# HAPPINESS IN THE UNITED STATES

## MEASURING THE SENTIMENT OF GEOLOCATED TWEETS

JOÃO FERREIRA LOFF

*SISTEMAS PARA INFORMAÇÃO GEO-REFERENCIADA*

*2013/2014*

# HAPPINESS (?)

# HAPPINESS (?)

- What is happiness?

# HAPPINESS (?)

- What is happiness?
- How do we measure happiness?

# HAPPINESS (?)

- What is happiness?
- How do we measure happiness?
- Self-evaluation?

# HAPPINESS (?)

- What is happiness?
- How do we measure happiness?
- Self-evaluation?
- Survey-methods?

# HAPPINESS (?)

- What is happiness?
- How do we measure happiness?
- Self-evaluation?
- Survey-methods?
- Social-economic factors?

# HAPPINESS (?)

- What is happiness?
- How do we measure happiness?
- Self-evaluation?
- Survey-methods?
- Social-economic factors?
- … what about **social media**?

# HAPPINESS (?)

- What is happiness?
- How do we measure happiness?
- Self-evaluation?
- Survey-methods?
- Social-e
- … what about **social media**?

**Analyse Twitter messages to evaluate happiness!**

# RELATED WORK

# HEDONOMETER

# HEDONOMETER

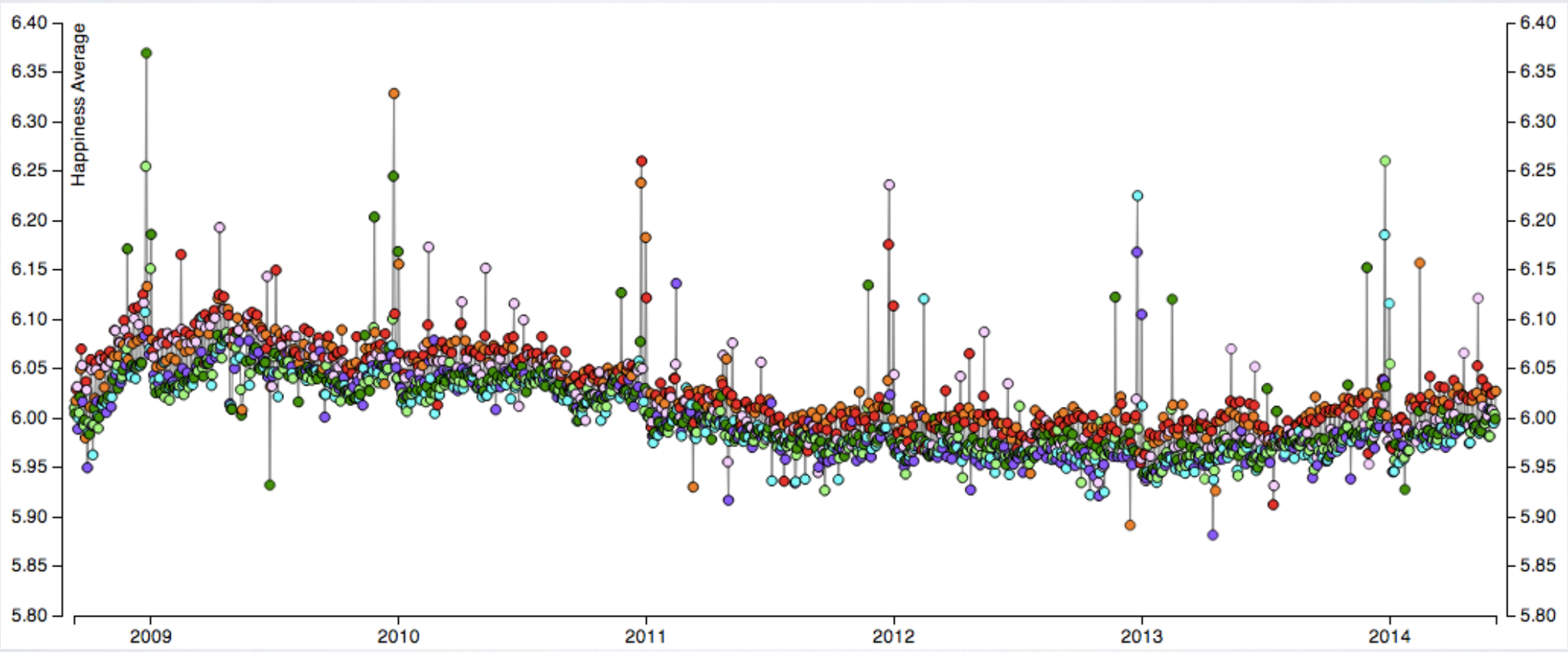
**Two main questions:**

- Measure the overall average happiness
- Explain the variation in happiness
  across different times and places

# HEDONOMETER

**Two main questions:**

- Measure the overall average happiness
- Explain the variation in happiness
  across different times and places



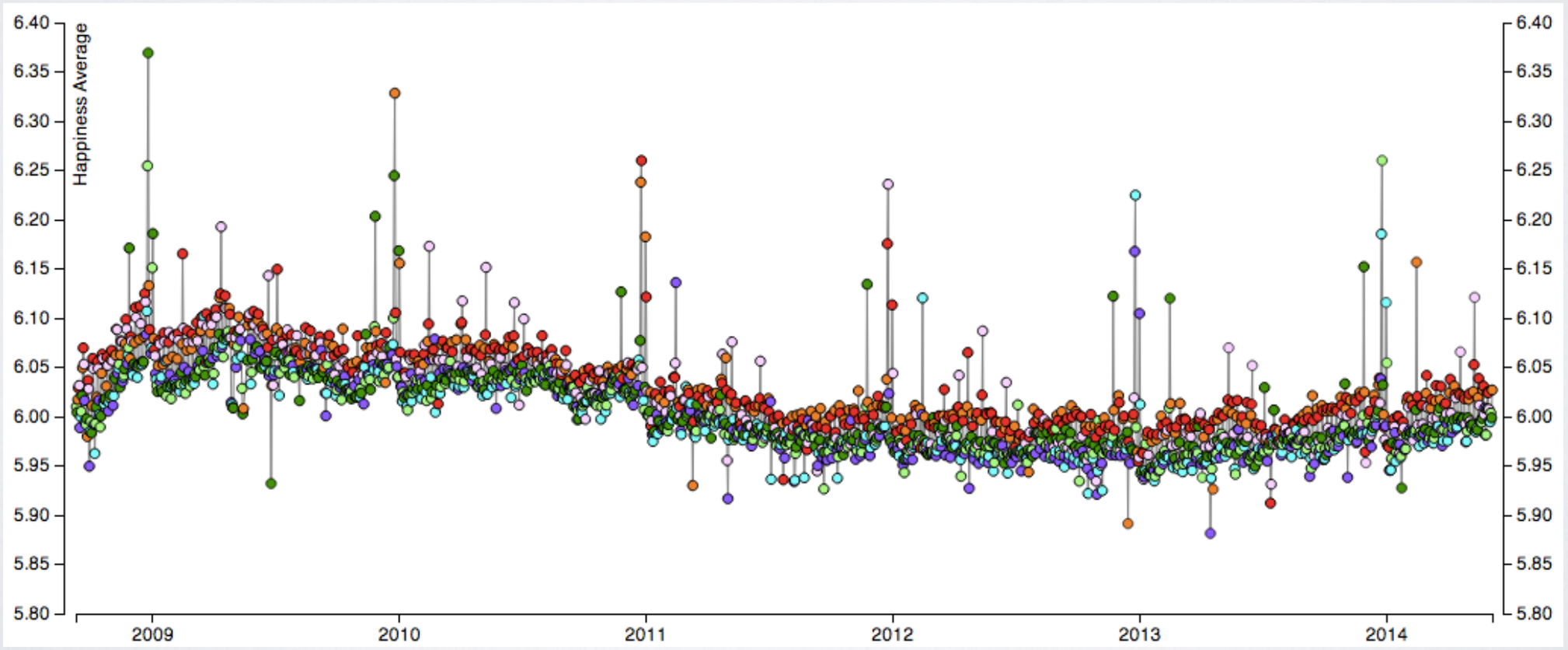
# HEDONOMETER

**Two main questions:**

- Measure the overall average happiness
- Explain the variation in happiness across different times and places

**Results**

- Well connected users write happier status updates
- The seven day week cycle is an historical and cultural artefact



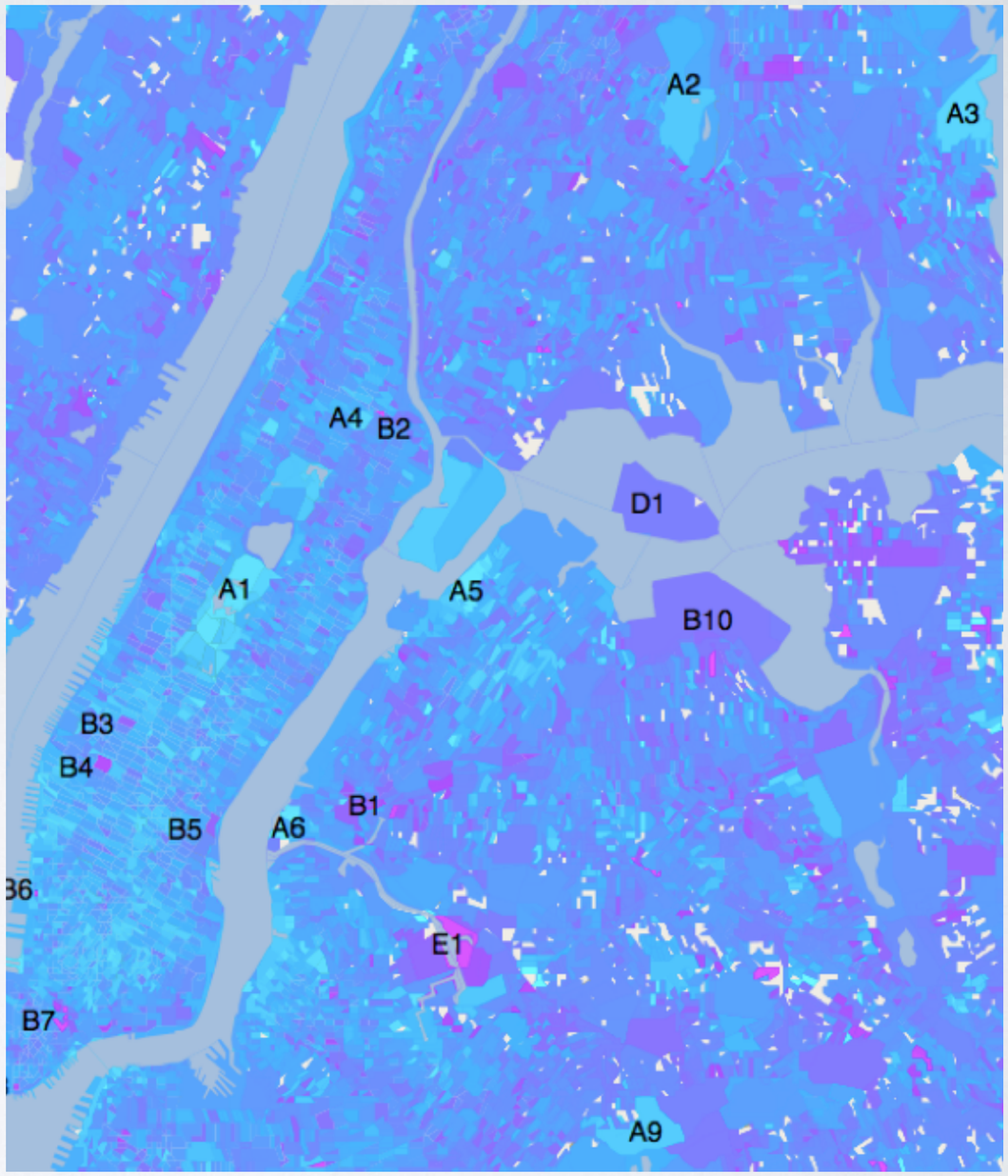
# SENTIMENT IN NYC

# SENTIMENT IN NYC

- Gauge public sentiment on extremely fine-grained spatial and temporal scales

# SENTIMENT IN NYC

- Gauge public sentiment on extremely fine-grained spatial and temporal scales



# SENTIMENT IN NYC

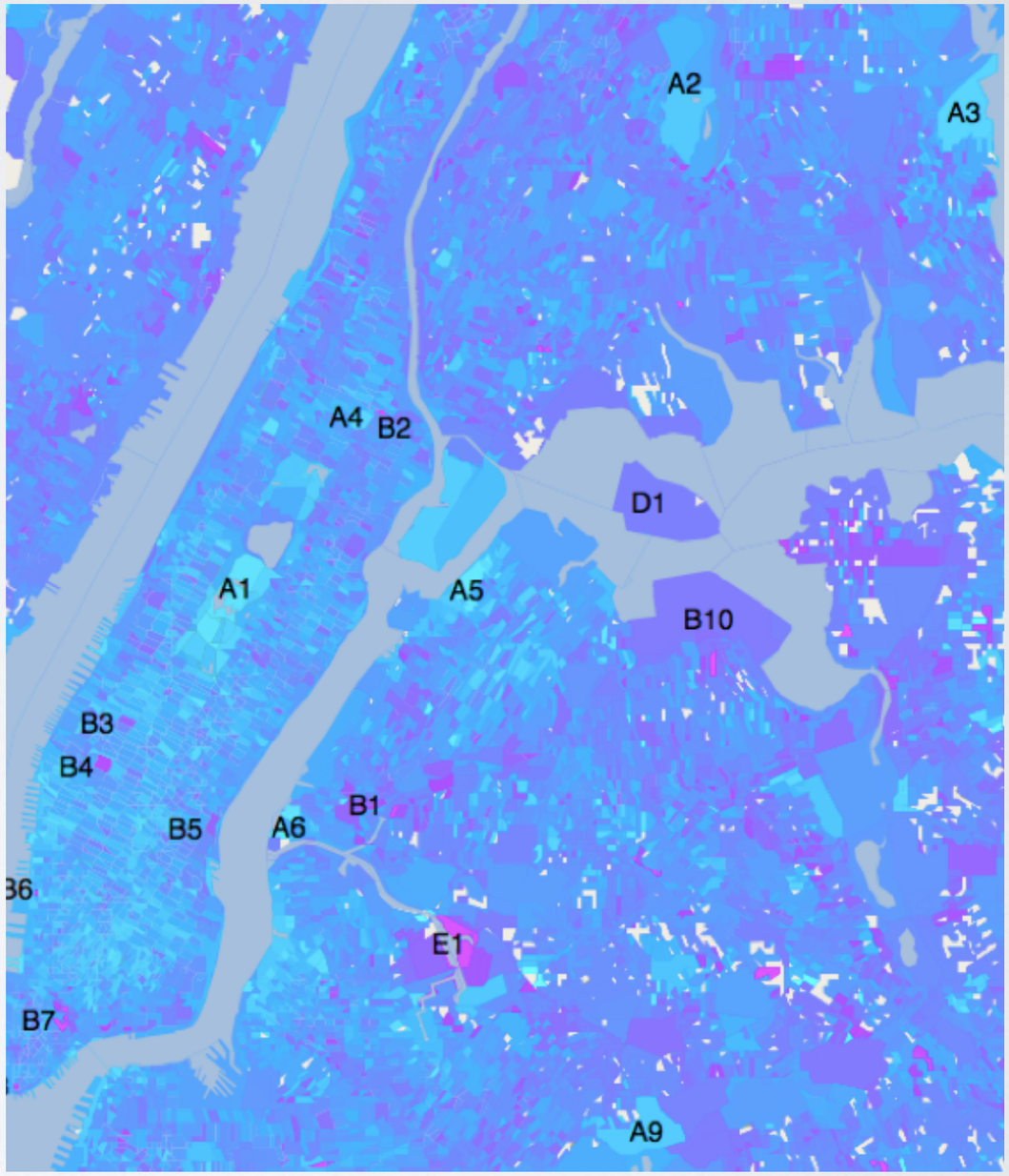
- Gauge public sentiment on extremely fine-grained spatial and temporal scales

- **Results:**

  - Sentiment progressively improves with proximity to Times Square.

  - Higher mood = parks

  - Lower mood = transportations / prisons



# FROM TWEETS TO POLLS

# FROM TWEETS TO POLLS

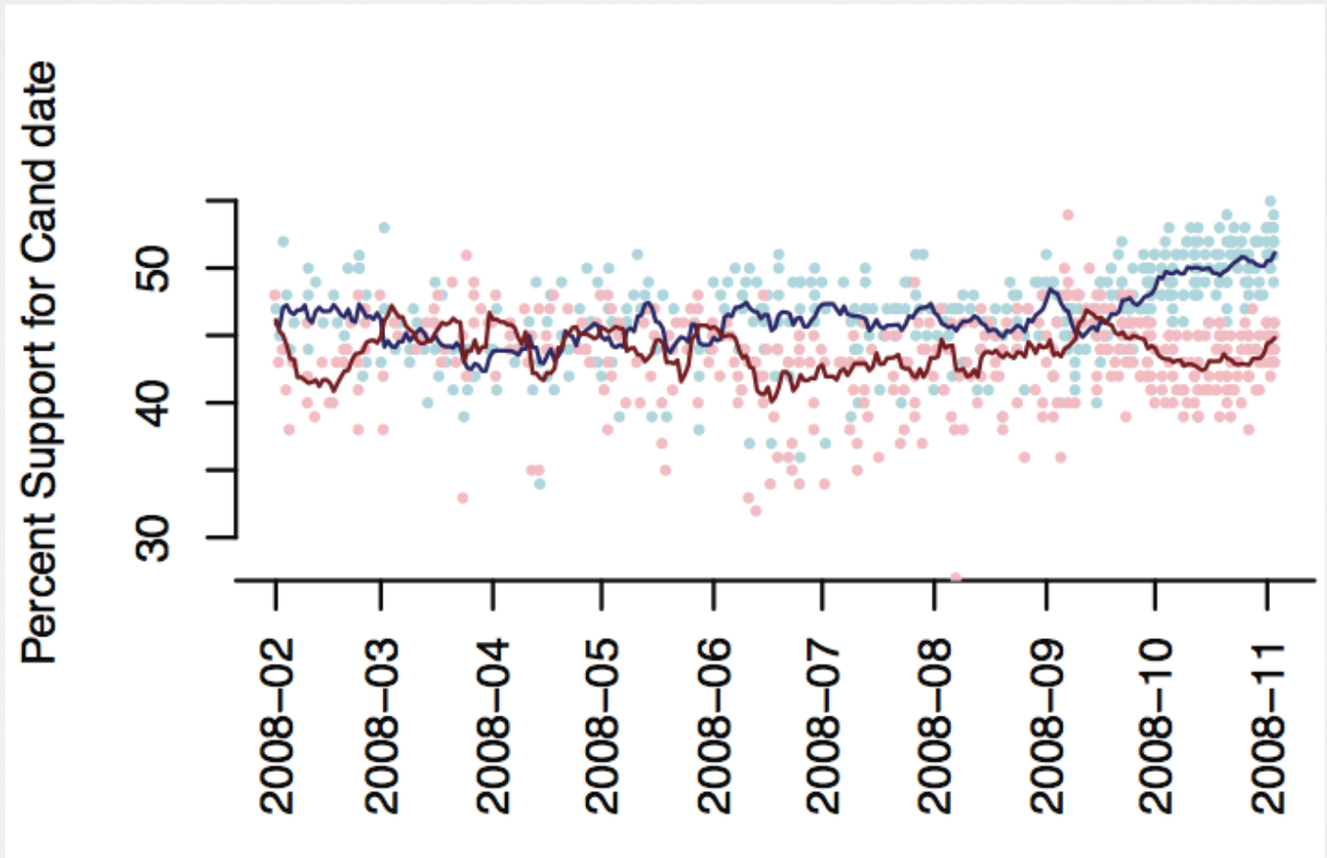
- Measures of public opinion derived from polls with sentiment

# FROM TWEETS TO POLLS

- Measures of public opinion derived from polls with sentiment
- **Result:** They demonstrated high correlation between polling data and political opinion

# FROM TWEETS TO POLLS

- Measures of public opinion derived from polls with sentiment
- **Result:** They demonstrated high correlation between polling data and political opinion



# CHARACTERISING GEOGRAPHIC VARIATION IN WELL- BEING

# CHARACTERISING GEOGRAPHIC VARIATION IN WELL- BEING

- They find the lexica and topics that correlate with life satisfaction

# CHARACTERISING GEOGRAPHIC VARIATION IN WELL- BEING

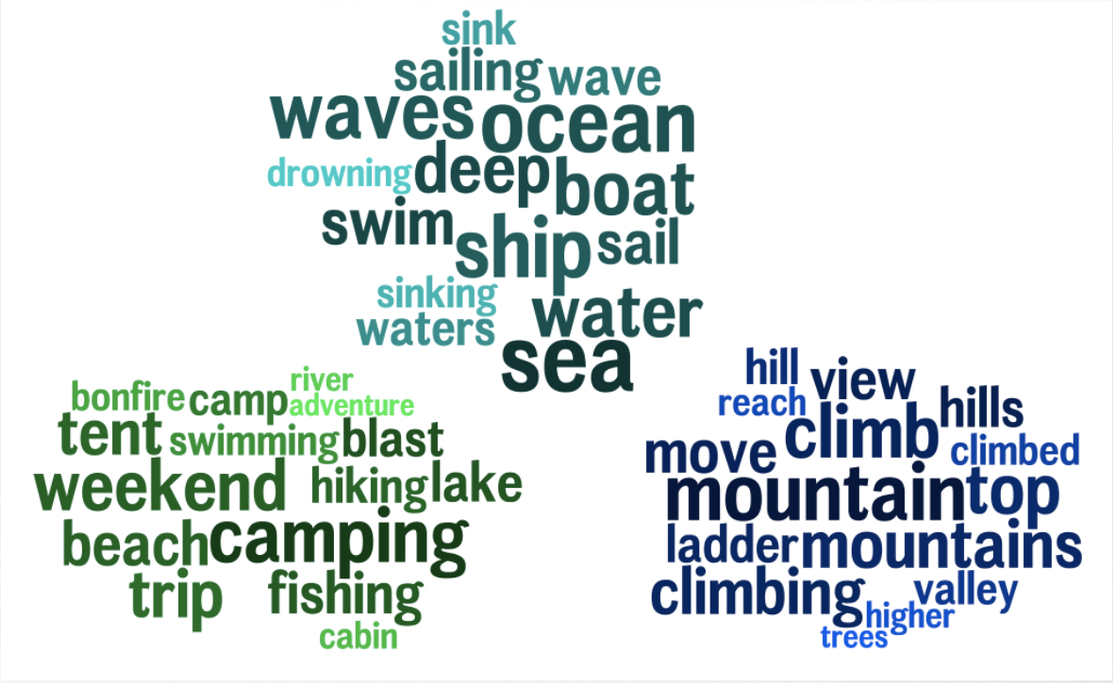
- They find the lexica and topics that correlate with life satisfaction

- They correlate the words within Twitter messages with the results for life satisfaction

# CHARACTERISING GEOGRAPHIC VARIATION IN WELL- BEING

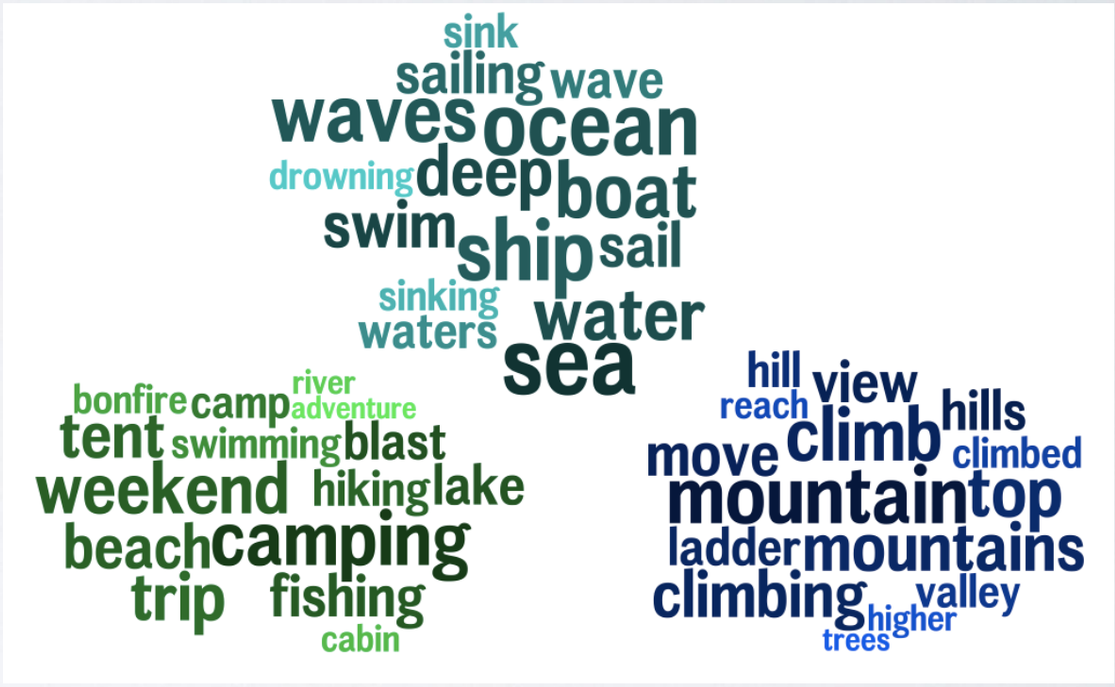
- They find the lexica and topics that correlate with life satisfaction

- They correlate the words within Twitter messages with the results for life satisfaction



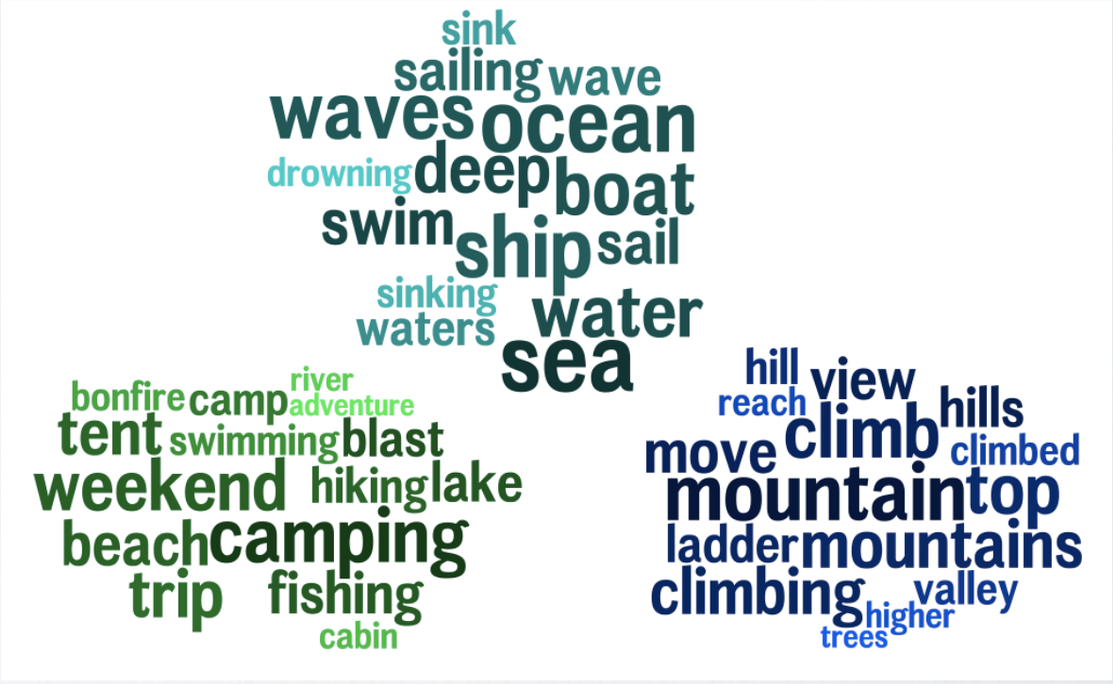
# CHARACTERISING GEOGRAPHIC VARIATION IN WELL- BEING

- They find the lexica and topics that correlate with life satisfaction
- They correlate the words within Twitter messages with the results for life satisfaction
- They run all the variables through a **linear regression**



# CHARACTERISING GEOGRAPHIC VARIATION IN WELL- BEING

- They find the lexica and topics that correlate with life satisfaction
- They correlate the words within Twitter messages with the results for life satisfaction
- They run all the variables through a **linear regression**
- **Result:** They can predict the happiness of one set of people from the tweets of other people



# METHODOLOGY
# &
# DATASETS

# TWEETS

# TWEETS

233 Million Tweets from 2012

# TWEETS

233 Million Tweets from 2012

# TWEETS

233 Million Tweets from 2012

1.8 Million Geotagged Tweets

# TWEETS

233 Million Tweets from 2012

1.8 Million Geotagged Tweets

# TWEETS

233 Million Tweets from 2012

1.8 Million Geotagged Tweets

0.5 Million U.S. Geotagged Tweets

# LABMT WORDS

# LABMT WORDS

- **10’233 words** rated form 1 (sad) to 9 (happy)
  - Approach used by Dodds et. al.
  - Each individual word rated by users

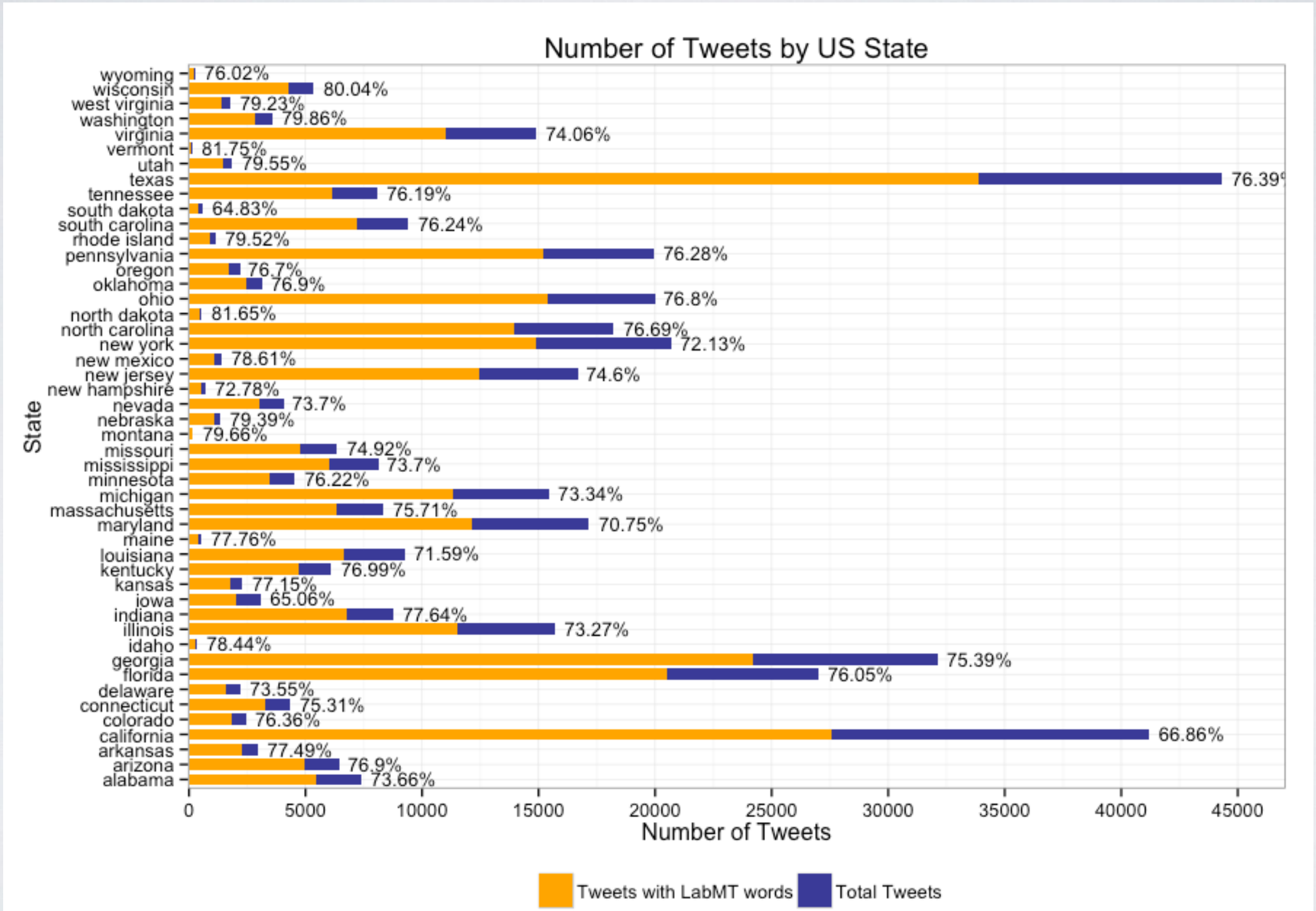
# LABMT WORDS

- **10’233 words** rated form 1 (sad) to 9 (happy)
  - Approach used by Dodds et. al.
  - Each individual word rated by users
- Approach used by Mitchell et. al.
  - Removed words in the middle ($4 < h_{avg} < 6$)
  - Removed ’N-word’
  - Removed state names

# LABMT WORDS

- **10’233 words** rated form 1 (sad) to 9 (happy)
  - Approach used by Dodds et. al.
  - Each individual word rated by users
  - Approach used by Mitchell et. al.
    - Removed words in the middle ($4 < h_{avg} < 6$)
    - Removed ’N-word’
    - Removed state names
- Final list with **3’722 words — Filter tweets!**

# TWEETS WITH LABMT WORDS



# TWEET HAPPINESS SCORE

Lyrics for
Michael Jackson's Billie Jean

"She was more like a beauty queen
from a movie scene.
⋮
And mother always told me,
be careful who you love.
And be careful of what you do
'cause the lie becomes the truth.

Billie Jean is not my lover,
She's just a girl who claims
that I am the one.
⋮

| ANEW words | $v_k$ | $f_k$ |
|---|---|---|
| $k$=1. love | 8.72 | 1 |
| 2. mother | 8.39 | 1 |
| 3. baby | 8.22 | 3 |
| 4. beauty | 7.82 | 1 |
| 5. truth | 7.80 | 1 |
| 6. people | 7.33 | 2 |
| 7. strong | 7.11 | 1 |
| 8. young | 6.89 | 2 |
| 9. girl | 6.87 | 4 |
| 10. movie | 6.86 | 1 |
| 11. perfume | 6.76 | 1 |
| 12. queen | 6.44 | 1 |
| 13. name | 5.55 | 1 |
| 14. lie | 2.79 | 1 |

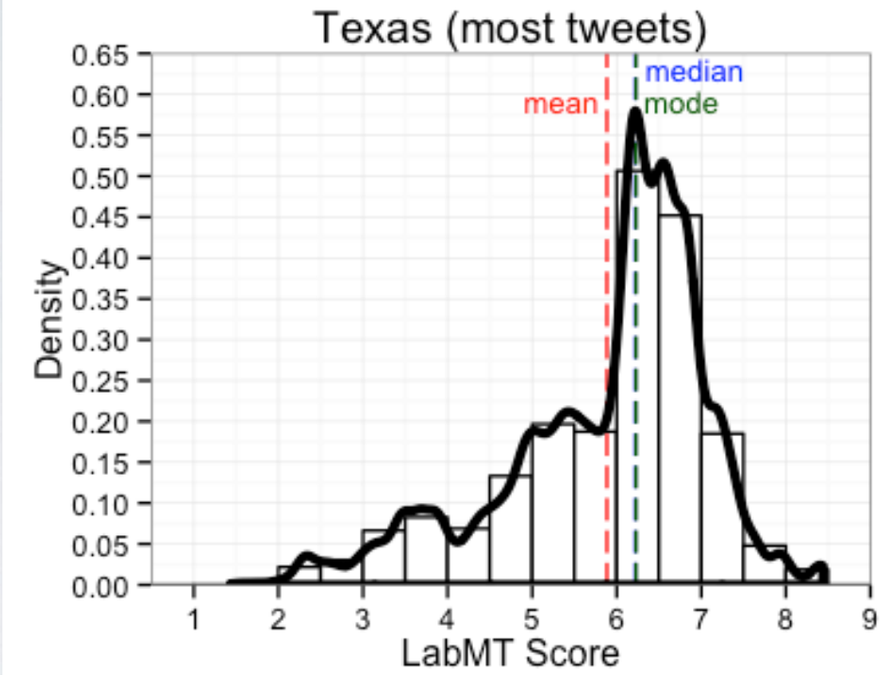
$$v_{\text{text}} = \frac{\sum_k v_k f_k}{\sum_k f_k}$$

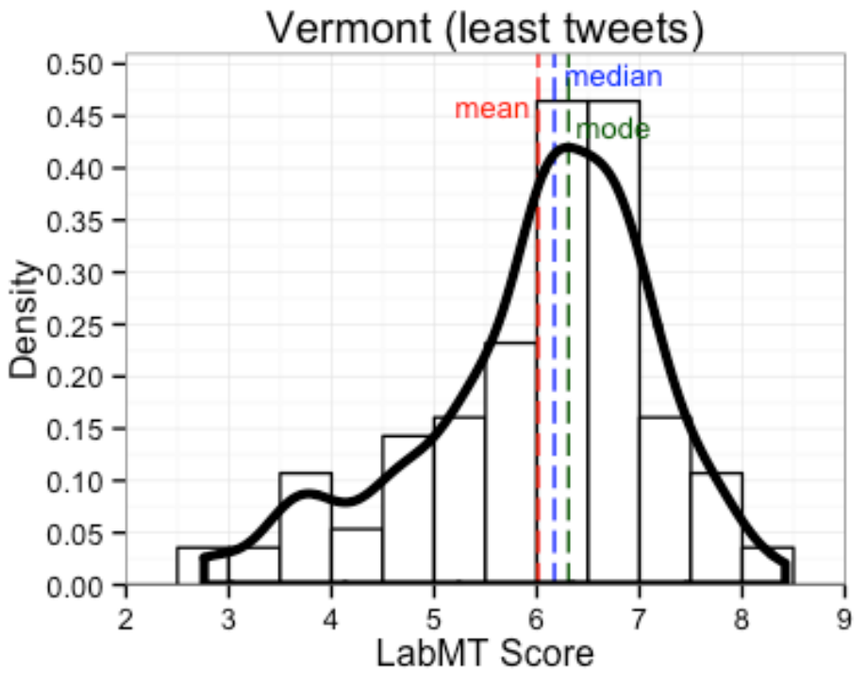
$$v_{\text{Billie Jean}} = 7.1$$

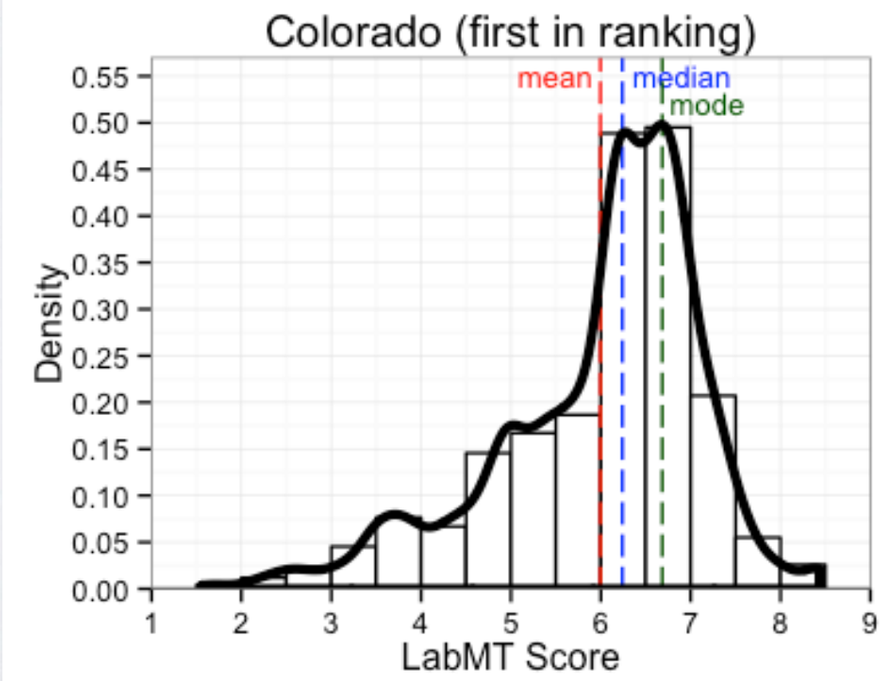
$$v_{\text{Thriller}} = 6.3$$

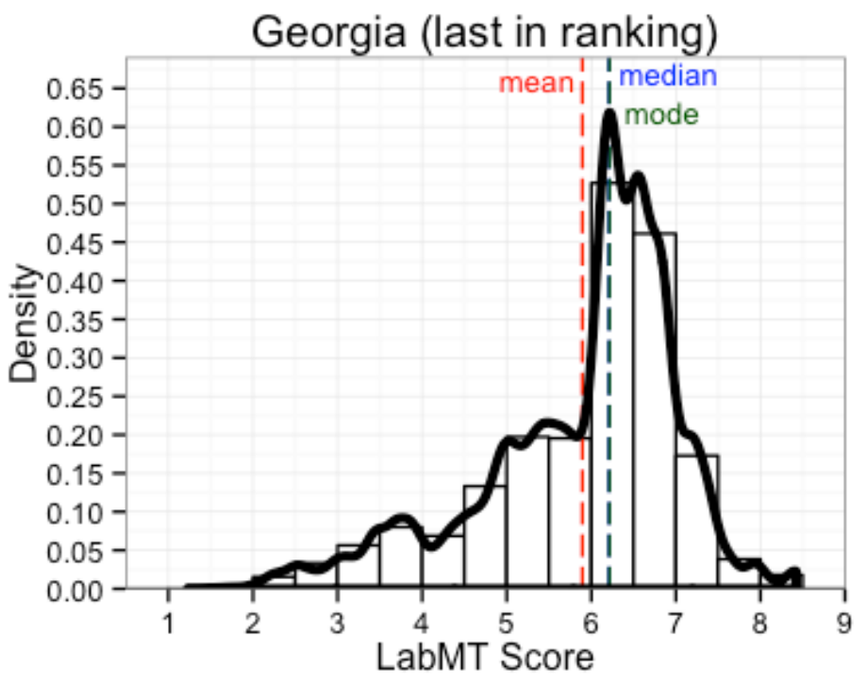
$$v_{\text{Michael Jackson}} = 6.4$$

# SCORES BY STATE



# SCORE BY STATE

# SCORE BY STATE

- *corr(gallup, mean) = 0.19*

# SCORE BY STATE

- *corr(gallup, mean) = 0.19*
- *corr(gallup, median) = 0.03*

# SCORE BY STATE

- *corr(gallup, mean) = 0.19*
- *corr(gallup, median) = 0.03*
- ***corr(gallup, mode) = 0.45***

# SCORE BY STATE

- *corr(gallup, mean) = 0.19*
- *corr(gallup, median) = 0.03*
- ***corr(gallup, mode) = 0.45***

# SCORE BY STATE

- *corr(gallup, mean) = 0.19*
- *corr(gallup, median) = 0.03*
- ***corr(gallup, mode) = 0.45***

- *mode* with:
  - range from 1 to 9 (LabMT word scale)
  - smooth factor of 0.80

# OTHER WORK

# OTHER WORK

- Also used an **Spanish** word list
  - Increased the total number of tweets in 40’000

# OTHER WORK

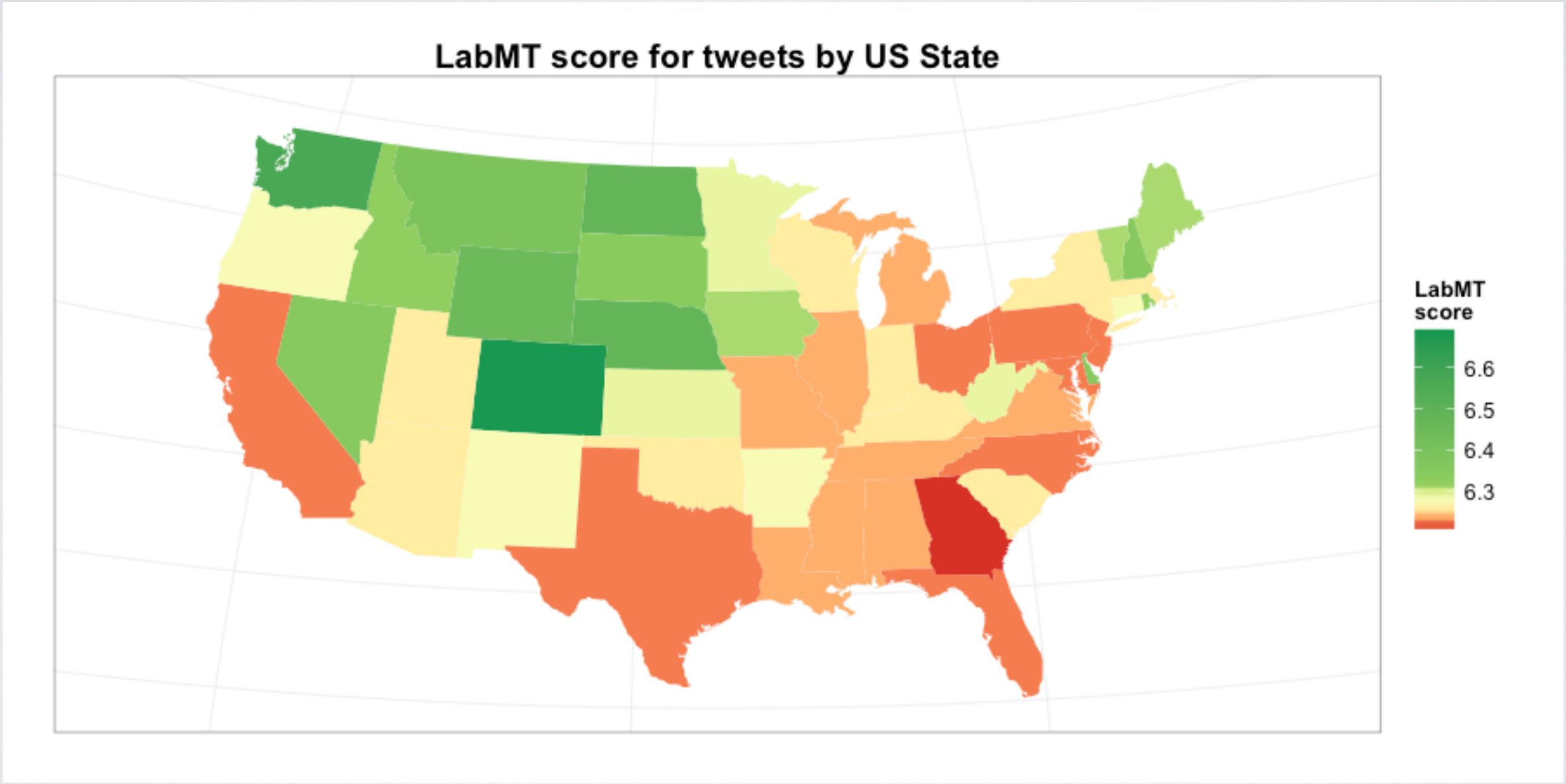
- Also used an **Spanish** word list
  - Increased the total number of tweets in 40'000
- U.S. States Polygon Areas — '*maps*' R package
- Compare lon/lat coordinates with polygons
- '*over*' function from '*sp*' R package

# OTHER WORK

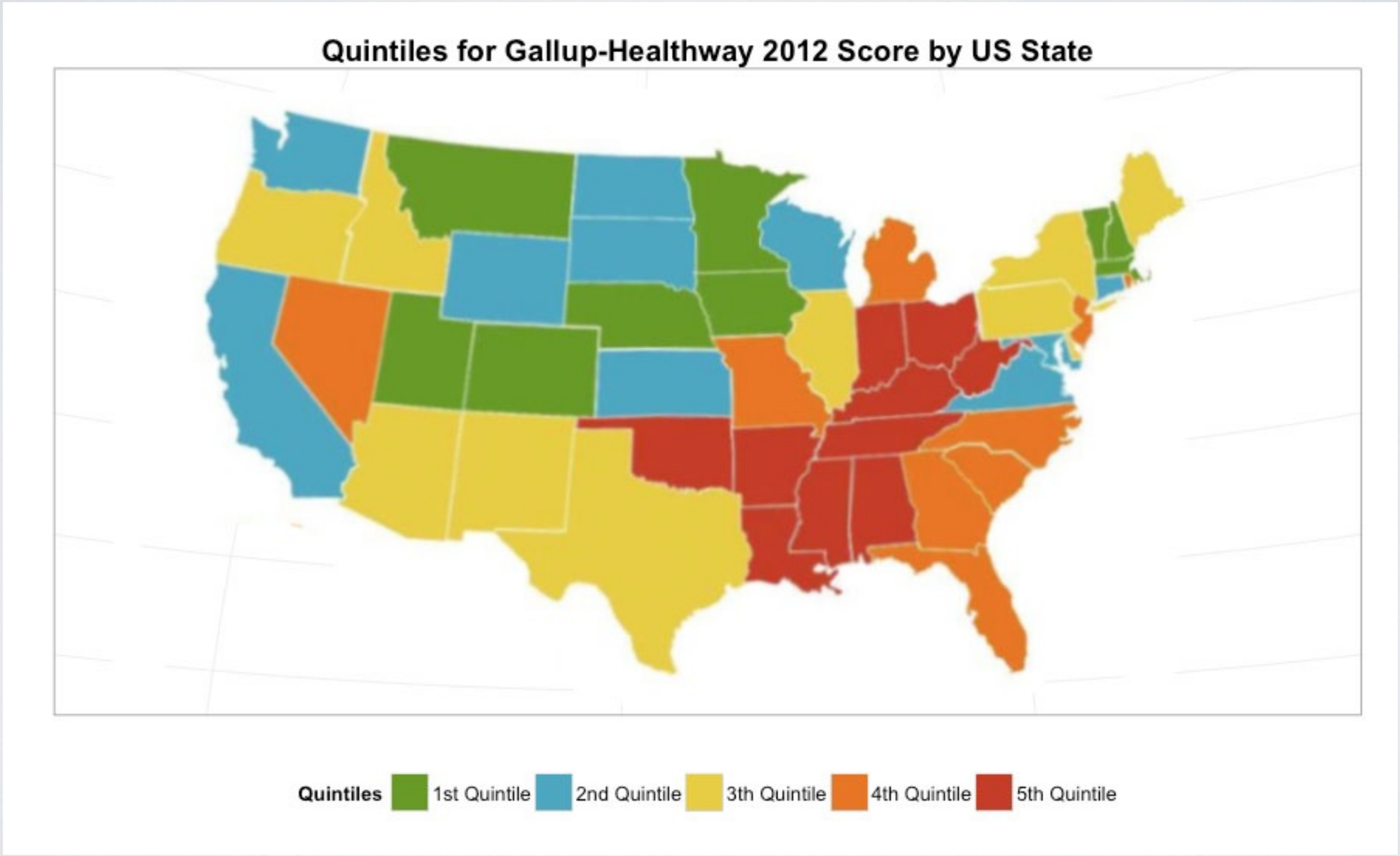
- Also used an **Spanish** word list
  - Increased the total number of tweets in 40'000
- U.S. States Polygon Areas — '*maps*' R package
  - Compare lon/lat coordinates with polygons
  - '*over*' function from '*sp*' R package
- *Gallup-Healthway* Well-Being Index 2012 Report
  - Rescaled from percentage scale (0-100) to 1-9 LabMT scale

# RESULTS

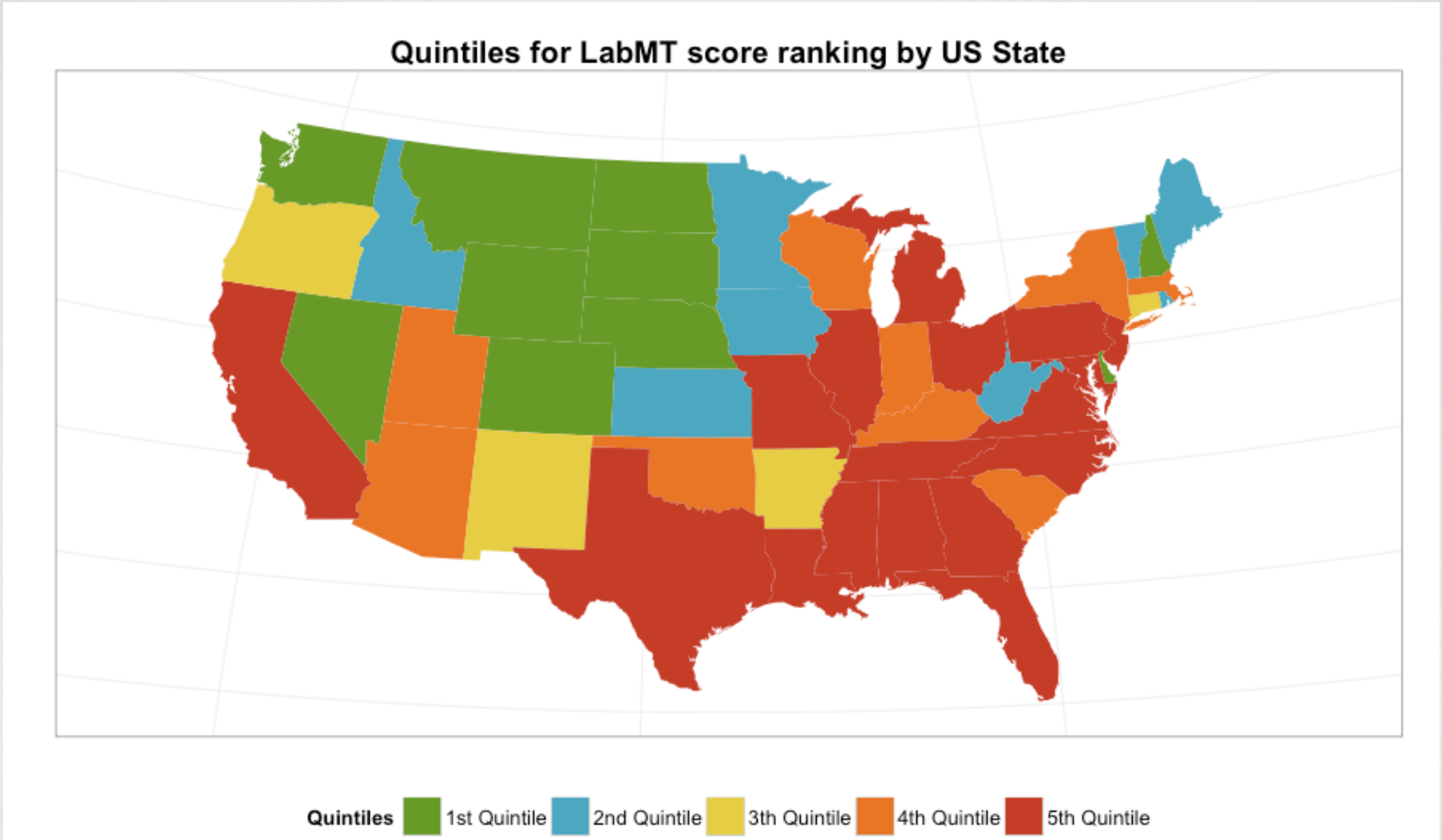
# SCORE BY STATE



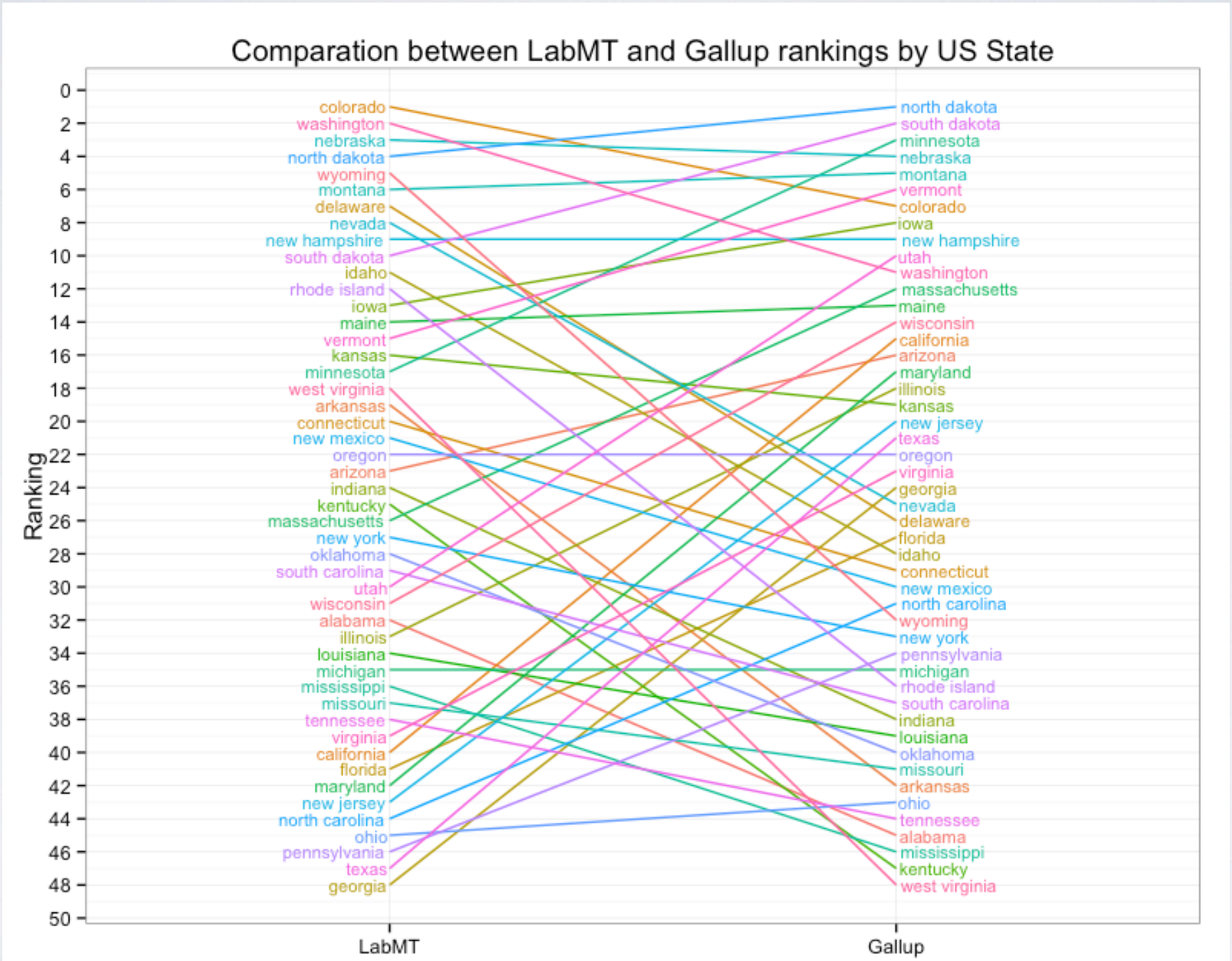
# SCORE WITH QUINTILES BY STATE



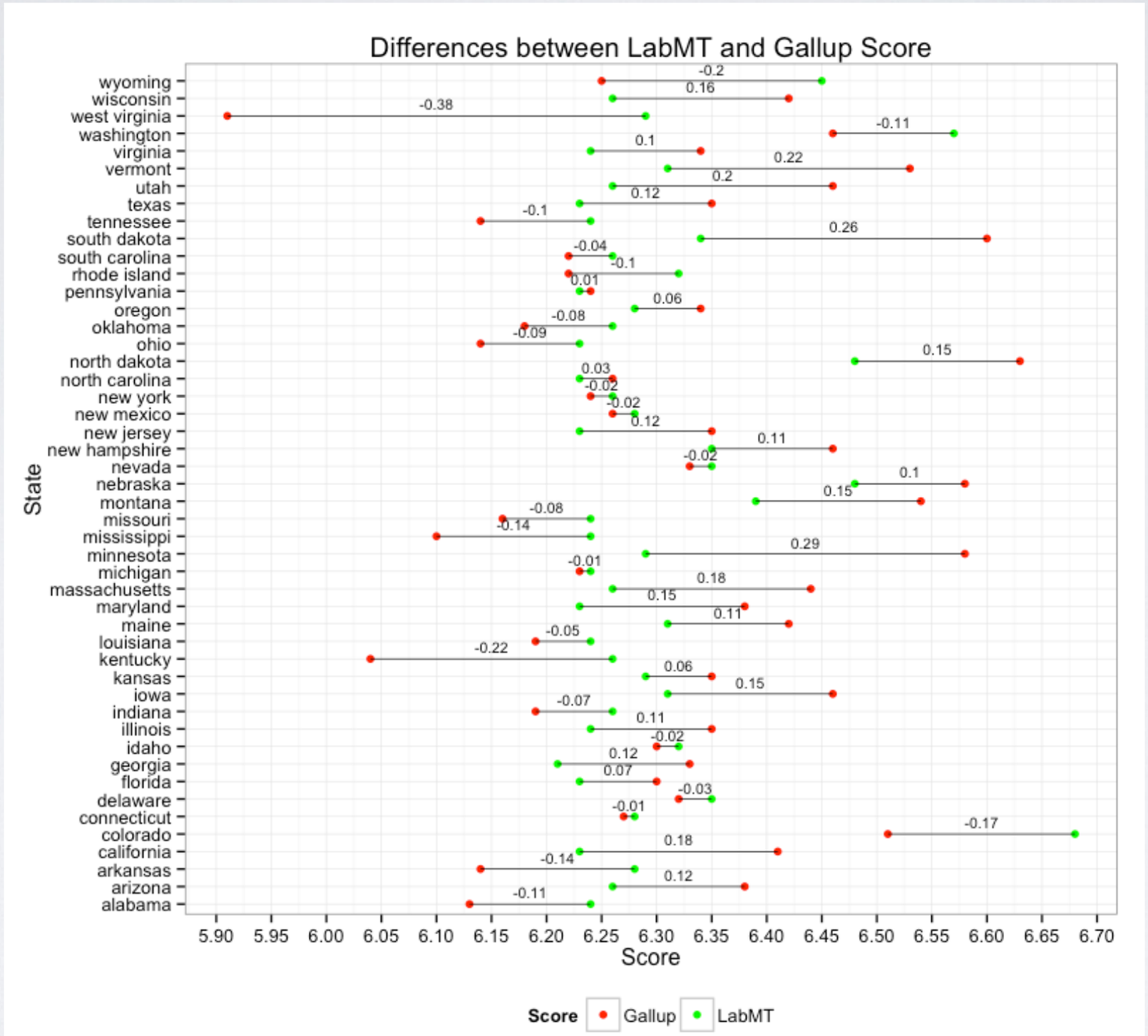
# SCORE WITH QUINTILES BY STATE



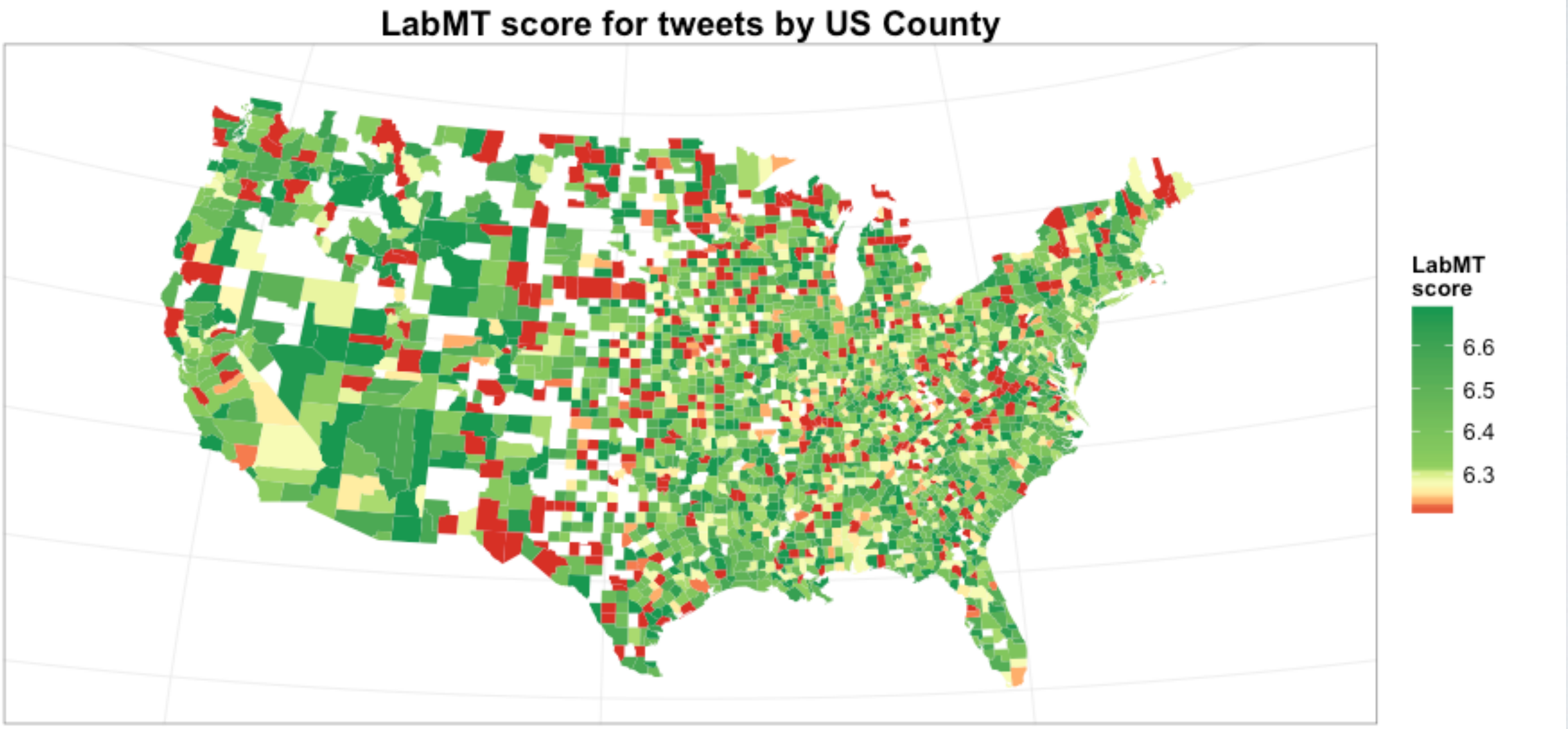
# LABMT VS GALLUP-HEALTHWAY



# LABMT VS GALLUP-HEALTHWAY



# SCORE BY COUNTY



# DISCUSSION
# &
# FUTURE WORK

# DISCUSSION

# DISCUSSION

- Allows more detailed results than survey methods

# DISCUSSION

- Allows more detailed results than survey methods
- Good correlation value between our study and Gallup-Healthway
  - ***corr(gallup, mode) = 0.45***

# DISCUSSION

- Allows more detailed results than survey methods
- Good correlation value between our study and Gallup-Healthway
  - ***corr(gallup, mode) = 0.45***
- Very high correlation value between ranking and number of tweets
  - ***corr(ranking, tweets) = 0.77***
  - ***corr(score, tweets) = -0.48***

# DISCUSSION

- Allows more detailed results than survey methods
- Good correlation value between our study and Gallup-Healthway
  - ***corr(gallup, mode) = 0.45***
- Very high correlation value between ranking and number of tweets
  - ***corr(ranking, tweets) = 0.77***
  - ***corr(score, tweets) = -0.48***

| Ranking | Avg. Tweets |
|---|---|
| 1-12th | 1376 |
| 13-24th | 4391 |
| 25-36th | 5344 |
| 37-48th | 14479 |

# FUTURE WORK

# FUTURE WORK

- Increase number of tweets
  - For non-geotagged tweets assume *user city*

# FUTURE WORK

- Increase number of tweets
  - For non-geotagged tweets assume *user city*
- Expand Schwartz et. al. work
  - Add number of tweets to regression model

# THANK YOU!

# QUESTIONS?