Citizen-Sourced Data for Public Health Modeling

Rumi Chunara, PhD
Assistant Professor
NYU Computer Science & Engineering
NYU Global Public Health
Citizen-Sourced Data for Public Health Modeling

• Public health intro and data overview
• Crowdsourcing and knowledge generation in public health
• Learning spatio-temporal features
• Other data opportunities
Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities
Emerging and Re-emerging Infections, 1996-2010

- Cryptosporidiosis
- Human Monkeypox
- E.Coli O157
- Venezuelan Equine Encephalitis
- Dengue Haemorrhagic Fever
- Ebola Haemorrhagic Fever
- Marburg Haemorrhagic Fever
- Ross River Virus
- Hendra Virus
- Reston Virus
- West Nile Virus
- Legionnaire’s Disease
- Severe Acute Respiratory Syndrome (SARS)
- Malaria
- Typhoid
- Cholera
- BSE
- Lassa Fever
- Yellow Fever
- Lyme Borreliosis
- Echinococcosis
- Diphtheria
- Influenza A (H5N1)
- Nipah Virus
- RVF/VHF
- O’Nyong-Nyong Fever
- Buruli Ulcer
- Multidrug Resistant Salmonella
- nvCJD
Current Public Health Surveillance

- Whole population: +
- Specificity: +
- Doctor/Nurse involvement: +
- Speed: -
- Sensitivity: -
- Cost: -
- Public engagement: -
<table>
<thead>
<tr>
<th></th>
<th>Current Public Health Surveillance</th>
<th>Crowdsourced Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole population</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Specificity</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Doctor/Nurse involvement</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Speed</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Cost</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Public engagement</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
Citizen-Sourced Data for Public Health Modeling

• Public health intro and data overview
• Crowdsourcing and knowledge generation in public health
• Learning spatio-temporal features
• Other data opportunities
Types of Crowdsourcing

- Contractual
- Contributory
- Collaborative
- Co-Created
- Collegial

Shirk et al. 2011
80% of citizen projects only harness participation in a limited form, such as for completion of tasks.
Findings from Our Efforts Generating Public Health Knowledge via Crowdsourcing

1. Unique incentives in public health: collective and intrinsic motivations are the most salient (aligns with Nov et al. 2011, Law et al. 2016)

2. Offline recruitment: important to improve external validity (Chunara et al. AJPM 2016)

3. Uniqueness of data generated: Spatio-temporal data Opportunities (Salathé et al. 2011, Relia et al., Rehman et al. 2017)

Chunara (under submission) 2017
Data Challenges

Observations, MAR or CAR, depends on PAR (denominator)

Value of interest, $y$

$y \sim \mathcal{N}(\phi(t) + f(t), \sigma^2)$

Variation:
- crowdsourced non-specificity (e.g. nlp classification)
- sample confidence
- other noise (uncorrelated) $\sigma^2$
Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities
Learning Features – Twitter, alcohol behavior example

• Social media (e.g. Twitter) recognized as useful source to learn about incidence and behaviors related to infectious and non-communicable diseases

• Much work relates to time-series, mapping and prediction

• We developed an NLP pipeline to specifically isolate individuals engaged in a behavior, and then examines relevant temporal representations (features) (Liu et al CSCW 2017, Huang et al CSCW 2018)
NLP Hierarchical Pipeline

All Tweets

Behavior

First Person
- Reflecting
- Current
- Planning
- Other
- Non-Behavior

Liu et al. ACM CSCW 2017
High-resolution Temporal Representations

L2 Normalized Tweets Over Week

Week 1

Week 2

Week 3

Week 4

Hour of Week (dotted lines at 8am local time)

red: overage, blue: underage

Huang et al. ACM CSCW 2018
Spatial Representations

Problem:
- Social attitudes like racism/homophobia can be predictors for health outcomes
- ZIP codes are defined to optimize mail delivery, need way to define exposure from a place based on the context

Approach:
- Develop a method for spatial representations based on SOMs and social media classification
- Use mobility data from a cohort of MSM to show that spatial representations matter
Learning Features – Twitter, social attitude example

- Use SOM’s a common neural network approach to learn embedded structure in the spatial distribution of classified Tweets

- Provides an interpretable output that can be mapped geographically

Illustration depicting grid cell (i,j) and its neighbors at threshold = 1 for formation of boundary between grid cells with different weights (value in second row)

Relia et al. (under review) 2017
Learning Features – Twitter, social attitude example

Distribution of Racism Exposure by ZIP

Distribution of Racism Exposure by SOM

Relia et al. (under review) 2017
How Good Are These Representations?

- Evaluate using common cluster evaluation techniques:
  - Robustness to missing data
  - Lower mean variance
  - Entropy

Overall SOM provide a more consistent geographical compartmentalization of each attitude
What Difference Does This Make?

Mean racism exposure difference using SOM versus Zip Codes was **40.3%** (SD: 18.8%).
Citizen-Sourced Data for Public Health Modeling

• Public health intro and data overview
• Crowdsourcing and knowledge generation in public health
• Learning spatio-temporal features
• Other data opportunities
Data Opportunities

- Predicting individual-level mobility a growing problem of importance
- Existing data (CDRs, GPS trackers) temporally rich, though expensive
  - High-resolution, and availability of social media
  - Sparse nature brings challenges
Intermediate Location Computing: Pre-processing

- 6 months of Twitter data from the API
- Pre-processed data by mapping to grids (0.1, 0.5 and 1 mi resolution)
- Inferred stay and home location
- Removed non-personal accounts
- Included users must have at least 1 location value present for each h during daytime hours (irrespective of day and week)
- 29,491, 4,947 and 1,119 users ($r_i = 1$ hour) and 45,710, 8,083 and 2,395 users ($r_i = 2$ hours) included from NYC, DC and SF
Intermediate Location Computing: Algorithm

\[ x_k \] location at position k

\[ P_{l,a} \] Next location probability

\[ P_{l,b} \] Previous location probability

\[ P_{l,c} \] Community location probability

\[ WS \] Week and hour specific

\[ Inter(F1,F2) = \begin{cases} \text{max}_\text{loc}(F1) & \text{if } \text{max}_\text{loc}(F1) \neq \text{NULL} \\ \text{max}_\text{loc}(F2) & \text{otherwise} \end{cases} \]

\[ P_{l,a}(x_k=j \mid x_{k-1}=18) \cdot (1-\alpha) \]

\[ P_{l,b}(x_k=j \mid x_{k+1}=32) \cdot (1-\alpha) \]
ILC: Accuracy versus baseline models

<table>
<thead>
<tr>
<th>City</th>
<th>$r_i$</th>
<th>Top 1</th>
<th>Top 3</th>
<th>Home-Work</th>
<th>Markov O(0)</th>
<th>Markov O(1)</th>
<th>POI</th>
<th>NextPlace</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York City</td>
<td>$r_i=1$</td>
<td>72.69</td>
<td>82.35</td>
<td>65.54</td>
<td>64.65</td>
<td>26.39</td>
<td>15.59</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>$r_i=2$</td>
<td>64.78</td>
<td>77.38</td>
<td>59.28</td>
<td>57.98</td>
<td>32.56</td>
<td>19.11</td>
<td>0.21</td>
</tr>
<tr>
<td>Washington, DC</td>
<td>$r_i=1$</td>
<td>75.08</td>
<td>83.61</td>
<td>66.91</td>
<td>65.76</td>
<td>27.75</td>
<td>31.27</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>$r_i=2$</td>
<td>68.85</td>
<td>79.57</td>
<td>62.35</td>
<td>60.64</td>
<td>34.13</td>
<td>34.56</td>
<td>0.19</td>
</tr>
<tr>
<td>San Francisco</td>
<td>$r_i=1$</td>
<td>77.20</td>
<td>86.28</td>
<td>67.74</td>
<td>67.21</td>
<td>16.78</td>
<td>35.49</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>$r_i=2$</td>
<td>70.78</td>
<td>82.06</td>
<td>63.66</td>
<td>62.91</td>
<td>19.52</td>
<td>32.69</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Transfer Learning to Improve Specificity

- Point value specificity challenging in crowdsourced data
- Manifests in public health through syndromic data
When Google Got Flu Wrong

Sources: Google Flu Trends (www.google.org/flu trends); CDC; Flu Near You
Types of Healthcare-facilitated Data Collection

Aliapoulios et al. (in prep) 2017
# From the Crowd to the Clinic

<table>
<thead>
<tr>
<th>Study</th>
<th>Location</th>
<th>Num. Observations (positive)</th>
<th>Symptoms Recorded</th>
<th>Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYUMC</td>
<td>New York</td>
<td>278 (23)</td>
<td>cough, diarrhea, fatigue, fever, headache, muscle, nausea, sorethroat, vomit</td>
<td>Clinical (emergency room)</td>
</tr>
<tr>
<td>GoViral</td>
<td>New York</td>
<td>899 (201)</td>
<td>body aches, chills, cough, diarrhea, fatigue, fever, leg pain, nausea, runny nose, shortness of breath, sorethroat, vomit</td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td>Ilorin, Nigeria</td>
<td>98 (23)</td>
<td>body ache, chills, cough, fever, nausea, runny nose, shortness of breath, sorethroat, vomit</td>
<td></td>
</tr>
<tr>
<td>Flu Watch</td>
<td>United Kingdom</td>
<td>2759 (844)</td>
<td>fever, cough, sorethropat, runny nose, blocked nose, sneeze, diarrhea, muscle, headache, rash, earache, wheezy, chills, joint aches, loss of appetite, fatigue, vomit, nausea</td>
<td></td>
</tr>
<tr>
<td>Hong Kong</td>
<td>Hong Kong</td>
<td>4379 (917)</td>
<td>cough, fever, headache, muscle, phlegm, runny nose, sorethroat</td>
<td>Secondary infections recorded by a health worker</td>
</tr>
<tr>
<td>Hutterite</td>
<td>Canada</td>
<td>1897 (628)</td>
<td>blocked nose, chills, cough, earache, fatigue, fever, headache, muscle, runny nose, sorethroat</td>
<td>Health-worker facilitated</td>
</tr>
</tbody>
</table>
Transfer Learning Paradigm

\[ \mathcal{D} = \{(x_{ji}, y_{ji}) \mid x_{ji} \in \mathcal{X}_j, y_j\}_{i=1}^{n_j} \]

\[ y(x_i) = \frac{1}{1 + \exp(-(b_0 + w x_i))} \]

1. Blind transfer
2. Additive transfer
3. Projection on latent space (tbd)
Performance so far...

<table>
<thead>
<tr>
<th>Study</th>
<th>Nigeria</th>
<th>Hong Kong</th>
<th>Hutterite</th>
<th>GoViral</th>
<th>FluWatch</th>
<th>NYUMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nigeria</td>
<td>0.56, 0.56</td>
<td>0.59*, 0.65</td>
<td>0.50, 0.56*</td>
<td>0.59, 0.65</td>
<td>0.50*, 0.56</td>
<td>0.50*, 0.50†</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>0.68†, 0.81</td>
<td>0.82, 0.82</td>
<td>0.55, 0.67</td>
<td>0.79*, 0.77†</td>
<td>0.55, 0.61</td>
<td>0.50, 0.68</td>
</tr>
<tr>
<td>Hutterite</td>
<td>0.53*, 0.50†</td>
<td>0.54*, 0.52†</td>
<td>0.55, 0.55</td>
<td>0.53*, 0.47</td>
<td>0.51*, 0.51†</td>
<td>0.50*, 0.50†</td>
</tr>
<tr>
<td>GoViral</td>
<td>0.68, 0.75†</td>
<td>0.79*, 0.78†</td>
<td>0.53, 0.55</td>
<td>0.79, 0.79</td>
<td>0.53, 0.57</td>
<td>0.50, 0.57</td>
</tr>
<tr>
<td>Flu Watch</td>
<td>0.52*, 0.43</td>
<td>0.54*, 0.53†</td>
<td>0.51*, 0.56†</td>
<td>0.53*, 0.55†</td>
<td>0.56, 0.56</td>
<td>0.51*, 0.52†</td>
</tr>
<tr>
<td>NYUMC</td>
<td>0.50, 0.81†</td>
<td>0.68, 0.67</td>
<td>0.68, 0.68</td>
<td>0.57, 0.85†</td>
<td>0.63, 0.85†</td>
<td>0.86, 0.86</td>
</tr>
</tbody>
</table>

Aliapoulis et al. (in prep) 2017
Citizen-Sourced Data for Public Health Modeling

• Public health intro and data overview
• Crowdsourcing and knowledge generation in public health
• Learning spatio-temporal features
• Other data opportunities
Acknowledgements

• Jason Liu
• Tom Huang
• Nabeel Rehman
• Kunal Relia
• Anas Elghafari
• Vishwali Mhasawade
• Mohammad Akbari, PhD
Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities

Questions?
rumi.chunara@nyu.edu