Geographical variation of happiness

as expressed by users of Twitter

Kameron Decker Harris

Peter S. Dodds
Isabel M. Klouman
Catherine Bliss
Christopher M. Danforth

April 26, 2011

Why measure happiness?

It's important:

- Choose: gross domestic... product or happiness?
- People $\neq$ rational . . . feelings influence everything. To model behavior, we must understand states of mind.

To this end we need an objective instrument, call it a *hedonometer*

Why measure happiness?

It's important:

- Choose: gross domestic... product or happiness?
- People $\neq$ rational . . . feelings influence everything. To model behavior, we must understand states of mind.

To this end we need an objective instrument, call it a *hedonometer*

Why measure happiness?

It's important:

- Choose: gross domestic... product or happiness?
- People $\neq$ rational . . . feelings influence everything. To model behavior, we must understand states of mind.

To this end we need an objective instrument, call it a *hedonometer*

Why measure happiness?

It's important:

- Choose: gross domestic... product or happiness?
- People $\neq$ rational . . . feelings influence everything. To model behavior, we must understand states of mind.

To this end we need an objective instrument, call it a *hedonometer*

Why measure happiness?

It's important:

- Choose: gross domestic... product or happiness?
- People $\neq$ rational ... feelings influence everything. To model behavior, we must understand states of mind.

To this end we need an objective instrument, call it a *hedonometer*

Hedonom-*a-what*?

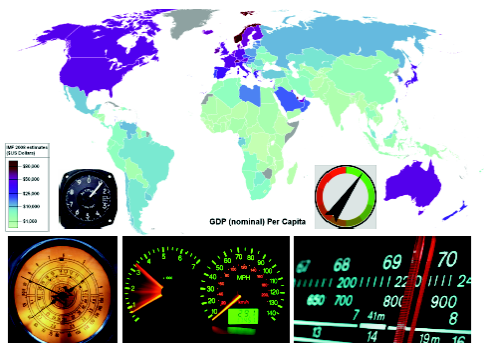
(actually GDP!)

'hedonometer' = device to measure well-being

Ideally:

- objective
- improvable
- text-based
- fast, for big data
  (Twitter data set has $\sim 30 \times 10^9$ words)

Hedonom-*a-what*?



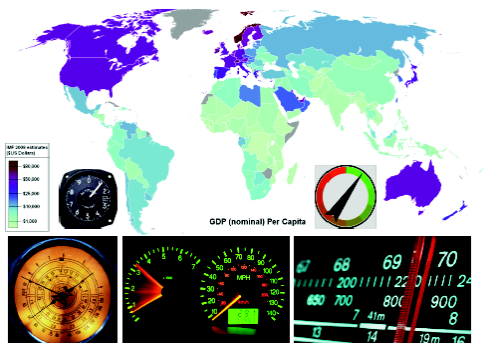
(actually GDP!)

'hedonometer' = device to measure well-being

Ideally:

- objective
- improvable
- text-based
- fast, for big data
  (Twitter data set has $\sim 30 \times 10^9$ words)

Hedonom-*a-what*?



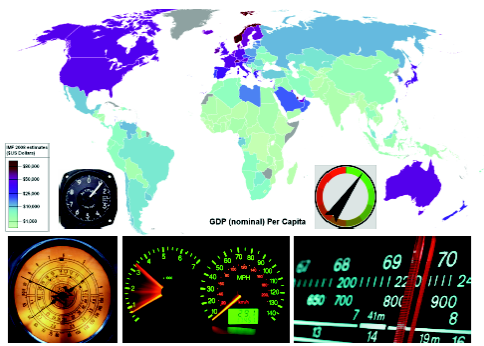
(actually GDP!)

'hedonometer' = device to measure well-being

Ideally:

- objective
- improvable
- text-based
- fast, for big data
  (Twitter data set has $\sim 30 \times 10^9$ words)

Hedonom-*a-what*?



(actually GDP!)

'hedonometer' = device to measure well-being

Ideally:

- objective
- improvable
- text-based
- fast, for big data
  (Twitter data set has $\sim 30 \times 10^9$ words)

Hedonom-*a-what*?

(actually GDP!)

'hedonometer' = device to measure well-being

Ideally:

- objective
- improvable
- text-based
- fast, for big data
  (Twitter data set has $\sim 30 \times 10^9$ words)

Hedonom-*a-what*?

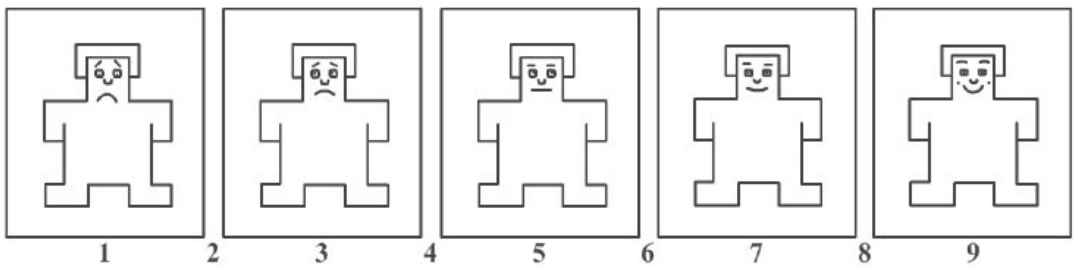
(actually GDP!)

'hedonometer' = device to measure well-being

Ideally:

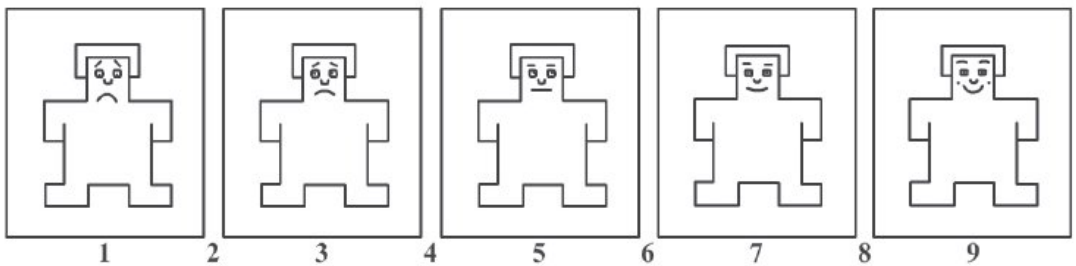
- objective
- improvable
- text-based
- fast, for big data
  (Twitter data set has $\sim 30 \times 10^9$ words)

ANEW = 'Affective Norms for English Words'



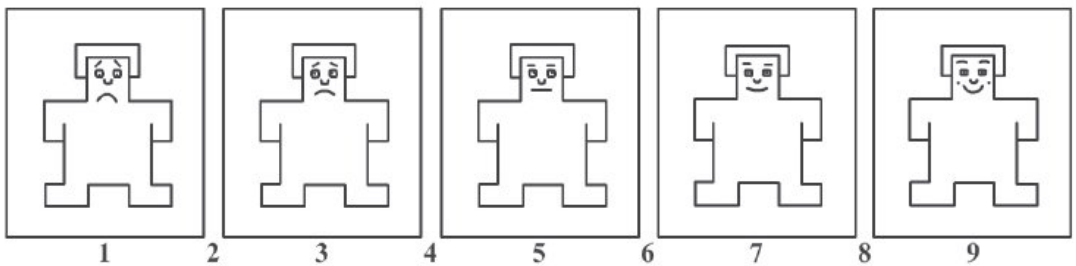
- 1034 words with scored on three dimensions:
  - valence (happiness)
  - arousal
  - dominance
- from Bradley and Lang (1999)

ANEW = 'Affective Norms for English Words'



- 1034 words with scored on three dimensions:
  - valence (happiness)
  - arousal
  - dominance
- from Bradley and Lang (1999)

ANEW = 'Affective Norms for English Words'



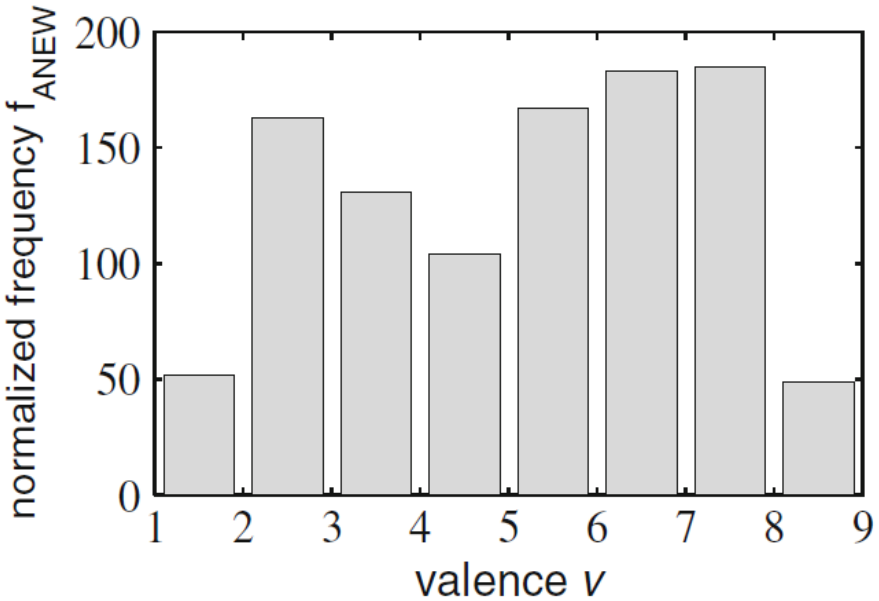
- 1034 words with scored on three dimensions:
  - valence (happiness)
  - arousal
  - dominance
- from Bradley and Lang (1999)

Scoring a text

- look from occurances of words in ANEW list, construct frequencies f_i, $i = 1, \ldots, 1034$
- create weighted average based on scores s_i

$$\text{happiness} = \frac{\sum_i s_i f_i}{\sum_i f_i}$$

Scores for typical ANEW words



triumphant (8.82) / paradise (8.72) / love (8.72)
luxury (7.88) / trophy (7.78) / glory (7.55)
optimism (6.95) / church (6.28) / pancakes (6.08)
street (5.22) / paper (5.20) / engine (5.20)
neurotic (4.45) / vanity (4.30) / derelict (4.28)
fault (3.43) / lawsuit (3.37) / corrupt (3.32)
disgusted (2.45) / hostage (2.20) / trauma (2.10)
funeral (1.39) / rape (1.25) / suicide (1.25)

Takeaway: ANEW sample a mixture of salient and neutral words

Example scores for certain texts

Text:	h_{avg}	Words with a similar score:
Soul/Gospel music lyrics [20]	6.9	chocolate (6.88), leisurely (6.88), penthouse (6.81)
Pop music lyrics [20]	6.7	dream (6.73), honey (6.73), sugar (6.74)
Dante's Paradise [33]	6.5	muffin (6.57), rabbit (6.57), smooth (6.58)
Tweets, 9/9/2008 to 12/31/2010 (present work)	6.4	thought (6.39), face (6.39), blond (6.42)
Rock music lyrics [20]	6.3	church (6.28), tree (6.32), air (6.34)
Enron Emails [34]	6.2	clouds (6.18), alert (6.20), computer (6.24)
State of the Union Messages [20]	6.1	grass (6.12), idol (6.12), bottle (6.15)
New York Times (1987–2007) [35]	6.0	hotel (6.00), tennis (6.02), wonder (6.03)
Blogs [20]	5.8	owl (5.80), whistle (5.81), humble (5.86)
Dante's Inferno [33]	5.5	glacier (5.50), repentant (5.53), mischief (5.57)
Metal/Industrial music lyrics [20]	5.4	lamp (5.41), elevator (5.44), truck (5.47)

So many possibilities!

One can examine:

- trends in time
- network characteristics
- beyond ANEW (distributions of all words, information theory)
- spatial trends

So many possibilities!

One can examine:

- trends in time
- network characteristics
- beyond ANEW (distributions of all words, information theory)
- spatial trends

So many possibilities!

One can examine:

- trends in time
- network characteristics
- beyond ANEW (distributions of all words, information theory)
- spatial trends

So many possibilities!

One can examine:

- trends in time
- network characteristics
- beyond ANEW (distributions of all words, information theory)
- spatial trends

So many possibilities!

One can examine:

- trends in time
- network characteristics
- beyond ANEW (distributions of all words, information theory)
- spatial trends

So many possibilities!

One can examine:

- trends in time
- network characteristics
- beyond ANEW (distributions of all words, information theory)
- spatial trends ⟵ this talk

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

Identifying the origin of tweets

Two routes:

1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

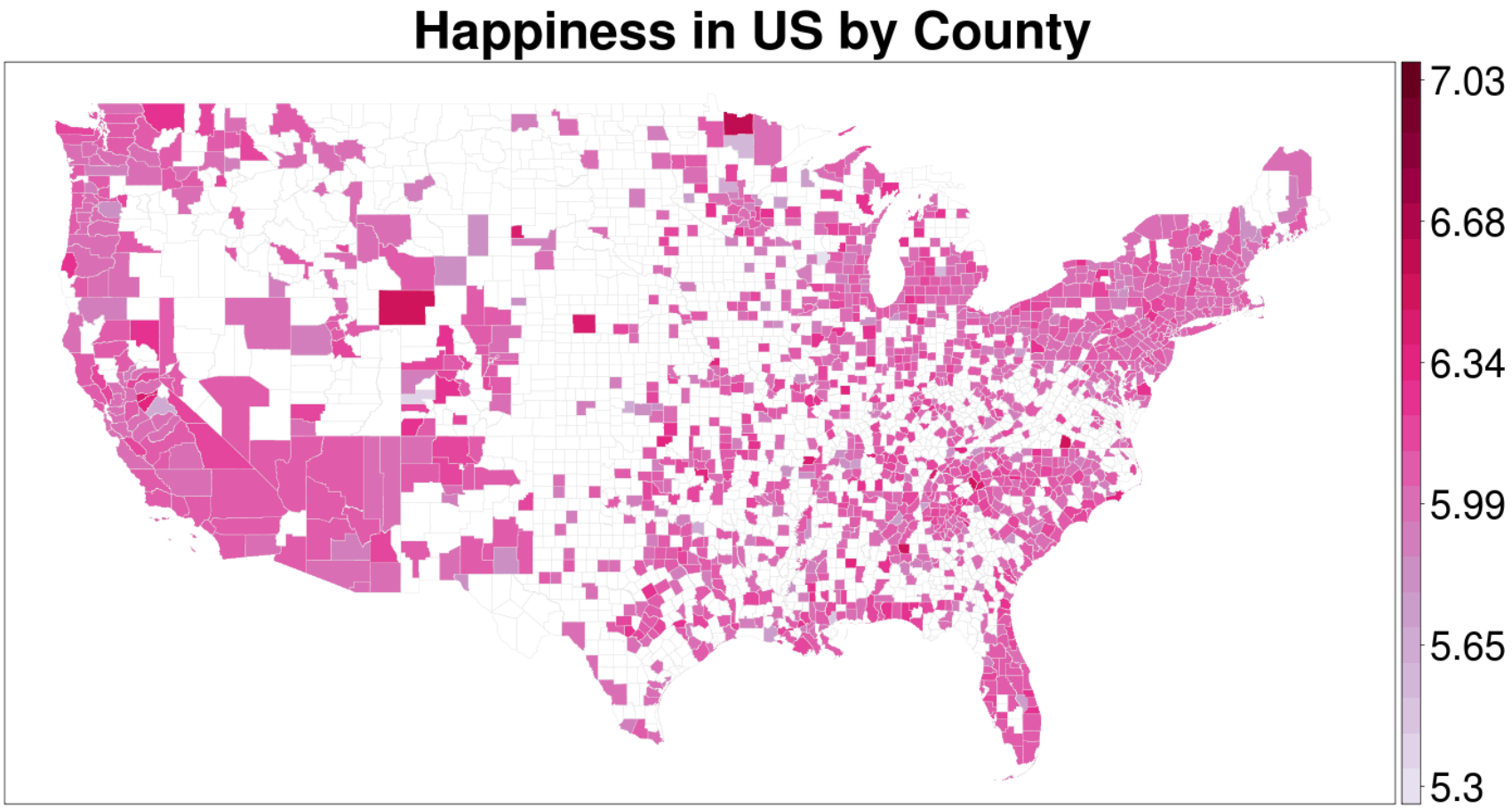
Identifying the origin of tweets

Two routes:

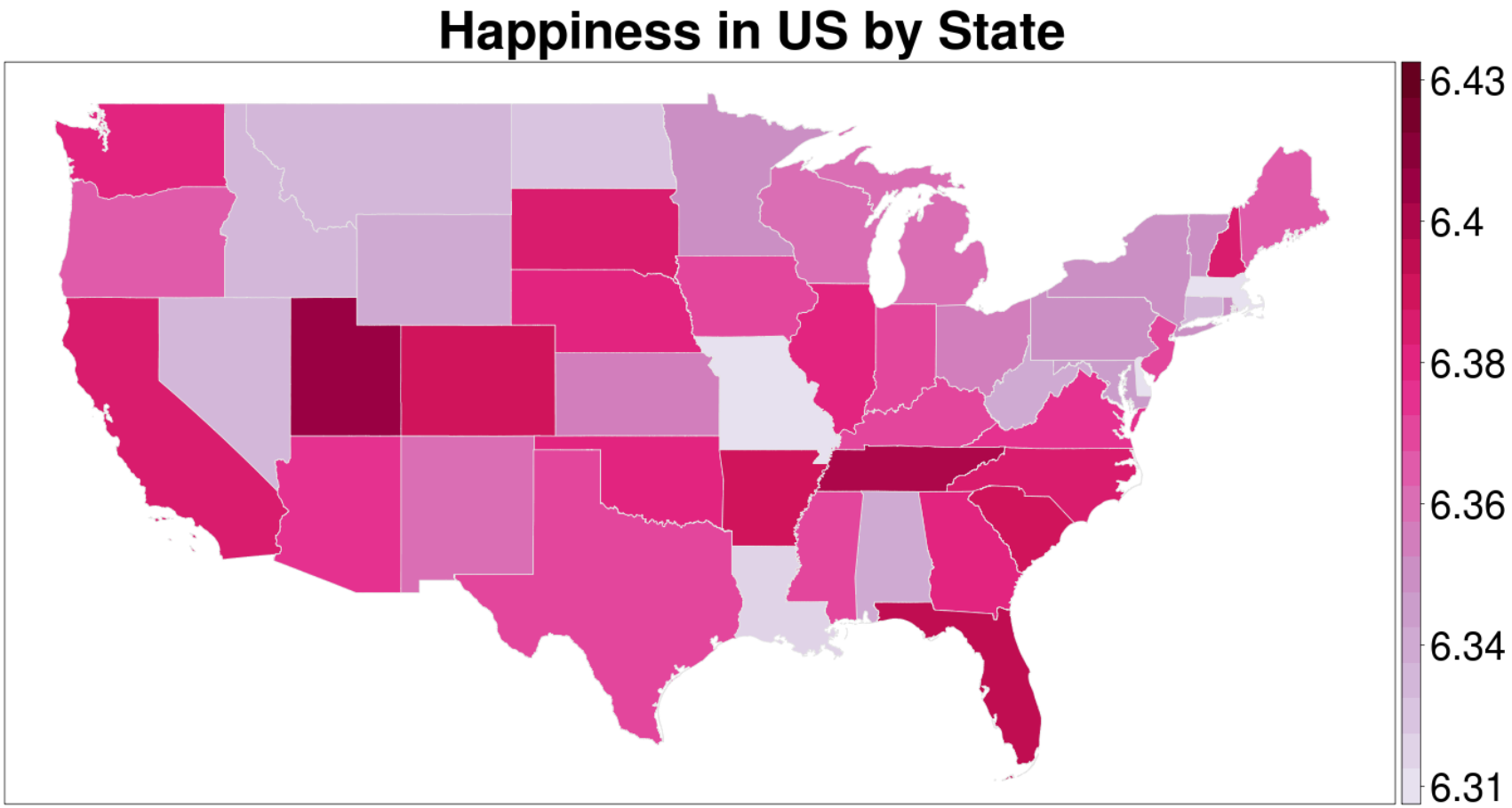
1. self-reporting
   - users accounts specify a "Location", e.g. "Burlington VT", "New York", "over the rainbow!! :-)"
   - parse out city & state, e.g. ["burlington", "vt"]
   - look up in database of places (USGS Geonames)
   - ∼5-20% identifiable this way
   - store at level of county
   - data since Sept. 2008
   - we do this now
2. location encoded in tweet

The big picture: maps

Counties with data for $>$ 100 unique users ('09-'10)

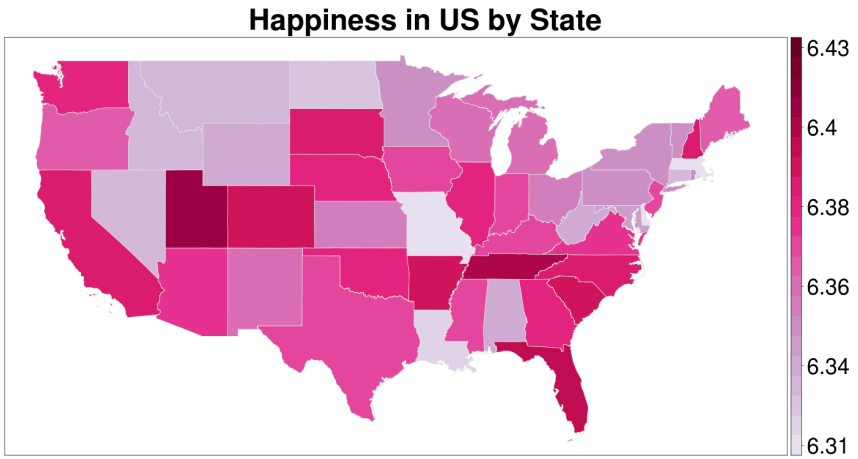


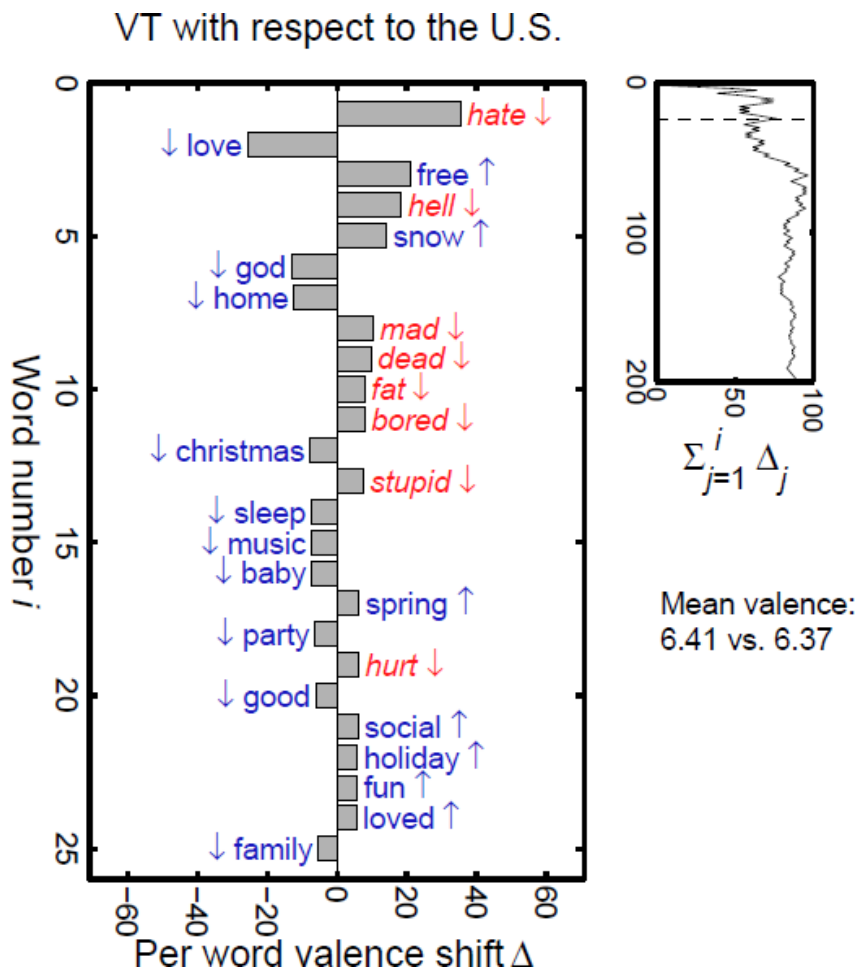
All states ('09-'10)



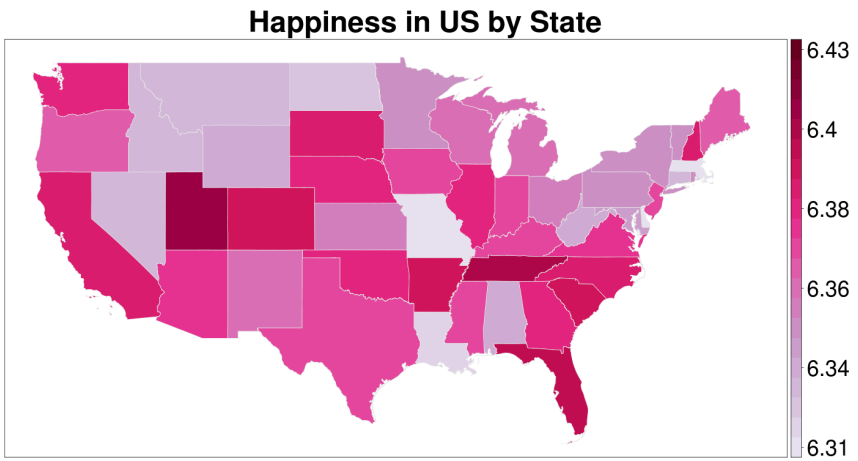
Getting into the details: word-shift plots

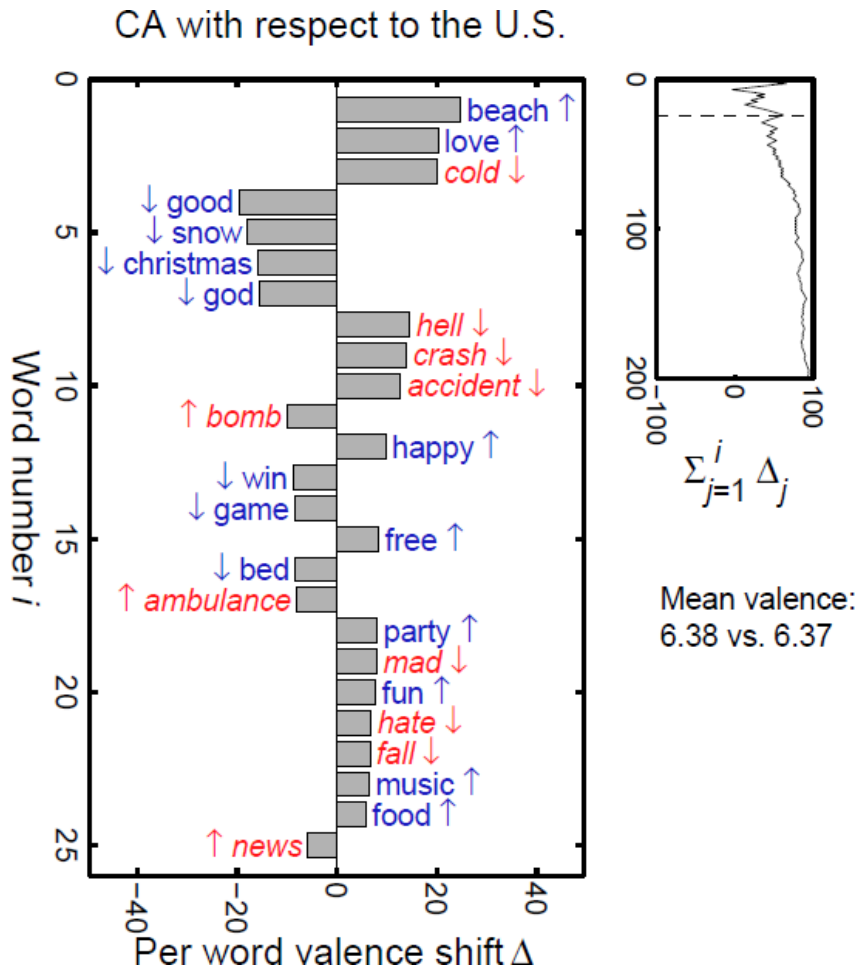
Why states' scores differ: specific words



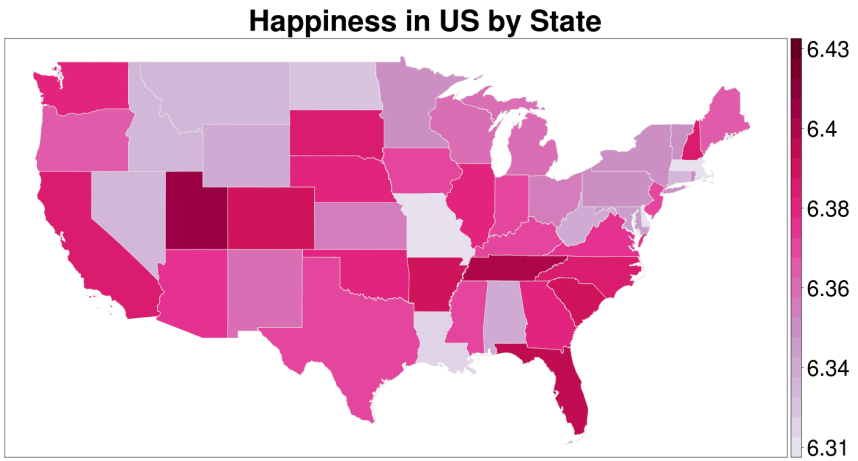


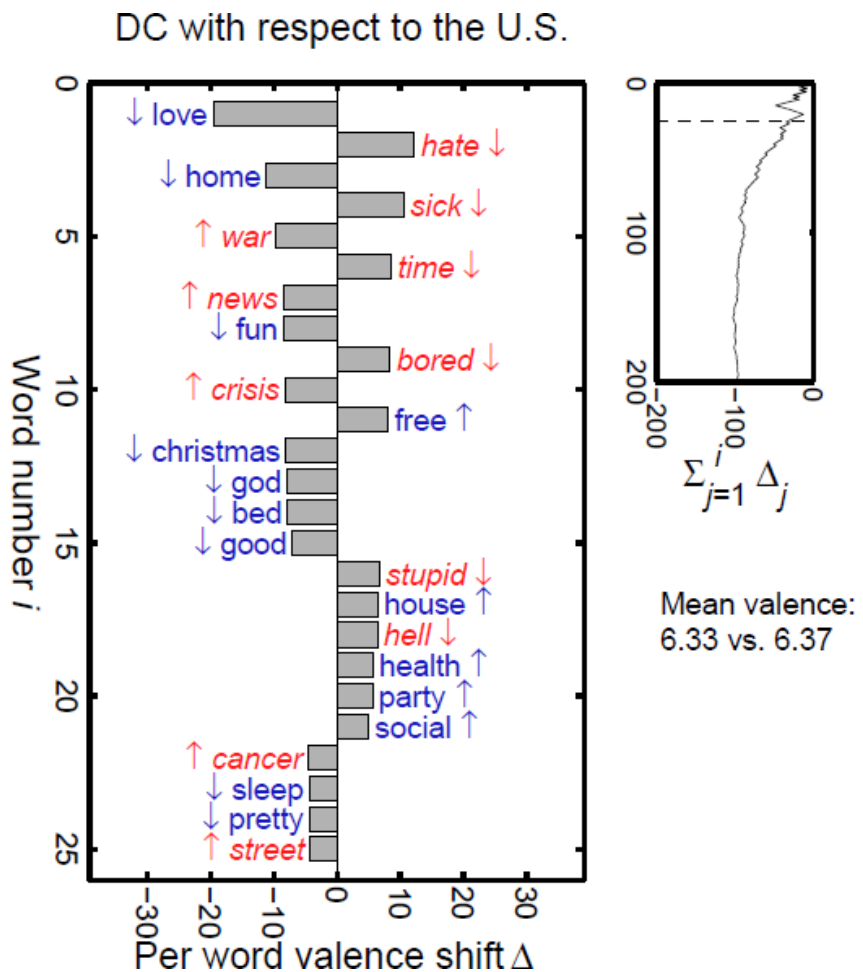
Why states' scores differ: specific words





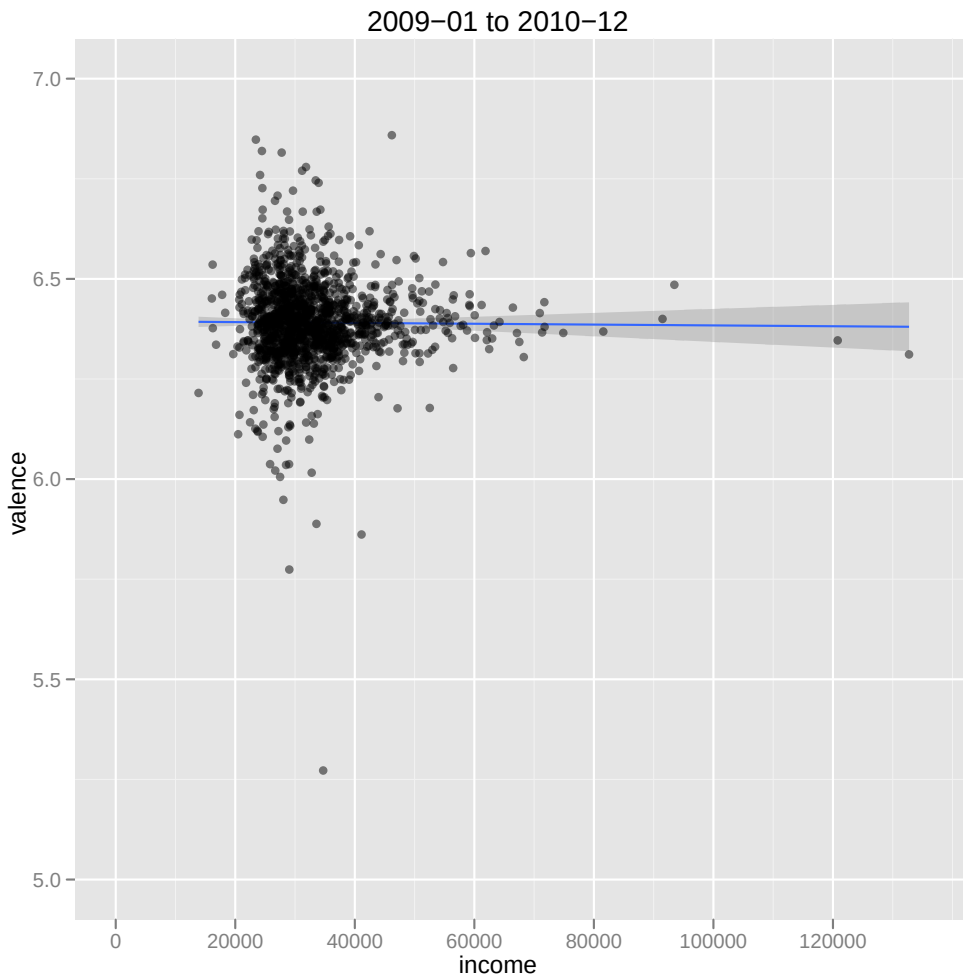
Why states' scores differ: specific words



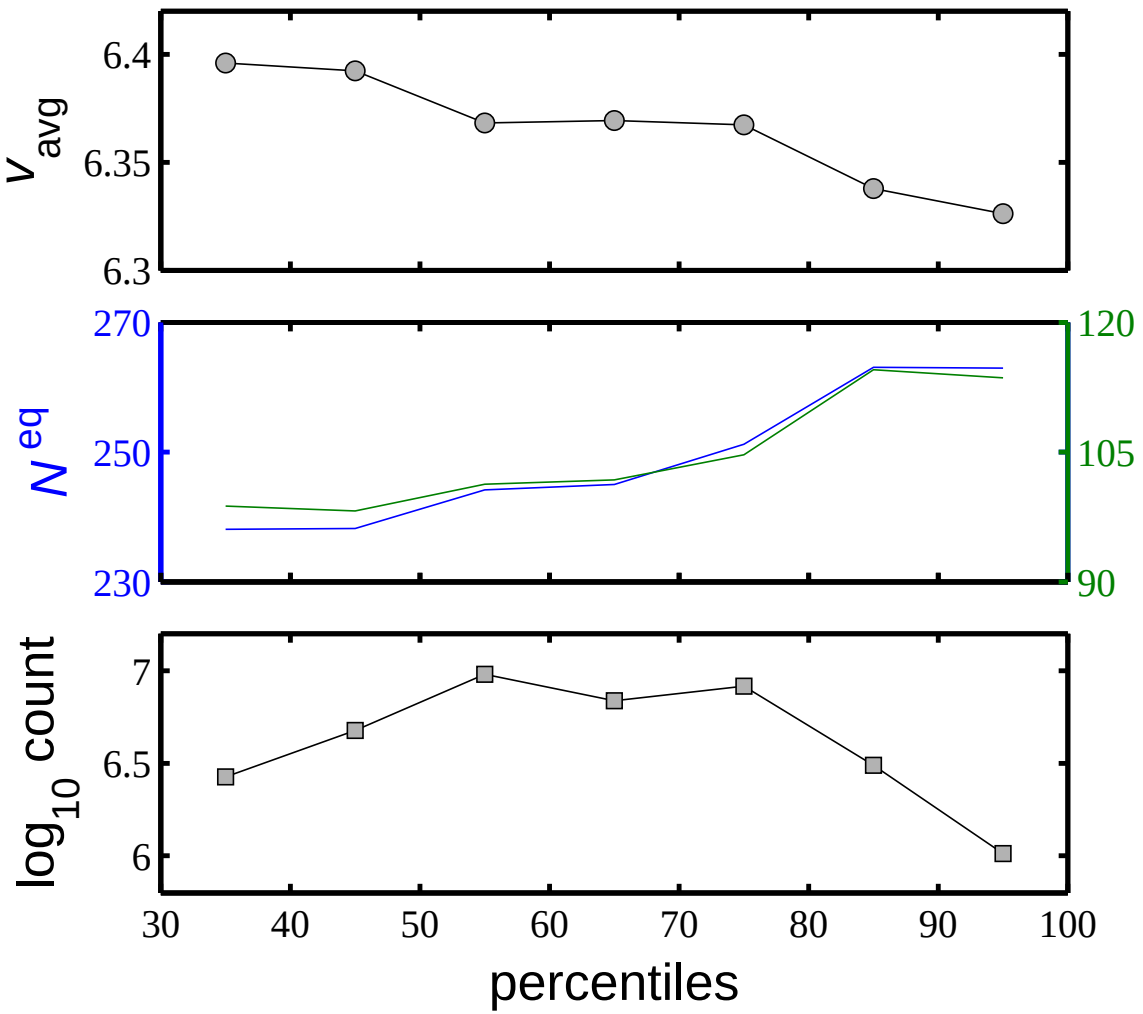


Spatial correlations with socioeconomic indices

Happiness and $$$



Happiness, information, and politics



Conclusions, future directions

- Fully investigate socioeconomic correlations
- Add in time analysis (deal with non-uniform sample sizes)
- Compare to detailed geocoding
- Kridgings, other elegant geostatistical analysis

Conclusions, future directions

- Fully investigate socioeconomic correlations
- Add in time analysis (deal with non-uniform sample sizes)
- Compare to detailed geocoding
- Kridgings, other elegant geostatistical analysis

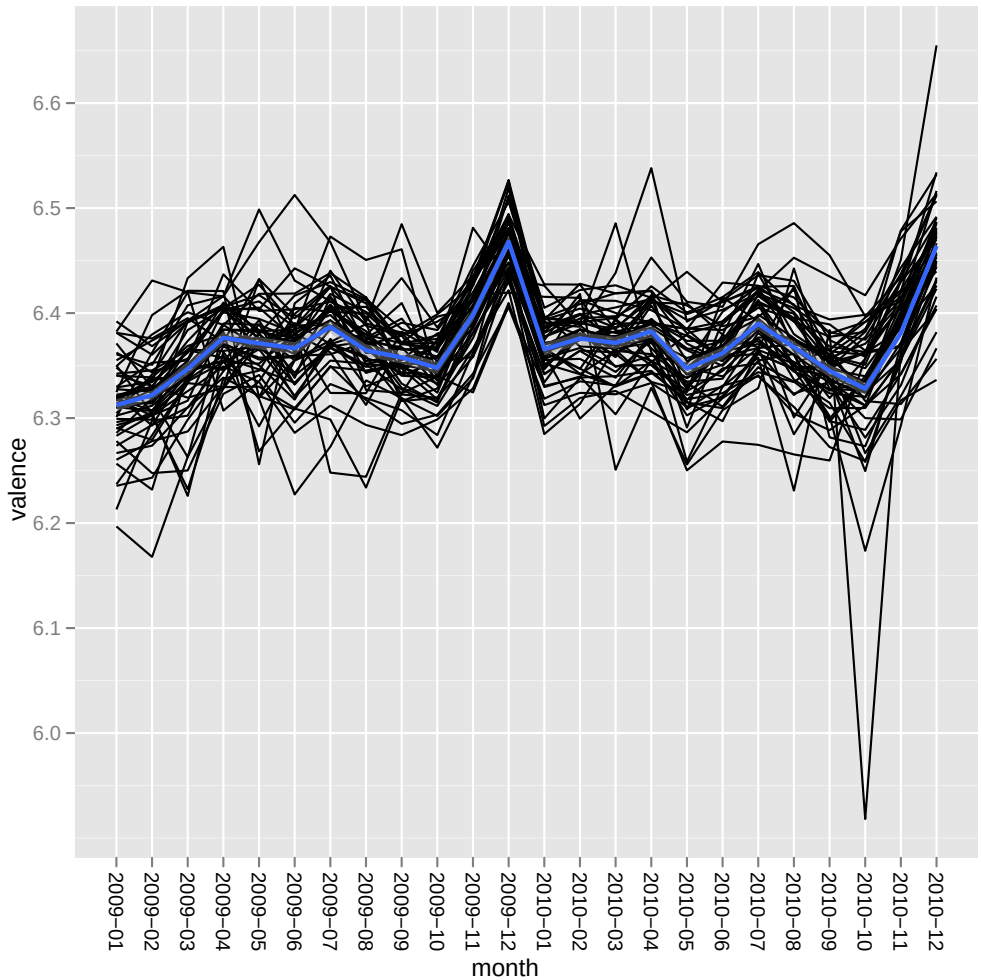
Conclusions, future directions

- Fully investigate socioeconomic correlations
- Add in time analysis (deal with non-uniform sample sizes)
- Compare to detailed geocoding
- Kridgings, other elegant geostatistical analysis

Conclusions, future directions

- Fully investigate socioeconomic correlations
- Add in time analysis (deal with non-uniform sample sizes)
- Compare to detailed geocoding
- Kridgings, other elegant geostatistical analysis

Happiness trajectories of individual states



Information trajectories of individual states