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

...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

300mil. tweets/day
4bil. messages/day
100mil. (Sina) weibos/day

PEOPLE:

150mil. (2014)
1bil. (2014)
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 concerns.
Social Media

Largest dataset(s) of everyday human behavior and concerns.
Social Media

1. Measurement

Largest dataset(s) of everyday human behavior and concerns.
Social Media

1. Measurement

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

Largest dataset(s) of everyday human behavior and concerns.
Measurement: Personality

Predicting Personality Traits:
Language vs. Friends

Accuracy/Correlation ($r$) with self-report

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism
Measurement: Personality

Predicting Personality Traits: Language vs. Friends

Accuracy/Correlation ($r$) with self-report

- Openness
- Conscientiousness
- Extraversion
- Agreeableness
- Neuroticism

Language

Friend
Measurement: Personality

Test-Retest Reliability

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>July-December</td>
<td>January-June</td>
<td>July-December</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.72</td>
<td>.72</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>.68</td>
<td>.66</td>
<td>.64</td>
</tr>
</tbody>
</table>

Extraversion
Social Media

1. Measurement

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

Largest dataset(s) of everyday human behavior and concerns.
Social Media

1. Measurement
   To what extent can we replace traditional survey-based methods?

2. Data-driven discovery
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 concerns.
Data-driven Social Science: Extraversion

sociable, assertive, active, energetic, talkative, outgoing

Data-driven Social Science: Introversion

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

Emotional stability
Data-Driven Social Science: Neuroticism

moody, anxious, fearful, worry-prone, depressive

Neuroticism

correlation strength

relative frequency

prevalence in topic
Data-Driven Social Science: Neuroticism
Data-Driven Social Science: Neuroticism

Neuroticism

- crap, sickness, ugh, hate, feel, sick, feels, miserable, worse
- anxiety, bored, crying, lonely, depressed, pissed:
  - I hate, depressed, fuck
  - anger, annoyed, mad, freaking
- worse, dead, alone

Emotional stability

- today, start, great, day
- volleyball, game, soccer, practice
- calm, peaceful, serene
- sun, beach, ocean, waves

Correlation strength

Relative frequency

Prevalence in topic
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 discover 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 discover 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

<table>
<thead>
<tr>
<th>microblogging</th>
<th>social interaction</th>
<th>messaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>mostly public</td>
<td>somewhat private</td>
<td>private</td>
</tr>
</tbody>
</table>
# Social Media Sources

<table>
<thead>
<tr>
<th>microblogging</th>
<th>social interaction</th>
<th>messaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>Facebook</td>
<td>Text Messages</td>
</tr>
<tr>
<td>Weibo</td>
<td>Renren</td>
<td>SnapChat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WeChat</td>
</tr>
<tr>
<td>mostly public</td>
<td>somewhat private</td>
<td>private</td>
</tr>
</tbody>
</table>

- microblogging: Twitter, Weibo
- social interaction: Facebook, Renren
- messaging: Text Messages, SnapChat, WeChat
## Social Media Sources

<table>
<thead>
<tr>
<th>microblogging</th>
<th>social interaction</th>
<th>messaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>Facebook</td>
<td>Text Messages</td>
</tr>
<tr>
<td>Weibo</td>
<td>Renren</td>
<td>SnapChat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WeChat</td>
</tr>
<tr>
<td>mostly public</td>
<td>somewhat private</td>
<td>private</td>
</tr>
<tr>
<td>big</td>
<td>bigger</td>
<td>biggest</td>
</tr>
</tbody>
</table>
# Social Media Sources

<table>
<thead>
<tr>
<th>microblogging</th>
<th>social interaction</th>
<th>messaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>Facebook</td>
<td>Text Messages</td>
</tr>
<tr>
<td>Weibo</td>
<td>Renren</td>
<td>Snapchat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WeChat</td>
</tr>
<tr>
<td>mostly public</td>
<td>somewhat private</td>
<td>private</td>
</tr>
<tr>
<td>big</td>
<td>bigger</td>
<td>biggest</td>
</tr>
</tbody>
</table>

**Other social media**
- Instagram
- YouTube
- Yelp
- Pinterest
- Tumblr
- Reddit
## Social Media Sources

<table>
<thead>
<tr>
<th>microblogging</th>
<th>social interaction</th>
<th>messaging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>Facebook</td>
<td>Text Messages</td>
</tr>
<tr>
<td>Weibo</td>
<td>Renren</td>
<td>SnapChat</td>
</tr>
<tr>
<td></td>
<td></td>
<td>WeChat</td>
</tr>
<tr>
<td>mostly public</td>
<td>somewhat private</td>
<td>private</td>
</tr>
<tr>
<td>big</td>
<td>bigger</td>
<td>biggest</td>
</tr>
</tbody>
</table>

**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, ...
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

Volunteer or Public Data

- social media messages
- personality age
- gender health
- well-being county

- visualization or predictive model

linguistic feature extraction

a) words and phrases
b) topics

correlation or model learning
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.
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

Automatic content analysis

Closed-vocabulary

- Manual dictionaries
- Crowdsourced dictionaries

Open-vocabulary

- Derived dictionaries
- Topics
- Words & phrases

Hand-driven

Data-driven

+ / - Emotion from LIWC
(Pennebaker et al., 2001)

Thumbs up... (Pang and Lee, 2004)

NRC Canada
(Mohammad et al., 2013)
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

Volunteer or Public Data

social media messages

personality age
gender health
well-being county
...

Visualization or predictive model

linguistic feature extraction

a) words and phrases

b) topics

correlation or model learning

:); love day everyone excited for
great blessed beautiful
amazing

praise psalm

today tomorrow
friends happy awesome prayers church
thankful fun praygood tonight
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

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

CDC-reported ACHD Mortality

Percentile of ACHD Mortality
Example: Community Heart Disease Mortality
Example: Community Heart Disease Mortality
Example: Community Heart Disease Mortality

Performance of Twitter-Based and Traditional Risk Factor-Based Regression Models of County-Level Atherosclerotic Coronary Heart Disease (ACHD) Mortality

Accuracy of County-Level ACHD Predictions (Pearson r with CDC-reported ACHD)
Example: Community Heart Disease Mortality

Performance of Twitter-Based and Traditional Risk Factor-Based Regression Models of County-Level Atherosclerotic Coronary Heart Disease (ACHD) Mortality

Accuracy of County-Level ACHD Predictions (Pearson r with CDC-reported ACHD)
Example: Community Heart Disease Mortality

Performance of Twitter-Based and Traditional Risk Factor-Based Regression Models of County-Level Atherosclerotic Coronary Heart Disease (ACHD) Mortality

Accuracy of County-Level ACHD Predictions (Pearson r with CDC-reported ACHD)
Example: Community Heart Disease Mortality

Performance of Twitter-Based and Traditional Risk Factor-Based Regression Models of County-Level Atherosclerotic Coronary Heart Disease (ACHD) Mortality

Accuracy of County-Level ACHD Predictions (Pearson r with CDC-reported ACHD)
Example: Community Heart Disease Mortality

Performance of Twitter-Based and Traditional Risk Factor-Based Regression Models of County-Level Atherosclerotic Coronary Heart Disease (ACHD) Mortality

Accuracy of County-Level ACHD Predictions (Pearson r with CDC-reported ACHD)
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

lexica by county
(LIWC, PERMA dictionary (WB))

topics by county
(based on 2000 social media LDA topics)

life satisfaction by county
(as surveyed by Gallup)

controls
(demographics, socio-economic)

train
(L1 penalized regression: "the Lasso")

life satisfaction model
Example: County Life Satisfaction

![Diagram showing life satisfaction with respect to county topic use, controlling for various factors.

- **controls**: Pearson r = 0.435
- **controls, topics & lexica**: Pearson r = 0.535

The graphs illustrate the relationship between county topic use and life satisfaction across different quartiles, controlling for various factors.
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

When Google got flu wrong

US outbreak foxes a leading web-based method for tracking seasonal flu

Declan Butler

13 February 2013

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

Estimated % of US population with influenza-like illness

Google’s algorithms overestimated peak flu levels this year
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
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 misclassification
Facebook status update.

<table>
<thead>
<tr>
<th>category</th>
<th>label</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Ambiguity</td>
<td>Wrong POS</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Wrong WS</td>
<td>38</td>
</tr>
<tr>
<td>Signal Negation</td>
<td>Strict Negation</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Desiring</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>Stem Issue</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>24</td>
</tr>
</tbody>
</table>
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?

Tweet!

Surveyed well-being from representative sample.
Representative Sample?

Tweet!

Surveyed well-being from representative sample.
Representative Sample?

Tweet!

fit model

well-being language model

Surveyed well-being from representative sample.
Representative Sample?

Tweet!

fit model

well-being language model

Surveyed well-being from representative sample.

Predicted well-being

$r = ?$

predict

Tweet!
Representative Sample?

- Alternative: 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
Why Social Media and Language?

unobtrusive
potential for real-time
longitudinal / look back in time
often personal / everyday concerns
Thank You! Questions?

hansens@seas.upenn.edu

Big Data for Official Statistics

Penn | World Well-Being Project | wwwbp.org
APPENDIX
Method: County-Mapping

Rule-Based Mapping

Tweet

- location string
- geo-coordinates

Tokenize

contains country? yes no

- is US?
  - no
  - contains state?
    - yes
    - contains valid city prior to state?
      - yes
      - city, state
      - map to most-probable county
      - no
      - not mapped
    - no
    - contains city with > 90% chance of being in one state?
      - yes
      - not mapped
      - no
    - no
  - no

if tweet has coordinates (< 2% of tweets), then map directly to county

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:

The overall MapReduce word count process

Input: Deer Bear River, Car Car River, Deer Car Bear

Splitting:
- Deer Bear River
- Car Car River
- Deer Car Bear

Mapping:
- Deer, 1
- Bear, 1
- River, 1

Shuffling:
- Bear, 1
- Bear, 1

Reducing:
- Bear, 2
- Car, 3
- Deer, 2
- River, 2

Final result:
- Bear, 2
- Car, 3
- Deer, 2
- River, 2

image: http://xiaochongzhang.me
Well-Being and Policy

OECD Guidelines on Measuring Subjective Well-being

=> Life Satisfaction (across domains)
What topics matter for all counties (that we have data for) in the United States?

Evidence for moderation
• A moderator alters the strength or direction of a relationship
• Question of external validity – how universal is the effect?

Individual Well-Being

satisfaction with life
Individual Well-Being: message to user-level

<table>
<thead>
<tr>
<th>baselines</th>
<th>$r$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(mean)</td>
<td>.000</td>
</tr>
<tr>
<td>lexica: GNH</td>
<td>.210</td>
</tr>
<tr>
<td>lexica: Hedonometer</td>
<td>.108</td>
</tr>
<tr>
<td>Correlations between Personality Ratings</td>
<td>Self-Language</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>N =</td>
<td>5,000</td>
</tr>
<tr>
<td>Openness</td>
<td>.43</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.37</td>
</tr>
<tr>
<td>Extraversion</td>
<td>.42</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.34</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>.35</td>
</tr>
<tr>
<td>Criterion measures</td>
<td>N</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Number of friends</td>
<td>711</td>
</tr>
<tr>
<td>Number of doctor visits</td>
<td>736</td>
</tr>
<tr>
<td>Number of sick days</td>
<td>733</td>
</tr>
<tr>
<td>Politically liberal</td>
<td>756</td>
</tr>
<tr>
<td>Fair-mindedness</td>
<td>864</td>
</tr>
<tr>
<td>Self-disclosure</td>
<td>864</td>
</tr>
<tr>
<td>Self-monitoring</td>
<td>927</td>
</tr>
<tr>
<td>Satisfaction with life</td>
<td>1114</td>
</tr>
<tr>
<td>Barratt Impulsiveness Scale</td>
<td>549</td>
</tr>
<tr>
<td>Attention</td>
<td></td>
</tr>
<tr>
<td>Cognitive Instability</td>
<td></td>
</tr>
<tr>
<td>Motor</td>
<td></td>
</tr>
<tr>
<td>Perseverance</td>
<td></td>
</tr>
<tr>
<td>Self Control</td>
<td></td>
</tr>
<tr>
<td>Cognitive Complexity</td>
<td></td>
</tr>
</tbody>
</table>

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

Outcomes

- prediction (measurement)
- insights
Example JSON - Tweet

```json
{
    "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

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
What's in a “Tweet JSON”

- 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

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

manual
coding

messages
person
community
country

open-vocabulary

derived
dictionaries

topics

words &
phrases

crowdsourced
dictionaries

controls

controls, topics & lexica

Pearson r

.435

.535

*.
## Individual Traits in Facebook

### Test-Retest Reliability

<table>
<thead>
<tr>
<th></th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>July-December</td>
<td>January-June</td>
<td>July-December</td>
</tr>
<tr>
<td></td>
<td>.72</td>
<td>.72</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>.72</td>
<td>.68</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>.64</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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

love <3 hun
sweetheart
inlove
love
sweetie
<3
hun

shopping
with_my
mom
birthday
loves_her
so_happy
 hubsy
dress
boyfriend
soooyay_

wonderful
lovely

excited
baby
much fun
cute

bestie

happy
years
birthday
wishing
day

babe

<3

prevalence in topic
relative frequency
correlation strength
Individual Traits in Facebook: Male

Gender

correlation strength

relative frequency

prevalence in topic