**Public Health Surveillance in High-Dimensions with Supervised Learning**

**Poster by:** Saylisse Dávila  
**Research Advisor:** Dr. George Runger

---

**What is Public Health Surveillance?**  
The monitoring of statistics for evidence of a change through the continuous, systematic collection and analysis of measurements (Lawson and Kleinman, 2005).

**Motivation / Research Focus**  
Increments in disease incidence rates present a threat to the public, thus, it is important to have in place a surveillance system that rapidly detects these increments to prevent a disease from spreading through the population across a large region. We will develop a method that allows to detect changes in disease incidence rates that occur only in a region and/or subpopulation.

**Transform to Supervised Learning**  
Contrast data from different time periods and, then, use a supervised learner to generate a signal whenever there is a difference between the incidence rates for the background period and the new test period.

**Tree-Based Ensembles**  
Learners from tree-based ensembles were adopted since:
- 1) They are able to isolate regions of different proportions of events.
- 2) They provide embedded importance measures that can identify the covariates relevant to the rate changes in disease incidence.

---

**Why Feature Selection in Public Health Surveillance?**
- Can explore the entire covariate space and select only those covariates that drive the disease clusters in our data.
- These covariates are the key to identify disease clusters that occur only in a region, time period, and/or subpopulation.

**Feature Selection**  
The covariates relevant to a rate change in disease incidence were selected using the feature selection methodology proposed by Tuv et al. (2007). Their methodology uses artificial variables that allow to perform:

1) Statistical inference on the significance of importance measures.
2) Statistical inference on the significance of masking measures among covariates.

**Experiment Description**
- Data set: contains all 2006 methicillin-resistant staphylococcus aureus (MRSA) cases from Maricopa County recorded in AZHQ.
- Covariates: gender, age, race, latitude, and longitude.
- Class 1 cases randomly sampled from MRSA data set and superimposed with different types of clusters.
- Each case was replicated 10 times.

---

**Data Plots**

**Covariate Selection**

<table>
<thead>
<tr>
<th>Cases</th>
<th>Class 0 Cases</th>
<th>Class 1 Cases Out of Cluster</th>
<th>Class 1 Cases In Cluster</th>
<th>Covariates In Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2679</td>
<td>250</td>
<td>0</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>2679</td>
<td>225</td>
<td>25</td>
<td>Lat, Long</td>
</tr>
<tr>
<td>3</td>
<td>2679</td>
<td>235</td>
<td>15</td>
<td>Lat, Long</td>
</tr>
<tr>
<td>4</td>
<td>2679</td>
<td>200</td>
<td>50</td>
<td>Lat, Long</td>
</tr>
<tr>
<td>5</td>
<td>2679</td>
<td>150</td>
<td>100</td>
<td>Lat, Long</td>
</tr>
<tr>
<td>6</td>
<td>2679</td>
<td>235</td>
<td>15</td>
<td>Lat, Long, Age</td>
</tr>
</tbody>
</table>

**Cluster location**
- Partial dependence plots accurately signaled the latitude and longitude coordinates used to create the clusters.
- Partial dependence plots often missed to detect the disease cluster location in categorical covariates.
  - a) Accurate signal in latitude coordinates.
  - b) Accurate signal in longitude coordinates.
  - c) Missed signal in age group.
  - d) Accurate signal in age group.

**Future Work**
- To improve signal detection in categorical covariates.
- To account for changes in population.