Analyzing the Language of Food on Social Media

Daniel Fried, Mihai Surdeanu, Stephen Kobourov, Melanie Hingle, & Dane Bell

University of Arizona

September 16, 2014

Why Study The Language of Food?

- Diet is associated with individual and community identity
- 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.

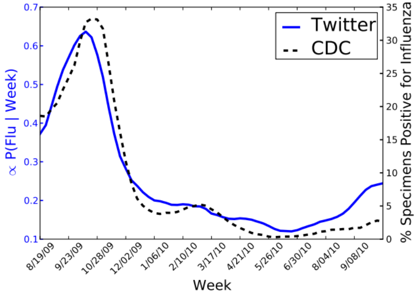
- political:

  Solongo
  @ssolongoo
  Follow

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

Why Twitter?

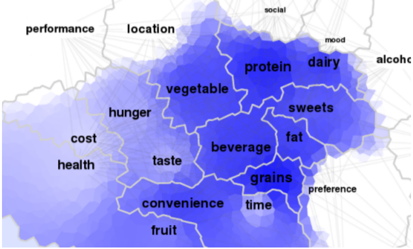
- Some limitations: short, sparse, slang, self-reported
- But, relatively broad usage across ethnic, gender, age, and socio-economic groups
- Tweets are freely available (in limited amounts) and easy to access in real-time
- Geographic linguistic analysis [Eisenstein et al. 2010]



Flu prediction using Twitter [Paul and Dredze 2011]

Twitter and Diet

- Social media interventions can reduce obesity modestly [Ashrafian et al. 2014]
- Twitter can be a greater source of positive influence for weight loss than family or friends [Pagoto et al. 2014]
- Previous work explores food logging and visualization via Twitter [Hingle et al. 2013]

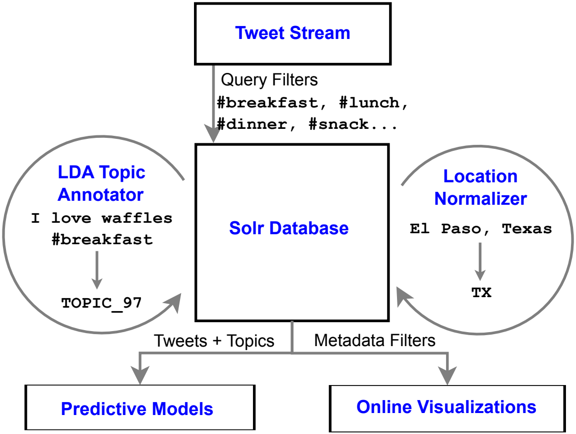


Heatmap of hashtag usage in food tweets, by Kobourov and Schneider

Goals of this Work

- Predict diabetes and obesity rates and social characteristics for communities
- Analyze and visualize predictive features of language and geographic variation in diet
- Move toward automatic risk identification (and intervention?) to prevent dietary-related diseases like diabetes

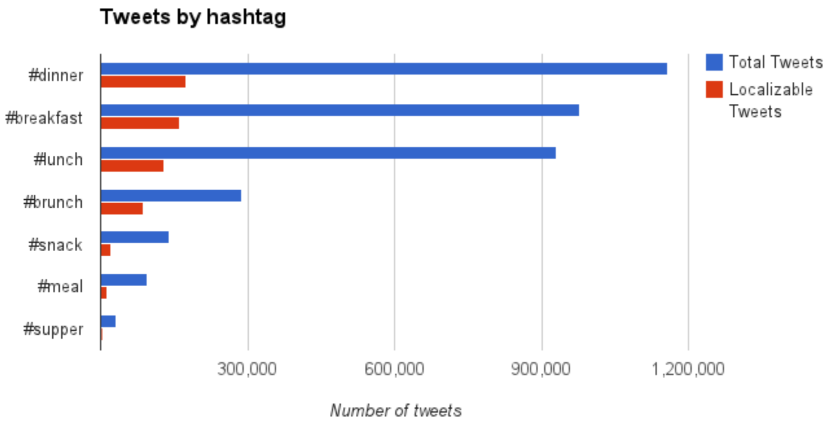
Collecting, Analyzing, and Visualizing Tweets



Collect tweets from the Twitter streaming API, store and query using Apache Solr

Tweet Corpus

- 3.5 million tweets collected from October 2013 - May 2014, worldwide
- Average tweet length: 8.7 words
- 30 million words, 1.5 million unique



State Location Normalization

- User can supply location for their account
- Regular expression matching on state names plus a few heuristics (e.g. “LA” + time zone → California or Louisiana)
- 560,000 tweets (16%) could be normalized to a US state

State Trends

Highest-ranked food word per state
using *term frequency – inverse document frequency*

Predictive Task Goals

- Predict diabetes and obesity rates and social characteristics on a state level
- Analyze and visualize predictive features of language and geographic variation in diet

Prediction Tasks

- Using the text of all tweets for a state, predict:
  - Diabetes rate: above or below the national median?
  - Overweight rate: above or below the national median for high BMI?
  - Political tendency: more Republican or Democratic votes?

Prediction Tasks

- Using the text of all tweets for a state, predict:
  - Diabetes rate: above or below the national median?
  - Overweight rate: above or below the national median for high BMI?
  - Political tendency: more Republican or Democratic votes?
- Location prediction
  - Predict geographic locale for a group of tweets
  - City: 15 largest in US by population
  - State: 50 states + Washington D.C.
  - Region: West, Midwest, South, or Northeast

Lexical Features

- Simple bag-of-words model: count number of times each word appears across all tweets
  - All words, food words, or hashtags

word	count in CA
dinner	36896
breakfast	32314
lunch	27616
brunch	16845
food	8748
...	...

word	count in NY
dinner	26617
brunch	23189
breakfast	22697
lunch	19142
food	6393
...	...

Example: Top food words, California vs New York

Topical Features

- Sparsity: tweets have short length and unique vocabulary
- Use Latent Dirichlet Allocation (LDA) to cluster words into 200 topics:

```
coffee, starbucks, cafe, morning, ...
vegan, vegetarian, healthy, ...
japanese, sushi, bento, ...
chicken, potatoes, fried, ...
```

- Train LDA model on all tweets
- Use the highest probability topic for each tweet as a feature

Classification Framework

- Support Vector Machine with linear kernel
- Diabetes, obesity, and political prediction
    - Predict labels for a single state using its tweets and all other states' data
    - Rotate through all states and average accuracy

Classification Framework

- Support Vector Machine with linear kernel
- Diabetes, obesity, and political prediction
  - Predict labels for a single state using its tweets and all other states' data
  - Rotate through all states and average accuracy
- Location prediction
  - Split into training data (80% of tweets) and testing data (20%) for each location
  - Train a classifier for each location
  - For a given set of testing tweets, predict the location

Diabetes, Obesity, and Political Accuracies

	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

Percentage of states classified correctly for each task and feature set.

Diabetes, Obesity, and Political Accuracies

	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

Percentage of states classified correctly for each task and feature set.

- All Words best on average, but Food words alone are nearly as good

Diabetes, Obesity, and Political Accuracies

	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

Percentage of states classified correctly for each task and feature set.

- All Words best on average, but Food words alone are nearly as good
- Best performance on overweight, but political and diabetes are well above the baseline

Diabetes, Obesity, and Political Accuracies

	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

Percentage of states classified correctly for each task and feature set.

- All Words best on average, but Food words alone are nearly as good
- Best performance on overweight, but political and diabetes are well above the baseline
- Topic modeling is 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 Advertising* (open, today, come, join), #date, happy, #dinner, 10

Feature Analysis - Diabetes

Rank individual features by the weights assigned by the SVM

Class	Highest-weighted features
diabetes: +	*Mexican* (mexican, tacos, burrito), *American Diet* (chicken, baked, beans, fried), #food, *After Work* (time, home, after, work), #pdx, my, lol, #fresh, *Delicious* (foodporn, yummy, yum), #fun, morning, special, good, cafe, #nola
diabetes: -	#dessert, *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, *Japanese* (ramen, japanese, noodles), some

Feature Analysis - Political

Rank individual features by the weights assigned by the SVM

Class	Highest-weighted features
Democrat	#vegan, #yum, w, served, #brunch, *Deli* (cheese, sandwich, soup), photo, #rvadine, *Restaurant Advertising* (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)

City, State, and Region 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**

Accuracy in predicting location for a group of tweets

City, State, and Region 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**

Accuracy in predicting location for a group of tweets

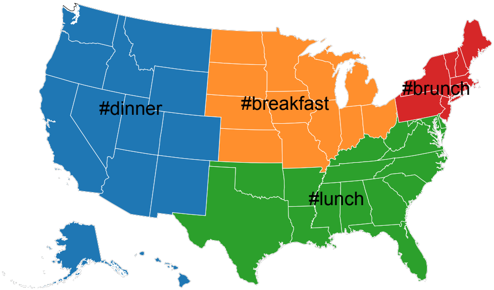
- Context is even more important here: All Words is best by a wider margin
- Food alone still improves substantially on baseline

City Prediction Features

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

Top five highest-weighted features for each city. LDA topics *italicized*

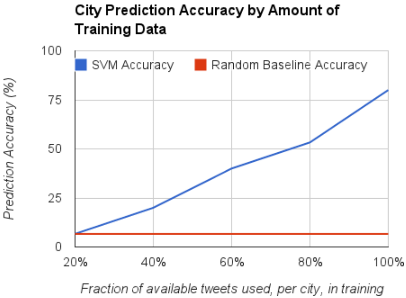
Region Prediction Features

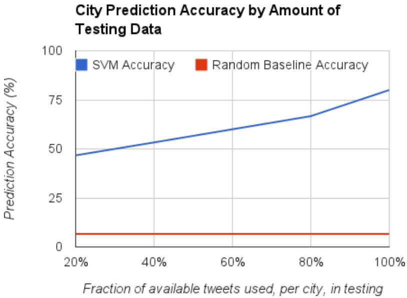


Region	Highest-weighted features
Midwest	#breakfast, i, #recipes, *After Work*, *Recipe*,
Northeast	#brunch, brunch, our, *Mixed Drinks*, we,
South	#lunch, *Mixed Drinks*, *After Work*, *American Diet*, chicken,
West	#dinner, #food, #foodporn, photo, dinner,

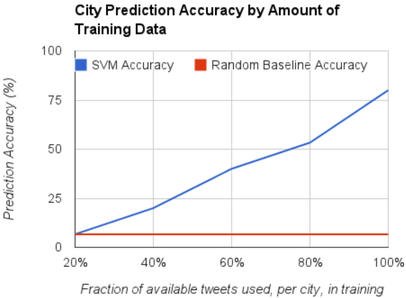
Top 5 highest-weighted features for predicting each region from its tweets.

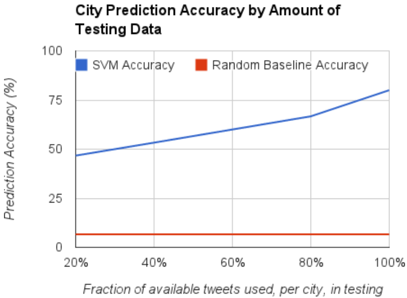
Learning Curves





Learning Curves

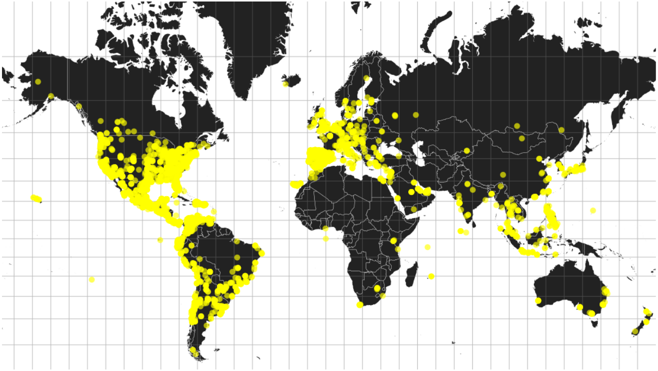




- Accuracy increases with size of training and testing data
- Can do relatively well with small testing set as long as training is large
- Same effect for state and region prediction

Tweet Location Visualization

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

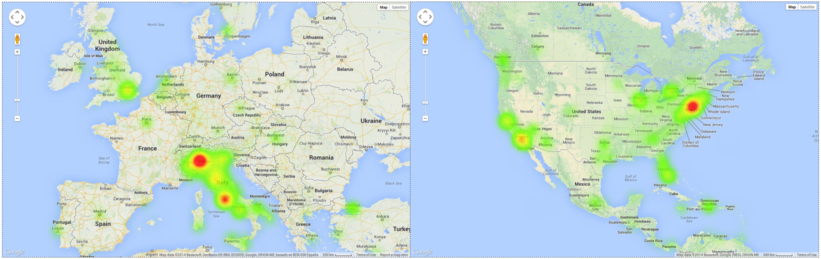


11,827 tweets from five *Spanish/Latin American food* topics (tacos, burritos, salsa, pollo, arroz, paella, ...)

[Link to live version]

Tweet Heatmaps

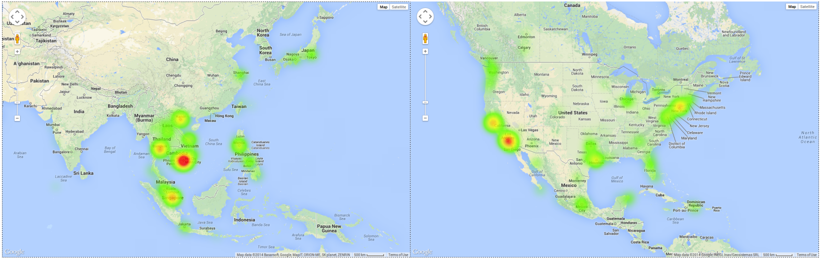
- Global trends possibly reflect migration patterns



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

Tweet Heatmaps

- Global trends possibly reflect migration patterns

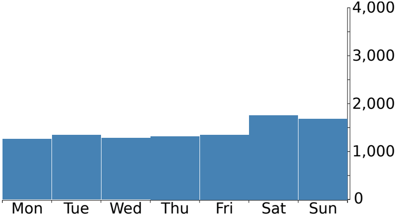


Heatmaps of 1,032 tweets from a *Vietnamese food* (pho, vietnamese, ...) topic.

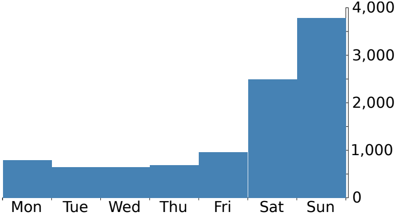
Link to live version

Temporal Histograms

- 71% of tweets (2.5 million) have a time zone
- Allow temporal analysis at varying granularities



(a) Tweets containing “breakfast”

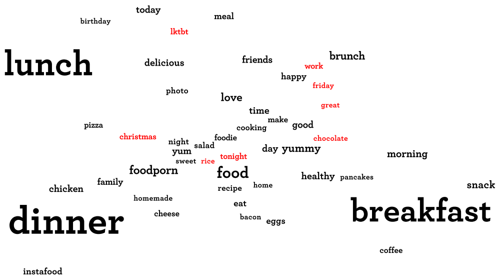


(b) Tweets containing “brunch”

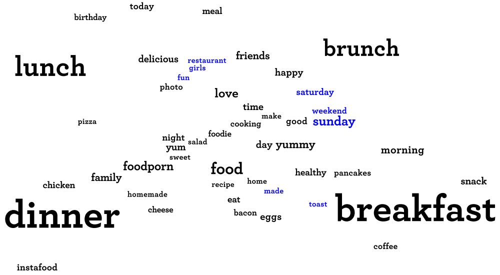
Link to live version

Parallel Semantic Word Clouds

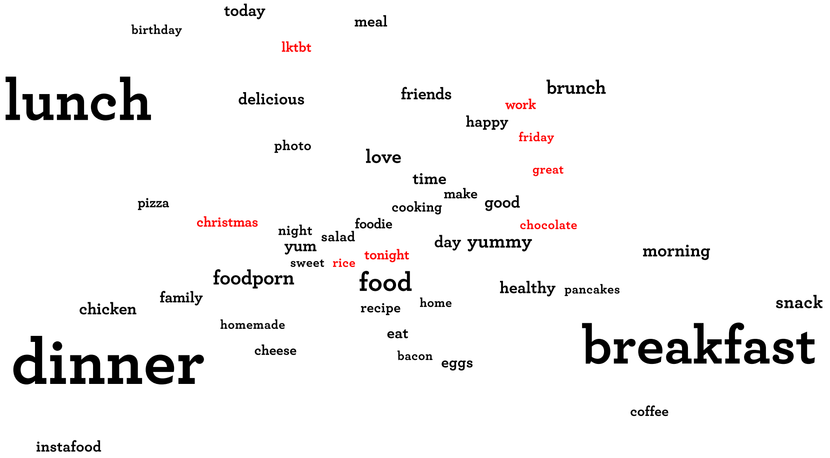
weekday:



weekend:

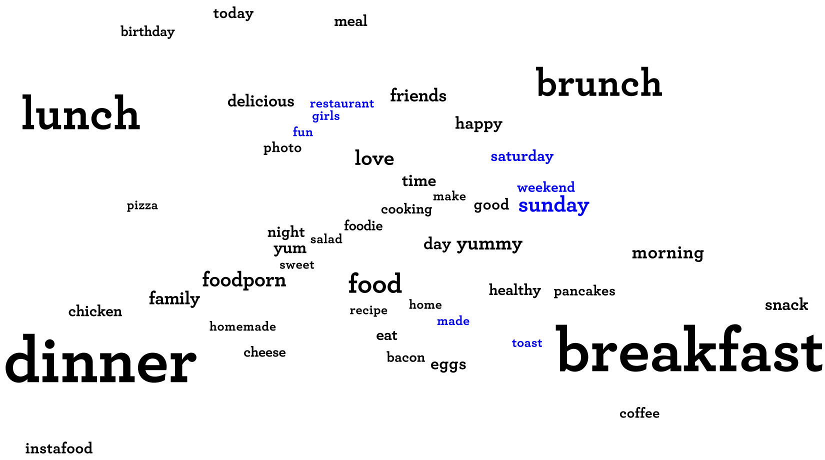


Parallel Semantic Word Clouds



Weekday Wordcloud
visualization by Jixian Li

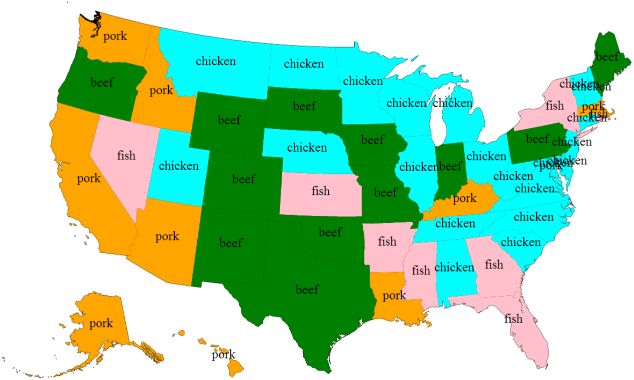
Parallel Semantic Word Clouds



Weekend Wordcloud
visualization by Jixian Li

State-Level Term Comparison

- Visualize the most popular term from a given set of words



Chicken, pork, beef, or fish?
visualization by Charlie Morfoot

[Link to live version]

Conclusions

- Language of food has predictive power for population characteristics, especially geographic locale
- Much of the predictive power comes from food words alone
- Future work:
  - Predict individual diabetes risk, obesity, etc. using diet
  - 20 Questions: can we guess where you live based on what you eat?
  - Analyze food-based communities and network effects
  - Use images and video in addition to text

Thanks!