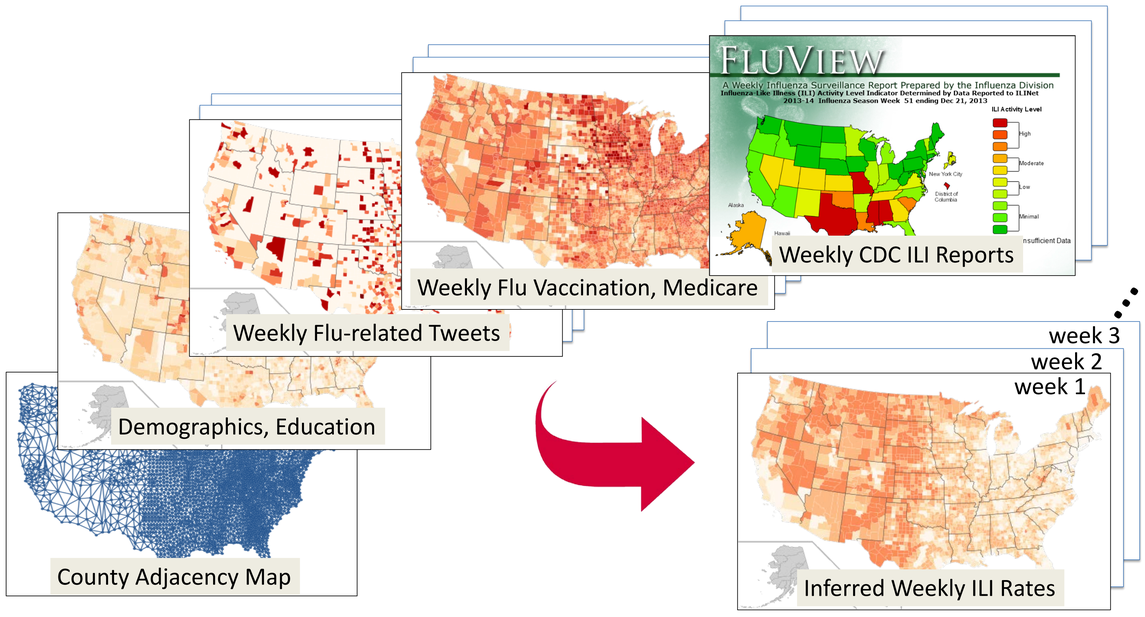
Inferring Spatio-temporal Spread of Flu Epidemics

Motivation

- Flu epidemics world wide cause 5 million cases of severe illness and 200,000-500,000 deaths annually. Predicting the spread of the epidemics through space and time can help government agencies and organizations better prepare and allocate resources.
- Influenza-like Illness (ILI) data aggregated by CDC has been valuable for researchers trying to develop models to forecast the spread of flu epidemics. Additional information such as demographics, vaccination rates and social network data also sheds light on the course of a flu epidemic. Considered together, they offer the opportunity to significantly improve our ability to assess and forecast flu epidemics both spatially and temporally.



Problem Specification

- **Phase 1** – **Reconstruct** weekly ILI rates at the county level by fusing multiple data sources. ILI data from HHS Regions (composed of multiple states) and some selected states, along with supporting social, geographic & demographic data, are available to infer the county-level ILI rates. The results will be aggregated and compared to a set of "Evaluation Regions" consisting of reported ILI rates from selected states and districts from 2 states.
- **Phase 2** – **Nowcast** weekly ILI rates at the county level by fusing multiple data sources. The ILI data from CDC are released after a delay of 1 to 2 weeks. The goal is to predict county ILI rates in week t using all data from previous weeks $t-1, t-2, \ldots$.

Problem Sets:
To facilitate development and testing of solutions, the problem and the corresponding data are divided into 3 levels of complexity – small, median and full – with increasingly larger geographical areas for reconstruction and nowcasting.

- Small – ILI rates for the state of Mississippi only
- Median – ILI rates for HHS Region 4, which includes 8 states.
- Full – ILI rates for all lower 48 states.

Evaluation Plan

Queries:

- **Reconstruction** – Given covariates such as weekly flu related tweets and cumulative percentage of Medicare flu vaccination claims of every county and weekly ILI rates of HHS regions and selected states/districts, output MAP and/or marginal MAP estimates for weekly ILI rates of individual counties.
- **Prediction** – Given covariates of every county for weeks $t = 1, \ldots, u$ and weekly ILI rates of HHS regions for weeks $t = 1, \ldots, u-1$, output MAP and/or marginal MAP estimates for weekly ILI rates of individual counties.

Metrics:

1. **Population-adjusted Squared Error**. Sum of squared errors compared to the actual district and state ILI rates where the truth data are available, weighted by the county populations.
2. **Start and Max of the epidemic**. The difference between the observed and predicted weeks when the ILI rate first exceeds 5% and that when the ILI rate reaches the peak for each region.

Available Data:

- Weekly ILI rates of HHS Regions, selected states and districts within some selected states
- Weekly counts of tweets from each county containing the keywords "flu" or "influenza"
- Weekly cumulative percentages of Medicare recipients filing flu vaccination claims for each county
- County demographics from US Census Bureau

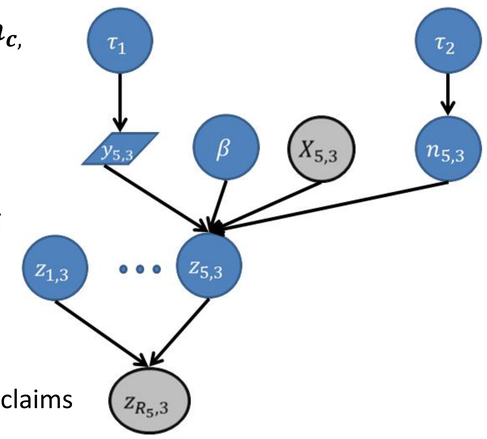
Reference Spatio-temporal Model

A Bayesian hierarchical spatio-temporal model and the corresponding Python code have been developed to describe both spatial clustering and temporal correlations of the ILI rates.

- Incorporate known covariates including flu-related tweets and vaccination in fixed-effects terms
- Capture latent spatio-temporal "propensity" with Gaussian Markov Random Field (GMRF)

Latent Flu Prevalence: $$\log\left(\frac{\boldsymbol{z_{c,t}} + \boldsymbol{\epsilon}_1}{1 - \boldsymbol{z_{c,t}} + \boldsymbol{\epsilon}_1}\right) = \boldsymbol{\beta}^T \boldsymbol{X_{c,t}} + \boldsymbol{y_{c,t}} + \boldsymbol{n_{c,}}$$

Aggregated Observations $$z_{R_i,t} = \sum_{c \in R_i} \left(\frac{N_c}{N_{R_i}}\right) z_{c,t}$$



c: county index; t: week index; R_i: set of counties in district/state i

$\boldsymbol{z_{c,t}}$: ILI rate (between 0 and 1) of county c in week t

$$\boldsymbol{X_{c,t}} = \left[\log\left(\frac{S_{c,t}+\epsilon_2}{\tilde{N}_c}\right),\quad \log\left(\frac{V_{c,t}+\epsilon_3}{1-V_{c,t}+\epsilon_3}\right)\right]^{\boldsymbol{T}}$$

$S_{c,t}$: number of flu-related tweets; $V_{c,t}$: cum. perc. of flu vacc. claims

$\boldsymbol{y_{c,t}}$: GMRF $p(Y) \propto \exp\left(-\frac{1}{2}\tau_1 Y^T (D_w - W)\, Y\right)$

W: $\begin{cases} w_{(c,t)(c,j)} = \rho & \text{where } j = t-1 \text{ or } t+1, \\ w_{(c,t)(i,t)} = 1 & \text{if } i \text{ is a neighboring county of } c, \\ w_{(c,t)(i,j)} = 0 & \text{otherwise,} \end{cases}$ $\quad (D_w)_{(c,t)(c,t)} = \Sigma_{(i,j)} w_{(c,t)(i,j)}$

$\boldsymbol{n_{c,t}}$: zero-mean Gaussian noise