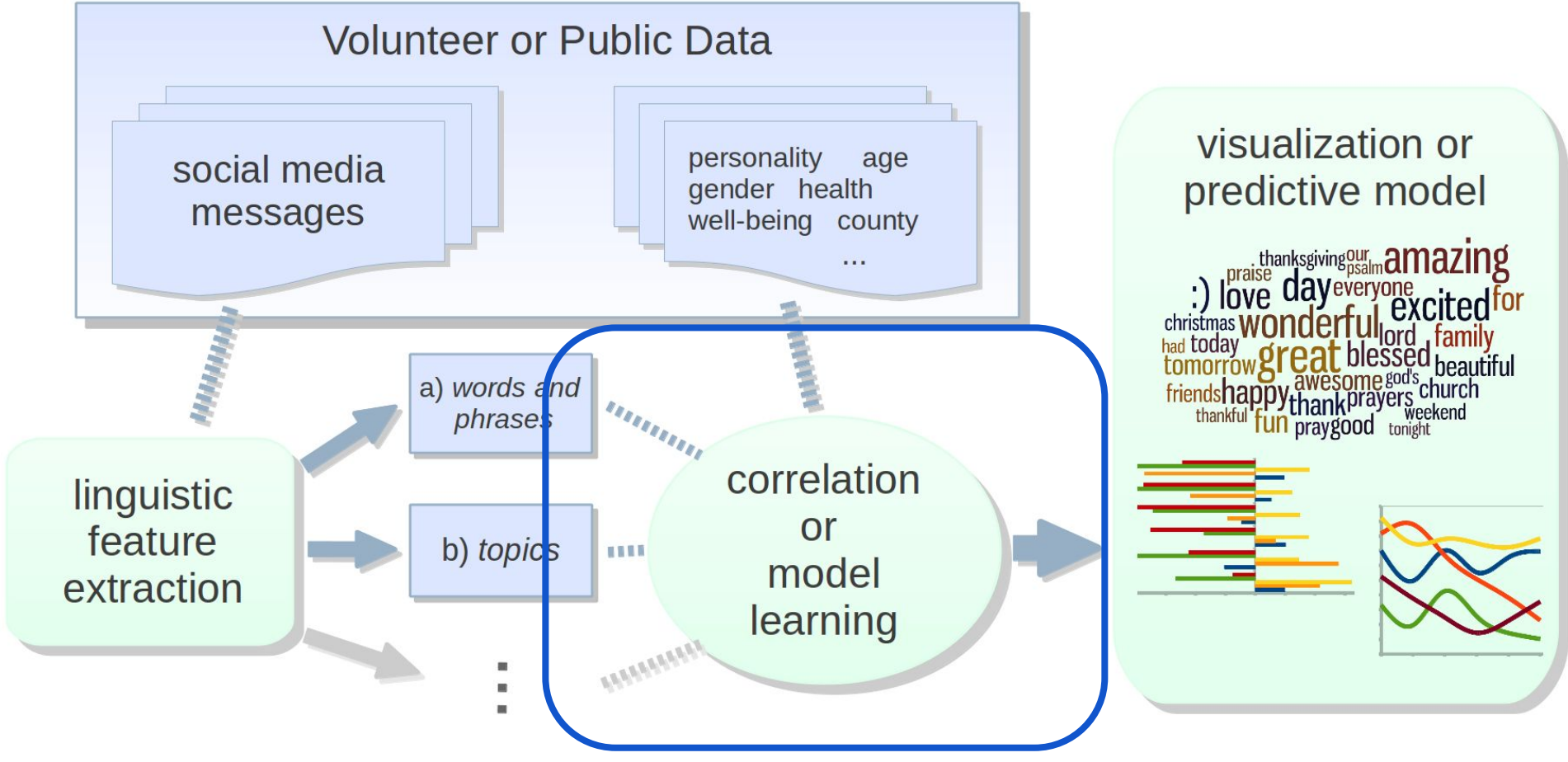
Challenge: Even Worse

Language:
High-dimensional,
sparse, and noisy.

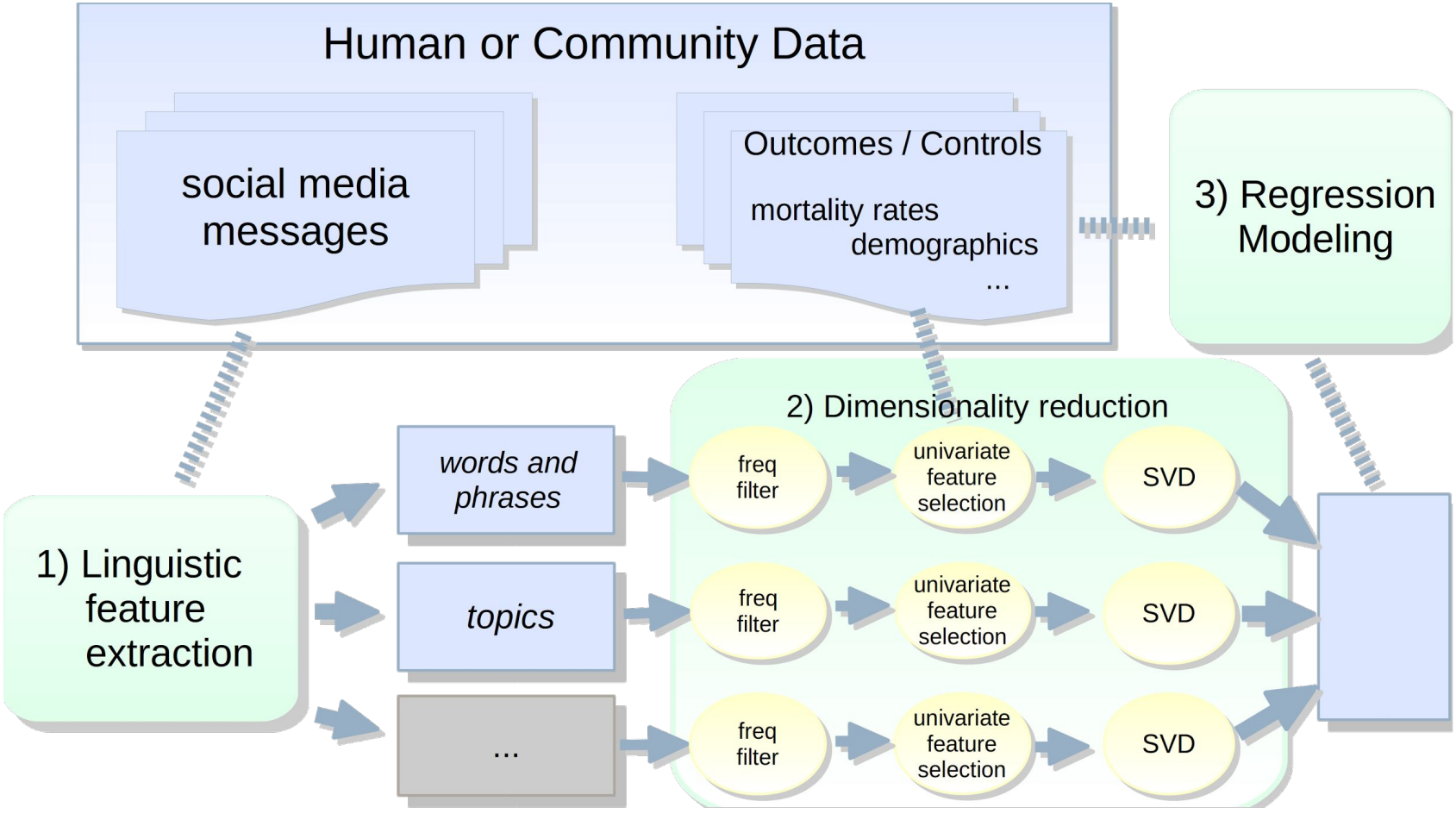
Socio-economics:
few and well estimated

VS

PART II: How?


Volunteer or Public Data
social media messages
personality age gender health well-being county ...
linguistic feature extraction
a) words and phrases
b) topics
correlation or model learning
visualization or predictive model
thanksgiving our psalm amazing praise :) love day everyone excited for christmas wonderful lord family had today great blessed beautiful tomorrow awesome god's church friends happy thank prayers weekend thankful fun pray good tonight

Prediction


Human or Community Data
social media messages
Outcomes / Controls
mortality rates
demographics
...
3) Regression Modeling
2) Dimensionality reduction
1) Linguistic feature extraction
words and phrases
topics
...
freq filter
univariate feature selection
SVD
freq filter
univariate feature selection
SVD
freq filter
univariate feature selection
SVD

Marginal gain from Socioeconomics:

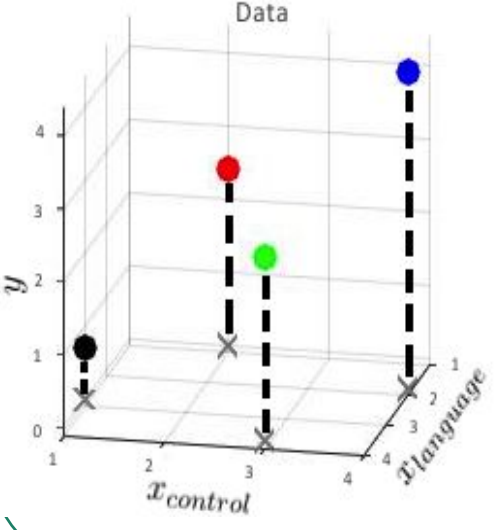
	Foreclosure	Increased-price
language	0.38	0.48
combined	0.40	0.49

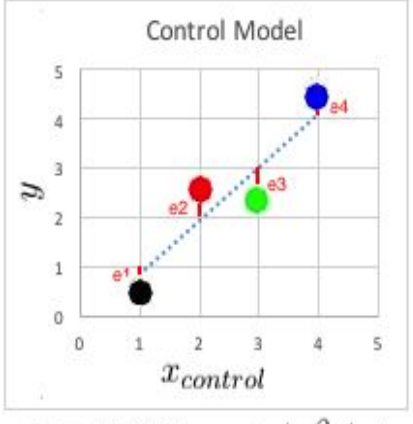
Out of sample Pearson r

Residualized Control Model

- Effectively use both low dimensional control features and high-dimensional, noisy language features:
  - Train a control model using the control values
  - Calculate the residual error and consider it as the new label
  - Train a language model over the new labels

Residualized Control Model

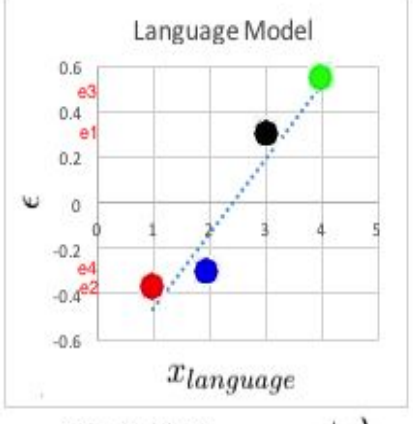




$y = \alpha \times x_{control} + \beta + \epsilon$

$\alpha = 1, \beta = 0$

error $\longrightarrow$



$\epsilon \simeq \gamma \times x_{language} + \lambda$

$\gamma = 0.3, \lambda = -0.7$

$\Rightarrow$

$$y \simeq \alpha \times x_{control} + \gamma \times x_{language} + \beta + \lambda$$

$$\beta + \lambda = -0.7$$

Zamani M, Schwartz HA. Using Twitter Language to Predict the Real Estate Market. EACL 2017. 2017 Apr 3:28.

Residualized Control vs. Combined Model

$$Y = \alpha x_1 + \beta x_2 + \epsilon$$

Both learn same linear model.

- Different learning algorithm per variable
- As if different penalization.

Marginal gain from Socioeconomics:

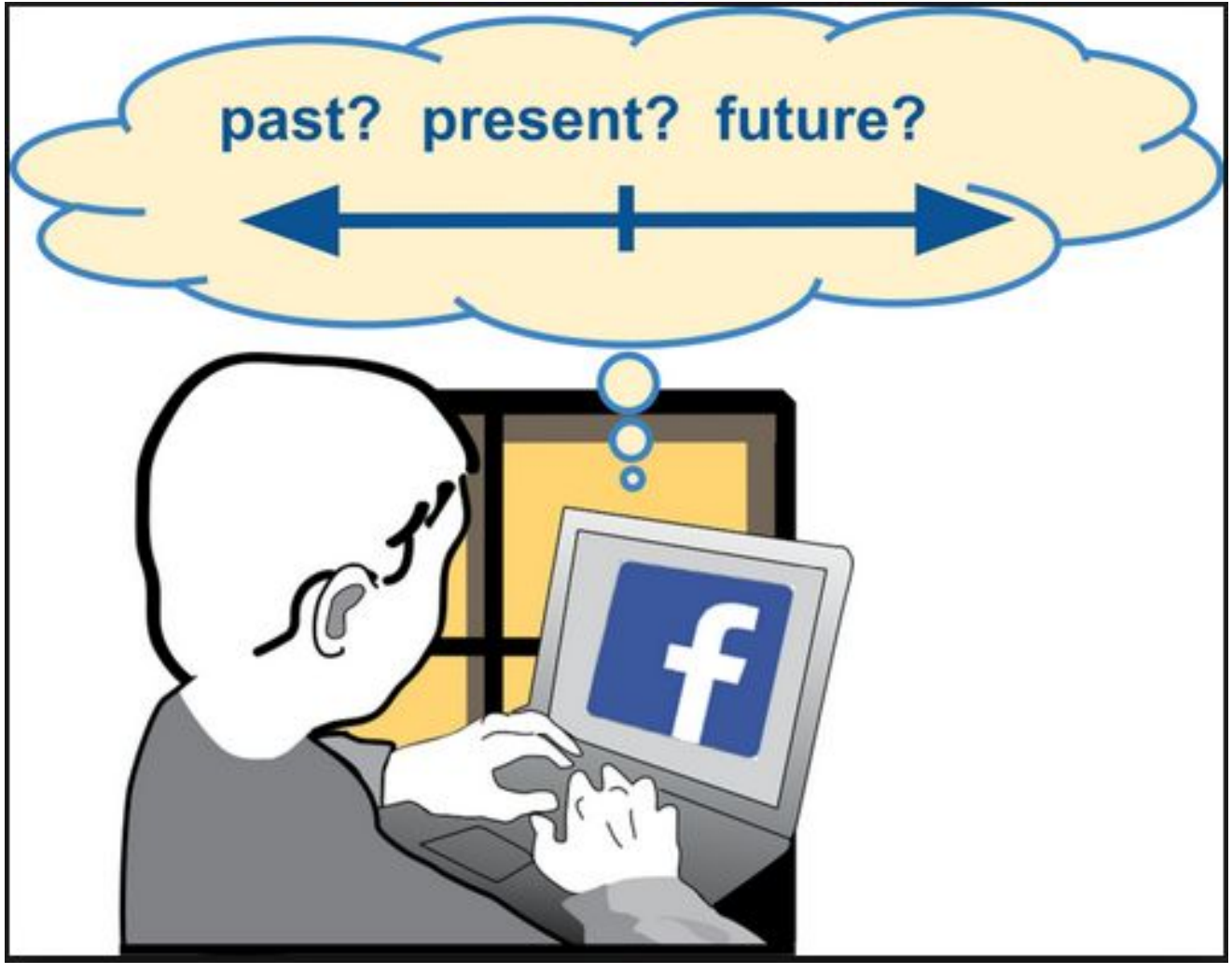
	Foreclosure	Increased-price
language	0.38	0.48
combined	0.40	0.49

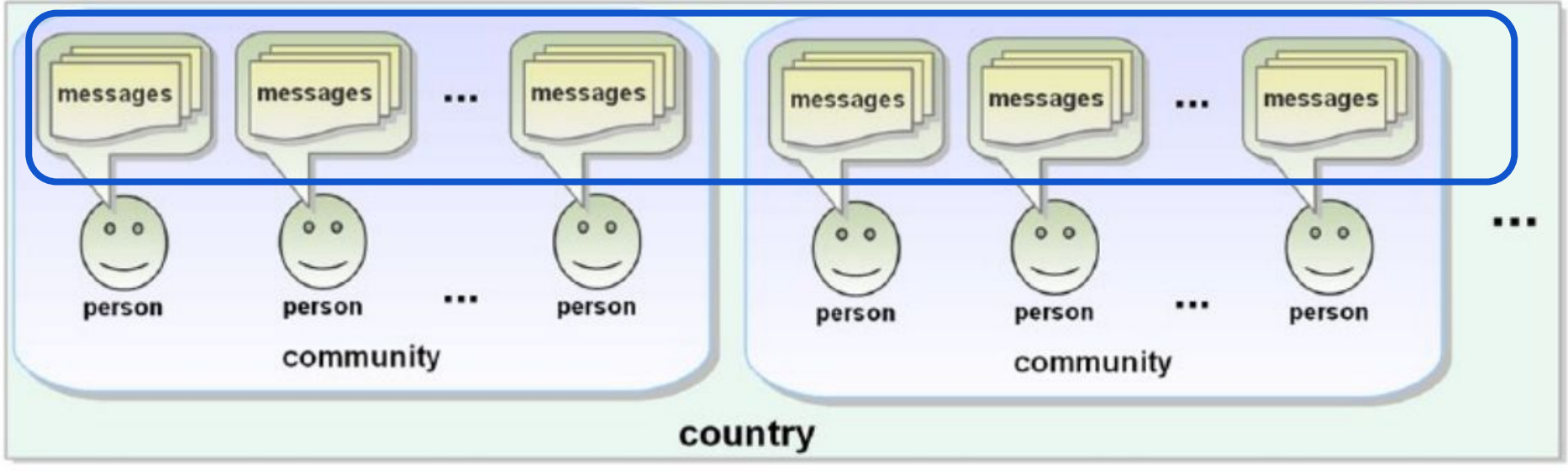
Out of sample Pearson r

Marginal gain from Residulaized Control

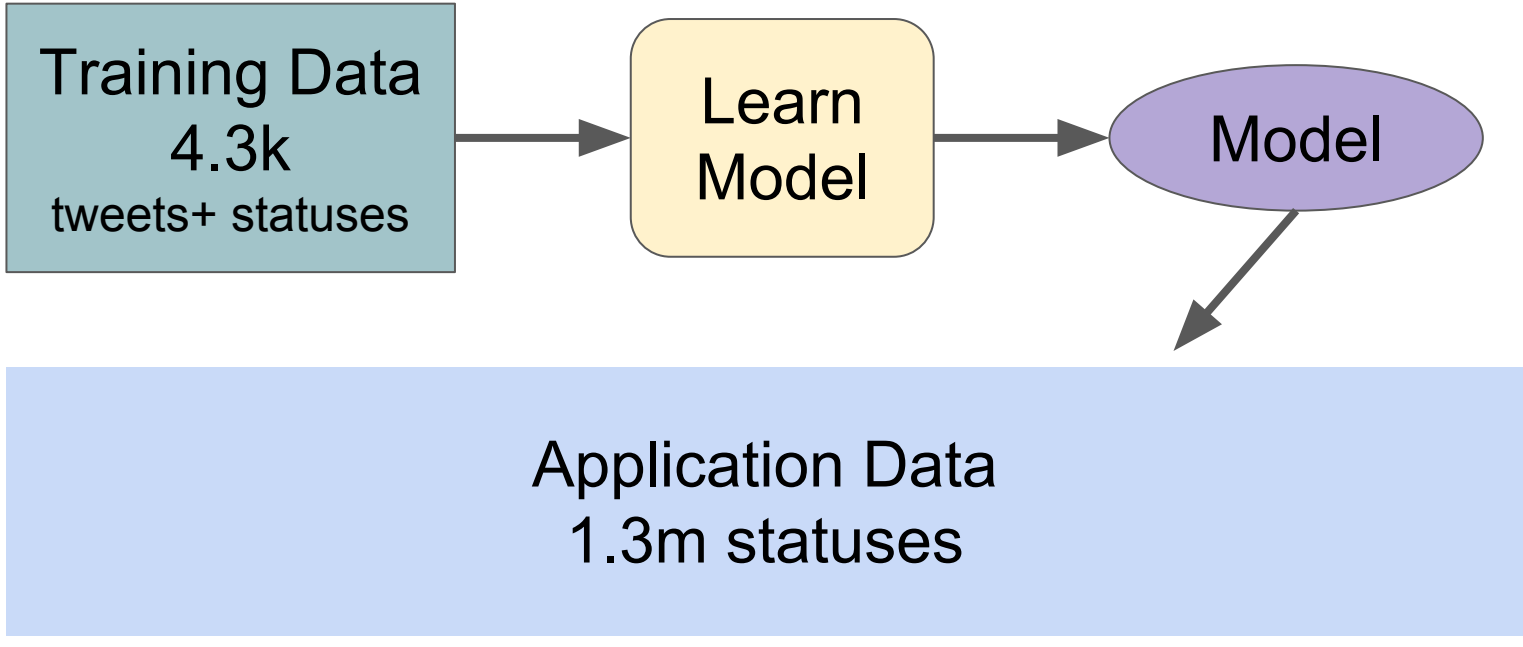
	Foreclosure	Increased-price
language	0.38	0.48
combined	0.40	0.49
residualized control	**0.42**	**0.59**

Out of sample Pearson r


past? present? future?

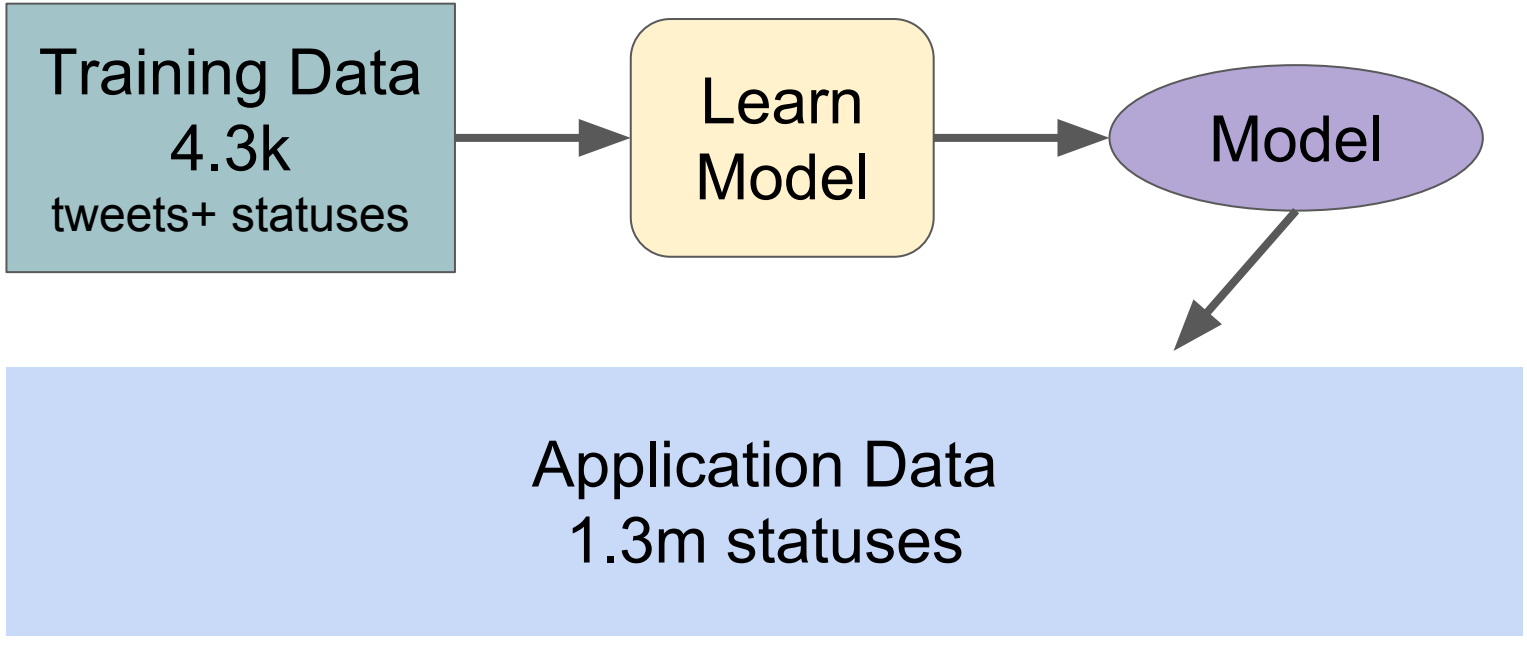

messages
messages
...
messages
messages
messages
...
messages
person
person
...
person
person
person
...
person
community
community
country

Building a model



Building a model

message	R1	R2	R3	m	class
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pancake day tomorrow pancake day tomorrow xxxxx	.50	.50	1	.67	future



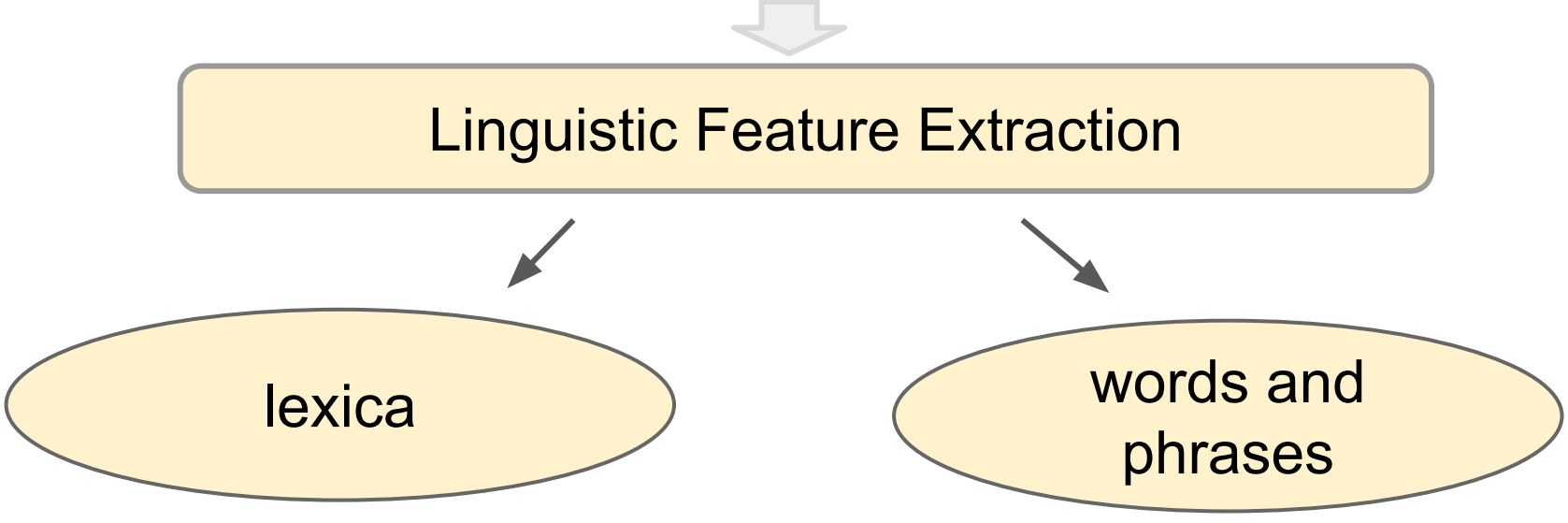
Building a model

message	R1	R2	R3	m	class
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pancake day tomorrow pancake day tomorrow xxxxx	.50	.50	1	.67	future

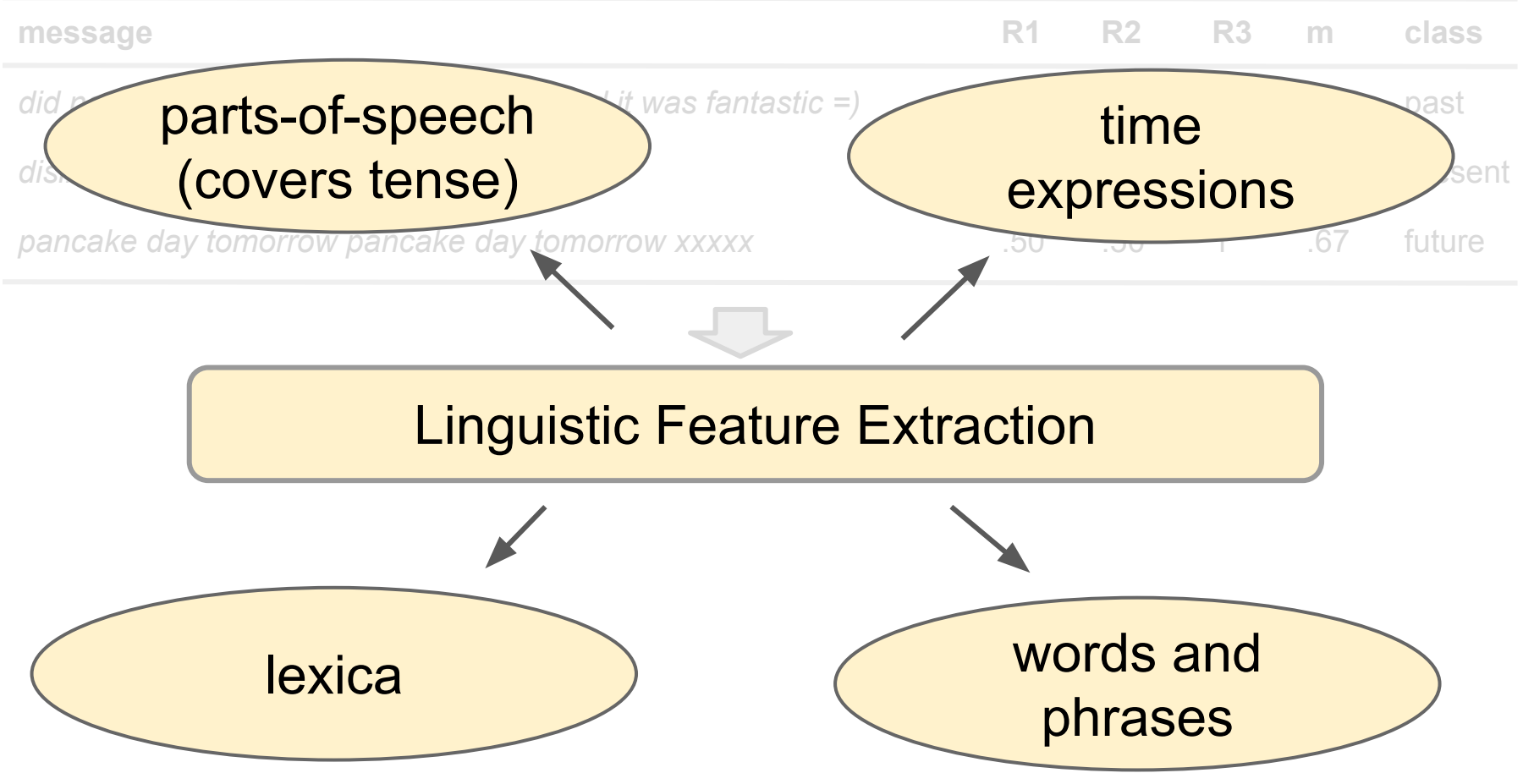
Linguistic Feature Extraction

Building a model

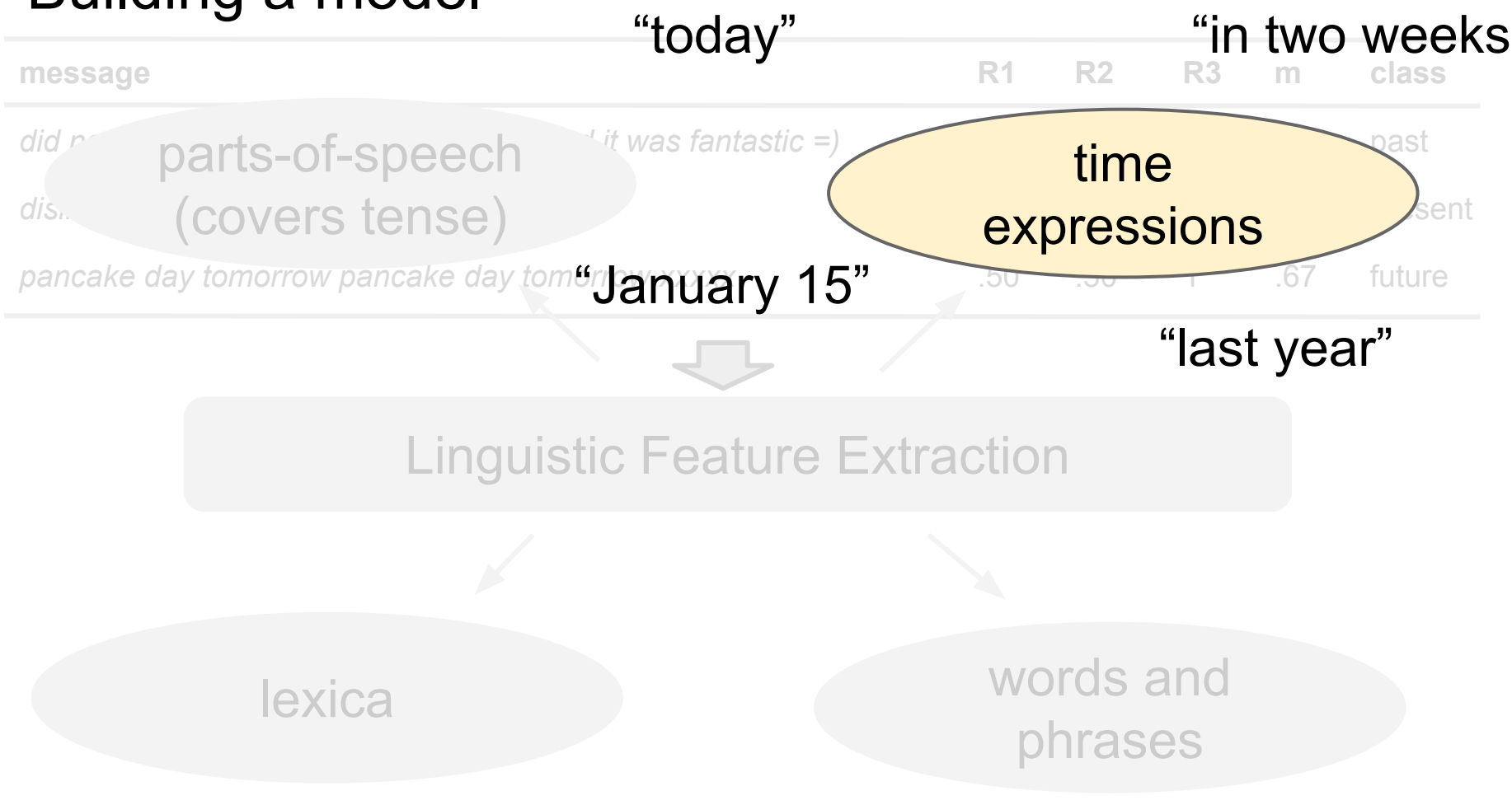
message	R1	R2	R3	m	class
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pancake day tomorrow pancake day tomorrow xxxxx	.50	.50	1	.67	future



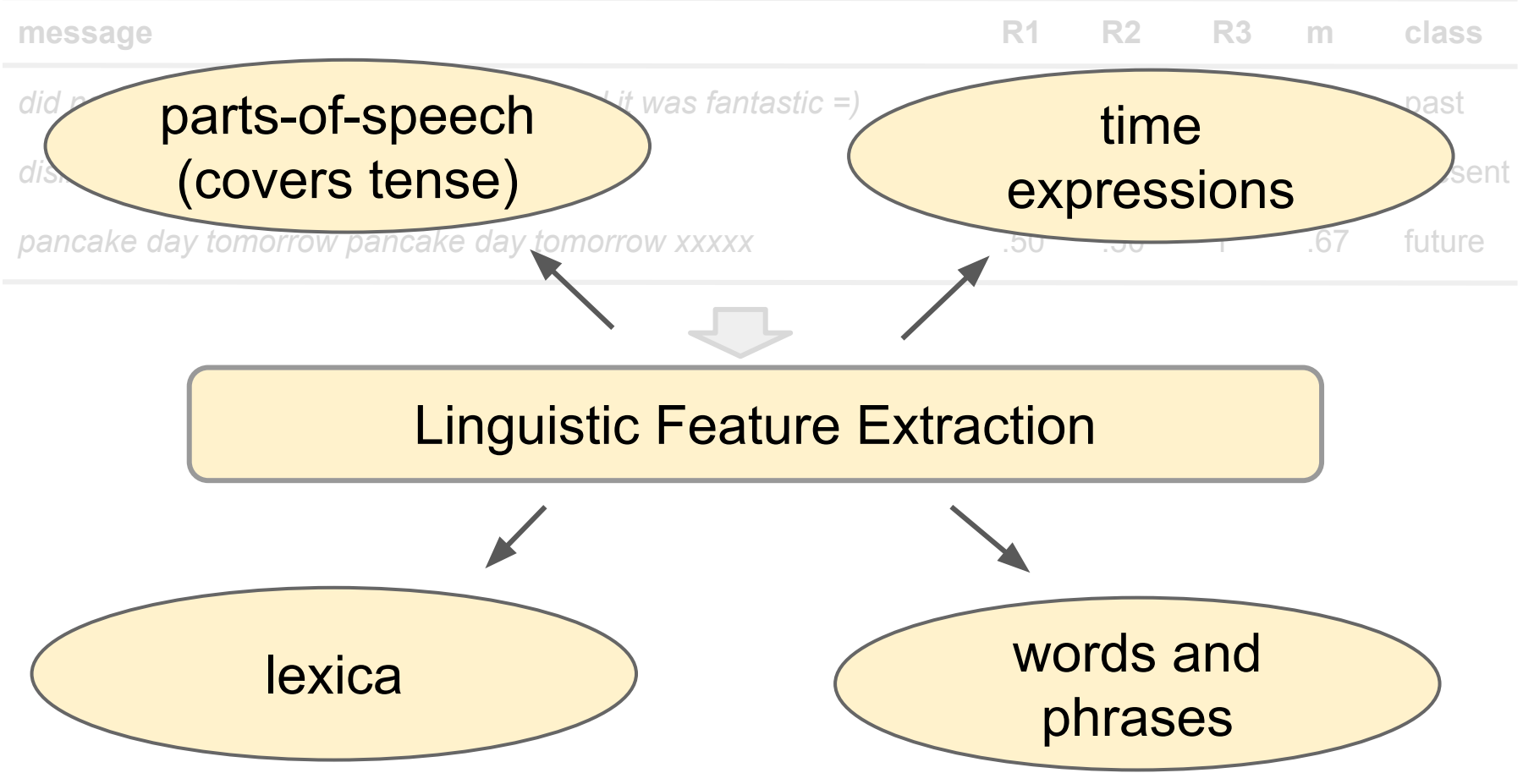
Building a model



Building a model



Building a model



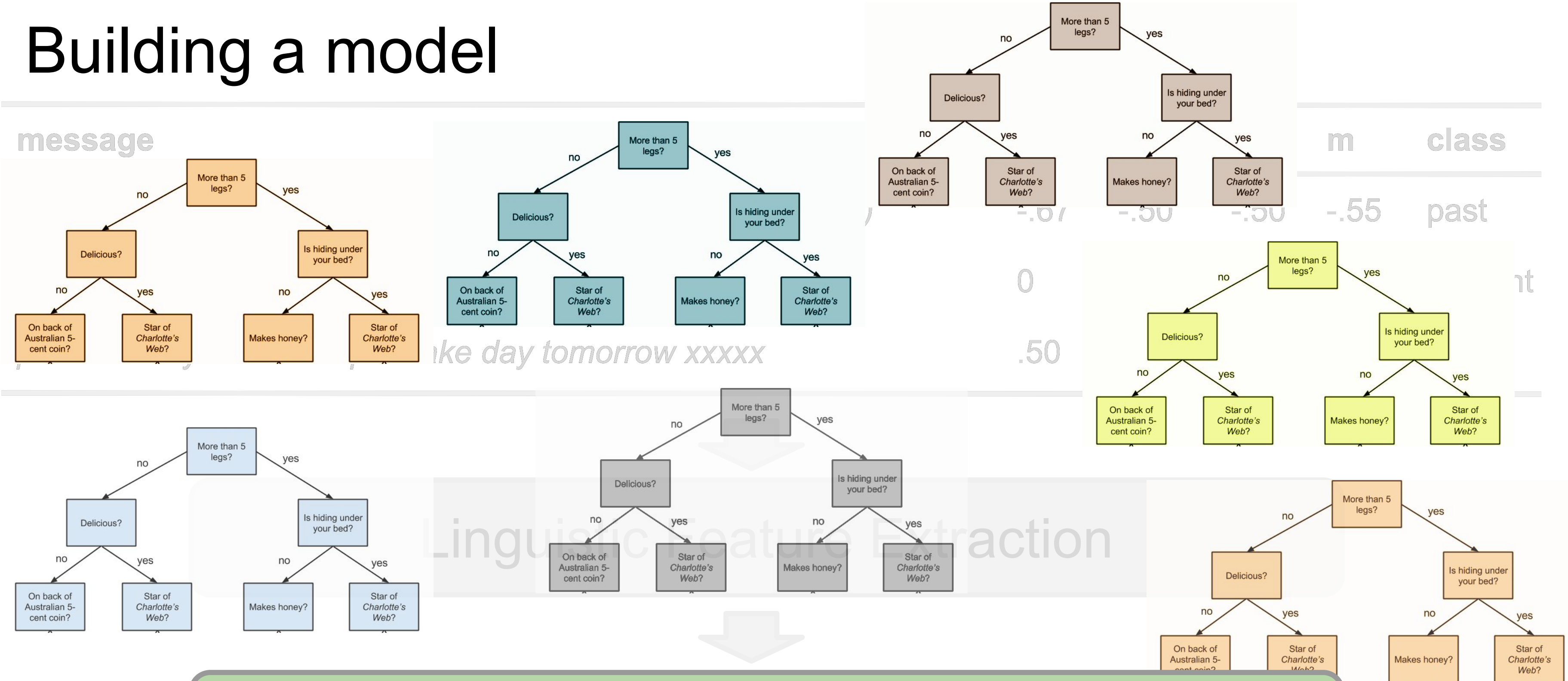
Building a model

message	R1	R2	R3	m	class
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pancake day tomorrow pancake day tomorrow xxxxx	.50	.50	1	.67	future

Linguistic Feature Extraction

Learn Message-Level Model

Building a model



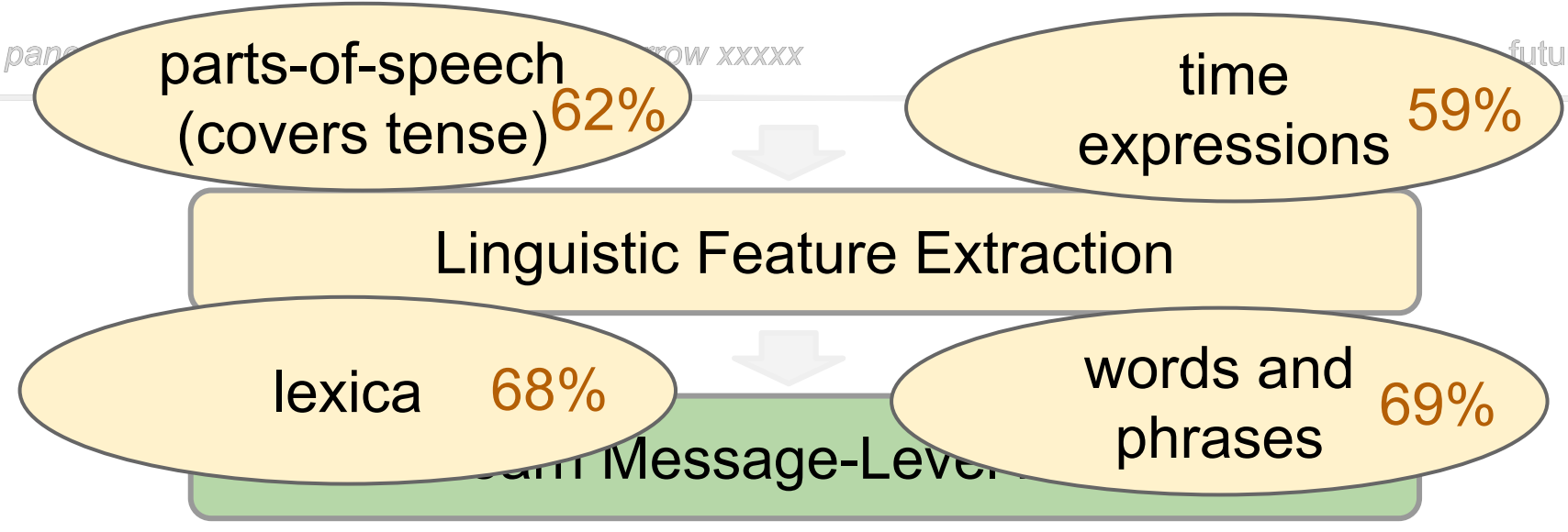
Learn Message-Level Model

Accuracy over a held-out set: 72%;   baseline: 53%

Schwartz, H. A., Park, G., Sap, M., ..., & Ungar, L. (2015). Extracting Human Temporal Orientation from Facebook Language. *NAACL-2015: Conference of the North American Chapter of the Association for Computational Linguistics*

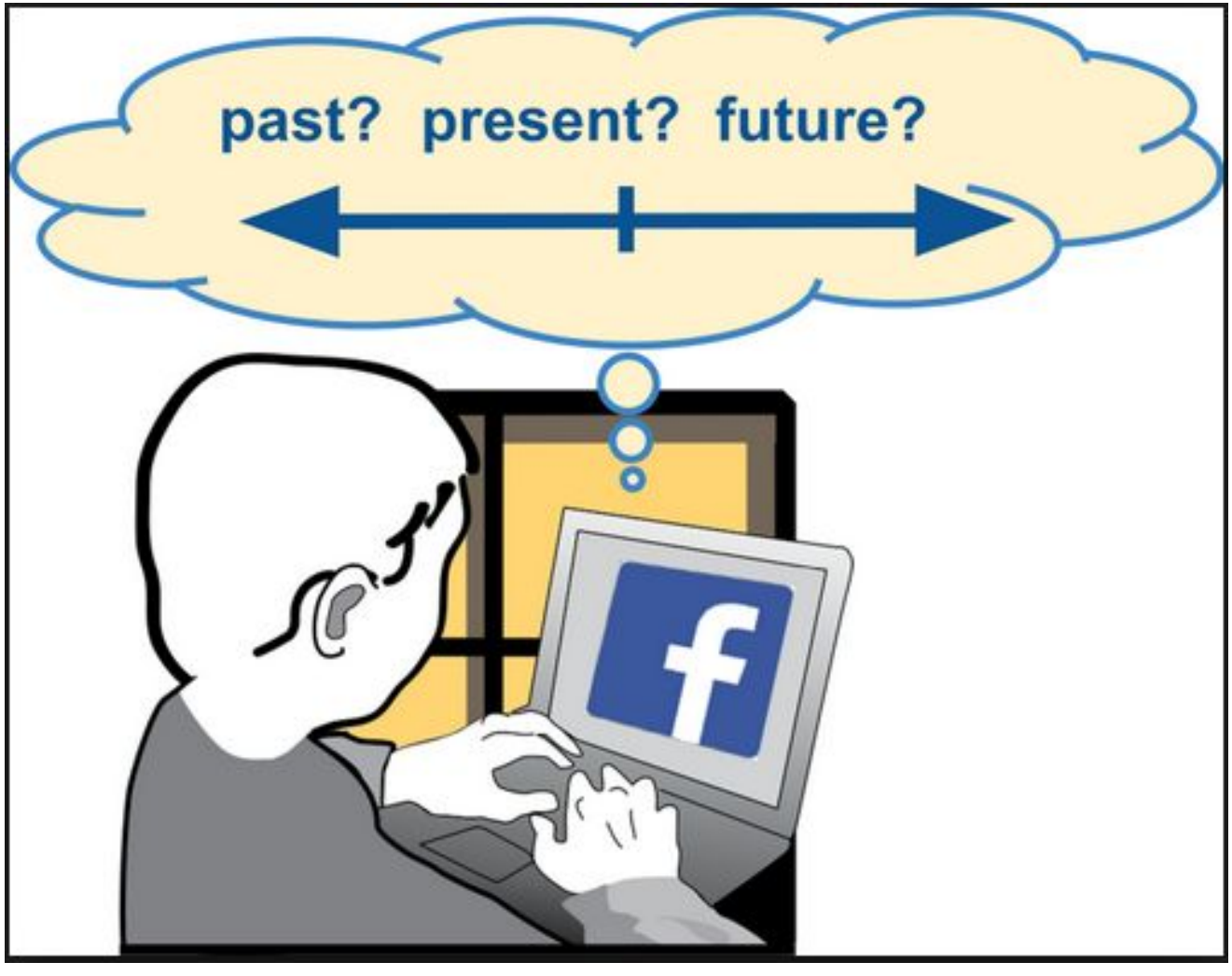
Building a model

message	R1	R2	R3	m	class
did nothing this morning but watch TV and it was fantastic =)	-.67	-.50	-.50	-.55	past
dislikes being sick.... and misses her bf	0	0	0	0	present
pan… …row xxxxx					future



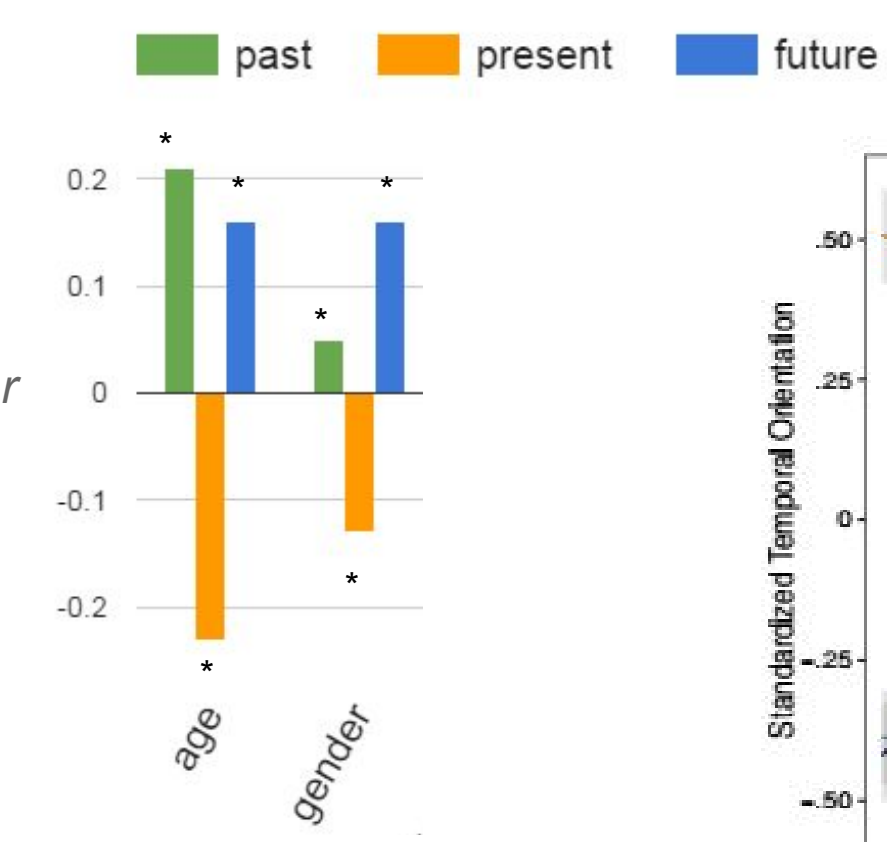
Accuracy over a held-out set: 72%; baseline: 53%

Schwartz, H. A., Park, G., Sap, M., ..., & Ungar, L. (2015). Extracting Human Temporal Orientation from Facebook Language. *NAACL-2015: Conference of the North American Chapter of the Association for Computational Linguistics*

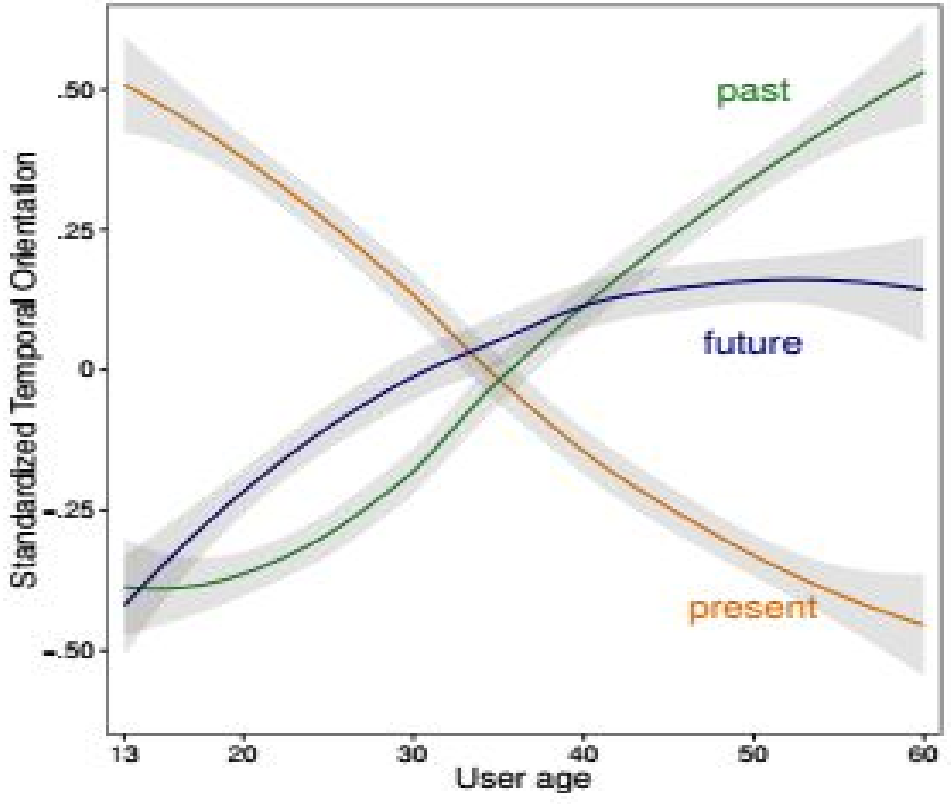


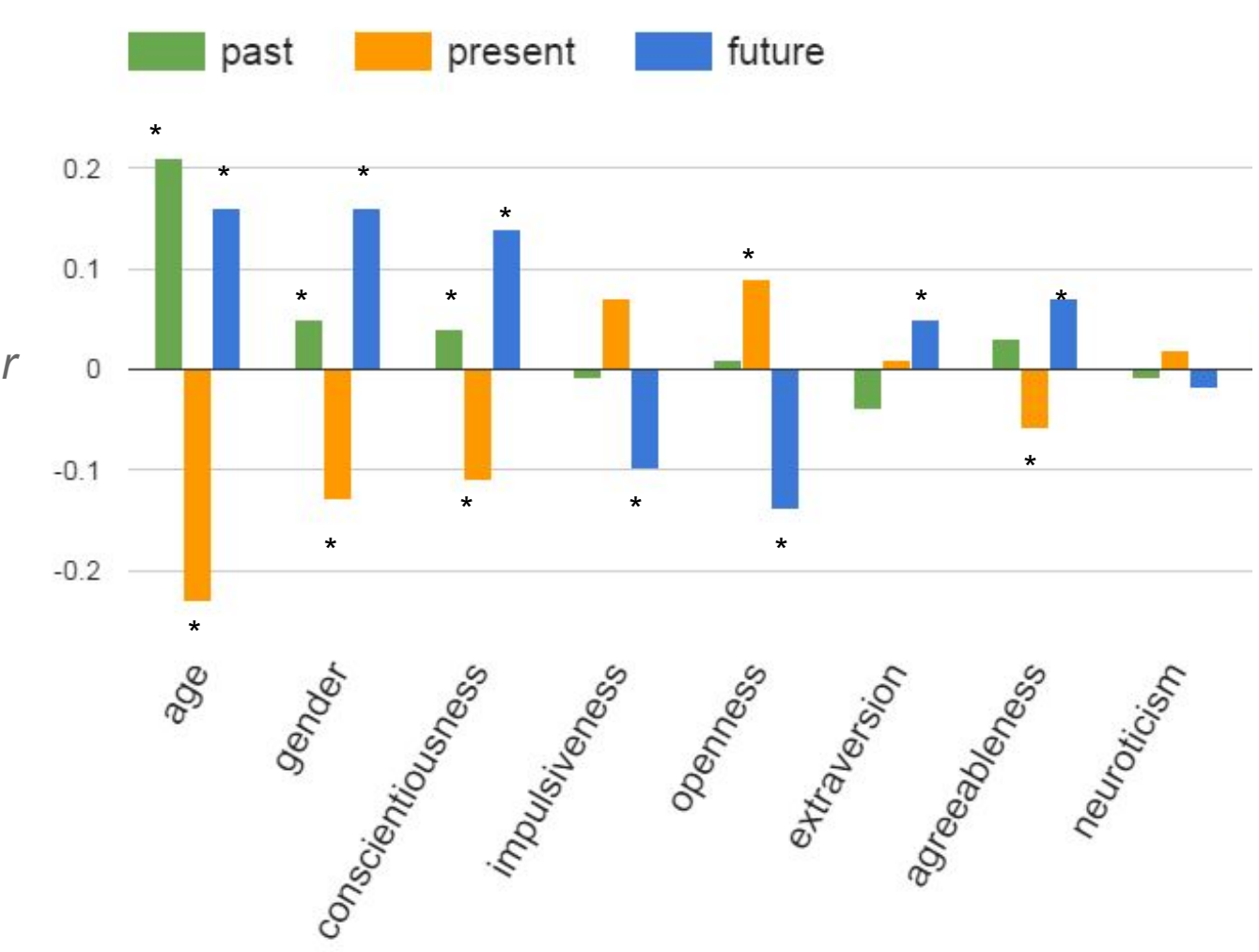
1.3m statuses
from
4,833 Participants

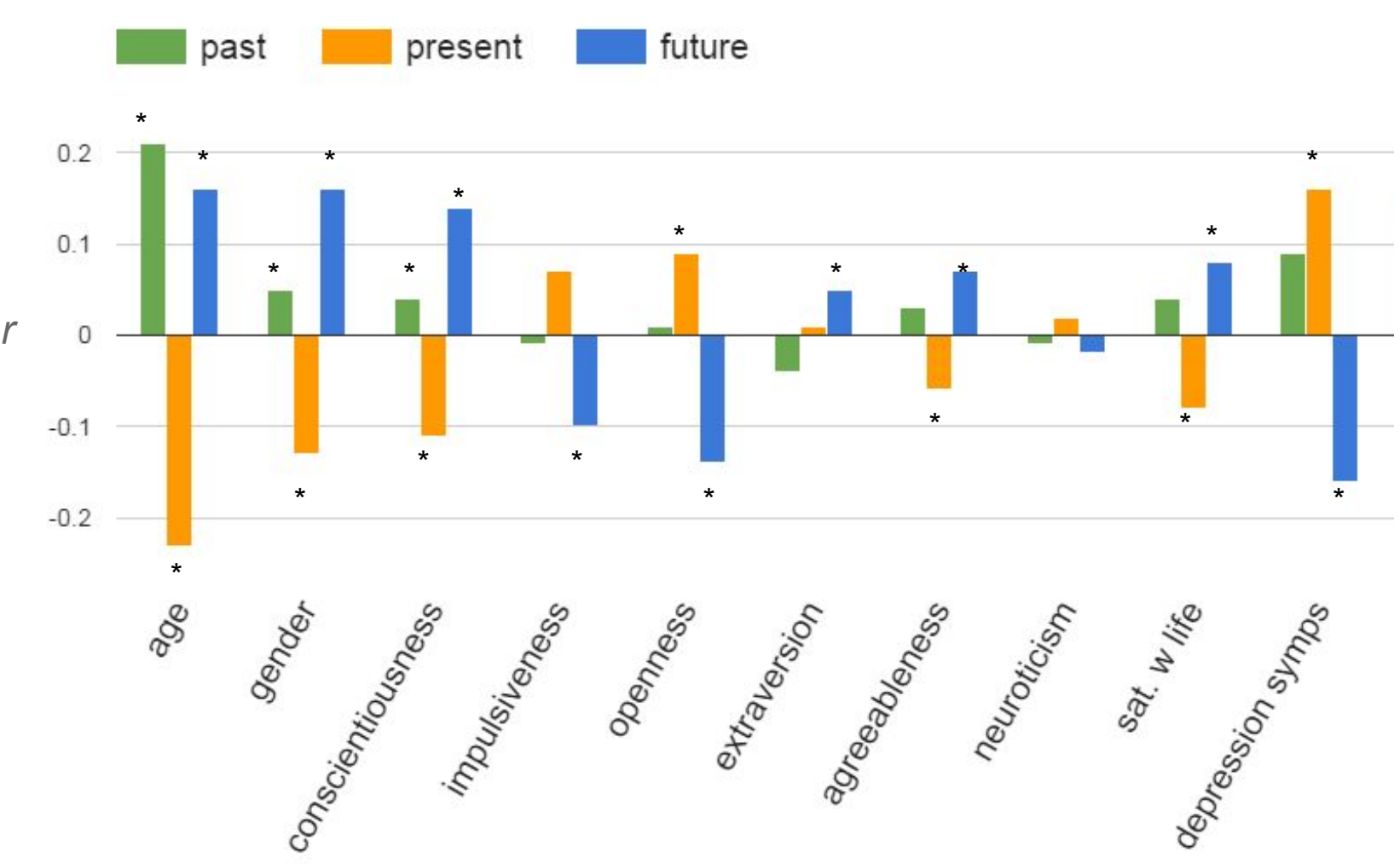
(3,240 AG Stratified)

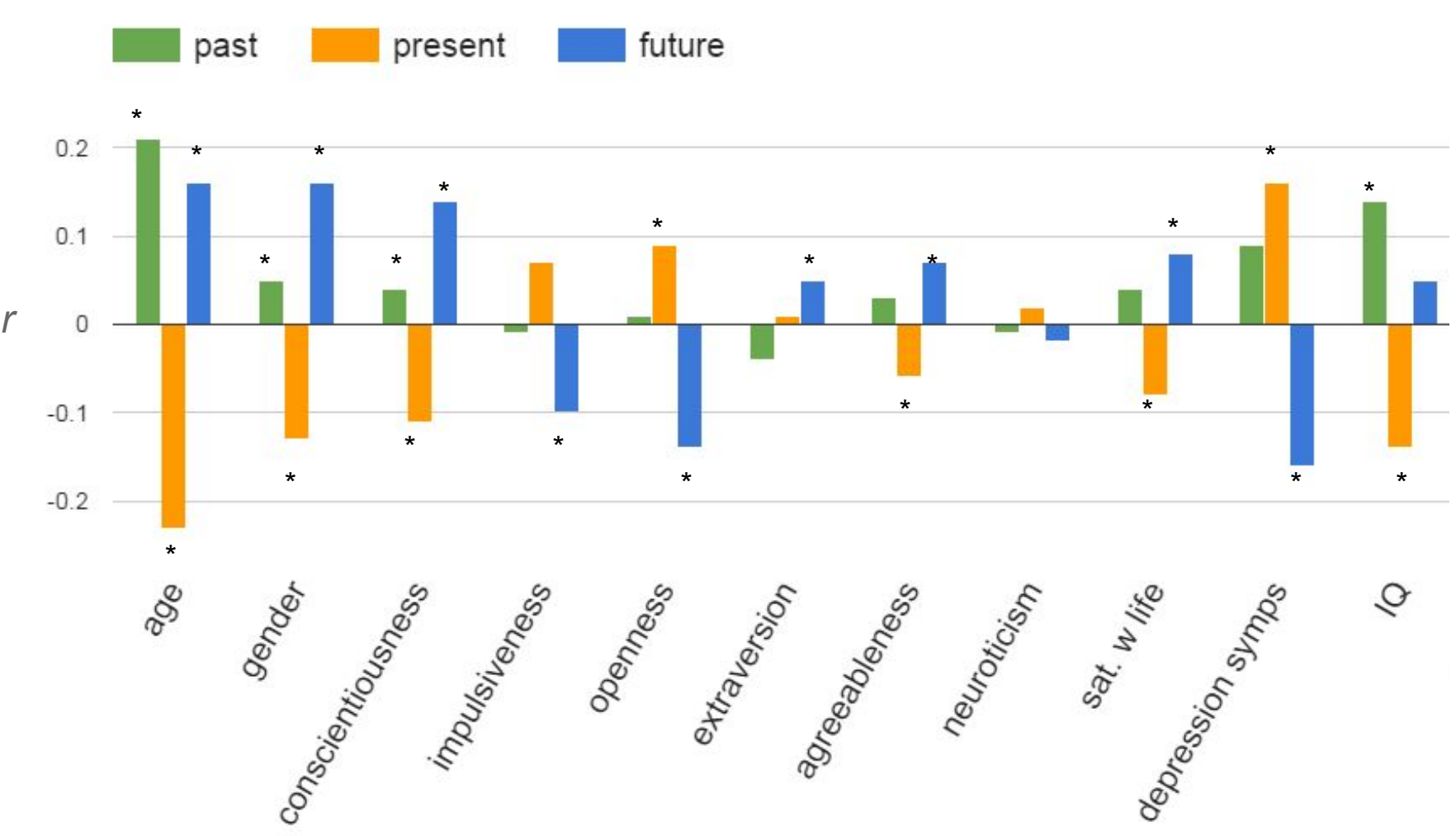


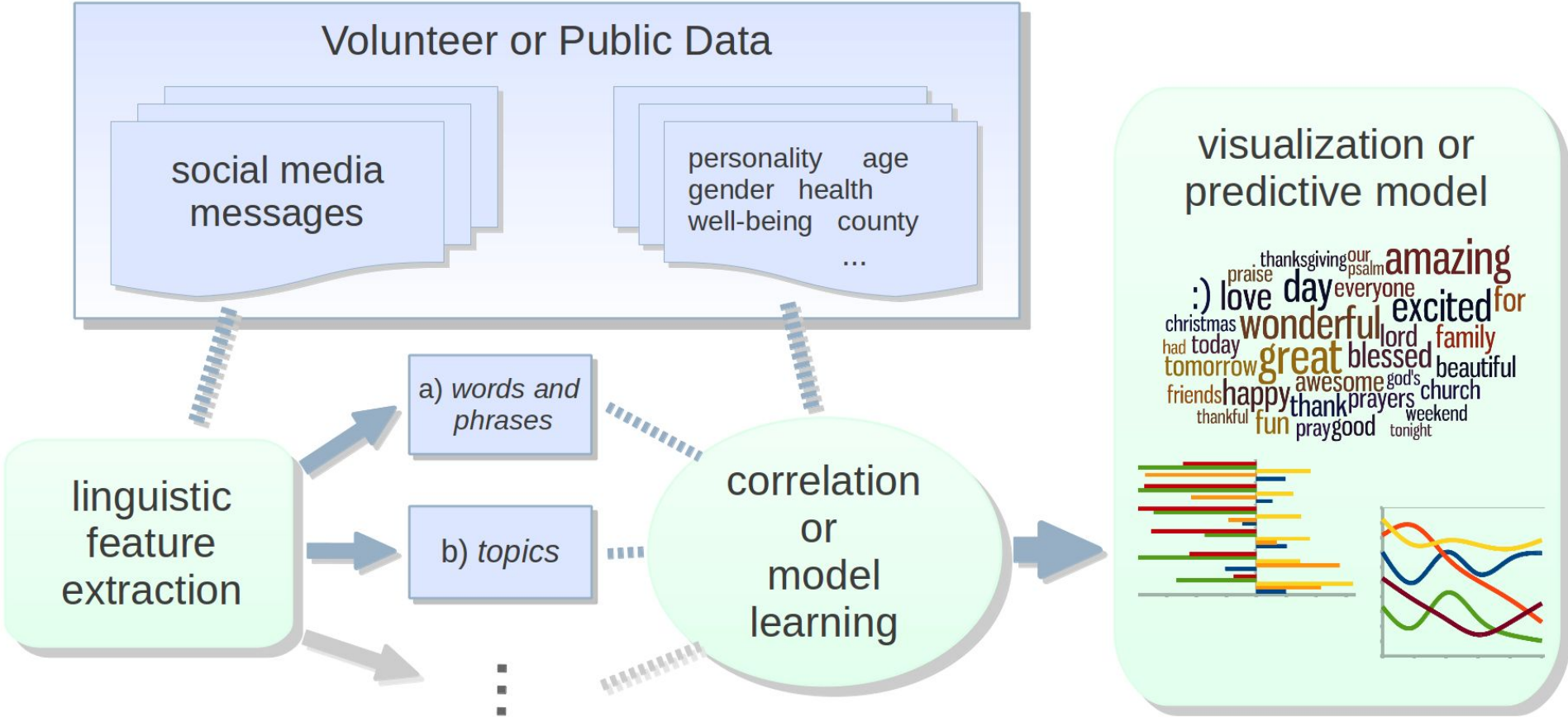
1: female
0: male




past
present
future
0.2
0.1
0
-0.1
-0.2
r
age
gender
conscientiousness
impulsiveness
openness
extraversion
agreeableness
neuroticism


past
present
future
r
0.2
0.1
0
-0.1
-0.2
age
gender
conscientiousness
impulsiveness
openness
extraversion
agreeableness
neuroticism
sat. w life
depression symps


past
present
future
r
0.2
0.1
0
-0.1
-0.2
age
gender
conscientiousness
impulsiveness
openness
extraversion
agreeableness
neuroticism
sat. w life
depression symps
IQ


Volunteer or Public Data
social media messages
personality age
gender health
well-being county
...
linguistic feature extraction
a) words and phrases
b) topics
correlation or model learning
visualization or predictive model
thanksgiving
our
psalm
amazing
praise
:) love
day
everyone
excited
for
christmas
wonderful
lord
family
had
today
great
blessed
beautiful
tomorrow
awesome
god's
friends
happy
prayers
church
thankful
fun
thank
pray
good
weekend
tonight

Resources

Resources


```
#feature extraction:
dlatkInterface.py
 -d <data> -t <corpus>
 -c <group_data_column>
 --add_ngrams -n 1 2 3
 --combine_feat_tables
  1to3gram
```

Resources


```
#feature extraction:
dlatkInterface.py
 -d <data> -t <corpus>
 -c <group_data_column>
 --add_ngrams -n 1 2 3
 --combine_feat_tables
  1to3gram
```

```
#run differential language analysis (linear regression)
dlatkInterface.py -d <data> -t <corus> -c user_id
 -f 'feat$cat_LIWC2007$msgs_xxx$user_id$16to16'  --outcome_table
 blog_outcomes  --group_freq_thresh 1000 --outcomes extraversion
 --controls age gender  --output_name xxx_output --make_wordcloud
```

extraversion -- *sociable, assertive, active, energetic, talkative, outgoing*

Penn | World Well-Being Project | wwbp.org

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*

Resources


```
#feature extraction:
dlatkInterface.py
 -d <data> -t <corpus>
 -c <group_data_column>
 --add_ngrams -n 1 2 3
 --combine_feat_tables
  1to3gram
```

```
#run differential language analysis (linear regression)
dlatkInterface.py -d <data> -t <corus> -c user_id
 -f 'feat$cat_LIWC2007$msgs_xxx$user_id$16to16'  --outcome_table
 blog_outcomes  --group_freq_thresh 1000 --outcomes extraversion
 --controls age gender  --output_name xxx_output --make_wordcloud
 --[logistic_reg|AUC|IDP]
```

Resources


```
#feature extraction:
dlatkInterface.py
 -d <data> -t <corpus>
 -c <group_data_column>
 --add_ngrams -n 1 2 3
 --combine_feat_tables
  1to3gram
```

```
#run differential language analysis (linear regression)
dlatkInterface.py -d <data> -t <corus> -c user_id
 -f 'feat$cat_LIWC2007$msgs_xxx$user_id$16to16'  --outcome_table
 blog_outcomes  --group_freq_thresh 1000 --outcomes extraversion
 --controls age gender  --output_name xxx_output --make_wordcloud
```

Resources


```
#feature extraction:
dlatkInterface.py
 -d <data> -t <corpus>
 -c <group_data_column>
 --add_ngrams -n 1 2 3
 --combine_feat_tables
  1to3gram
```

```
#create and cross validate a predictive model
dlatkInterface.py -d <data> -t <corus> -c user_id
 -f 'feat$cat_LIWC2007$msgs_xxx$user_id$16to16'  --outcome_table
 blog_outcomes  --group_freq_thresh 1000 --outcomes extraversion
 --controls age gender --model ridge --combo_test_regression
```

Resources

Intended for anyone
(still under develoment)

LEXHUB
LANGUAGE ANALYSIS X SOCIAL SCIENCE

lexhub.org

Thank You!

...the largest data set of **who we are.**

The most interesting results may never be hypothesized.

has@cs.stonybrook.edu

Stony Brook University

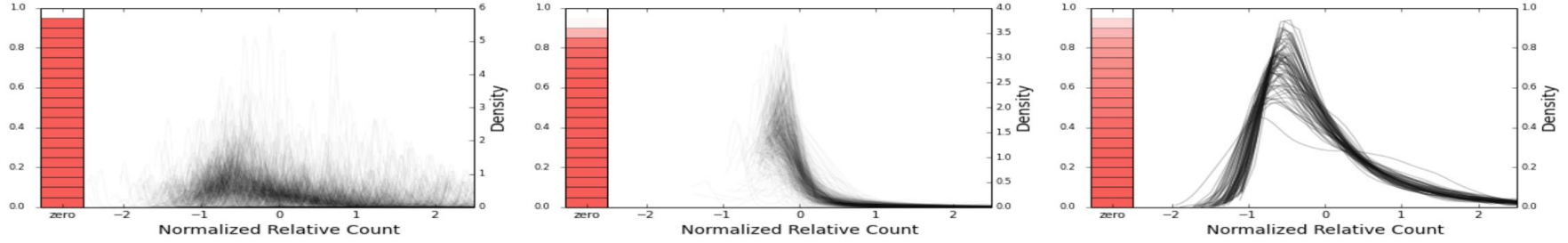
Thank You!

The most 

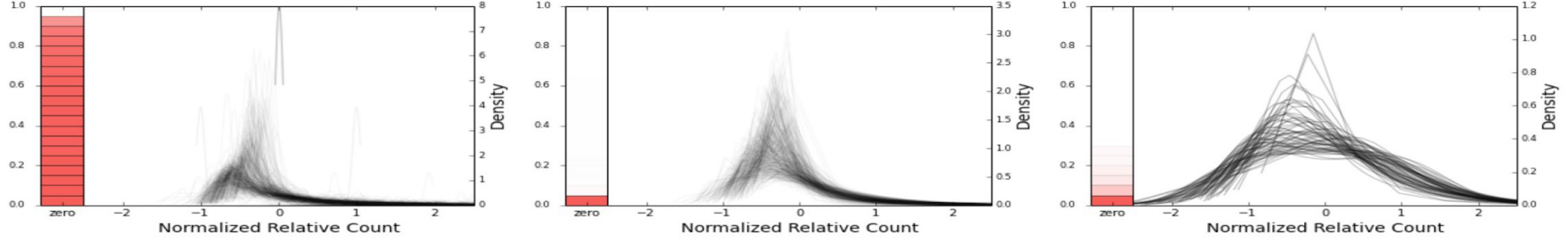
thesized.

Stony Brook University

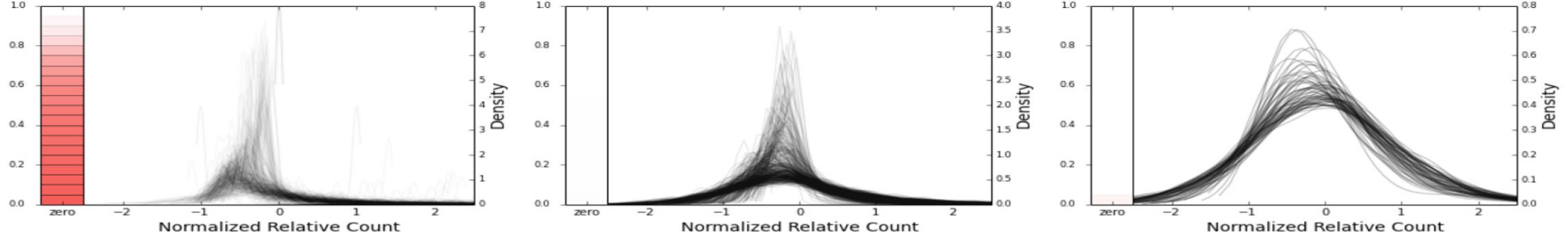
Supplemental Material



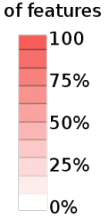
(a) Message - 1gram (b) Message - Topic (c) Message - Lexicon



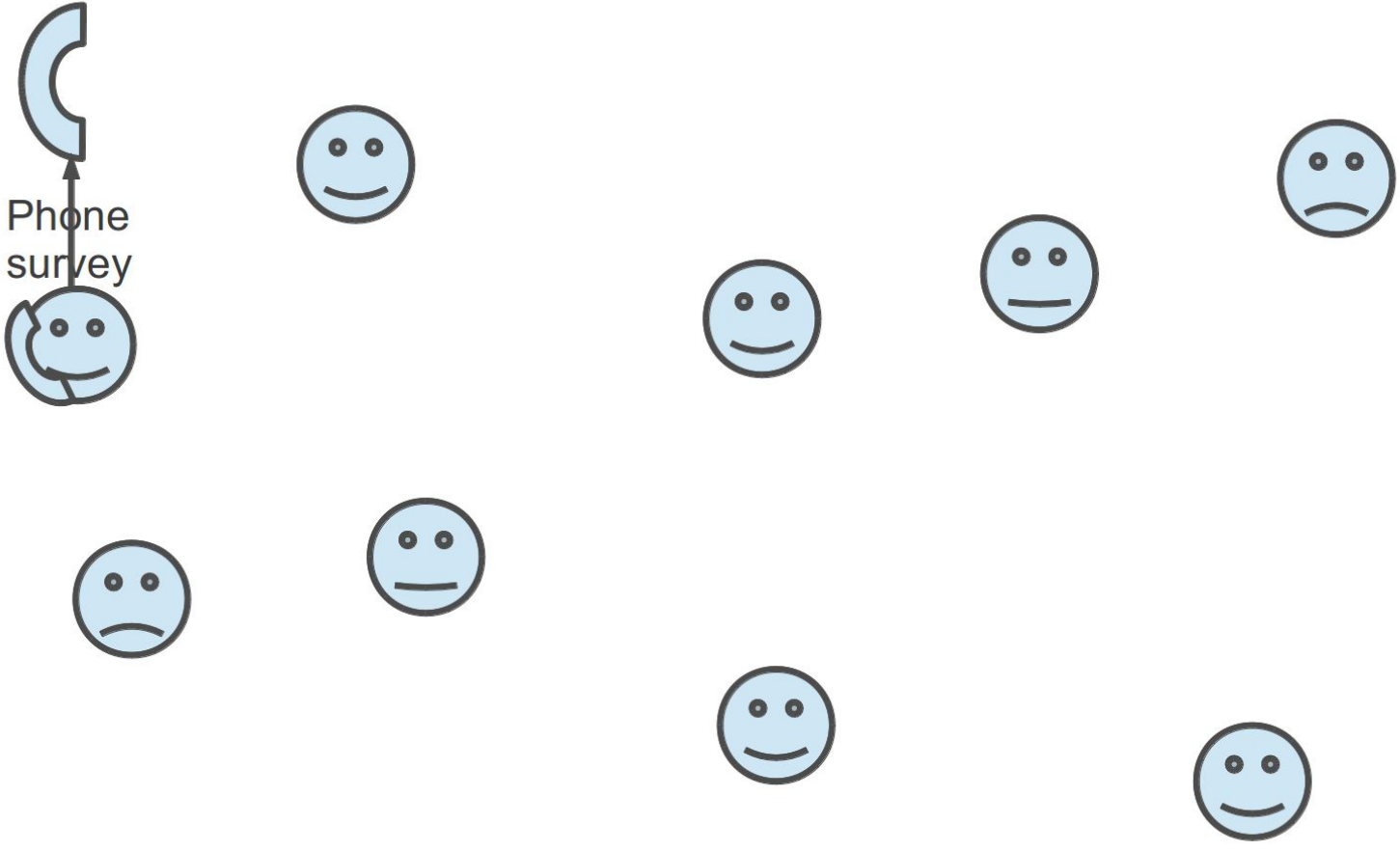
(d) User - 1gram (e) User - Topic (f) User - Lexicon



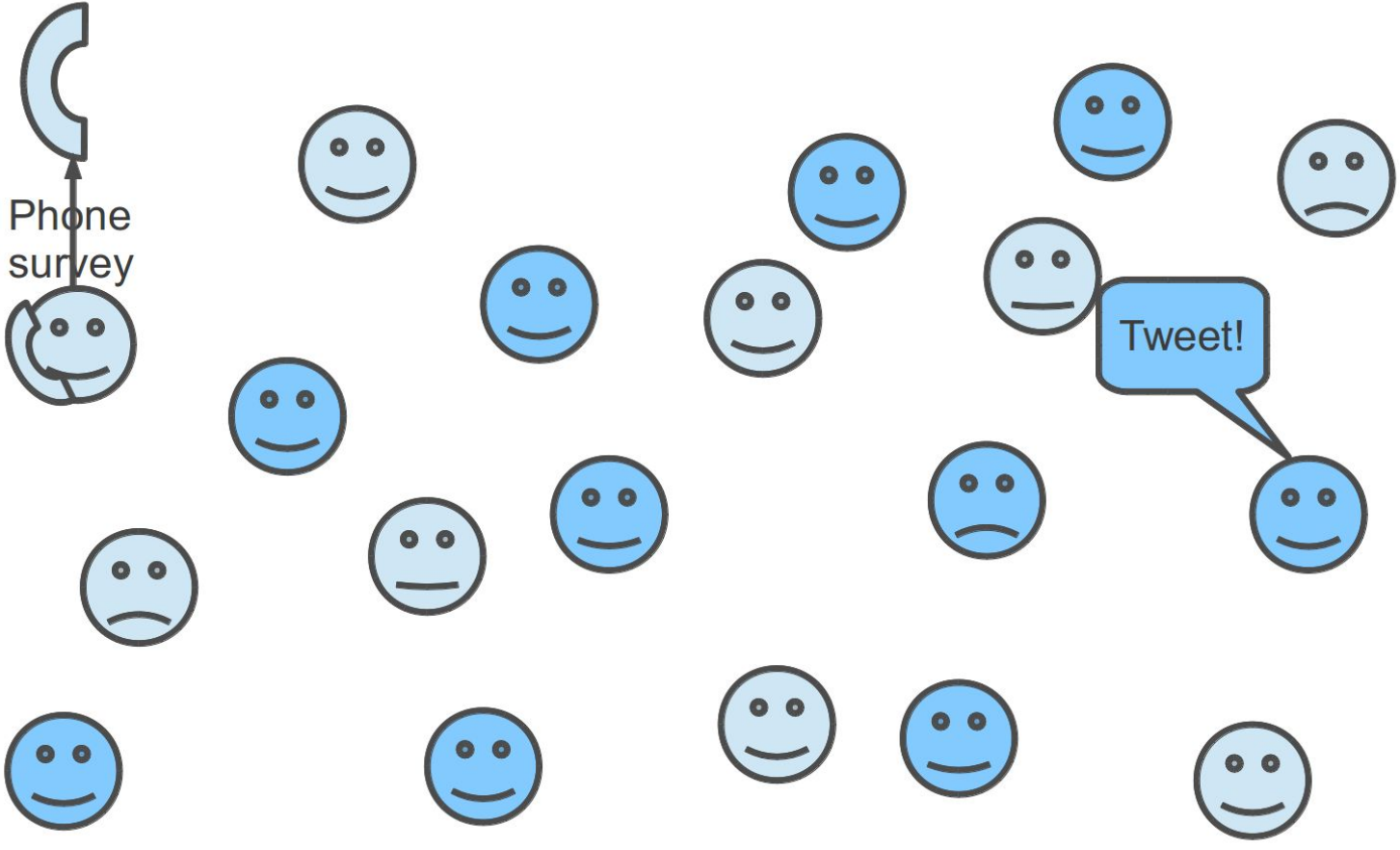
(g) County - 1gram (h) County - Topic (i) County - Lexicon



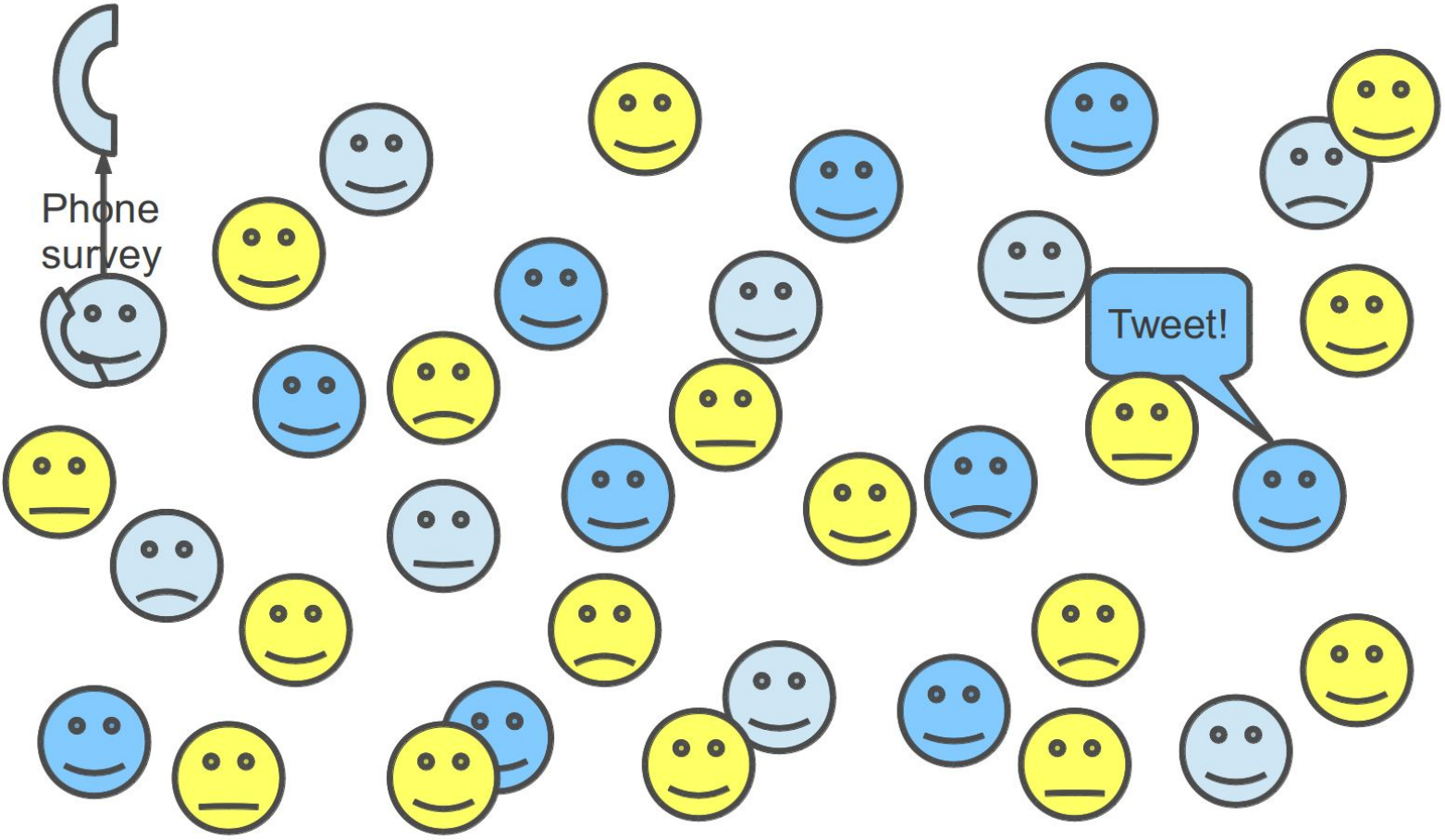
Predicting based on a different sample



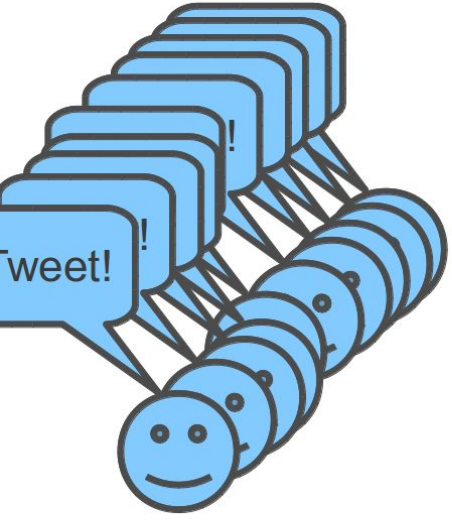
Predicting based on a different sample

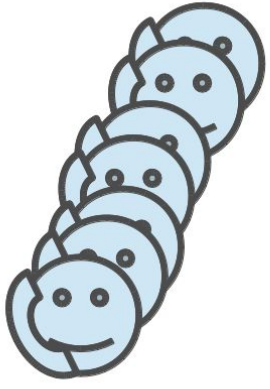


Predicting based on a different sample



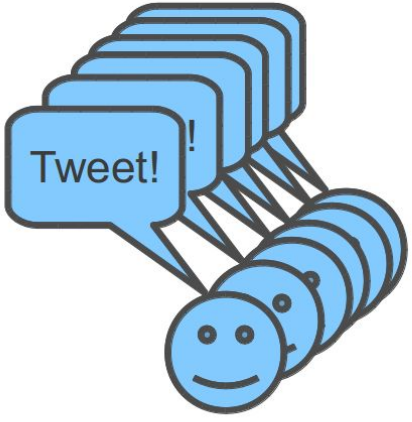
Representative Sample?

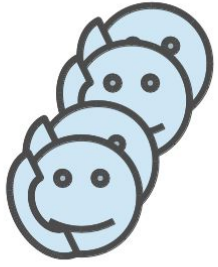




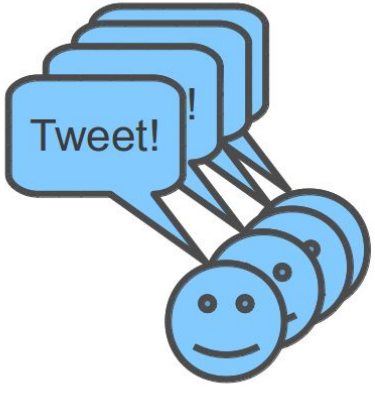
Surveyed well-being
from
representative sample.

Representative Sample?

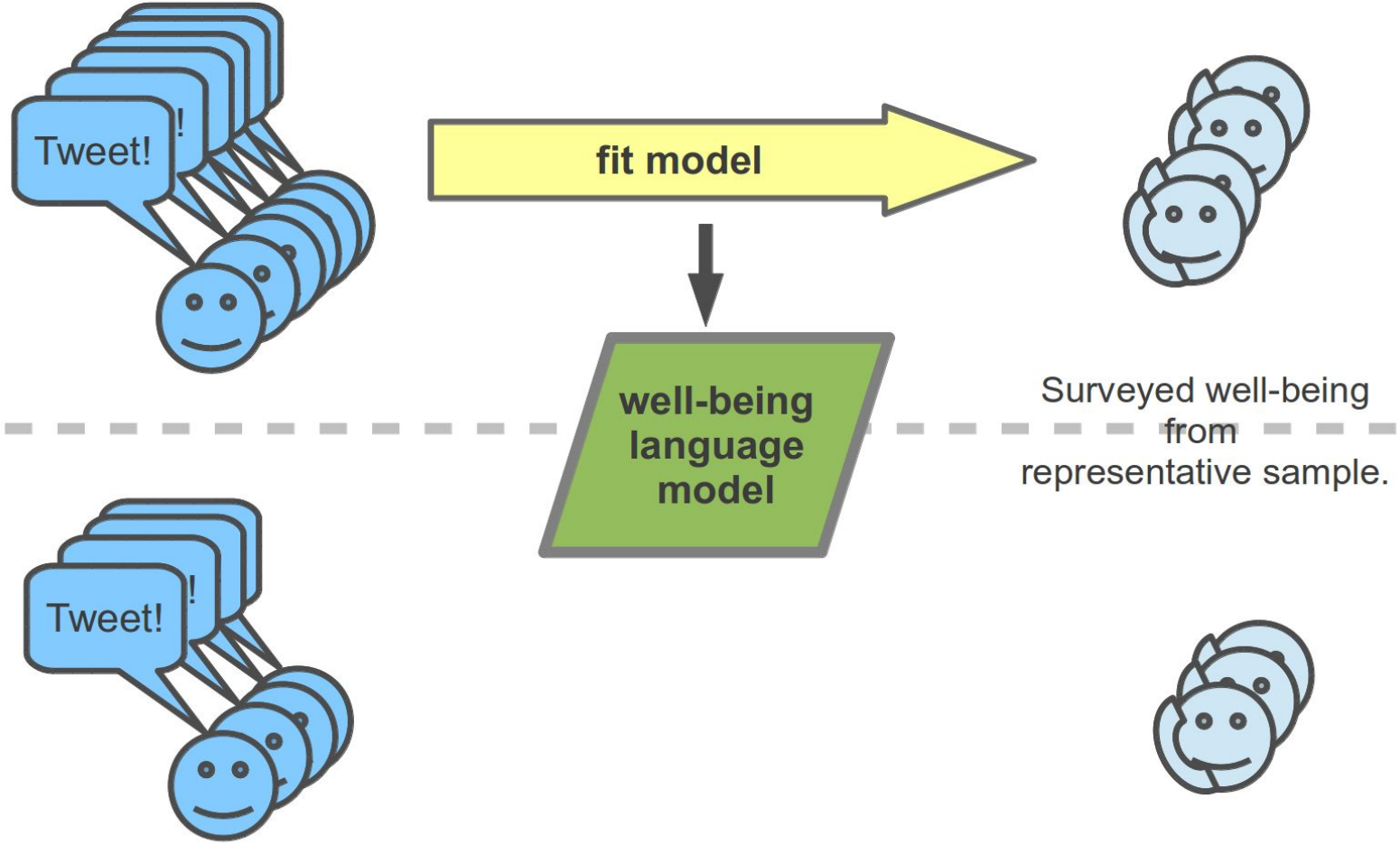




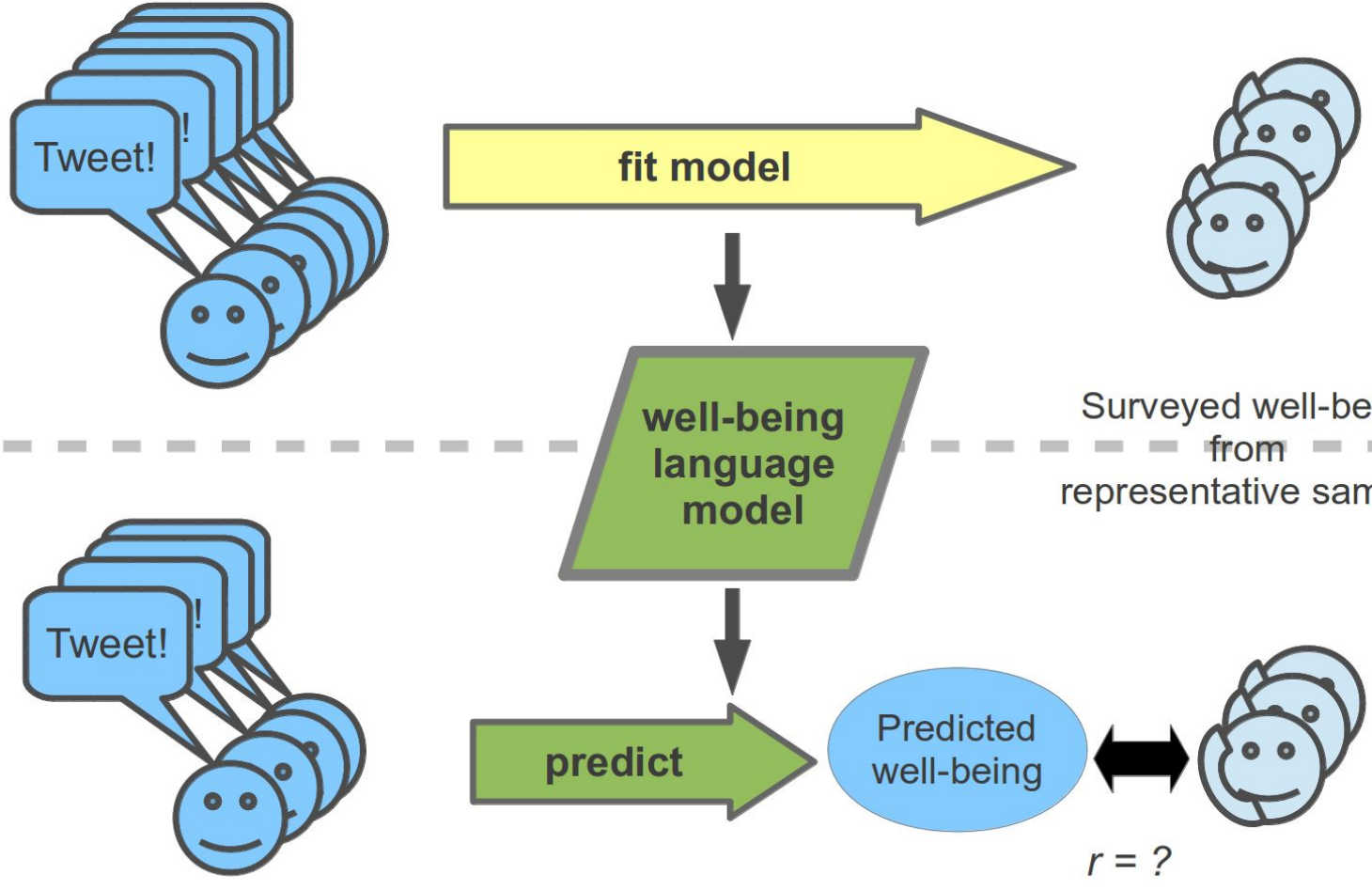
Surveyed well-being from representative sample.



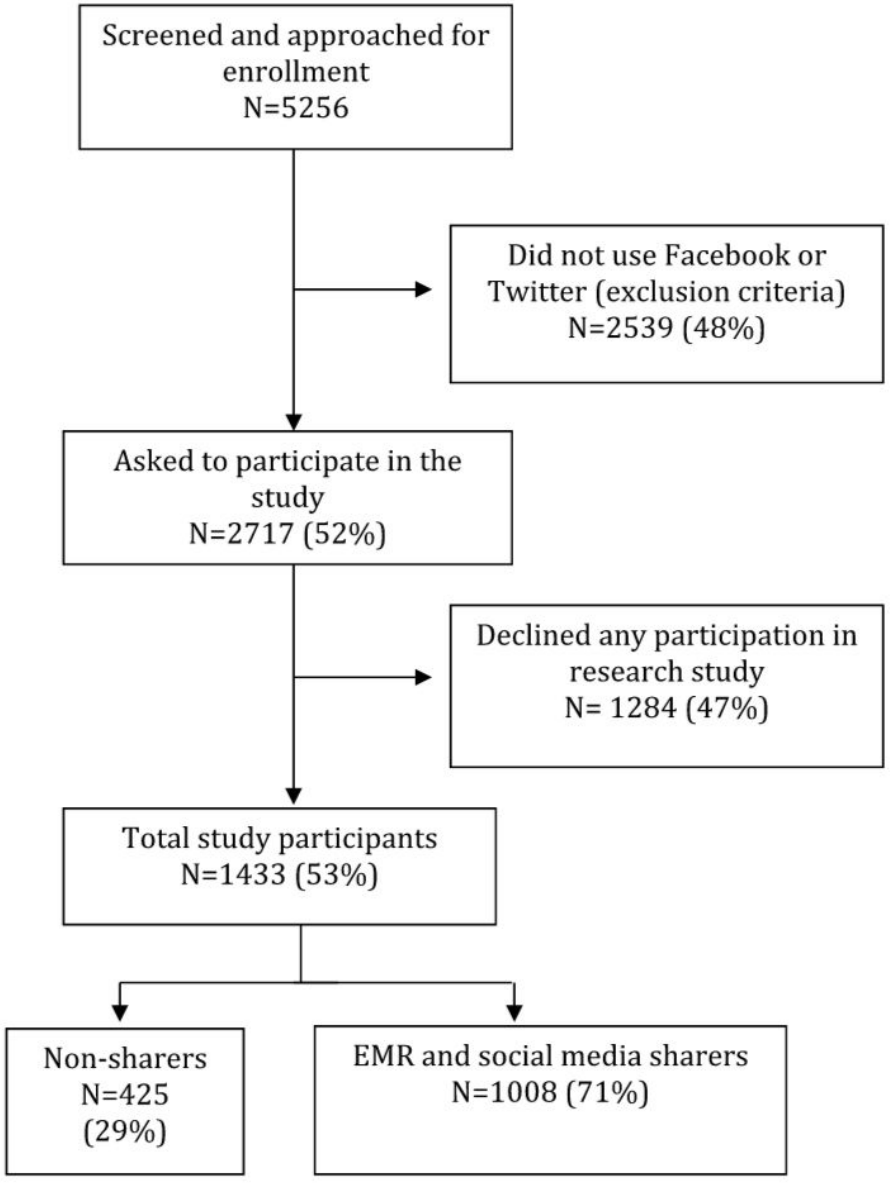
Representative Sample?



Representative Sample?



Kevin A Padrez, L Ungar, HA Schwartz, RJ Smith, S Hill, T Antanavicius, DM Brown, P Crutchley, D Asch, Merchant. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ Quality & Safety* | 2015

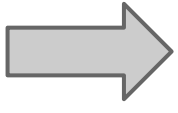


ICD-9 Diagnosis Code	Terms searched	n (%) of patients with diagnosis who used term	n (%) of patients without diagnosis who used term
Abdominal Pain (789)	abdominal pain, stomach pain, belly pain, tummy pain, stomach hurts, belly hurts, tummy hurts, tummyache, stomachache, bellyache	81 (21)	21 (8)[a]
Nausea/Vomiting (787)	nausea, vomiting, vomit, throwing up, spitting up, threw up, puke, puked, vomited	101 (29)	69 (22)[c]
Headache (339&784.0)	headache, migraine, head hurts	141 (59)	192 (46)[b]
Pain in limb (729.5)	leg hurts, arm hurts, finger hurts, toe hurts	5 (3)	5 (1)
UTI (599.0)	uti, urinary tract infection	1 (1)	4 (1)
Back pain (724.1&724.2&724.3&724.4&724.5)	back pain, backache, back hurts	29 (15)	51 (11)
Cough (786.2)	cough, coughing, coughed	40 (26)	109 (22)
Normal delivery (V27.0)	giving birth, gave birth	62 (33)	148 (10)[a]
Anemia (280&281&282&283&284&285)	anemia	3 (2)	2 (0)[c]
Dizziness (780.4)	dizzy, dizziness, vertigo	28 (22)	79 (15)
Asthma (493)	asthma	36 (28)	36 (7)[a]
Acute URI (465)	caught a cold, have a cold	7 (7)	24 (4)
Throat Pain (784.1)	sore throat, throat hurts	26 (24)	62 (11)[a]
Depression (311)	depression, depressed	35 (38)	169 (30)

Big Data and Emergency Care:

Social Media as a Source for Data-Driven Health Prediction and Insight

The Big Data Promise: enabling links with unprecedented amount of **new information.**

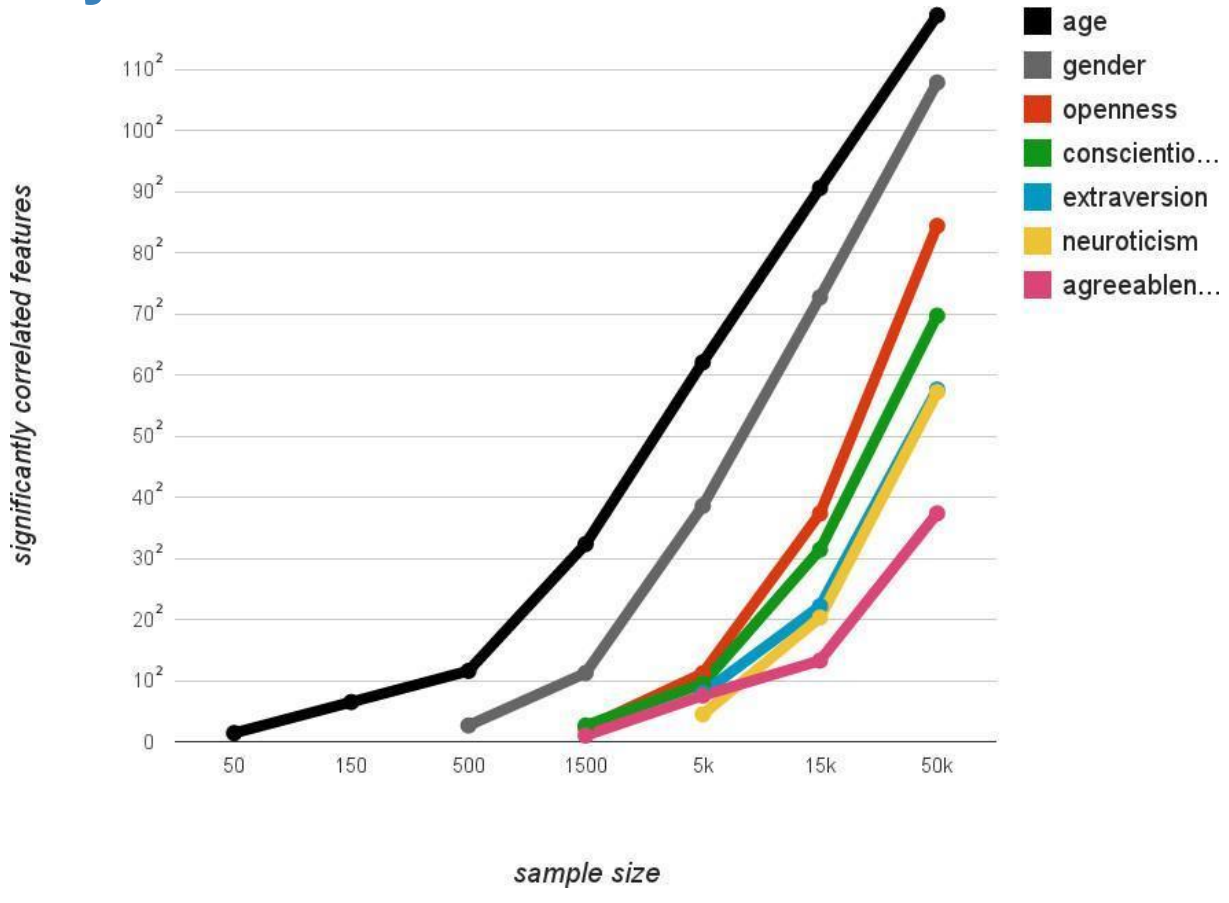


Better health care

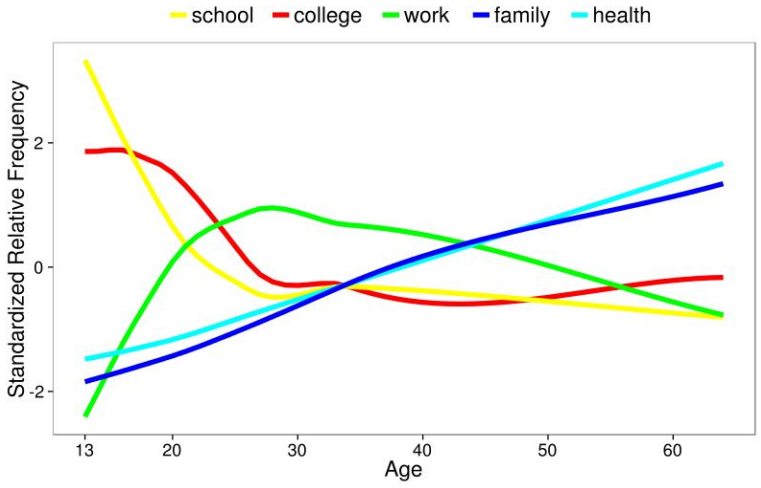
Personalized Health: **Patients' Everyday Lives**

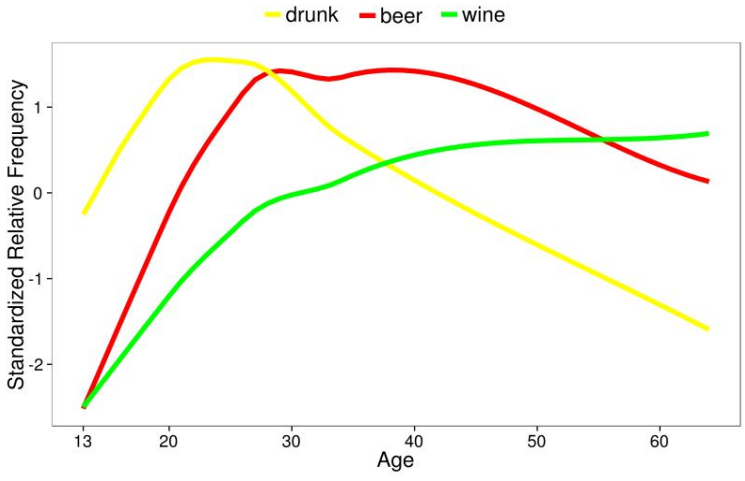
Individual Traits in Facebook

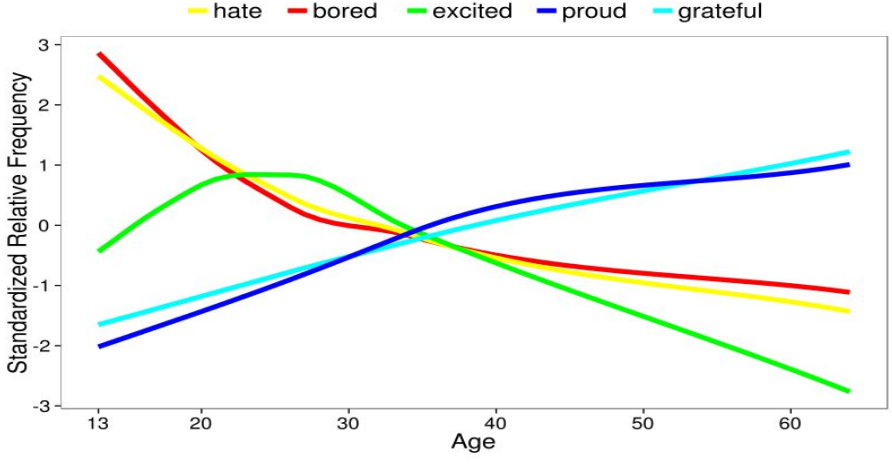
Power Analyses



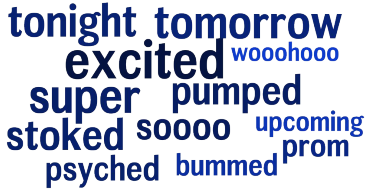
Development

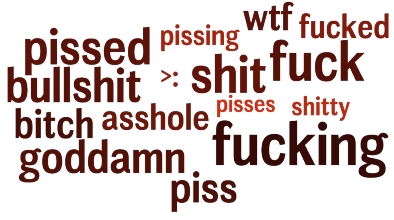






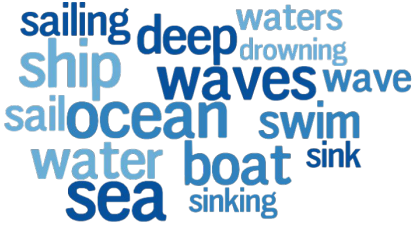
Individual Well-Being



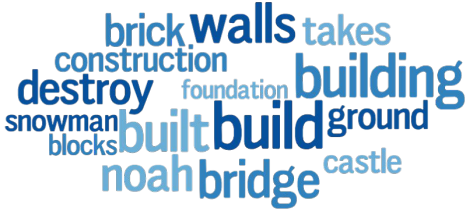


satisfaction with life

Community Well Being

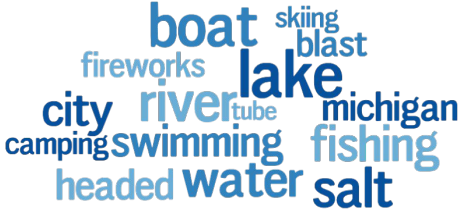


$r' = .15$



$r' = .13$

$r' = .12$

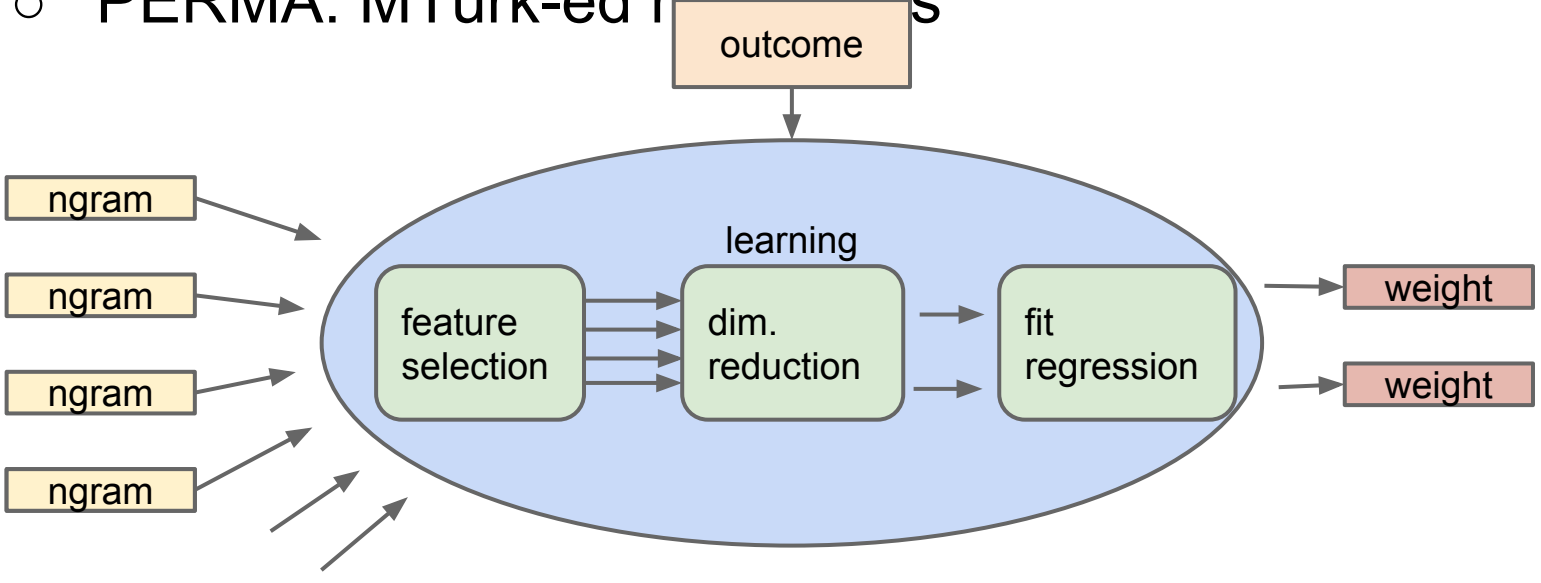


$r' = .12$

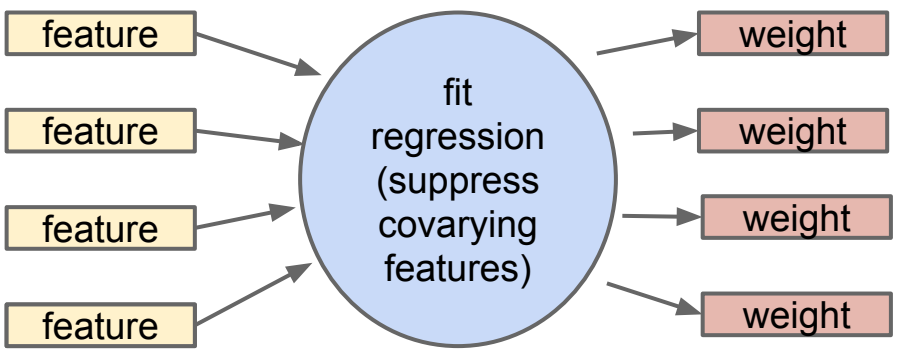
$r' = .12$

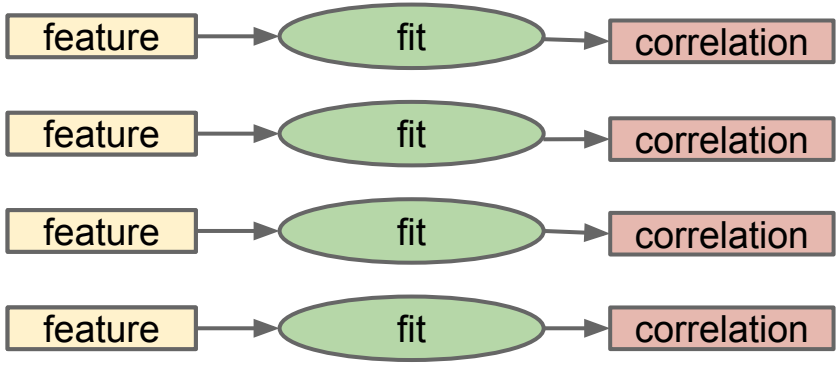
Generating Lexica from *most* Supervised n-gram Models

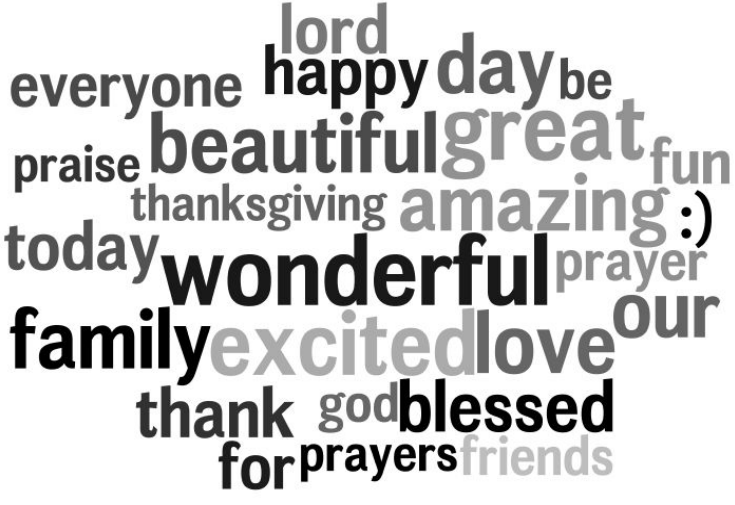
- Generalize multi-variate regression model into lexica.
- Works at multiple levels:
  Hand annotated messages or users
  - OCEAN: User-level Cambridge data set
  - PERMA: MTurk-ed messages



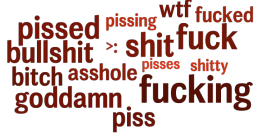
Multivariate or Univariate for *Insights*?

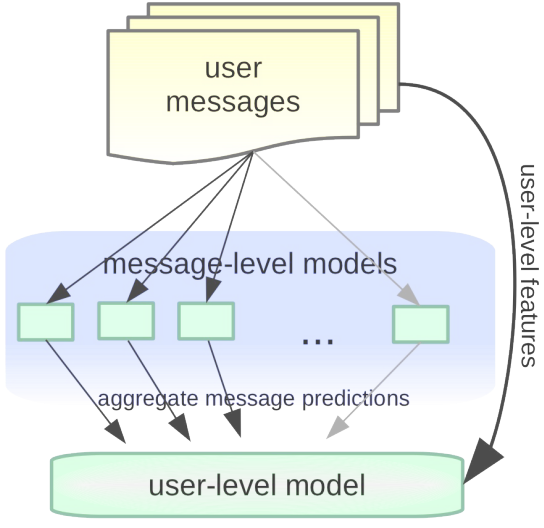






Individual Well-Being: message to user-level





baselines	r
(mean)	.000
lexica: GNH	.210
lexica: Hedonometer	.108

message and user-level

writing size

sample size / populations
(Gosling 2004; 2010)

self-descriptive variables

Why Social Media and Language?

unobtrusive

longitudinal / look back in time

often personal /
everyday concerns

potential for real-time