Carmen:
A Twitter Geolocation System with Applications to Public Health

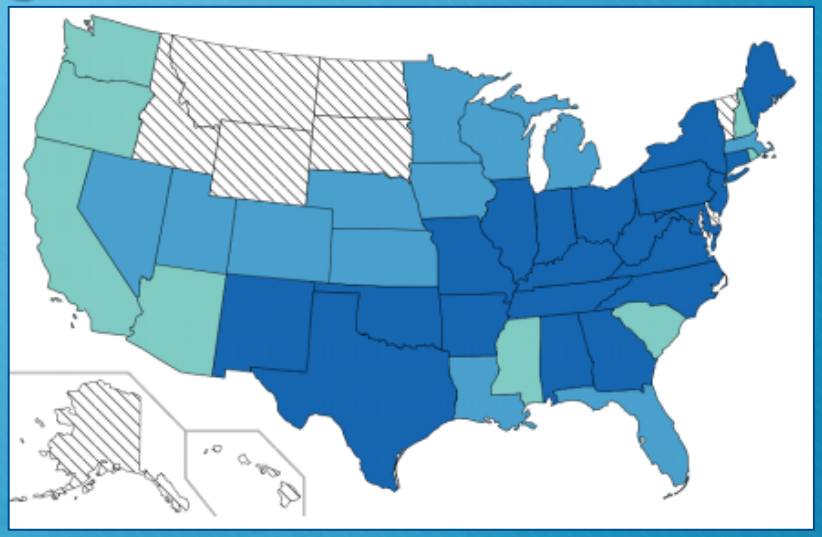
Mark Dredze, Michael Paul,
Shane Bergsma, Hieu Tran

Twitter and Public Health

- Many studies have shown that health information can be extracted from Twitter
  - Influenza prevalence
    - Culotta, 2010;
      Sadilek and Kautz, 2012
  - Depression
    - Choudhury et al, 2013
  - Medication use (Paul and Dredze, 2012)

Twitter and Public Health

- Such analyses often require information about the location of the tweets and users
  - e.g. allergy prevalence is higher in some U.S. states
    - Paul and Dredze, 2011



Twitter Geolocation

- Most studies rely on location metadata
  - GPS coordinates associated with tweets
  - Structured locations resolved by Twitter
  - Covers less than 3% of users
    - And we are already working with a 1% sample…

Twitter Geolocation

- User profiles
  - Self-reported locations
  - 56% of users fill this in
- Tweet content
  - Language analysis
  - More involved – we don't do this here

Carmen

- Returns structured object for each tweet
  - City
  - County
  - State
  - Country
- Fast and simple
  - 27,000 tweets / sec
  - Code available on Github

Carmen

- Uses GPS data when available
  - Location information was obtained from Yahoo Maps API
- Mapping of user profile strings to places
  - e.g. “NYC”, “New York” -> {city: New York, state: NY, …}
  - Manually curated
  - Automatically added aliases using location clusters created from social network structure
    - Bergsma et al, 2013
  - 4,811 unique places in our mapping

Evaluation

- Treated GPS locations as ground truth
- Evaluated geolocation from user profiles against the ground truth
  - Accuracy (precision)
  - Coverage (recall)
- Test set: 56,000 tweets (plus 10,000 dev)

Evaluation

- Accuracy:
  - Country: 91%
  - State+Country: 65%
  - Within 250 miles: 75%
  - Within 25 miles: 55%
- Coverage:
  - 44% (38% without automatic alias extensions)
  - 22% of tweets can be geolocated

Influenza Surveillance

- Classifier to detect tweets about flu infection
  - Lamb, Paul, Dredze, 2013
- Estimate flu prevalence in the US and UK
  - Evaluation: compare to government data

Influenza Surveillance

- Pearson correlation:

	US	UK	US	UK
Location	2009		2011	
All Twitter	.9604	.5138	.6993	.6010
US Only	**.9714**	.1982	**.7792**	**.6312**
UK Only	.9231	**.8827**	.6277	.5123

Thank You

- If you want to try Carmen, go to:
  - https://github.com/mdredze/carmen