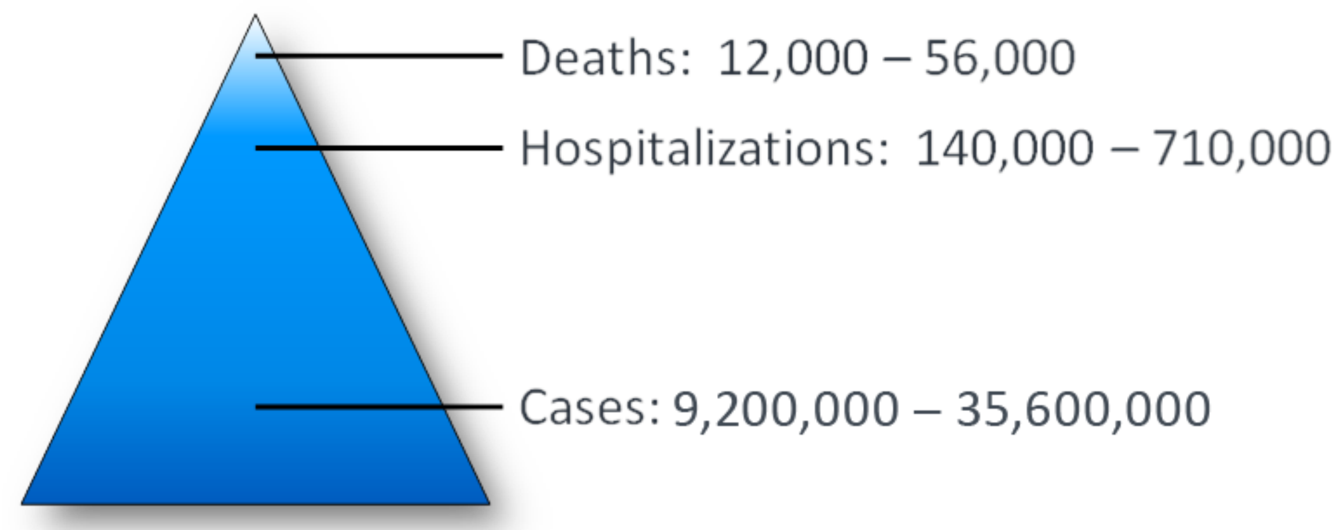
# Outline

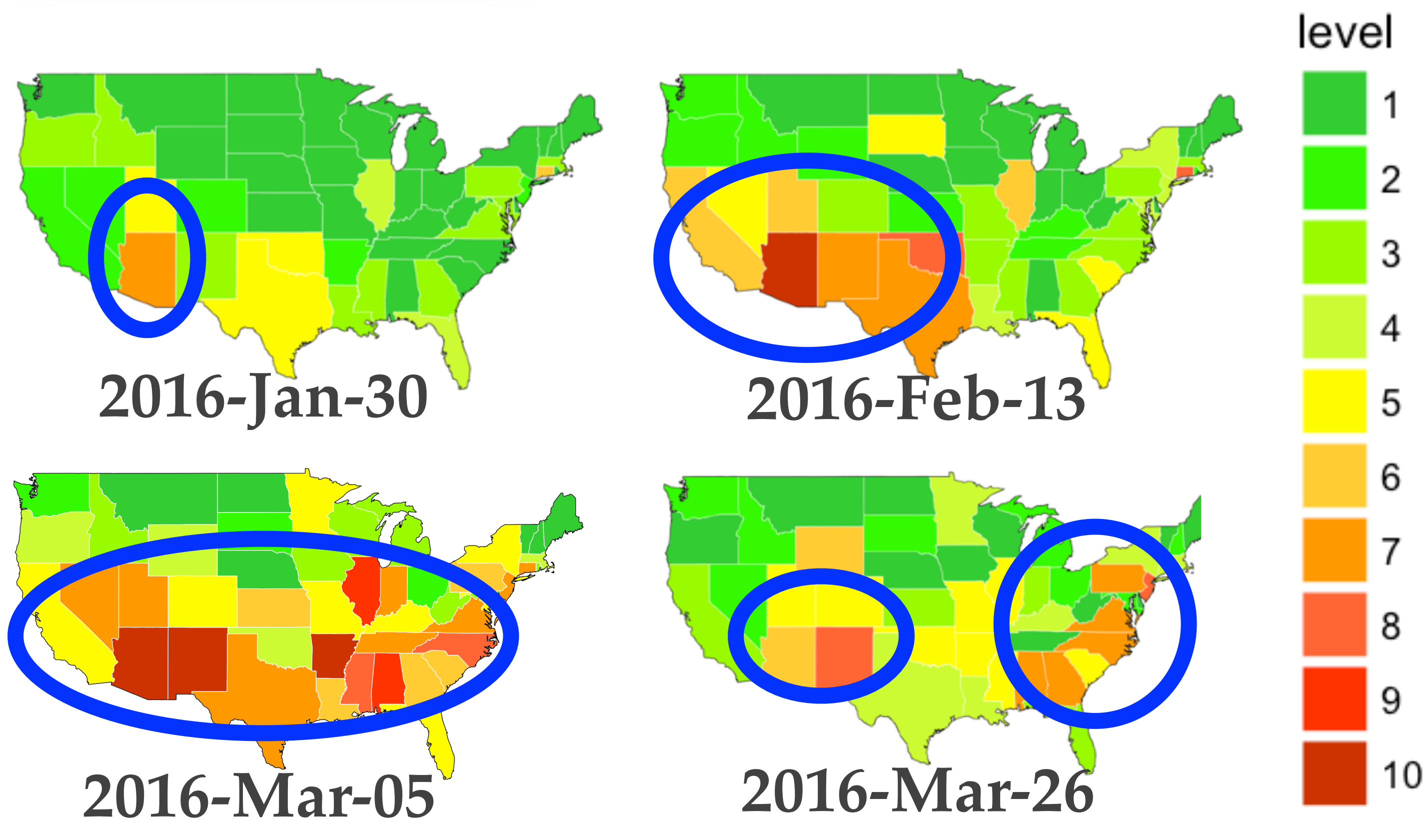
- Motivation
- Overview
- Model & Inference
- Experiments
- Conclusion
- Tasks and Schedule

Virginia Tech

# Motivation

- **Flu is deadly**
  - Population Loss
  - Economic Loss
- **Flu is predictable**
  - Spatial pattern
  - Temporal pattern



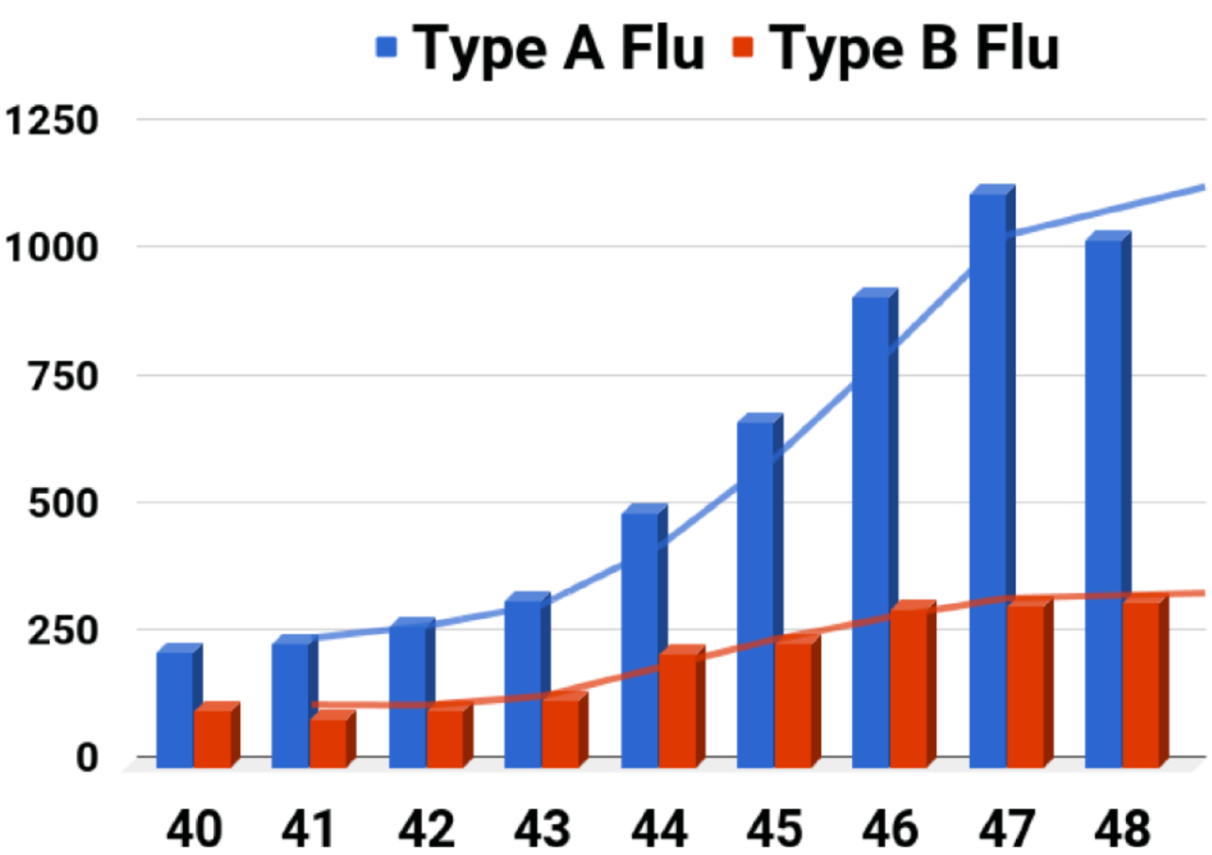

Virginia Tech

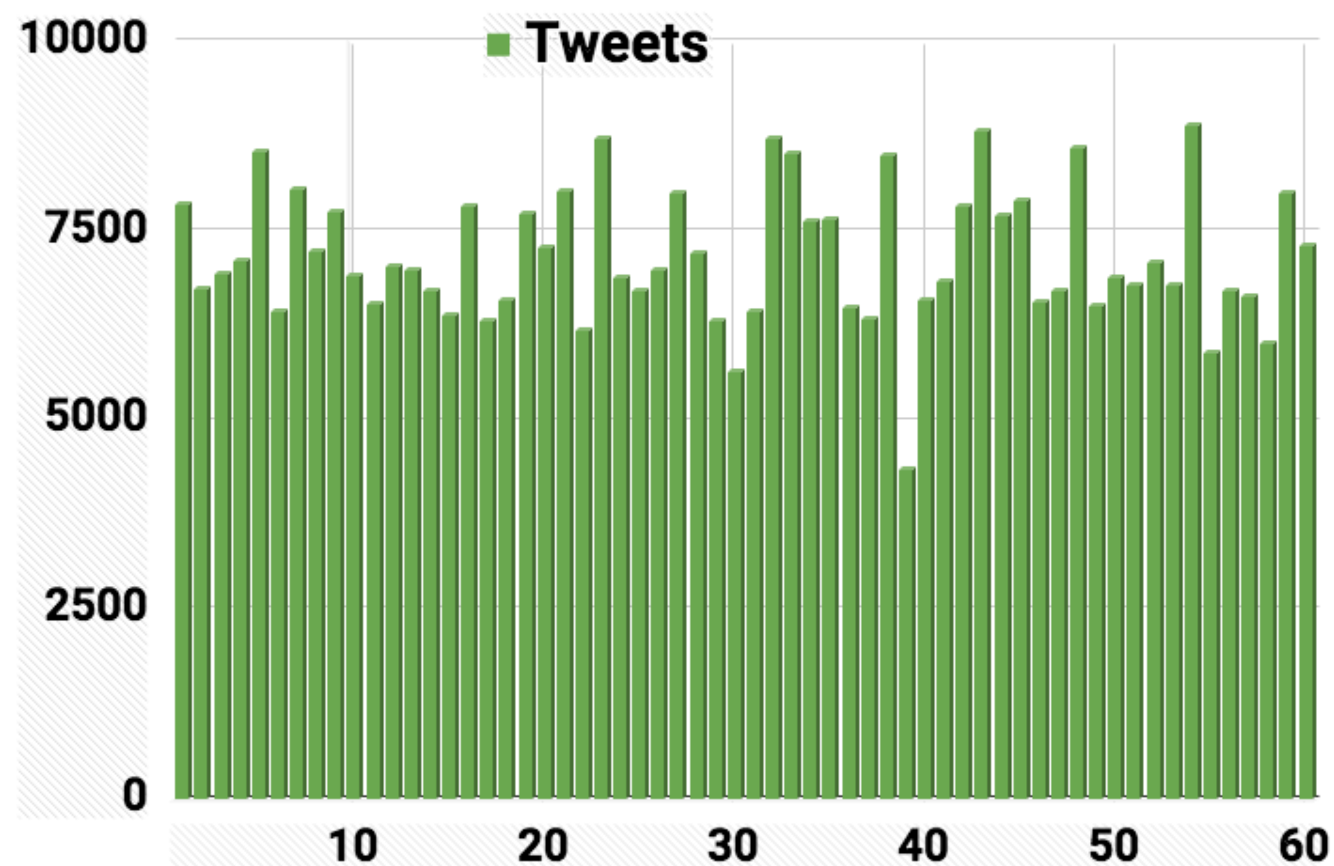
# Outline

- Motivation
- Overview
- Model & Inference
- Experiments
- Conclusion
- Tasks and Schedule

Virginia Tech

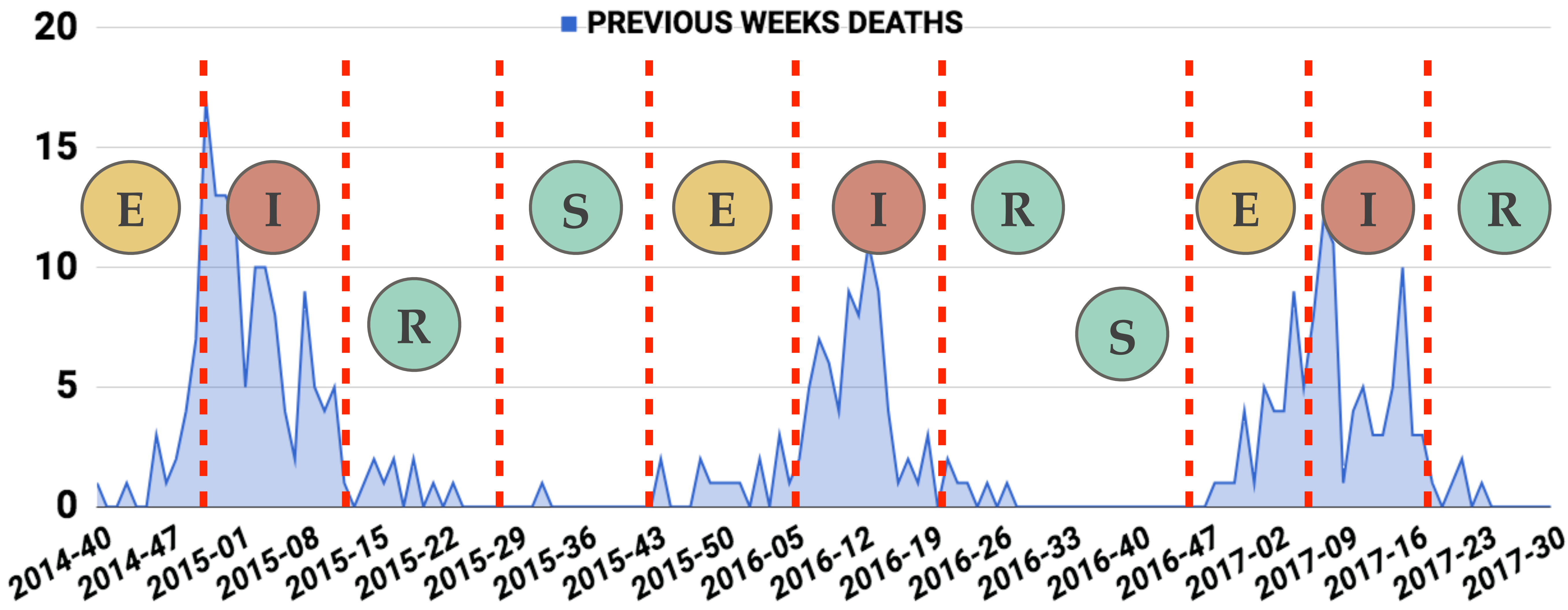
# Challenge 1: Accuracy in Short Time Prediction




- CDC data updated:
  - State-level
  - Weekly
- Twitter data:
  - Huge: 6,500 per second
  - Sparse: content related to flu

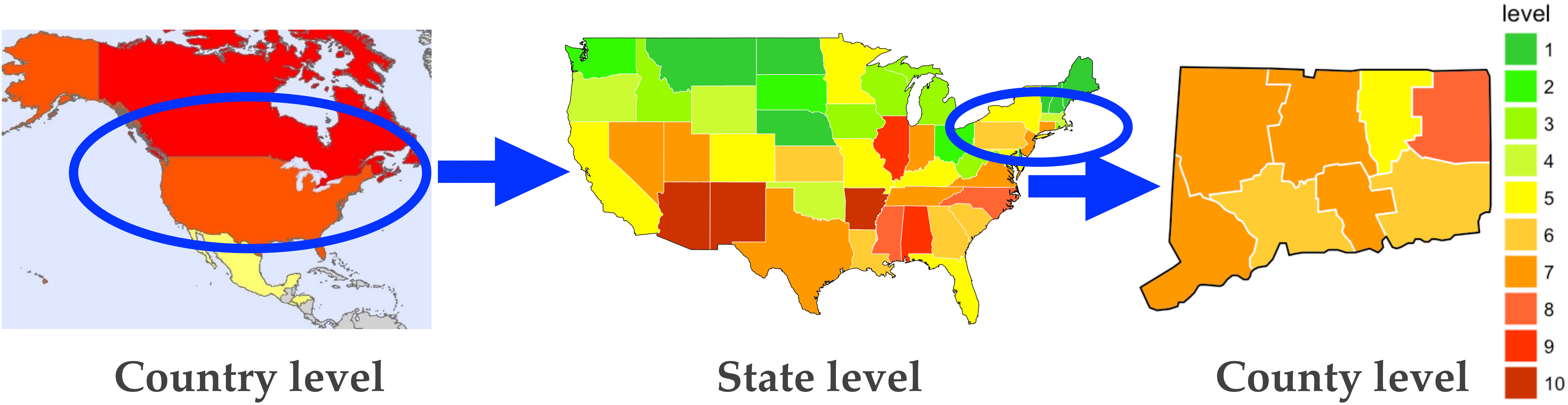
# Challenge 2: Stability in Long Time Prediction




- Flu cost price of lives
- Flu appears periodically
- Predict beforehand that can reflect trending

Virginia Tech

# Challenge 3: Flexible Geographical Levels




**Country level** **State level** **County level**

- Flu has obvious geography feature
  - Central
  - Distance
- Predict in different geographical level
  - Coarse-level
  - Find-level

Virginia Tech

# Idea Overview

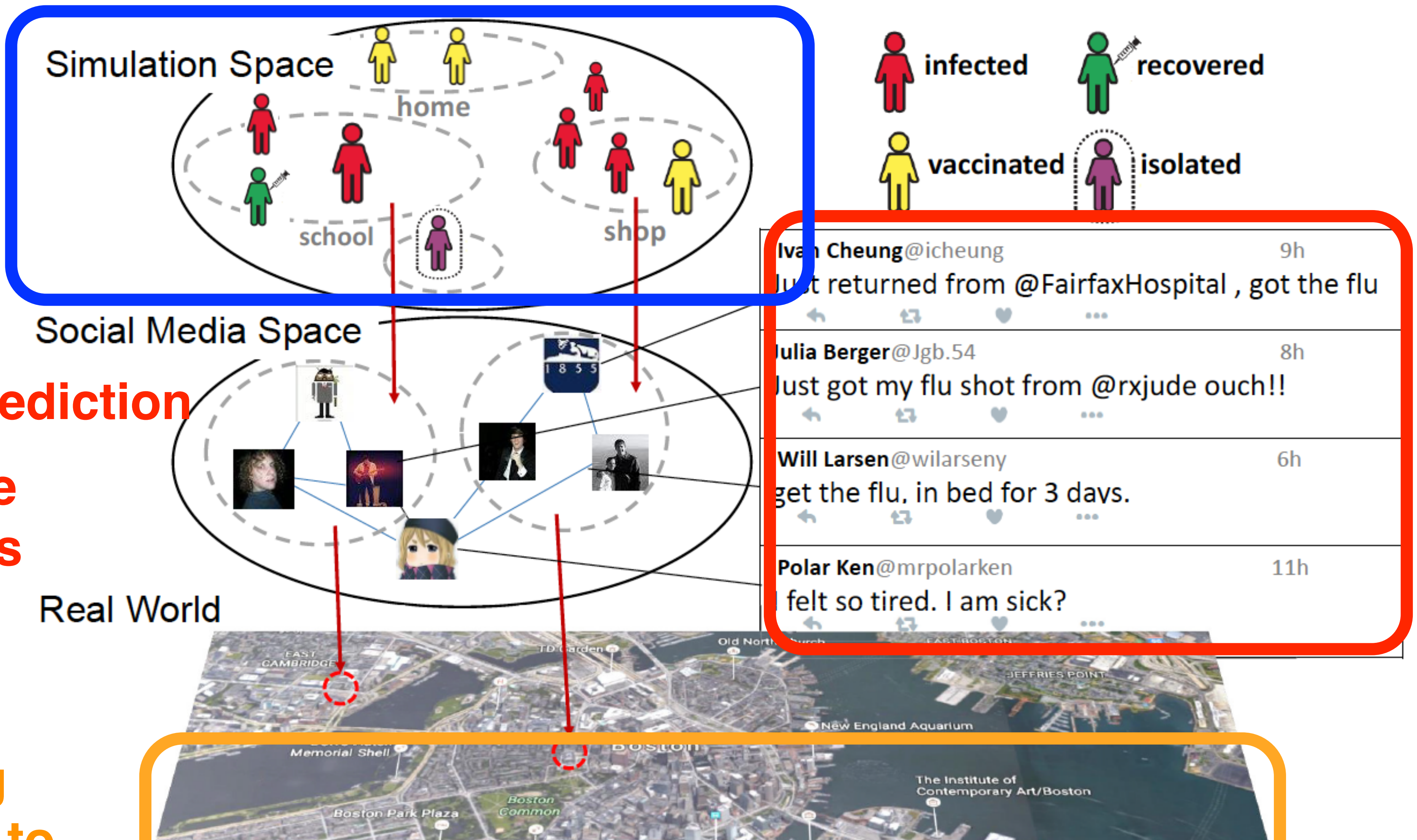
- Combine Computational models and data-driven methods

**Challenge2—Long time prediction**

1. Build contact network
2. Model flu via SEIR model

**Challenge1—Short time prediction**

1. Obtain data in real time
2. Change into parameters for simulation system




1. Geocoding
2. Aggregate to different levels

**Challenge3—Geographical Scalability**

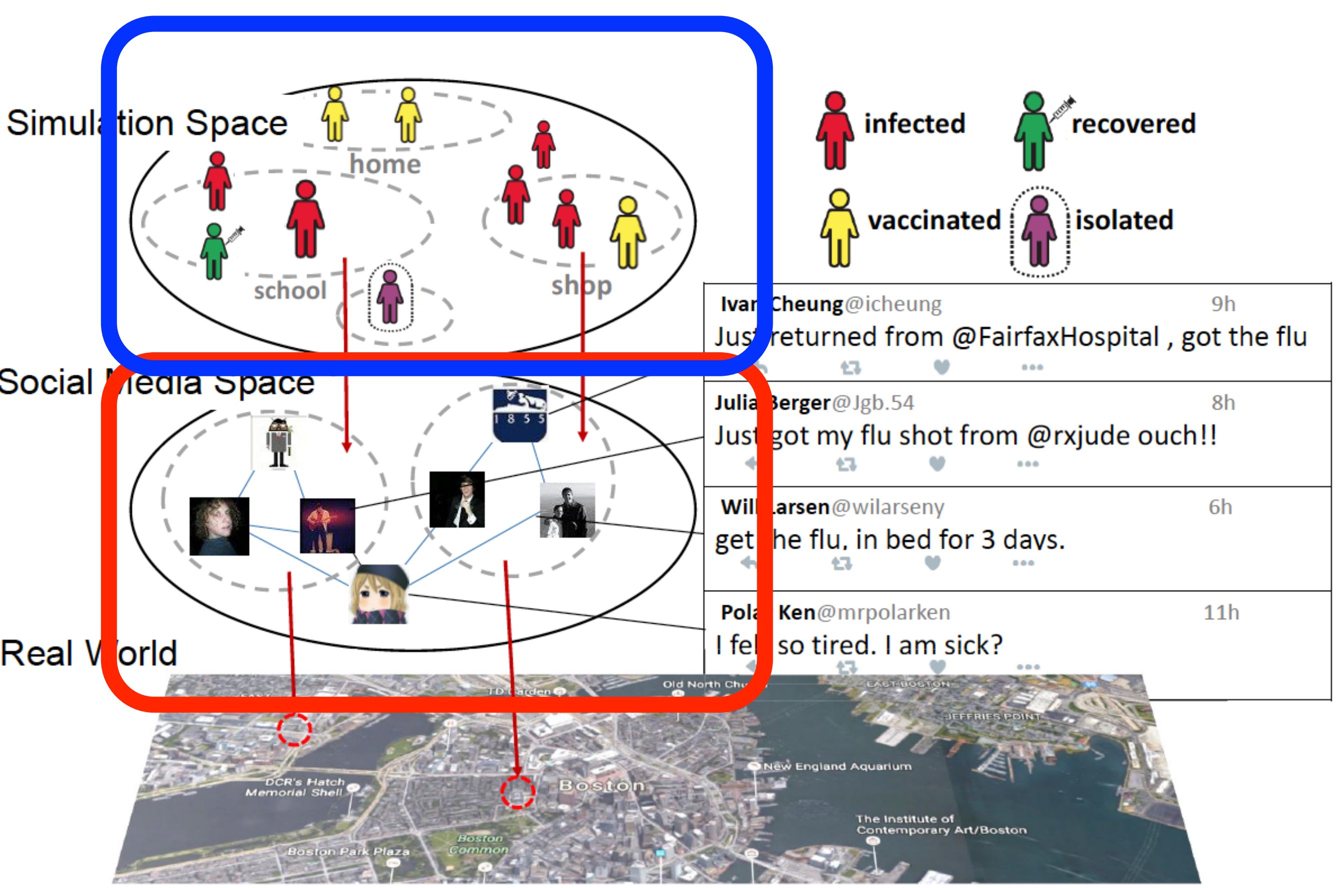
Virginia Tech

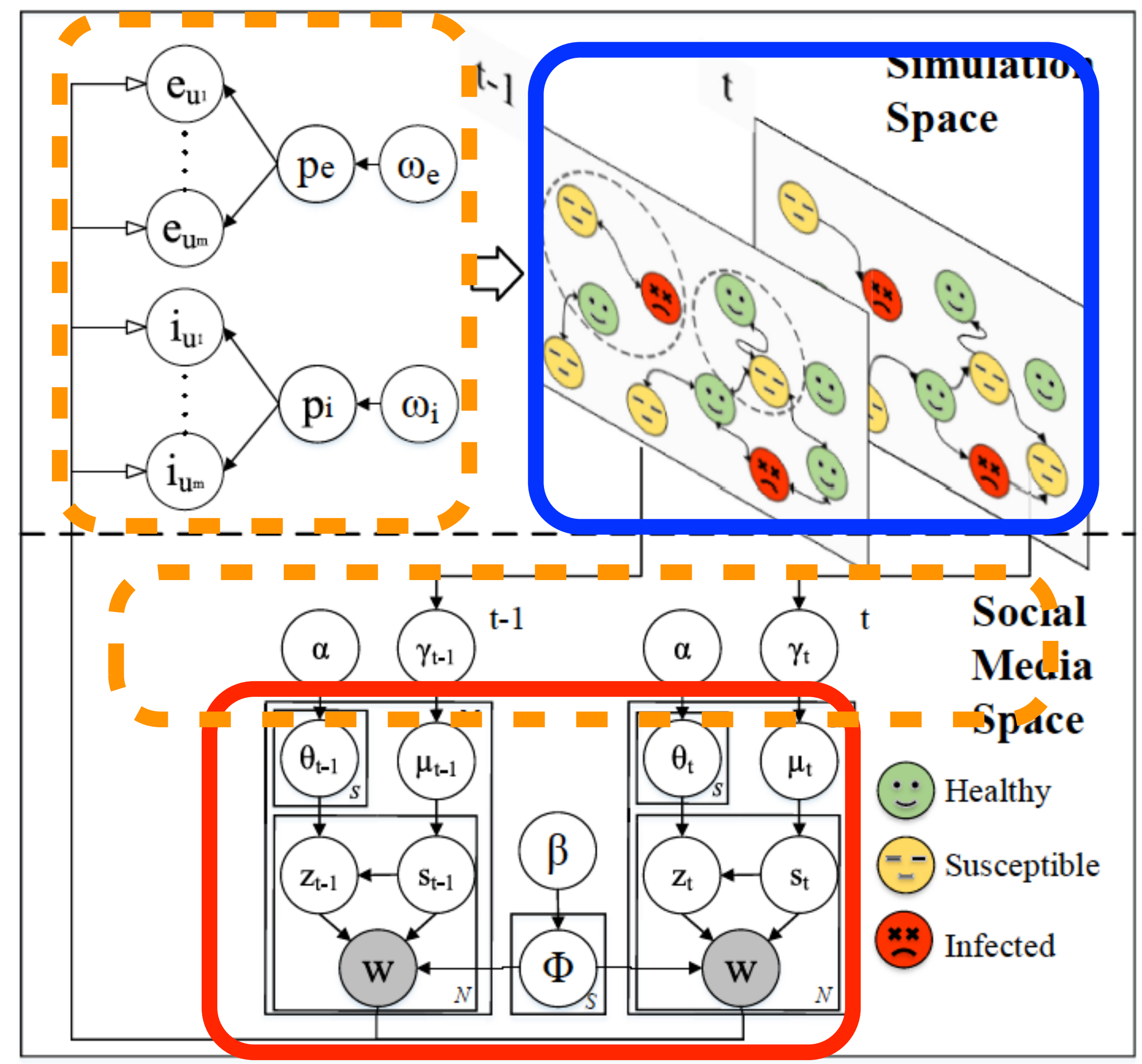
# Outline

- Motivation
- Overview
- Model & Inference
- Experiments
- Conclusion
- Tasks and Schedule

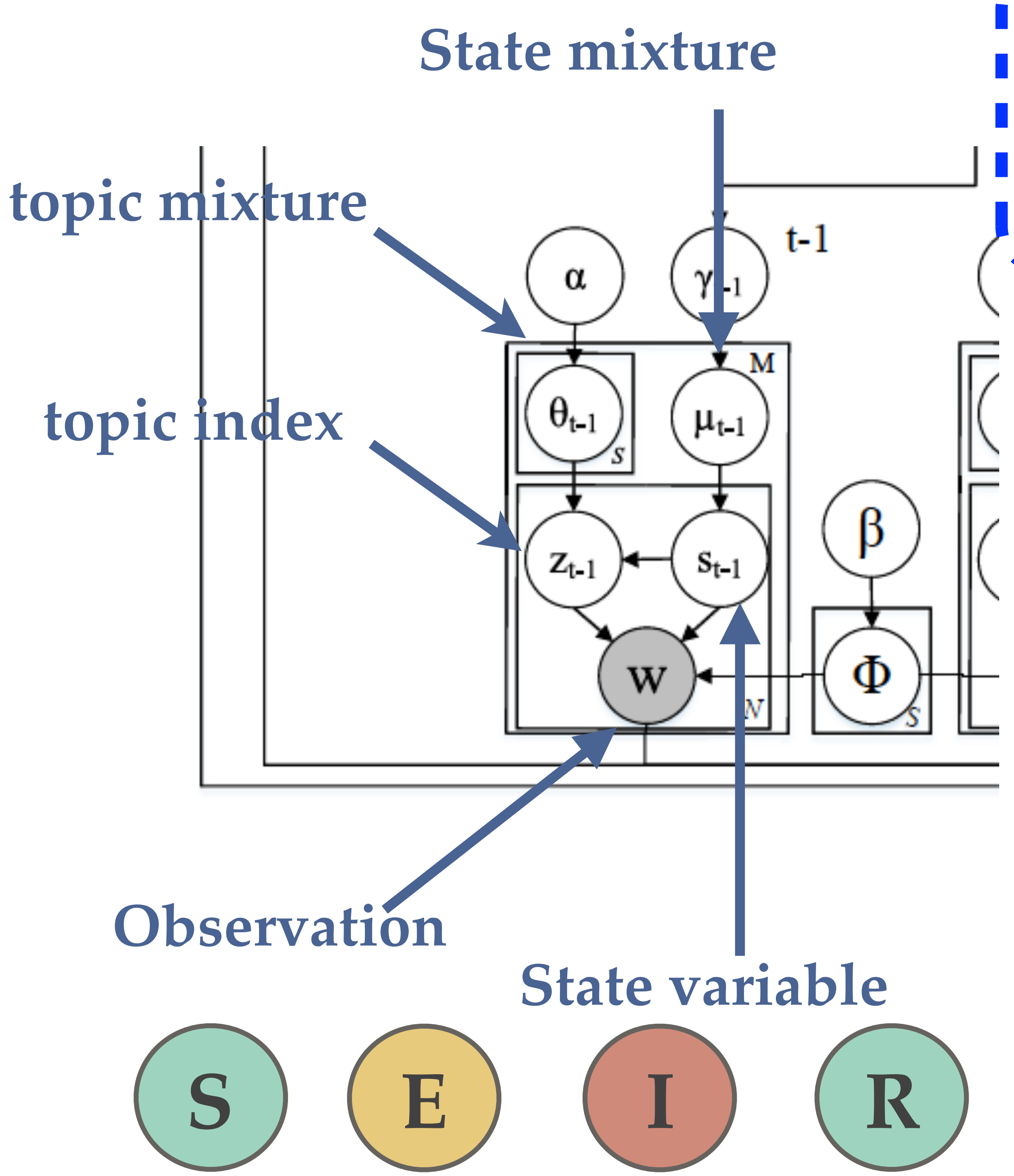
# Model

- Combine Computational models and data-driven methods



# Model: Social Media Part




**for** *each label* $s = 1, 2, \ldots, S$ **do**
  **for** *each topic* $z = 1, 2, \ldots, K$ **do**
    Draw $\phi_{s,z} \sim Dir(\beta)$;
  **end**
**end**

Global

**for** *each time stamp* $t = 1, 2, \ldots, T$ **do**
  **for** *each document* $D_{u,t} = 1, 2, \ldots, U$ **do**
    Draw $\mu_{u,t} \sim Dir(\gamma)$ ;
    **for** *each label* $s = 1, 2, \ldots, S$ **do**
      Draw $\theta_{u,t,s} \sim Dir(\alpha)$ ;
    **end**
    **for** *each word* $w$ *in document* $D_{u,t}$ **do**
      Draw $s \sim Multi(\mu_{u,t})$ ;
      Draw $z \sim Multi(\theta_{u,t,s})$ ;
      Draw $w \sim Multi(\phi_{s,z})$ ;
    **end**
  **end**
**end**

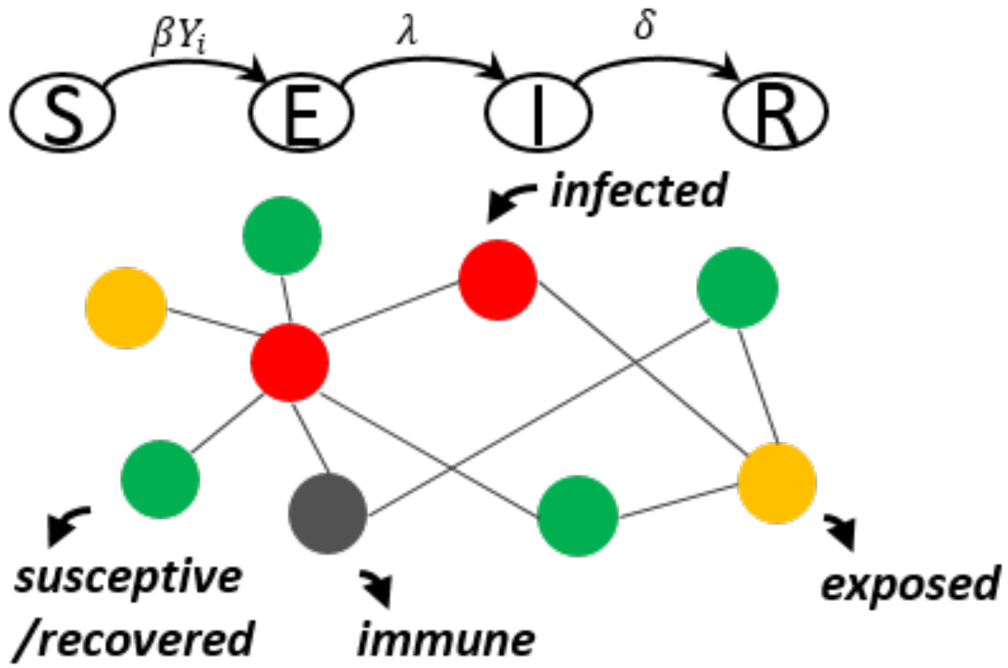
Time stamps

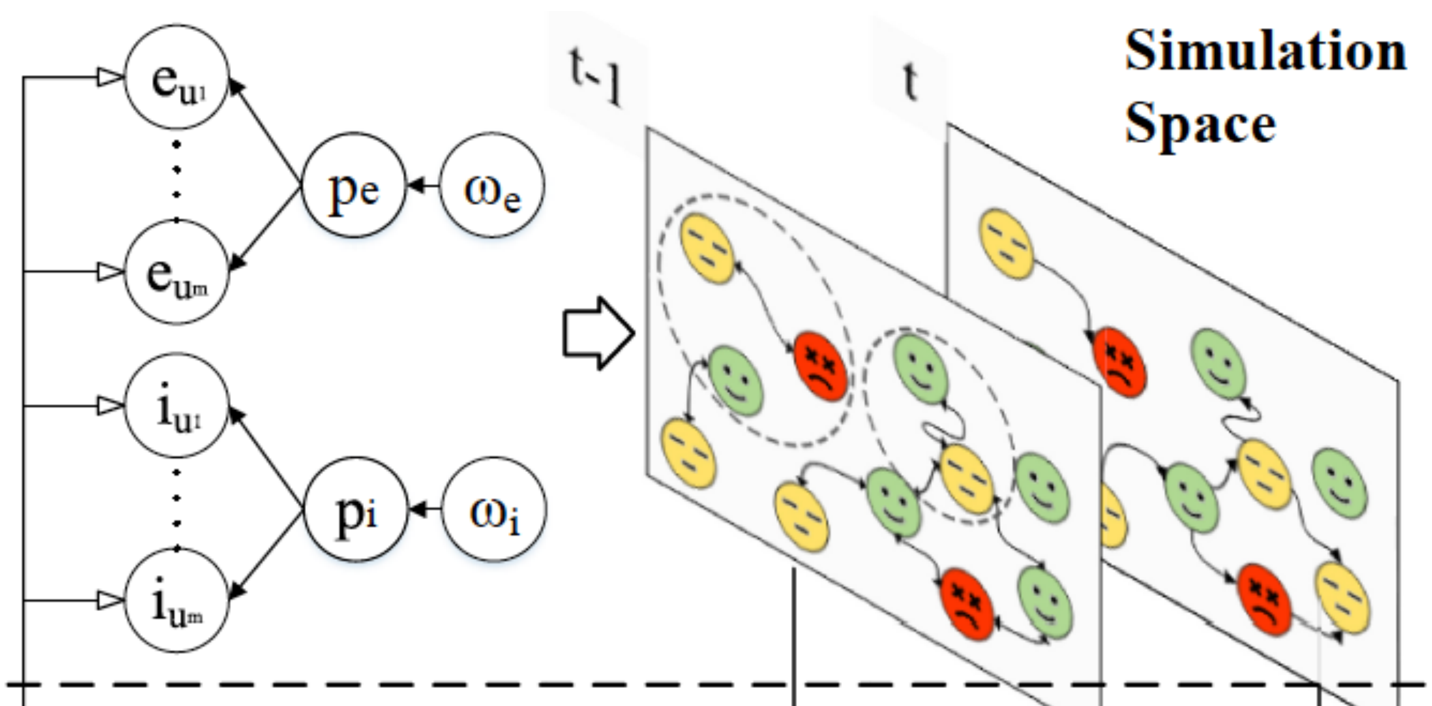
doc-level

word-level

Virginia Tech 1872

# Model: Simulation Part



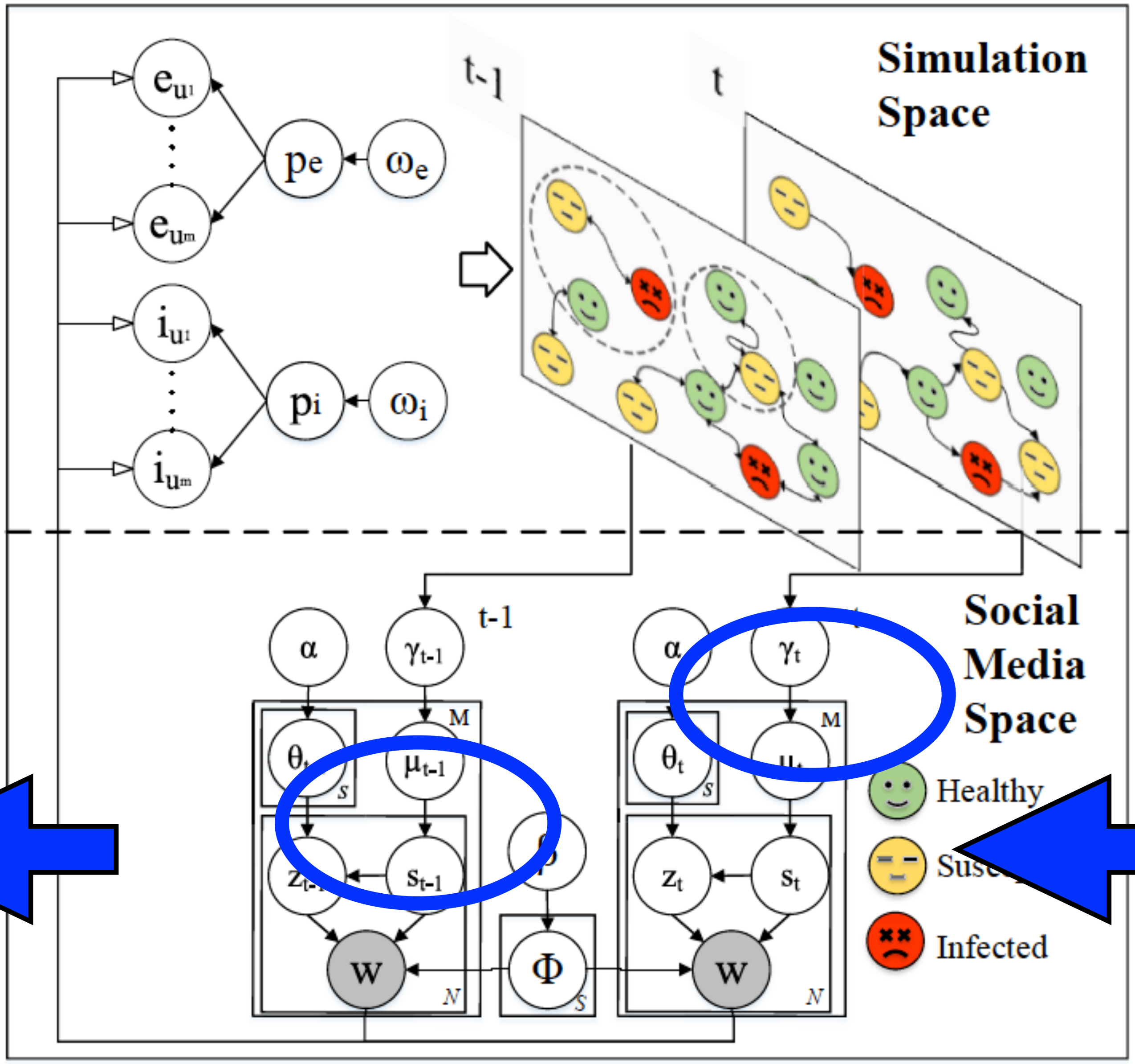




1. **Demographical features**
   - Each person has information such as age, gender, income …
2. **Disease features**
   - Infected rate
   - Recovered rate
3. **Contact network**
   - Detailed schedule for each person, such as daily activity
4. **SEIR model**
   - Each person is one of the 4 state in any time: susceptible, exposed, infectious, and removed.

# Model: Connection between Two Spaces

Parameters of simulation system
1. Incubation period
2. Infectious period
3. Transmissibility

Multinomial Posteriors



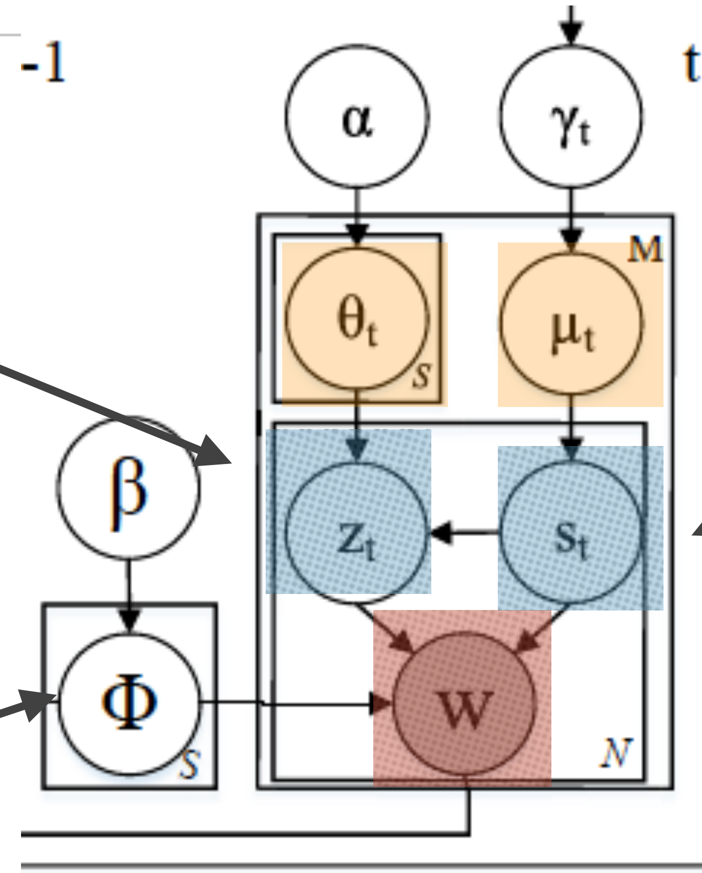

Simulation system Output
1. Person ID
2. Infectious duration
3. Incubation duration

Dirichlet Priors at next times tamp

# Inference: Joint Distribution

$$p(z_{mn} = k | \mathbf{w}, \mathbf{z}_{\neg i}, \mathbf{s}) = \frac{n^{v}_{sz_{\neg i}} + \beta}{\sum_{v=1}^{V} (n^{v}_{sz_{\neg i}} + \beta)} \frac{n^{z_{\neg i}}_{ms} + \alpha}{\sum_{z=1}^{K} (n^{z_{\neg i}}_{ms} + \alpha)}$$

$$\Phi_{szv} = \frac{n^{v}_{sz} + \beta}{\sum_{v=1}^{V} (n^{v}_{sz} + \beta)},$$




$$p(s_{mn} = s | \mathbf{w}, \mathbf{z}, \mathbf{s}_{\neg i}) = \frac{n^{v}_{s_{\neg i}z} + \beta}{\sum_{v=1}^{V} (n^{v}_{s_{\neg i}z} + \beta)} \frac{n^{z}_{ms_{\neg i}} + \alpha}{\sum_{z=1}^{K} (n^{z}_{ms_{\neg i}} + \alpha)} (n^{s_{\neg i}}_{m} + \gamma).$$

$$P(\mathbf{w}, \mathbf{z}, \mathbf{s} | \alpha, \gamma, \beta)$$

**Observation** $= \prod_{m=1}^{M} \prod_{n=1}^{N} p(w_{mn} | s_{mn}, z_{mn})$

**Multinational variable** $\prod_{m=1}^{M} \prod_{n=1}^{N} p(z_{mn} | \theta^{s_{mn}}_{m}) \prod_{m=1}^{M} \prod_{n=1}^{N} p(s_{mn} | \mu_m)$

**Multinational priors** $\prod_{m=1}^{M} p(\mu_m | \gamma) \prod_{s=1}^{S} \prod_{m=1}^{M} p(\theta^{s_{mn}}_{m} | \alpha) p(\gamma | \varphi, \sigma).$

# Model: Connection between Two Spaces

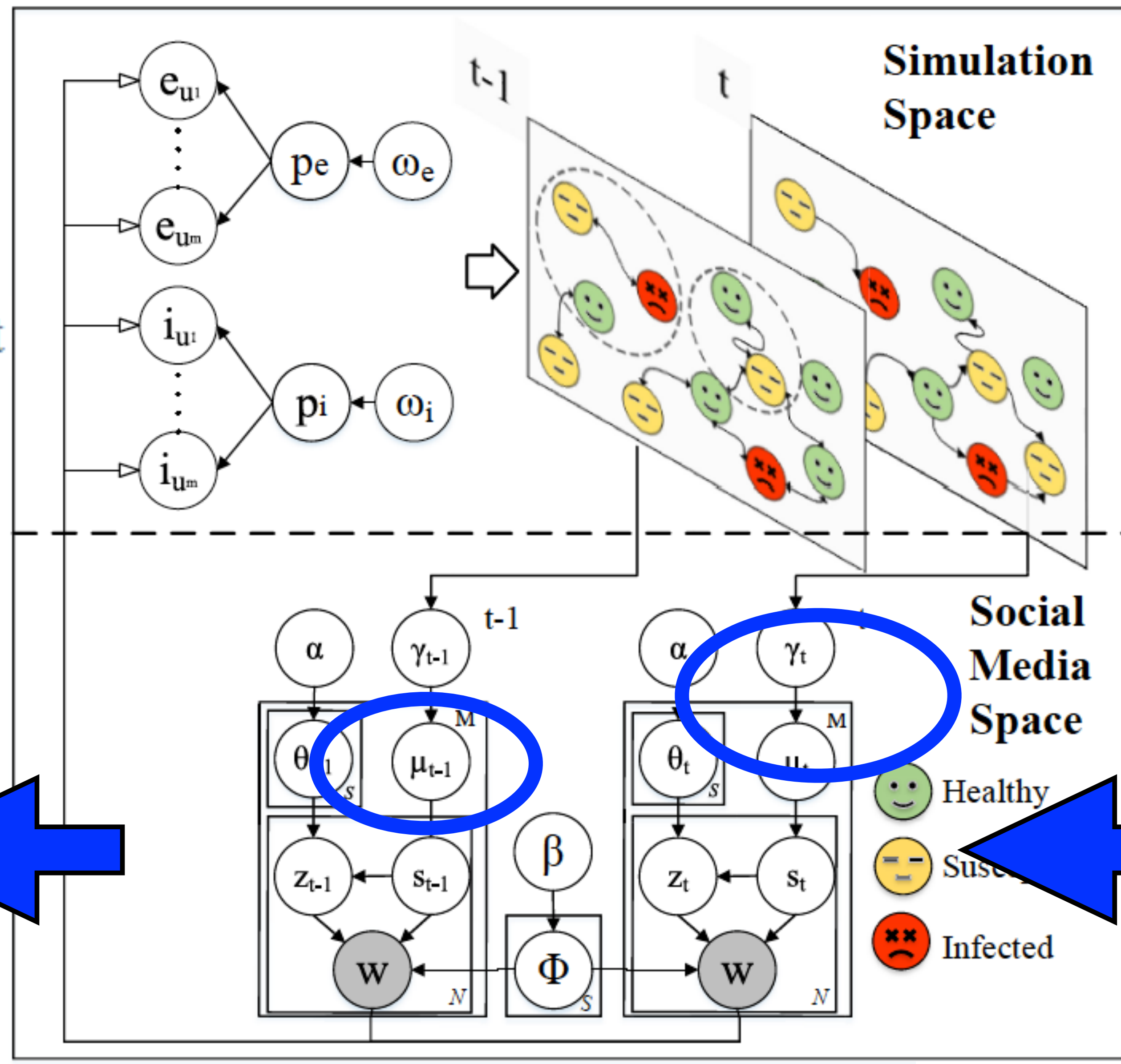
**Parameters of simulation system**

$S_{u,t} = (h_{u,t}, e_{u,t}, i_{u,t})$

the $s$-th ($s = 0, 1, 2$) element is 1, if and only if $argmax(\mu_{u,t})$ equals to $s$

**Multinomial Posteriors**

$$\mu_{ms} = \frac{n_m^s + \gamma}{\sum_{s=1}^{S} (n_m^s + \gamma)}.$$




**Simulation system Output**

$$\rho_{t,e} = \sum_{v=1}^{\mathcal{V}} E_{v,t}(\mathcal{G}, p_E, p_I, \tau)/\mathcal{V},$$

$$\rho_{t,i} = \sum_{v=1}^{\mathcal{V}} I_{v,t}(\mathcal{G}, p_E, p_I, \tau)/\mathcal{V},$$

**Dirichlet Priors at next times tamp**

$$p(\gamma|\varphi, \sigma) = \prod_s \frac{\sigma^{\sigma\varphi_s} \gamma_s^{\sigma\varphi_s - 1} \exp(-\sigma\gamma_s)}{\Gamma(\sigma\varphi_s)},$$

# Outline

- Motivation
- Overview
- Model & Inference
- Experiments
- Conclusion
- Tasks and Schedule

Virginia Tech

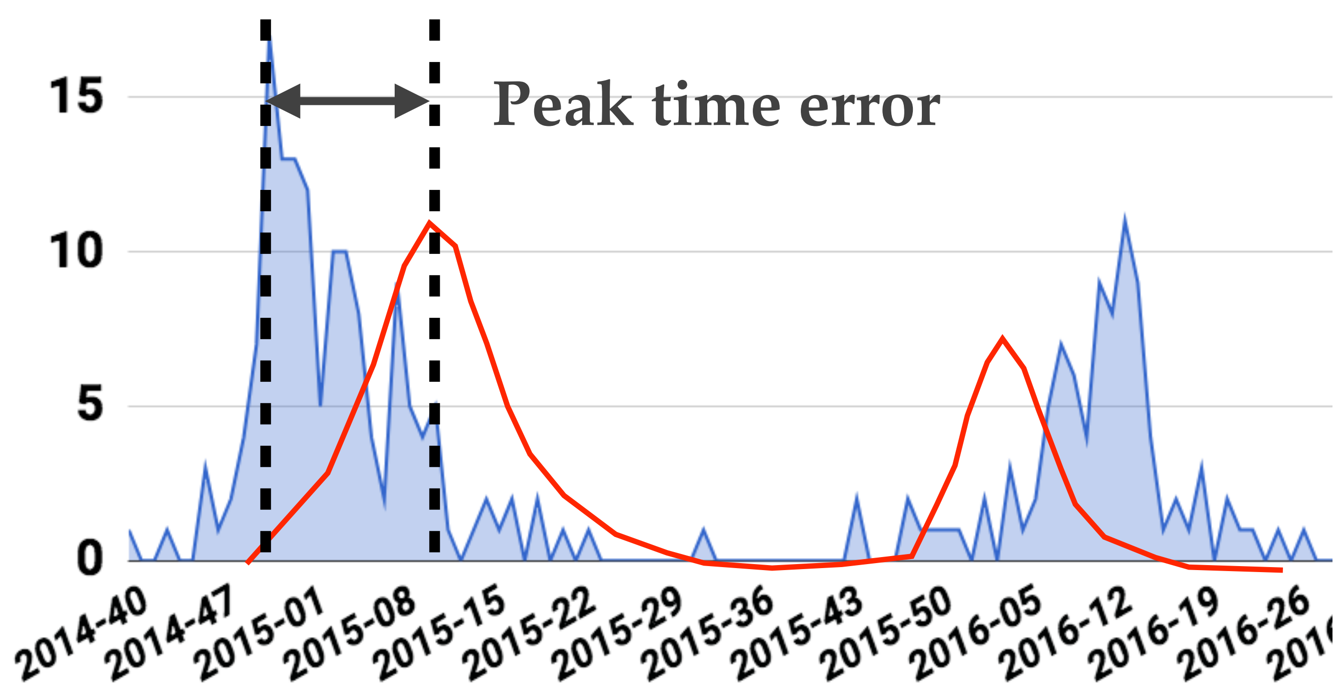
# Baselines

- **Computational epidemiology models**
  - SEIR
  - EpiFast [Bisset et al., ICS 2009]
  - Weakness: Relay too much on the CDC data
- **Social media mining methods**
  - LinARX [Achrekar et al., INFOCOM WKSHPS 2011]
  - LogARX [Achrekar et al., BIOSTEC 2012]
  - Weakness: No understanding of epidemic modeling

Virginia Tech

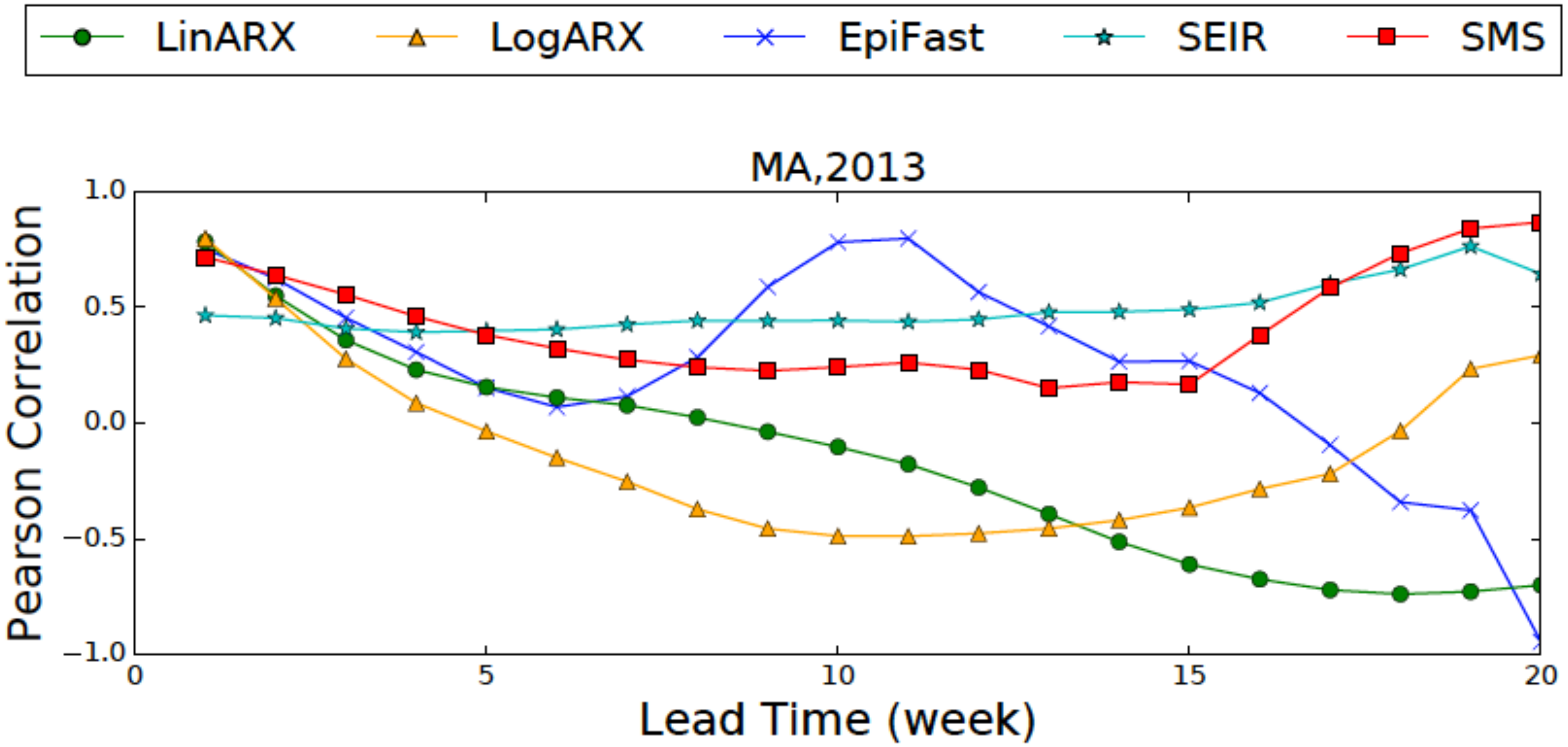
# Metrics

- **Pearson correlation**
  - Measure the relation between two variables (curves)
  - Range from -1 to 1
  - The bigger the better
- **Mean square error**
  - Measure the error between prediction results and ground truth
  - Range from 0 to 1
  - The smaller the better
- **Peak-time error**
  - Measure the error for the peak time
  - Average effort
  - The smaller the better



# Pearson correlation

Pearson Correlation of methods in MA 2013




Pearson Correlation values are in the range of (-1,1):

1. **Positive**: the trending of two variables are similar
2. **Zero**: no obvious relation observed
3. **Negative**: the trending of two variables are inversely

- **LinARX and LogARX**
  - Good in the initial
  - Drop quickly
  - End in negative
- **EpiFast**
  - Bad in the initial
  - End in the positive
  - But not stable
- **SEIR**
  - End in the positive
  - But not stable
- **SMS**
  - Best end performance
  - Maintain stable performance

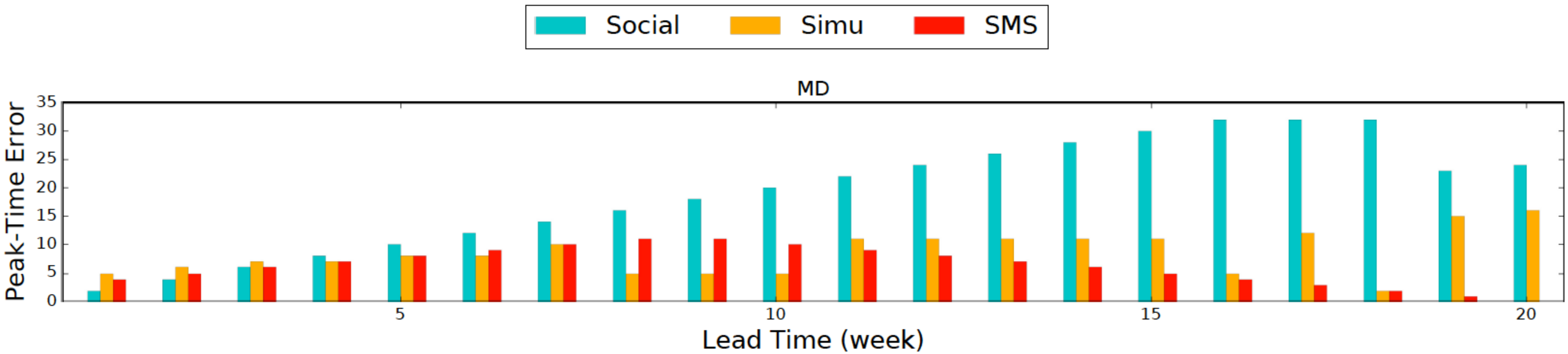
# Mean Square Error

| | 2013 MA | 2013 MD |
|---|---|---|
| LinARX | 6.65E-04 ± 3.74E-04 | 8.19E-04 ± 4.01E-04 |
| LogARX | 5.51E-04 ± 2.39E-04 | 5.00E-04 ± 1.79E-04 |
| EpiFast | 2.24E-03 ± 9.24E-04 | 5.14E-03 ± 5.57E-03 |
| SEIR | <u>3.73E-04</u> ± <u>5.38E-05</u> | <u>4.61E-04</u> ± <u>1.53E-04</u> |
| SMS | **2.38E-04** ± **6.16E-05** | **2.63E-04** ± **7.51E-05** |

Table 1: Performance in terms of mean square error in MA and MD states for 2013 data. The best performers are marked in bold, the corresponding second best performers are marked with underlines.

Virginia Tech

# Peak Time Error




- **Peak Time Comparison among SMS Vs Simu and Social variance methods**
  - Leas time is smaller than 5 weeks
    - Simu is the worst: no data for the fist week, worst in performance
  - Leas time is bigger than 5 weeks
    - Simu goes better and better
    - Social  go worse and worse

Virginia Tech

# Experiment Summary

- **SMS model yields the best results**
    - Good stability
    - Less errors
        - comparable to social media mining methods in the short time prediction
        - comparable to the computational methods in the long time prediction
    - Support inner-state report
- **Guideline for baselines**
    - Social media methods are good in short time prediction
    - Computational models are good in long time prediction

Virginia Tech

# Conclusion

- Motivation
- Overview
- Model & Inference
- Experiments
- Conclusion
- Tasks and Schedule

Virginia Tech

# Our Contributions

- Proposed  a novel SMS model  to combine the strength of social media mining models and computational methods within one framework.
- Provided an efficient inference for the propose SMS model.
- Demonstrated the effectiveness of SMS model through extensive experiments.
- Compared SMS model with other baselines and discussed their best application scenarios

Virginia Tech

# Thank You

**Questions or Comments?**