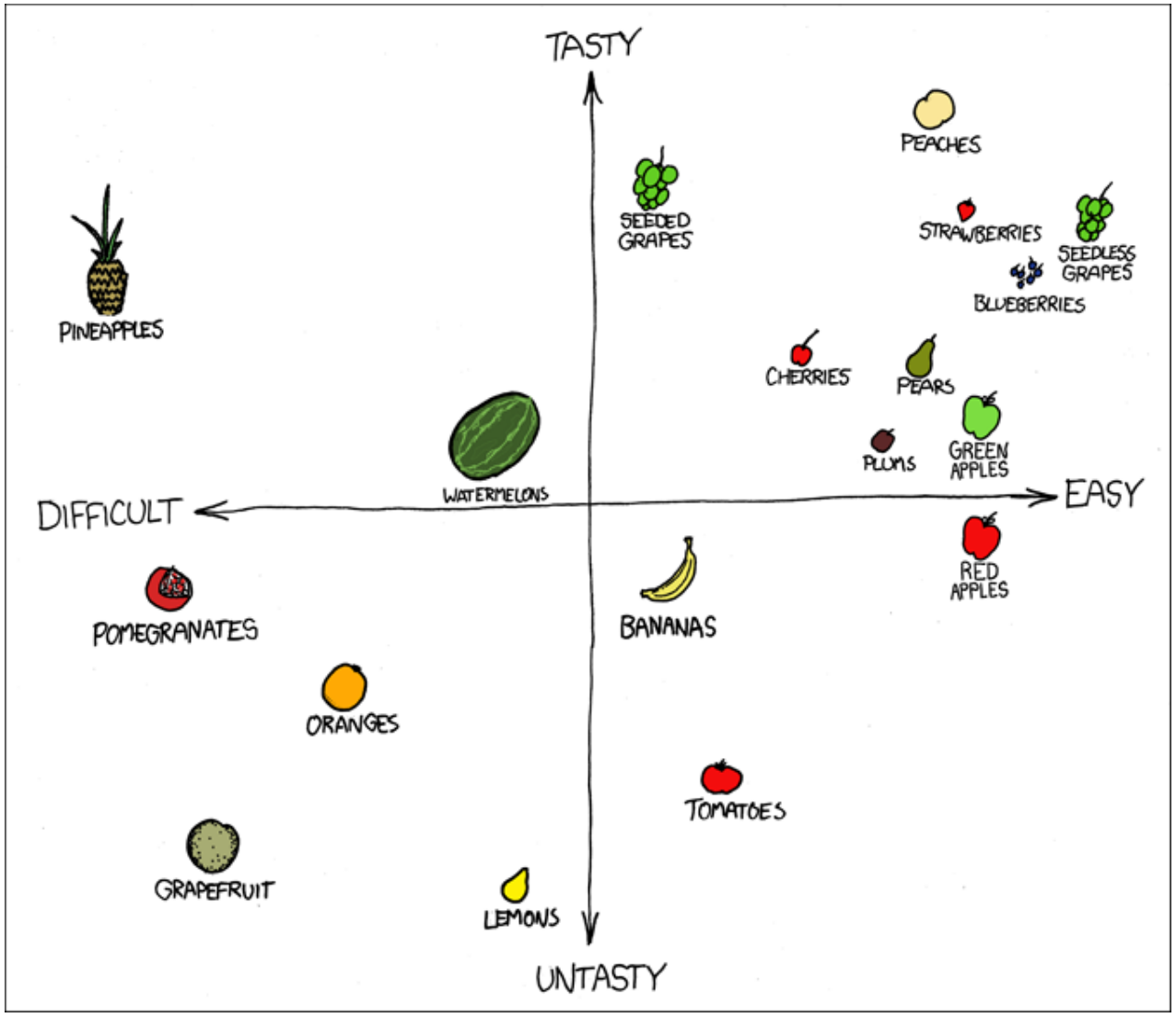
Analyzing the Language of Food on Social Media

Stephen Kobourov
Department of Computer Science
University of Arizona

Why Food?

Research inspired by xkcd comics



Food is

- Art [Lernert & Sander, *Cubes*, 2015]

Food is

- Art [Lernert & Sander, *Cubes*, 2015]



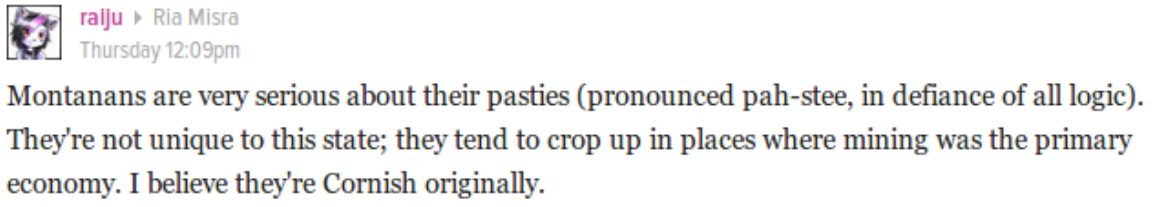
- #foodporn



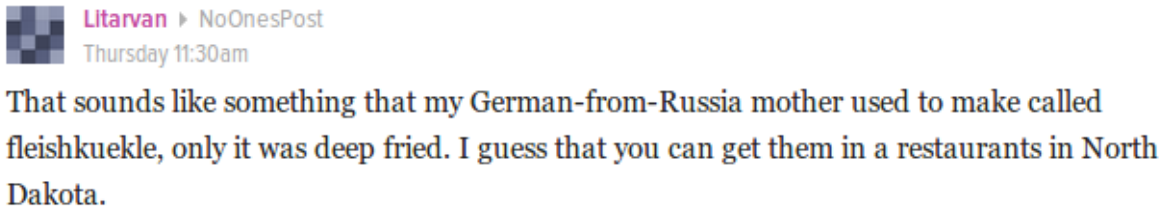
Why Study The Language of Food?

Our diets reflect lifestyles, habits, upbringing and cultural heritage:

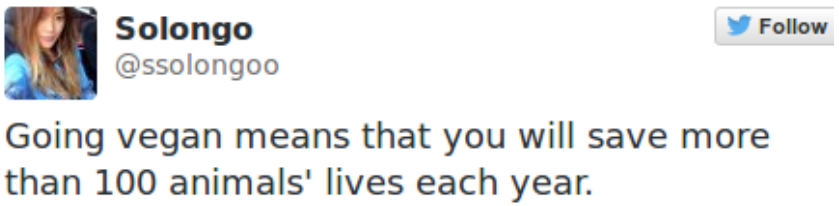
- geographic



- cultural



- political



But our diets also shape who we will be, by impacting health, well-being, lifestyles, ...

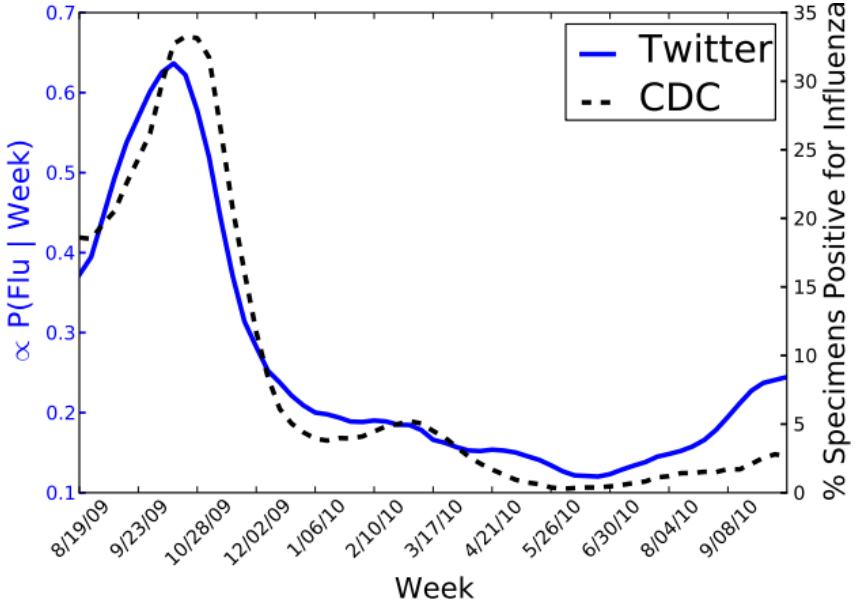
Why Twitter?

Serious shortcomings

- short, sparse, slang, self-reported, ...

But very tempting, very large, freely available, accessible dataset

- used across ethnic, gender, age, socio-economic groups
- geographic linguistic analysis [Eisenstein et al. 2010]
- flu, allergies prediction using Twitter [Paul and Dredze 2011]



Twitter and Food

- Obesity is socially contagious [Christakis and Fowler 2007]

Twitter and Food

- Obesity is socially contagious [Christakis and Fowler 2007]

Twitter and Food

- Obesity is socially contagious [Christakis and Fowler 2007]

Twitter and Food

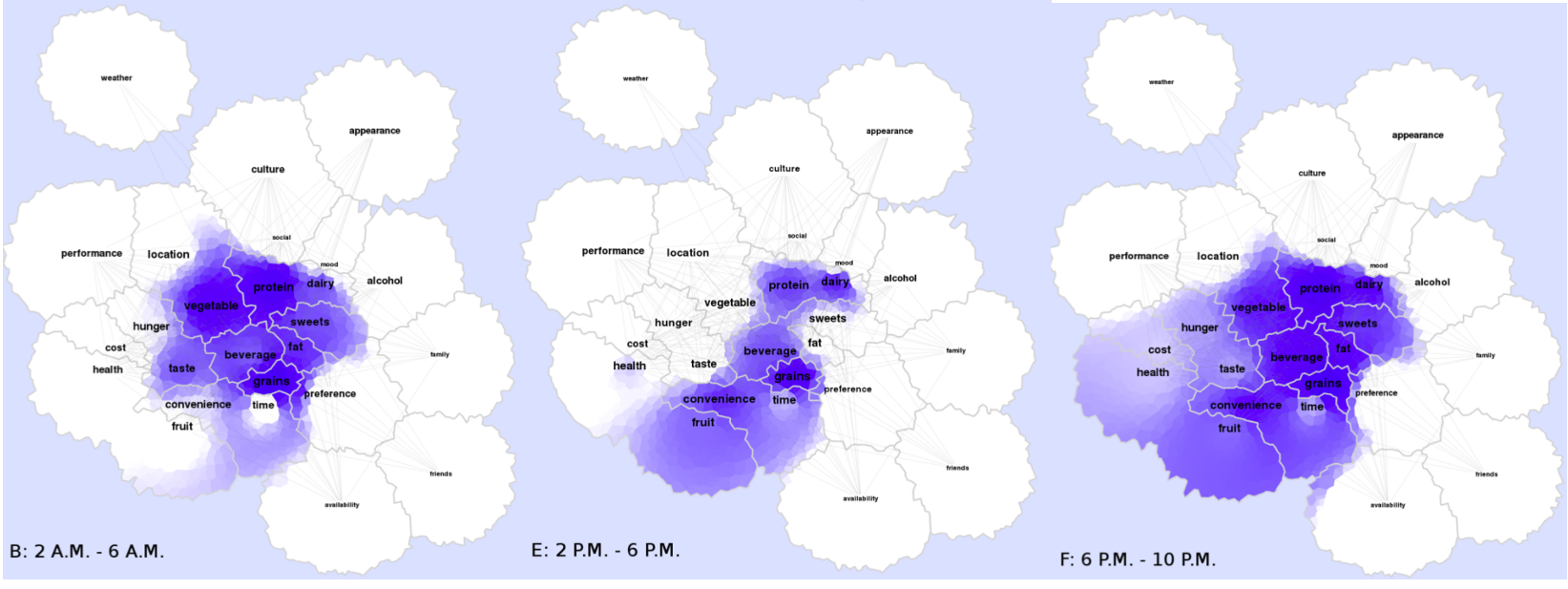
- Obesity is socially contagious [Christakis and Fowler 2007]

- 1 billion tweets used to measure and predict well-being [Schwartz et al. 2013]
- Social networking strategies can help people lose weight [Ashrafian et al. 2014]
- Twitter can be a greater source of positive influence for weight loss than family or friends [Pagoto et al. 2014]

Twitter and Food

Recent work explores food logging and visualization via Twitter [Hingle et al. 2013]

- 50 participants
- 2862 hashtags (foods and reasons for eating/overeating)
- reasons for overeating: #social, #convenience, #taste
- track patterns over time



Goals

- Use more data, albeit less specific
- Analyze predictive features of language of food
- Identify textual features with most predictive power
- Visualize results in geographical and temporal dimensions
- Predict diabetes and obesity rates for communities

*Rudolof II, by Arcimboldo c. 1590

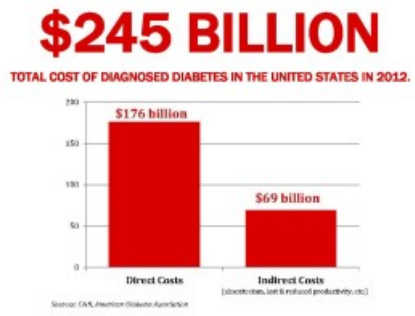
Motivation (scary statistics)

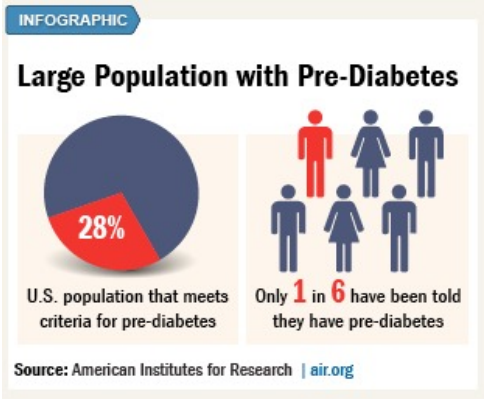
Why obesity and diabetes?

- 86 million Americans have pre-diabetes
- 70% of these pre-diabetics will develop Type 2 diabetes
- Yet 90% of these individuals are not aware of this risk
- Estimated annual diabetes costs $245 billion
- 33% of untreated diabetics die of it
- 80% of Type 2 diabetes is preventable!

What can we do?

- Predict diabetes for individuals
- Identify people at risk for diabetes
- Attempt to intervene to prevent diabetes

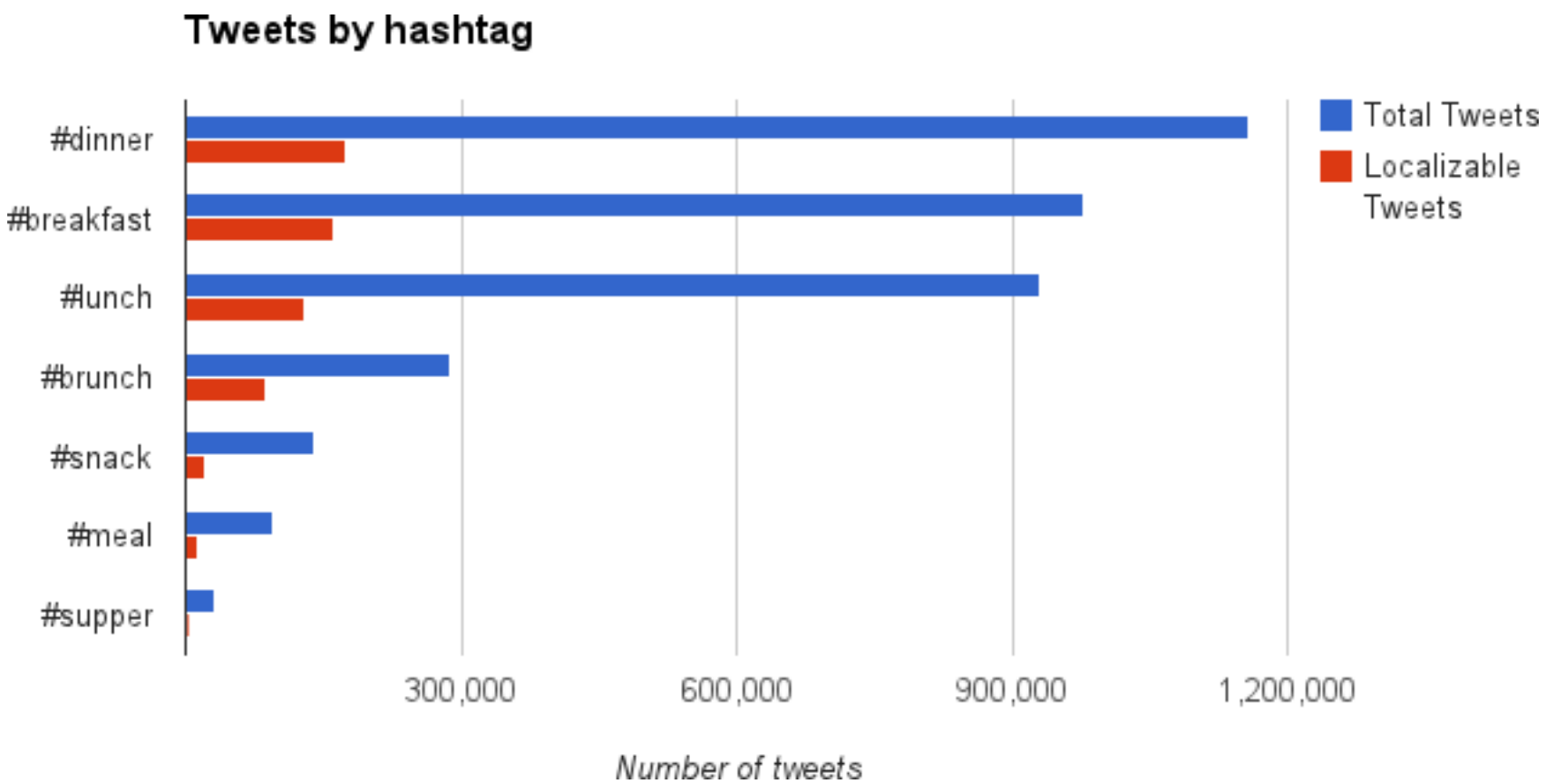




*Source: CDC

Tweet Corpus

- Collect meal-related tweets: `breakfast`, `lunch`, `dinner`, ...
- 3.5 million tweets (October 2013 - May 2014)
- Average tweet length: 8.7 words
- 30 million words, 1.5 million unique
- 560,000 tweets (16%) normalized to a US state



State Trends: not the most popular...

The most misinterpreted figure from our paper: from the Washington Post, the Guardian and Slate to Fox News and the Daily Mail

State Trends: the most popular is boring...

The most popular term in nearly every state is chicken...

TF-IDF Ranking

Term Frequency – Inverse Document Frequency

- measures importance of a word for a document in a collection
- importance grows with freq. of the word in the document (tf)
- but is offset by the freq. of the word in the corpus (idf)

TF-IDF and the Language Maps

What is the 2nd most commonly spoken language in the US?

*Data from Census Bureau American Community Survey and maps from Slate.

TF-IDF and the Language Maps

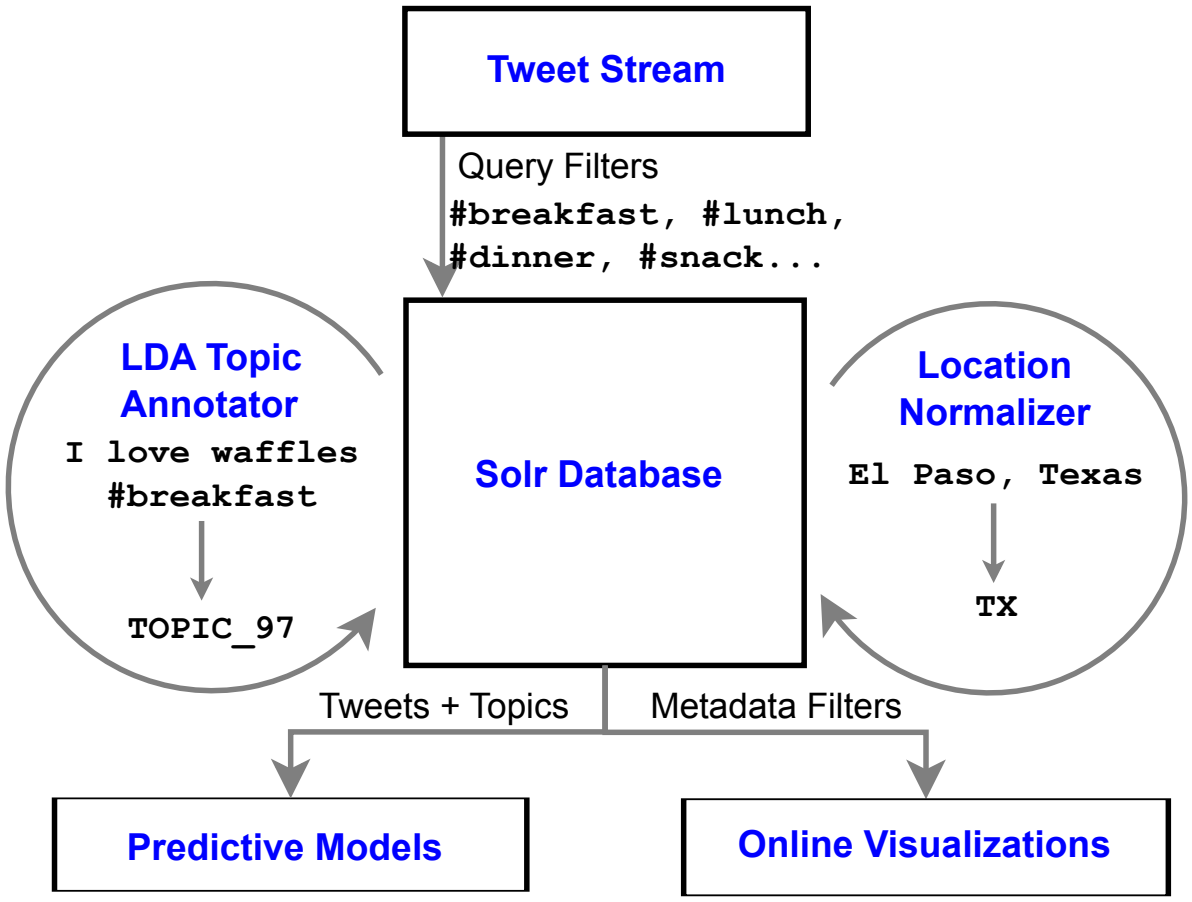
What is the 3rd most commonly spoken language in the US?

*Data from Census Bureau American Community Survey and maps from Slate.

Predictive Task Goals

- Predict diabetes and obesity rates for the US states
- Identify textual features with most predictive power
- Visualize results in geographical and temporal dimensions

Collecting, Analyzing, and Visualizing Tweets

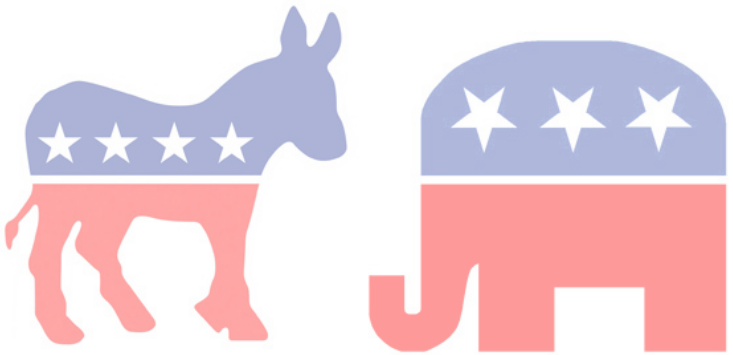


Collect tweets from Twitter API; store and query using Apache Solr

Prediction Tasks

Using the tweets for a state, predict:

- Diabetes rate: above or below US median?
- Overweight rate: above or below US median for high BMI?
- Political tendency: more Republican or Democratic votes?



Prediction Tasks

Using the tweets for a state, predict:

- Diabetes rate: above or below US median?
- Overweight rate: above or below US median for high BMI?
- Political tendency: more Republican or Democratic votes?

Diabetes data

- diabetes rate: min 7% (AK), median 9.7%, max 13% (WV)

.....[..x..].....

Overweight data (BMI)

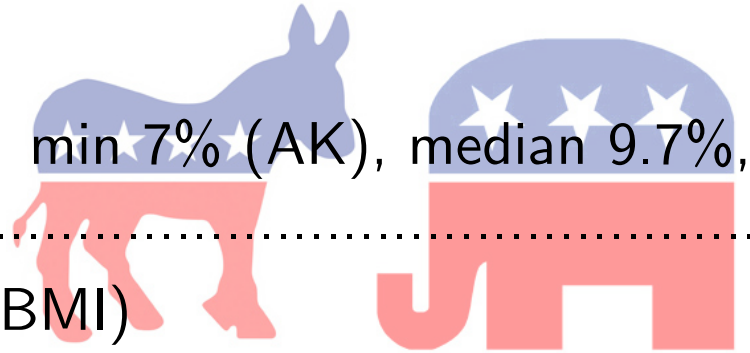
- overweight rate: min 52% (DC), median 64.2%, max 70% (LA)

.....[.....x.....].....

Democratic political tendency (2008-2013)

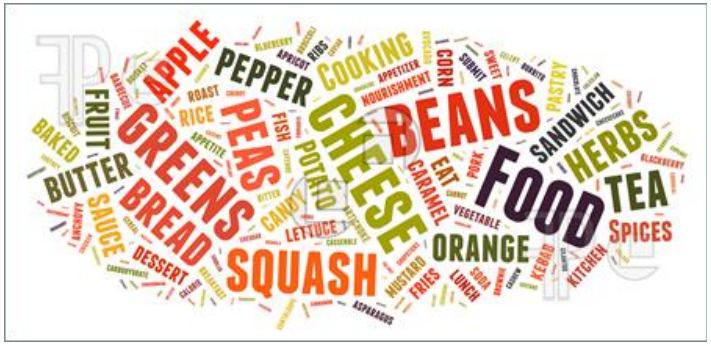
- Democratic rate: min 27% (WY), median 51%, max 92% (DC)

.....[.....x.....].....



Lexical Features

- All words (105,125 words that appear > 1 in localized tweets)
- Hashtags (64,037 words)
- Just food words (809 words related to food and meals)
- Topics via Latent Dirichlet Allocation (LDA)
- Various combinations



LDA Topics

- *Japanese:* `ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood`

LDA Topics

- *Japanese:* `ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood`
- *Mexican:* `mexican tacos burrito salsa nachos chicken homemade delicious guacamole chips enchiladas`

LDA Topics

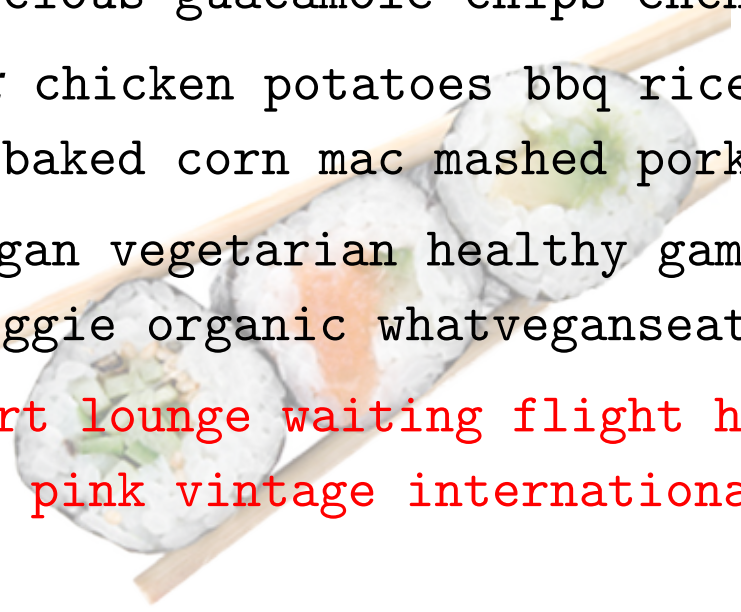
- *Japanese:* `ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood`
- *Mexican:* `mexican tacos burrito salsa nachos chicken homemade delicious guacamole chips enchiladas`
- *American Diet:* `chicken potatoes bbq rice cheese fried beans potato baked corn mac mashed pork steak`

LDA Topics

- *Japanese:* ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood
- *Mexican:* mexican tacos burrito salsa nachos chicken homemade delicious guacamole chips enchiladas
- *American Diet:* chicken potatoes bbq rice cheese fried beans potato baked corn mac mashed pork steak
- *Vegetarian:* vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum

LDA Topics

- *Japanese:* ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood
- *Mexican:* mexican tacos burrito salsa nachos chicken homemade delicious guacamole chips enchiladas
- *American Diet:* chicken potatoes bbq rice cheese fried beans potato baked corn mac mashed pork steak
- *Vegetarian:* vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum
- *Airport:* airport lounge waiting flight home my party yumyum purple pink vintage international modern sleepy



LDA Topics

- *Japanese:* ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood
- *Mexican:* mexican tacos burrito salsa nachos chicken homemade delicious guacamole chips enchiladas
- *American Diet:* chicken potatoes bbq rice cheese fried beans potato baked corn mac mashed pork steak
- *Vegetarian:* vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum
- *Airport:* airport lounge waiting flight home my party yumyum purple pink vintage international modern sleepy
- *After Work:* time so up just after work day my now home today out last all go night not some back

LDA Topics

- *Japanese:* ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood
- *Mexican:* mexican tacos burrito salsa nachos chicken homemade delicious guacamole chips enchiladas
- *American Diet:* chicken potatoes bbq rice cheese fried beans potato baked corn mac mashed pork steak
- *Vegetarian:* vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum
- *Airport:* airport lounge waiting flight home my party yumyum purple pink vintage international modern sleepy
- *After Work:* time so up just after work day my now home today out last all go night not some back
- *First Person Casual:* my i lol up wings time some bout da good bomb like chicken

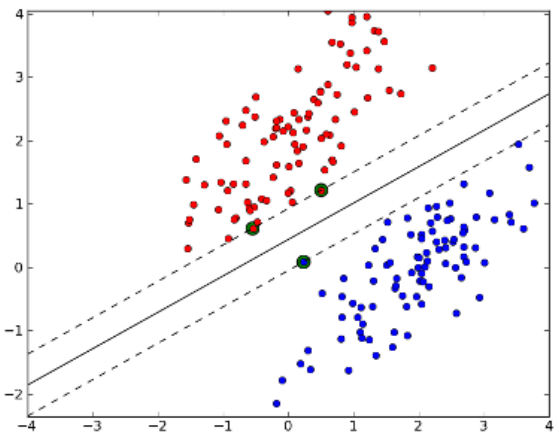
LDA Topics

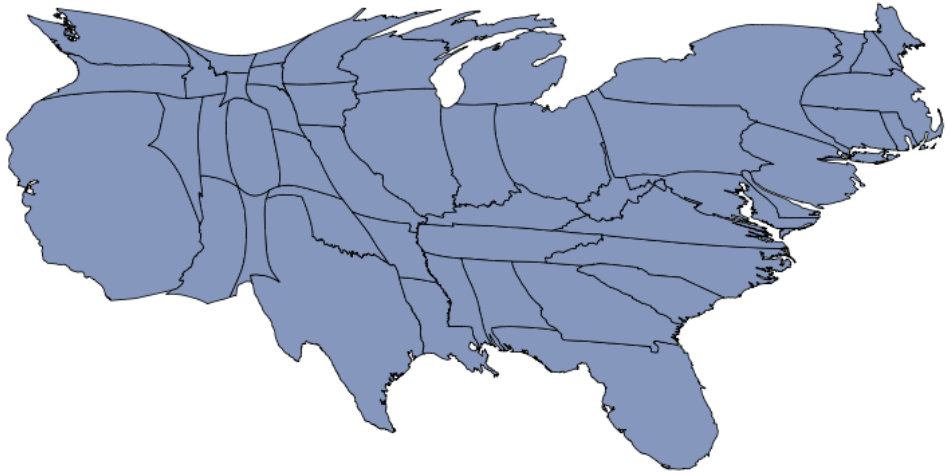
- *Japanese:* ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood
- *Mexican:* mexican tacos burrito salsa nachos chicken homemade delicious guacamole chips enchiladas
- *American Diet:* chicken potatoes bbq rice cheese fried beans potato baked corn mac mashed pork steak
- *Vegetarian:* vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum
- *Airport:* airport lounge waiting flight home my party yumyum purple pink vintage international modern sleepy
- *After Work:* time so up just after work day my now home today out last all go night not some back
- *First Person Casual:* my i lol up wings time some bout da good bomb like chicken
- *You-We:* you we your us today our see all come who great if time hope up thanks day good know

Classification Framework

Support Vector Machine (SVM) with linear kernel

- Large range in # of tweets per state: 339 (WY) - 83,670 (NY)
- Many features in prediction task (from all tweets in a state)
- But small number of data points (51 US states + DC)





Leave-One-Out Cross-Validation

- Each state is held out in turn
- Train SVM on features of tweets from the remaining 50 states
- Use SVM to predict the label of the held-out state
- Model accuracy: number of correct predictions divided by 51

Diabetes, Obesity, and Political Accuracies

- Percentage of states classified correctly

	overweight	diabetes	political	average
majority baseline	51.0	51.0	51.0	51.0
All Words	76.5	64.7	66.7	69.3
All Words + topics	**80.4**	64.7	68.6	**71.2**
Food	70.6	60.8	68.6	66.7
Food + topics	68.6	60.8	**72.6**	67.3
Hashtags	72.6	**68.6**	60.8	67.3
Hashtags + topics	74.5	**68.6**	62.8	68.6

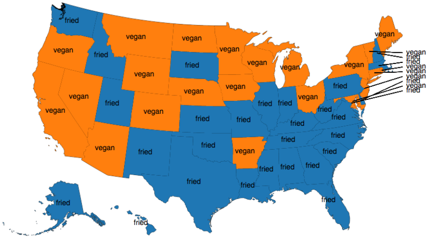
- All words best on average, but Food alone nearly as good
- Best performance on overweight
- Political and diabetes are well above baselines
- Topic modeling is often beneficial

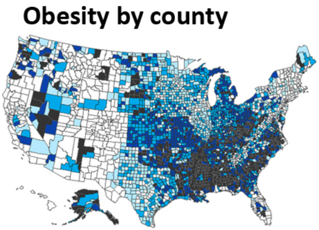
Feature Analysis: Overweight

Rank individual features by the weights assigned by the SVM

Class	Highest-weighted features
overweight: +	i, day, my, great, one, *American Diet* (chicken, baked, beans, fried), #snack, *First-Person Casual* (my, i, lol), cafe, *Delicious* (foodporn, yummy, yum), *After Work* (time, home, after, work), house, chicken, fried, *Breakfast* (day, start, off, right)
overweight: -	*You-We* (you, we, your, us), #rvadine, #vegan, make, photo, dinner, #meal, #pizza, *Giveaway* (win, competition, enter), new, *Restaurant Ads* (open, today, come, join), #date, happy, #dinner, 10

Fried (+ for overweight) vs Vegan (- for overweight)





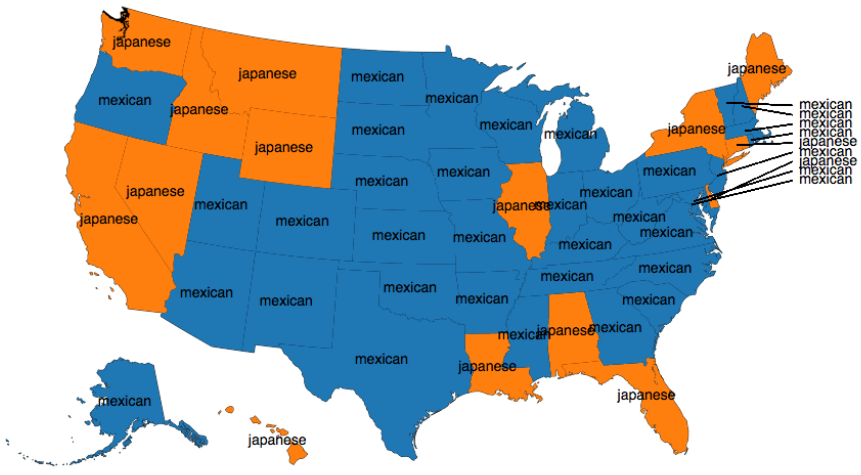
Relative usage of “fried” (7,254 tweets) and “vegan” (12,424 tweets) versus the 2010 CDC obesity map

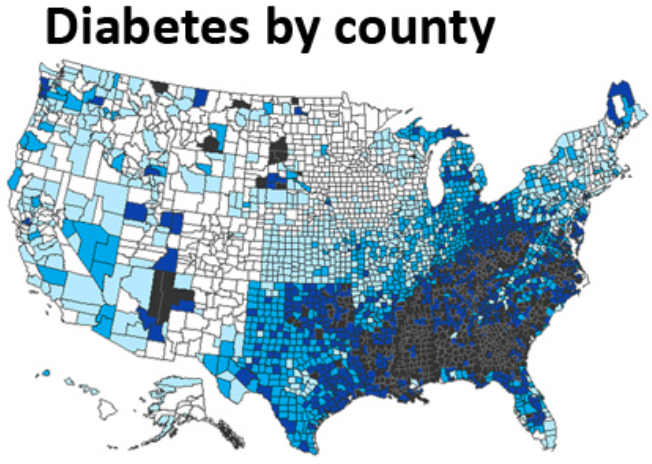
Feature Analysis: Diabetes

Rank individual features by the weights assigned by the SVM

Class	Highest-weighted features
diabetes: +	*American Diet* (chicken, baked, beans, fried), *Mexican* (mexican, tacos, burrito), #food, *After Work* (time, home, after, work), #pdx, my, lol, #fresh, *Delicious* (foodporn, yummy, yum), #fun, morning, special, good, cafe, #nola
diabetes: -	#dessert, *Japanese* (ramen, japanese, noodles), *Turkish* (turkish, kebab, istanbul), #foodporn, #paleo, #meal, *Paleo Diet* (paleo, chicken, healthy), i, *Giveaway* (win, competition, enter), *I, You* (i, my, you, your), your, new, today, #restaurant, some

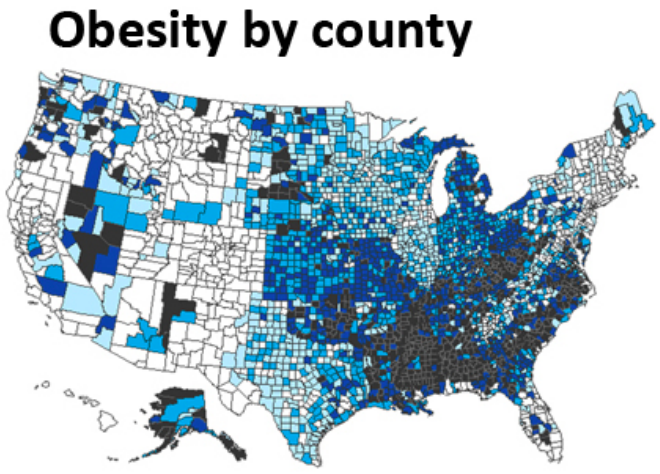
Mexican (+ for diabetes) vs Japanese (- for diabetes)

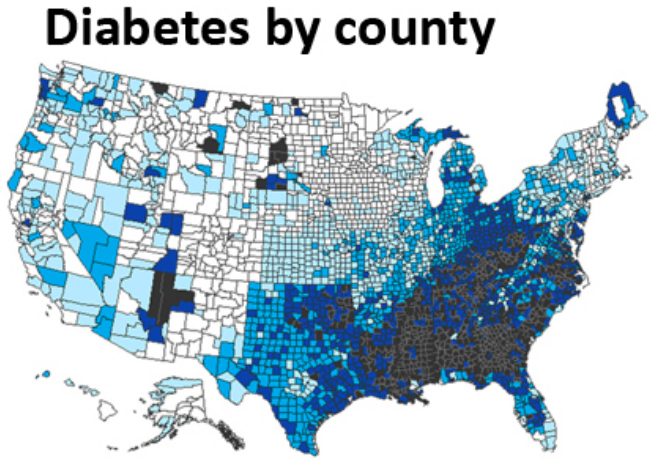




Relative usage of “Mexican” (4,438 tweets) and “Japanese” (2,464 tweets) versus the 2010 CDC diabetes map

Obesity vs Diabetes?





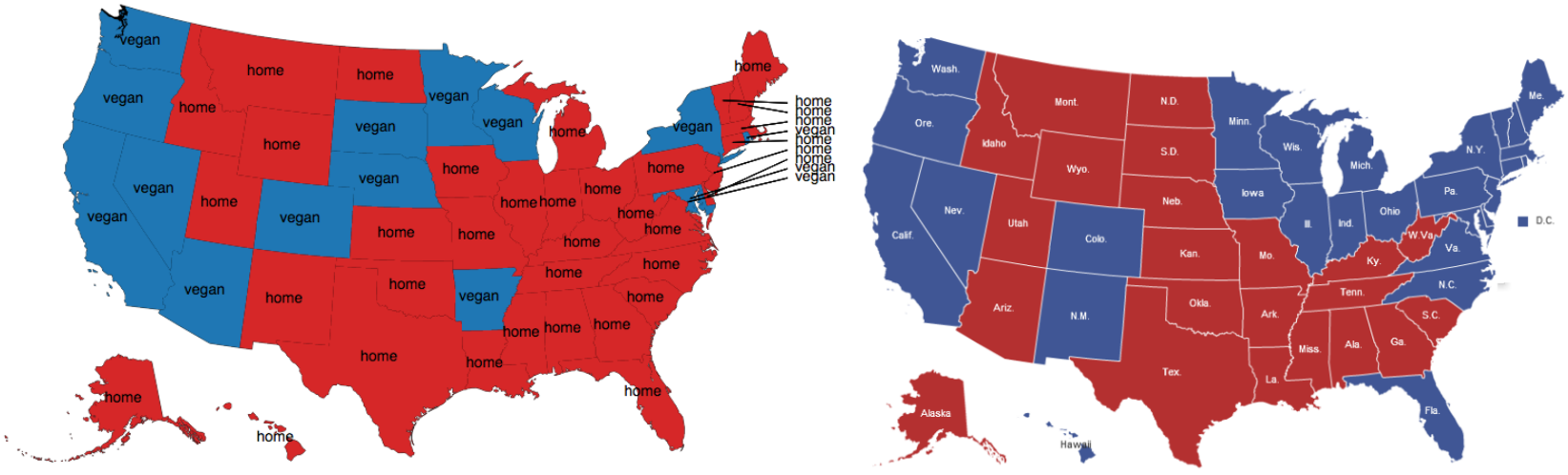
Although similar, the patterns are NOT the same!

Feature Analysis: Political

Rank individual features by the weights assigned by the SVM

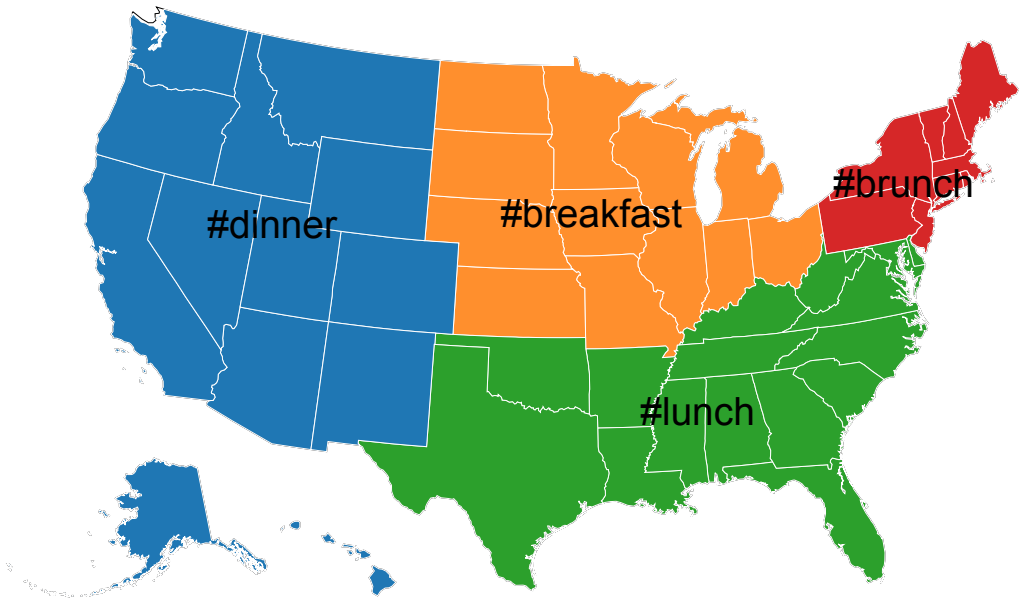
Class	Highest-weighted features
Democrat	#yum, #vegan, served, #brunch, *Deli* (cheese, sandwich, soup), photo, #rvadine, *Restaurant Ads* (open, today, come, join), #breakfast, #bacon, delicious, #food, #dinner, 21dayfix
Republican	my, #lunch, i, *Airport* (airport, lounge, waiting), easy, #meal, tonight, #healthy, #easy, us, sunday, *After Work* (time, home, after, work), #party, #twye, *First-Person Casual* (my, i, lol)

Vegan (+ for political) vs Home (- for political)



Relative usage of "home" (8,842) and "vegan" (12,424) tweets versus the 2012 Presidential election map

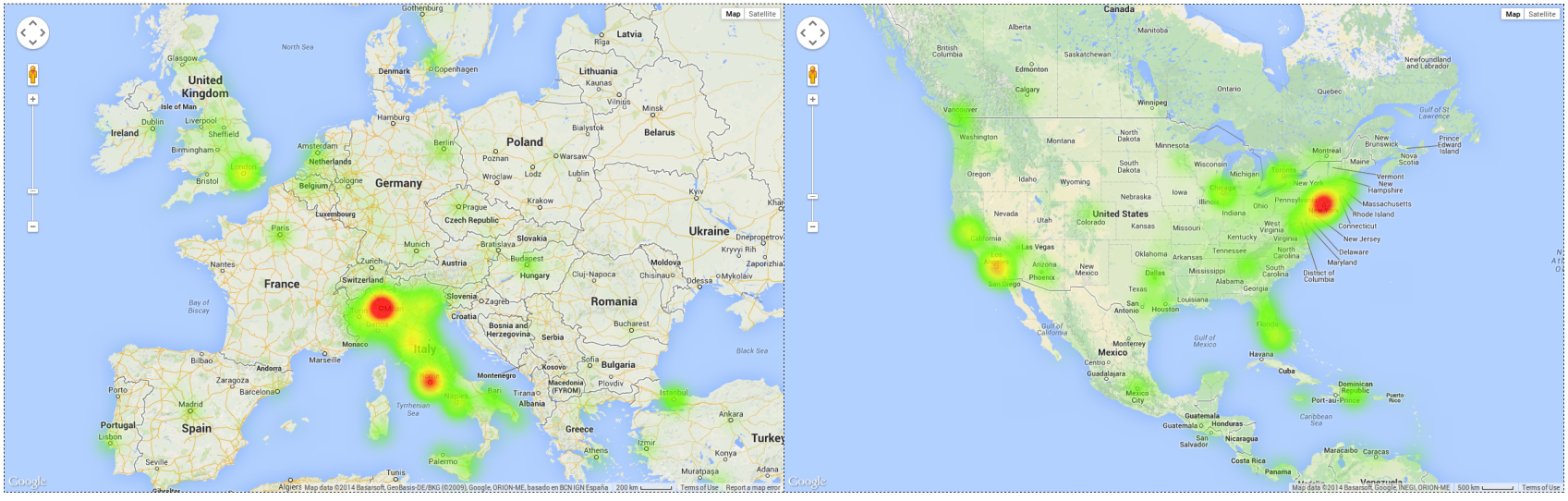
Region Prediction Features



Region	Highest-weighted features
Midwest	#breakfast, i, #recipes, *After Work*, *Recipe*,
Northeast	#brunch, brunch, our, *Mixed Drinks*, we,
South	#lunch, *Mixed Drinks*, *After Work*, *American Diet*, chicken,
West	#dinner, #food, #foodporn, photo, dinner,

Tweet Heatmaps

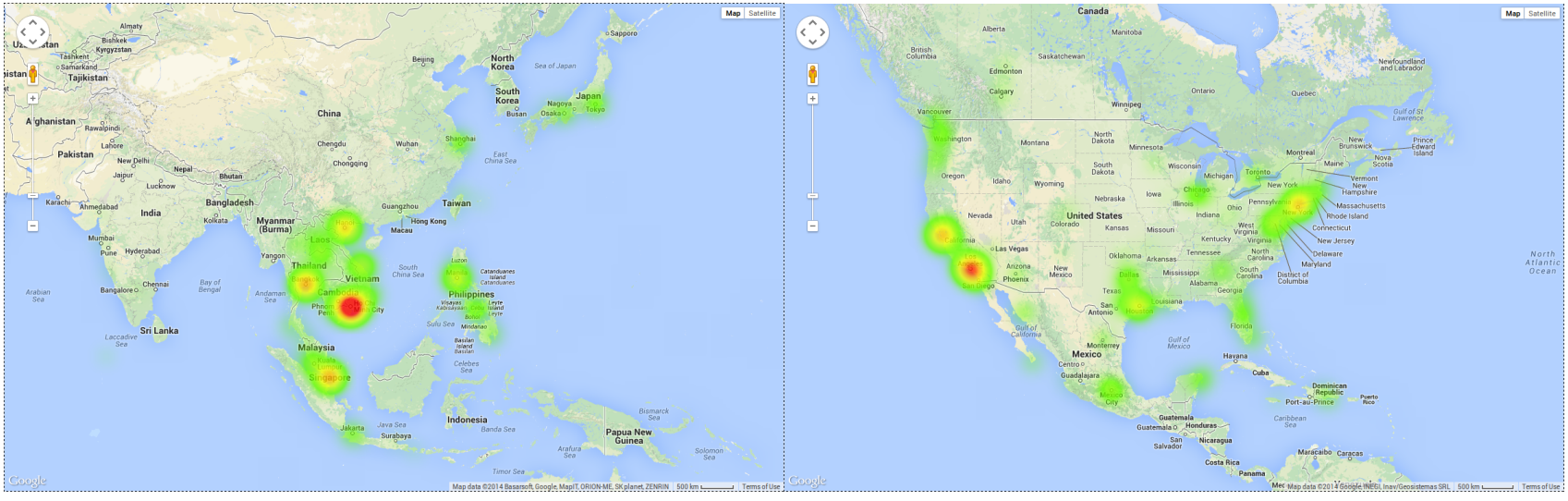
- Global trends possibly reflect migration patterns



Heatmaps of 7,372 tweets from three *Italian food* (pasta, pizza, italian, carbonara, lasagna, ...) topics.

Tweet Heatmaps

- Global trends possibly reflect migration patterns

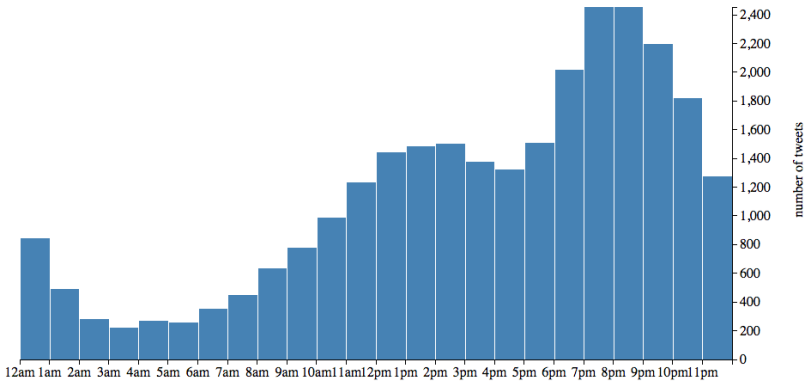


Heatmaps of 1,032 tweets from a *Vietnamese food* (pho, vietnamese, ...) topic.

Link to live version

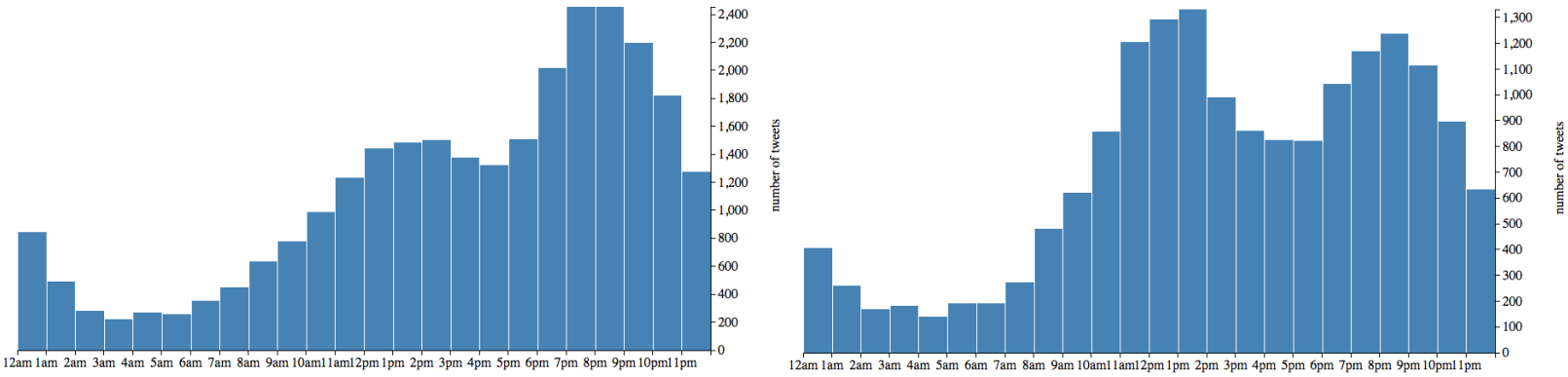
Temporal Histograms

- 71% of tweets (2.5 million) have a time zone
- Allow temporal analysis at varying granularities: hours
- Hourly tweets containing “wine” and “beer”



Temporal Histograms

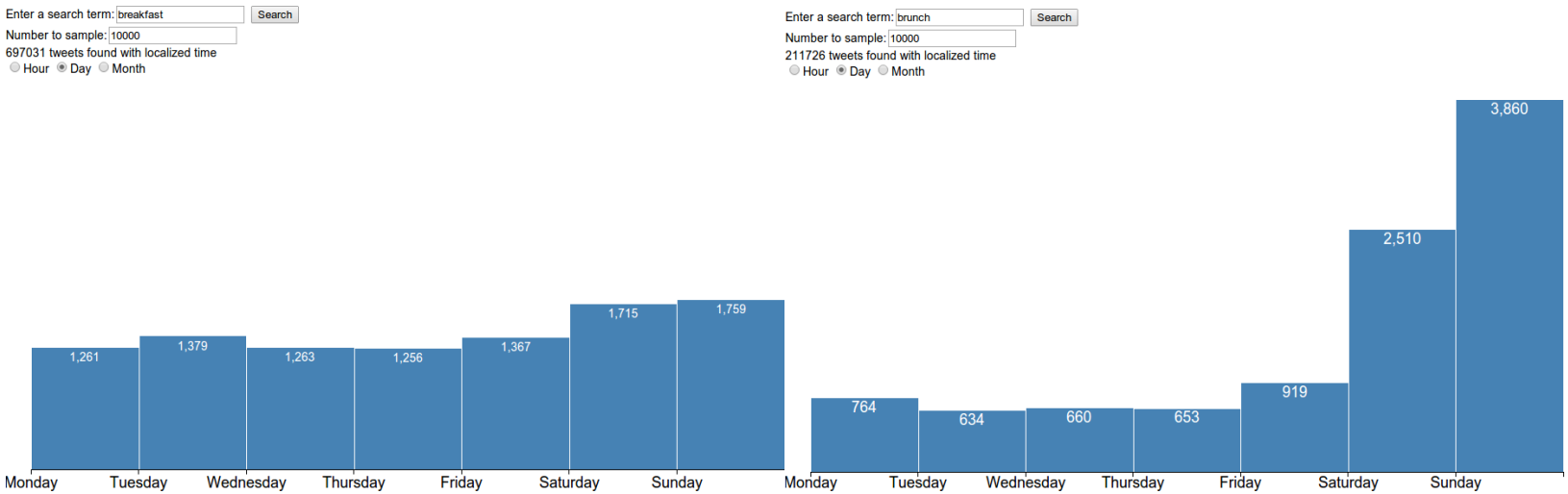
- 71% of tweets (2.5 million) have a time zone
- Allow temporal analysis at varying granularities: hours
- Hourly tweets containing “wine” and “beer”



Link to live version

Temporal Histograms

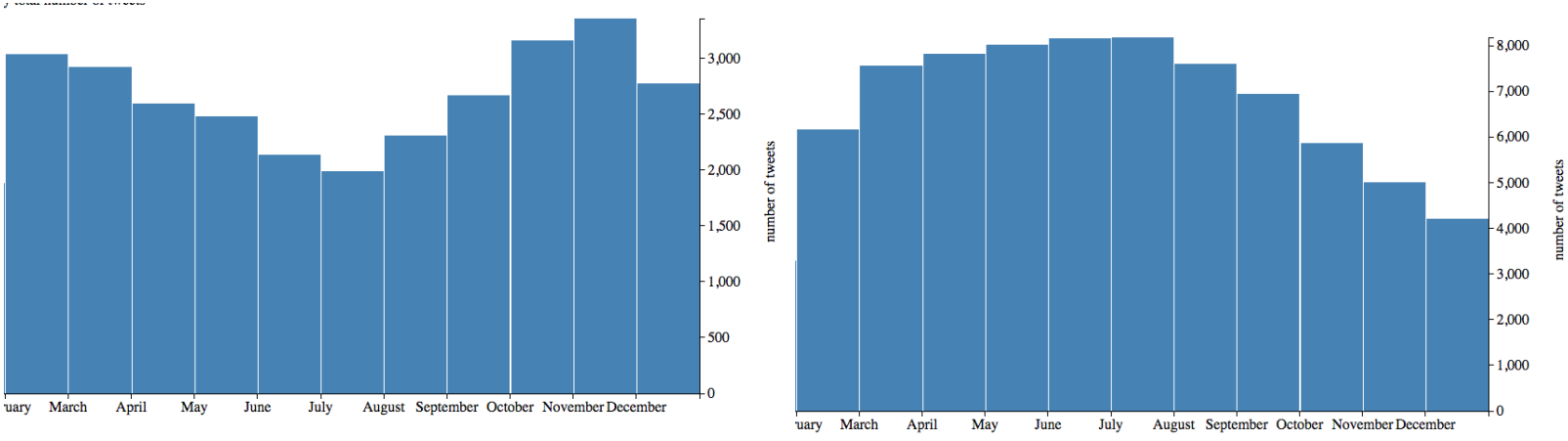
- Allow temporal analysis at varying granularities: days
- Daily tweets containing "breakfast" and "brunch"



Link to live version

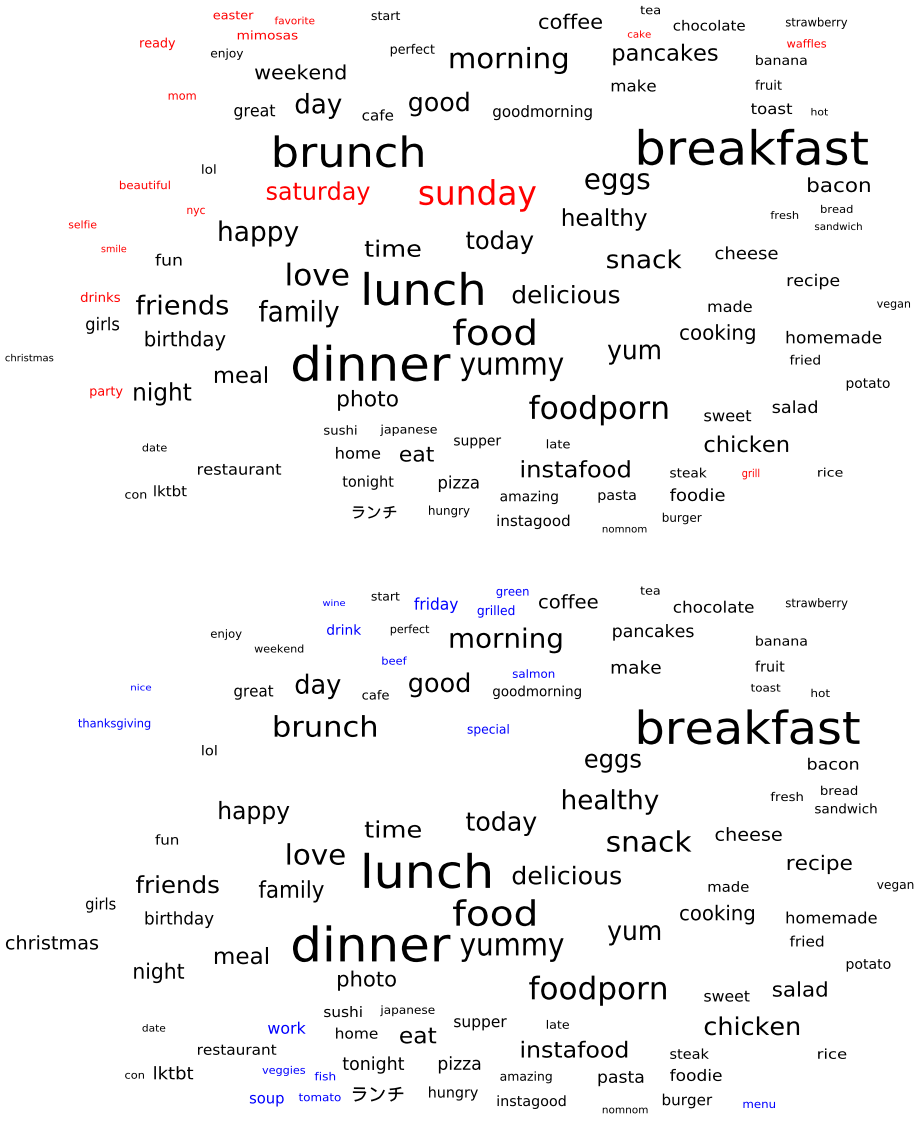
Temporal Histograms

- Allow temporal analysis at varying granularities: months
- Monthly tweets containing “soup” and “salad”



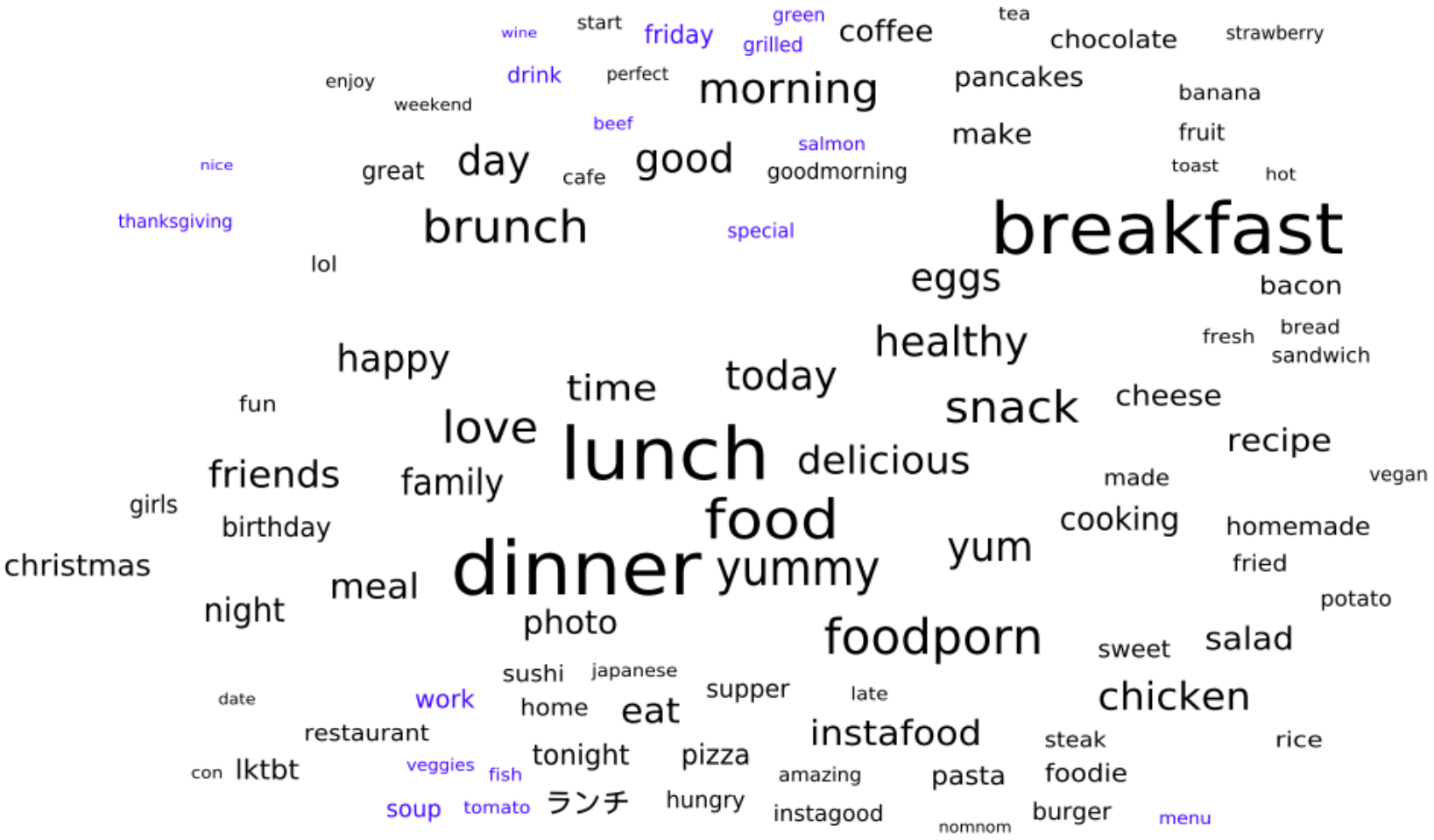
Link to live version

Parallel Semantic Word Clouds



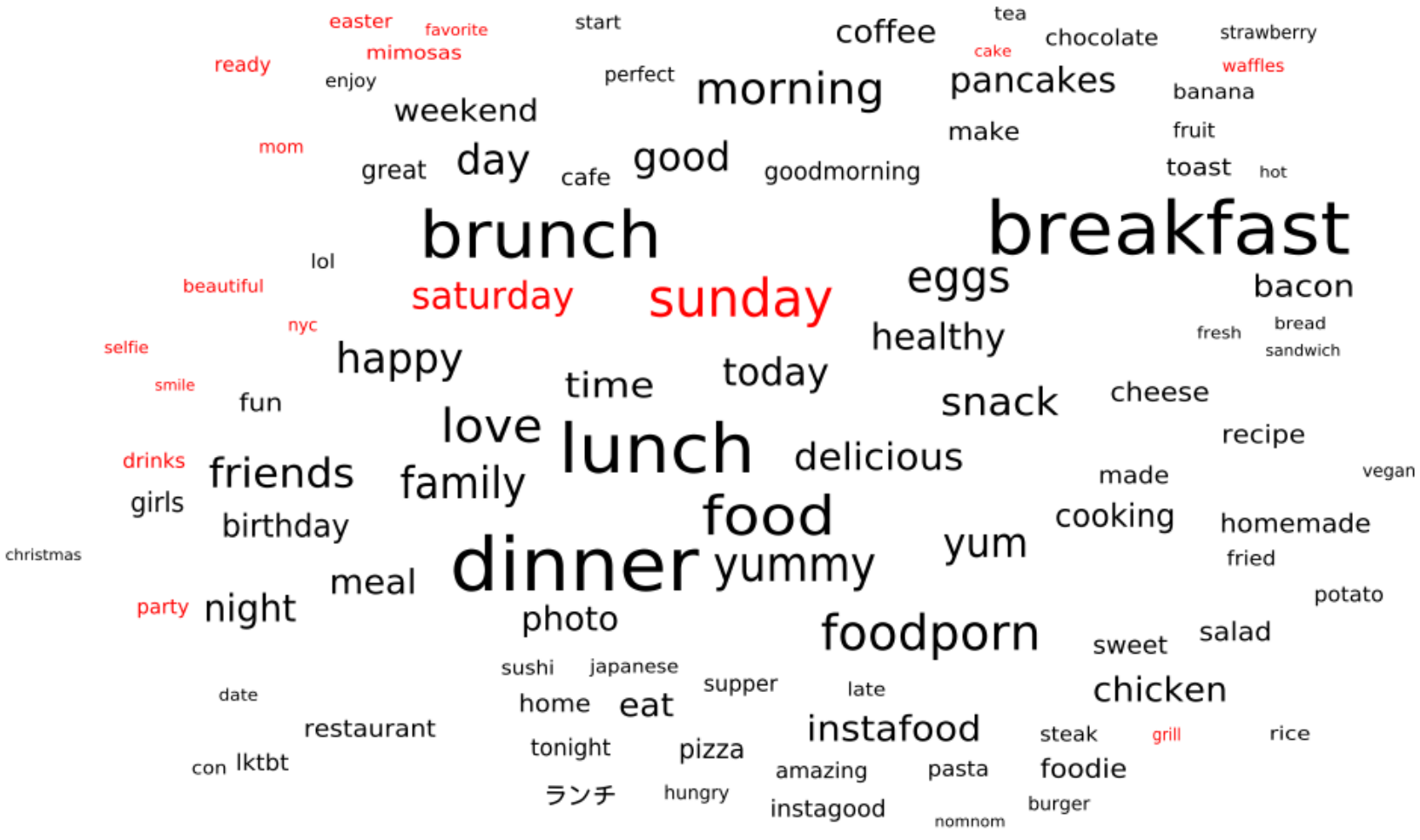
[Link to live version]

Parallel Semantic Word Clouds



Weekday Wordcloud

Parallel Semantic Word Clouds



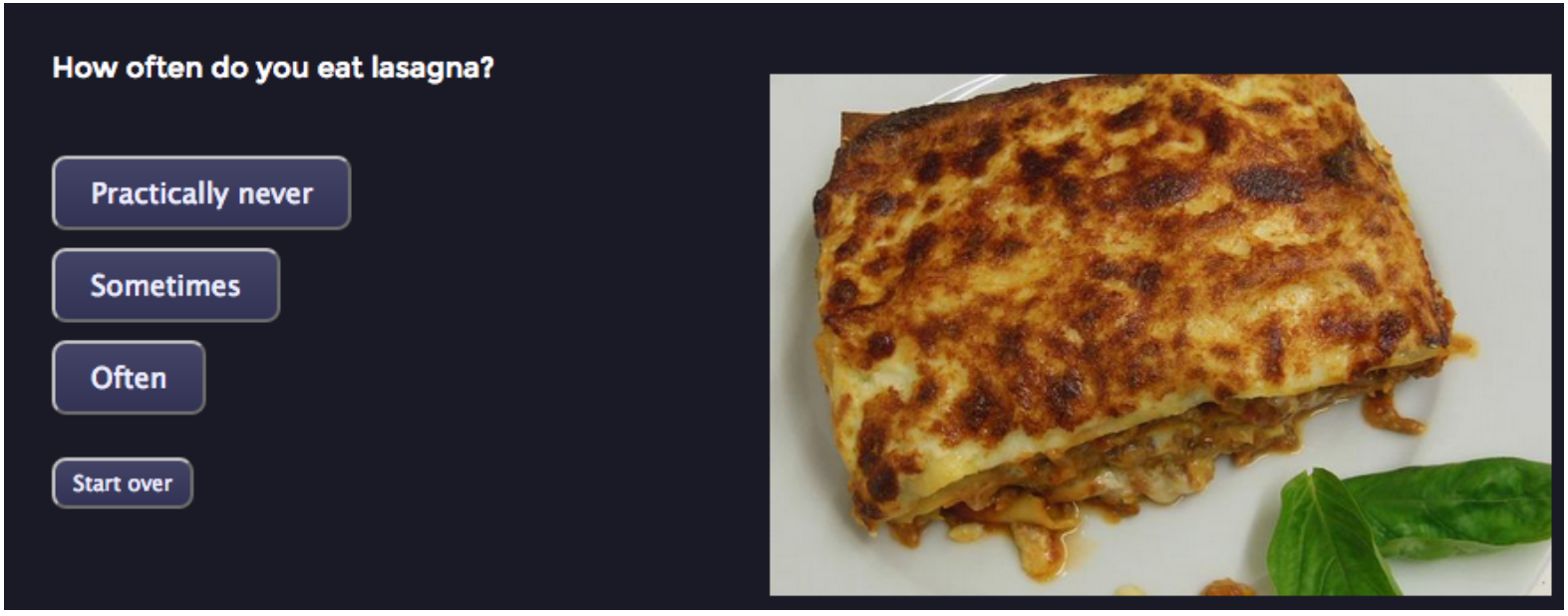
Weekend Wordcloud

20 Questions Quiz

- Use learned textual features to make individual predictions
- Our 20 questions quiz `http://52.4.173.235/owquiz.php`
- Goal: predict above or below US median BMI

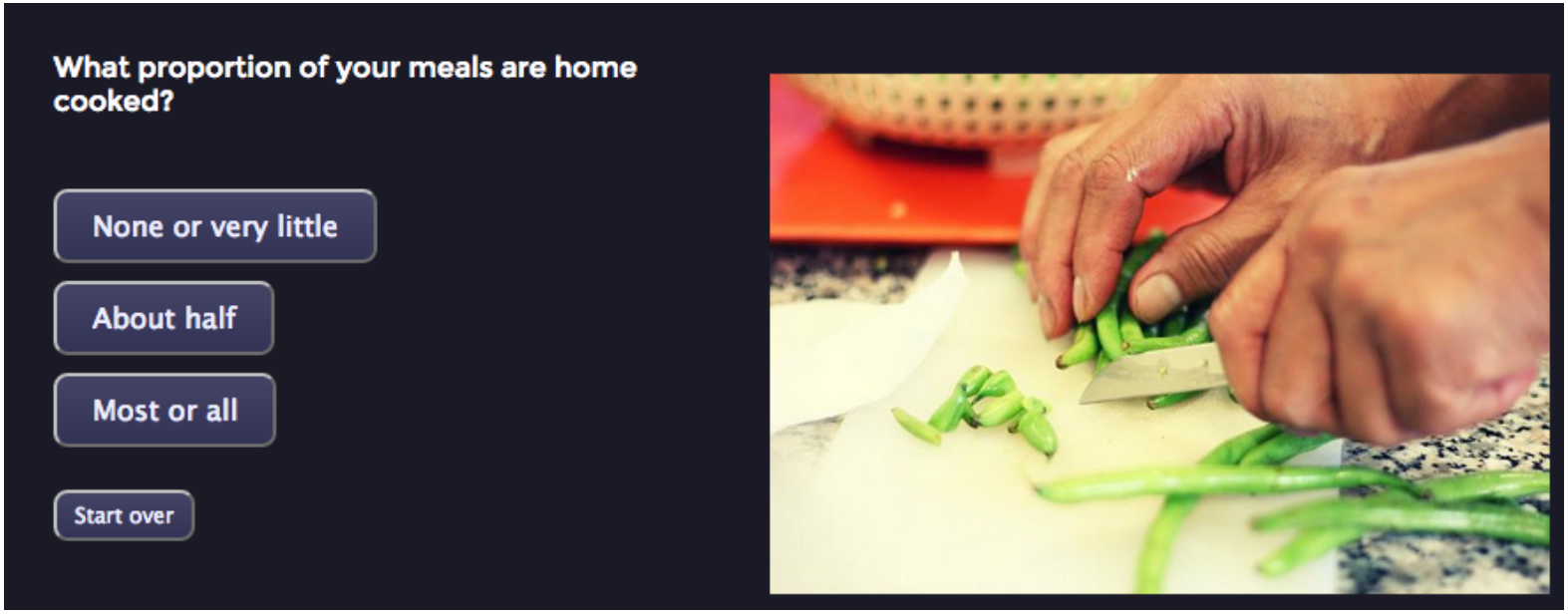
20 Questions Quiz

- Use learned textual features to make individual predictions
- Our 20 questions quiz `http://52.4.173.235/owquiz.php`
- Goal: predict above or below US median BMI



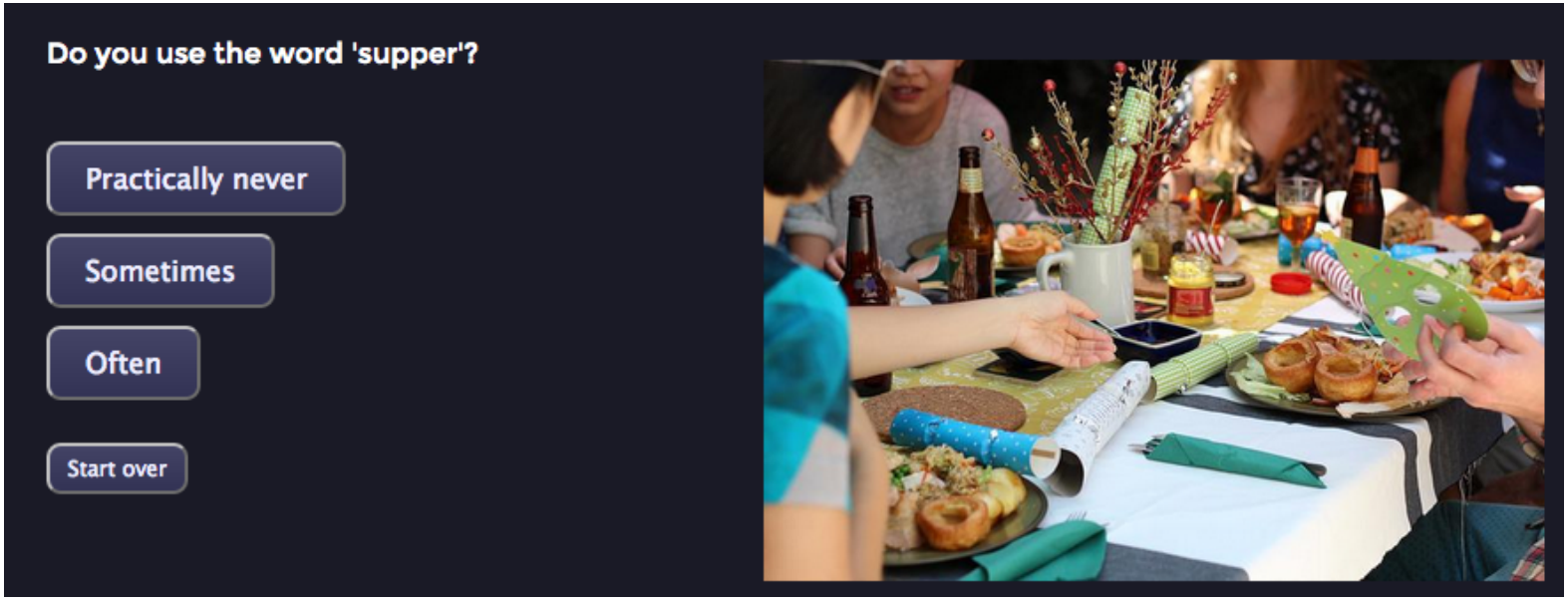
20 Questions Quiz

- Use learned textual features to make individual predictions
- Our 20 questions quiz `http://52.4.173.235/owquiz.php`
- Goal: predict above or below US median BMI



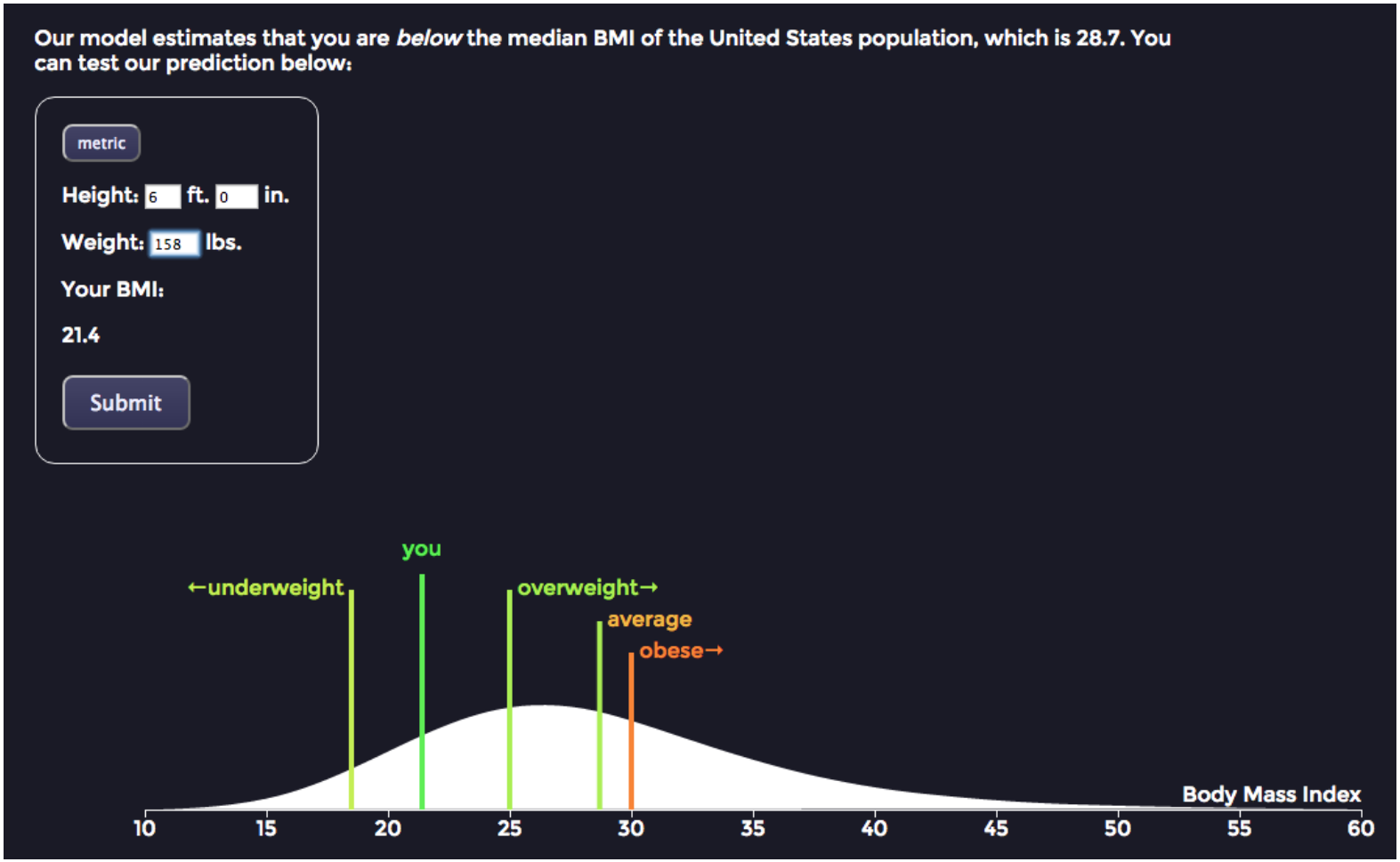
20 Questions Quiz

- Use learned textual features to make individual predictions
- Our 20 questions quiz `http://52.4.173.235/owquiz.php`
- Goal: predict above or below US median BMI



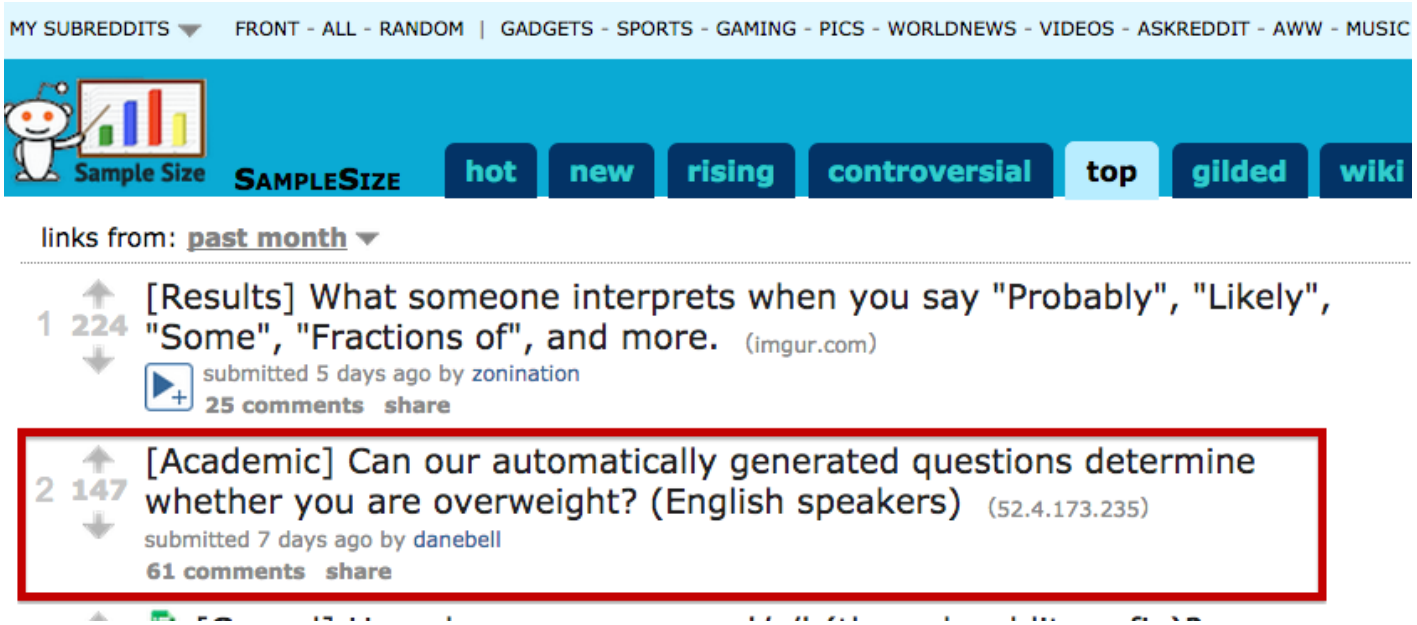
20 Questions Quiz

- Use learned textual features to make individual predictions
- Our 20 questions quiz `http://52.4.173.235/owquiz.php`
- Goal: predict above or below US median BMI



20 Questions Quiz

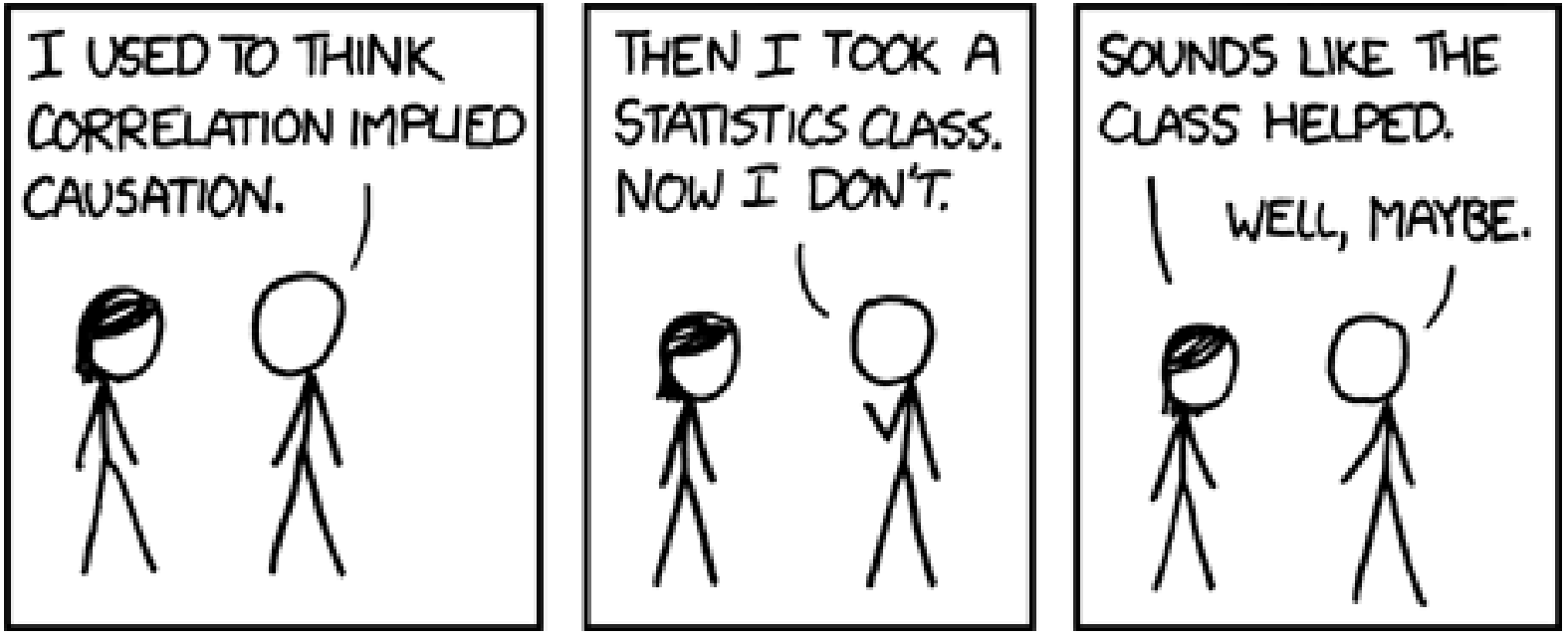
- Posted the quiz on reddit last week



- Good participation: 800+ participants in 48 hours
- Good engagement: 61 comments on reddit
- Good accuracy: 74%

Conclusions and Future Work

Important caveat: causation vs correlation [xkcd]



Conclusions and Future Work

- Language of food has predictive power
- Much of the predictive power comes from food words alone
- Paper at BigData'14 (also ArXiv)
- `https://sites.google.com/site/twitter4food`



Conclusions and Future Work

- Language of food has predictive power
- Much of the predictive power comes from food words alone
- Paper at BigData'14 (also ArXiv)
- `https://sites.google.com/site/twitter4food`



- Can we predict individual diabetes risk from Twitter
- Might visualization draw attention to risk factors?
- How to "nudge" pre-diabetics to see a doctor?

Acknowledgments

- Colleagues
  - Dane Bell, Arizona
  - Daniel Fried, Arizona $\rightarrow$ Berkeley
  - Melanie Hingle, Arizona
  - Mihai Surdeanu, Arizona
- Workshops
  - Dagstuhl
  - Bertinoro
  - Barbados
- Funding
  - NSF
  - USDA
  - ONR

Conclusions and Future Work

- Language of food has predictive power
- Much of the predictive power comes from food words alone
- Paper at BigData'14 (also ArXiv)
- `https://sites.google.com/site/twitter4food`



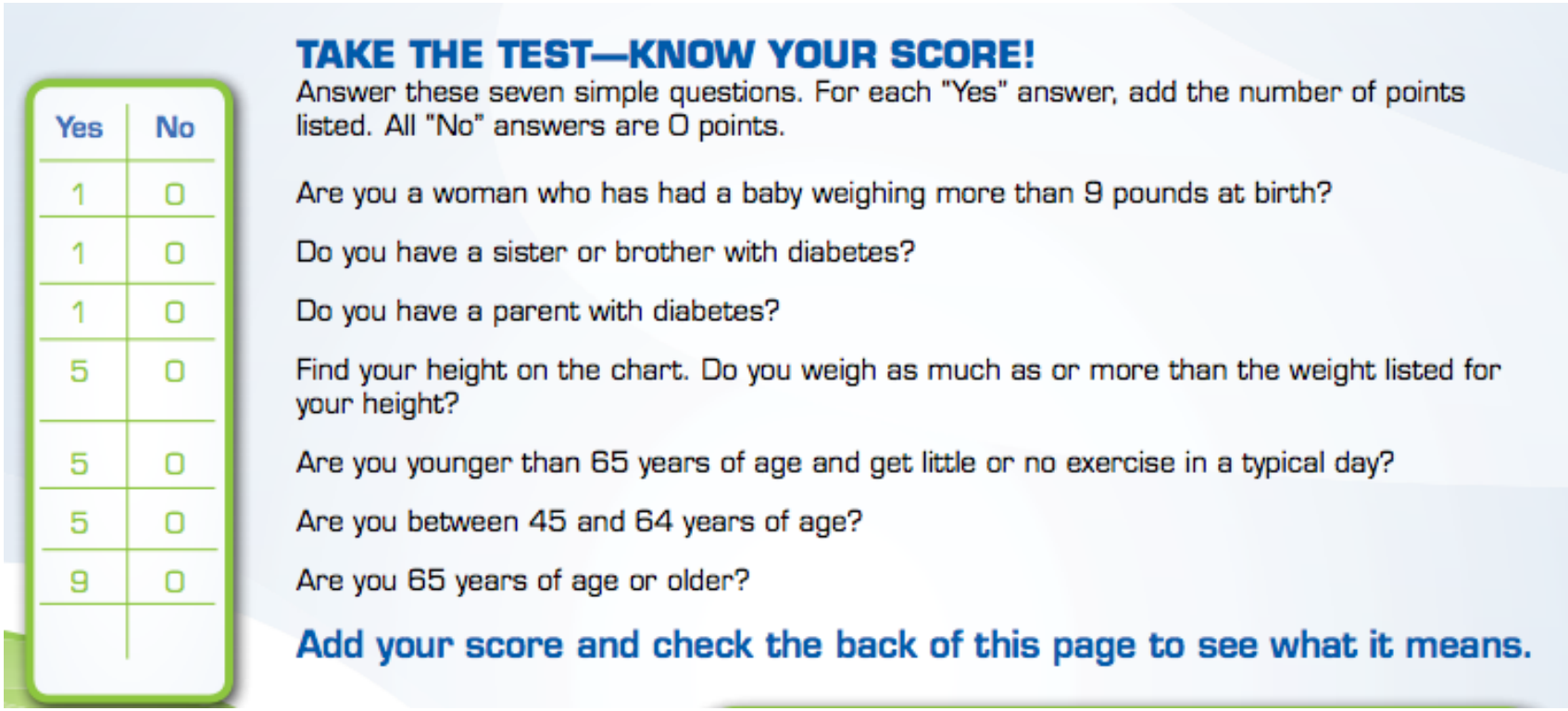
- Can we predict individual diabetes risk from Twitter
- Might visualization draw attention to risk factors?
- How to "nudge" pre-diabetics to see a doctor?

State Trends: not the most popular...

Link to live version The most misinterpreted figure from our paper: from the Washington Post, the Guardian and Slate to Fox News and the Daily Mail

CDC Pre-diabetes Test

The most common pre-diabetes test from the CDC



Yes	No
1	0
1	0
1	0
5	0
5	0
5	0
9	0

TAKE THE TEST—KNOW YOUR SCORE!

Answer these seven simple questions. For each "Yes" answer, add the number of points listed. All "No" answers are 0 points.

Are you a woman who has had a baby weighing more than 9 pounds at birth?

Do you have a sister or brother with diabetes?

Do you have a parent with diabetes?

Find your height on the chart. Do you weigh as much as or more than the weight listed for your height?

Are you younger than 65 years of age and get little or no exercise in a typical day?

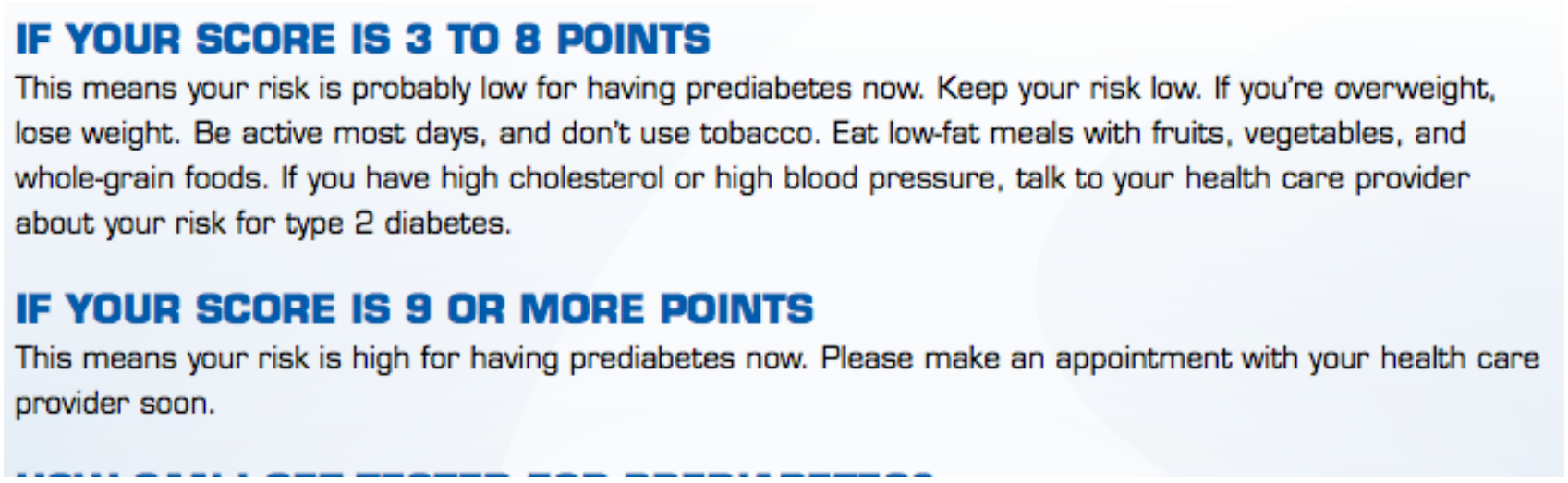
Are you between 45 and 64 years of age?

Are you 65 years of age or older?

Add your score and check the back of this page to see what it means.

CDC Pre-diabetes Test

The most common pre-diabetes test from the CDC



IF YOUR SCORE IS 3 TO 8 POINTS

This means your risk is probably low for having prediabetes now. Keep your risk low. If you're overweight, lose weight. Be active most days, and don't use tobacco. Eat low-fat meals with fruits, vegetables, and whole-grain foods. If you have high cholesterol or high blood pressure, talk to your health care provider about your risk for type 2 diabetes.

IF YOUR SCORE IS 9 OR MORE POINTS

This means your risk is high for having prediabetes now. Please make an appointment with your health care provider soon.