FROM FORECASTING THE FLU TO PREDICTING THE “NEXT” DISEASE

UP-STAT 2016 - Buffalo, NY

Shannon Gallagher
April 23, 2016

Carnegie Mellon University
Department of Statistics
Lee Richardson
Sam Ventura
Ryan Tibshirani
Bill Eddy
Department of Machine Learning
Roni Rosenfeld
We want to better predict of the spread of infectious diseases
Infectious diseases are often ...

- old
- deadly
- costly
- stochastic
We want to mitigate the effects of disease through prediction.

With accurate predictions, the infectious diseases are:

- old
- deadly → manageable
  - Resource allocation
  - Alert health officials
  - Issue warnings
- costly → feasible
  - Fewer sick days
  - More awareness
- stochastic → forecasted
WE FOCUS ON 10 CDC REGIONS

Figure: From cdc.gov
### Table: Cross Section of Available data.

<table>
<thead>
<tr>
<th>Region</th>
<th>Year</th>
<th>Week</th>
<th>wILI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2015</td>
<td>25</td>
<td>0.55</td>
</tr>
<tr>
<td>2</td>
<td>2015</td>
<td>25</td>
<td>1.45</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>2015</td>
<td>25</td>
<td>0.44</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>2015</td>
<td>45</td>
<td>0.76</td>
</tr>
<tr>
<td>2</td>
<td>2015</td>
<td>45</td>
<td>1.57</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>2015</td>
<td>45</td>
<td>0.89</td>
</tr>
</tbody>
</table>

**wILI** - Weighted Influenza Like Illness
WE WANT TO PREDICT THE WILI FOR THE REMAINING WEEKS OF A SEASON

Figure: Examples of WILI curves. From David Farrow’s FluV. epicast.org
**THE MODEL RELIES ON SEASONAL VARIABLES**

\[ Y_{t}^{(r,s)} \sim N(\mu_{t}^{(r,s)}, \sigma^2) \]

where

\[ \mu_{t}^{(r,s)} = [a_s \cdot \alpha_r] \cdot f(t - b_s - \beta_r) \]

for week \( t \), region \( r \), season \( s \) and priors:

- \( a_s \sim \text{Unif}(2, 10) \) - seasonal scaling
- \( b_s \sim \text{Unif}\{-6, -5, \ldots, 6\} \) - seasonal shifting
- \( \alpha_r \sim \text{Unif}(0.25, 1.25) \) - regional scaling
- \( \beta_r \sim \text{Unif}\{-3, -2, \ldots, 3\} \) - regional shifting
- \( f \sim \hat{F} \) - smoothed observed curves
- \( \sigma^2 \sim \text{Unif}(0.5, 2.5) \) - variance
ESTIMATING A POINT IS SIMPLER THAN ESTIMATING A CURVE

**Figure**: Leave-one-season-out cross validation for EB with regional effects (a) and a targeted regression (b). The x-axis is the weeks from the observed peak and the y-axis is the mean absolute error.
Figure: Image depicting of weighting curves whose peak values are closer to our estimated values. A thicker line represents a larger weight. The black dot is our estimated value of the peak height and week.
ULTIMATELY, WE WANT TO BIAS ESTIMATES/SHRINK POSTERIOR

Figure: Image depicting of weighting curves whose peak values are closer to our estimated values. A thicker line represents a larger weight. The black dot is our estimated value of the peak height and week.
Figure: Cross Validation error averaged over the different seasons for the different model types.
THE “NEXT” DISEASE
For past diseases like the flu, we have

- Years of data
- Knowledge of the disease
- Public awareness
- Specific models
THE TOOLS WE HAVE FOR NEW DISEASES ARE LIMITED

For **past** diseases like the flu, we have

- Years of data
- Knowledge of the disease
- Public awareness
- Specific models

But for **new** diseases we have

- Little data
- Less knowledge
- Frenzied awareness
- Few, if any, models
Agent-based models (ABMs) use agents, an environment, and update rules:

```plaintext
for (time in time steps)
    agents = update(agents, env)
end
```
Agent-based models (ABMs) use agents, an environment, and update rules.

```python
for (time in time steps)
    agents = update(agents, env)
end
```

ABMs are flexible and modular!
An ABM can incorporate:

- Transmission Type
- Reproduction Rate
- Cultural factors
- Prevention strategies
ABMS REQUIRE HIGH QUALITY AGENTS - SPEW

- ~ 4 billion agents
- 80+ countries
- Automatic diagnostic reports
- 2 custom populations from users
  - Canada (Data from CDs)
  - California (Hispanic Population)
- 2 location sampling modules
  - Uniform and Road-Based
- 4 sampling schemes
  - Uniform, Moment Matching, IPF, Density Estimation
- Open Source
  - https://github.com/leerichardson/spew

**Figure:** Synthetic Populations and Ecosystems of the World
IN SUMMARY
When we have data, we can build rich models (Flu)

Agent-Based Modeling can be used to simulate diseases previously unseen
Questions?