From the Twitter Stream to your Stats Screen:

Towards Working with Social Media Data for Official Statistics

H. Andrew Schwartz

@

International Conference and Global Working Group meeting on Big Data for Official Statistics
29 October, 2014, Beijing, China

Penn | World Well-Being Project

...shedding light on psychosocial phenomena through big language analysis.

Thank You

United Nations Statistics Division (UNSD)

National Bureau of Statistics of China (NBS)

Social Media

Social Media

300mil. tweets/day

Social Media

300mil. tweets/day

4bil. messages/day

Social Media

300mil. tweets/day

4bil. messages/day

100mil. (Sina) weibos/day

Social Media

300mil. tweets/day

4bil. messages/day

100mil. (Sina) weibos/day

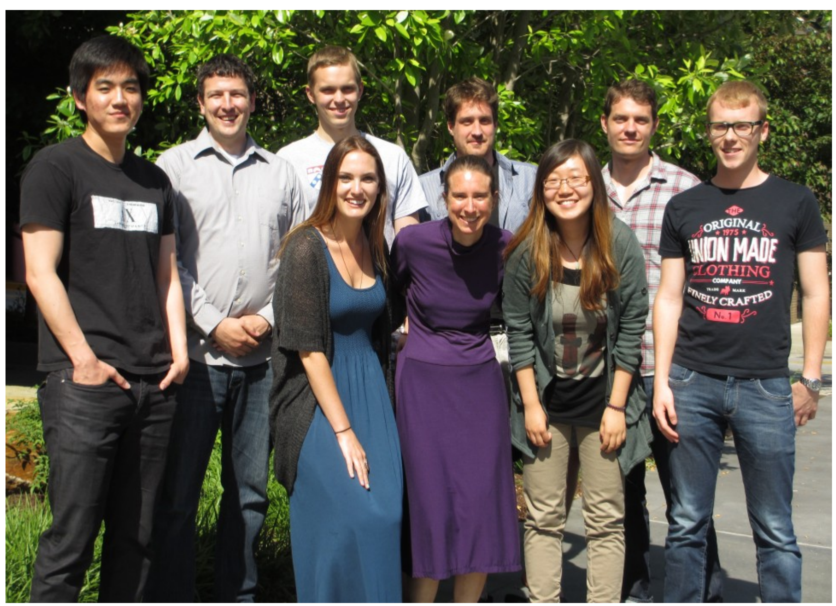
BIGGER DATA

Social Media

300mil. tweets/day

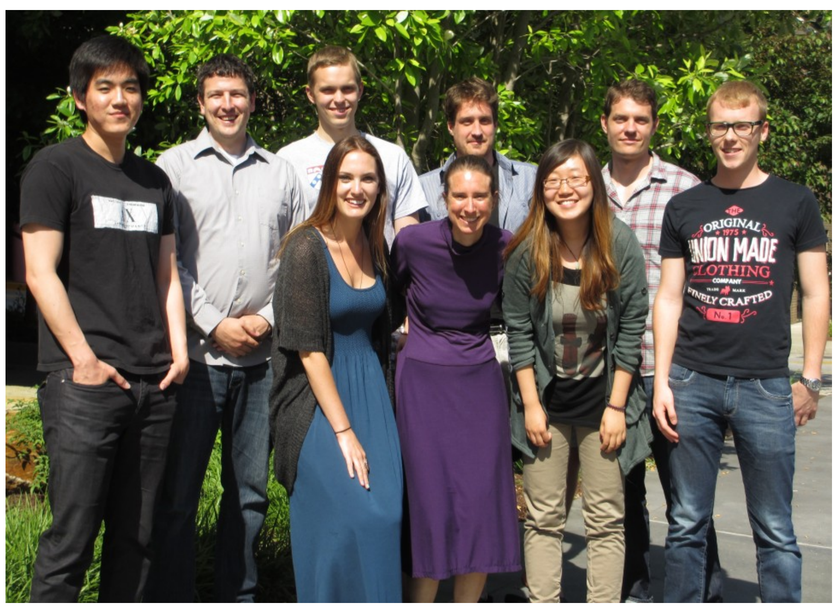
4bil. messages/day

100mil. (Sina) weibos/day



Social Media

	PEOPLE:	
300mil. tweets/day	150mil.	(2014)
4bil. messages/day	1bil.	(2014)
100mil. (Sina) weibos/day	75mil.	(2014)



Social Media

		PEOPLE:	
	300mil. tweets/day	150mil.	(2014)
	4bil. messages/day	1bil.	(2014)
	100mil. (Sina) weibos/day	75mil.	(2014)

Largest dataset(s) of everyday human behavior and conerns.

Social Media

Applications

Largest dataset(s) of everyday human behavior and conerns.

Social Media

1. Measurement

Largest dataset(s) of everyday human behavior and conerns.

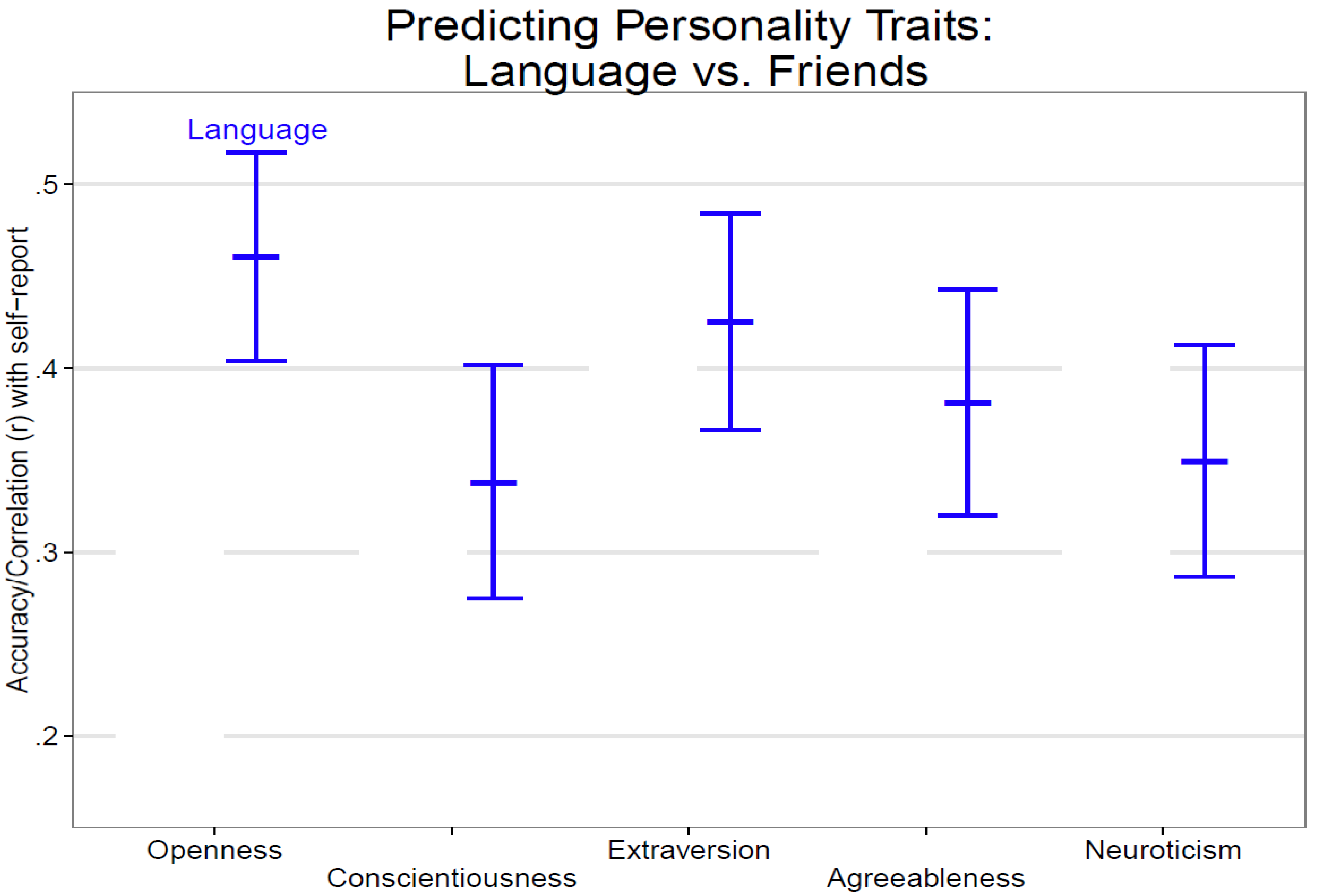
Social Media

1. Measurement

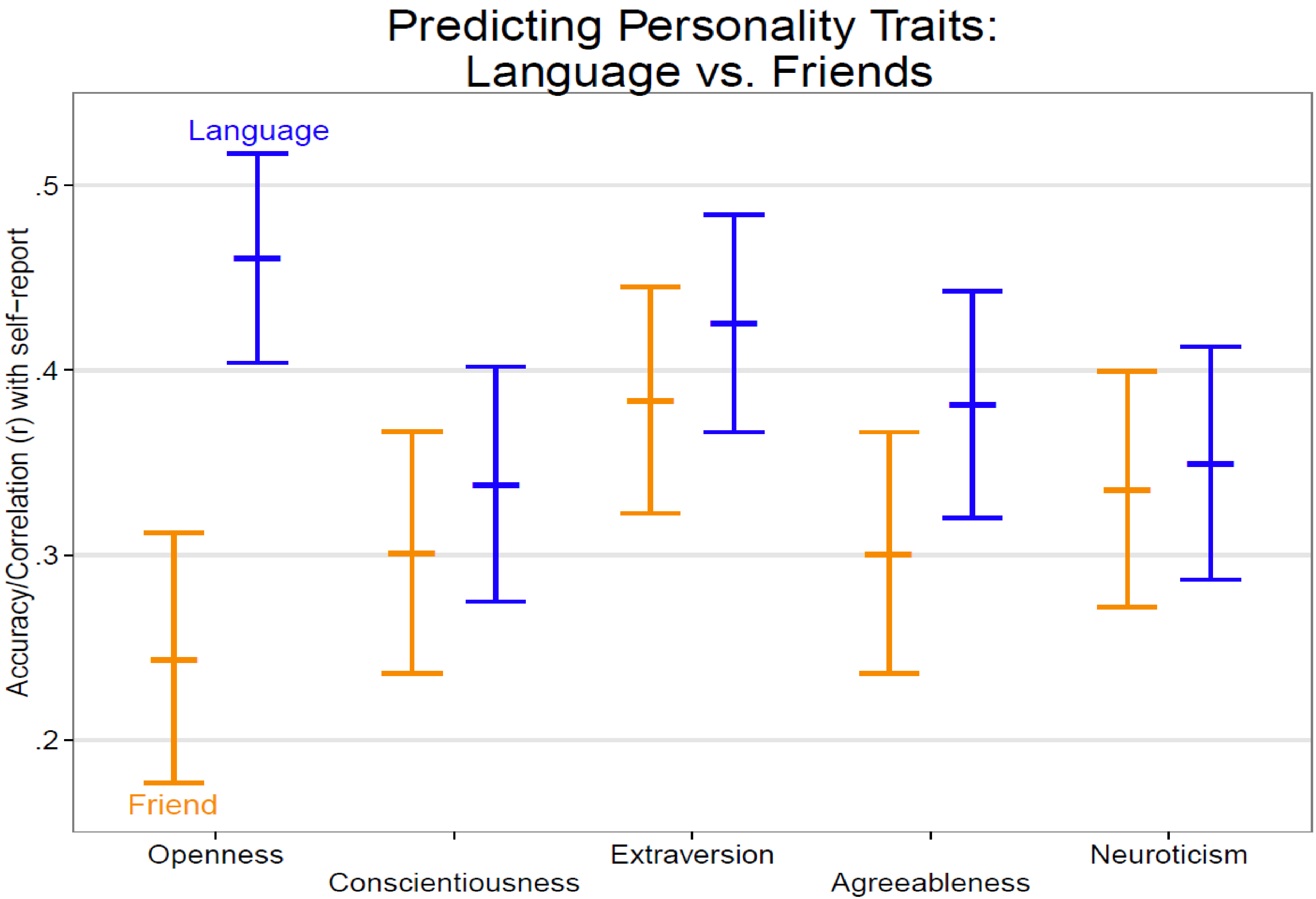
*To what extent **can we replace traditional survey-based methods?***

Largest dataset(s) of everyday human behavior and conerns.

Measurement: Personality

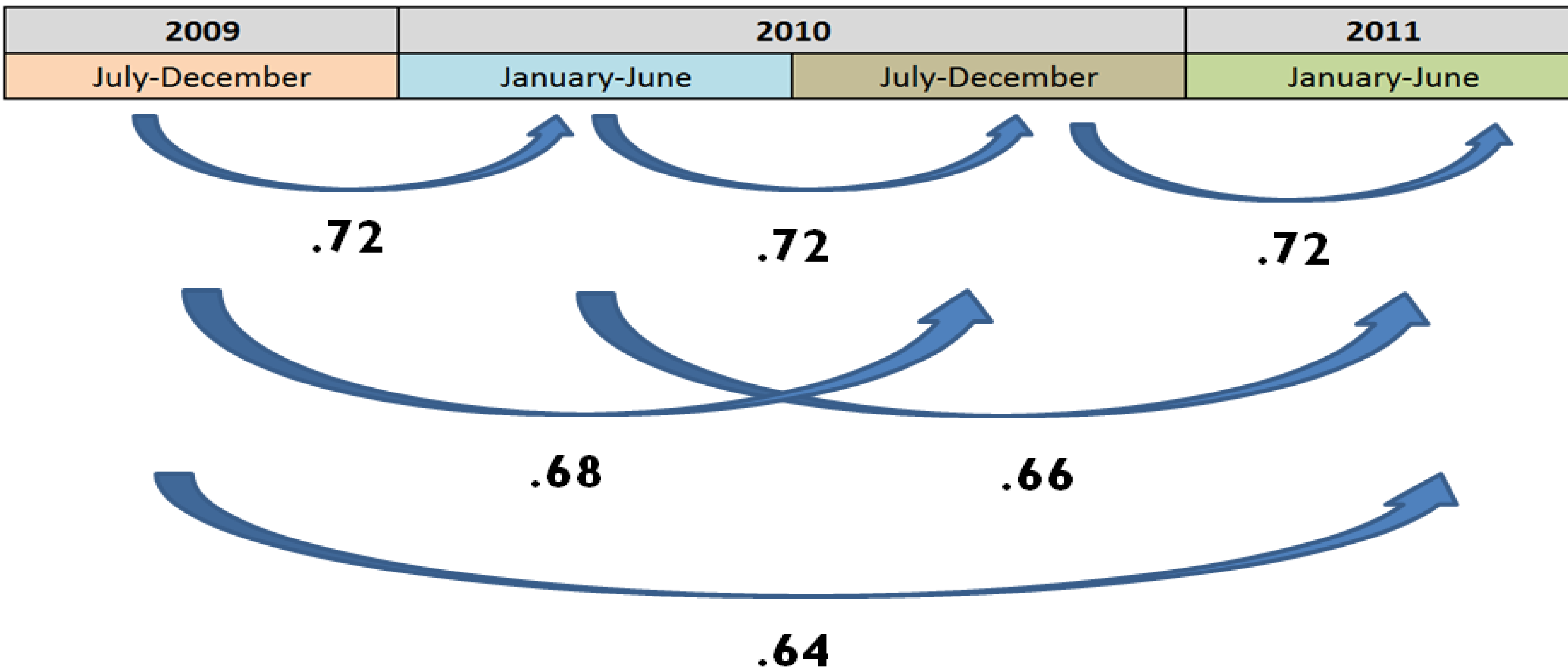


Measurement: Personality



Measurement: Personality

Test-Retest Reliability



Extraversion

Social Media

1. Measurement

To what extent ***can we replace traditional survey-based methods?***

Largest dataset(s) of everyday human behavior and conerns.

Social Media

1. Measurement

To what extent ***can we replace traditional survey-based methods?***

2. Data-driven discovery

Largest dataset(s) of everyday human behavior and conerns.

Social Media

1. Measurement

*To what extent **can we replace traditional survey-based methods?***

2. Data-driven discovery

*Can we discovery **new links with outcomes?***
What is driving a trend?

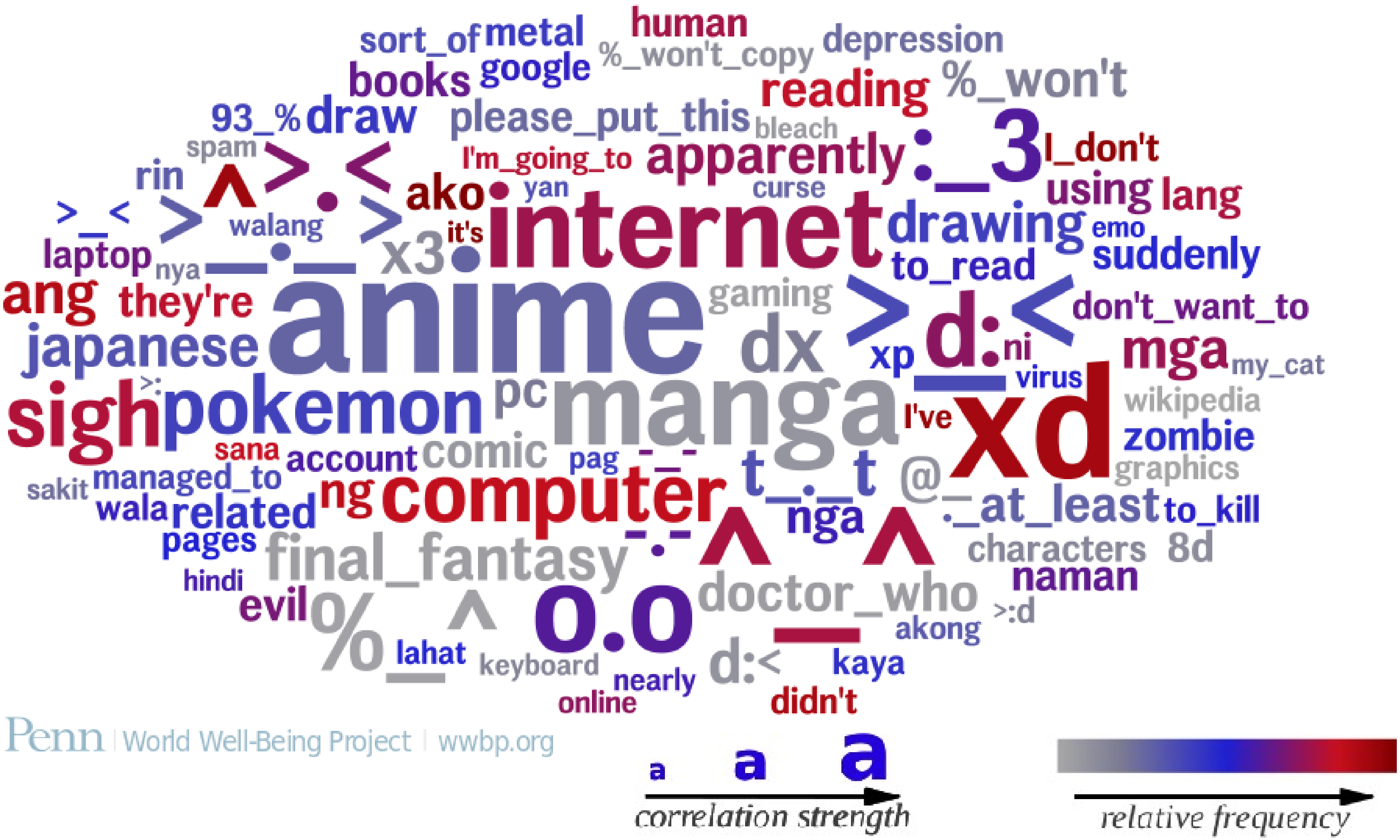
Largest dataset(s) of everyday human behavior and conerns.

Data-driven Social Science: Extraversion

sociable, assertive, active, energetic, talkative, outgoing

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)*.

Data-driven Social Science: 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).*

Data-Driven Social Science: Neuroticism

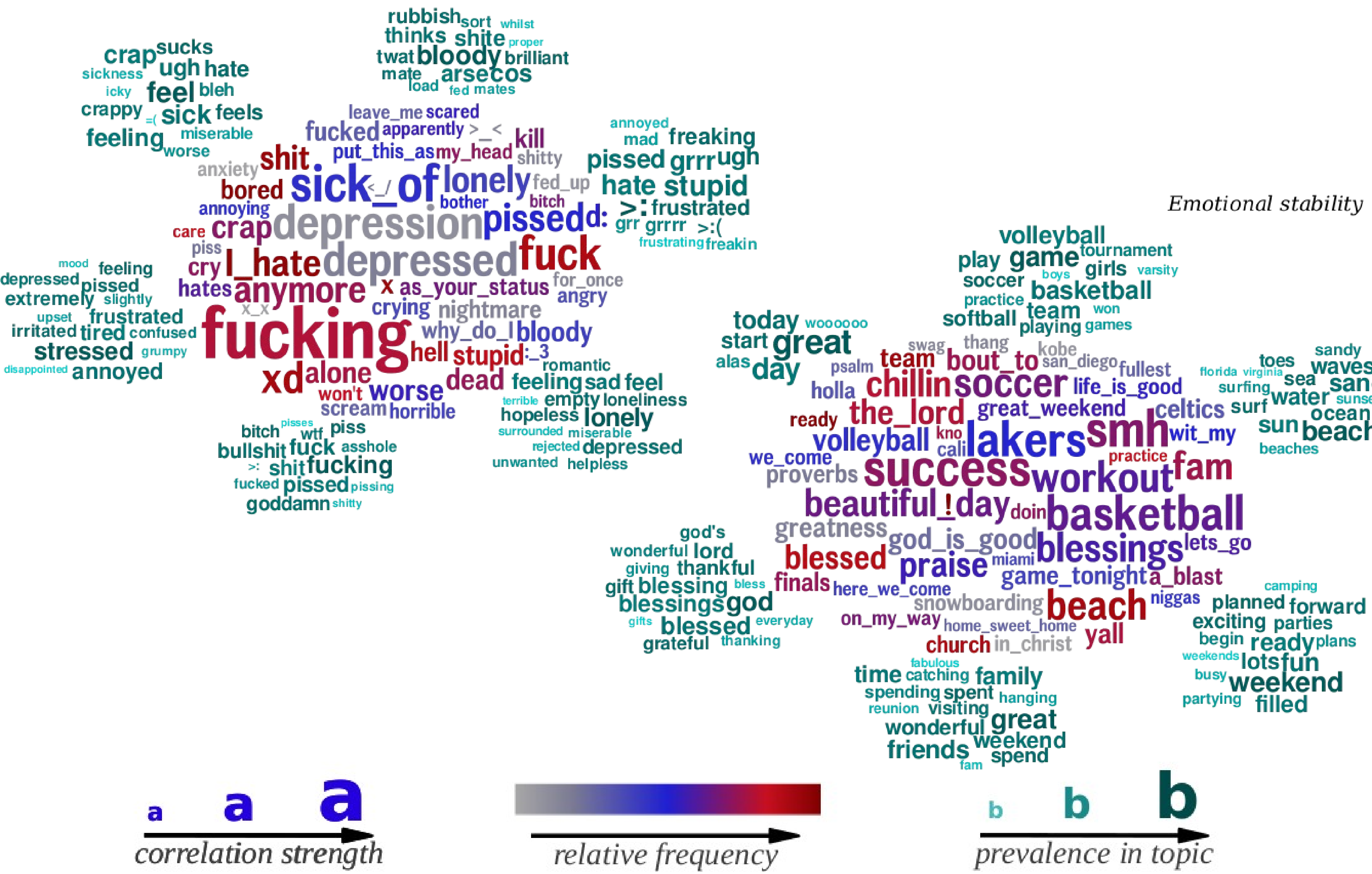
moody, anxious, fearful, worry-prone, depressive

Explicit Language Warning

Data-Driven Social Science: Neuroticism

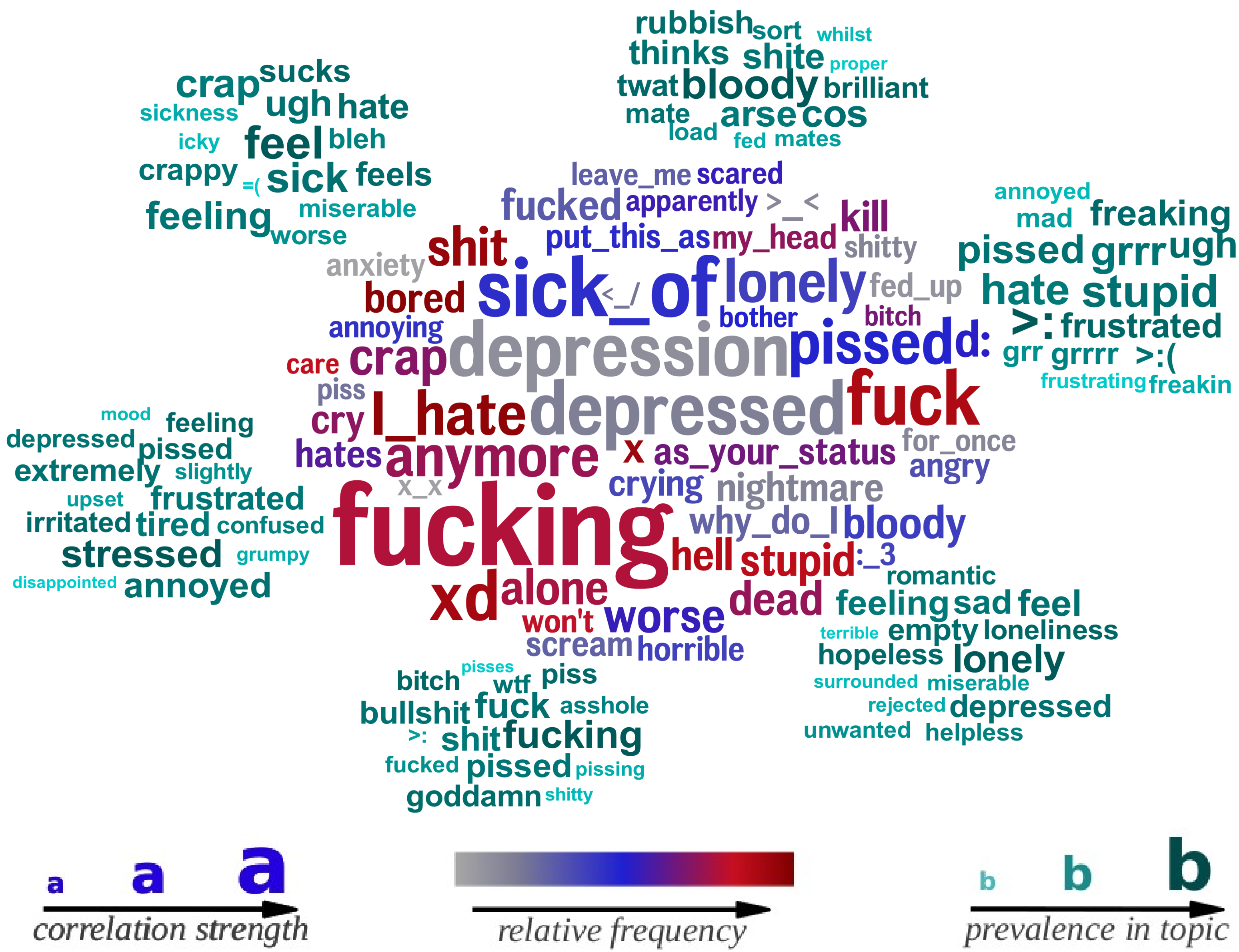
Neuroticism

moody, anxious, fearful, worry-prone, depressive

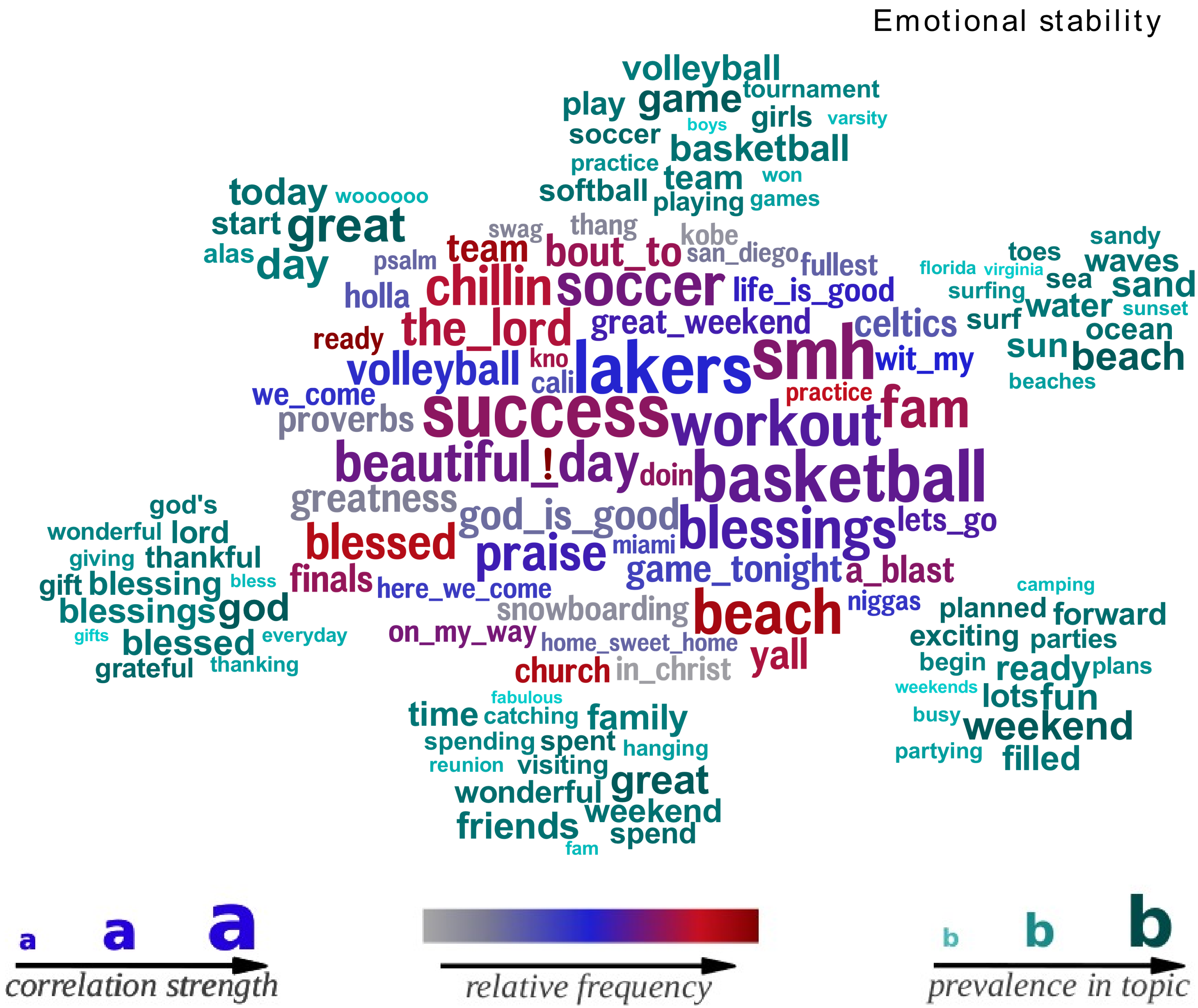


Data-Driven Social Science: Neuroticism

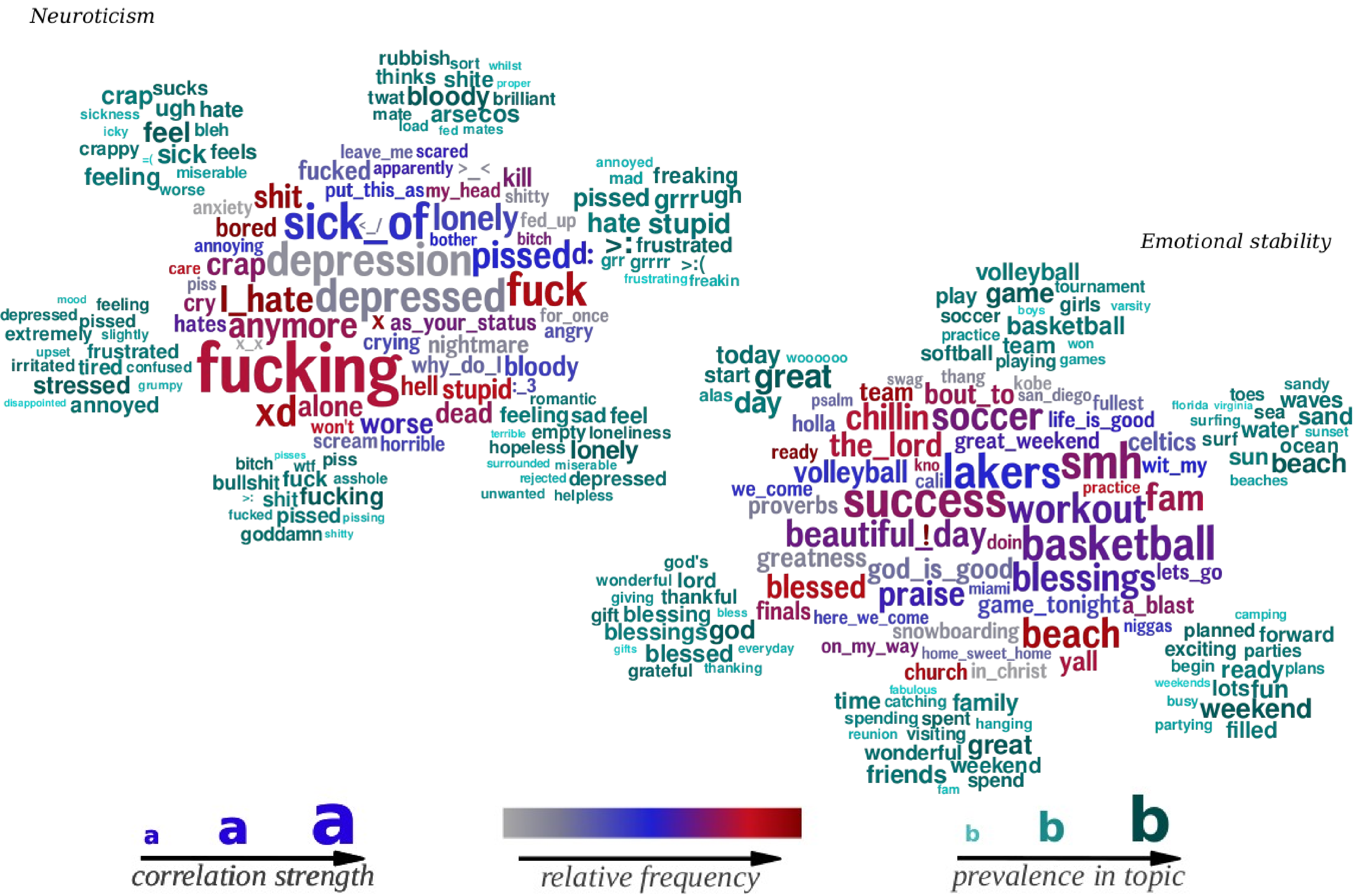
Neuroticism moody, anxious, fearful, worry-prone, depressive



Data-Driven Social Science: Neuroticism



Data-Driven Social Science: Neuroticism



Social Media

1. Measurement

To what extent ***can we replace traditional survey-based methods?***

2. Data-driven discovery

Can we discovery ***new links with outcomes?***

What is driving a trend?

Social Media

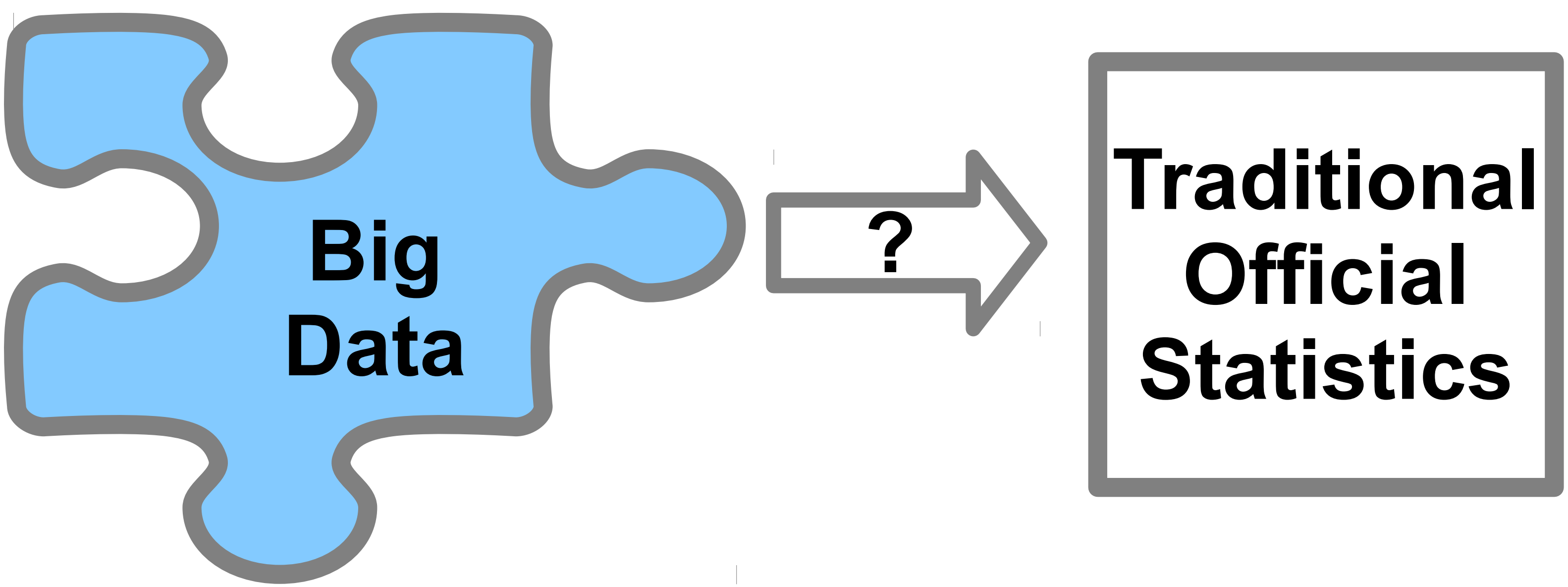
1. Measurement

To what extent ***can we replace traditional survey-based methods?***

2. Data-driven discovery

Can we discovery ***new links with outcomes?***

What is driving a trend?



Social Media

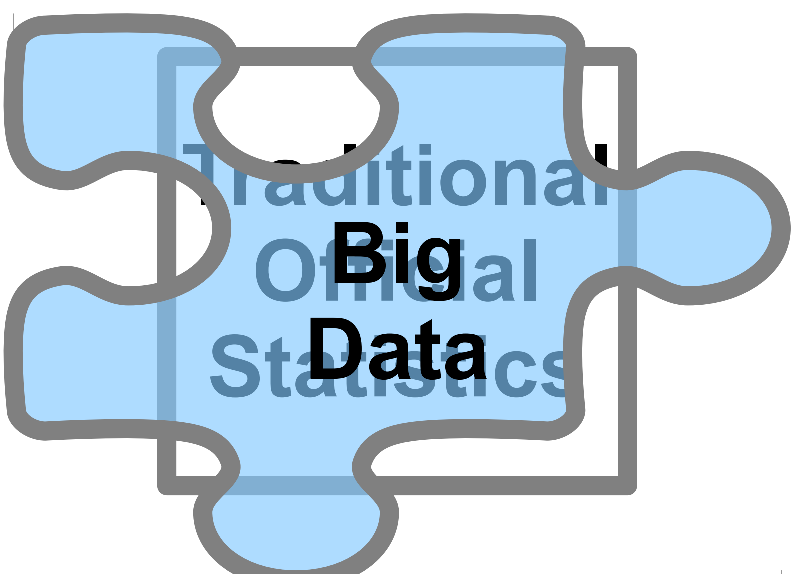
1. Measurement

To what extent ***can we replace traditional survey-based methods?***

2. Data-driven discovery

Can we discovery ***new links with outcomes?***

What is driving a trend?



Overview

- Introduction
- Background on Social Media Data
- Examples
- Challenges
- Summary

Overview

- Introduction
- **Background on Social Media Data**
  - **Sources**
  - **Types**
  - **Acquisition**
  - **Analysis Methodology**
- Examples
- Challenges
- Summary

Social Media Sources

microblogging	social interaction	messaging
mostly public	*somewhat private*	*private*

Social Media Sources

microblogging	social interaction	messaging
Twitter Weibo	Facebook Renren	Text Messages SnapChat WeChat
mostly public	*somewhat private*	*private*

Social Media Sources

microblogging	**social interaction**	**messaging**
Twitter Weibo	Facebook Renren	Text Messages SnapChat WeChat
mostly public	*somewhat private*	*private*
big	*bigger*	*biggest*

Social Media Sources

microblogging	social interaction	messaging
Twitter Weibo	Facebook Renren	Text Messages SnapChat WeChat
mostly public	*somewhat private*	*private*
big	*bigger*	*biggest*

<u>Other social media</u>

Instagram
YouTube
Yelp
Pinterest
Tumblr
Reddit

Social Media Sources

microblogging	social interaction	messaging
Twitter Weibo	Facebook Renren	Text Messages SnapChat WeChat
mostly public	*somewhat private*	*private*
big	*bigger*	*biggest*

Other social media
Instagram
YouTube
Yelp
Pinterest
Tumblr
Reddit

Search
Google
Yahoo
Baidu
Bing

Social Media Data Types:

Text!

Social Media Data Types:

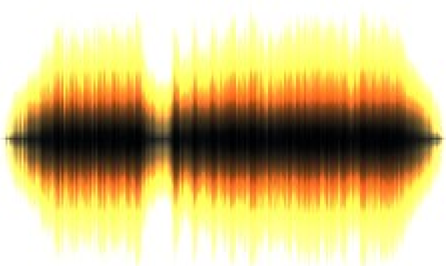
Text!

Social Media Data Types:

Text!

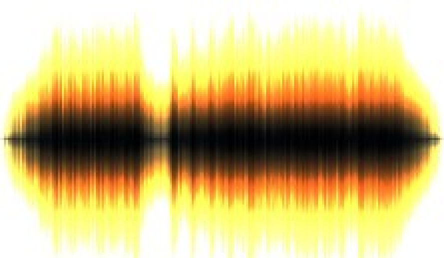
Social Media Data Types:

Text!



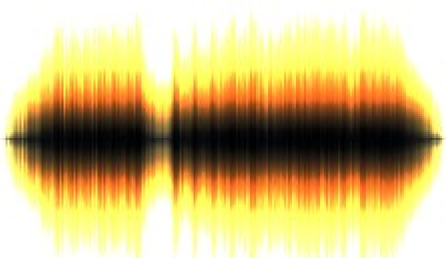
Social Media Data Types:

Text!



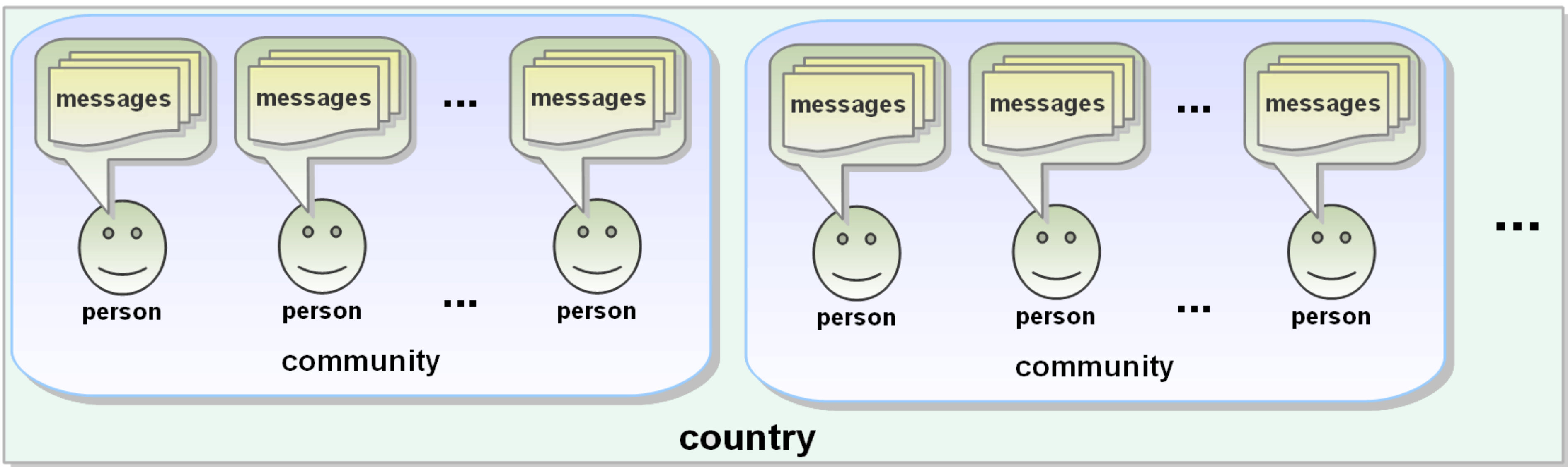
Social Media Data Types:

Text!

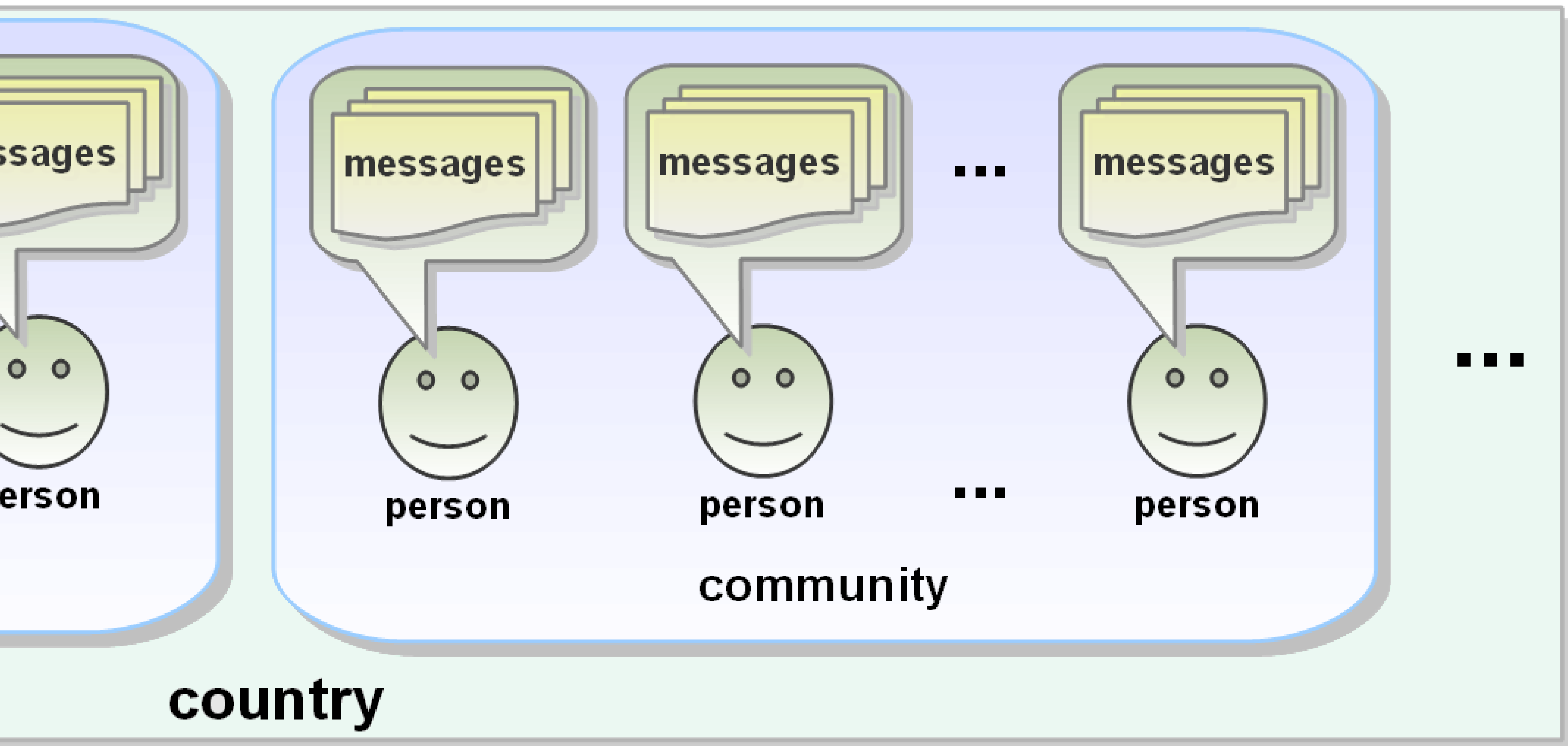


Social Media Data Types:

Levels of Analysis



Social Media Data Types:



Acquiring Social Media

Twitter

- Application Programming Interfaces (APIs)
  - random stream (1% daily = ~2 to 3.5m)
  - filter stream (1%; not random sample)
  - search API (180 queries per 15 minutes)

Acquiring Social Media

Twitter

- Application Programming Interfaces (APIs)
  - random stream (1% daily = ~2 to 3.5m)
  - filter stream (1%; not random sample)
  - search API (180 queries per 15 minutes)
- More data provided by third parties (Datasift, Gnip, ...)

Acquiring Social Media

JSON encoding

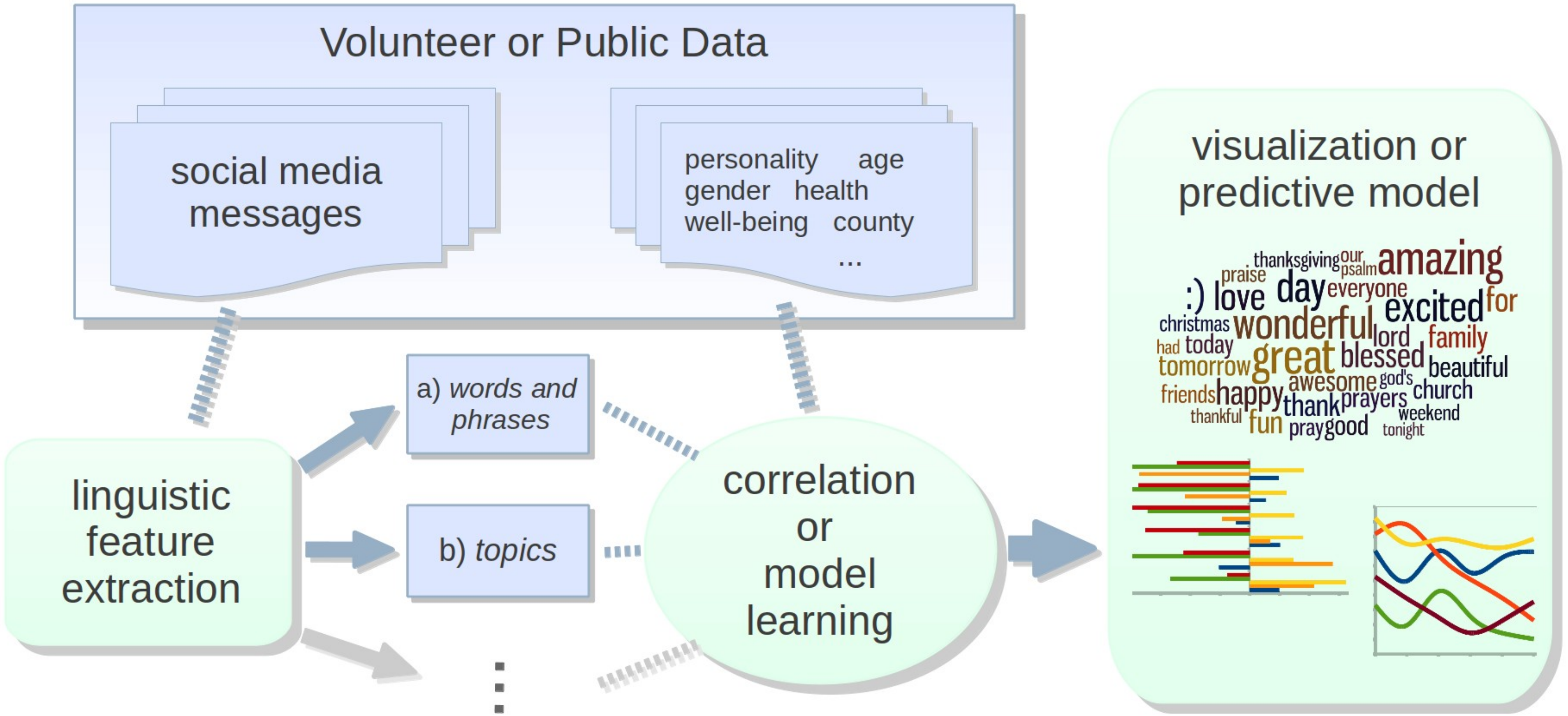

```
{
 "coordinates": None,
 "created_at": "Wed Jan 29 22:58:50 +0000 2014",
 "favorite_count": 19,
 "favorited": False,
 "geo": None,
 "id": 428663556889145344,
 "lang": "en",
 "place": None,
 "retweet_count": 14,
 "retweeted": False,
 "text": "Wow, where did January go? Was I in
Tulsa or Yemen? Or Vermont?",
 ...
}
```


Acquiring Social Media

Facebook

- Graph API
- Limited public data
- Consent participants to share private data through Facebook App.

Analysis / Methodology



Analysis / Methodology

Features

words and phrases: 1 to 3 word sequences more likely to occur together than chance.

- Words identified from text via social-media aware *tokenization.*
- usually restricted to those used more than a few times
- e.g. 'day', 'the beautiful day', 'Mexico City', etc...

Analysis / Methodology

Features

words and phrases: 1 to 3 word sequences more likely to occur together than chance.

- Words identified from text via social-media aware *tokenization.*
- usually restricted to those used more than a few times
- e.g. 'day', 'the beautiful day', 'Mexico City', etc...

topics: Clusters of semantically-related words found via *latent Dirichlet allocation*

e.g.

Method:Data-driven language analysis

Features

words and phrases: 1 to 3 word sequences more likely to occur together than chance.

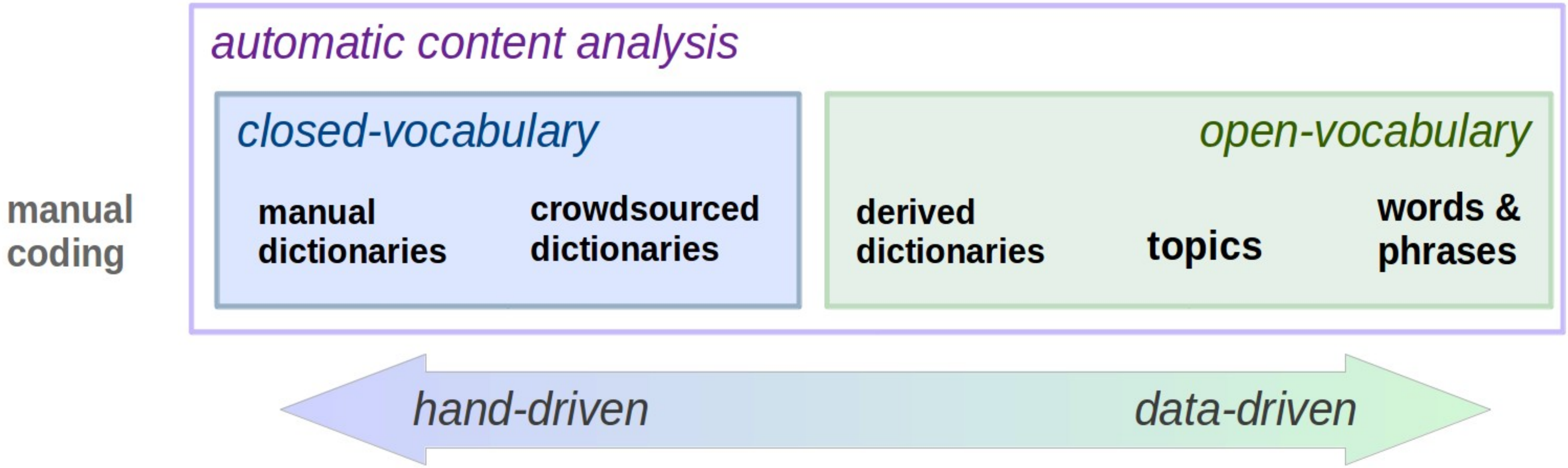
topics: Clusters of semantically-related words found via *latent Dirichlet allocation*

lexica: Manually-created clusters of words

e.g. *positive emotion:* happy, joyous, like, etc…

negative emotion: sad, hate, terrible, etc…

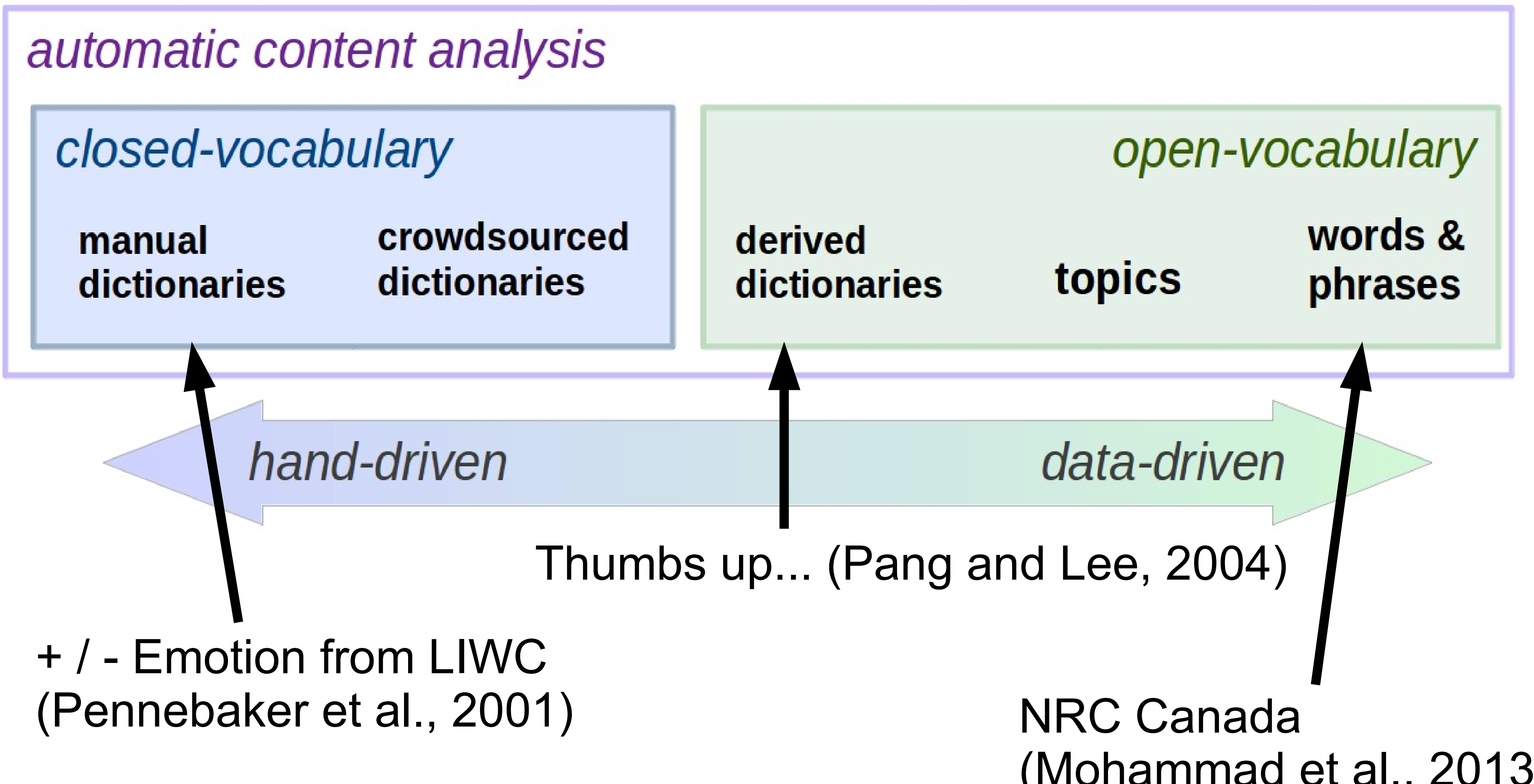
Analysis / Methodology



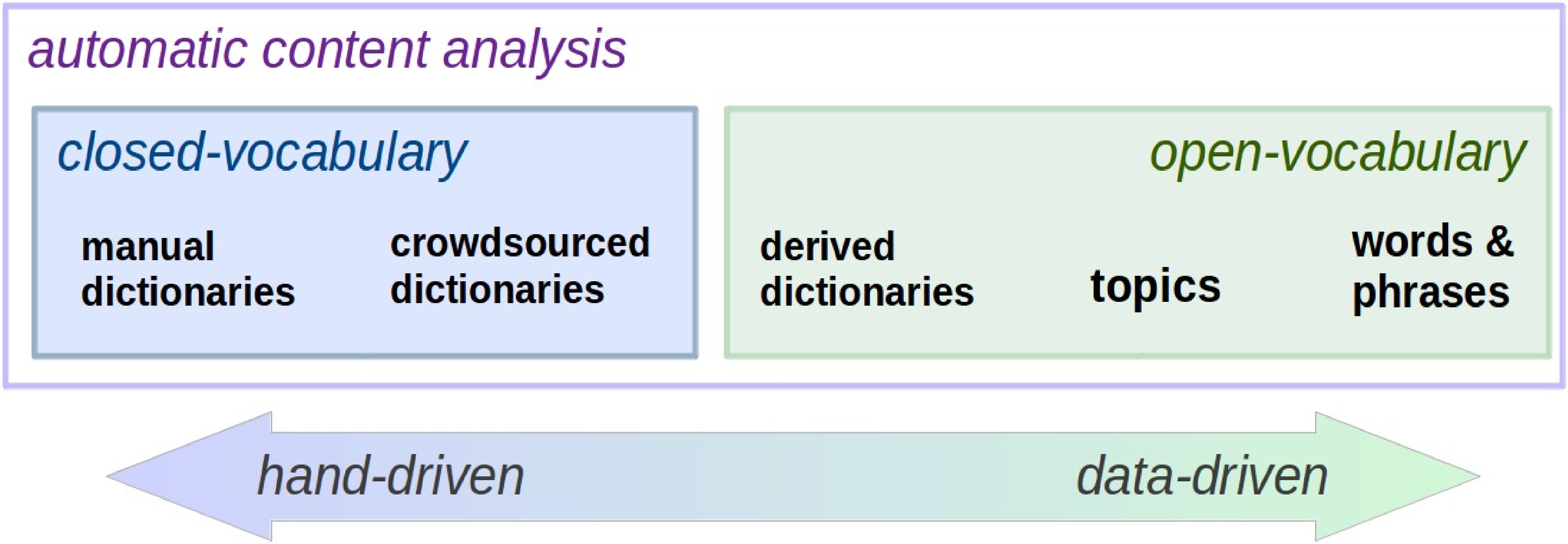
open-vocabulary : Not restricted to predefined lists of features.

Analysis / Methodology

Example: Sentiment Analysis



Analysis / Methodology



All require validation in new domain.
(e.g., new platform, time-frame, or level of analysis)

Analysis / Methodology

Prediction

How to fit a single model on lots of language variables?
(e.g. 25,000 words and phrases)

Methods from Machine Learning:

- discrete outcomes: *support vector machines* (*SVM*)
- continuous outcomes: *ridge regression*

Analysis / Methodology

Prediction

Issues with words as variables:

- sparseness: most words do not occur very often
- high co-variance: e.g. people that say “soccer” often are also more likely to say “goal”

Analysis / Methodology

Prediction

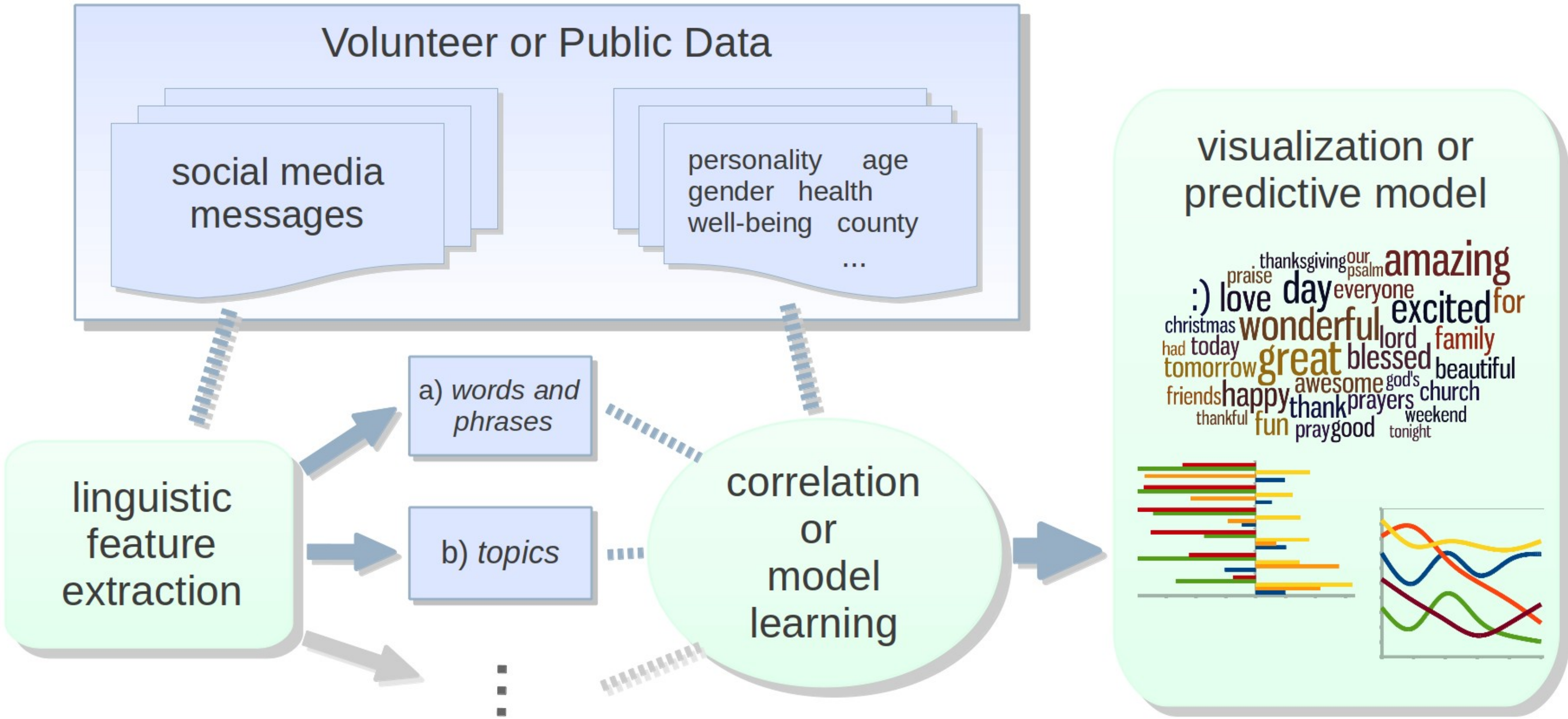
Issues with words as variables:

- sparseness: most words do not occur very often
- high co-variance: e.g. people that say “statistics” often are also more likely to say “variable”

Solutions:

- L1 penalized fit (lasso regression)
- Use principal components analysis before fit

Analysis / Methodology



Some Available Resources

MALLET: Machine Learning Language Toolkit
Good for topic modeling
http://mallet.cs.umass.edu/
GUI: http://code.google.com/p/topic-modeling-tool/

Lightside: Point and Click Machine Learning
http://ankara.lti.cs.cmu.edu/side/download.html

WWBP Resources
wwbp.org/data.html
Coming this January:
"LexHub: Language Analysis X social science"
email to get on list: hansens@seas.upenn.edu

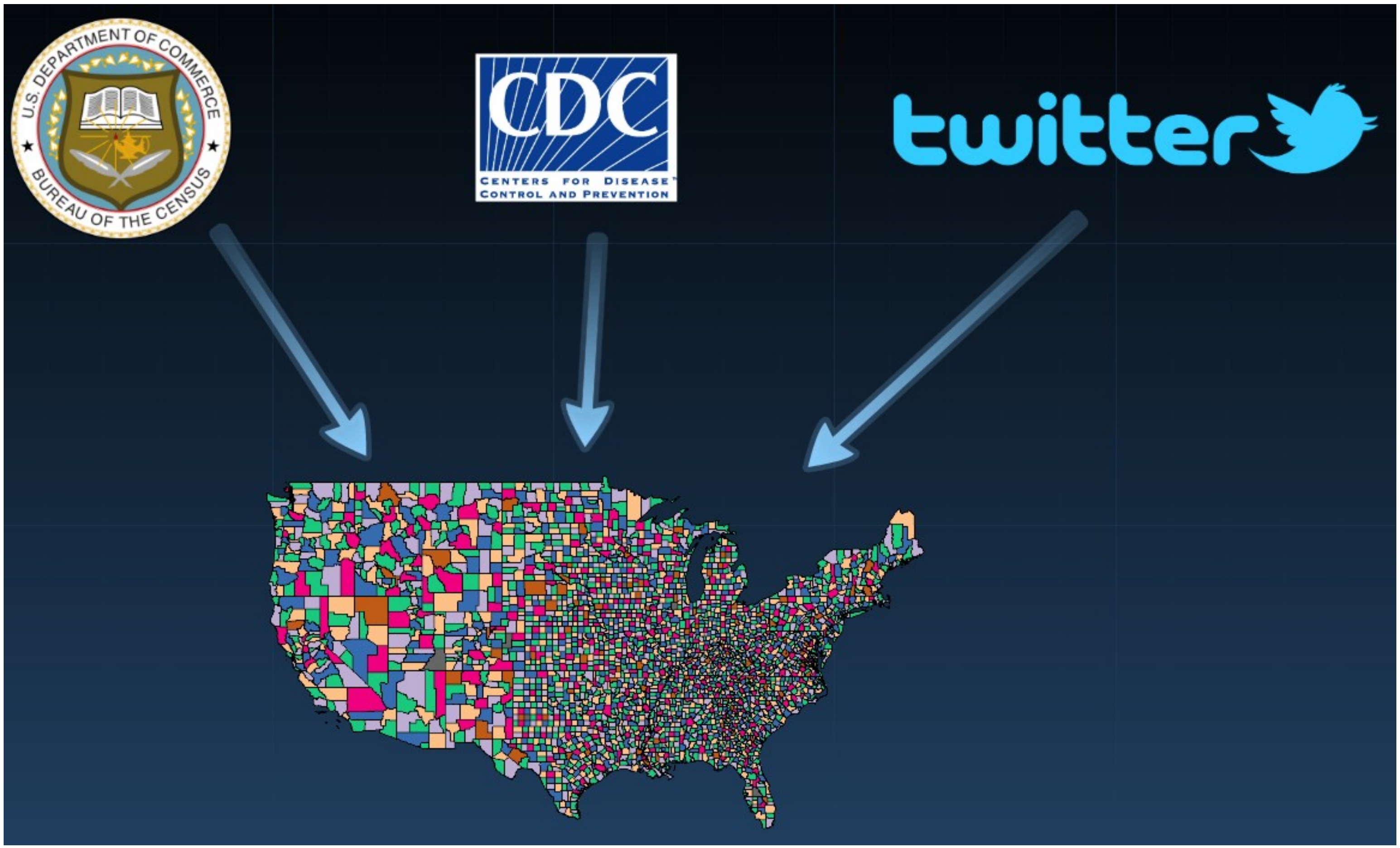
Overview

- Introduction
- **Background on Social Media Data**
  - **Sources**
  - **Types**
  - **Acquisition**
  - **Analysis Methodology**
- Examples
- Challenges
- Summary

Overview

- Introduction
- Background on Social Media Data
- **Examples**
  - **Heart Disease Mortality**
  - **HIV Prevalence**
  - **Life Satisfaction**
  - **Flu Tracking**
- Challenges
- Summary

Example: Community Heart Disease Mortality



Eichstaedt, Schwartz, Park, Kern, … Ungar, Seligman. (2014; in press)

Example: Community Heart Disease Mortality

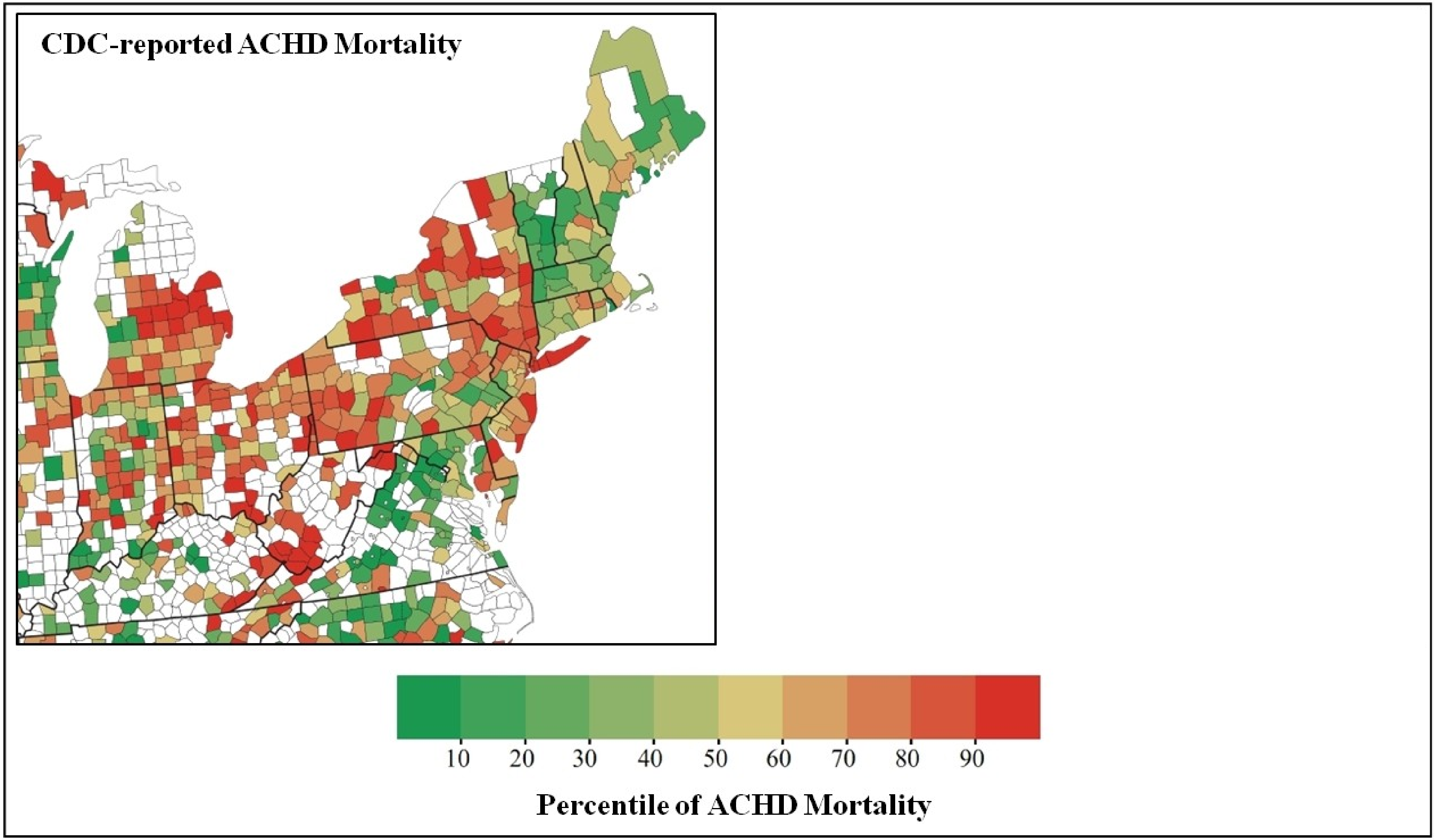
Twitter Dataset Studied:

10% of tweets from June 2009 to March 2010
(826 million tweets)

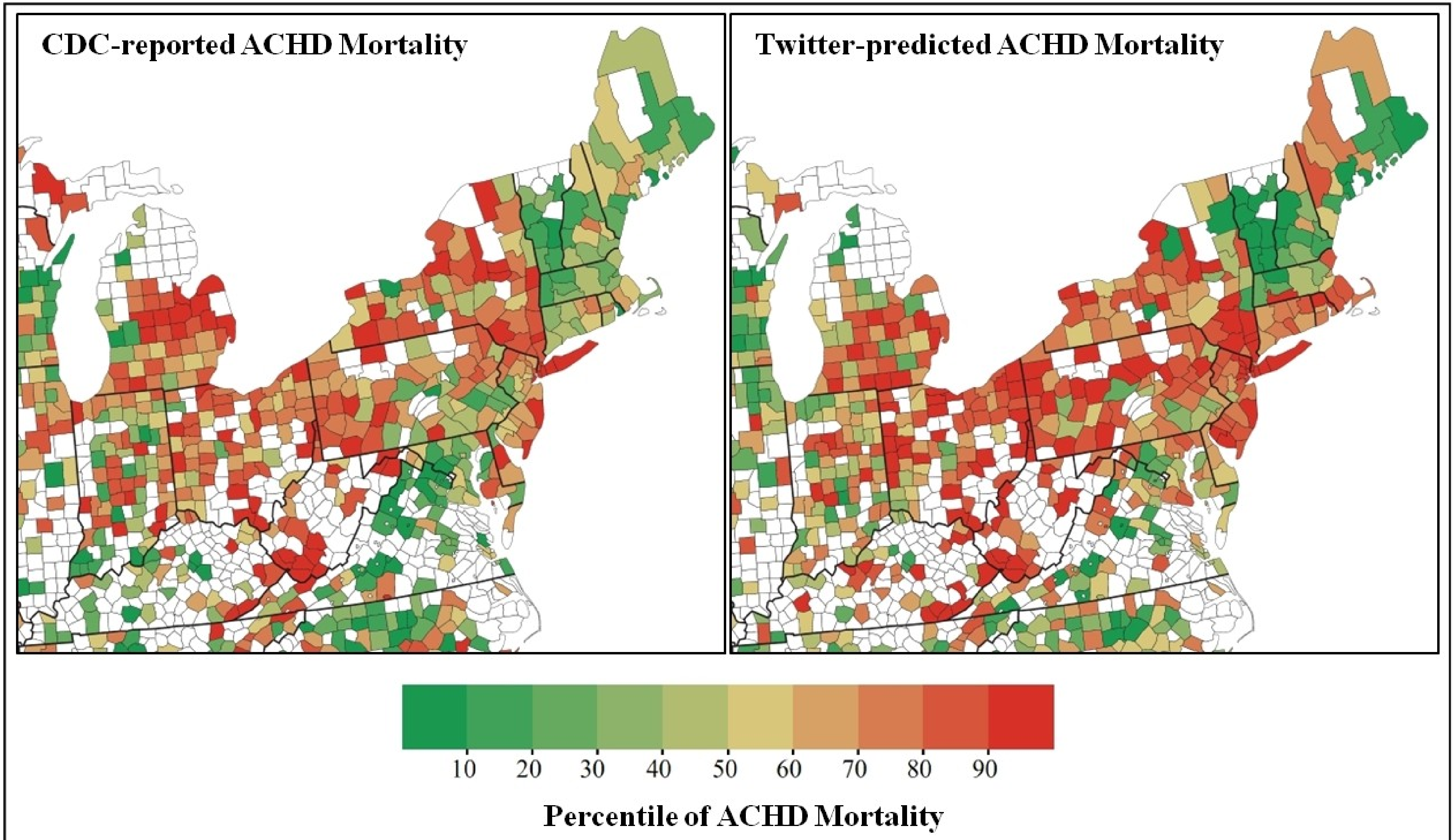
United States CDC data:

2009-2011 Atherosclerotic Heart Disease Mortality

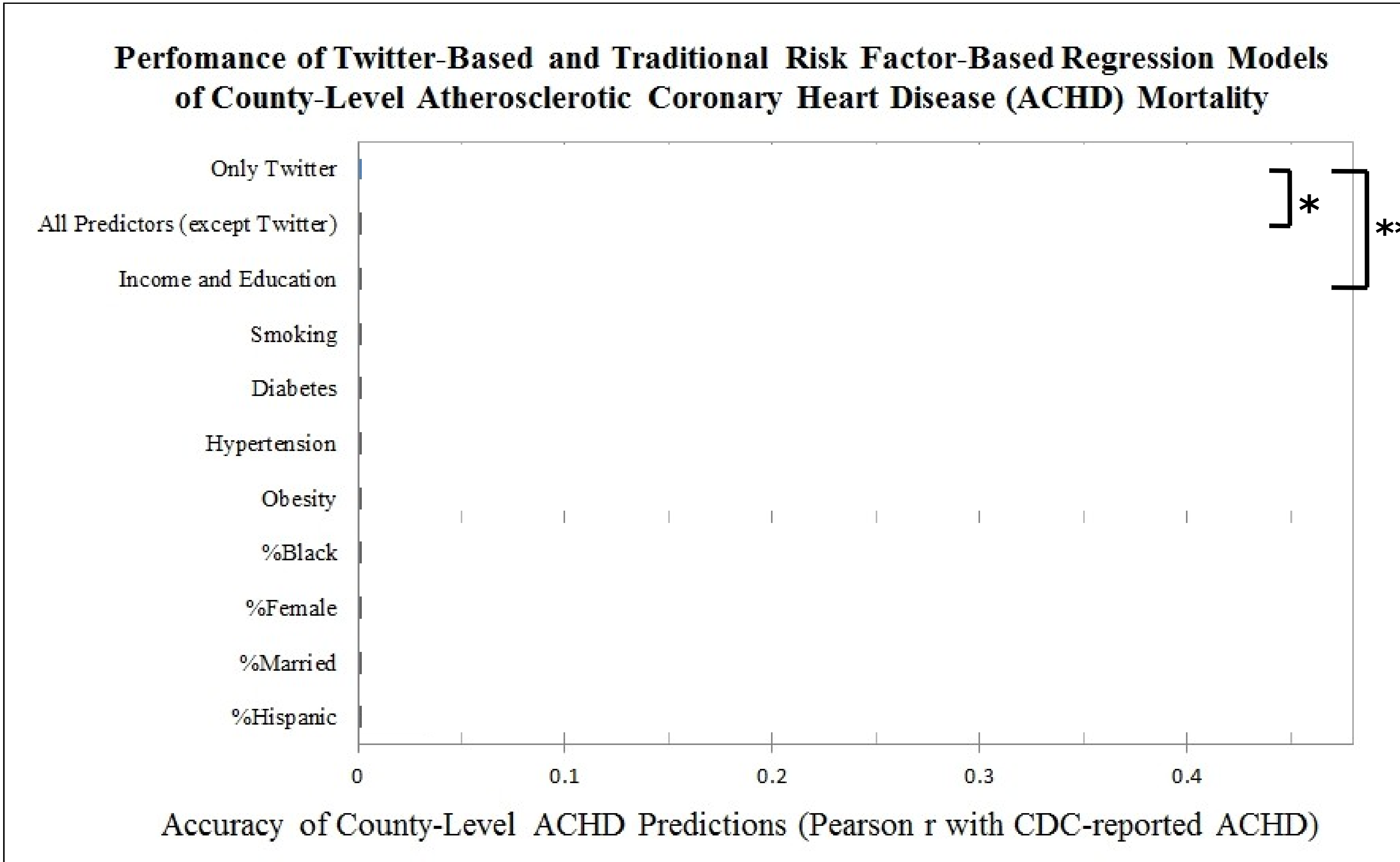
Example: Community Heart Disease Mortality



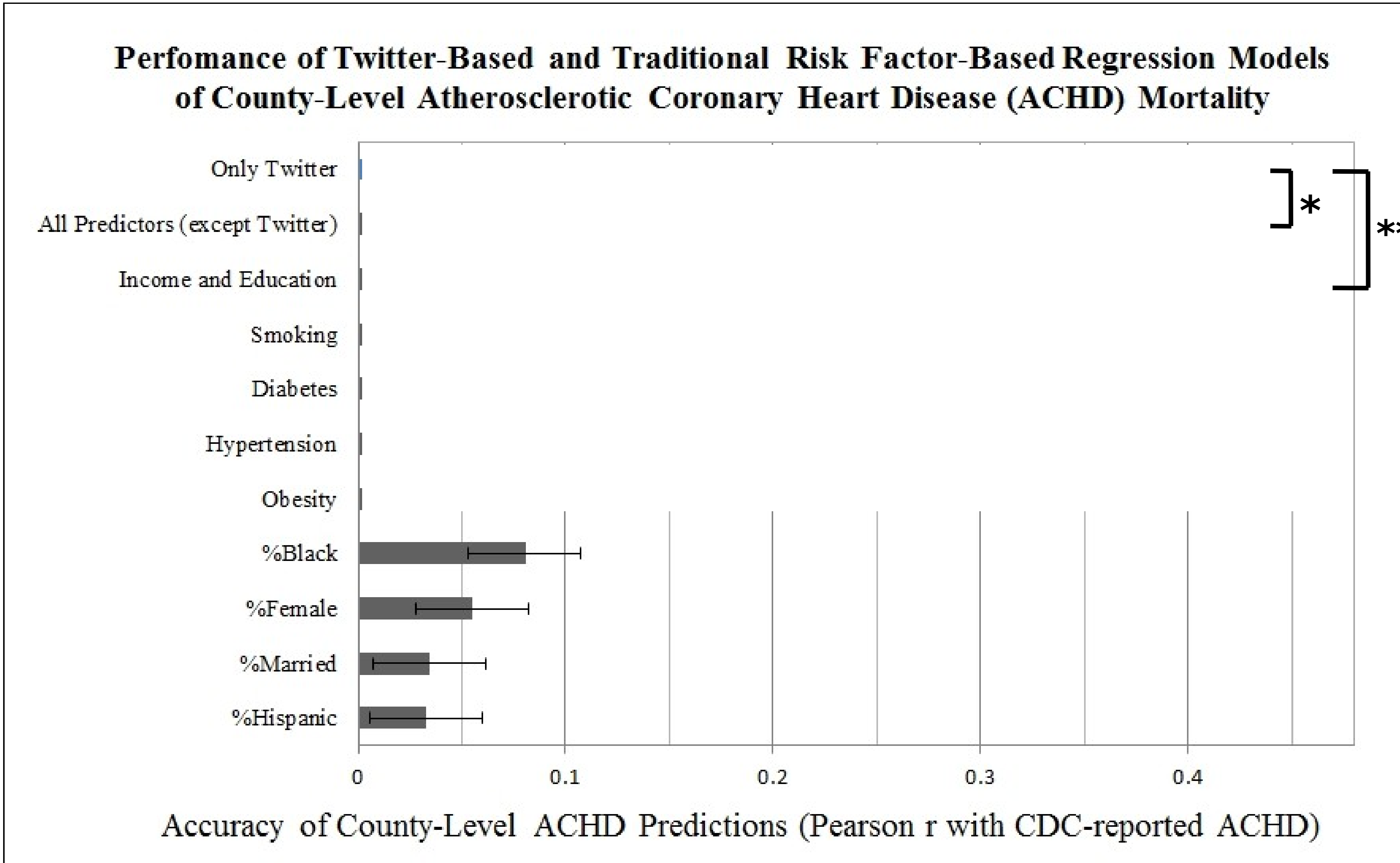
Example: Community Heart Disease Mortality



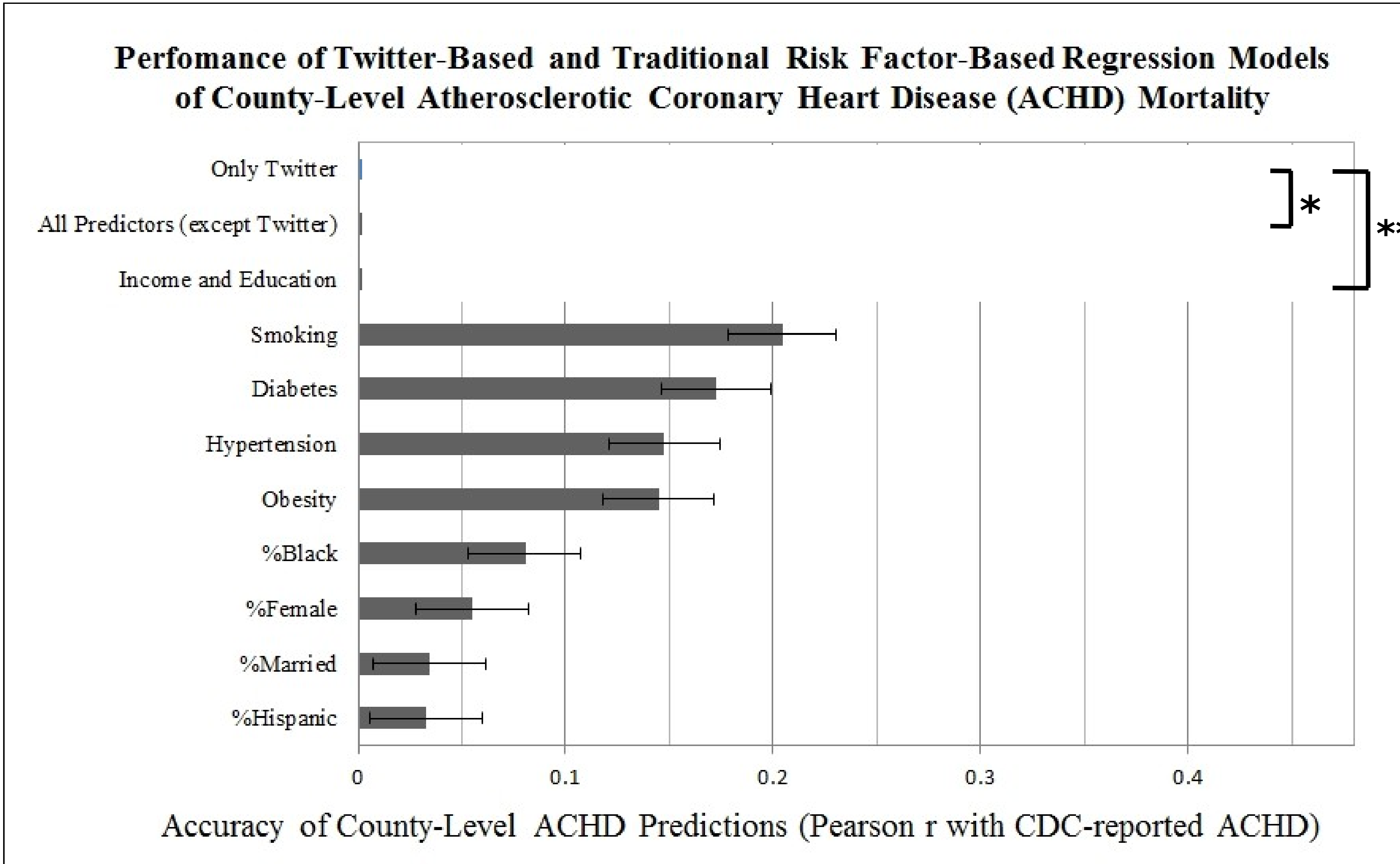
Example: Community Heart Disease Mortality



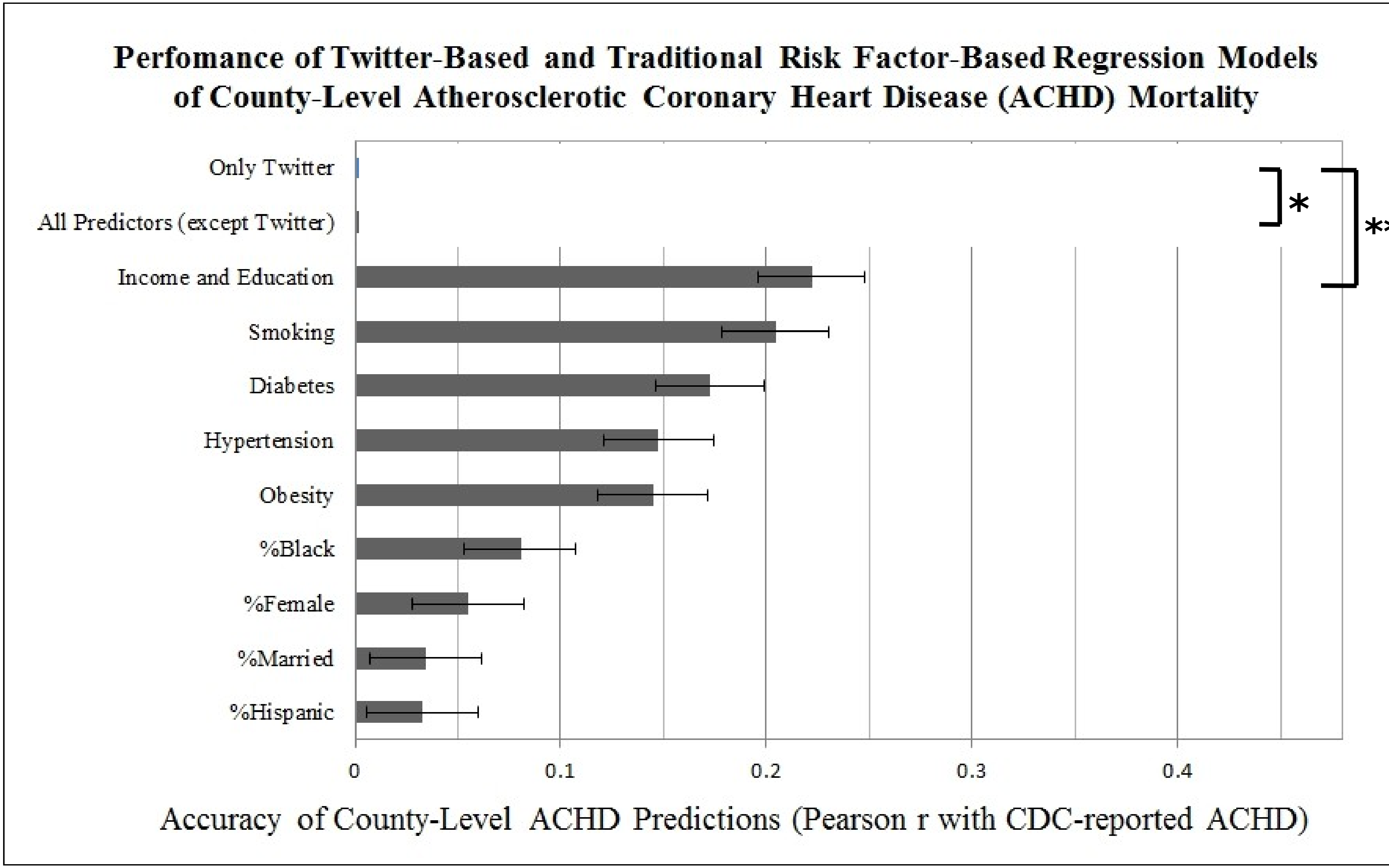
Example: Community Heart Disease Mortality



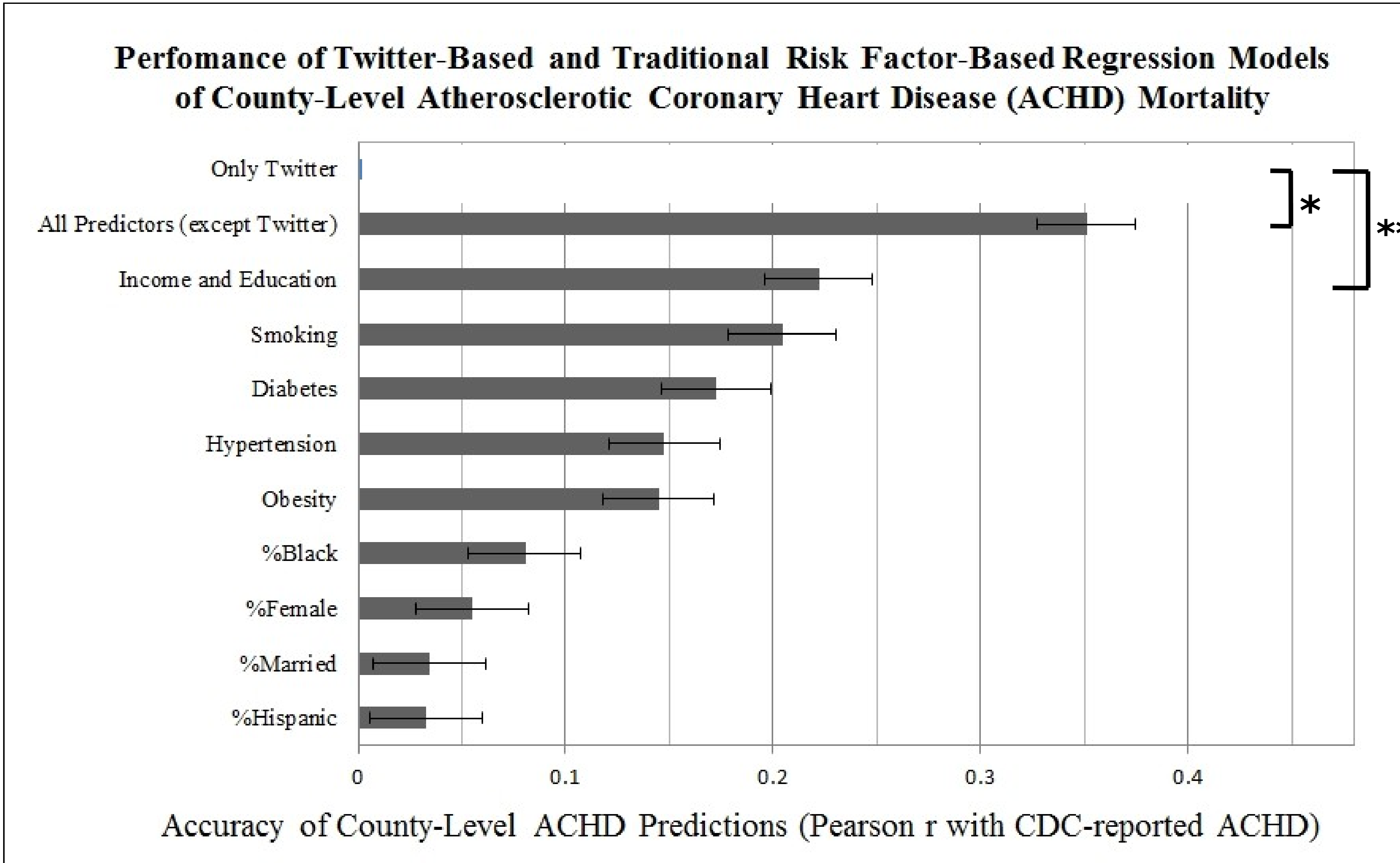
Example: Community Heart Disease Mortality



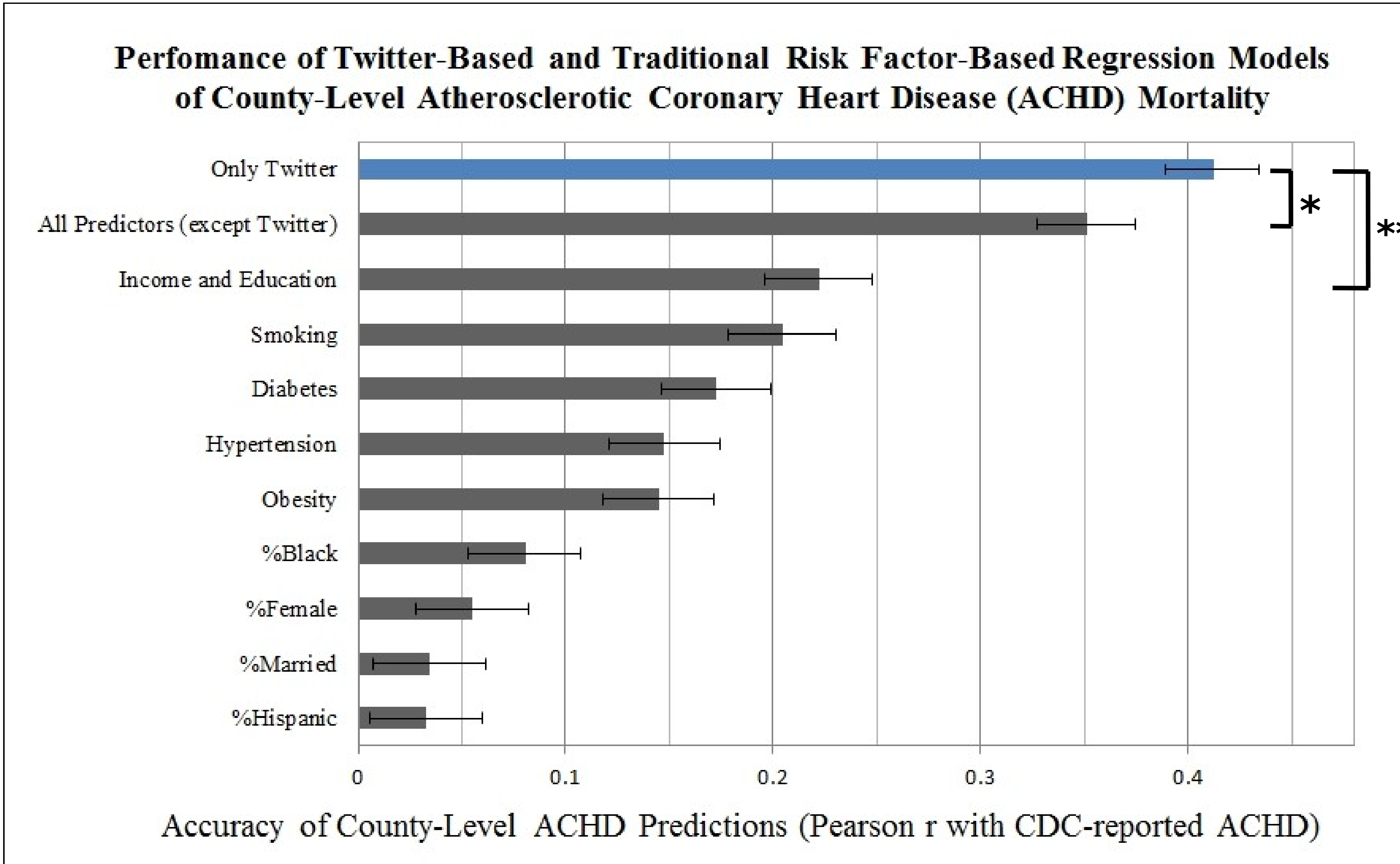
Example: Community Heart Disease Mortality



Example: Community Heart Disease Mortality



Example: Community Heart Disease Mortality



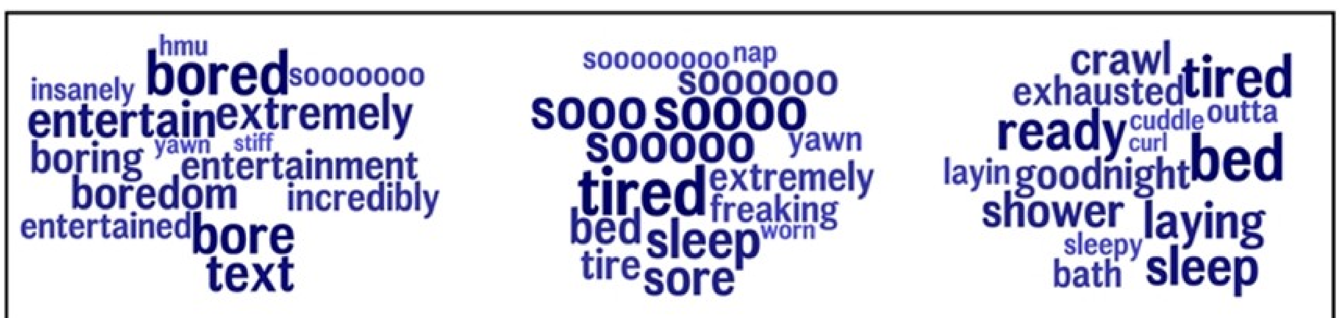
Language **positively** correlated with US-county-level Heart Disease

Anger, Hostility, Aggression

RR=1.43 to *RR*=1.74

Negative Relationships

RR=1.37 to *RR*=1.53



Disengagement

RR=1.43 to *RR*=1.49

Language **negatively** correlated with US-county-level Heart Disease

Higher Status Occupations

RR =.70 to *RR*=.75

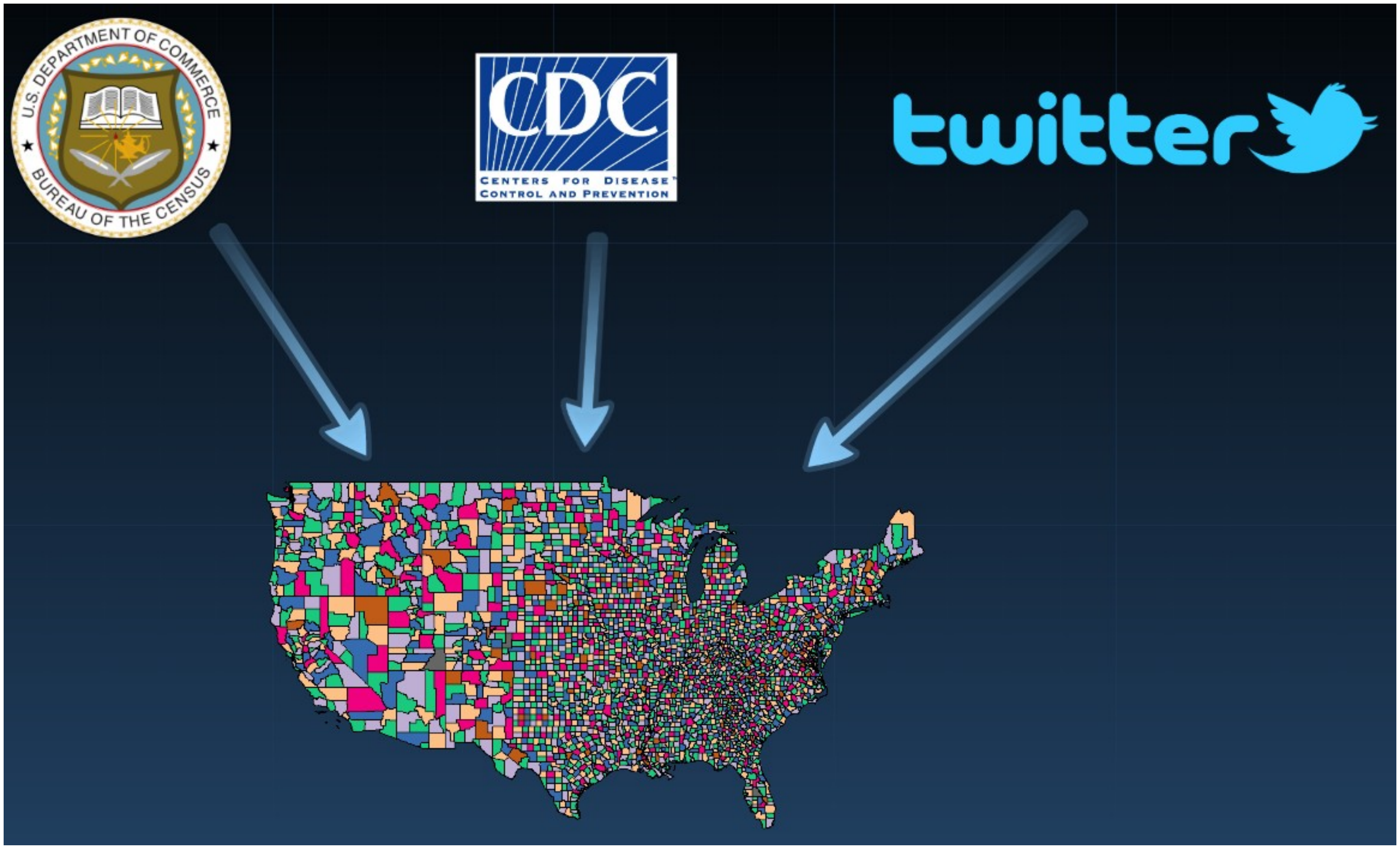
Positive Emotions, Engagement

RR=.73 to *RR*=.76

Optimism

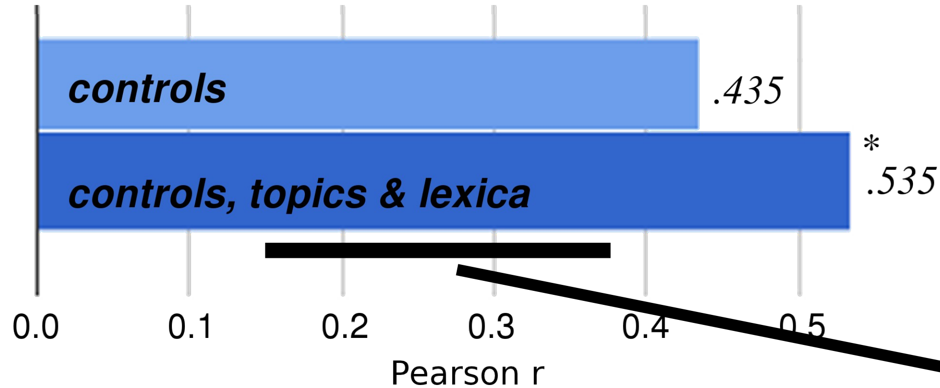
RR=.76 to *RR*=.78

Example: County Life Satisfaction



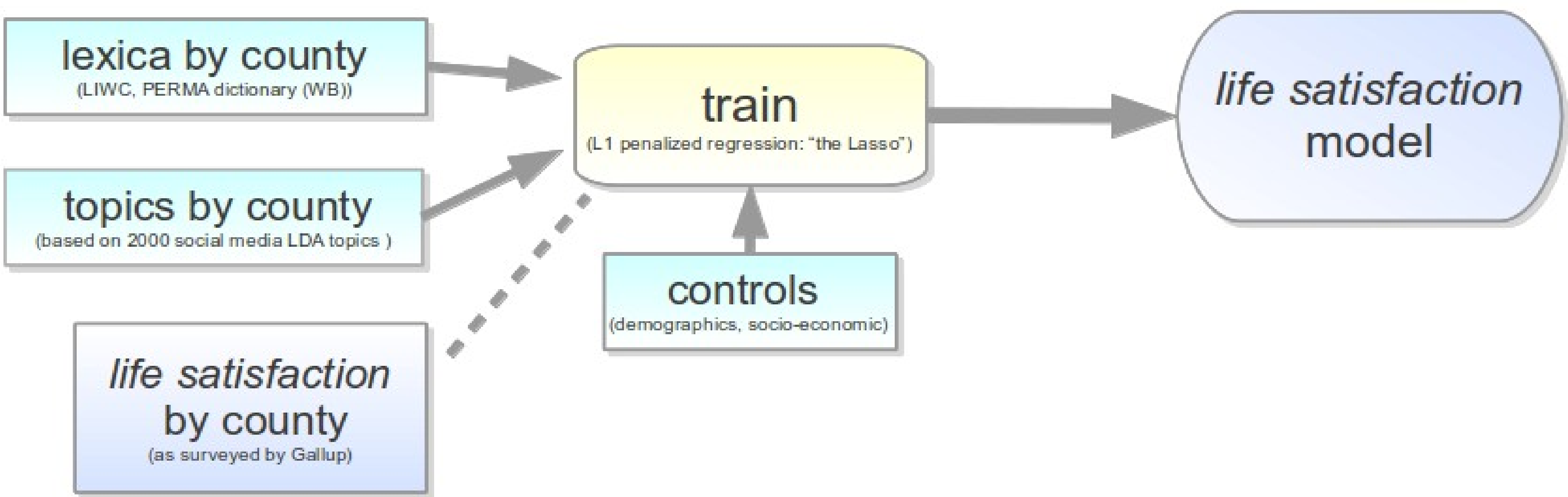
In collaboration with Molly Ireland and Dolores Albaraccin

Example: County Life Satisfaction

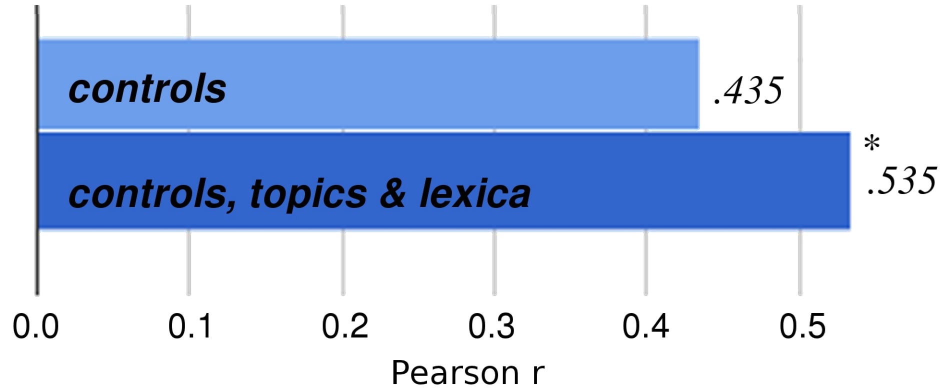


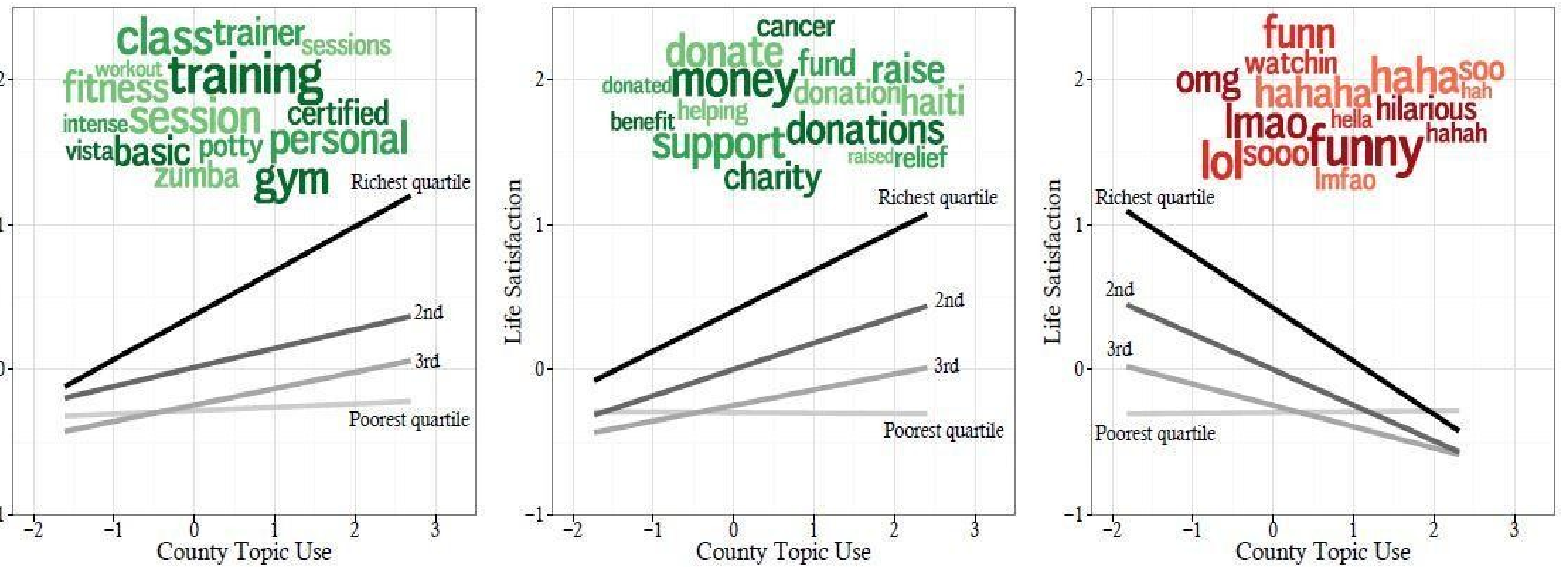
education level, income, demographics, ethnicity

Twitter

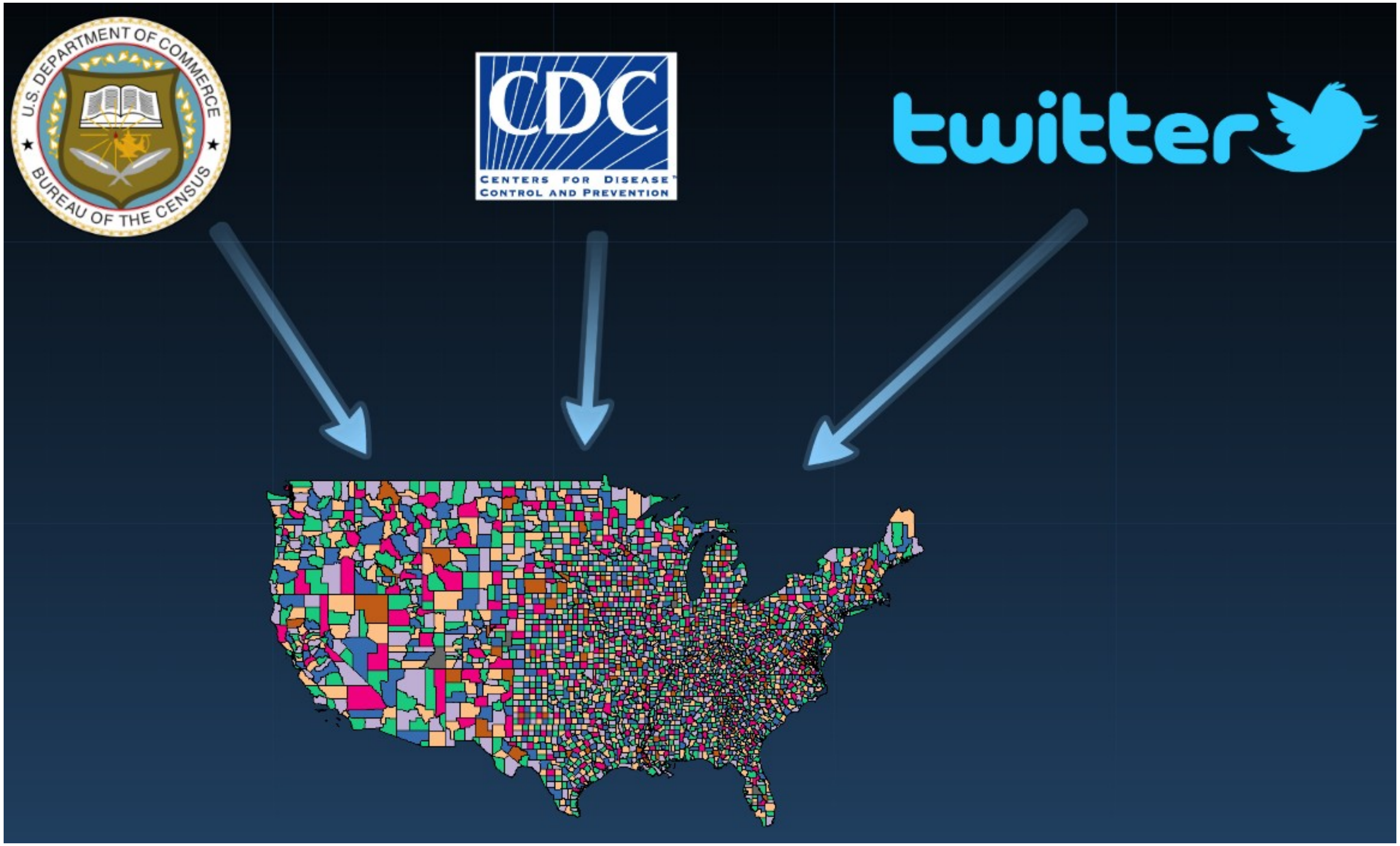


Example: County Life Satisfaction





Example: County HIV Prevalence



In collaboration with Molly Ireland and Dolores Albaraccin

Example: County HIV Prevalence

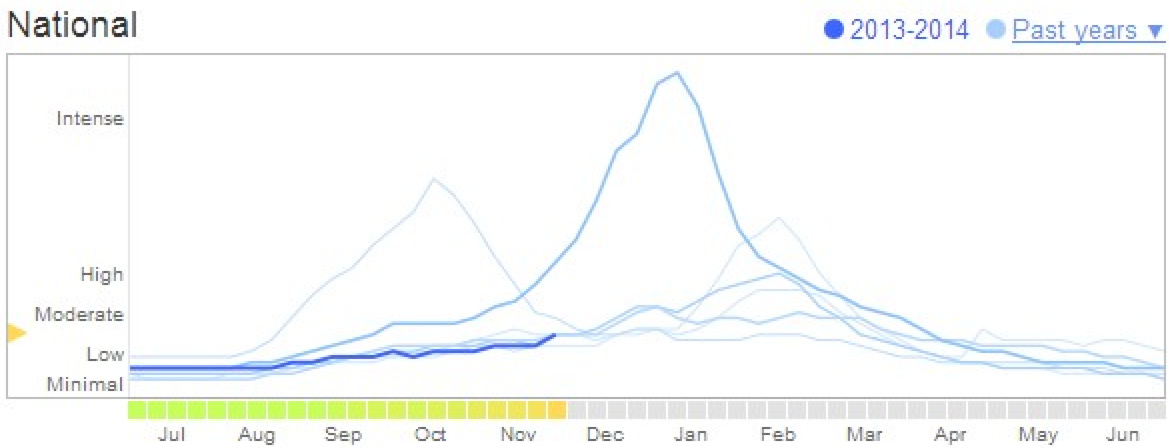
Example: County HIV Prevalence

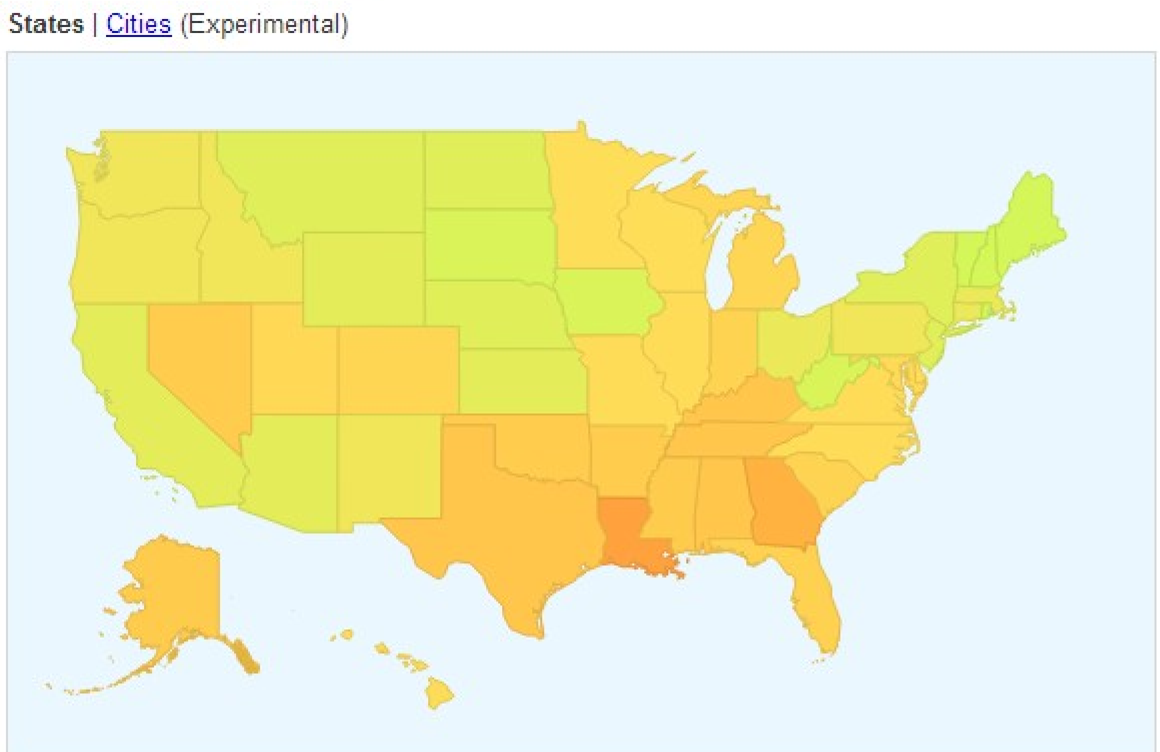
HIV prevalence is higher in counties with less future tense in...

all 1375 qualifying counties
(*Beta* = -0.48, *p* <.001)

top 200 most populated counties
(*Beta* = -0.27, *p* <.001)

Example: Flu Trends





Google Flu Trends

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue

Archive > Volume 494 > Issue 7436 > News > Article

NATURE | NEWS

عربي

When Google got flu wrong

US outbreak foxes a leading web-based method for tracking se

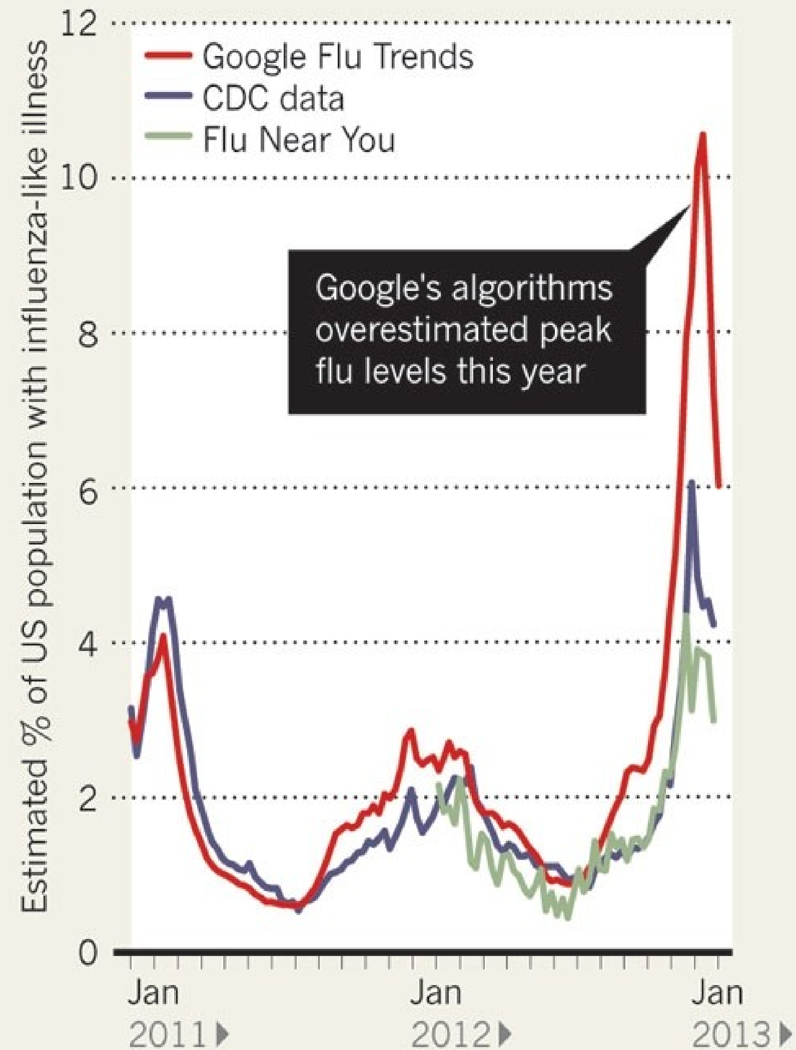
Declan Butler

13 February 2013

PDF | Rights & Permissions

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



Health Tweets

http://www.healthtweets.org/

(Mark Dredze and Michael Paul; Johns Hopkins University)

narrows in on health-related tweets

Overview

- Introduction
- Background on Social Media Data
- **Examples**
  - **Heart Disease Mortality**
  - **HIV Prevalence**
  - **Life Satisfaction**
  - **Flu Tracking**
- Challenges
- Summary

Overview

- Introduction
- Background on Social Media Data
- Examples
- **Challenges**
- Summary

Challenges

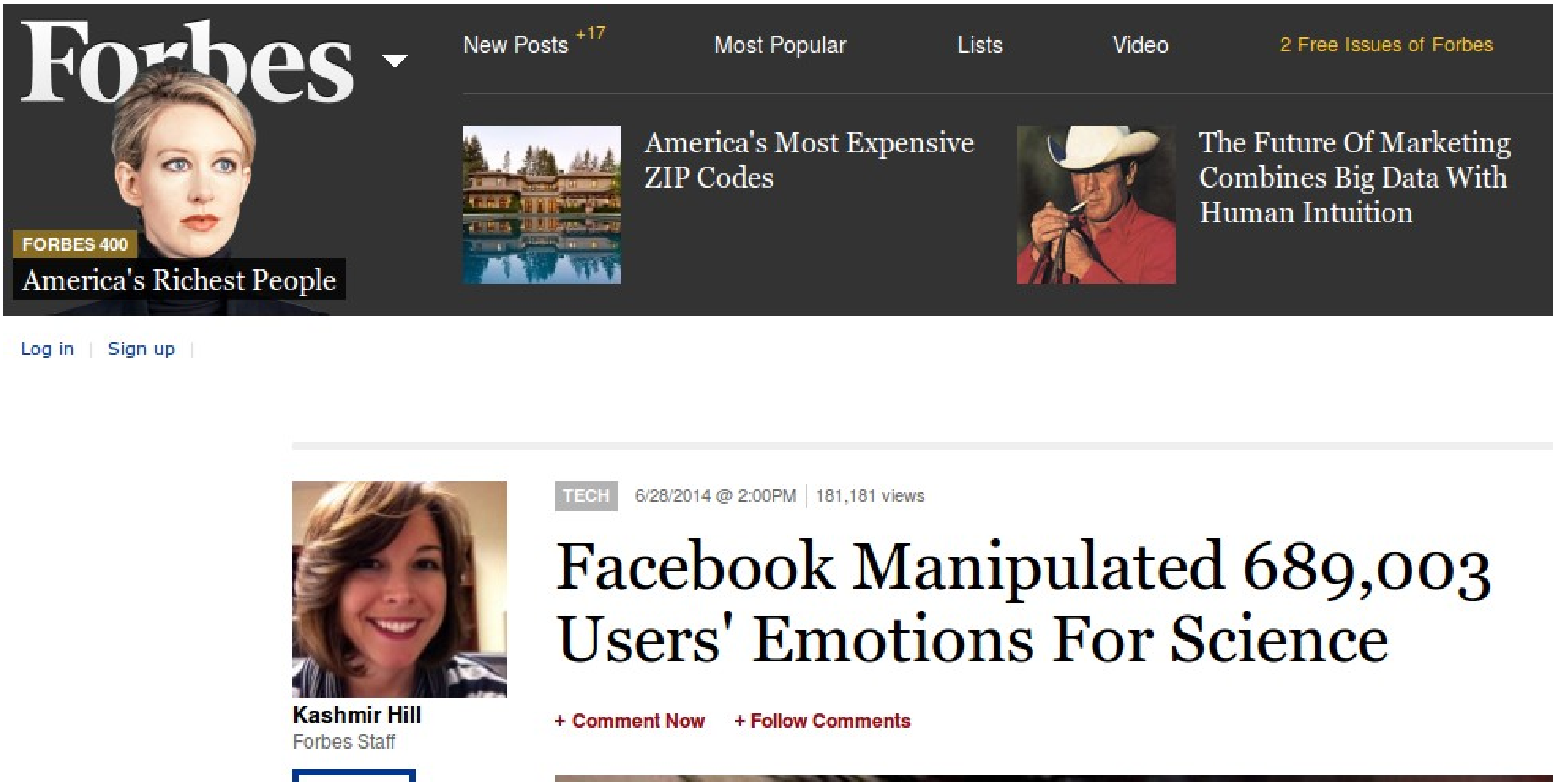
- Ethical / Privacy

- Technical

- Methodological

  –

Challenges


Forbes
New Posts +17
Most Popular
Lists
Video
2 Free Issues of Forbes
America's Most Expensive ZIP Codes
The Future Of Marketing Combines Big Data With Human Intuition
FORBES 400
America's Richest People
Log in | Sign up
TECH
6/28/2014 @ 2:00PM | 181,181 views
Facebook Manipulated 689,003 Users' Emotions For Science
Kashmir Hill
Forbes Staff
+ Comment Now
+ Follow Comments

Challenges

- **Ethical / Privacy**
  - Public Awareness / Participant Consent
- **Technical**
- **Methodological**

Challenges

- **Ethical / Privacy**
  - Public Awareness / Participant Consent
- **Technical**
  - Data Storage and Analysis Infrastructure
  - Evolving APIs
- **Methodological**

Challenges

- **Ethical / Privacy**
  - Public Awareness / Participant Consent
- **Technical**
  - Data Storage and Analysis Infrastructure
  - Evolving APIs
- **Methodological**
  - Word meaning / domains
  - Correlation versus Causation
  - Sample Bias
  - Self-presentation Bias

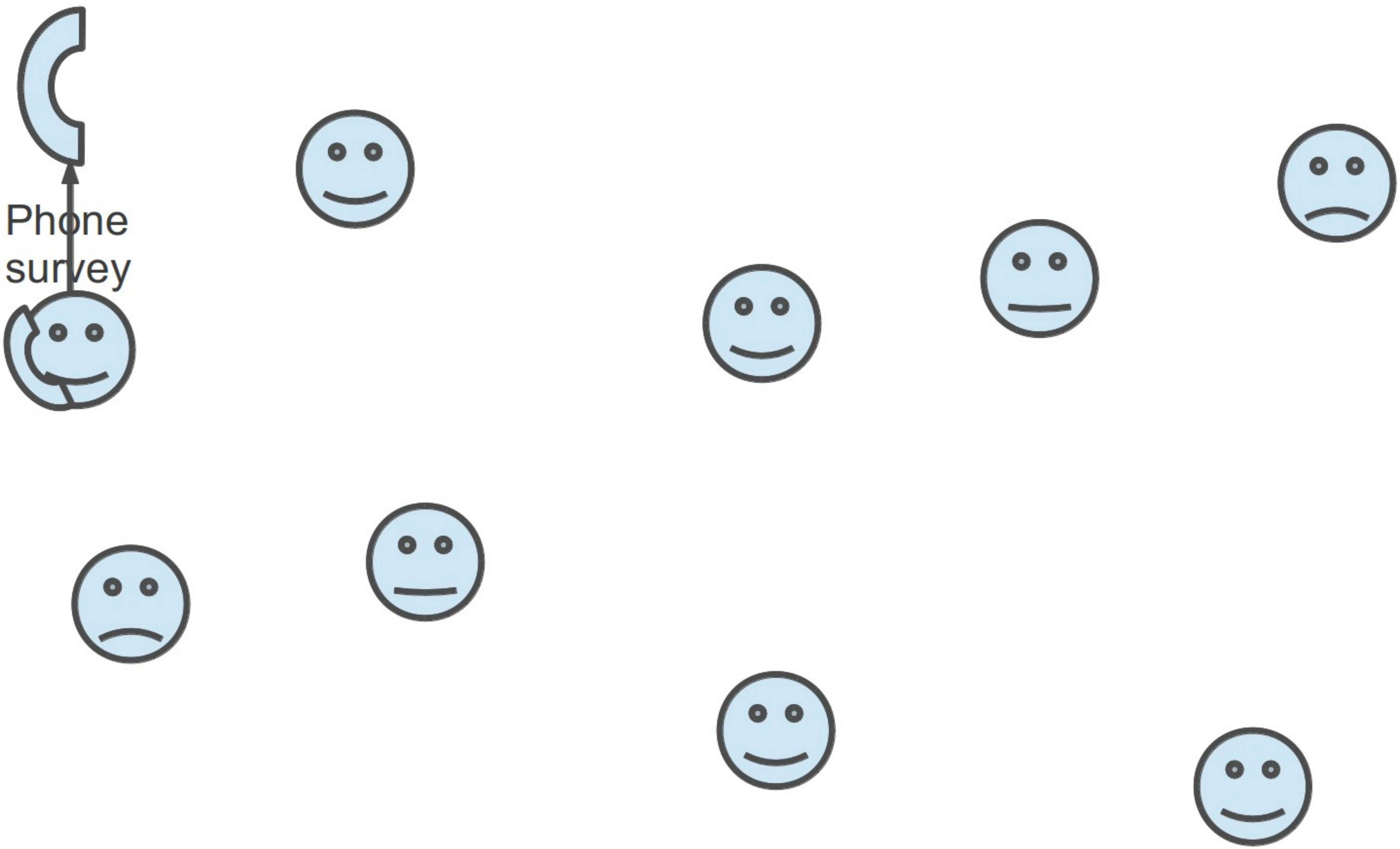
Issues attributed to missclassification Facebook status update.

category	label	frequency
Lexical Ambiguity	Wrong POS	15
	Wrong WS	38
Signal Negation	Strict Negation	16
	Desiring	6
Other	Stem Issue	5
	Other	24

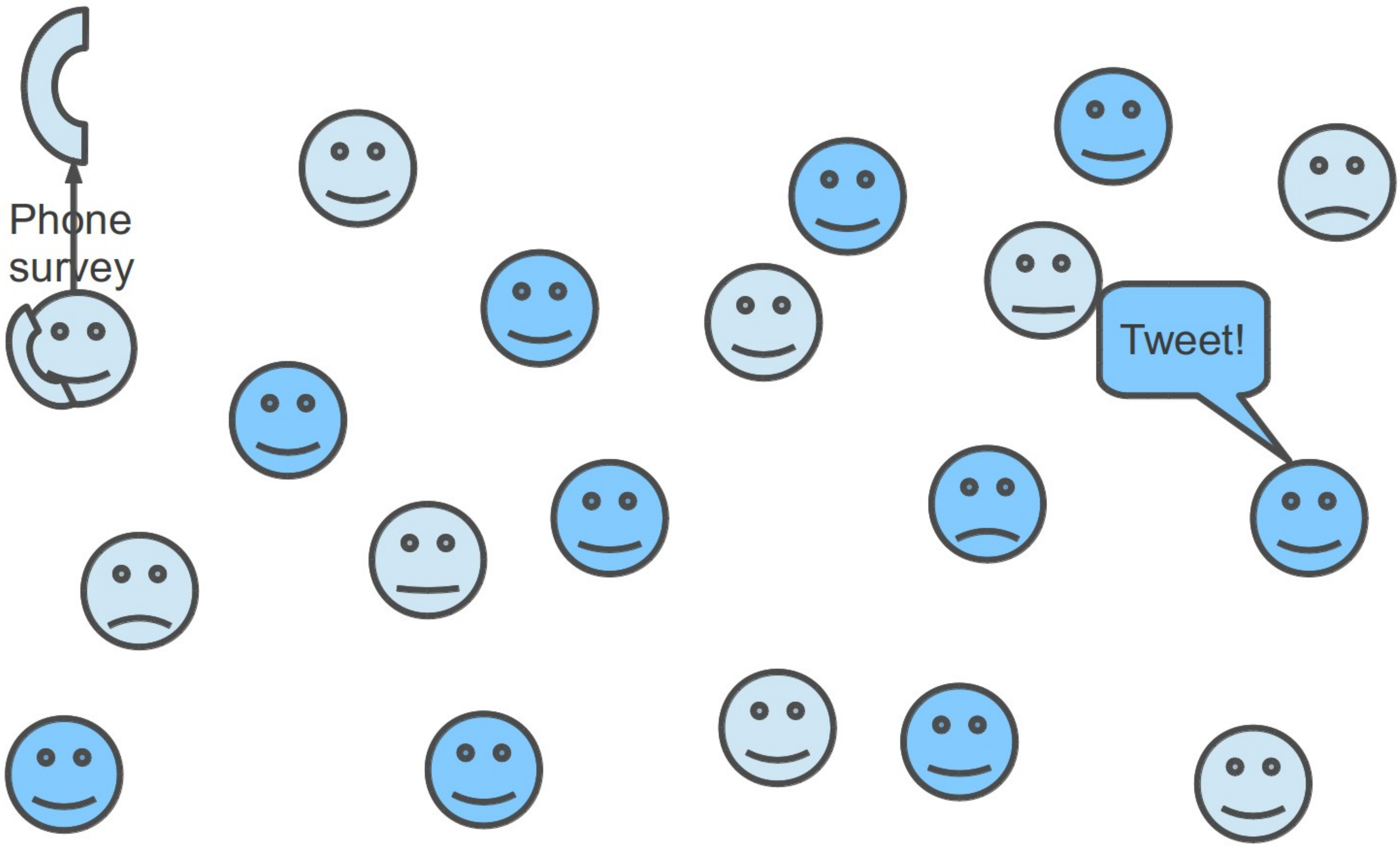
Challenges

- **Ethical / Privacy**
  - Public Awareness / Participant Consent
- **Technical**
  - Data Storage and Analysis Infrastructure
  - Evolving APIs
- **Methodological**
  - Word meaning / domains
  - Correlation versus Causation
  - Sample Bias
  - Self-presentation Bias

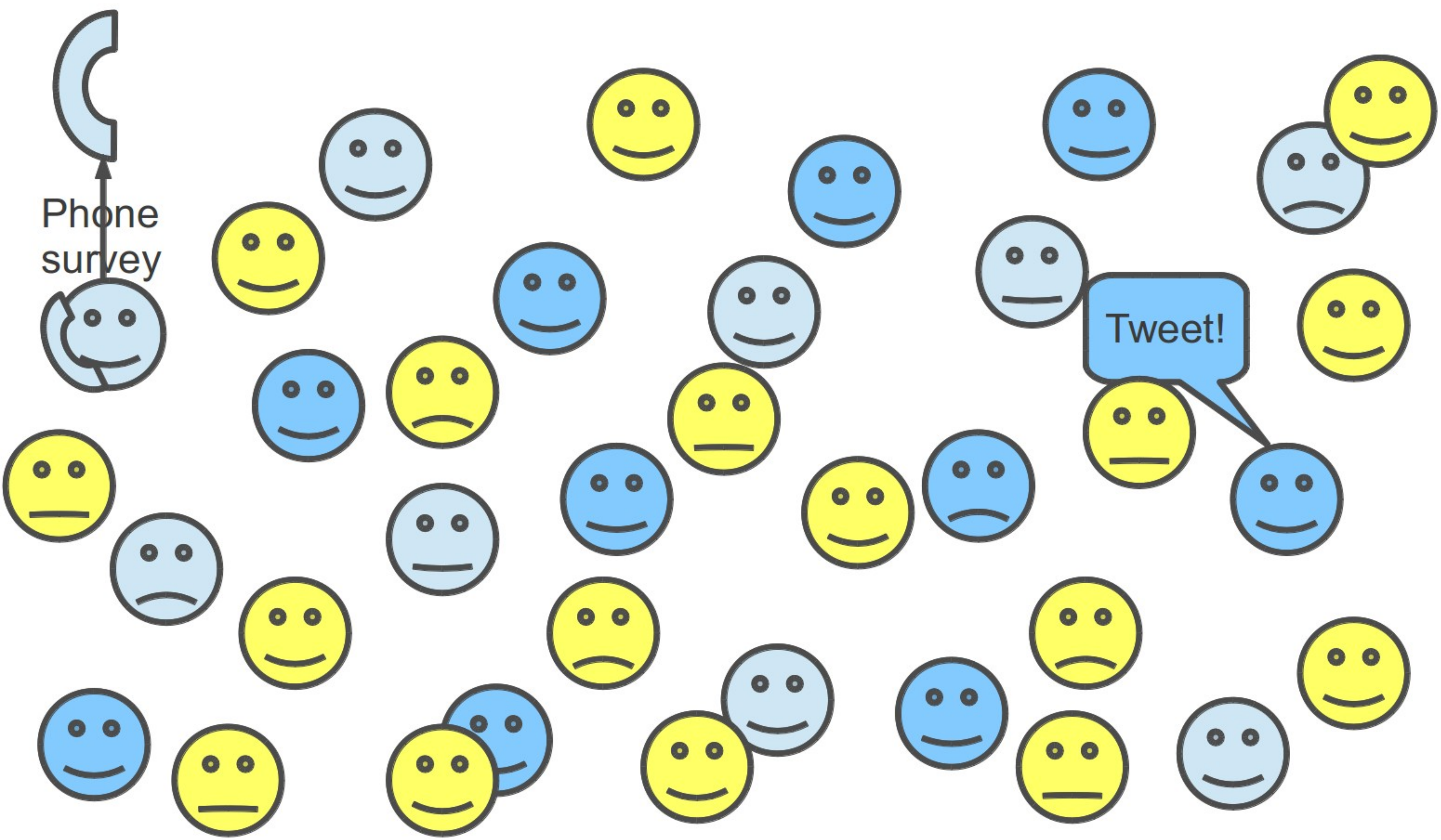
Predicting based on a different sample



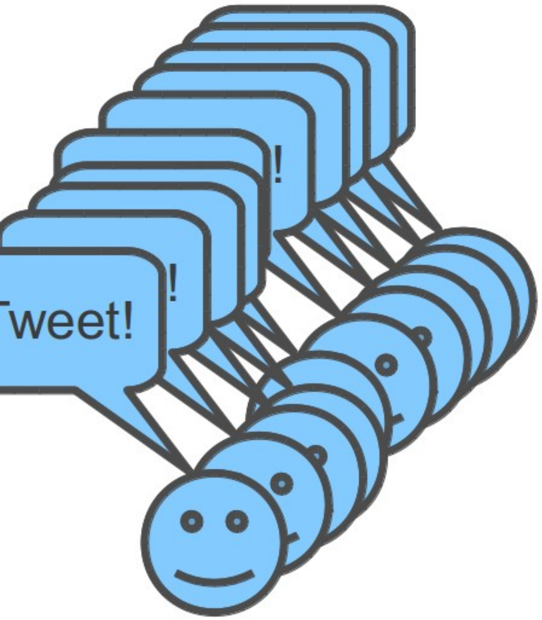
Predicting based on a different sample

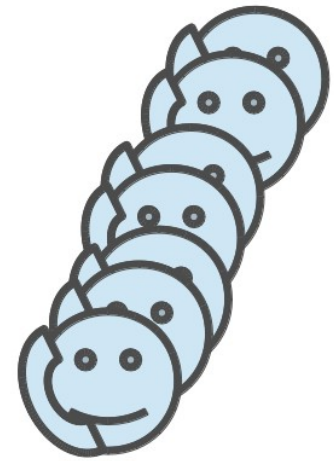


Predicting based on a different sample



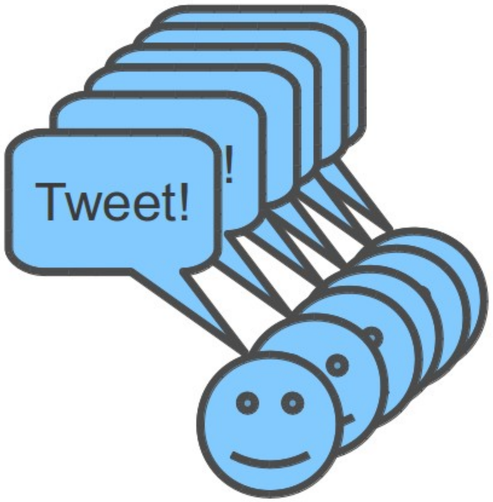
Representative Sample?

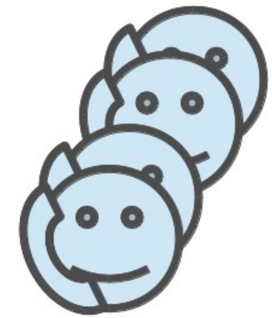




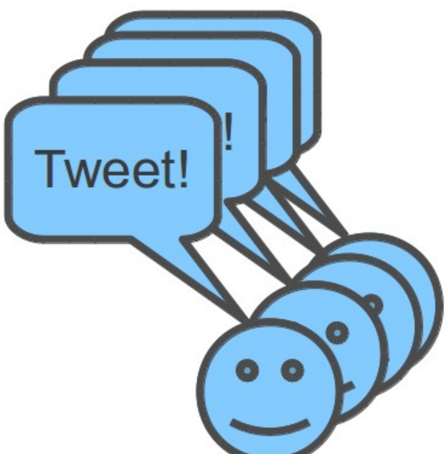
Surveyed well-being
from
representative sample.

Representative Sample?

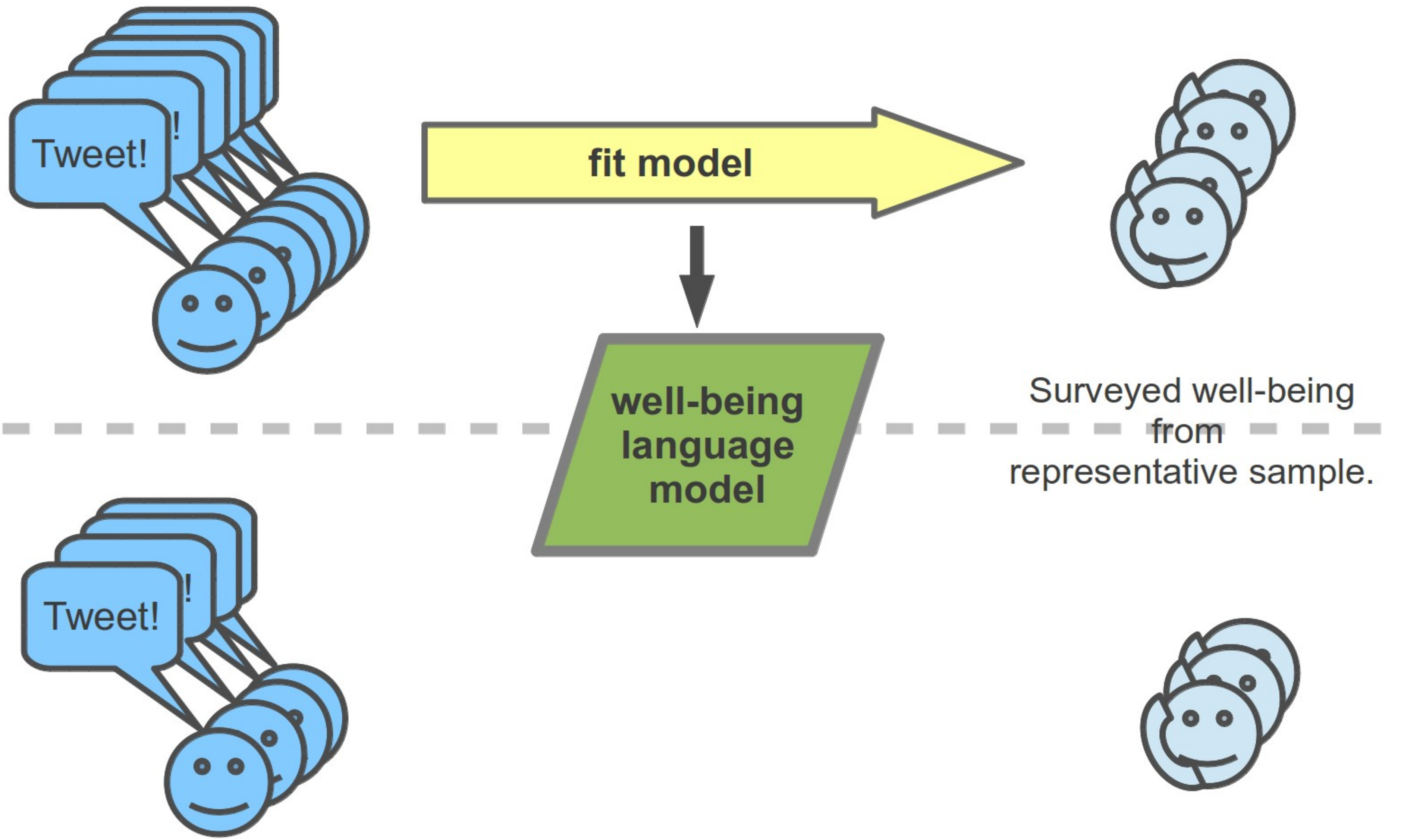




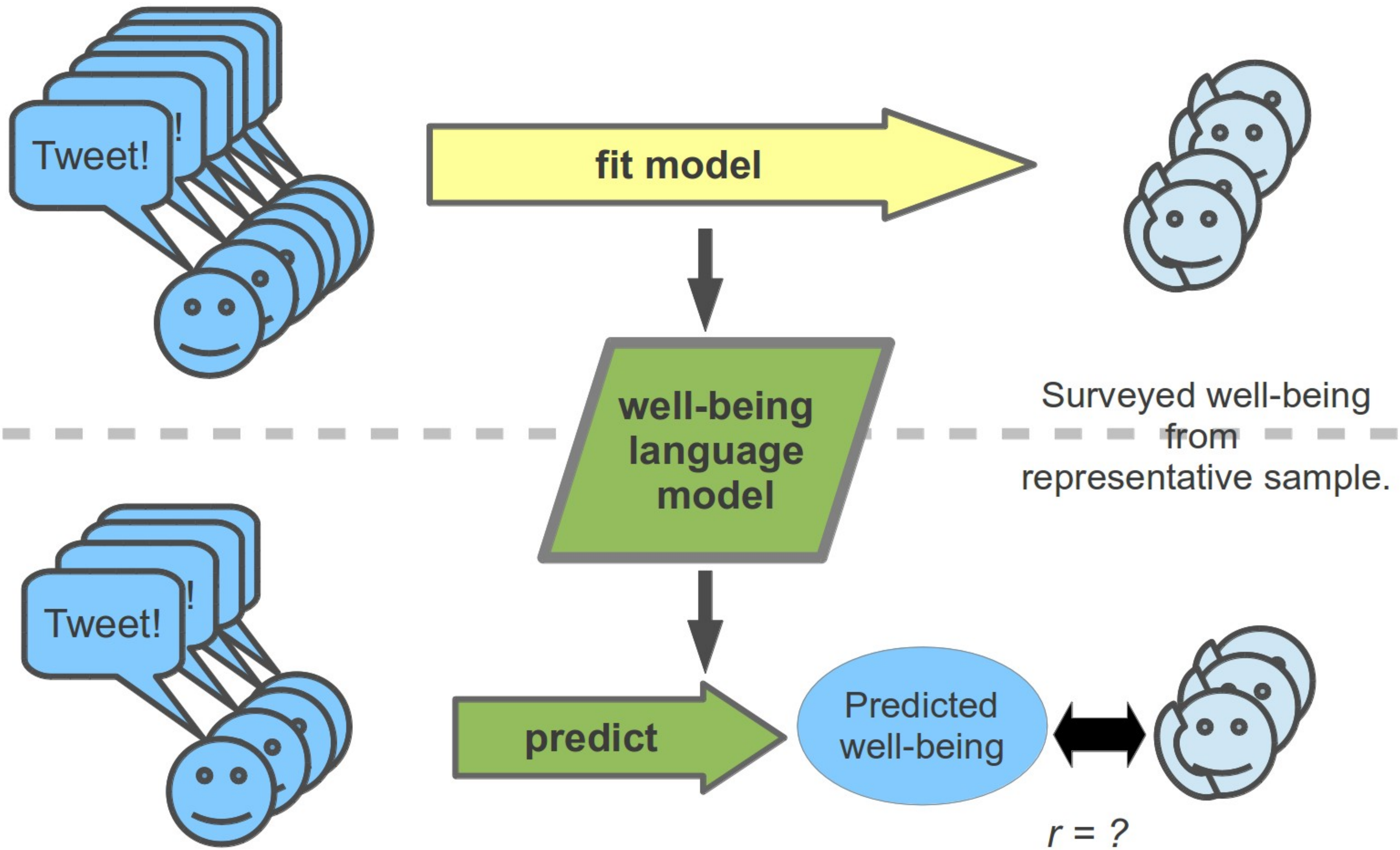
Surveyed well-being
from
representative sample.



Representative Sample?



Representative Sample?



Representative Sample?

- **Alternaitve: Post-stratification**
  - Demographics are one of the most accurately predicted from language
    - gender 92% accuracy
    - age 0.86 correlation

Challenges

- **Ethical / Privacy**
  - Public Awareness / Participant Consent
- **Technical**
  - Data Storage and Analysis Infrastructure
  - Evolving APIs
- **Methodological**
  - Word meaning / domains
  - Correlation versus Causation
  - Sample Bias
  - Self-presentation Bias

Challenges

- **Ethical / Privacy**
  - Public Awareness / Participant Consent
- **Technical**
  - Data Storage and Analysis Infrastructure
  - Evolving APIs
- **Methodological**
  - Word meaning / domains
  - Correlation versus Causation
  - Sample Bias
  - Self-presentation Bias

validate

writing size

sample size / populations

(Gosling 2004; 2010)

self-descriptive variables

Why Social Media and Language?

unobtrusive

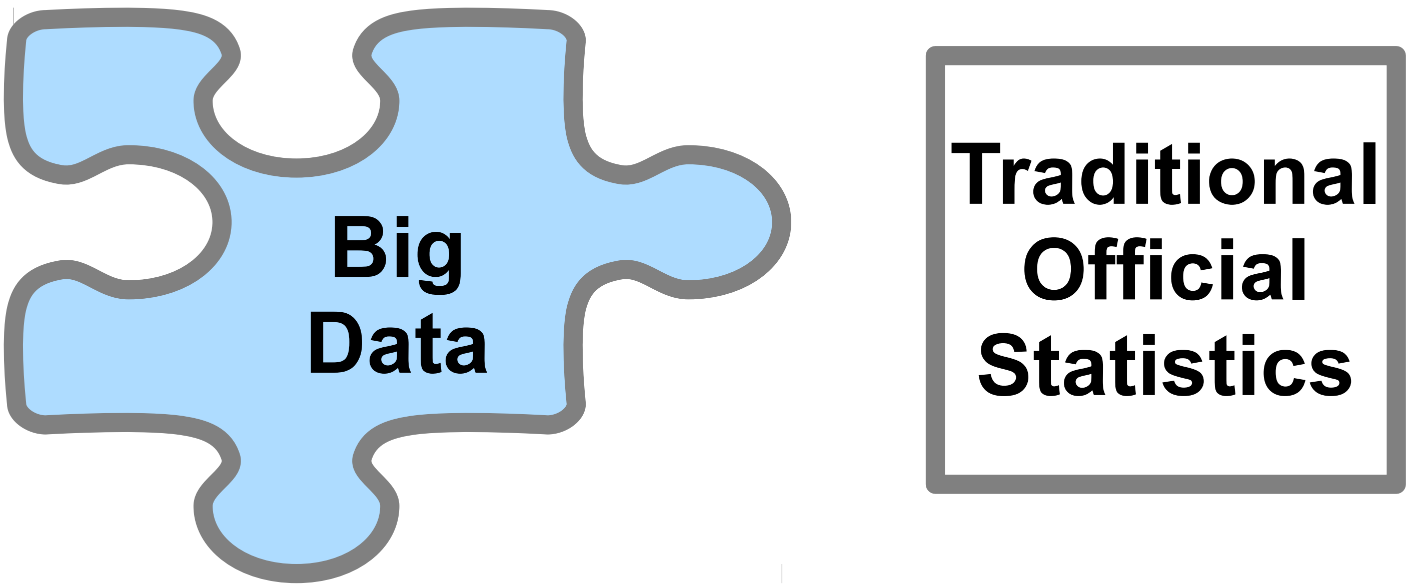
longitudinal / look back in time

often personal /
everyday concerns

potential for real-time

Thank You! Questions?

hansens@seas.upenn.edu



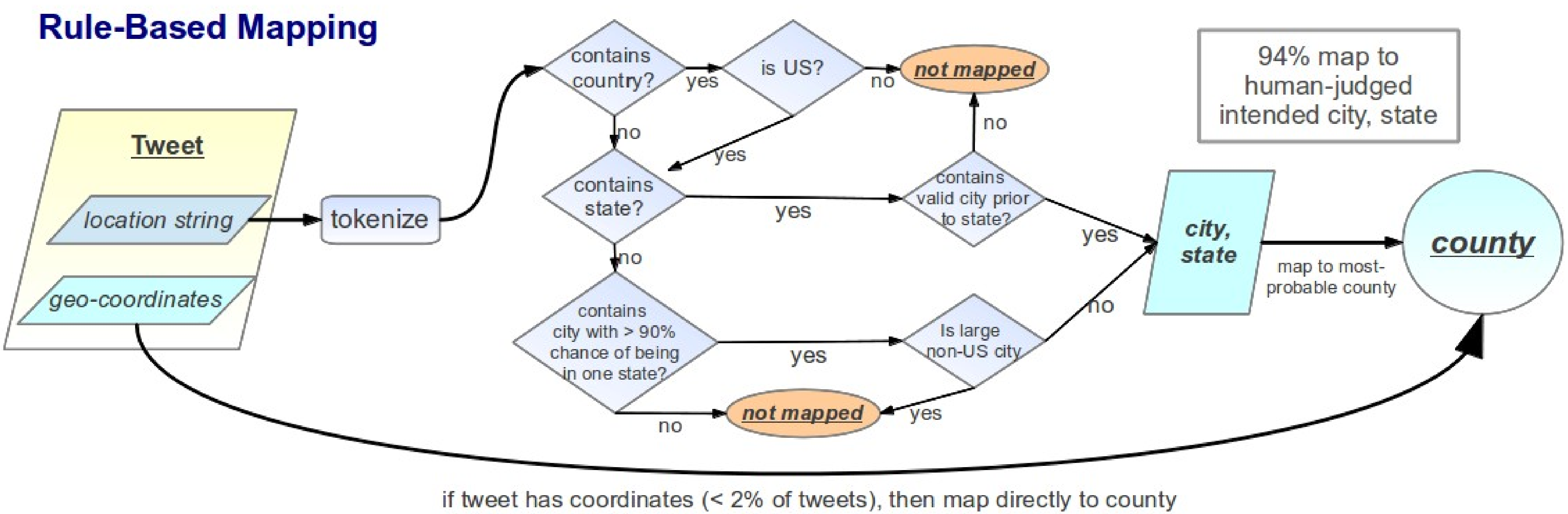
Thank You! Questions?

hansens@seas.upenn.edu

Big Data for Official Statistics

APPENDIX

Method: County-Mapping



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

Distributed Computing

- approximately 1 billion tweets
  - Too much for single computer system
- Utilize map-reduce in a “Hadoop” style cluster:

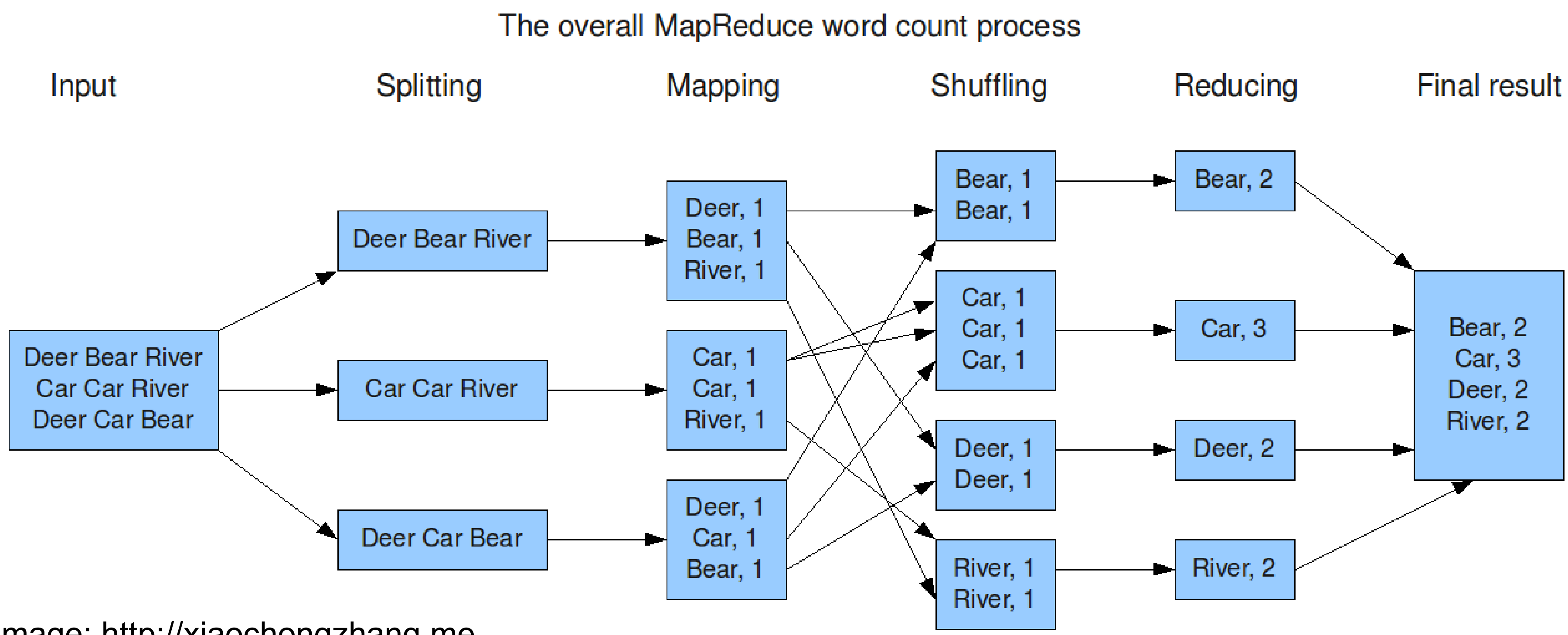
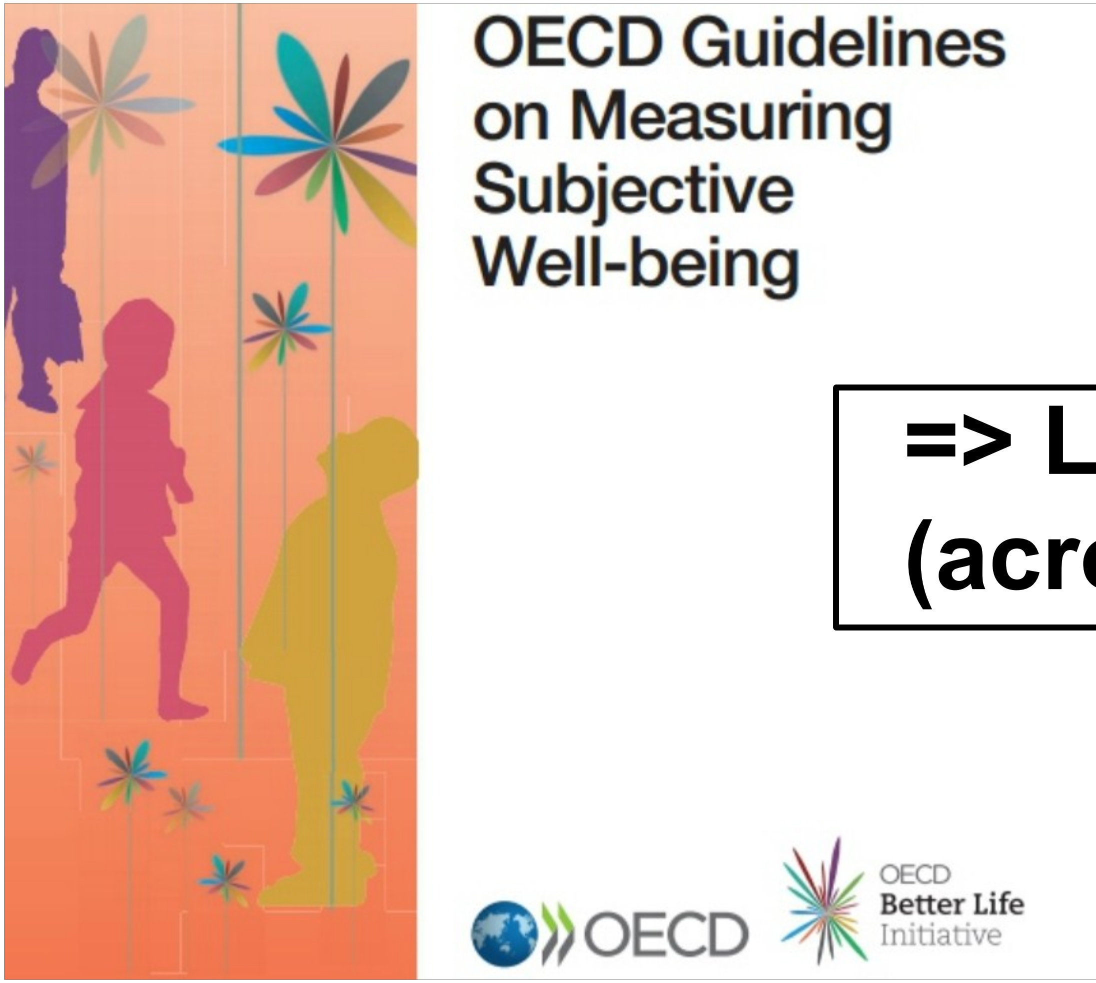


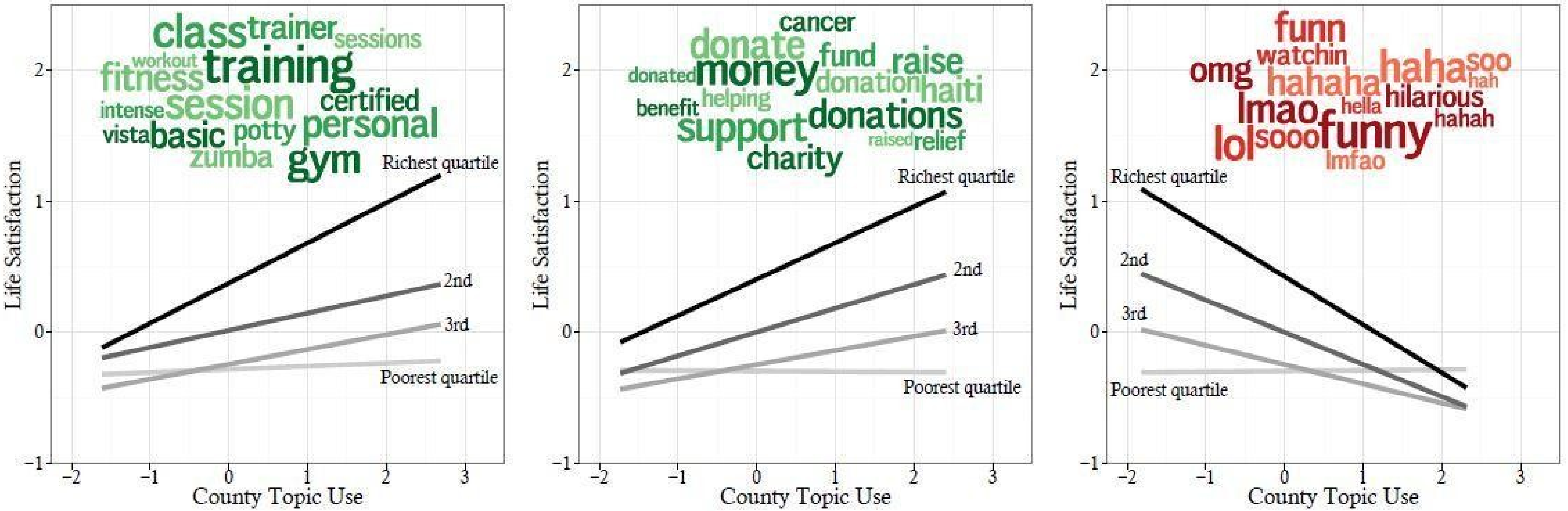
image: http://xiaochongzhang.me

Well-Being and Policy



=> Life Satisfaction (across domains)

What topics matter for all counties (that we have data for) in the United



Evidence for moderation

- A moderator alters the strength or direction of a relationship
- Question of external validity – how universal is the effect?

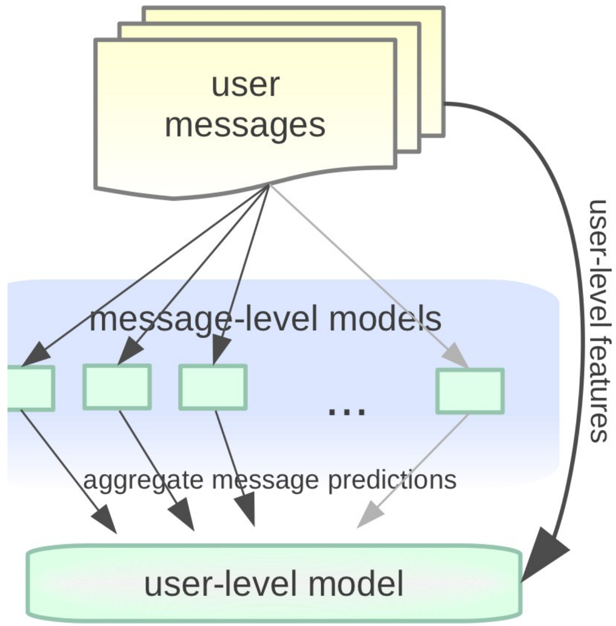
Daivd Kenny – Moderator Variables: Introduction, http://davidakenny.net/cm/moderation.htm

What topics matter for the poorest 25% of counties in

Individual Well-Being

satisfaction with life

Individual Well-Being: message to user-level



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

message and user-level

	Correlations between Personality Ratings		
	Self-Language	Self-Friend	Friend-Language
N =	5,000	745	745
Openness	.43	.25	.30
Conscientiousness	.37	.30	.20
Extraversion	.42	.39	.24
Agreeableness	.34	.30	.24
Neuroticism	.35	.34	.20

Criterion measures	N	Extraversion Qx	Extraversion Language
Number of friends	711	.23	.22
Number of doctor visits	736	.05	.10
Number of sick days	733	-.01	.03
Politically liberal	756	.07	.03
Fair-mindedness	864	.24	.10
Self-disclosure	864	.15	.14
Self-monitoring	927	.36	.15
Satisfaction with life	1114	.24	.13
Barratt Impulsiveness Scale	549		
Attention		-.08	-.01
Cognitive Instability		-.02	.00
Motor		.30	.15
Perseverance		-.01	.05
Self Control		.00	.06
Cognitive Complexity		.05	.09

Column-vector correlations	**.83**

Criterion measures	N	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
		Qx	Language	Qx	Language	Qx	Language	Qx	Language	Qx	Language
Number of friends	711	.05	-.05	-.01	-.15	.23	.22	.04	.03	-.13	-.09
Number of doctor visits	736	.00	-.01	-.05	.12	.05	.10	.02	.03	.14	.08
Number of sick days	733	.01	.07	-.07	-.01	-.01	.03	-.02	.02	.22	.11
Politically liberal	756	.32	.22	-.13	-.14	.07	.03	-.01	-.19	.05	.08
Fair-mindedness	864	.17	.03	.33	.23	.24	.10	.28	.17	-.35	-.19
Self-disclosure	864	-.02	-.07	.37	.29	.15	.14	.37	.28	-.28	-.16
Self-monitoring	927	.18	.08	-.03	-.09	.36	.15	-.03	-.01	-.10	-.05
Satisfaction with life	1114	.05	-.03	.29	.19	.24	.13	.24	.21	-.45	-.19
Barratt Impulsiveness Scale	549										
Attention		-.08	.03	-.42	-.15	-.08	-.01	-.18	-.17	.31	.13
Cognitive Instability		.24	.14	-.22	-.18	-.02	.00	-.15	-.17	.16	.09
Motor		.09	-.03	-.17	-.02	.30	.15	.06	-.07	-.04	-.02
Perseverance		.01	.00	.00	.01	-.01	.05	-.11	-.02	.09	-.04
Self Control		-.04	.04	-.47	-.12	.00	.06	-.09	-.10	.24	.07
Cognitive Complexity		-.03	-.04	-.23	-.03	.05	.09	-.01	-.07	.10	.05
Column-vector correlations		**.74**		**.82**		**.83**		**.89**		**.95**	

Representative Sample?

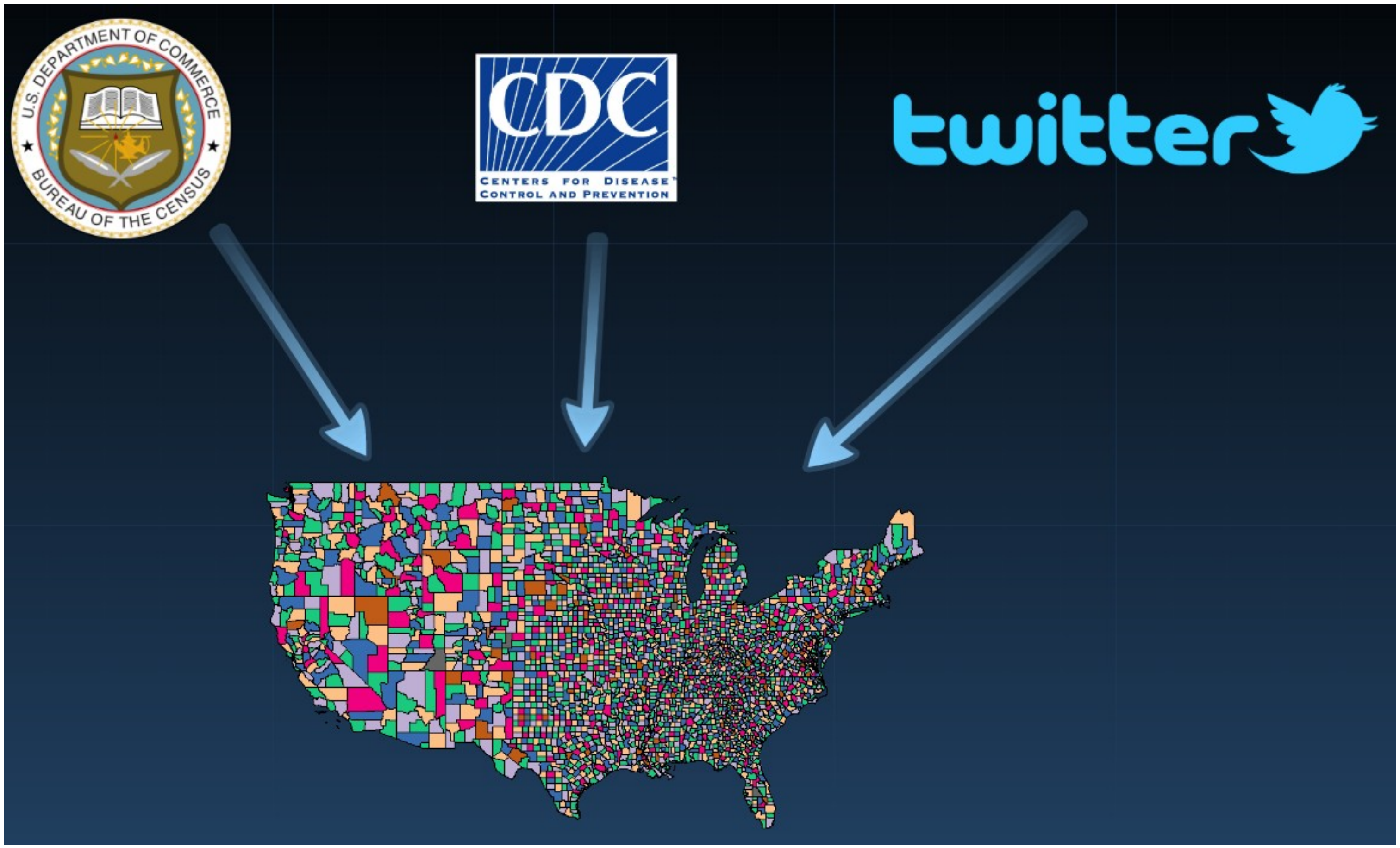
- Fit unrepresentative sample to representative sample results (i.e. implicitly maps unrepresentative sample to representative)
- **In the end we are validating against representative data.**

Individual Traits in Facebook

MyPersonality Dataset

- Facebook application to take "Big-5" personality survey.
- Approximately 75,000 users of the app:
  - shared their status updates for research
  - wrote at least 1,000 words
  - share their age and gender

Community Well-Being Through Twitter

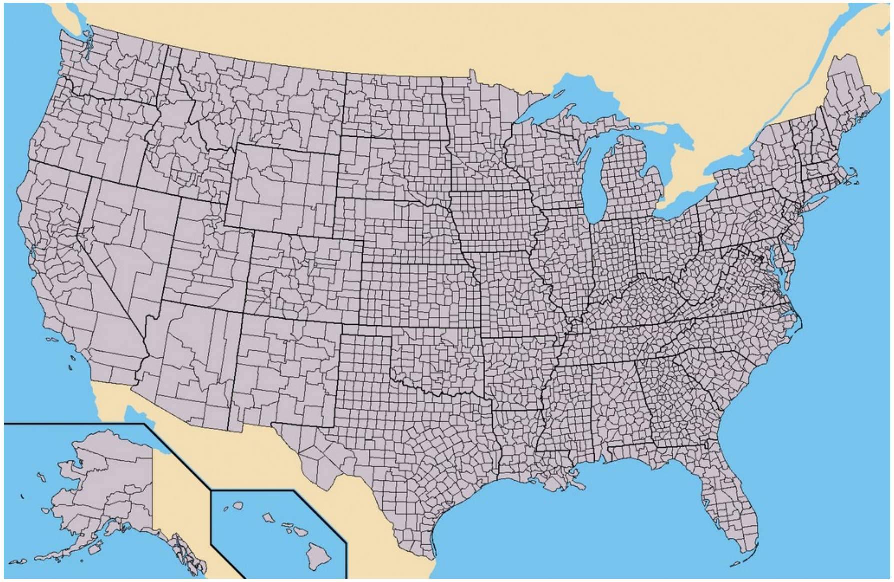


Community Well-Being through Twitter

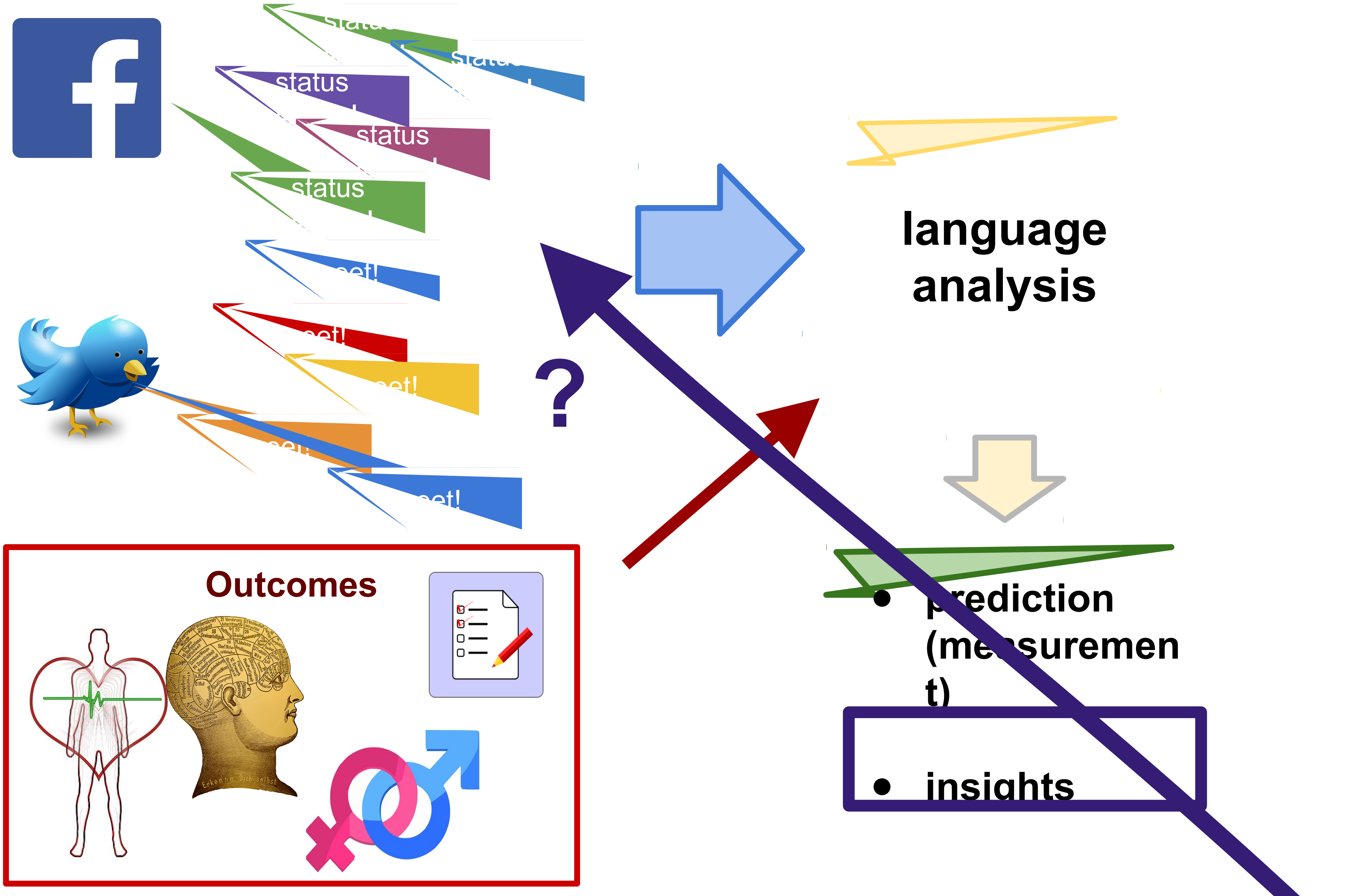
Twitter

> 150 million active monthly users
> 350 million messages a day

often list a location or geo-coordinates

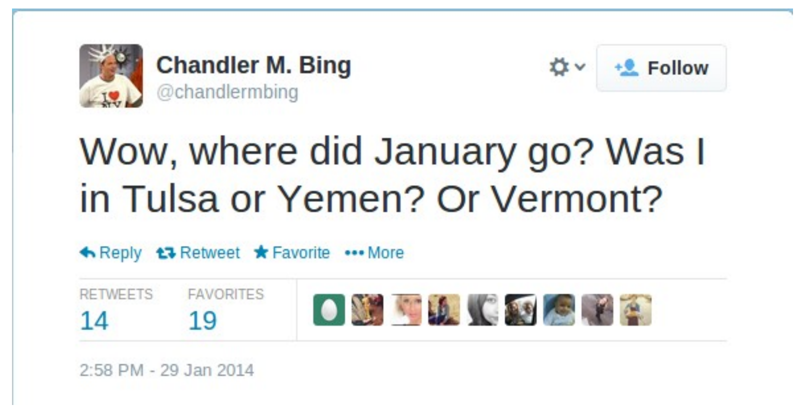


You Are What You Tweet



Example JSON - Tweet

```
{
 "coordinates": None,
 "created_at": "Wed Jan 29 22:58:50 +0000 2014",
 "favorite_count": 19,
 "favorited": False,
 "geo": None,
 "id": 428663556889145344,
 "lang": "en",
 "place": None,
 "retweet_count": 14,
 "retweeted": False,
 "text": "Wow, where did January go? Was I in Tulsa or Yemen? Or Vermont?",
 ...
}
```



- REST APIs
    - Twitter App building (e.g. smartphone apps)
    - Search API
- Streaming APIs
    - Firehose
    - public random sample
    - “user” and “site” streams

https://dev.twitter.com/docs

Sample Stream

- 1 % of all public tweets
- real time
- useful for representative language sample
    - less than 40% of tweets are in English
    - can be useful for frequencies of terms looked at

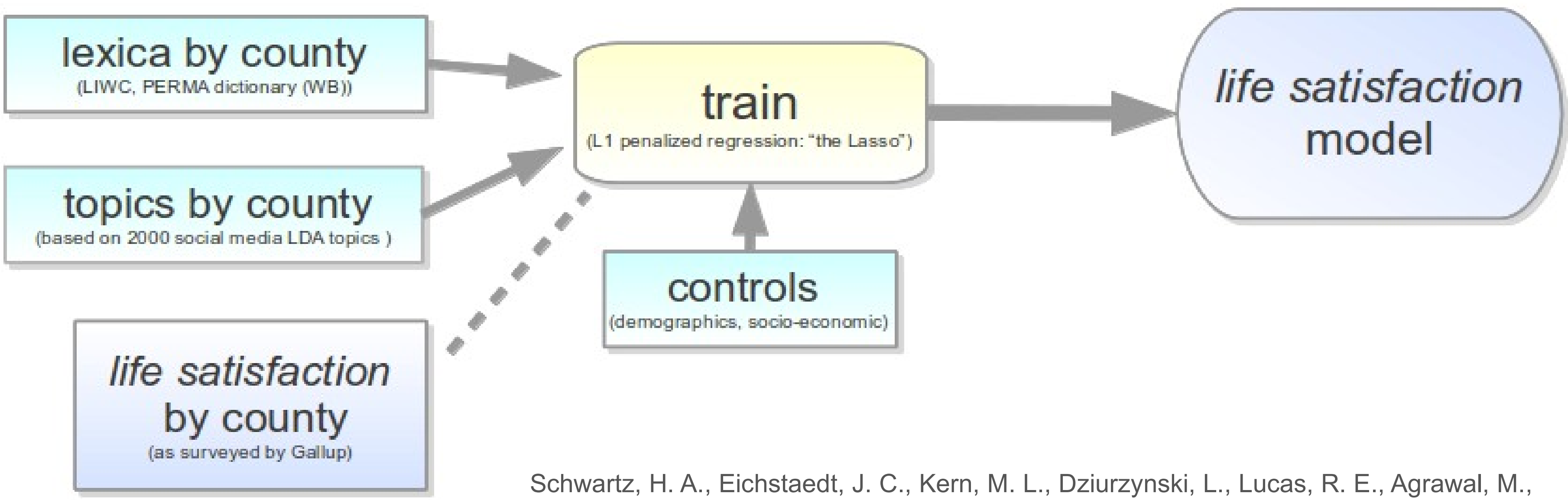
Search Stream

- Specific to what you're looking for
- same content as the web search
  https://twitter.com/search?q=obama
- parameters include
  - Recent vs Top tweets
  - Geolocalization
  - Language filter (Twitter's algorithm is "best effort")
  - time ranges (limited)
  - more:
    https://dev.twitter.com/docs/api/1.1/get/search/tweets

Community Heart Disease through Twitter

Method: Prediction

- Lasso, L1 penalized, regression
- Controls:
  - *demographics:* age, gender, ethnicity
  - *socio-economic status:* income, education



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.

Search Stream

- Specific to what you’re looking for
- same content as the web search
  https://twitter.com/search?q=obama
- parameters include
  - Recent vs Top tweets
  - Geolocalization
  - Language filter (Twitter’s algorithm is “best effort”)
  - time ranges (limited)
  - more:
    https://dev.twitter.com/docs/api/1.1/get/search/tweets

- Twitter uses OAuth2 for authentication
- Not a “username, password” authentication
- Need a “Twitter App” (and a Twitter account)
    - Anyone can create a blank app
    - Go to https://apps.twitter.com/app/new
    - Generate API key, API secret, access token & access secret on this page: [https://apps.twitter.com/app/*YOUR_APP_ID*/keys](https://apps.twitter.com/app/YOUR_APP_ID/keys)

- text of the tweet
- unique Twitter id
- created date & time
- replies:
  - user id & tweet id of tweet replied to
- retweets:
  - Tweet JSON of the original tweet
- favorited & retweeted counts
- entities
  - expanded links, hashtags, media & user mentions
- user info:
  - unique Twitter id
  - screen name, handle, location, description
  - nb tweets, favourites, followers
  - profile picture & background information

Find a complete list of fields at:
https://dev.twitter.com/docs/platform-objects/tweets &
https://dev.twitter.com/docs/platform-objects/users

!! Some fields are optional !!

Example Tweet JSON: https://gist.github.com/gnip/764239

Limitations of Twitter API

Sample Stream:

- only 1 % of all tweets
- terms that aren't frequent enough might not even appear in your dataset

Search:

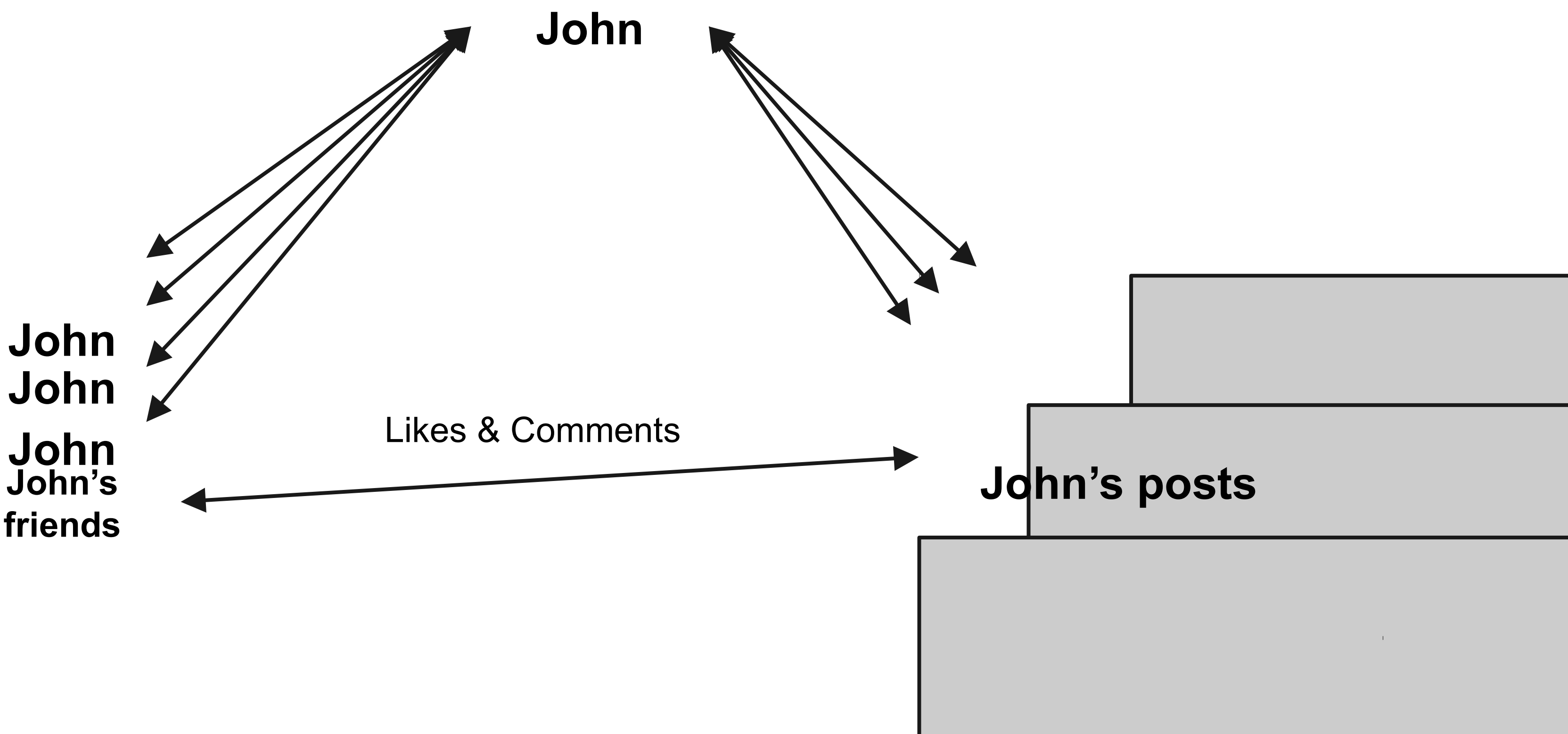
- 180 "queries" limit in every 15 minute window
- each search query can only contain 10 terms

- Free APIs
  - Graph API
  - Chat API
  - FQL API
- Third party APIs
  - Public Feed API
  - Keywords Insights API

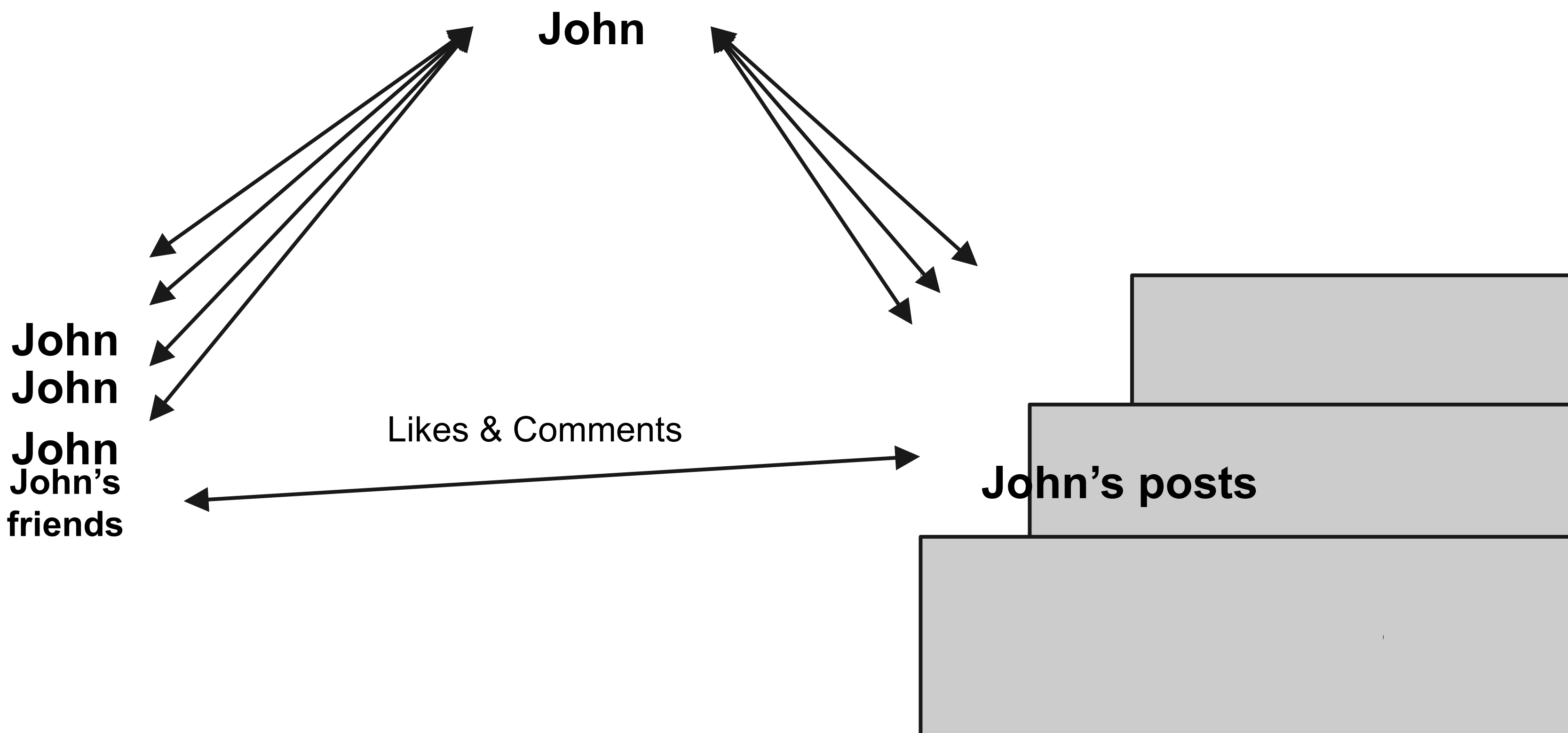
- Free APIs
  - Graph API
  - Chat API
  - FQL API
- Third party APIs
  - Public Feed API
  - Keywords Insights API

That's where the data is

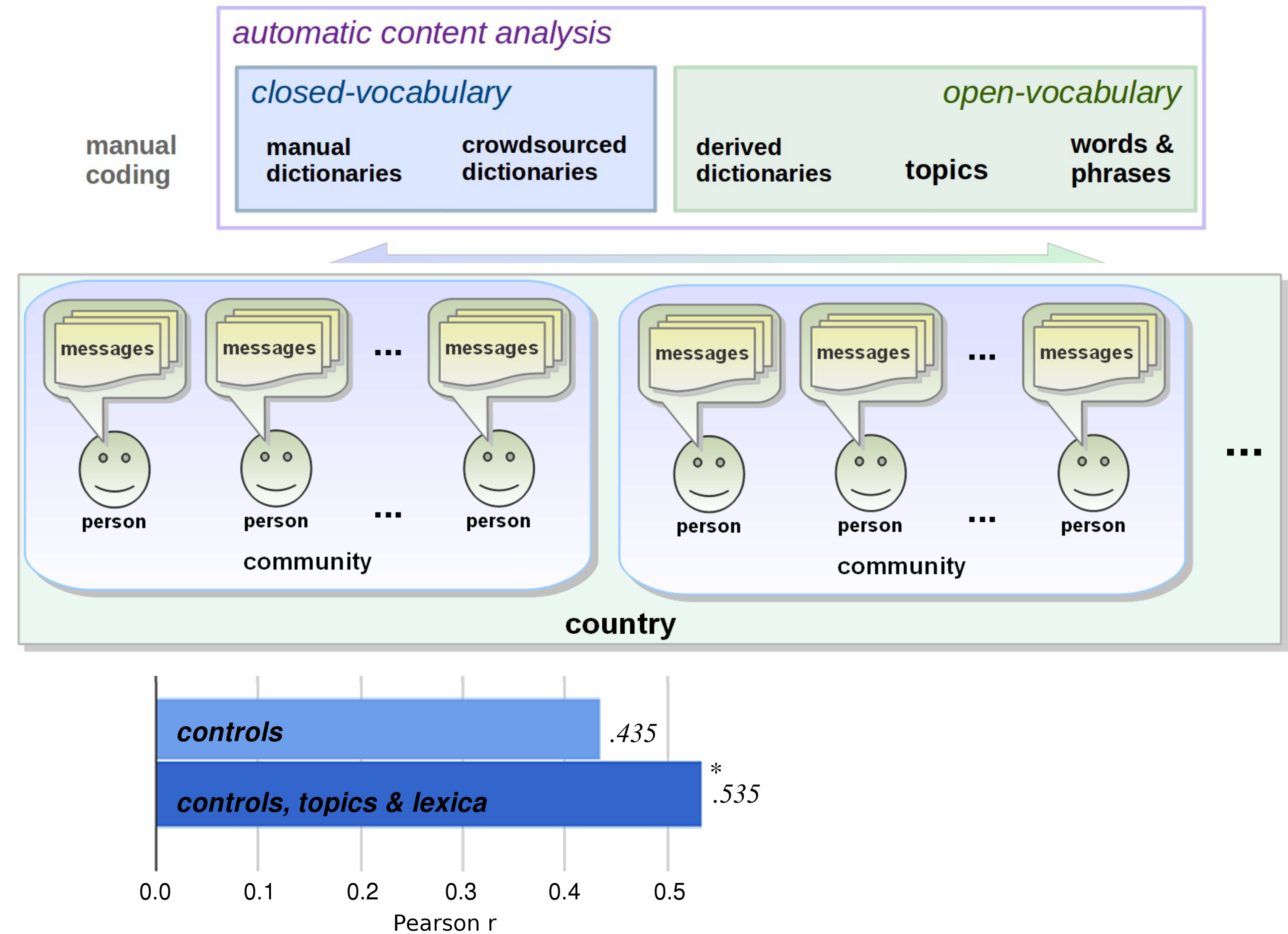
- Every data point is a node in a graph



- Every data point is a node in a graph



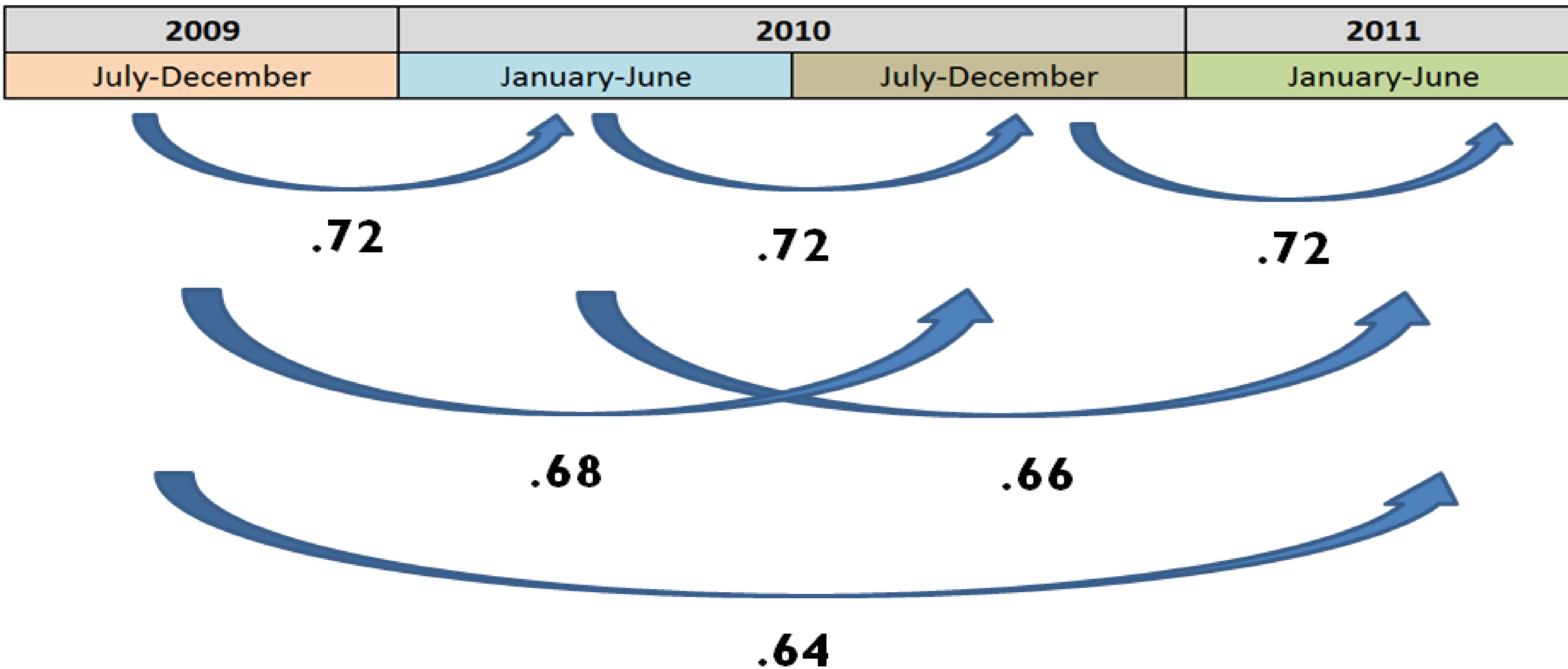
- API = Application Programming Interface
- Easier for huge amounts of data
- Twitter has multiple APIs
- So does Facebook
- How to use the Graph API to post/delete a status
- You might want to ask your programmer for help



automatic content analysis
closed-vocabulary
open-vocabulary
manual coding
manual dictionaries
crowdsourced dictionaries
derived dictionaries
topics
words & phrases
messages
person
community
country
controls
.435
controls, topics & lexica
*
.535
0.0
0.1
0.2
0.3
0.4
0.5
Pearson r

Individual Traits in Facebook

Test-Retest Reliability



Extraversion

Individual Traits in Facebook

MyPersonality Dataset

- Facebook application to take “Big-5” personality survey.
- Approximately 75,000 users of the app:
  - shared their status updates for research
  - wrote at least 1,000 words
  - share their age and gender

Individual Traits in Facebook: **Female**

Gender

Individual Traits in Facebook: **Male**

Gender