Course: Statistics and Modeling with Novel Data Streams  
Instructors: Alex Vespignani, Elaine Nsoesie, and Mauricio Santillana  
When: July 16, 17, 18; 2018  
Where: University of Washington, Seattle. Summer Institute in Statistics and Modeling in Infectious Diseases  

Software Requirements: Python, R and/or Matlab  

Total of hours in the course: 15 hours  

General plan

Monday

8:30am-10am Part 1: Predicting Dengue using Google searches. A tutorial (Mauricio)  
- Short Introduction to using Google searches as a way to monitor disease activity  
- Hands-on exercise 1.0: Coding a simple version of Google Dengue Trends  
- Single variable approach. Static vs dynamic  

10:30am - 12pm Part 2: Multivariable linear models and Google searches and News alerts. A tutorial (Mauricio)  
- Hands-on exercise 1.1: Coding a simple version of Google Dengue Trends  
- Multiple variable approach. Static vs dynamic  
- Incorporating historical information (seasonality)  
- Hands-on exercise 2.0: Early estimation of Zika cases in Colombia (2016 outbreak) using news alerts. Data download

LUNCH

1:30pm - 3pm Using non-traditional data sources for foodborne illness and unsafe food surveillance (Elaine)  
- Introduction to data streams: Yelp, Twitter, Amazon and Crowdsourced surveillance  
- Machine learning and natural language processing  
- Developing surveillance dashboards for Departments of Health  
- Hands-on exercise: Coding a machine learning classifier for foodborne illness tweets

3:30am - 5pm General introduction (Alex)
- Prediction
- What we need (input)
- Big data streams: Passive (Google, Twitter); Active (RFID TAGS: contact network patterns, Flu Near You)
- Models

Models introduction (Alex)
- Generative models
- Compartmental disease representation
- Stochastic models
- Individual based & Networks models
- Network data
- RFID experiments

<table>
<thead>
<tr>
<th>Tuesday</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>8:30am-10am</strong></td>
</tr>
<tr>
<td><strong>Data streams introduction: beyond Google searches, and beyond Flu (Mauricio)</strong></td>
</tr>
<tr>
<td>- Flu Trends and Dengue Trends</td>
</tr>
<tr>
<td>- Summary: using ARGO to predict disease outbreaks</td>
</tr>
<tr>
<td>- News reports (success stories with Ebola and Zika)</td>
</tr>
<tr>
<td>- <strong>Hands on exercise 2.1</strong>: Early estimation of Zika cases in Colombia (2016 outbreak) using news alerts</td>
</tr>
<tr>
<td>- Continuation of Zika exercise/ Using Twitter to track flu (Mauricio)</td>
</tr>
<tr>
<td>- <strong>Hands on exercise 3</strong>: exploring flu-related tweets</td>
</tr>
<tr>
<td>- Real-time prediction. Development of a website to scrape/centralize information</td>
</tr>
<tr>
<td>- Beyond Google searches (Flu).</td>
</tr>
<tr>
<td>- UpToDate and Electronic Health Records as proxies for disease incidence</td>
</tr>
<tr>
<td>- Participatory surveillance as a way to track diseases.</td>
</tr>
<tr>
<td>- “Together we are stronger”: ensemble approaches lead to more robust systems.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10:30am - 12pm</th>
<th><strong>Human mobility Data and Modeling (Alex)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Human Mobility (mobile call records, airline data etc.)</td>
</tr>
</tbody>
</table>
- Spatially structured models
- Metapopulation networks
- Data-driven large scale simulations

**LUNCH**

**1:30pm - 3pm**

**Image Processing and Disease Surveillance (Elaine)**
- Sources for image data
- **Hands-on exercise**: Acquiring image data
- Overview on image processing with deep learning and other methods
- Use of image data for disease surveillance

**3:30pm - 5pm**

**Flu Near You and participatory surveillance (Mauricio)**
- What is Flu Near You?
- Demographics of FNY.
- Consistent users vs sporadic users.
- How many reports do we need to see a signal?
- Getting access to alternative data sources (Google searches, Twitter)
- Mobility patterns in Brazil (determined by airline data, bus data and Twitter)

**Final Considerations: Building a tool to track diseases in real time. Challenges**
- Google correlate, Google trends, Twitter API,
- Multivariable models (variable aggregation vs individual contribution)
- Adding auto-regressive information: ARGO
- Technicalities: CDC (Data acquisition, revisions), Google (Sampling issues)
- Google Correlate
- Coding ARGO a near-time and forecasting model
- Natural Language processing
- Beyond Flu. Dengue and Zika prediction using Google searches
- Human mobility built from airline data, bus data and twitter data.
- Role of mobility in the spread of Dengue in Brazil
Wednesday

8:30am – 10am  Global Epidemic and Mobility Platform  
Data-Driven Simulations and Forecast (Alex)  
- Near-time and Real-time forecast  
- Seasonal Flu  
- Emerging infectious diseases

10:30am - 12pm  Data limitations, representation and ethics (Elaine)  
- Methods for identifying and addressing limitations in novel data streams  
- Machine learning and data matching for demographic inference  
- Incorporating demographic differences in statistical models  
- Ethics and digital data