Data-driven Computational Epidemic Forecasting

Alexander Rodríguez
Harshavardhan Kamarthi
B. Aditya Prakash

College of Computing
Georgia Institute of Technology

December 1, 2021
AdityaLab @ Georgia Tech

• One of our lab’s focus: explore performance of data-driven methods in epidemiology/public health (surveillance, interventions, vaccination,...)
  • Data from multiple source is often more sensitive to what is happening ‘on the ground’
  • Complementary helpful perspective to other traditional methods
About us

• PI: B. Aditya Prakash
  • Assoc. Professor
  • PhD. CMU, 2012.
  • Data Mining, Applied ML
  • Networks and Sequences
  • Applications:
    • Epidemiology and Public Health
    • Urban Computing
    • The web
    • Security

• Homepage: https://www.cc.gatech.edu/~badityap/
About us

• Alexander Rodríguez
  • 4th year PhD student, graduating May 2023
  • Data science/ML in time series and networks
  • Motivated by impactful problems
    • Critical infrastructure networks
    • Epidemic forecasting
  • PhD thesis topic: ML for epidemic forecasting
  • Homepage: https://www.cc.gatech.edu/~acastillo41/
About us

• Harshavardhan Kamarthi
  • 2\textsuperscript{nd} year PhD student
  • Research Interests
    • Epidemic forecasting
    • Probabilistic forecasting and uncertainty quantification
    • Deep Probabilistic models
  • Homepage: https://harsha-pk.com/
Workshop Webpage

Workshop on Data-driven Computational Epidemic Forecasting for West African Countries

We have been invited by the Forecasting for Social Good (F4SG) Research Network to lead an online workshop on epidemic forecasting. The target audiences are researchers and practitioners from West African Countries, but anyone is welcome until we reach the capacity.

Abstract
Our vulnerability to emerging infectious diseases has been illustrated with the devastating impact of the COVID-19 pandemic. Forecasting epidemic trajectories (such as future incidence over the next four weeks) gives policymakers a valuable input for designing effective healthcare policies and optimizing supply chain decisions; however, this is a non-trivial task with multiple open questions. In this workshop, we will go...

- [https://adityalab.cc.gatech.edu/workshops/21-forecasting-f4sg.html](https://adityalab.cc.gatech.edu/workshops/21-forecasting-f4sg.html) or [b.gatech.edu/3cBPfQ7](b.gatech.edu/3cBPfQ7)

- All Slides will be posted there. Talk video as well (later).

- **License**: for education and research, you are welcome to use parts of this presentation, for free, with standard academic attribution. For-profit usage requires written permission by the authors.
Outline

1. Epidemic forecasting (30 min)
2. Mechanistic models (1 hrs)
3. Statistical models (1.5 hrs)
4. Hybrid models (20 min)
5. Ensembles (10 min)
6. Epidemic forecasting in practice (30 min)

• 15 min breaks after Part 2 and Part 3
  • We’ll be available for questions
Plan for the Workshop

• Theory and research
  • Setting up the epidemic forecasting problem
  • General epidemiology: key concepts and models
  • Statistical modeling and deep learning
    • Research innovations

• Practice
  • US real-time forecasting experiences
  • Coding examples
    • Mechanistic models
    • Statistical models
  • Demo session
    • Statistical correction of forecasts

Workshop focus:
  • Computational data-driven methods
  • Short-term forecasting (up to 4 weeks ahead)
Forecasting Infectious Diseases

• Why? Allocate resources/budget, inform public policy, improve preparedness

• Background:
  • Traditional methods are based on ODEs and agent-based models
  • Data collection has increased
    • Methods have difficulties ingesting these data sources
Real-time Epidemic Forecasting

Possible near future:
- Goes down
- Stays still
- Goes up

Depends on:
- Current number of infections
- Interventions in place
- Contact patterns
- Exposure to disease

Oklahoma Incidence Mortality

Incidence death counts

epidemic week

Ground truth data from JHU
Why Computational Data-driven Forecasting?

• Epidemic spread is a spatiotemporal phenomena over multi-scale networks

• New end-to-end methods available capable of modeling data with minimal assumptions

• Before and after the COVID-19 pandemic: Explored **performance** and **utility** of data-driven models in short-term forecasting
Our Participation on CDC Forecasting Initiatives

**Target 1:** Influenza like illness per week

**Target 2:** Weekly Covid Mortality

**Target 3:** Daily Covid Hospitalizations

Since April End 2020

Last few years Also in COVID-ILI (March 2020)
Our Impact

Only individual Deep Learning model in top-5 accuracy in the CDC-led evaluation for 1+ year

FiveThirtyEight
1 of 11 shown on their page

1st Prize
The COVID-19 Symptom Data Challenge

facebook
Carnegie Mellon University

2nd Prize
C3.ai COVID-19 Grand Challenge

Out of 115 global participants

43 Countries
777 Participants
COVID Response

Visualizing impact of nonpharmaceutical interventions

On-campus Mobility and Data-driven Interventions

Adaptive surveillance

Hospital Acquired Infections

... and others like vaccine allocation algorithms etc.

Rodriguez, Kamarthi, and Prakash 2021

https://www.cc.gatech.edu/~badityap/covid.html
Recent Publications


Coming up soon

• Survey paper on Data-driven Computational Epidemic Forecasting.
  • Workshop material based on this survey
• Preprint soon in medRxiv.
• Link will be posted in workshop website.
Part 1: Epidemic Forecasting
Epidemic Forecasting Pipeline

A. Data Processing

Raw data

Processing: delays, anomalies, revisions
Exploratory analysis

Input Data

Behavioral
Environmental
Digital
CDC Clinical
Genomics
Nextstrain

Epi-indicators
Real-valued
Event-based

Targets

Sample 1
Sample N

B. Model Training & Validation

Validation and Model Selection

Multiscale dynamics
Uncertainty quantification
Feature engineering and selection
Interpretability
Robustness to noisy data
Scenario selection

Model

Neural models
Mechanistic models
Hybrid models
Ensembles

Training

Log Score
MAE, WIS

Hyper Param Tuning

C. Utilization & Decision Making

Real-Time Forecasting

Dashboards
CDC Initiatives

Ensemble of Real-Time Predictions

Resource Allocation
Risk Assessment
Forecast Communication

Decision Making

Feedback

Rodríguez, Kamarthi, and Prakash 2021
Epidemic Forecasting Setting

1. Forecasting Tasks
2. Targets of interest
3. Spatial and temporal scales
4. Datasets
5. Model evaluation
[1] Common Forecasting Tasks

- Used in annual CDC Flu forecasting challenge

Short-term forecasting: Up to 4 weeks ahead
[2] Targets of Interest

- Influenza
  - %ILI: symptomatic outpatients
    - Syndromic surveillance
  - Lab-tested hospitalizations
- COVID-19
  - Mortality
  - Hospitalizations
  - Cases
[3] Spatial and Temporal Scales

- **Spatial scales:**
  - National
  - Region/state/province
  - County/city (less common)

- **Temporal scales:**
  - Weekly
  - Daily

---

Rodríguez, Kamarthi, and Prakash 2021
Datasets: surveillance pyramid
Line-list data

• Who, when and where a person was infected

Hospital records

Lab surveys

Population surveys
Digital Epidemiology

Marcel Salathé¹,²*, Linus Bengtsson³, Todd J. Bodnar¹,², Devon D. Brewer⁴, John S. Brownstein⁵, Caroline Buckee⁶, Ellsworth M. Campbell¹,², Ciro Cattuto⁷, Shashank Khandelwal¹,², Patricia L. Mabry⁸, Alessandro Vespignani⁹

¹ Center for Infectious Disease Dynamics, Penn State University, University Park, Pennsylvania, United States of America, ² Department of Biology, Penn State University, University Park, Pennsylvania, United States of America, ³ Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden, ⁴ Interdisciplinary Scientific Research, Seattle, Washington, United States of America, ⁵ Harvard Medical School and Children’s Hospital Informatics Program, Boston, Massachusetts, United States of America, ⁶ Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, United States of America, ⁷ Institute for Scientific Interchange (ISI) Foundation, Torino, Italy, ⁸ Office of Behavioral and Social Sciences Research, NIH, Bethesda, Maryland, United States of America, ⁹ College of Computer and Information Sciences and Bouvé College of Health Sciences, Northeastern University, Boston, Massachusetts, United States of America
Surveillance pyramid and datasets

Final patient outcome
- Johns Hopkins
- The COVID Tracking Project

Intensive Care
- HealthData.gov

Hospitalized
- Health Services Data
- OpenSafely

Sought Healthcare
- Community-based/syndromic surveillance
- Sero-Epidemiology

Symptomatic/sick
- Community Surveys

Infected (some asymptomatic)
- General Population

Waste water
- Behavioral surveys
- Policy
- Demographics

Search engines
- Google
- flu near you
- kinsa.

Social media
- Twitter

Digital medical and wellness devices
- RS Metrics
- OpenTable

Satellite images

Pharmaceutical orders

Mobility and exposure

Contact tracing

Retail and commerce data

American Community Survey
- U.S. Census Bureau

CDC

NHS

Nextstrain

Georgia Tech

Rodríguez, Kamarthi, and Prakash 2021
Search Engines and Social Media

- Search activity
  - Ad-hoc search engines
  - Specialized search engines

- Social media
  - Tweets
  - RSS feed
Online Surveys

• Symptomatic surveys

• Behavioral surveys
  • Adoption of public health recommendations
  • Mask wearing
  • Social distance
Mobility

• Quantify contact patterns within and across communities

• Sources:
  • Mobile call records
  • Mobile apps
Satellite Images

[NEWS EXCLUSIVE] ZHONGNAN HOSPITAL OF WUHAN UNIVERSITY

10/10/18 | 506 CARS
10/17/19 | 640 CARS

[Brownstein+ 2020]

- **Point Forecasts**: Single value per forecast
- **Probabilistic Forecasts**: Probability distribution of forecast
  - Captures uncertainty, useful for decision making
Evaluation of Point Forecasts

- **RMSE:** $\sqrt{\frac{\sum_{i=1}^{T}(y_i - \tilde{y}_i)^2}{T}}$
- **MAE:** $\frac{\sum_{i=1}^{T}|y_i - \tilde{y}_i|}{T}$
- **MAPE:** $\sum_{i=1}^{T} \frac{|y_i - \tilde{y}_i|}{|y_i|}$
- **Others:** WAPE, NMSE
Evaluation of Probabilistic Forecasts

• Log Score: \[ \frac{1}{T} \sum_{i=1}^{T} \ln(p_i(y_i)) \]
  • Log probability of ground truth outcome (binned)

• Other metrics
  • Coverage score
  • Interval score & Weighted Interval Score (WIS) 
    [Bracher+ 2021]

\[
IS_{\alpha}(F, y) = (u - l) + \frac{2}{\alpha} (l - y) \mathbb{1}(y < l) + \frac{2}{\alpha} (y - u) \mathbb{1}(y > u)
\]

\[
WIS_{\alpha_{\{0,K\}}}(F, y) = \frac{1}{K + 1/2} \times |y - m| + \sum_{k+1} \{w_k \times IS_{\alpha_k}(F, y)\}
\]
How to choose eval. metrics?

- Based on decision making
  - Uncertainty and calibration are important
  - Probabilistic evaluation metrics are more desirable
- Log score for influenza
  - %ILI are within some bounds
- WIS for COVID-19
  - Unbounded values for mortality, cases, hosp
Part 2: Mechanistic Models
Mechanistic models

• Intuition:
  • People move from compartments based on the disease progression
  • Differential equations describe movement

• Modeling approaches:
  1. Mass-action models (ODE models)
  2. Metapopulation models
  3. Agent-based networked models
[1] ODE Models: SIR

- One of the most simplest models
  - Susceptible: healthy, can get infected
  - Infected: can infect others through contact
  - Recovered: can not infect others
Assumptions

• Perfect mixing
  • Any infected person can infect any susceptible person

• No birth or deaths (no ‘demography’)
  • Total population is constant

• Deterministic!
SIR Model

\[
\begin{align*}
\frac{dS}{dt} &= -\beta SI \\
\frac{dI}{dt} &= \beta SI - \delta I \\
\frac{dR}{dt} &= \delta I
\end{align*}
\]

Number of new infections = \( \beta \) * # infection attempts

Number of infected nodes curing
Solving SIR

• No closed form solution!
SIR: numerical output

![SIR numerical output graph](image_url)

- **S(t)**: Susceptible
- **I(t)**: Infected
- **R(t)**: Recovered

Number of nodes vs. time
Online interactive example

Data-Driven Computational Epidemic Forecasting / R-Notebook

```r
---
title: "SIR"
output: html_document
---

```r
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
```n
```{r}
### Credits

This code has been adapted from [http://epirecip.es/epicookbook/chapters/sir/r_desolve](http://epirecip.es/epicookbook/chapters/sir/r_desolve)

Author: Simon Frost

---

```{r}
library(deSolve)
library(reshape2)

sir_code <- function(times, init, parms){
  with(as.list(c(parms,init)), {
    # ODEs
    dS <- -beta*S*I
    dI <- beta*S*I-gamma*I
    dR <- gamma*I
  })
}

SIR modelos
```
Many many extensions

- With birth/death rates (‘vital dynamics’)
- Variable contact rates
- Age-structured models
- Make things stochastic
- Multiple viruses/diseases
- .......... 
- .......... 
SIR: implicit solution

\[ S(t) = S(0)e^{-R_0(R(t) - R(0))} \]

\[ R_\infty = 1 - S(0)e^{-R_0(R_\infty - R(0))} \]

\[ R_0 = N\beta/\delta \]

Reproductive Number
Threshold Phenomenon: $R_0$

\[
\frac{dI}{dt} = \beta SI - \delta I = I(\beta S - \delta)
\]

- This implies

\[
\frac{dI}{dt} < 0 \text{ if } S(0) < \frac{\delta}{\beta}
\]

- So $R_0 = \frac{\beta}{\delta}$

  - Basic Reproductive number: average number of secondary cases caused by one individual
Threshold Phenomenon

- If $S(0) < \delta / \beta = 1 / R_0$
  - Epidemic dies out
  - Large epidemic if and only if $R_0 > 1$
  - Hence estimating $R_0$ very important!
    - Why?
      - Immunization: reduce $S(0)$ to below $1/R_0$
R0 and disease dynamics

Source: Dimitrov and Meyers, INFORMS 2010
# R0 of various diseases

<table>
<thead>
<tr>
<th>Disease</th>
<th>Transmission</th>
<th>$R_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measles</td>
<td>Aerosol</td>
<td>12–18$^{[20][30]}$</td>
</tr>
<tr>
<td>Chickenpox (varicella)</td>
<td>Aerosol</td>
<td>10–12$^{[31]}$</td>
</tr>
<tr>
<td>Mumps</td>
<td>Respiratory droplets</td>
<td>10–12$^{[32]}$</td>
</tr>
<tr>
<td>Rubella</td>
<td>Respiratory droplets</td>
<td>6–7$^{[b]}$</td>
</tr>
<tr>
<td>COVID-19 (Delta variant)</td>
<td>Respiratory droplets and aerosol</td>
<td>5–8$^{[37]}$</td>
</tr>
<tr>
<td>Polio</td>
<td>Fecal–oral route</td>
<td>5–7$^{[b]}$</td>
</tr>
<tr>
<td>Pertussis</td>
<td>Respiratory droplets</td>
<td>5.5$^{[36]}$</td>
</tr>
<tr>
<td>Smallpox</td>
<td>Respiratory droplets</td>
<td>3.5–6.0$^{[39]}$</td>
</tr>
<tr>
<td>COVID-19 (Alpha variant)</td>
<td>Respiratory droplets and aerosol</td>
<td>4–5$^{[37]}$</td>
</tr>
<tr>
<td>HIV/AIDS</td>
<td>Body fluids</td>
<td>2–5$^{[43]}$</td>
</tr>
<tr>
<td>COVID-19 (ancestral strain)</td>
<td>Respiratory droplets and aerosol</td>
<td>2.9 (2.4–3.4)$^{[42]}$</td>
</tr>
<tr>
<td>SARS</td>
<td>Respiratory droplets</td>
<td>2–4$^{[43]}$</td>
</tr>
<tr>
<td>Diphtheria</td>
<td>Saliva</td>
<td>2.6 (1.7–4.3)$^{[44]}$</td>
</tr>
<tr>
<td>Common cold</td>
<td>Respiratory droplets</td>
<td>2–3$^{[45]}$</td>
</tr>
<tr>
<td>Ebola (2014 outbreak)</td>
<td>Body fluids</td>
<td>1.8 (1.4–1.8)$^{[46]}$</td>
</tr>
<tr>
<td>Influenza (2009 pandemic strain)</td>
<td>Respiratory droplets</td>
<td>1.6 (1.3–2.0)$^{[2]}$</td>
</tr>
<tr>
<td>Influenza (seasonal strains)</td>
<td>Respiratory droplets</td>
<td>1.3 (1.2–1.4)$^{[47]}$</td>
</tr>
<tr>
<td>Andes hantavirus</td>
<td>Respiratory droplets and body fluids</td>
<td>1.2 (0.8–1.6)$^{[48]}$</td>
</tr>
<tr>
<td>Nipah virus</td>
<td>Body fluids</td>
<td>0.5$^{[49]}$</td>
</tr>
<tr>
<td>MERS</td>
<td>Respiratory droplets</td>
<td>0.5 (0.3–0.8)$^{[50]}$</td>
</tr>
</tbody>
</table>

- Takes time to estimate!
  - Not as easy
- E.g. SARS was estimated in hospitals
  - Where perfect mixing was a reasonable assumption
- NOT homogenous in several situations
- COVID-19
  - Still under investigation for novel variants


• Spatially structured
• For example: modeling COVID-19 and influenza, Zika, Ebola...

• Model heterogeneity by using travel data
  • But assume homogeneity at ‘right’ granularities

\[ \sigma_{ij}: \text{ daily passenger flow from city } i \text{ to city } j \]
\[ n_i: \text{ population of city } i, \text{ assumed to be fixed} \]
\[ X_i(t), Y_i(t), Z_i(t): \text{ number of people in S/I/R states in city } i \text{ at time } t \]

\[
X_i^{\text{eff}}(t) = X_i(t) + \left[ \sum_j X_j(t) \frac{\sigma_{ji}}{n_j} - \sum_j X_j(t) \frac{\sigma_{ij}}{n_i} \right]
\]

Similarly, \( Y^{\text{eff}} \) and \( Z^{\text{eff}} \)
Metapopulation Models contd.

\[ X_i(t + 1) = X_i(t) + \sum_j X_i^{\text{eff}}(t) \beta \frac{I_j^{\text{eff}}(t)}{N_j} \]

- Written in terms of \( X^{\text{eff}}, Y^{\text{eff}}, Z^{\text{eff}} \)
But... Human contact patterns are not random

Source: Mi Jin Lee at petterhol.me
How to Capture Them?

Example: Using Call Data Records

- Many recent studies on this topic
  
  #raw data


[Image: Diagram showing user locations and activity times]
Numerous COVID-19 examples

- Apple (maps/directions)
- Google (location history)
- Facebook (using high resolution imagery)
- Safegraph (poi access)
- Cubeiq (mobile phones etc)
- ....
Agent-based networked models

- Each individual is an agent in a simulation
- Disease spread over contact networks
  - Model heterogeneous interactions between agents
- Concepts:
  - Social contact networks
  - Twin cities
First principles Approach for Constructing Social Contact Networks

• For individuals in a population
  • Demographics (who)
  • Sequences of their activities (what)
  • Times of their activities (When)
  • Places/locations of their activities (where)
  • Reasons for their activities (Why)

• No explicit datasets available
• Synthesize multiple datasets and domain knowledge
• Can model behavioral changes as well

[Marathe and Vullikanti, CACM 2013]
First principles Approach for Constructing Social Contact Networks

[Marathe and Vullikanti, CACM 2013]
Example: COVID-19 in MA

Calibration of Mechanistic Models

- Estimate parameters
  - Beta, delta, initial conditions

\[ \{\beta^*, \delta^*\} = \arg \min (R(t) - R_{\text{observed}}(t))^2 \]

- Typical data includes
  - Time-series of new cases from surveillance
  - Lots of data problems (missing data, biases, lags)

- For example for COVID-19
  - Calibration on infected cases is unlikely to be robust
  - On mortality and hospitalizations likely to be better
Typically

- Ranges of parameters
  - From epidemiological data

- Try to model uncertainty in the data
  - Multiple stochastic calibrations
Pros/Cons Mechanistic Models

• Workhorse of epidemiology
  • Many success stories over 100 years
  • Easy to extend and build (e.g. see COVID-19 work)
  • Good numerical solvers exist
    • Some can also be handled analytically
    • Long history of ODE and Dynamical theory
    • See Strogatz: Nonlinear Dynamics and Chaos

• Useful to get intuition and some broad principles
  • More qualitative rather than quantitative
Pros/Cons contd.

• Sometimes does not reflect reality
  • SARS example
    • High R0 (2.2-3.6)
    • Estimates were based on hospital wards, where full mixing was reasonable

• Calibration is challenging
  • Small deviations in parameters can lead to very different results
Remarks

• A lot more to say about mechanistic models
  • Only reviewed some concepts and models

• Other resources:
  • N. Dimitrov and L. Meyers. 2010. Mathematical approaches to infectious disease prediction and control. INFORMS, 1–25
Part 3: Statistical Models
Statistical Models

• Also known as phenomenological models.

• Intuition:
  • Find the best function from a family of functions that approximate forecast target given input data.
  • Best approximate is found using past training data.

\[
\min_{f \in \mathcal{H}} \sum_{i=1}^{T} \mathcal{L}(f(x_i) - y_i)
\]

• Modeling approaches:
  1. Regression models
  2. Language models
  3. Neural models
  4. Density estimation models
Regression Models

• Assume a linear relationship between input features and future forecast: \( \tilde{y} = w_0 + w^T x \)

• The features \( x \) can be high-dimensional set of multi-modal features
  
  • Eg: Past values of epidemic curve (called AutoRegressive models), Search query volumes, word occurrence in text, etc.
AutoRegressive Models

• Use past values of epidemic cures as features to predict future values
• Eg:

\[ y_t = \sum_{j=1}^{p} \phi_j y_{t-j} + \phi_0 + \epsilon \]

• We can also add difference between values as features (like in ARIMA)
Google Flu Trends

- Simple linear model for nowcasting ILI
- Use search logits of query fractions as features

\[ \text{logit}(P) = \beta_0 + \beta_1 \cdot \text{logit}(Q) + \epsilon \]

High-impact work, media coverage

However,…

• Didn’t capture changing trends in keyword correlates, i.e. didn’t handle data drift
• Failed to capture H1N1 pandemic, overestimate 2012-13 season
ARGO

- ARGO: AutoRegression with Google search data
- Auto Regressive: past N ILI values are used
- Uses separate variables for multiple search queries
- Search data: Of current time \( t \)

\[
y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + \epsilon
\]
ARGO2

- Simultaneously predict HHS and national level ILI
- Capture interdependencies across regions
- Step 1: Region-level independent prediction
- Step 2: Refining prediction using increments modelled as multi-variate Gaussian with inter-region covariates

[Ning+ 2019 Sci. Reports]
2. Language models: using Tweets to forecast H1N1 pandemic

- Topic modelling approach
  - Cluster tweets

- Combines
  - Information propagation on Twitter
  - Epidemiological model

[Chen+ ICDM '17]
States of infection cycle

- Model states of infection cycle using tweets
Forecasting

- Hidden states model flu-state (SEIR)
- Learn topic model that
  - models vocabulary for hidden state and
  - transition probabilities across states
HFSTM Model

- Generating tweets
  - Generate state for tweet
  - Generate topic for word

S: This restaurant is really good
E: The movie was good but it was freezing
I: I think I have flu

Topic: [Background, Non-flu, State]

State: [S, E, I]

Use EM Algorithm for learning parameters
Online interactive example
[3] Neural Models

• Why deep learning?
  • Capture non-linear patterns in high-dimensional data with minor assumptions
  • Flexible learning of rich representations
  • Leverage multiple sources of data of variety of modalities
  • Composite signals are challenging for calibration
    • E.g. %ILI is a mix of multiple flu strains and others
Modeling considerations for neural models

<table>
<thead>
<tr>
<th>Aspect</th>
<th>DISEASE SPREAD</th>
<th>DATA</th>
<th>UTILIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Challenges</td>
<td>Spatial Transmission</td>
<td>Sparse data</td>
<td>Interpretability</td>
</tr>
<tr>
<td></td>
<td>Mobility</td>
<td>Data revisions</td>
<td>Uncertainty quantification</td>
</tr>
<tr>
<td></td>
<td>Mask adoption</td>
<td>Anomalies</td>
<td>Actionable forecasts</td>
</tr>
<tr>
<td></td>
<td>Social distancing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Modeling ideas

1. Model temporal dynamics via similarity
   • Overcome data sparsity
   • Enable interpretability

2. Transfer knowledge representations
   • Learn from other relevant domains

3. Incorporate spatial structure
   • Model the spread over adjacent regions
   • Propagation over networks
Modeling idea 1: Model temporal dynamics via similarity

• Idea: clustering for prediction
Model temporal dynamics via similarity CONTD.

- **Idea:** Dynamic clustering for prediction

![Graph showing incidence over time with different years and comparison markers](image)
Model temporal dynamics via similarity CONTD.

- Idea: **Dynamic deep clustering for prediction with limited data**
Find similarity to historical seasons

- Embed the historical seasons to capture the similarity with the current season
- Current season is observed only till week \( t \)
- Use snippets of historical seasons till week \( t \) to learn embedding

![Graph showing current season and historical seasons](image)
Data-driven approach: EpiDeep

- Deep approach for forecasting ILI based on historical data
- Forecasts multiple targets
- One of the first deep learning-based approach for influenza forecasting
- Performs pretty well in real-time forecasting

[Adhikari+, KDD’19]
Experiments: Baselines

- **EB**: an empirical Bayesian approach. [Brooks+, PLOS ComBio 2015]
  - published and publicly available version
- **ARIMA**: an auto-regressive method for making predictions on time-series data.
- **HIST**: historical average of all previous seasons.
- **KNN**: selects the top k closest historical seasons to the current season, and make predictions on their average. [Nsoesie+, Stats Com in Infectious Diseases 2011]
- **LSTM**: a version of [Venna+, IEEE Access 2017] without climate and geographical data.
• How well does EPI\textsc{Deep} perform in different tasks for the national region?

EpiDeep outperforms baselines in most settings.
Other examples of modeling temporal similarity

- Temporal and geo. similarity (adjacent regions)

  [Wang et al., BigData 2020]

- Inter-series similarity

  [Jin et al., SDM 2021]
Detour: Incorporating guidance in Epidemic Forecasting

- Epidemiological experts may notice unideal behavior exhibited by statistical approaches.

- How to enforce epidemic forecasting models to incorporate expert’s guidance to show desirable properties?

[Rodríguez+, epiDAMIK @ KDD 2020]
Modeling idea 2: Transfer knowledge representations

- Neural model automatically learn what to transfer
  - Not everything is relevant! Needs selection

- Examples:
  - From one country to another country
    - Even in different continents
    - In Panagopoulos et al., AAAI 2020
  - From a historical scenario to a novel scenario
    - From pre-COVID flu to COVID-contaminated flu counts
    - In Rodríguez et al., AAAI 2020
Influenza Surveillance in the Early COVID Pandemic

• March 2020:
  • Flu counts are syndromic (symptomatic)
  • COVID-Flu are symptomatic similar
  • COVID was being captured by flu surveillance systems

ILINet surveillance network

CDC

Region 2 (NY)

Region 10 (WA)

Epidemiological Week (EW)
A Novel Forecasting Setting

- Influenza counts may be affected by
  - COVID “contamination”
  - Shift in healthcare seeking behavior
- This new scenario lead us a novel forecasting problem
- Historical flu models unable to adapt to new trends
New COVID-related signals correlate with new trends

- Line-list based
- Testing
- Crowdsourced
- Mobility
- Exposure
- Social Media surveys
Attentive transfer learning for heterogeneous domains

- CALI-Net: steer a historical flu model (EpiDeep, KDD 2019) with new COVID-related signals
Modeling idea 3: Incorporate spatial structure

- Pathogens propagate to adjacent regions
  - And then to new adjacent regions
- Propagation over spatial graphs
Graph message passing for spatial propagation

- ColaGNN:
  - Graph neural network for spatial structure
  - Dilated convolution for temporal modeling

[Deng+, CIKM 2020]
Density Estimation Models

- Directly model the forecast distribution
- **Parametric**: parameters of distribution as function of features
- **Non-parametric**: Function of training datapoints leveraging similarity
Empirical Bayes

- Idea: Current season’s epidemic curve is a probabilistic distribution of features
- Model parameters:
  - Similarity is shape to past sequences
  - Peak height, week
  - Scaling factor of the curve
- All modelled as priors of forecast distribution
- Use Bayesian Inference to calibrate for current season

[Brooks+ 2015 PLoS]
Delta Density

• Use kernel density estimation to leverage similarity with historical seasons
• One of the top models in Flusight 2017 challenge

[Brooks+ 2017 PLoS]
Gaussian Process

- Used Gaussian Process over incidence values of previous seasons
- Showed reasonable confidence intervals and state-of-art log score over past models
Neural models for calibrated forecasts

- Density Estimation models don’t focus on well-calibrated forecasts
  - Can’t adapt to provide reliable forecast uncertainty on novel patterns
EpiFNP: Neural non-parametric model for better calibration

• Leverage Neural Sequential models to capture long term sequential patterns

• Non-parametric Gaussian Process
  • Flexibly model forecast distribution
  • Leveraging similarities with past historical sequences

准确和校准的神经预测模型
EpiFNP: Architecture

Sequential representations + neural Gaussian processes

Rodríguez, Kamarthi, and Prakash 2021
Results

Well calibrated predictions

Adapt to novel patterns

Explaining predictions

Most similar seasons chosen by EpiFNP

Rodríguez, Kamarthi, and Prakash 2021
Pros/Cons Statistical Models

- State of the art in multiple forecasting tasks
  - Short-term forecasting
  - Uncertainty quantification
- Bring a complementary perspective closer to data
- Unaware of epidemic spread mechanisms
  - Poor performance in long-term
  - Unable of evaluating what-if scenarios
Part 4: Hybrid Models
Hybrid Models

• Use both mechanistic and statistical components as complementary pieces.

• Modeling approaches:
  1. Discrepancy modeling
  2. Parameter estimation
[1] Discrepancy modeling

• Statistical model resolves the discrepancies between a model (often mechanistic) and ground truth data.
• In other words, statistical model refines/corrects another model.
Hierarchical Bayesian Model for Mechanistic Discrepancy

[Osthus et al. 2019, Bay. Analysis]

- DBM refines mechanistic predictions with a hierarchical Bayesian model.

- Refinement components:
  - State-specific deviation
  - Season-specific deviation
  - Trends

Figure credit: Sara Del Valle, LANL
[2] Parameter estimation

- Hierarchical two-layer Gaussian process (GP).
- Upper-layer GP uses country-specific features + policies in place to estimate $R_0$
- Lower-layer GP refines predictions

[Qian+ NeurIPS 2020]
Counterfactual based on new set of policies
Part 5: Ensembles
Ensembles

• Combining models into an "ensemble" often provides more robust forecasts than any single model

• Consistently found across multiple epidemic forecasting efforts
  • Flu: Reich et al. 2019, PLOS Comp Bio
  • Dengue: Johansson et al. 2019, PNAS
  • Ebola: Viboud et al. 2018, Epidemics
Policy makers needed >1 model

Early April 2020

Slide credit: Nicholas Reich, UMass Amherst
Diversity of COVID-19 models

- **IHME-CurveFit**: "hybrid modeling approach to generate our forecasts, which incorporates elements of statistical and disease transmission models."

- **MOBS-GLEAM_COVID**: "The GLEAM framework is based on a metapopulation approach in which the world is divided into geographical subpopulations. Human mobility between subpopulations is represented on a network."

- **UMass-MechBayes**: "classical compartmental models from epidemiology, prior distributions on parameters, models for time-varying dynamics, models for partial/noisy observations of confirmed cases and deaths."

- **UT-Mobility**: "For each US state, we use local data from mobile-phone GPS traces made available by [SafeGraph] to quantify the changing impact of social-distancing measures on 'flattening the curve.'"

- **GT-DeepCOVID**: "This data-driven deep learning model learns the dependence of hospitalization and mortality rate on various detailed syndromic, demographic, mobility and clinical data."

- **Google Cloud AI**: "a novel approach that integrates machine learning into compartmental disease modeling to predict the progression of COVID-19"

- **Facebook AI**: "recurrent neural networks with a vector autoregressive model and train the joint model with a specific regularization scheme that increases the coupling between regions"

- **CMU-TimeSeries**: "A basic AR-type time series model fit using lagged values of case counts and deaths as features. No assumptions are made regarding reopening or governmental interventions."

Slide credit: Nicholas Reich, UMass Amherst
What is the optimal ensemble?

<table>
<thead>
<tr>
<th>&quot;Robust&quot; (i.e. ensemble does not &quot;blow up&quot;)</th>
<th>&quot;Trained&quot; (i.e. component forecasts are weighted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- Median of best 5 or 10 individual models
- Weighted median, weights from a weighted mean ensemble
- Weighted median, weights based on relative WIS

• Takeaway: use a robustly trained ensemble

Slide credit: Nicholas Reich, UMass Amherst
Results in COVID-19

[Craemer+, medRxiv 2021]
All models are useful

• No model is always good
• Top models in COVID Forecast Hub:
  • Mechanistic
  • Statistical
• Usefulness may depend on
  • Epidemic stage: uptrend, downtrend, near peak
  • Geographical region
  • But largely an open research question
Super-ensembles

[Adiga+, medRxiv 2021]
Part 6: Epidemic Forecasting in Practice
Epidemic Forecasting Pipeline

A. Data Processing
Raw data
- Processing: delays, anomalies, revisions
- Exploratory analysis

B. Model Training & Validation
- Multiscale dynamics
- Uncertainty quantification
- Feature engineering and selection
- Interpretability
- Robustness to noisy data
- Scenario selection

C. Utilization & Decision Making
- Real-Time Forecasting
  - Ensemble of Real-Time Predictions
  - Dashboards
  - CDC Initiatives

Decision Making
- Resource Allocation
- Risk Assessment
- Forecast Communication

Feedback

Rodríguez, Kamarthi, and Prakash 2021
Forecasting in Practice

• Topics:
  1. US CDC initiatives
  2. Real time experiences
  3. Decision making
[1] Forecasting Initiatives

- CDC’s Epidemic Prediction Initiative
  - 2014-2020 Influenza – US National
  - 2015 Dengue – Iquitos, Peru & San Juan, PR
  - 2015-2020 Influenza – US HSS Regions
  - 2017-2020 Influenza – US States
  - 2019-2020 Department of Defense Influenza – US military facilities
  - 2020 West Nile neuroinvasive disease – US counties

Slide credit: Matt Biggerstaff, US CDC
COVID-19 Forecast Hubs

Source: Johannes Bracher, KIT Karlsruhe and HITS Heidelberg
Standardization efforts of real-time forecast submissions

A Forecast Archive

Welcome to the Zoltar forecast archive, an open-source web application that facilitates the storage, retrieval, evaluation, and visualization of point and probabilistic forecasts. Zoltar was developed to assist with many kinds of real-time forecasting projects.

Learn more »

Project: COVID-19 Forecasts

<table>
<thead>
<tr>
<th>Summary:</th>
<th>111 models, 5424 forecasts, 79,213,246 predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owner:</td>
<td>covid19hub</td>
</tr>
<tr>
<td>Model Owners:</td>
<td>ydh28, vrushti-mody</td>
</tr>
<tr>
<td>Time Interval Type</td>
<td>Week</td>
</tr>
</tbody>
</table>
[2] Real-time Experience and Challenges
Operational Deep Learning Framework [Rodríguez+, IAAI 2021]
Highlights of results

Anticipate Trend Changes

Capture finer-grain patterns

Excels in short-term forecasting

Provides explanations

Lower is better

Main epidemic driver: mobility
Top Ranked Model

- Cramer et al. evaluated model predictions submitted to the CDC.

- Evaluation:
  - 1 to 4 week ahead
  - May 2020 - Oct 2021 (1+ year)
  - 51 locations (national + states)

- DeepCOVID ranked **top 5** out of 25 individual models.
Data Challenges: Don't Underestimate!

(C1) Multiple data sources and formats
  • Format varies over time

(C2) Select signals with epidemiological significance

(C3) Temporal misalignment
  • Delays, pause in reporting, differ in granularity

(C4) Spatial misalignment
  • Differ in granularity: county vs state vs national

(C5) Data quality and missing data
  • Noisy and unreliable for some states
  • New hospitalizations (target) is not reported by all states
Data Quality issues: Data Revisions

Revision of mortality for TX released on week 28

Stability Time = 24 weeks
Mortality finally stabilize to around 400 at week 28+24 = week 52!

Backfill Error = |404-223|/|404| = 44.8%

Human error, data instability, delays, disasters

Initial/real-time value = 223

Revision of mortality for TX released on week 28
Data revisions are significant

- Over half the signals show backfill error over 32%
- Targets revised by 5%
- Stability time average around 3-4 weeks
Model performance is affected by data revision

Model Prediction \[\rightarrow\] Initial (real-time) target

Not same!

Model Prediction \[\leftarrow\] Stable label
Refining predictions due to backfill

Given
- Bseqs of all past signals from all regions
- History of model’s predictions due to training on real-time data
- Model’s current week’s prediction

Output
- Refined current week prediction of model that is closer to (unknown) revised target

Current week prediction

Refined Prediction
Back2Future

- Learns from past revision patterns of all features
- Refines model predictions of *any* model given prediction history
Back2Future: Results

- Improves predictions of top-models by 6.65% with over 10% in some US states

Takeaway: data quality issues can be helped with statistical correction
Demo

Link: https://github.com/AdityaLab/Back2Future
[3] Decision making

- Leverage predictions to inform decision making for policymakers, public health workers, supply chains, etc.

- Types:
  - Strategic: Large-scale policies
  - Tactical: Small-scale, high density action space, to accomplish a narrow goal
Strategic Interventions for mitigating foot and mouth disease

[Probert+ PloS 2018, RS 2019]

• Use simulations based on past outbreak data.

• Control measures:
  • Vaccinate animals
  • Cull farm animals

• Can be solved as Sequential Decision making problem (leverage Reinforcement Learning)
Tactical Interventions for ventilator allocation

- Bertsimas et al. (2021) leverage future case forecasts to model optimal resource-allocation

- Tradeoff:
  - Satisfy future demand for ventilators
  - Reduce inter-state transport cost
Final Remarks
[1] All models are useful

- We have provided a toolkit of methods
  - Ensembles are often the most robust
- Mechanistic often better for qualitative insights rather than quantitative accuracy
  - Especially agent-based models
- Statistical models have SOTA performance in multiple short-term forecasting tasks
- Hybrid models are gaining traction
[2] Asking when, where, who

- When and where did the outbreak start? Who got infected?
  - Requires accurate and timely data from the ground
  - Reports from public health agencies e.g. CDC, WHO, PAHO,…

- Very challenging!

- What to expect as it is spreading? What kinds of people are likely to get infected? When will it peak?
  - Many outbreaks die out on their own
  - Need data plus models to understand how the disease will spread
    - Roles: short term, long term prediction vs understanding
    - Conflicting goals: accuracy, transparency, flexibility

- Important objective: forecast how the outbreak will spread for resource planning and decision making
  - Many ‘forecasting challenges’ recently! E.g. flu, COVID etc.
  - How big will the peak be?
  - When will it peak?
  - Public Communication

Data + Models + Efficient Algorithms + Simulations
Studying epidemics in *real time*

  - Five areas:
    - Pandemic risk,
    - vulnerable populations,
    - available interventions,
    - implementation possibilities
    - pitfalls, and public understanding
Studying epidemics in *real time*

- **Modeling Before the epidemic**
  1. Determine the (non)medical interventions required,
  2. feasibility of containment
  3. optimal size of stockpile
  4. best use of pharmaceuticals once a pandemic begins

- **Modeling After/During the epidemic**
  1. Quantifying transmission parameters,
  2. Interpreting real-time epidemiological trends,
  3. measuring antigenic shift
  4. assessing impact of interventions.

Data Science is very important for all of these!
Why data science?

• IN ADDITION to increasing data collection:
  • Questions about epidemic spread naturally have a large spatial and temporal scale
    • And multiple such scales!
  • Small and big data, noisy and incomplete
  • New tools can help epidemiologists
  • New data science and AI techniques which can handle end-to-end learning
  • New Stochastic optimization techniques
Data Science for Epidemiology

ML & Stats.

Theory & Algo.

Comp. Systems

Biology

Physics

Social Science

Econ.

Big Picture

Rodríguez, Kamarthi, and Prakash 2021
Reminder on Workshop Webpage

- [https://adityalab.cc.gatech.edu/workshops/21-forecasting-f4sg.html](https://adityalab.cc.gatech.edu/workshops/21-forecasting-f4sg.html) or [b.gatech.edu/3cBPfQ7](b.gatech.edu/3cBPfQ7)

- All Slides will be posted there.

- Talk video as well (later).

- **License**: for education and research, you are welcome to use parts of this presentation, for free, with standard academic attribution. For-profit usage requires written permission by the authors.
Stay tuned

• Survey paper coming soon

• Epidemiology meets Data Science Workshop
  • https://epidamik.github.io/
  • Hosted at KDD 2021

• And more exciting research and tools!
We recently organized the National PREVENT symposium (Feb 22/23): Cross-cutting disciplines and scales for pandemic prevention and prediction

Videos and handouts: prevent-symposium.org
Thanks!

• To F4SG for the invitation
• CDC COVID-19 Forecasting Hub
• Data collection volunteers
• Collaborators
• Funding agencies

Stay in touch!
Alexander Rodríguez
• email: arodriguezc@gatech.edu
• web: cc.gatech.edu/~acastillo41
  @arodriguezca

Harsha Kamarthi
• email: hkamarthi3@gatech.edu
• web: www.harsha-pk.com/
  @harsha_64

B. Aditya Prakash
• email: badityap@cc.gatech.edu
• web: cc.gatech.edu/~badityap/
  @badityap

Fill survey: https://forms.gle/5JuSSTQde8FV3PVq9