Analyzing the Language of Food on Social Media

Stephen Kobourov
University of Arizona

Why Study The Language of Food?

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

- geographic

  ralju ▸ Ria Misra
  Thursday 12:09pm

  Montanans are very serious about their pasties (pronounced pah-stee, in defiance of all logic). They're not unique to this state; they tend to crop up in places where mining was the primary economy. I believe they're Cornish originally.

- cultural

  Litarvan ▸ NoOnesPost
  Thursday 11:30am

  That sounds like something that my German-from-Russia mother used to make called fleishkuekle, only it was deep fried. I guess that you can get them in a restaurants in North Dakota.

- political

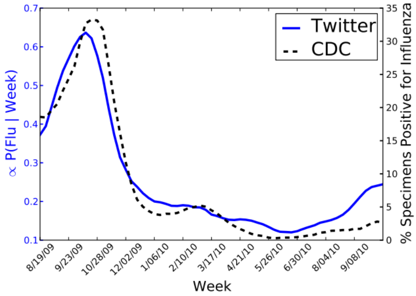
  Solongo
  @ssolongoo
  Follow

  Going vegan means that you will save more than 100 animals' lives each year.

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

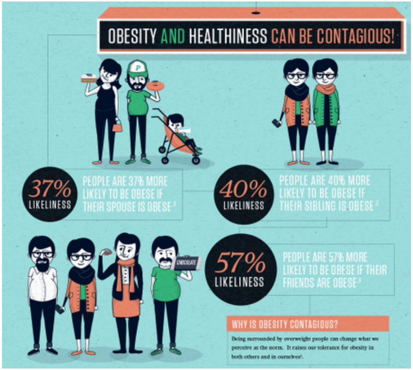
Why Twitter?

- Serious limitations: short, sparse, slang, self-reported, ...
- But used across ethnic, gender, age, socio-economic groups
- Tweets are freely available and easy to access
- Geographic linguistic analysis [Eisenstein et al. 2010]
- Flu, allergies prediction using Twitter [Paul and Dredze 2011]



Twitter and Food

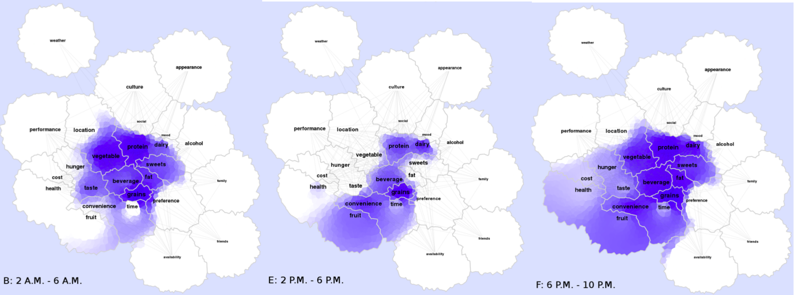
- Obesity is socially contagious [Christakis and Fowler 2007]
- 1 billion tweets used to measure and predict well-being [Schwartz et al. 2013]
- Social networking strategies can help people lose weight [Ashrafian et al. 2014]
- Twitter can be a greater source of positive influence for weight loss than 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 (1756 foods and 1106 reasons for eating)
- reasons for eating: #social (122), #taste (146), #convenience (173)
- track patterns over time



Motivation

- Analyze predictive features of language of food
- Identify textual features with most predictive power
- Find connections b/n the language of food and geography
- Visualize results in geographical and temporal dimensions
- Predict diabetes and obesity rates for communities

*Rudolof II, by Arcimboldo c. 1590

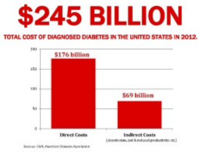
Motivation, cont.

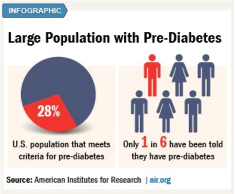
Why obesity and diabetes?

- 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 costs of diabetes exceed $245 billion annually
- 33% of untreated diabetics die of it
- 80% of Type 2 diabetes is preventable!

What can we do?

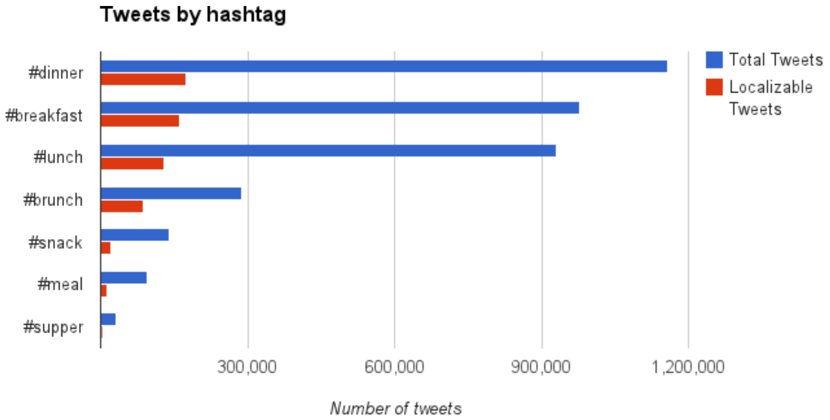
- Predict diabetes for individuals
- Identify people at risk for diabetes
- Attempt to intervene to prevent diabetes





Tweet Corpus

- Collect meal-related tweets: `breakfast`, `lunch`, `dinner`, ...
- 3.5 million tweets over October 2013 - May 2014
- Tweets limited by 140 characters
- Average tweet length: 8.7 words
- 30 million words, 1.5 million unique



State Location Normalization

- Location can be specified in the Twitter account
- Regular expression matching on state name
- Heuristics (e.g., "LA" + time zone $\rightarrow$ California or Louisiana)
- 560,000 tweets (16%) can be 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 is chicken in nearly every state...

TF-IDF Ranking

Term Frequency Inverse Document Frequency

- the tf-idf weight is used in inf. retrieval and text mining
- 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

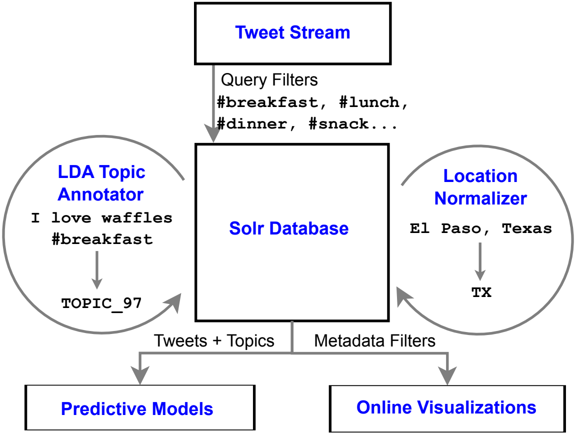
What is the 3rd most commonly spoken language in the US?

*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
- Find connections b/n the language of food and geography
- Visualize results in geographical and temporal dimensions

Collecting, Analyzing, and Visualizing Tweets

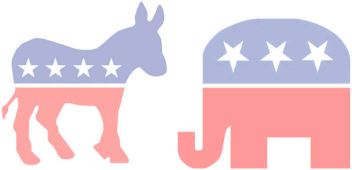


Collect tweets from Twitter API; store and query using Apache Solr

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

- median diabetes rate is 9.7%
- min 7.0% (Alaska) and max 13.0% (West Virginia)

Overweight data (BMI)

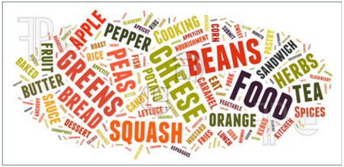
- median overweight rate is 64.2%
- min 51.9% (Washington, D.C) and max 69.6% (Louisiana)

Democratic Political Tendency (2008-2013)

- median Democratic rate is 51.6%
- min 27.0% (Wyoming) and max 92.4% (Washington D.C.)

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)
- Sparsity issue: tweets have short length and unique vocabulary
- Latent Dirichlet Allocation (LDA) to identify topics
- Various combinations



Topical Features

- LDA treats each tweet as a mixture of latent topics
- Each topic is itself a probability distribution over words
- Words in tweet: samples from mixture of distributions
- Train LDA model from all available tweets in the corpus
- # topics selected to capture patterns in diet & specific diets
- Use the highest probability topic for each tweet as a feature

LDA Topics

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

LDA Topics

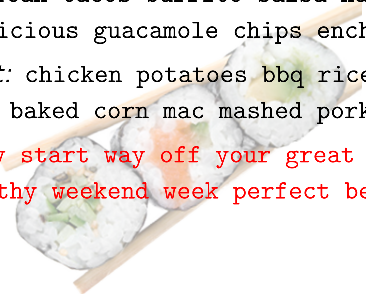
- *Japanese:* ramen japanese food noodles noodle yummy japan byob takeout spicy open katsu pork japanesefood
- *Mexican:* mexican tacos burrito salsa nachos chicken homemade delicious guacamole chips enchiladas

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`

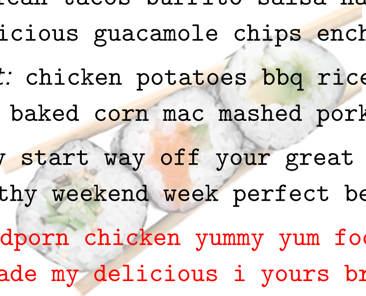
LDA Topics

- *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`
- *Breakfast:* `day start way off your great right morning my good healthy weekend week perfect best begun hunt`



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`
- *Breakfast:* `day start way off your great right morning my good healthy weekend week perfect best begun hunt`
- *Delicious:* `foodporn chicken yummy yum food homemade steak rice made my delicious i yours broccoli truly`



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`
- *Breakfast:* `day start way off your great right morning my good healthy weekend week perfect best begun hunt`
- *Delicious:* `foodporn chicken yummy yum food homemade steak rice made my delicious i yours broccoli truly`
- *Giveaway:* `win competition enter our 2 giveaway follow chance meal free you us your winner two today http`

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`
- *Breakfast:* `day start way off your great right morning my good healthy weekend week perfect best begun hunt`
- *Delicious:* `foodporn chicken yummy yum food homemade steak rice made my delicious i yours broccoli truly`
- *Giveaway:* `win competition enter our 2 giveaway follow chance meal free you us your winner two today http`
- *Restaurant Ad:* `we us open today come our join you day until all serving sunday now pm tomorrow weekend till`

LDA Topics

- *Paleo:* `paleo chicken peppers healthy eggs bacon mushrooms potatoes stuffed kale asparagus salmon`

LDA Topics

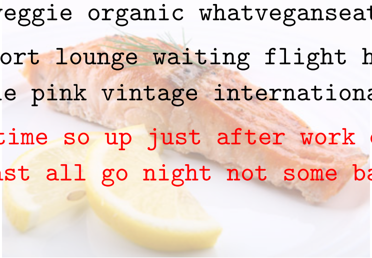
- *Paleo:* `paleo chicken peppers healthy eggs bacon mushrooms potatoes stuffed kale asparagus salmon`
- *Vegetarian:* `vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum`

LDA Topics

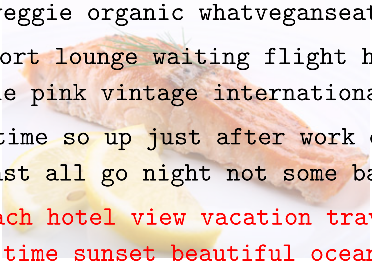
- *Paleo:* paleo chicken peppers healthy eggs bacon mushrooms potatoes stuffed kale asparagus salmon
- *Vegetarian:* vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum
- *Airport:* airport lounge waiting flight home my party yumyum purple pink vintage international modern sleepy

LDA Topics

- *Paleo:* paleo chicken peppers healthy eggs bacon mushrooms potatoes stuffed kale asparagus salmon
- *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

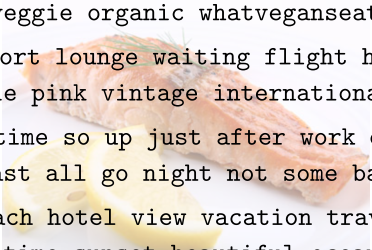


- *Paleo:* `paleo chicken peppers healthy eggs bacon mushrooms potatoes stuffed kale asparagus salmon`
- *Vegetarian:* `vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum`
- *Airport:* `airport lounge waiting flight home my party yumyum purple pink vintage international modern sleepy`
- *After Work:* `time so up just after work day my now home today out last all go night not some back`
- *Vacation:* `beach hotel view vacation travel resort sea holiday sun time sunset beautiful ocean love island`

LDA Topics

- *Paleo:* paleo chicken peppers healthy eggs bacon mushrooms potatoes stuffed kale asparagus salmon
- *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
- *Vacation:* beach hotel view vacation travel resort sea holiday sun time sunset beautiful ocean love island
- *First Person Casual:* my i lol up wings time some bout da good bomb like chicken

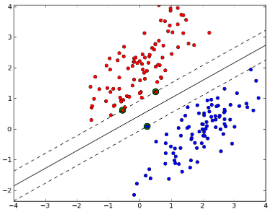
LDA Topics

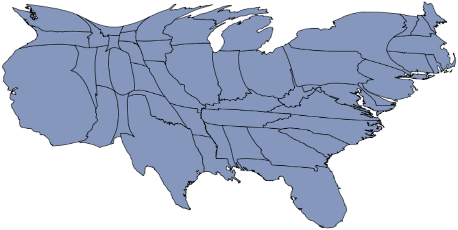


- *Paleo:* paleo chicken peppers healthy eggs bacon mushrooms potatoes stuffed kale asparagus salmon
- *Vegetarian:* vegan vegetarian healthy game fun raw tofu glutenfree veggie organic whatveganseat salad yum
- *Airport:* airport lounge waiting flight home my party yumyum purple pink vintage international modern sleepy
- *After Work:* time so up just after work day my now home today out last all go night not some back
- *Vacation:* beach hotel view vacation travel resort sea holiday sun time sunset beautiful ocean love island
- *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)
  - Normalization: scale all the features collected for each state
- Diabetes, obesity, and political prediction
  - Many features in prediction task (from all tweets in a state)
  - But small number of data points (51 US states + DC)
- Use 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 dataset's 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

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

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

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

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

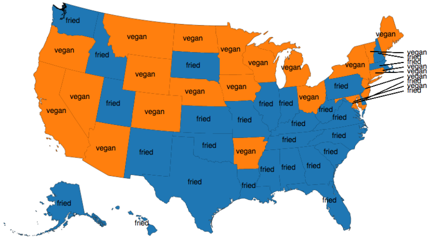
- All words best on average, but Food alone nearly as good
- Best performance on overweight
- Political and diabetes are well above baselines
- Topic modeling is often beneficial

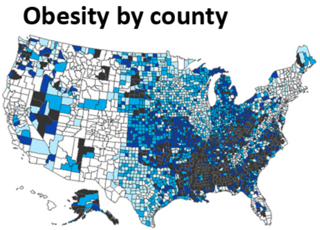
Feature Analysis: Overweight

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)





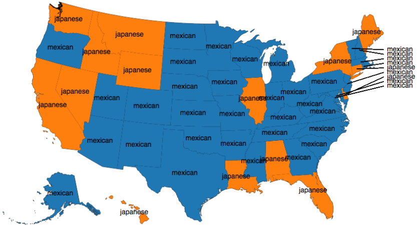
Relative usage of “fried” (7,254 tweets) and “vegan” (12,424 tweets) versus the 2010 CDC obesity map

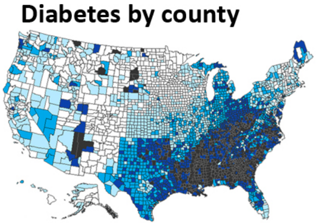
Feature Analysis: Diabetes

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

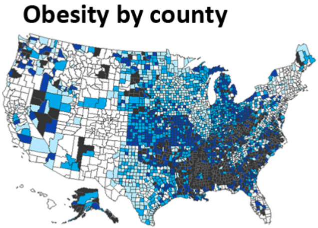
Mexican (+ for diabetes) vs Japanese (- for diabetes)

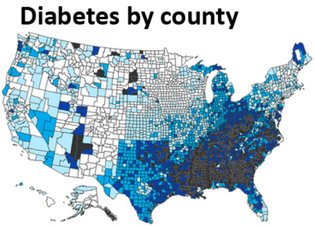




Relative usage of “Mexican” (4,438 tweets) and “Japanese” (2,464 tweets) versus the 2010 CDC diabetes map

Obesity vs Diabetes?





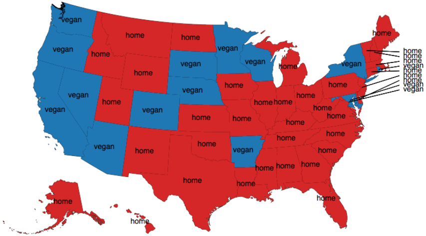
Although similar, the patterns are NOT the same!

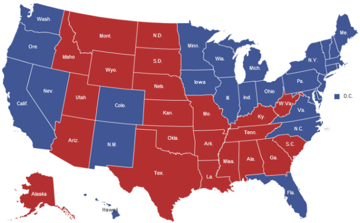
Feature Analysis: Political

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)

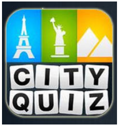




Relative usage of “home” (8,842) and “vegan” (12,424) tweets versus the 2012 Presidential election map

Classification Framework

- Location prediction
  - Predict the location for a given set of testing tweets
  - Support Vector Machine (SVM) with linear kernel
  - Split: training (80%) and testing (20%) data for each location
  - Train 1-vs-many multi-class SVM classifier for each location
  - Model accuracy: percentage of correct guesses
- Reduce cheating
  - remove state name, city name, acronyms from tweets
  - remove URLs, usernames (preceded by an @ symbol)



City, State, and Region Prediction

- Percentage of accurate location prediction

model	city acc.	state acc.	region acc.
Random Baseline	$1/15 = 6.7$	$1/51 = 2.0$	$1/4 = 25$
All Words	66.7	60.8	50
All Words + topics	**80.0**	**66.7**	**75**
Food	40.0	33.3	50
Food + topics	40.0	35.3	50
Hashtags	53.3	62.8	50
Hashtags + topics	66.7	56.9	**75**

City, State, and Region Prediction

- Percentage of accurate location prediction

model	city acc.	state acc.	region acc.
Random Baseline	$1/15 = 6.7$	$1/51 = 2.0$	$1/4 = 25$
All Words	66.7	60.8	50
All Words + topics	**80.0**	**66.7**	**75**
Food	40.0	33.3	50
Food + topics	40.0	35.3	50
Hashtags	53.3	62.8	50
Hashtags + topics	66.7	56.9	**75**

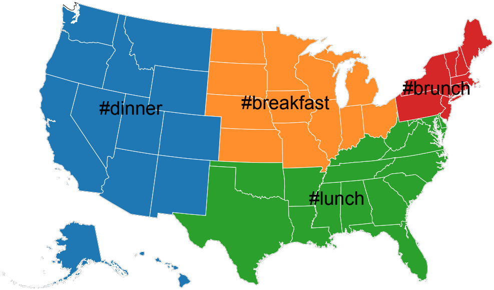
- Context is very important here: All words wins consistently
- Food alone still improves substantially on baseline

City Prediction Features

Top five highest-weighted features for each city

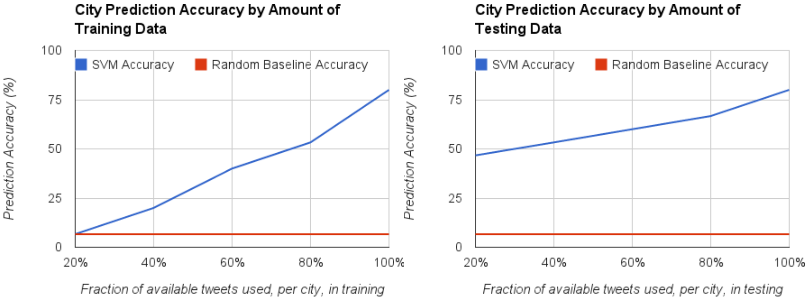
City	Highest-weighted features
Austin	we, come, tacos, #tacos, *Mixed Drinks*
Chicago	*Giveaway*, jerk, #breakfast, #bbq, #foodie
Columbus	#breakfast, #asseenincolumbus, *Directions*, #cbus, #great
Dallas	#lunch, my, lunch, porch, come
Houston	*After Work*, #lunch, #snack, i, #breakfast
Indianapolis	you, our, delicious, *You & We*, side
Jacksonville	#dinner, #ebaymobile, #food, kitchen, #yum
Los Angeles	my, #foodie, *Directions* , #timmynolans, #tolucalake
New York City	#brunch, *Mixed Drinks*, our, *Eggs and Bacon*, #sarabeths
Philadelphia	cafe, day, #fishtown, shot, #byob
Phoenix	#lunch, #easy, *Wine*, st, we
San Antonio	my, i, 1, bottomless, our
San Diego	*Restaurant*, #bottomless, *Mixed Drinks*, *Vacation*, your
San Francisco	#vegetarian, #dinner, #foodie, brunch, *Vacation*
San Jose	#foodporn, #dinner, bill, #bacon, #goodeats

Region Prediction Features

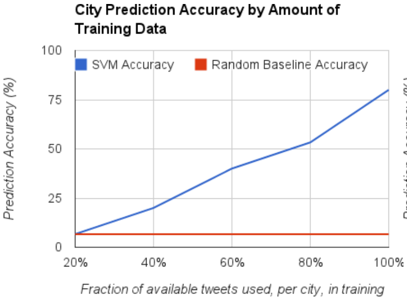


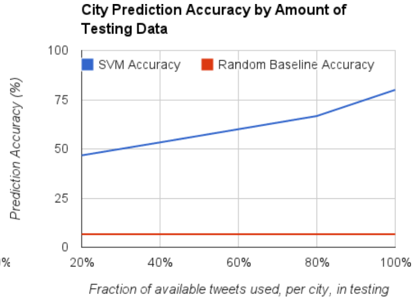
Region	Highest-weighted features
Midwest	#breakfast, i, #recipes, *After Work*, *Recipe*,
Northeast	#brunch, brunch, our, *Mixed Drinks*, we,
South	#lunch, *Mixed Drinks*, *After Work*, *American Diet*, chicken,
West	#dinner, #food, #foodporn, photo, dinner,

Learning Curves



Learning Curves

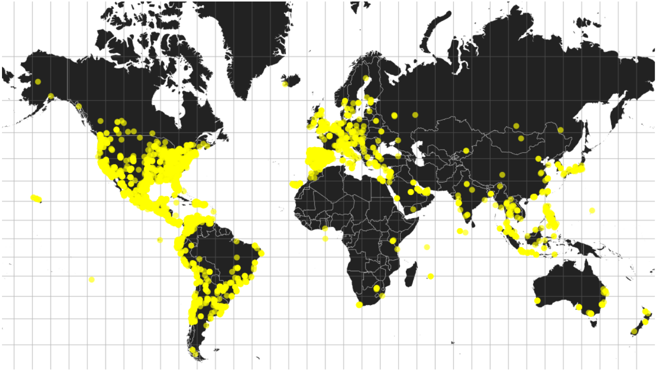




- Accuracy increases with size of training and testing data
- Can do well with small testing set as long as training is large
- Same effect for state and region prediction

Tweet Location Visualization

- 10% of tweets (360,000) have a GPS location

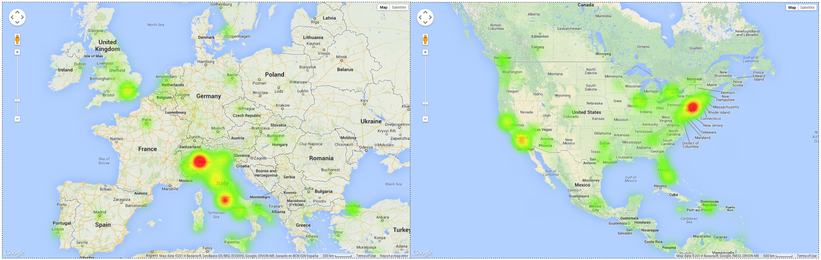


11,827 tweets from *Spanish/Latin American food* topics: tacos, burritos, salsa, pollo, arroz, paella, ...

[Link to live version]

Tweet Heatmaps

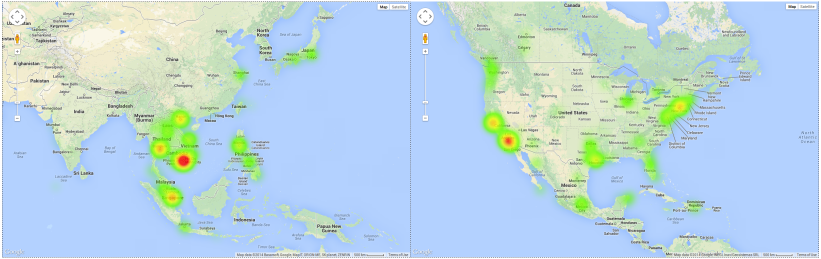
- Global trends possibly reflect migration patterns



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

Tweet Heatmaps

- Global trends possibly reflect migration patterns

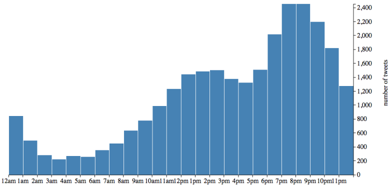


Heatmaps of 1,032 tweets from a *Vietnamese food* (pho, vietnamese, ...) topic.

[Link to live version]()

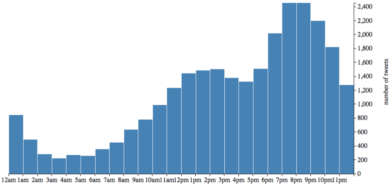
Temporal Histograms

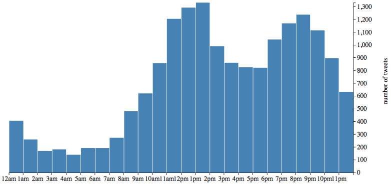
- 71% of tweets (2.5 million) have a time zone
- Allow temporal analysis at varying granularities: hours
- Hourly tweets containing "wine" and "beer"



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”

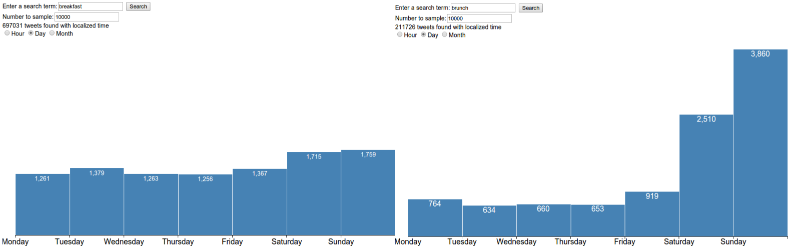




Link to live version

Temporal Histograms

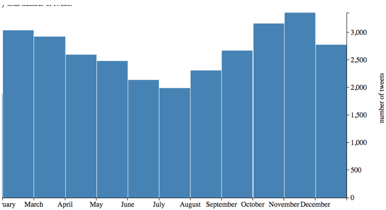
- Allow temporal analysis at varying granularities: days
- Daily tweets containing "breakfast" and "brunch"

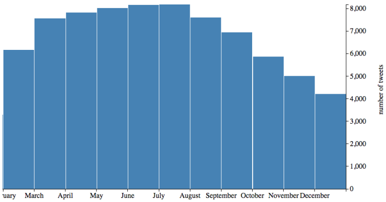


[Link to live version]

Temporal Histograms

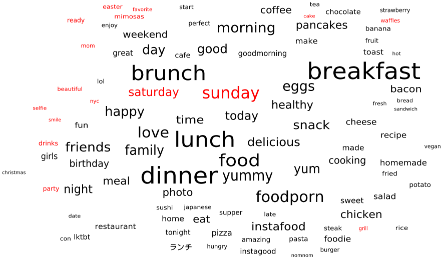
- Allow temporal analysis at varying granularities: months
- Monthly tweets containing “soup” and “salad”

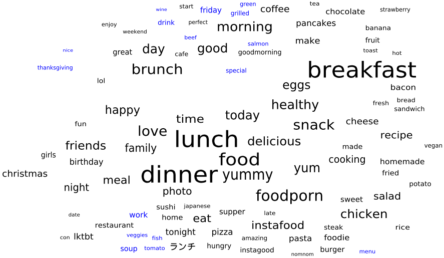




[Link to live version]

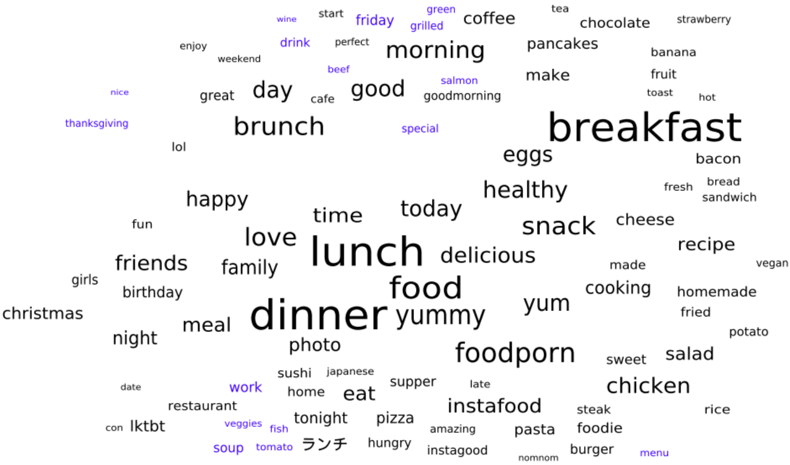
Parallel Semantic Word Clouds





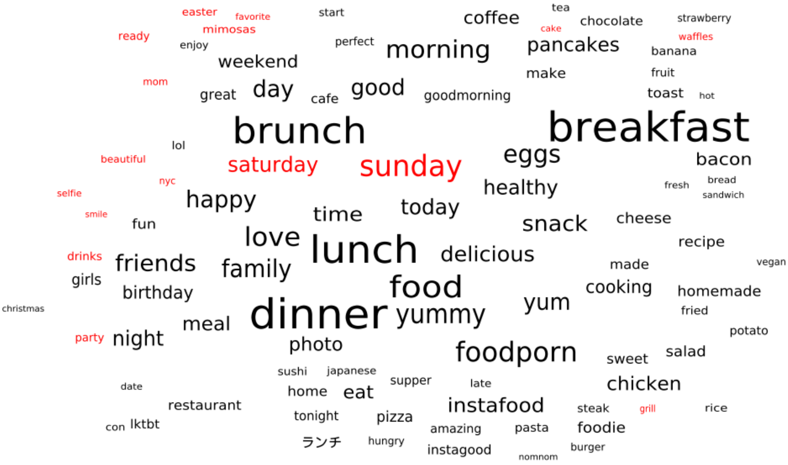
Link to live version

Parallel Semantic Word Clouds



Weekday Wordcloud

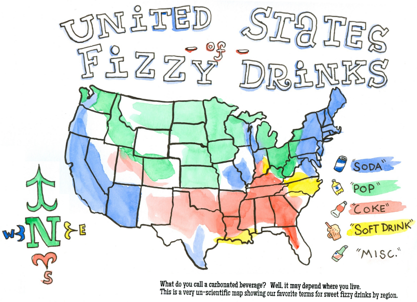
Parallel Semantic Word Clouds

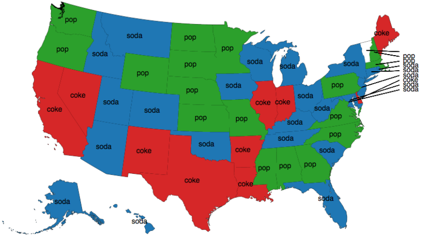


Weekend Wordcloud

State-Level Term Comparison

- Visualize the most popular term from a given set of words
- Soda, pop, or coke?





[Link to live version]

Conclusions

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

Conclusions

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

- Future work
  - Can we predict individual diabetes risk from Twitter
  - Will adding more social media inputs (images, video) help?
  - How can we get pre-diabetics to see a doctor?
  - Might visualization draw attention to risk factors?
  - 20 Questions: guess where you live based on what you eat?

Acknowledgments

- Colleagues
  - Daniel Fried, Oxford
  - Melanie Hingle, Arizona
  - Mihai Surdeanu, Arizona
- Workshops
  - Dagstuhl
  - Bertinoro
  - Barbados
- Funding
  - NSF
  - USDA
  - ONR