

AI for Data-Centric Epidemic Forecasting

Alexander Rodríguez
Harshavardhan Kamarthi
B. Aditya Prakash

College of Computing
Georgia Institute of Technology

February 8, 2023



Rodríguez, Kamarthi, and Prakash 2023



THE 37TH AAAI CONFERENCE ON
ARTIFICIAL INTELLIGENCE

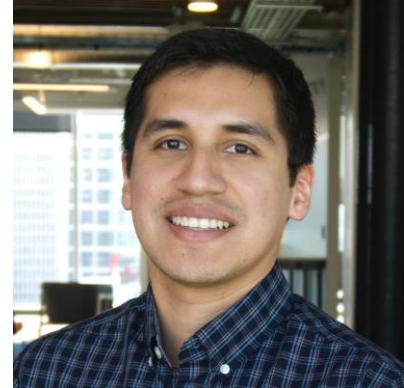
FEBRUARY 7-14, 2023 • WASHINGTON, DC, USA
WALTER E. WASHINGTON CONVENTION CENTER

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: cc.gatech.edu/~badityap/



About us



- Alexander Rodríguez
 - 5th year PhD candidate, graduating July 2023
 - Data science/ML in time series and networks
 - Motivated by impactful problems
 - Critical infrastructure networks
 - Epidemic forecasting
 - PhD thesis topic: AI for epidemic forecasting
 - Homepage: cc.gatech.edu/~acastillo41/

About us



- Harshavardhan Kamarthi
 - 3rd year PhD student
 - Research Interests
 - Epidemic forecasting
 - Probabilistic forecasting and uncertainty quantification
 - Deep Probabilistic models
 - Homepage: harsha-pk.com

Tutorial Webpage

The screenshot shows a GitHub repository interface for a tutorial. At the top left is a navigation bar with a three-line menu icon, the text "README.md", and a pencil icon for editing. The main content area has a title "AAAI-23 Tutorial: AI for Data-Centric Epidemic Forecasting". Below the title are links: "Survey paper companion: PDF", "Slides: PDF", and "Website: adityalab.cc.gatech.edu/talks/aaai-23-ai4epi-tutorial.html". A section titled "Tutorial abstract" is visible at the bottom.

AAAI-23 Tutorial: AI for Data-Centric Epidemic Forecasting

Survey paper companion: [PDF](#)

Slides: [PDF](#)

Website: adityalab.cc.gatech.edu/talks/aaai-23-ai4epi-tutorial.html

Tutorial abstract

- github.com/AdityaLab/aaai-23-ai4epi-tutorial
- All Slides will be posted there. Link to 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.

Survey companion arxiv.org/abs/2207.09370

- Tutorial largely based on recent survey paper

Data-Centric Epidemic Forecasting: A Survey

Alexander Rodríguez*, Harshavardhan Kamarthi*, Pulak Agarwal,
Javen Ho, Mira Patel, Suchet Sapre, and B. Aditya Prakash†

College of Computing, Georgia Institute of Technology, USA

Abstract

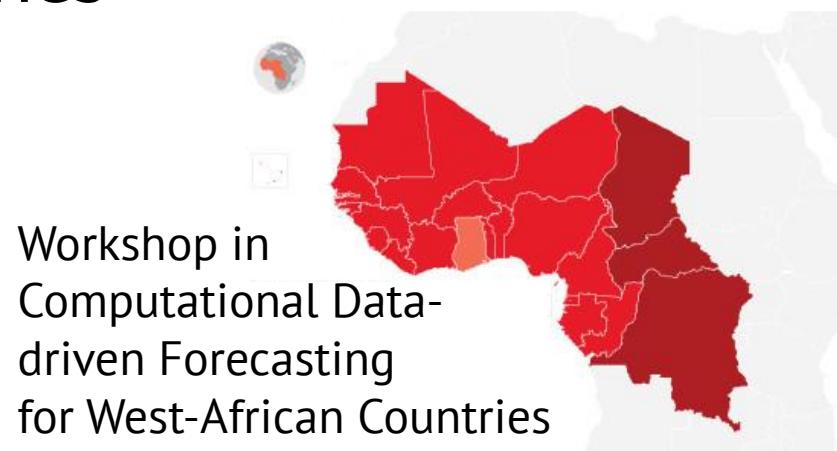
The COVID-19 pandemic has brought forth the importance of epidemic forecasting for decision makers in multiple domains, ranging from public health to the economy as a whole. While forecasting epidemic progression is frequently conceptualized as being analogous to weather forecasting, however it has some key



All citations in this tutorial can be found there

Data-centric epidemic forecasting for practitioners

- Invited by Forecasting for Social Good (F4SG) Research Network
- **Target audience:** researchers and practitioners from West African Countries
- **Today's focus:** ML/data science innovations



Outline

1. Epidemic forecasting: data and setup (40 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (60 min)
 - 30 min break at 3:15 PM, feel free to catch us for coffee
5. Hybrid models (45 min)
 - 5 min break
6. Epidemic forecasting in practice (25 min)
7. Open challenges and final remarks (20 min)

Plan for the Tutorial

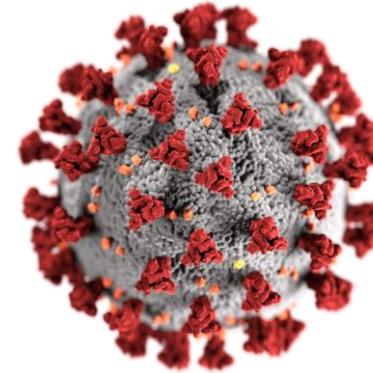
- Theory and research
 - Setting up the epidemic forecasting problem
 - General epidemiology: key concepts and models
 - Statistical modeling and deep learning
 - Research innovations
- Practice
 - Public health initiatives
 - US real-time forecasting experiences
- Open challenges

AI Topics Covered

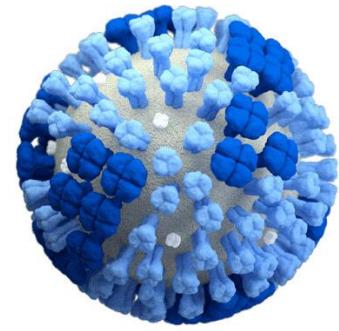
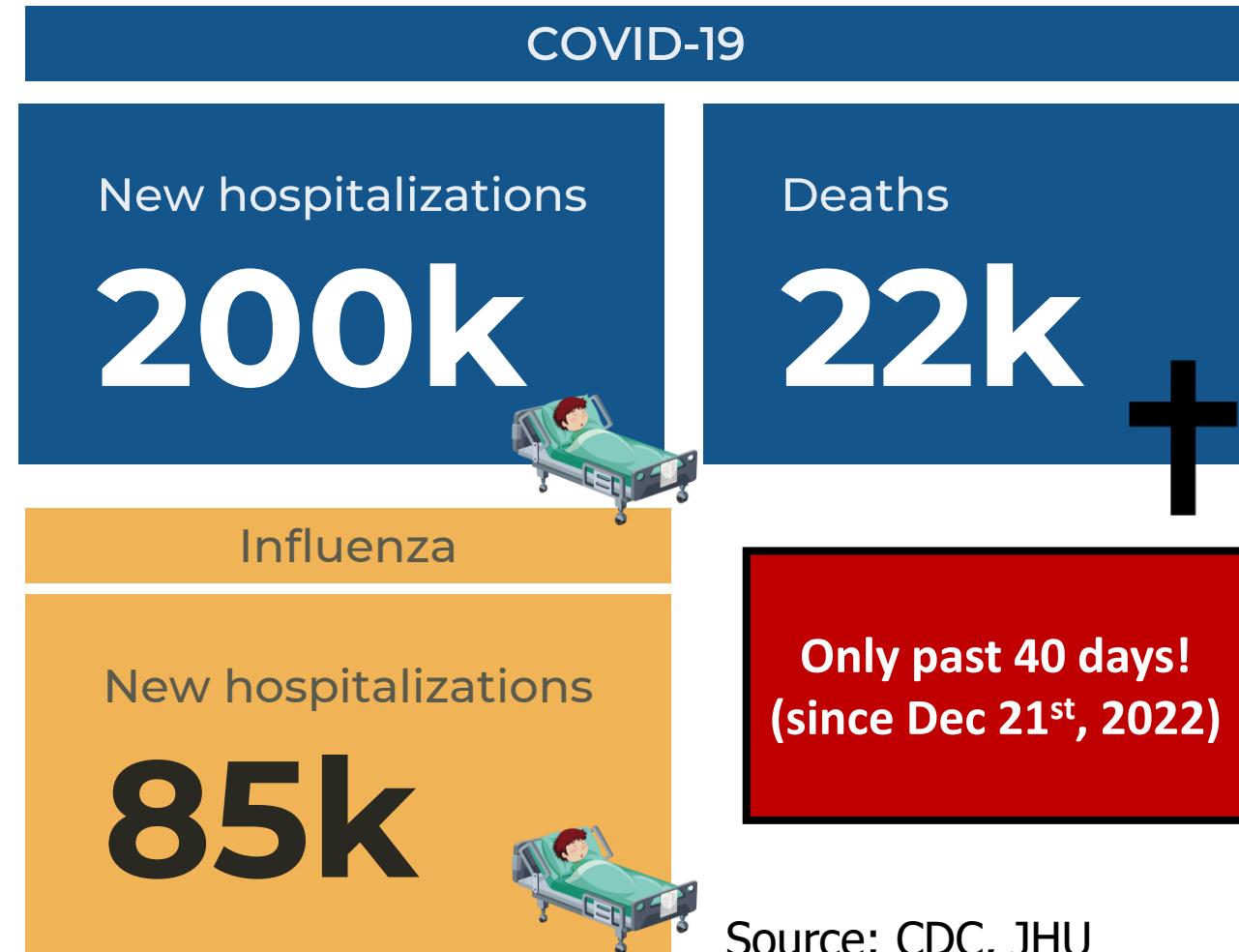
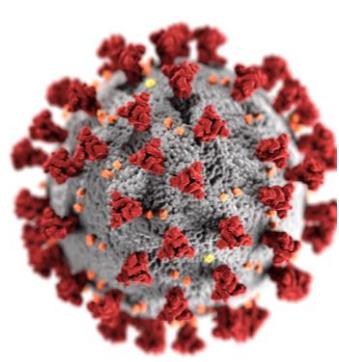
- Deep sequential modeling
- Graph neural networks
- Deep uncertainty quantification
- Multi-view forecasting
- Hierarchical forecasting
- Transfer learning & multi-task learning
- Scientific ML
- AI for agent-based models

Motivation: COVID-19 pandemic

- **Global pandemic**
 - 500+ million cases
 - 6+ million deaths
- Hard to imagine any aspect of life not been affected
- Never before have epi. concepts captured public attention and imagination so vividly! Examples:
 - Reproduction number
 - Non-pharmaceutical interventions
 - Social distancing
 - Surveillance
 - Contact-tracing



Current Toll of Epidemics in the US

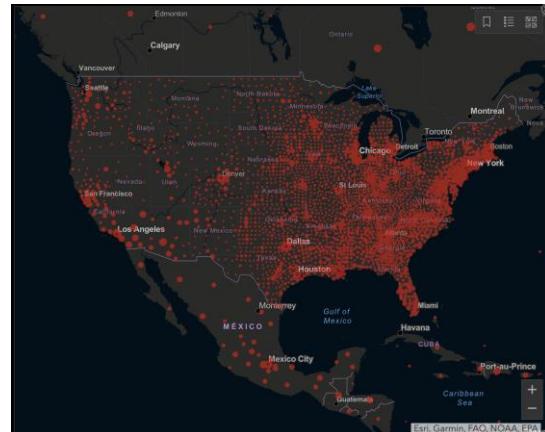


Source: CDC, JHU

COVID-19 trajectories

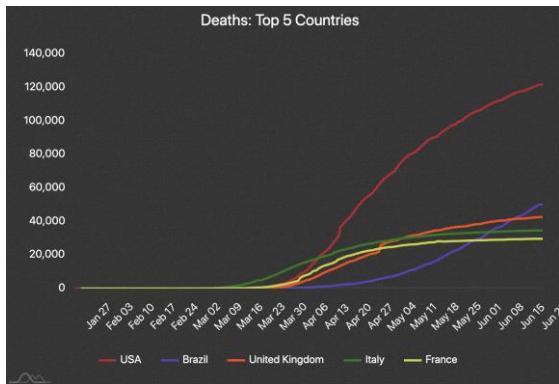


Global spatial incidence distribution

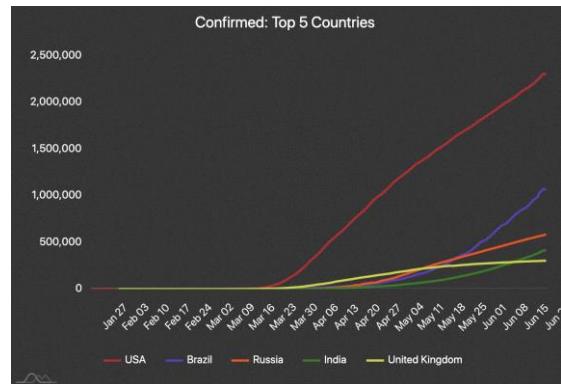


Spatial incidence distribution in USA

[source: <https://gisanddata.maps.arcgis.com>]



Cumulative Mortality



Confirmed cases Cumulative

[source: <https://nssac.bii.virginia.edu/covid-19/dashboard/>]

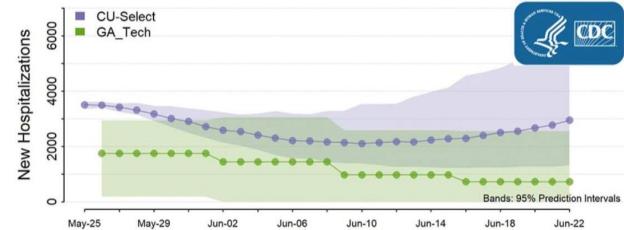
Why Forecasting?

An outlook to the future allow communities to

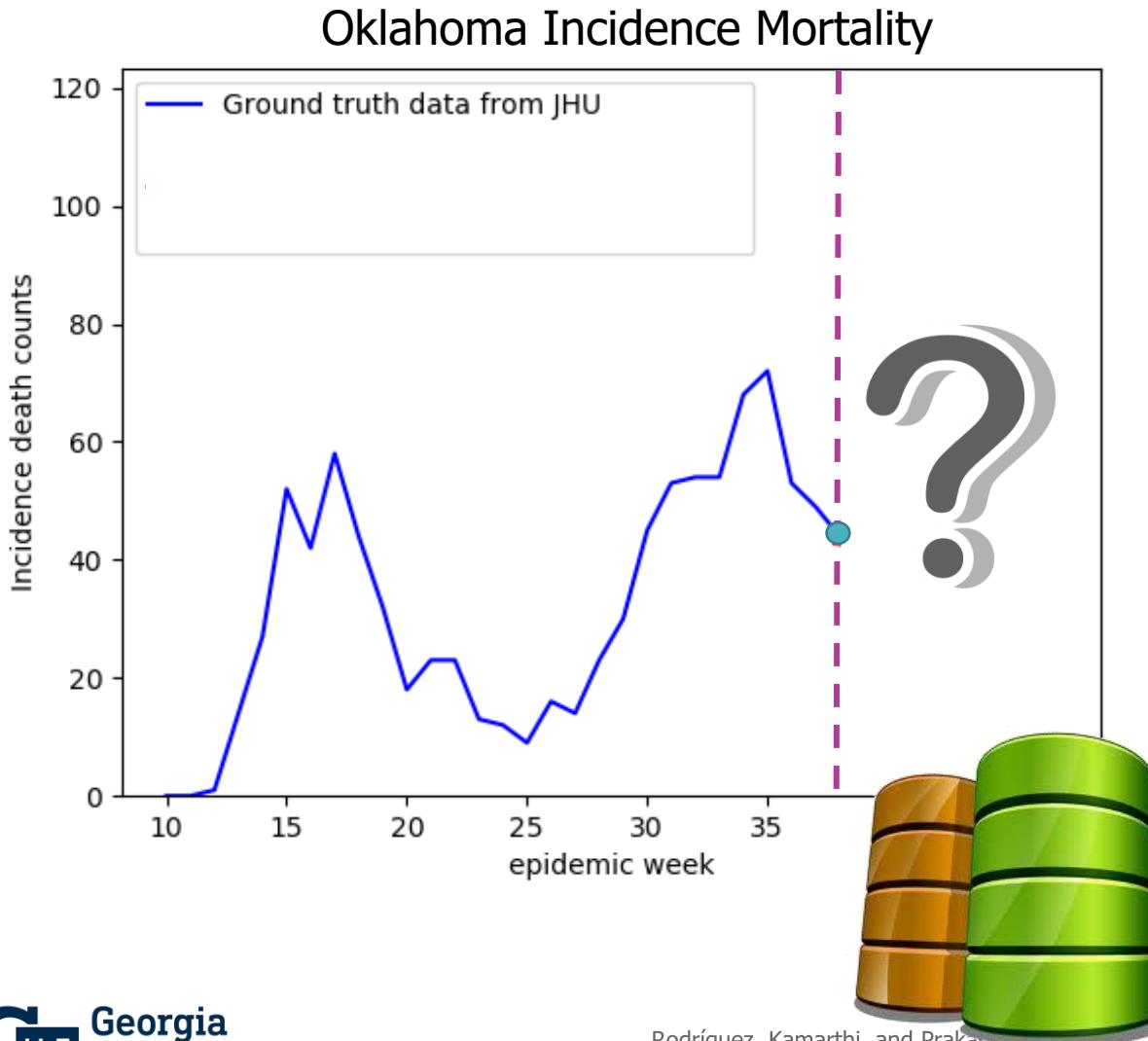
- Allocate resources/budget
 - Ventilators, enable more ICU beds
- Inform public policy
 - E.g., mandate shelter in place?
- Improve preparedness
- Public Communication
- ...



National Forecasts



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



SAFE GRAPH



 Google  kinsa.

Increasing data collection

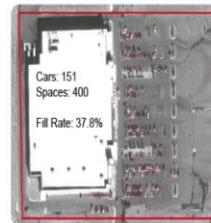
- Mobility
- Point of care
- Line lists
- Surveys
- Social Media
- Genomic
- ...



Medical record



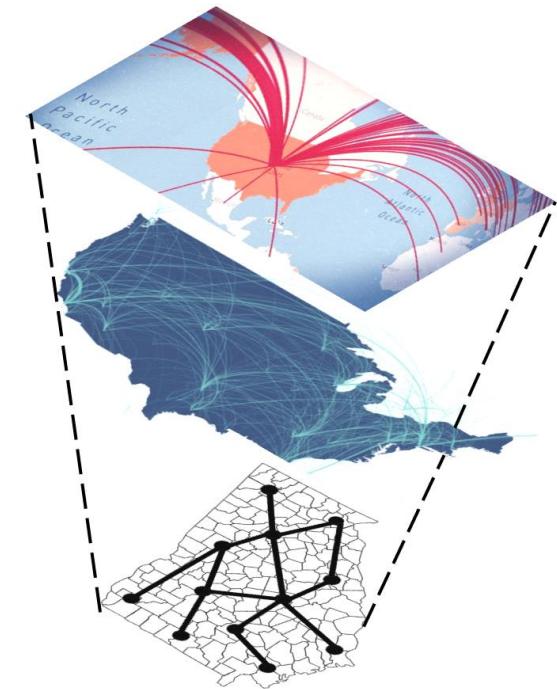
Estimating
"fill rate"
in parking lots



GAATTCATACCAGATCAC **CGGATTCCCGA** CTCCAAATGTGTCCCCCTCACAC
TCCC **CCGATTACCGT** CTTCTGCTCTAGACCACTCTACCCATTCCCCACACT
CACCGGAGCAAAGCCGCCCTCCGTT **CCGATTACCGA** AAAGACCCCA
CCCGTAGGTGGCAAGCTAGCTTAAGTAACGCCACT **TCGATTAACGA** GGAAA
AATACATAACTGA **CCTATTATCGA** GTTCAGATCAAGGTAGGAACAAAGAA
ACA **CCGATTACCGT** AACCGTAAGATARTGGTATCGATACGTAGACAGTTA akash 2

Why Computational Data-centered Forecasting?

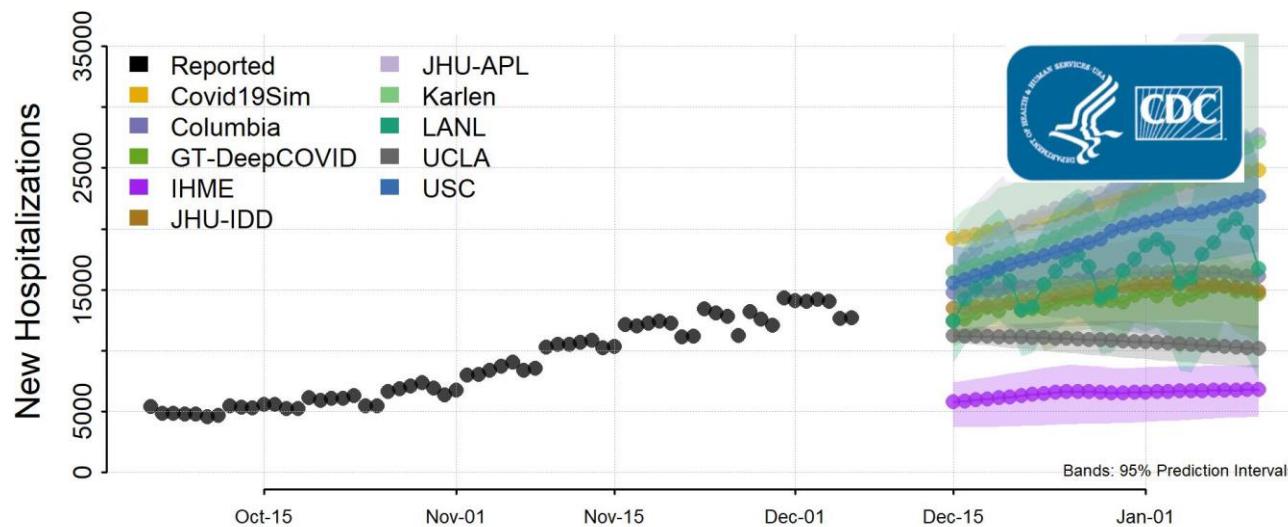
- Epidemic spread is a spatiotemporal phenomena over multi-scale networks
- New end-to-end methods available capable of modeling data with minimal assumptions
- However, traditional methods have difficulties ingesting these data sources
 - Based on ODEs and agent-based models



Our approach

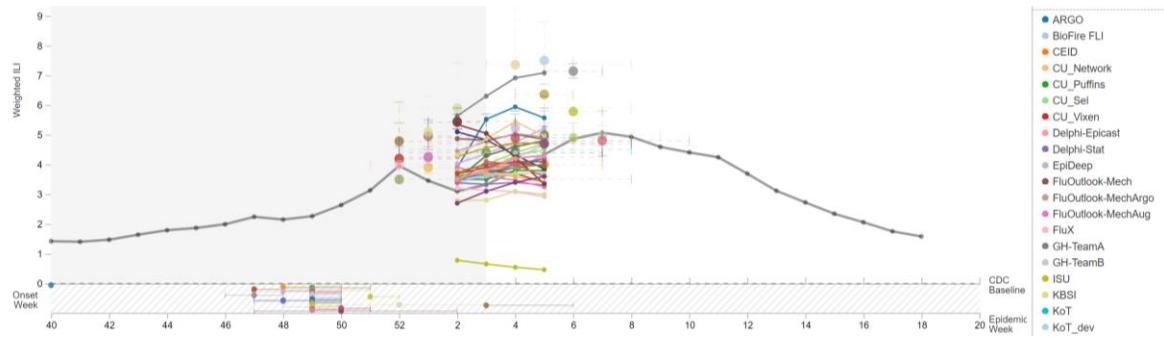
- Before and after the COVID-19 pandemic: Explored **performance** and **utility** of data-driven models in short-term forecasting

National Forecast



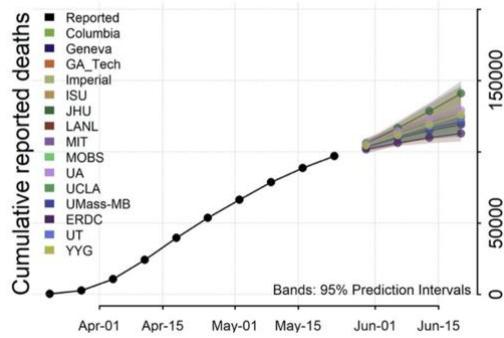
Our Participation in CDC Forecasting Initiatives

Target 1: Influenza like illness per week



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

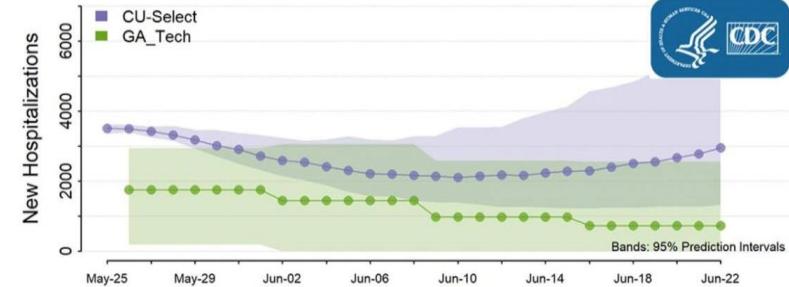
Target 2: Weekly Covid Mortality



Since April End 2020

Target 3: Daily Covid Hospitalizations

National Forecasts



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

facebook

Carnegie
Mellon
University



2nd Prize

C3.ai COVID-19 Grand Challenge



43

Countries



777

Participants

Out of 115 global participants

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

COVID response projects:
cc.gatech.edu/~badityap/covid.html



At AAAI 2023



- This tutorial.
- Two papers:
 - 8809 EINNs: Epidemiologically-informed Neural Networks
(Thursday, February 9, 11:15am – Room 202B)
 - 9579 Detecting Sources of Healthcare Associated Infections
(Friday, February 10, 9:30 am – Room 144A)
- Keynote at the Graphs and More Complex Structures for Learning and Reasoning (GCLR).
 - Tuesday, February 14

Recent Publications

- A. Rodríguez, N. Muralidhar, B. Adhikari, Anika Tabassum, N. Ramakrishnan, B. A. Prakash. Steering a Historical Disease Forecasting Model Under a Pandemic: Case of Flu and COVID-19. In AAAI-21.
- H. Kamarthi, A. Rodríguez, B. A. Prakash. Back2Future: Leveraging Backfill Dynamics for Improving Real-time Predictions in Future. In ICLR 2022.
- A. Rodríguez, A. Tabassum, J. Cui, J. Xie, J. Ho, P. Agarwal, B. Adhikari, B. A. Prakash. DeepCOVID: An Operational DL-driven Framework for Explainable Real-time COVID-19 Forecasting. In IAAI-21.
- A. Chopra*, A. Rodríguez*, J. Subramanian, B. Krishnamurthy, B. A. Prakash, R. Raskar. Differentiable Agent-based Epidemiological Modeling for End-to-end Learning. In AI4ABM @ ICML 2022
- A. Rodríguez, J. Cui, B. Adhikari, N. Ramakrishnan, B. A. Prakash. EINNs: Epidemiologically-Informed Neural Networks. Under review.
- H. Kamarthi, L. Kong, A. Rodríguez, C. Zhang, B. A. Prakash. When in Doubt: Neural Non-Parametric Uncertainty Quantification for Epidemic Forecasting. In NeurIPS 2021.
- A. Rodríguez, B. Adhikari, N. Ramakrishnan, and B. A. Prakash. Incorporating Expert Guidance in Epidemic Forecasting. In epiDAMIK @ KDD 2020.
- H. Kamarthi, L. Kong, A. Rodríguez, C. Zhang, B. A. Prakash. CAMUL: Calibrated and Accurate Multi-view Time-Series Forecasting. In submission (available as arXiv preprint).
- P. Sambaturu, B. Adhikari, B. A. Prakash, S. Venkatramanan, A. Vullikanti. Designing Near-Optimal Temporal Interventions to Contain Epidemics. In AAMAS 2020
- B. Adhikari, X. Xu, N. Ramakrishnan and B. A. Prakash. EpiDeep: Exploiting Embeddings for Epidemic Forecasting. In SIGKDD 2019
- B. Adhikari, B. Lewis, A. Vullikanti, J. Jimenez, and B. A. Prakash. Fast and Near-Optimal Monitoring for Healthcare Acquired Infection Outbreaks. In PLoS Computational Biology. 2019.
- J. Cui, A. Haddadan, A. Haque, Bi. Adhikari, A. Vullikanti and B. A. Prakash. Information Theoretic Model Selection for Accurately Estimating Unreported COVID-19 Infections. In submission (available as medRxiv preprint).
- V. Swain, J. Xie, M. Madan, S. Sargolzaei, J. Cai, M. De Choudhury, G. Abowd, L. Steimle and B. A. Prakash. WiFi mobility models for COVID-19 enable less burdensome and more localized interventions for university campuses. In submission (available as medRxiv preprint).

Outline

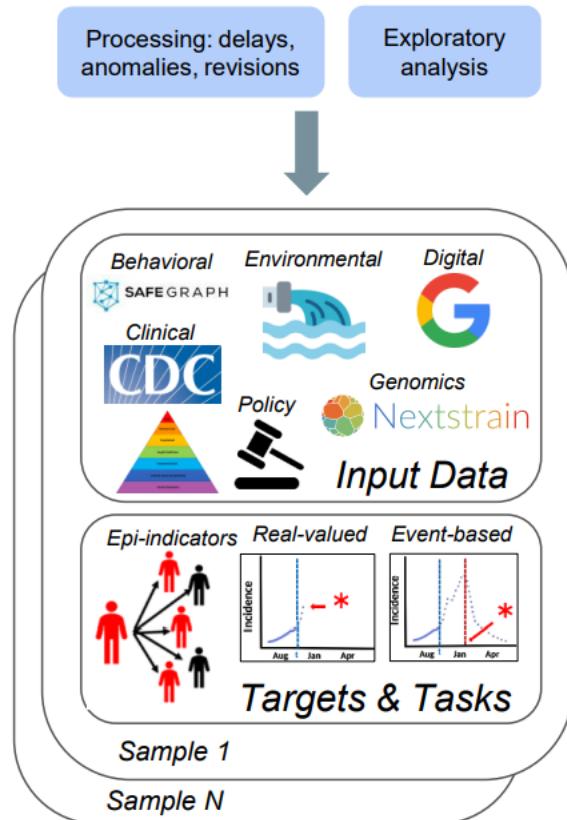
- 1. Epidemic forecasting: data and setup (40 min)**
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (60 min)
 - 30 min break at 3:15 PM, feel free to catch us for coffee
5. Hybrid models (45 min)
 - 5 min break
6. Epidemic forecasting in practice (25 min)
7. Open challenges and final remarks (20 min)

Part 1: Epidemic Forecasting

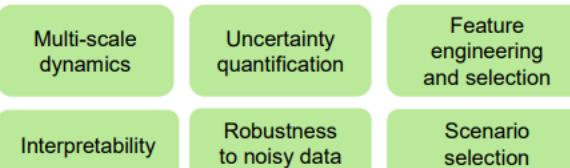
Epidemic Forecasting Pipeline

A. Data Processing

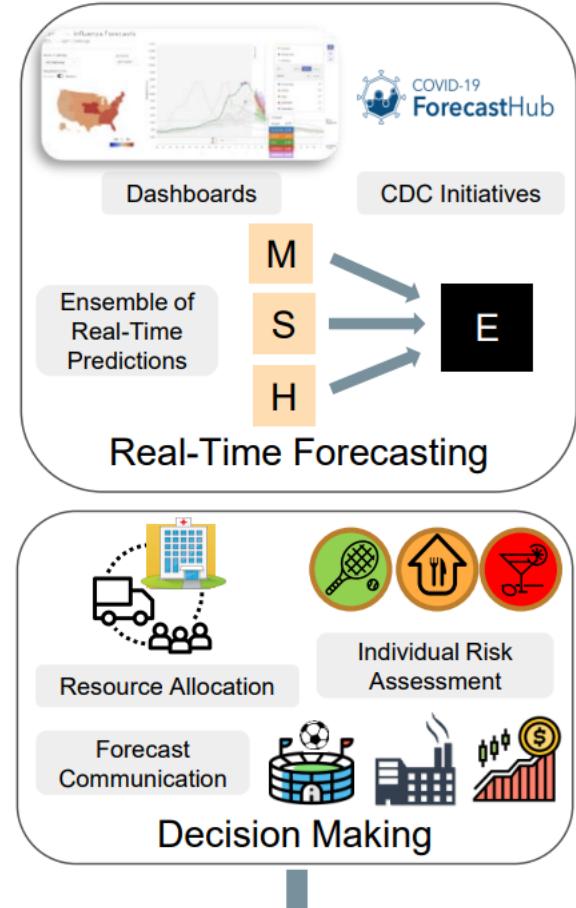
Raw data



B. Model Training & Validation



C. Utilization & Decision Making

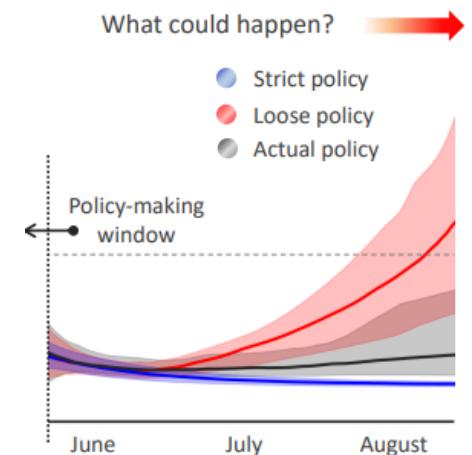


Epidemic Forecasting Setting

1. Forecasting Tasks
2. Targets of interest
3. Spatial and temporal scales
4. Real-time setup
5. Model evaluation
6. Datasets

Preliminaries: Projections vs Predictions

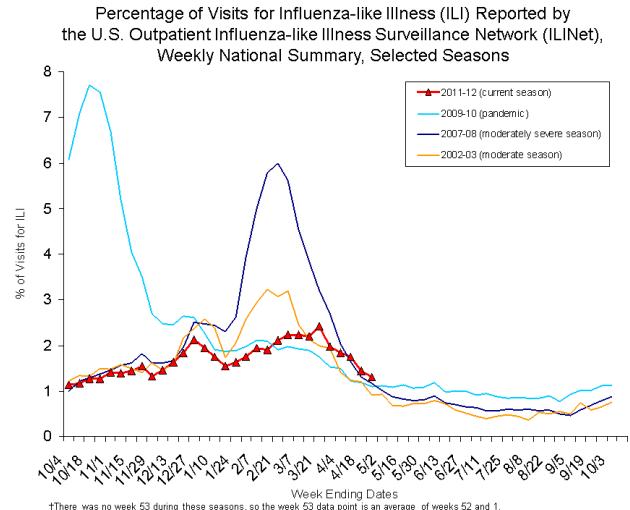
- Projections:
 - Outcomes for **specific scenario**. E.g.: Mask Mandates, Lockdowns
 - Require mechanistic assumptions/domain knowledge
- Predictions:
 - **Most likely outcome** based on past data.
- This tutorial: Mostly focus on predictions
 - But models from projections can be extended to predictions



Courtesy of [Qiann+ NeurIPS 2020]

Preliminaries: Prediction Seasons

- Fixed periods where predictions are gathered
- Typically coincides with prevalence of disease
- E.g.: CDC ILI Predictions for Flu
 - Week 40 of start year to Week 20 of next year (Aug-Apr)



Epidemic Forecasting Setting

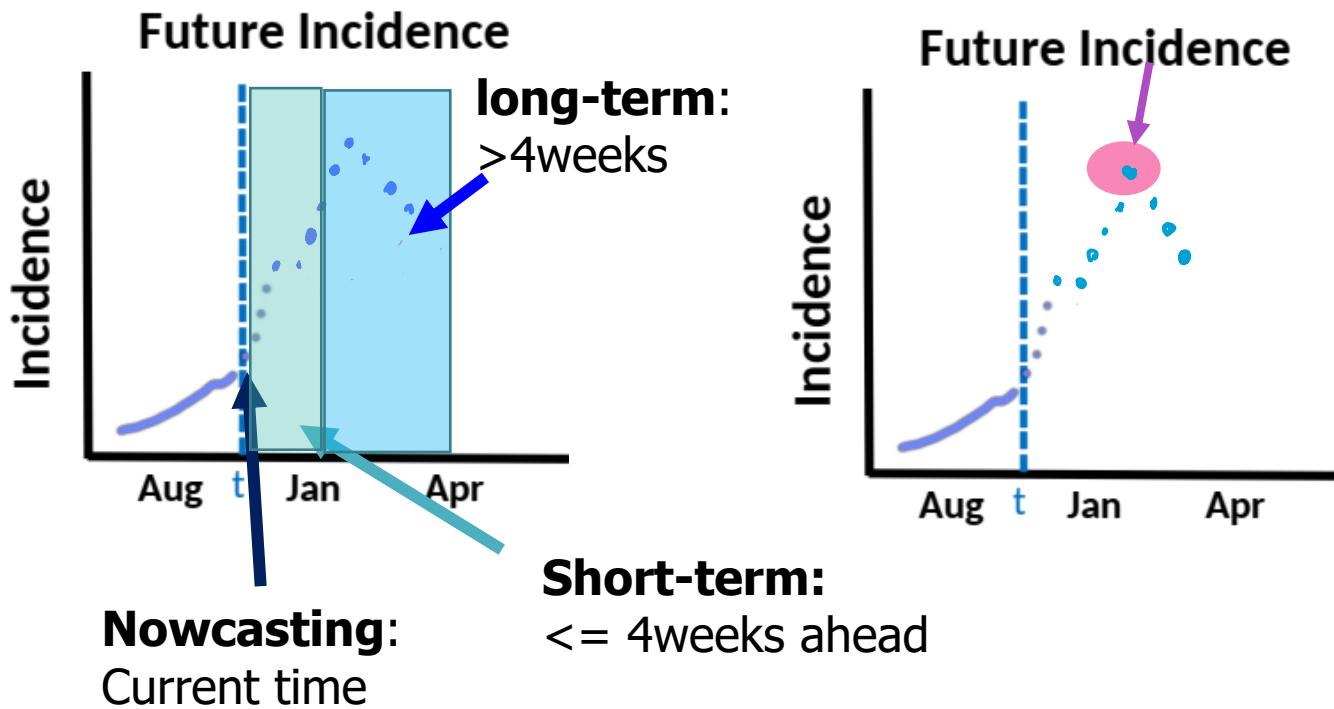
- 1. Forecasting Tasks**
2. Targets of interest
3. Spatial and temporal scales
4. Real-time setup
5. Model evaluation
6. Datasets

Different forecasting tasks

- We identified three categories of tasks
 - Real-valued predictions
 - Event-based predictions
 - Epidemiological indicator predictions

[1.1] Real-valued predictions

- **Future incidence:** Future values of indicators
- **Peak Intensity:** Max value through full season

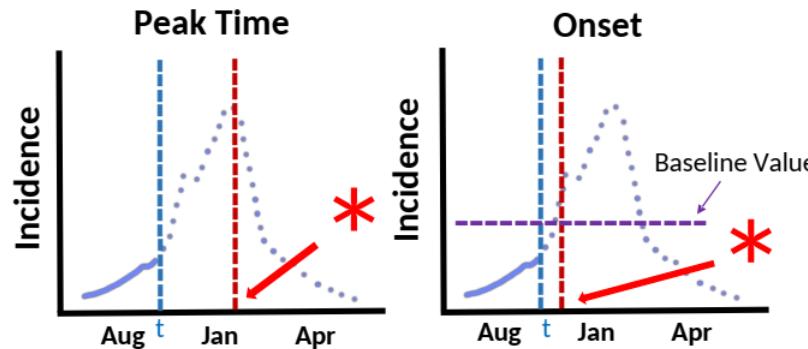


Why predicting the current value (nowcasting)?

- Delays in data reporting
 - Right truncation: Many datasets updated up to 1-3 weeks in past
- Usefulness:
 - Predict current targets from past (**prediction**)
 - Widely used in economics [Reichlin+ OECD 2019, Varian+ SSRN 2010]

[1.2] Event-based predictions

- **Peak-time:** time when peak value is observed
- **Onset:** time when indicator first increase above baseline
 - Baseline: decided by forecast organizers
 - E.g.: CDC set baseline (for each region) as average incidence value during non-flu season from past 3 years



[1.3] Epidemiological Indicators

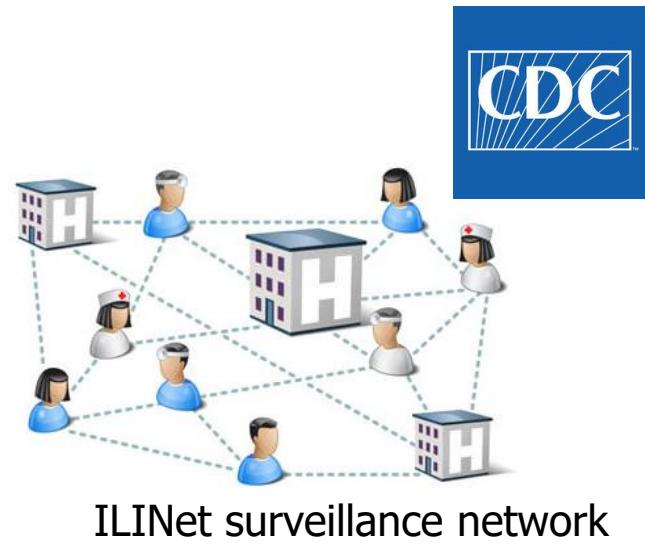
- Predict widely used epidemiological indicators that characterize the behavior of the epidemic
- Examples:
 - **Reproduction number:** Expected no. of secondary infections caused by one infected individual
 - **Final size:** Total fraction of population that will be infected over course of epidemic.

Epidemic Forecasting Setting

1. Forecasting Tasks
2. **Targets of interest**
3. Spatial and temporal scales
4. Real-time setup
5. Model evaluation
6. Datasets

[2] Targets of Interest

- Important indicators:
 - Cases (e.g., West Nile virus)
 - Mortality
 - Hospitalizations
- Influenza
 - %ILI: symptomatic outpatients
 - Syndromic surveillance
 - Lab-tested hospitalizations
- COVID-19
 - Reported deaths, hospitalizations, cases

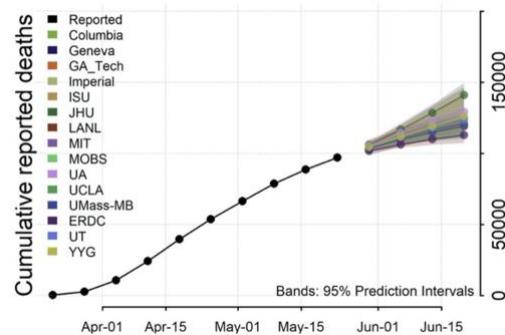
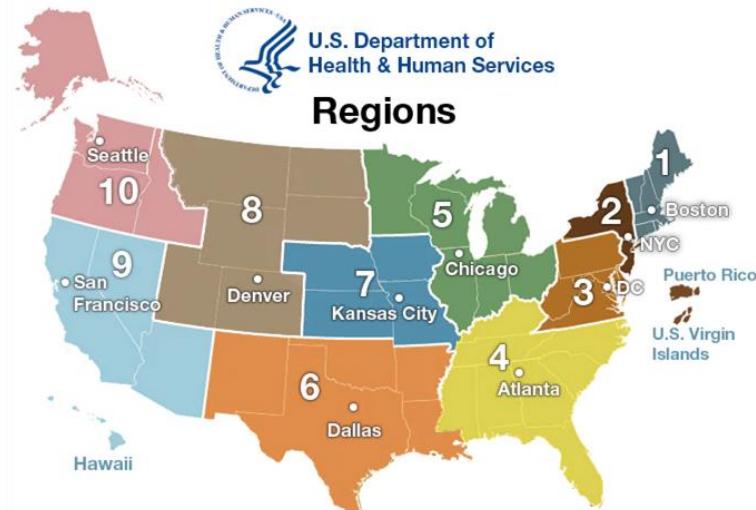


Epidemic Forecasting Setting

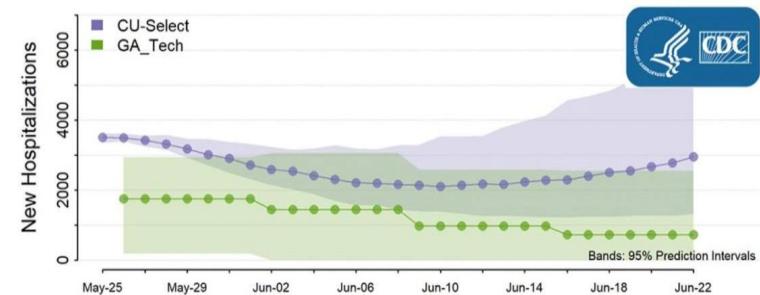
1. Forecasting Tasks
2. Targets of interest
- 3. Spatial and temporal scales**
4. Real-time setup
5. Model evaluation
6. Datasets

[3] Spatial and Temporal Scales

- Spatial scales:
 - National
 - Region/state/province
 - County/city (less common)
- Temporal scales:
 - Weekly
 - Daily



National Forecasts



Epidemic Forecasting Setting

1. Forecasting Tasks
2. Targets of interest
3. Spatial and temporal scales
- 4. Real-time setup**
5. Model evaluation
6. Datasets

[4] Real-time Forecasting Setup

- Use data till current week t to train the model
- **Input:** Prediction weeks W , forecasting horizon K in weeks, time series of features X_t until time t , time series of target Y_t until time t .

FOR w in W : // for each prediction week

 FOR k in K : // for each week ahead

1. Pre-process data X_t and Y_t
2. Train model M with gradient-based optimization
3. Forecast target with M

 ENDFOR

ENDFOR

[4] Real-time Forecasting Setup (Contd.)

- We process $X_t = \{x_1, x_2, \dots, x_t\}$ in time series windows of same length L

Input sequence of length L	Output/target	
$[x_1, \dots, x_{L-1}, x_L]$	y_{L+k}	
$[x_2, \dots, x_L, x_{L+1}]$	y_{L+1+k}	
...	...	
$[x_{t-L-k}, \dots, x_{t-1-k}, x_{t-k}]$	y_t	
...	...	
$[x_{t-L}, \dots, x_{t-1}, x_t]$	Not available	Test set

Note: If x_i can be only the target value (i.e. $x_i \in \mathbb{R}$) or a number d of features (i.e. $x_i \in \mathbb{R}^d$)

Epidemic Forecasting Setting

1. Forecasting Tasks
2. Targets of interest
3. Spatial and temporal scales
4. Real-time setup
- 5. Model evaluation**
6. Datasets

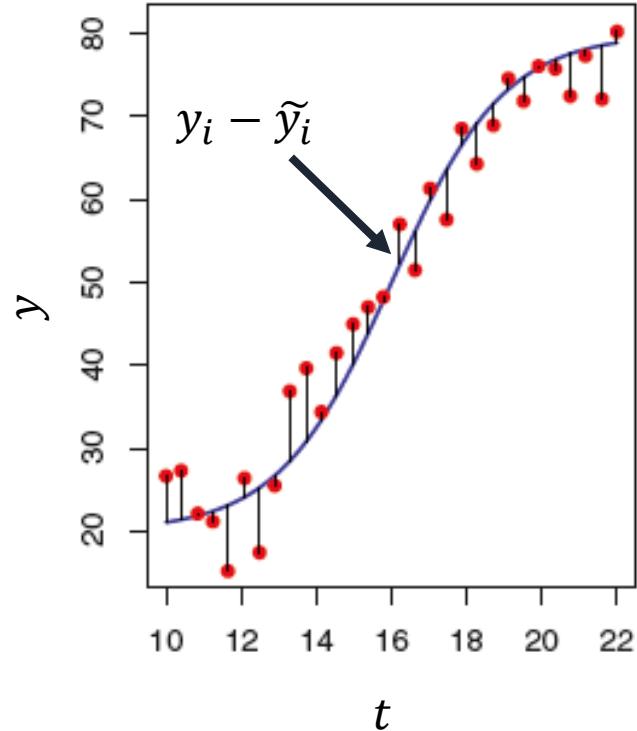
[5] Evaluation: success metrics

- 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, ...



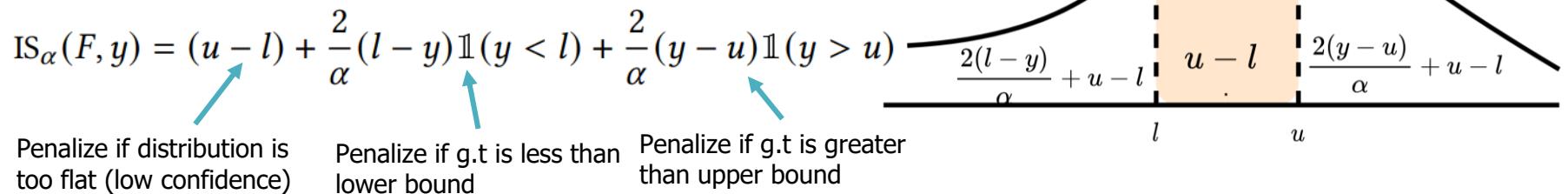
Eval. of probabilistic forecasts (1)

- Consider both accuracy and uncertainty of distributions/confidence intervals
- Log Score:
 - Log probability of ground truth outcome (binned)
$$\frac{1}{T} \sum_{i=1}^T \ln(p_i(y_i))$$
 - Each term clipped at -10 for stability and interpretability

Eval. of probabilistic forecasts (2)

- Interval Score

- Penalize for how far ground truth (g.t) is farther from α confidence intervals



- Weighted interval Score [Bracher+ 2021]

- Aggregates for multiple α

$$\text{WIS}_{\alpha_{\{0:K\}}}(F, y) = \frac{1}{K + 1/2} \times |y - m| + \sum_{k=1}^K \{w_k \times \text{IS}_{\alpha_k}(F, y)\}$$

$[l, u]$ cover $1-\alpha$ interval around the mean

Eval. of probabilistic forecasts (3)

- Other metrics
 - Coverage Score: fraction of g.t covering a confidence interval
- Probabilistic measures from general probabilistic forecasting literature
 - CRPS, Quantile loss,... [Gneiting+ RSA 2014]

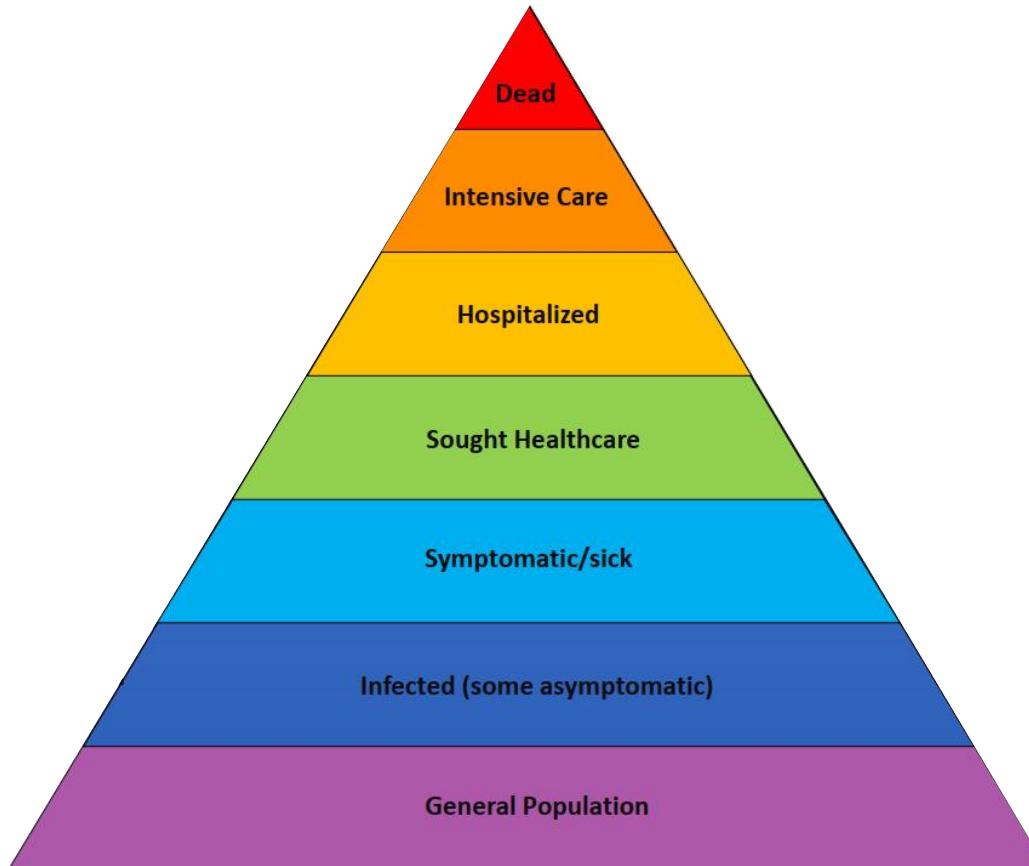
How to choose eval. metrics?

- Based on decision making
 - Uncertainty and accuracy are both 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

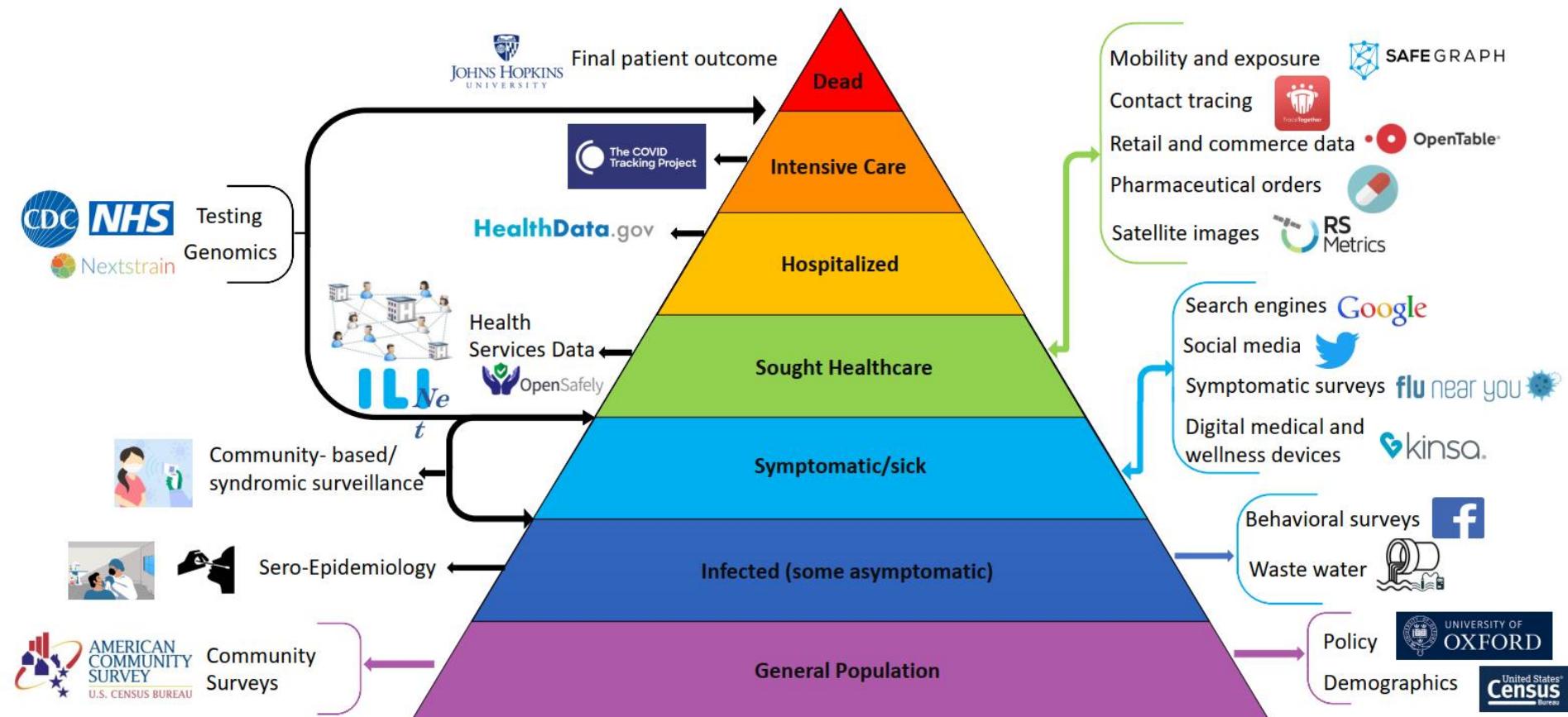
Epidemic Forecasting Setting

1. Forecasting Tasks
2. Targets of interest
3. Spatial and temporal scales
4. Real-time setup
5. Model evaluation
- 6. Datasets**

[6] Datasets: Disease surveillance pyramid



Surveillance pyramid and datasets



Sources of Data

- Clinical Surveillance
 1. Line List
 2. Health Service Records
 3. Electronic Health Records (EHR)
- Digital Surveillance
 4. Social media, search engines
 5. Online surveys
 6. Mobility and contact tracing
- Novel data sources
 7. Satellite Images
 8. Genomics
 9. Environmental

(1) Line-list data

- Who, when and where a person was infected



Hospital records



Lab surveys



Population surveys

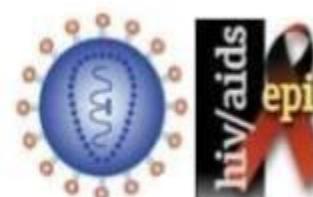
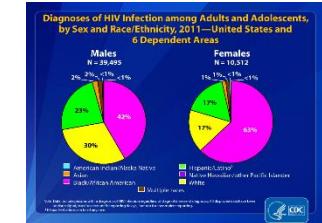
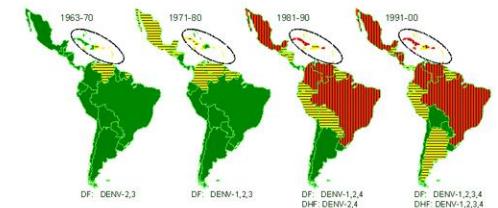
NHS



CDC

CENTERS FOR DISEASE
CONTROL AND PREVENTION

Surveillance
Reports



surveillance +
epidemiology



(1) Line-list data (contd.)

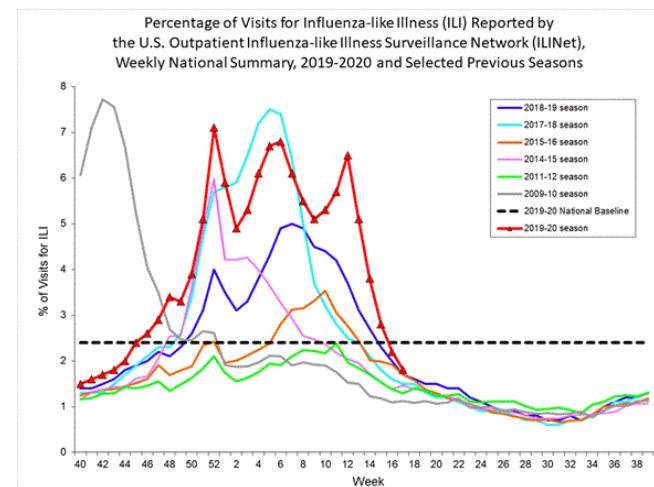
- COVID-19:
 - Disease burden: confirmed cases, hospitalizations, deaths
 - Public concern: negative tests
- Other infectious diseases:
 - Reported to the CDC's National Notifiable Disease Surveillance System (NDSS)
 - Ex: Tuberculosis, Dengue, Herpes, and Botulism

(2) Health services records

- Aggregate records collected from health-service providers
- Include inpatient and outpatient records
- Syndromic surveillance focuses on one or more symptoms rather than a physician-diagnosed or laboratory-confirmed disease

Example: Syndromic surveillance for Flu

- The burden of flu:
 - Millions get infected
 - Hundreds of thousands hospitalized
 - Thousands die
- Testing expensive and only performed in special cases (e.g., serious cases, hospitalized)
 - Instead, use symptomatic incidences of Influenza-Like Illnesses (ILI)
- Data collected by CDC from a sample of healthcare providers and reported at state and HHS region level



Traditional data sources

- Advantages
 - Very detailed
- Limitations
 - Biases
 - Very expensive
 - Take long time

(3) Electronic Health Records (EHR)

- Digital health records collected by healthcare providers
- Individual level information
- E.g.: OpenSafely (NHS - UK)
- Pros:
 - Temporally dense data at individual levels
 - Automatically collected
- Cons
 - Privacy



New data sources: Sky is the limit

- Data created for sharing (e.g., tweets) or not (e.g., search)
- Types of platforms
 - General purpose
 - Blogs, microblogs
 - Social networks, e.g., Facebook: not used as much
 - Media sharing platforms: YouTube, Reddit, Digg
- Domain specific
 - Review websites: RateMDs, Drugs.com
 - Patient communities: PatientsLikeMe, discussion forums
 - Group chats on Twitter

Digital epidemiology

OPEN  ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Review

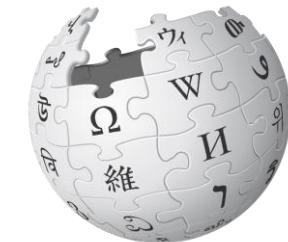
Digital Epidemiology

Marcel Salathé^{1,2*}, Linus Bengtsson³, Todd J. Bodnar^{1,2}, Devon D. Brewer⁴, John S. Brownstein⁵, Caroline Buckee⁶, Ellsworth M. Campbell^{1,2}, Ciro Cattuto⁷, Shashank Khandelwal^{1,2}, 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

(4) Digital Surveillance: Search Engines

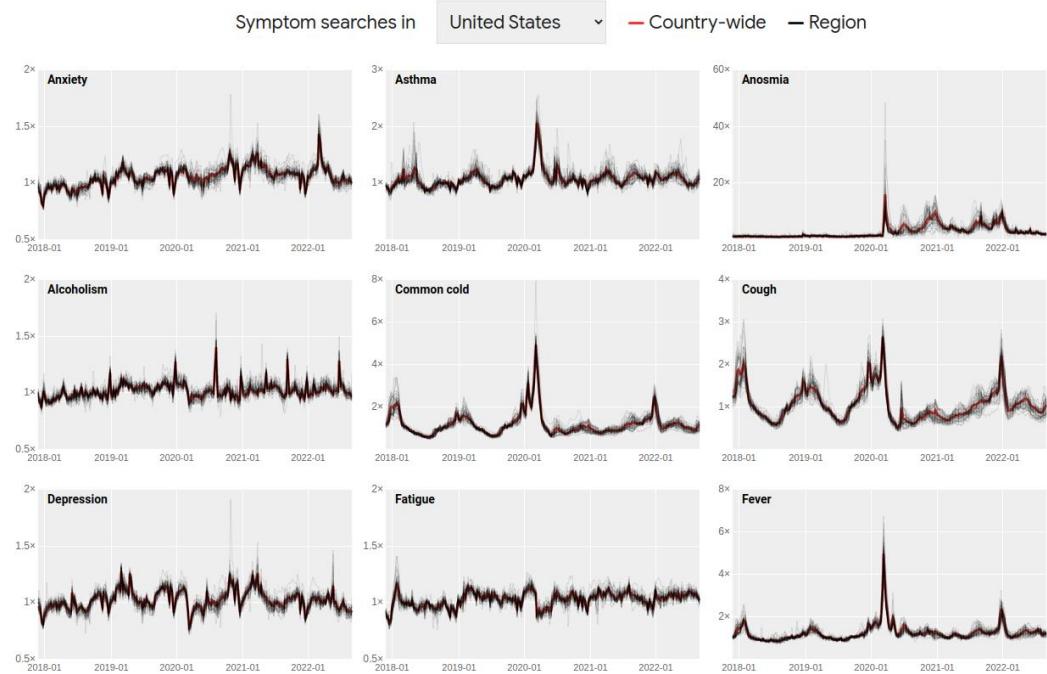
- Search activity
 - Track search volumes of specific epidemic related keywords [Polgreen+ 2008 Nature, Ginsberg+ 2009 Nature]
- Specialized Search Engines
 - UpToDate: Used by health practitioners
 - Wikipedia



Example: Google Symptom Search Trends

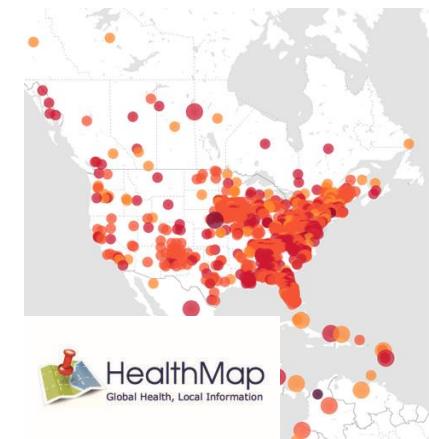
[Bavadekar+ arXiv 2020]

- Search volume per symptom, for ~500 top symptoms
 - Since 2017
 - Spatial resolution: US county & state levels
 - Temporal resolution: Daily OR Weekly



(5) Digital Surveillance: Social Media

- News, Opinions, Tweets, Blogs, etc.
- Twitter
 - Track tweets with keywords [Cullotta+ 2008]
- Health-specific Social media
 - E.g.: HealthMap: RSS feed of health-related contents.



(6) Digital Surveillance: Online Surveys

- Symptomatic surveys
- Examples
 - FluNearYou (US)
 - Dengue na Web (Brazil)

The screenshot shows a mobile application interface for 'flu near you'. At the top, there's a header bar with icons for signal strength, battery, and time (10:24). Below the header, the app logo 'flu near you' is displayed, featuring a blue sun-like icon with a white dot.

The main content area is titled 'Select Symptoms' in bold blue text. A message below it says 'Thanks! Report for Monday, August 18 through Sunday, August 24.' In the 'Last week, I experienced:' section, there are two columns of symptoms with checkboxes:

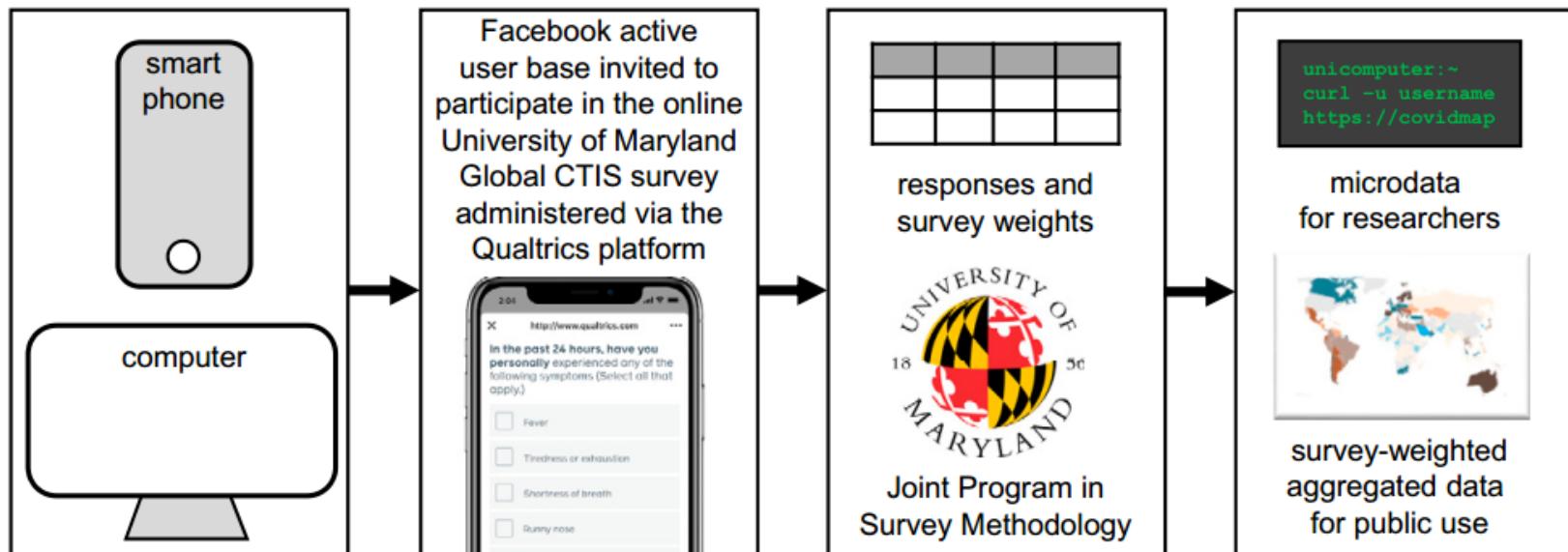
Symptom	Symptom
<input type="checkbox"/> Fever	<input type="checkbox"/> Fatigue
<input type="checkbox"/> Cough	<input type="checkbox"/> Nausea
<input type="checkbox"/> Sore throat	<input type="checkbox"/> Diarrhea
<input type="checkbox"/> Short breath	<input type="checkbox"/> Body aches
<input type="checkbox"/> Chills	<input checked="" type="checkbox"/> Headache

Below this, a question asks 'Did you receive the flu vaccine after July 31, 2013?' with three radio button options: 'Yes', 'No', and 'Don't know'. A large blue 'Submit' button is at the bottom of the form. The bottom of the screen has standard Android navigation icons for back, home, and recent apps.

Example: Facebook Survey

[Astley+ PNAS 2021]

- Request FB users to participate in a survey
 - Survey of symptoms + behavior + accessibility
- Adjust time series to improve representativeness
 - Weights adjust for sampling bias, nonresponse bias, and country/territory-level demographics



Pros/cons of digital surveillance

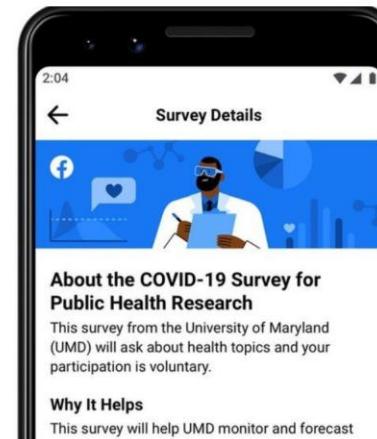
- Pros:
 - Easy to track at fine-grained spatial and temporal
 - Large population sample, diverse features
- Cons:
 - Spurious correlations
 - Varying participation across time or regions
 - Susceptible to misinformation (social media)
 - Non-uniform demographic representation

(7) Behavioral Data: Digital Surveys

- Internet social media/
Phone-based surveys
- Examples
 - Adoption of public health recommendations
 - Mask wearing
 - Social distance

Coronavirus: Facebook launches UK Covid-19 symptom survey

22 April 2020 · 0 Comments



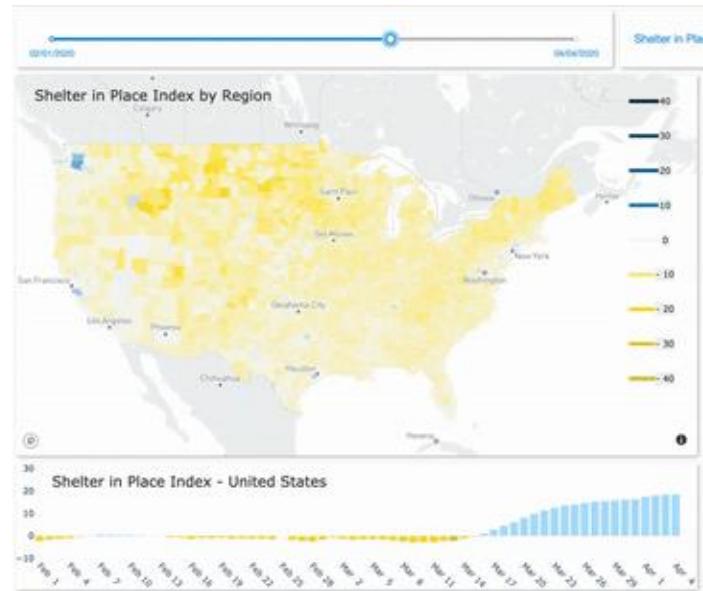
(8) Behavioral Data: Mobility

- Quantify movements within and across communities
- Sources:
 - Mobile call records
 - GPS location
 - Google mobility, SafeGraph
 - Travel data



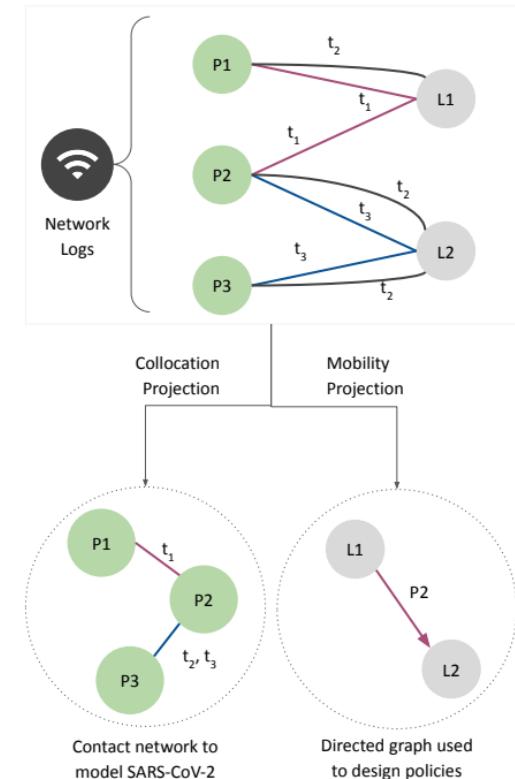
Examples in Covid-19

- Google:
 - Trends of commute to point of interest (POI)
 - Based on Google Maps
- Safegraph:
 - Visitors per hour in POI; dwell time; proportion of people staying at home
 - Based on multiple cellphone apps



(9) Behavioral Data: Contact Tracing

- Track spread of infections among individuals via proximal contact
- Build contact networks based on
 - Bluetooth, GPS
- WiFi logs to detect colocation of individuals [Swain+ 2021]



Mobility and Contact Tracing: Pros & Cons

- Pros:
 - Covers large demographics
 - Large-scale movements
- Cons:
 - Privacy, security risks
 - Representativeness



Google/Apple's contact-tracing apps susceptible to digital attacks

Researchers find way to fix privacy flaw



Tatyana Woodall
Ohio State News
woodall.52@osu.edu

West Australians' highly sensitive personal data put at risk as COVID-19 contact tracing system lacks security

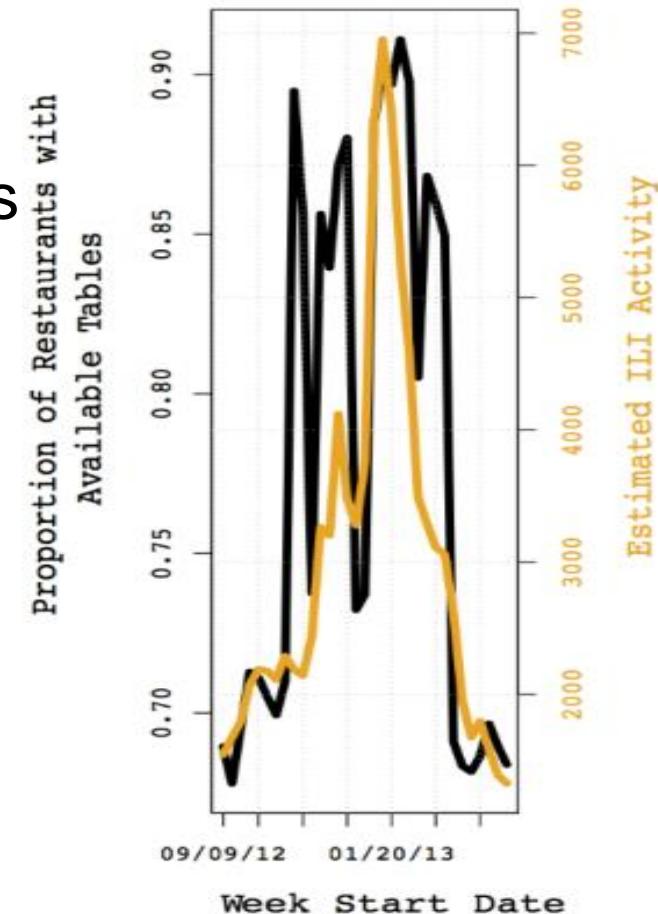
By [Herlyn Kaur](#)

Posted Wed 18 May 2022 at 7:04pm

(10) Retail and Commerce Data

[Nsoesie+ J Med Internet Res., 2014]

- Example: OpenTable
 - Daily search performed for restaurants with available tables for 2 at the hour and half past the hour for 22 distinct times
 - Increase in restaurant table availabilities was associated with an increase in disease incidence, specifically influenza-like illness (ILI)



(11) Satellite Images

[Butler+ IEEE Computing 2014, Nsoesis+ arXiv 2020]



- Pros:
 - Easy to collect at scale
- Cons:
 - Confounding factors like seasonal events, disasters

(12) Genomics



- Use pathogen genomic data to model transmission
- E.g.: Seasonal patterns, contingent environmental conditions
- Genomic Datasets
 - NextStrain: tracks pathogen genomes and mutations
 - Genomic repositories: GSAID, GenBank, COG-UK
- Pros:
 - Study past and novel epidemic spread at genomic level
 - Track mutations through to prepare for subsequent outbreaks
- Con:
 - Novel dataset with limited access

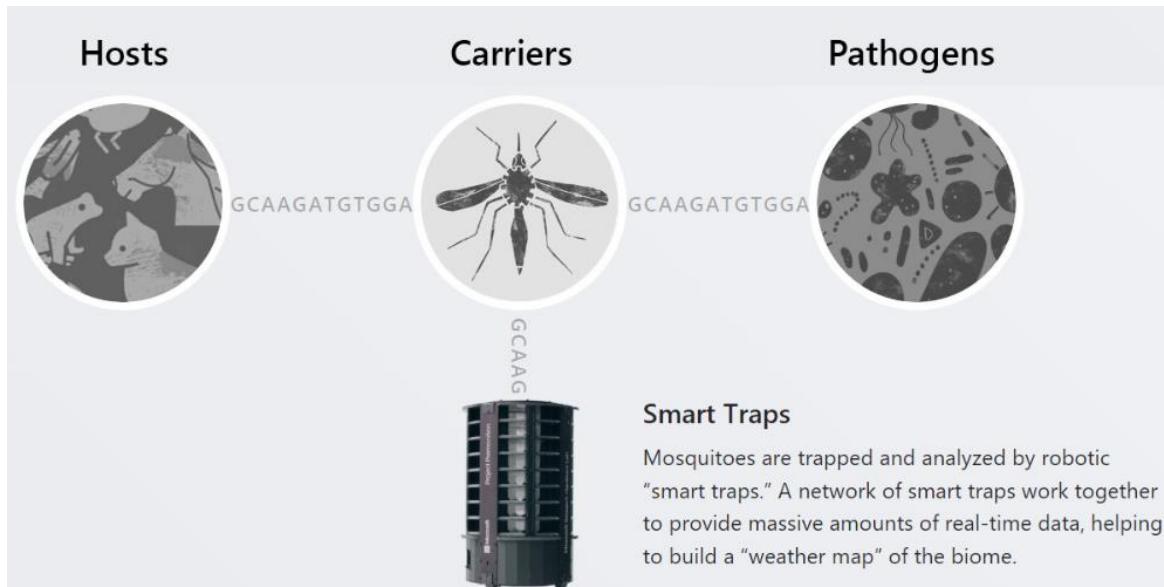
(13) Environmental Sources



- Meteorological
 - Temperature, humidity, etc. Influence transmission
- Zoonotic
 - Track diseases born in animals and transmitted to humans (E.g.: Bats for Covid-19)
 - E.g.: Microsoft Premonition project (for mosquitoes)

Example: Microsoft Premonition

- Goal: scalable **monitoring of the biome** to detect disease threats early, using robotics and genomics.



Example: Wastewater data

- Study genetic remnants (RNA) from wastewater sludge
- Useful for early detection of outbreaks [Peccia+ 2020 Nature]
 - Pros:
 - Doesn't require extensive human involvement
 - Cost-effective in long-run
 - Cons:
 - Require cutting-edge infrastructure



Epidemic Forecasting Setting

1. Forecasting Tasks
2. Targets of interest
3. Spatial and temporal scales
4. Model Evaluation
5. Datasets

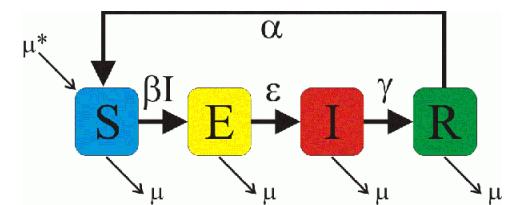
Outline

1. Epidemic forecasting: data and setup (40 min)
2. **Modeling paradigms - Overview**
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (60 min)
 - 30 min break at 3:15 PM, feel free to catch us for coffee
5. Hybrid models (45 min)
 - 5 min break
6. Epidemic forecasting in practice (25 min)
7. Open challenges and final remarks (20 min)

Part 2: Modeling Paradigms

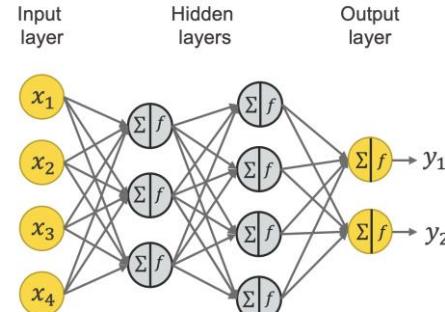
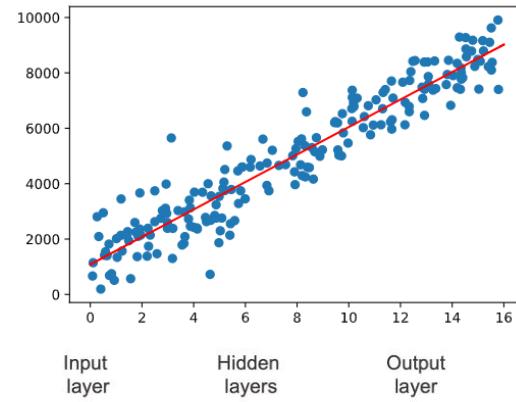
Mechanistic Models

- Encode mechanism of epidemic spread
 - Based on domain-based constraints
- 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



Statistical/ML/AI Models

- More recent data-driven models
- Leverage wide variety of large datasets
- Require lesser modelling constraints for flexible modelling
- Approaches
 - Regression-based
 - Language, Vision models
 - Neural Models
 - Density Estimation models



Hybrid Models

- Combine best of both worlds
 - Domain-based priors, expert knowledge of mechanistic models
 - Flexible modelling data-driven approach of statistical/ML methods
- Approaches
 - Statistical models estimate Mechanistic parameters
 - Mechanistic priors inform statistical models
 - Discrepancy modeling
 - Wisdom of Crowd and ensemble models

Model vs Data sources

- Stat. and Hybrid models can use recent complex large data sources from Digital Surveillance

	Clinical Surveillance	Digital Surveillance	Behavioral	Genomics	Environmental	Crowd-sourced Predictions	Policy data
Mech.	✓		✓				✓
Statistical	✓	✓	✓		✓	✓	
Hybrid	✓	✓	✓	✓	✓	✓	✓

Model vs Data sources

- Genomics, crowd-sourced predictions not widely adopted yet.

	Modelling Paradigms	Clinical Surveillance	Digital Surveillance	Behavioral	Genomics	Environmental	Crowd-sourced Predictions	Policy data
Mech.	Mass-Action Models	✓						
	Metapopulation Model	✓			✓			
	Agent-Based	✓			✓			✓
Statistical	Regression	✓	✓	✓		✓	✓	
	Vision/Language	✓	✓					
	Neural Models	✓	✓	✓		✓		
	Density Estimation	✓	✓			✓		
Hybrid	Statistical models estimate Mechanistic parameters							
	Mechanistic priors inform stat. models	✓	✓	✓	✓	✓		
	Discrepancy modelling	✓		✓				
	WoC, Ensembles	✓	✓				✓	✓

Modeling Paradigms		Data					Tasks		Model Features												
		Clinical surveillance data	Electronic surveillance data	Behavioral data	Genomics data	Environmental data	Crowd-sourced predictions	Policy data	Real-valued prediction	Event-based prediction	Epidemiological indicators	Deep learning	Geographical granularity (Cou=County, C= County, S=State, R=Region)	Temporal granularity (D=Days, W=Weeks)	Gradient-based learning	Uncertainty estimation	Handle data quality issues	Spatio-temporal modeling	Interpretability	Transfer learning	Expert in the loop
Mech	Mass-Action Models	✓							✓	✓	✓		C/S	D/W							
	Metapopulation Models	✓	✓						✓	✓			C/S/Cou/R	D/W							
	Agent-Based Models	✓	✓				✓	✓	✓				C/Cou	D/W							
Statistical	Regression Models								✓	✓											
	Sparse Linear Models	✓	✓						✓	✓			C/S	W	✓					✓	
	Auto-regressive Models	✓	✓				✓		✓	✓			S/C/Cty/R	W	✓					✓	
	Complex Regression Models	✓	✓	✓		✓			✓				R/C/S	W	✓						
	Hierarchical Models	✓	✓				✓		✓				S/C	D/W	✓	✓			✓	✓	
	Vision and Language Models									✓											
	Vision Models	✓	✓							✓			C	D	✓						
	Language-based Models	✓	✓							✓			C	D/W	✓						
	Probabilistic topic models	✓	✓							✓			C	W	✓						
	Neural Models																				
Statistical	Off the Shelf	✓	✓	✓		✓			✓			✓	C/R	W	✓						
	Similarity modeling	✓		✓					✓	✓		✓	C/R	W	✓	✓					
	Transfer Learning	✓		✓					✓			✓	C/R	D/W	✓	✓					
	Multimodal Data	✓		✓					✓			✓	C/R	D/W	✓	✓					
	Spatial Modeling	✓	✓	✓					✓			✓	C/R/S	W	✓	✓					
Density Estimation	Density Estimation									✓	✓										
	Kernel density estimation	✓							✓	✓			C/R	W		✓					
	Parametric Bayesian inference	✓							✓	✓			C/S	W		✓					
	Non-parametric methods	✓	✓			✓			✓				C	W		✓					
Mechanistic	Neural uncertainty quantification	✓							✓			✓	C/R	W	✓	✓	✓	✓	✓	✓	
	Mechanistic with Statistical Components									✓	✓			S and C	W		✓	✓			
	Data Assimilation	✓	✓		✓	✓			✓	✓											
	Statistical estimation of mechanistic parameter	✓	✓	✓		✓			✓	✓	✓	✓	C/S	D/W	✓		✓	✓	✓	✓	
	Discrepancy Modeling	✓		✓					✓	✓		✓	C/S	W	✓	✓	✓	✓	✓	✓	
Mechanistic	Mechanism informs statistical model																				
	Learning from synthetic and simulation data	✓	✓	✓		✓			✓	✓		✓	C/R and S	W	✓	✓	✓				
	Learning with mechanistic constraints	✓							✓			✓	S and C	D	✓	✓	✓		✓		
Wisdom of Crowds	Experts and prediction markets	✓					✓		✓	✓										✓	
	Ensembles	✓	✓						✓	✓		✓	Cty/S/C/R	W	✓		✓		✓	✓	

Detailed table in survey

102 methods

250+ references

Dating back to 2000

Outline

1. Epidemic forecasting: data and setup (40 min)
2. Modeling paradigms - Overview
3. **Mechanistic models** (15 min)
4. Statistical/ML/AI models (60 min)
 - 30 min break at 3:15 PM, feel free to catch us for coffee
5. Hybrid models (45 min)
 - 5 min break
6. Epidemic forecasting in practice (25 min)
7. Open challenges and final remarks (20 min)

Part 3: Mechanistic Models

Mechanistic Models

- Explicitly model the mechanisms of epidemic spread
- A lot of important work here
- Resources for 101 course on epidemiology:
 - N. Dimitrov and L. Meyers. 2010. Mathematical approaches to infectious disease prediction and control. INFORMS, 1–25
 - H. Hethcote. 2000. The mathematics of infectious diseases. SIAM review 42, 4 (2000), 599–653
 - M. Marathe and A. Vullikanti. 2013. Computational epidemiology. Commun. ACM 56, 7 (2013), 88–96.

Mechanistic Models (Outline)

- Approaches:
 1. Mass-action models
 2. Metapopulation models
 3. Agent-based models

Note: 1 and 2 are also known as compartmental models

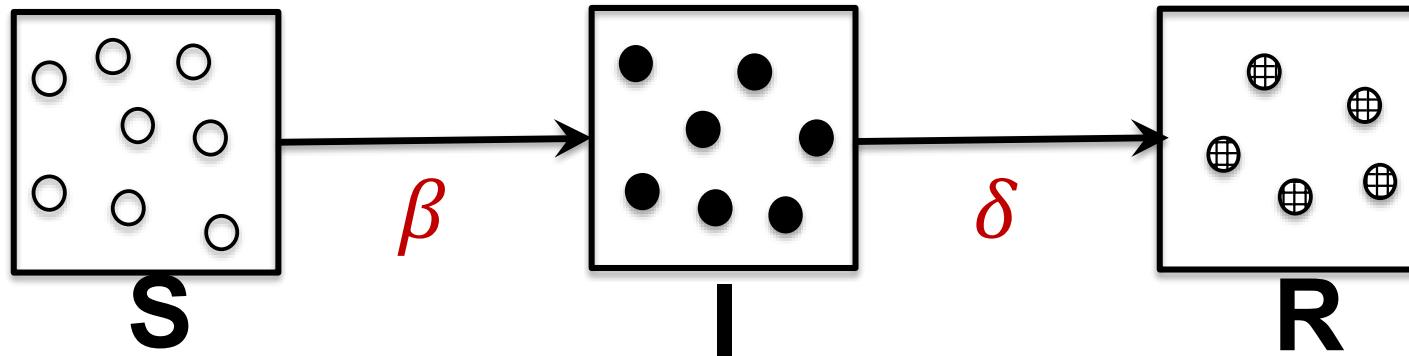
Mechanistic Models (Outline)

- Approaches:
 1. **Mass-action models**
 2. Metapopulation models
 3. Agent-based models

[M1] Mass-action models

[Hethcote, SIAM Review 2000]

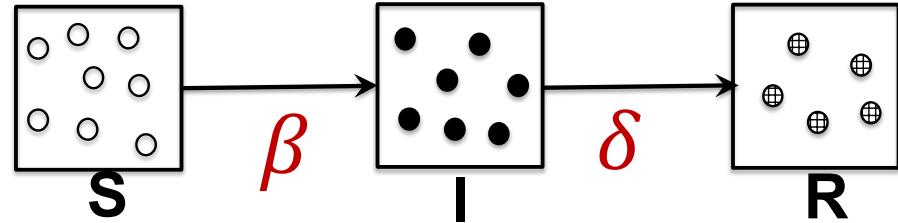
- One of the simplest models
 - Susceptible: healthy, can get infected
 - Infected: can infect others through contact
 - Recovered: cannot infect others



Assumptions

- Perfect mixing
 - Any infected person can infect any susceptible person
- No birth or deaths (no 'demography')
 - Total population is constant
- Deterministic!

SIR Model



$$\frac{dS}{dt} = -\beta SI$$

Number of new infections =

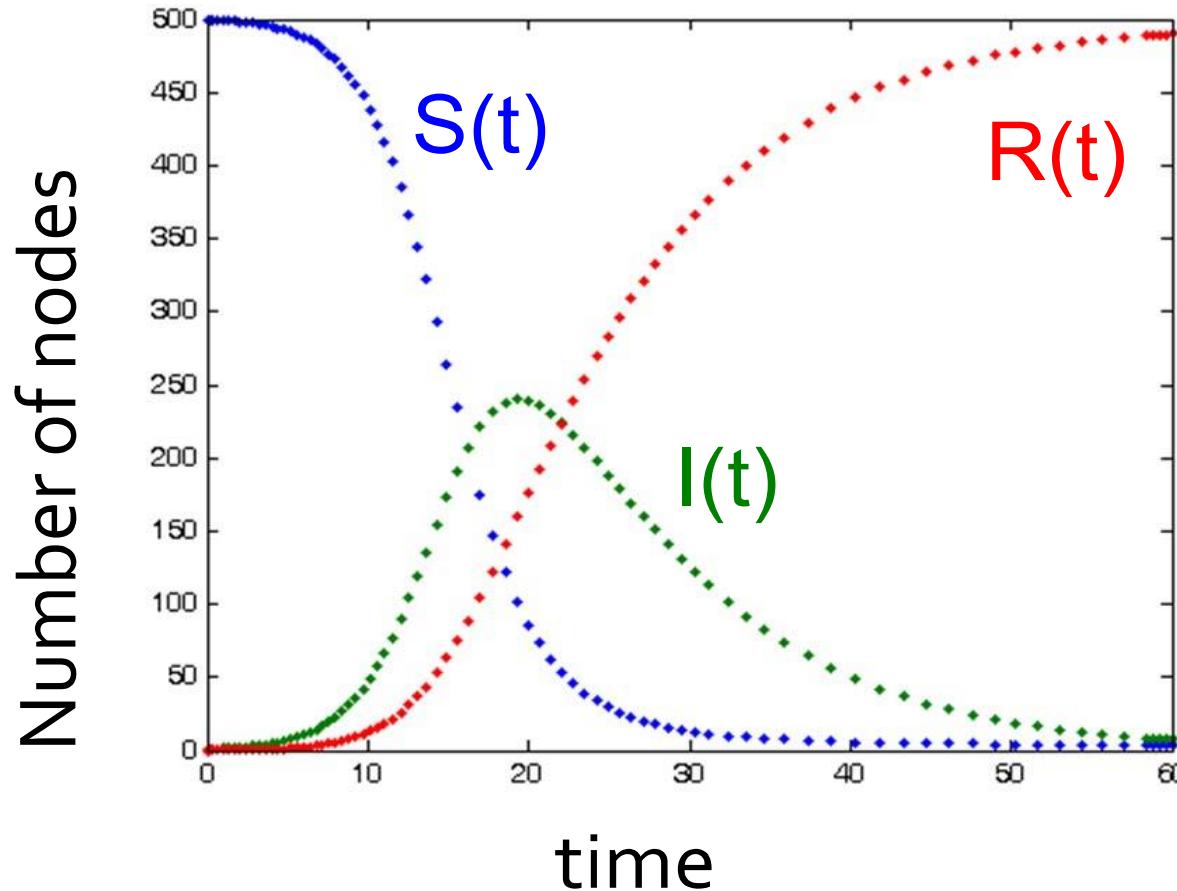
$$\frac{dI}{dt} = \underbrace{\beta SI}_{\text{Number of new infections}} - \underbrace{\delta I}_{\text{Number of infected nodes curing}}$$

$$\frac{dR}{dt} = \delta I$$

Solving SIR

- No closed form solution!

SIR: numerical output



Many many extensions

- With birth/death rates ('vital dynamics')
- Time-varying contact rates
- Make things stochastic
- Multiple viruses/diseases
-
- See Hethcote 2000, and the book by May and Anderson 1992

Threshold Phenomenon: R₀

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

- This implies

$$\frac{dI}{dt} < 0 \quad \text{if} \quad S(0) < \delta/\beta$$

Threshold Phenomenon

- So, $R_0 = \beta/\delta$
 - Basic Reproductive number: average number of secondary cases caused by one individual
- 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 $1/R_0$

Mechanistic Models (Outline)

- Approaches:
 1. Mass-action models
 2. **Metapopulation models**
 3. Agent-based models

[M2] Metapopulation Models

[Dimitrov and Meyers, INFORMS 2010]

- Considers heterogeneity of population
 - E.g., epidemic dynamics in location A \neq location B.
 - But assume homogeneity at 'right' granularities
 - One mass-action model per population
- Ex. Model heterogeneity using travel data

σ_{ij} : daily passenger flow from city i to city j

n_i : population of city i , assumed to be fixed

$X_i(t)$, $Y_i(t)$, $Z_i(t)$: number of people in S/I/R states in city i at time t

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

Similarly, Y^{eff}
and Z^{eff}

But... Human contact patterns are not random

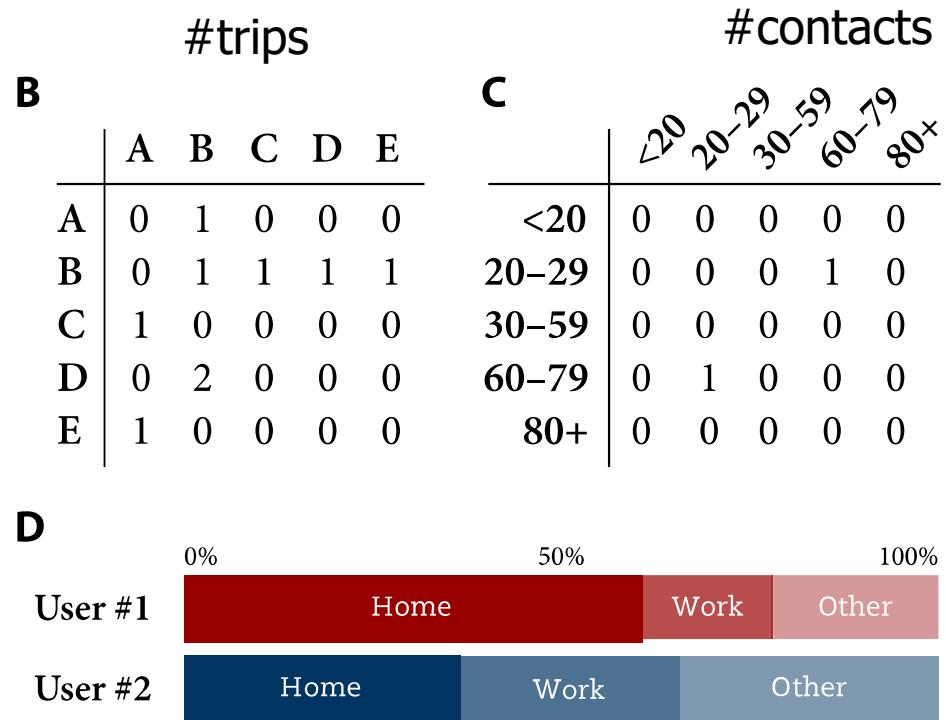
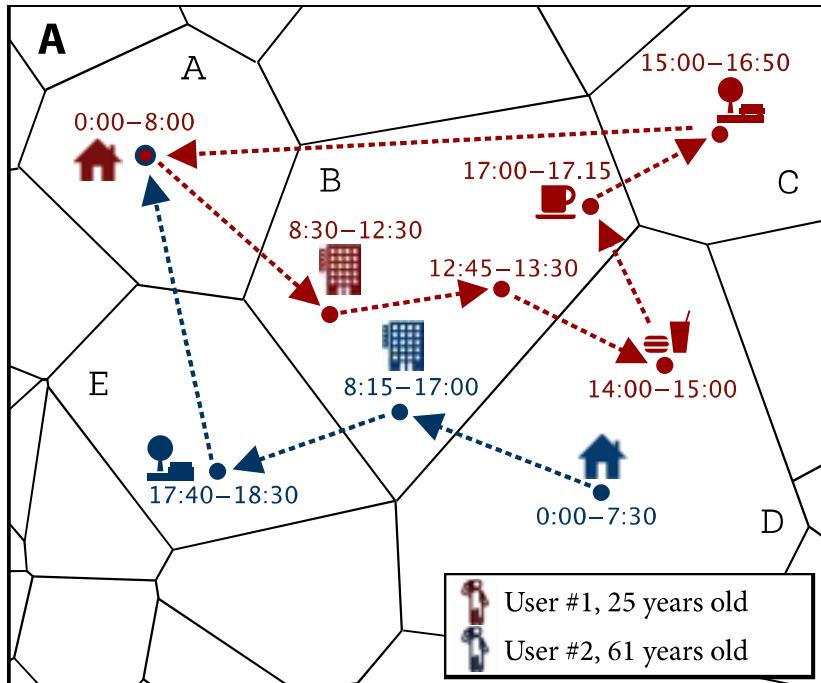


Source: Mi Jin Lee at petterhol.me

How to Capture Them?

Example: Using Call Data Records

- Many recent studies on this topic [Oliver et al, Sci. Adv. 2020]
#raw data



Numerous COVID-19 examples

- Apple (maps/directions)
- Google (location history)
- Safegraph (POI access)
- CubeIQ (mobile phones etc)
-

Mechanistic Models (Outline)

- Approaches:
 1. Mass-action models
 2. Metapopulation models
 - 3. Agent-based models**

[M3] Agent-based networked models

[Marathe and Vullikanti, CACM 2013]

- Each individual is an agent in a simulation
- Disease spread over contact networks
 - Model heterogeneous interactions between agents
- Concepts:
 - Social contact networks
 - Twin cities

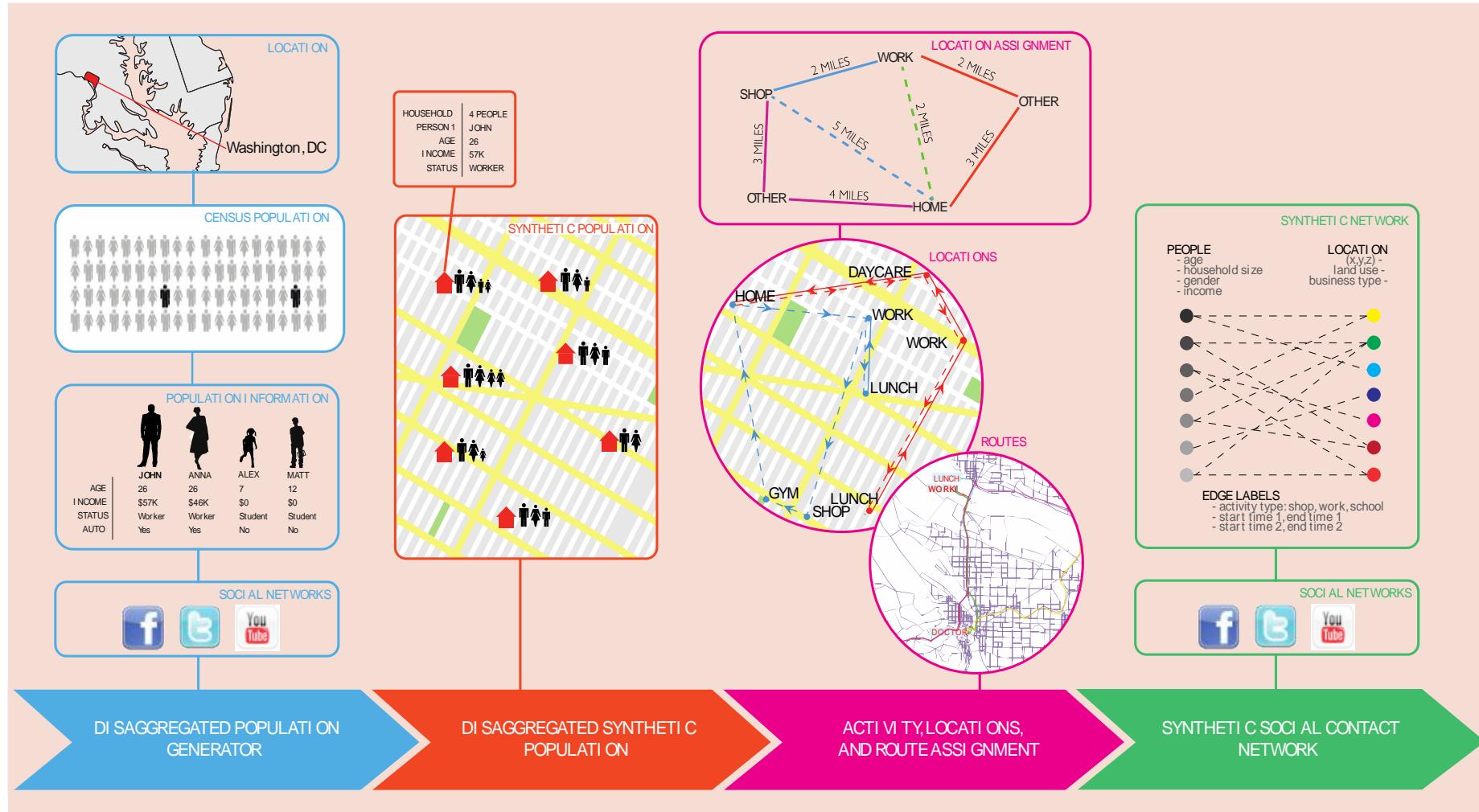


First principles Approach for Constructing Social Contact Networks

- 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

First principles Approach for Constructing Social Contact Networks

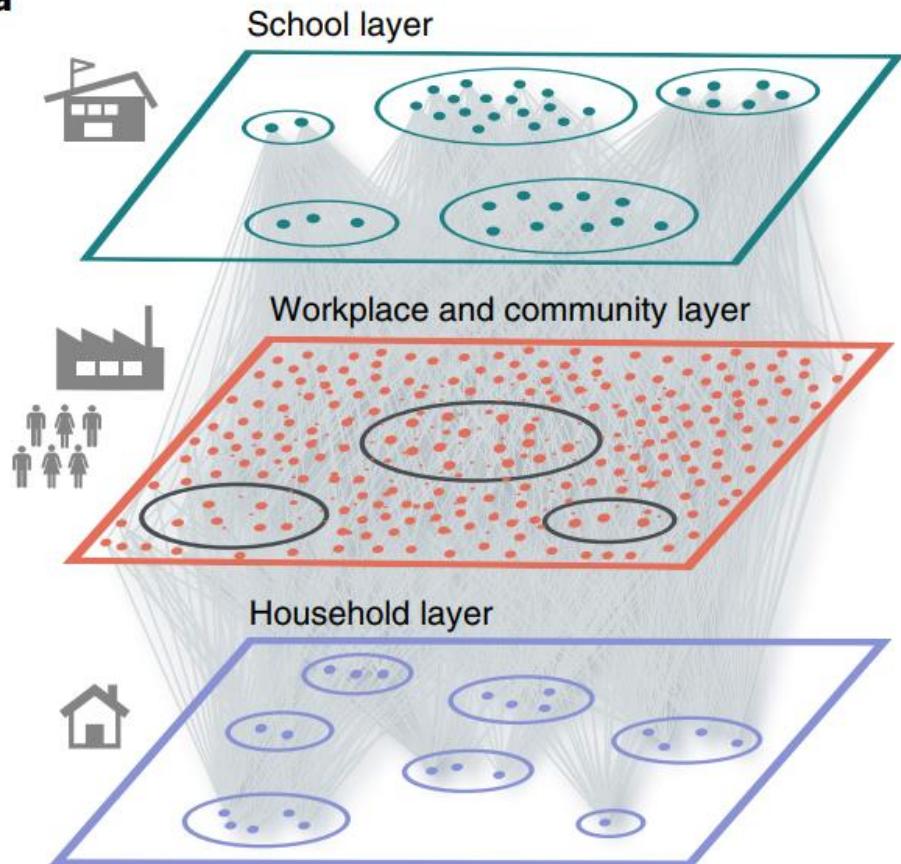
[Marathe and Vullikanti, CACM 2013]



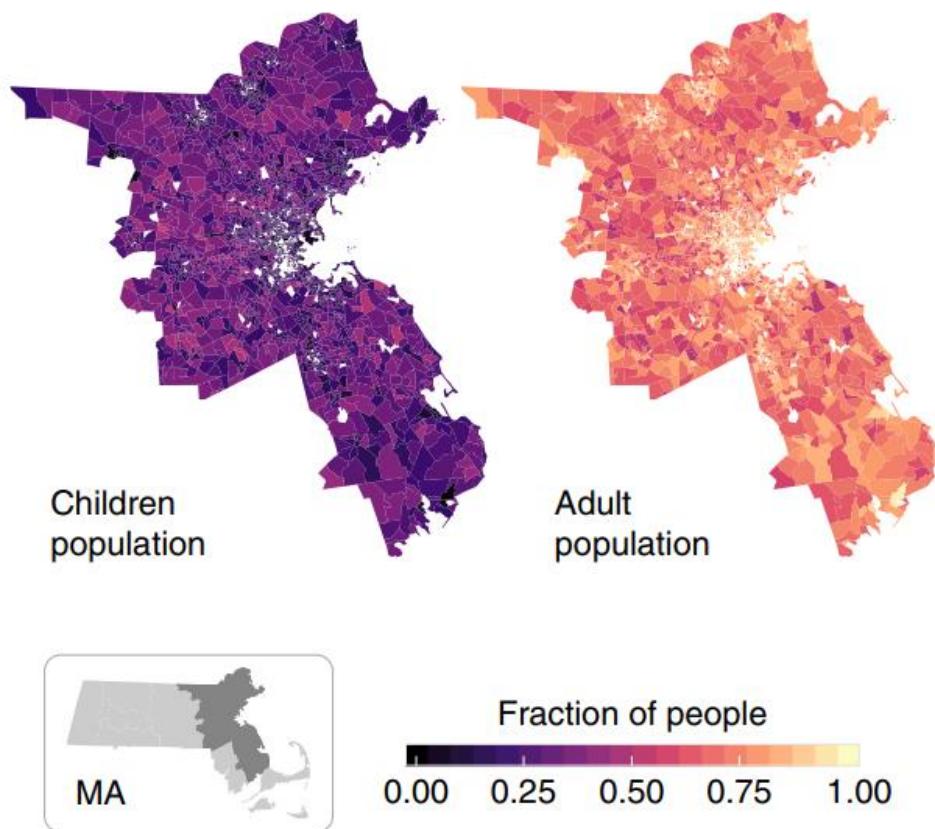
Example: COVID-19 in MA

[Aleta et al, Nature Human Behavior 2020]

a



b



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

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 R_0 (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
 - H. Hethcote. 2000. The mathematics of infectious diseases. SIAM review 42, 4 (2000), 599–653
 - M. Marathe and A. Vullikanti. 2013. Computational epidemiology. Commun. ACM 56, 7 (2013), 88–96.

Outline

1. Epidemic forecasting: data and setup (40 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. **Statistical/ML/AI models** (60 min)
 - 30 min break at 3:15 PM, feel free to catch us for coffee
5. Hybrid models (45 min)
 - 5 min break
6. Epidemic forecasting in practice (25 min)
7. Open challenges and final remarks (20 min)

Part 4: Statistical, Machine-learning/AI models

Statistical/Machine Learning 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.

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

Choose best function f from family \mathcal{H}

Prediction from function f

Ground truth

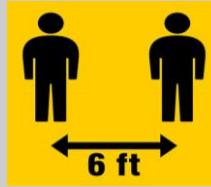
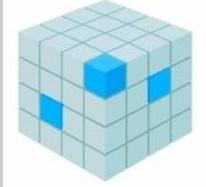
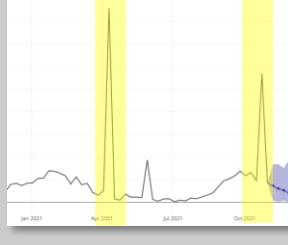
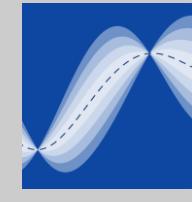
Loss function \mathcal{L}

The diagram shows the optimization equation for finding the best function f from a family \mathcal{H} . The equation is $\min_{f \in \mathcal{H}} \sum_{i=1}^T \mathcal{L}(f(x_i) - y_i)$. Four purple arrows point to different parts of the equation: one to the term $f \in \mathcal{H}$ with the label 'Choose best function f from family \mathcal{H} ', one to the term $f(x_i)$ with the label 'Prediction from function f ', one to the term y_i with the label 'Ground truth', and one to the symbol \mathcal{L} with the label 'Loss function \mathcal{L} '.

Why Stat./ML models?

- Doesn't aim to model generative mechanics of epidemics
- Can handle wide variety of datasets
 - Languages, Images, time-series, etc.
- Flexible modelling with powerful family of functions \mathcal{H} , optimization algorithms for learning underlying patterns

Overview of challenges for Stat./ML models

Aspect	DISEASE SPREAD	DATA	UTILIZATION
Challenges	    Spatial Transmission Mobility Mask adoption Social distancing	   Sparse data Data revisions Anomalies	   Interpretability Uncertainty quantification Actionable forecasts

Statistical, ML/AI Models (Outline)

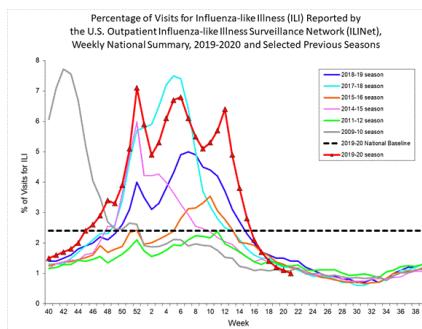
- Approaches:
 1. Regression Models
 2. Language and Vision Models
 3. Neural Models
 4. Density Estimation

Statistical, ML/AI Models (Outline)

- Approaches:
 1. **Regression Models**
 2. Language and Vision Models
 3. Neural Models
 4. Density Estimation

[S1] Regression Models

- Assume a linear relationship between input features and future forecast $\tilde{y} = w_0 + \mathbf{w}^T \mathbf{x}$
- The features \mathbf{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, features from satellite images, etc.



Idea 1: AutoRegressive Models

- Use past values of epidemic cures as features to predict future values
- E.g.:

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

Future target to predict Past values of epi curve Noise

- $\{\phi_j\}_{j=0}^p$ are parameters to learn

Ex. 1: Google Flu Trends

[Ginsberg+ Nature 2009]

- Simple linear model for nowcasting ILI
- Use search logits of query fractions as features
$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \epsilon$$
 - P = ILI (physician visits)
 - Q = Fraction of search queries that are ILI-related



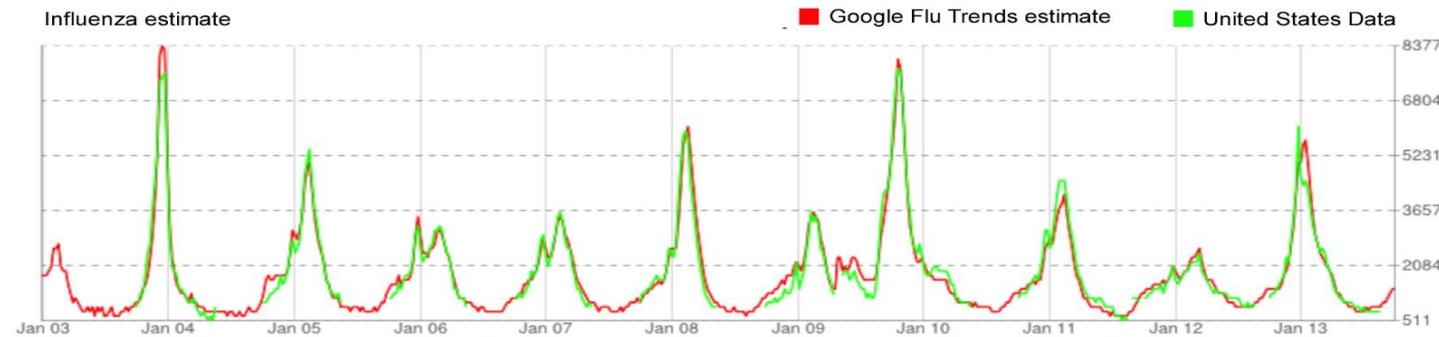
The figure is a screenshot of a New York Times article. The header reads 'The New York Times'. The main title is 'Google Uses Searches to Track Flu's Spread' by Miguel Heftt, published on Nov. 11, 2008. Below the title are social sharing icons for Facebook, Twitter, and Email. The text discusses how Google tracks flu outbreaks through search queries.

The figure is a screenshot of a CNN news article. The headline is 'Google tool uses search terms to detect flu outbreaks' by Elizabeth Landau, CNN, updated at 6:51 p.m. EST, Tuesday December 9, 2008. Below the headline are buttons for 'READ', 'VIDEO', and 'QUIZ'.

CNN 2008

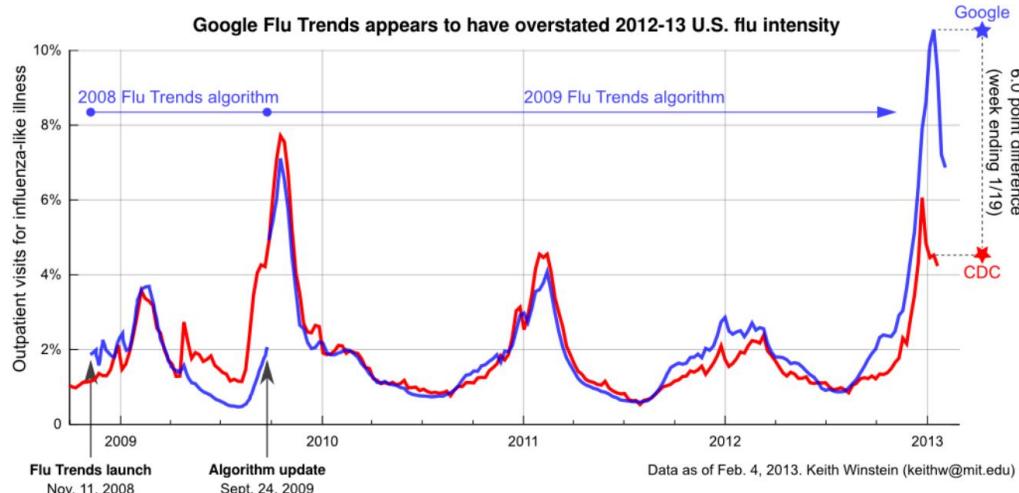
Google Flu Trends (Contd.)

- ILI (Flu) -related queries
 - Automated selection based on proprietary set of keywords
- Effective up to 2009 H1N1 pandemic
 - 0.94 PCC with CDC ILI data [Ortiz+ PLoS 2011]



However,...

- Didn't capture changing trends in keyword correlates, i.e. didn't handle data drift
 - Failed to capture H1N1 pandemic [Olson+ PLoS Comp. Bio 2013]



Sources: <http://www.google.org/flutrends/us>, CDC ILInet data from <http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>, Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic.

 Georgia Tech

Rodríguez, Kamarthi, and Prakash 2023

The Conversation

Academic rigor, journalistic flair

COVID-19 Arts + Culture Economy Education Environment + Energy Ethics + Religion Health Politics + Society

Patterns and proofs

Taking a mathematical look at life

Google's flu fail shows the problem with big data

POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,¹ Ryan Kennedy,¹ King Kong,² Alessandro Vespignani^{1,3,4}

¹ Florida Center for Medical Image Computing, University of Florida; ² Department of Mathematics, Northeastern University; ³ Department of Physics, Northeastern University; ⁴ Department of Earth and Atmospheric Sciences, University of Colorado Boulder

February 13, 2013 Google Flu Trends (GFT) was born. It was not for a reason that Google imagined it would be useful. A tracking system would have been developed by CDC to predict the percentage of people more than doubles the previous year who had sought medical attention like those (II) that the Centers for Disease Control and Prevention (CDC), which has its estimation system running in laboratories across the United States (I). The system was built so that GFT was built to predict CDC reports given that I had no data. It is an example of big data (3, 4) as well as what lessons we can draw from it.

The problems we identify are not limited to Google Flu. They are in whether search and social media can predict disease outbreaks with mathematical precision and hypotheses. Although these studies have shown some promise, they can be misleading. They can suggest more traditional methods are unnecessary. They also issue that contributed to GFT's big data success and offer lessons for moving forward in the future.

ability and dependencies among data (1,2). The core challenge is that most big data that have received popular attention are not the result of carefully designed experiments that yield valid and reliable data amenable for scientific analysis.

The initial version of GFT was a prediction system that used Google search data as input data. Essentially, the methodology involved fitting a mathematical model to historical Google search data and then using this model to predict the number of Google users who had influenza-like illness (ILI) symptoms. The predictions were based on a linear regression model that assumed a constant rate of growth in the number of Google users who had ILI symptoms over time. This model was updated weekly as new data became available, the comparative value of the algorithm was assessed, and the results were published online. A study in 2010 demonstrated that GFT's accuracy was comparable to that of a fairly simple computer program, forward using a mathematical model to predict the number of Google users who had ILI symptoms (5). This model significantly outperformed GFT in terms of accuracy and reliability.

Although the initial version of GFT was updated weekly, the comparative value of the algorithm was assessed, and the results were published online. A study in 2010 demonstrated that GFT's accuracy was comparable to that of a fairly simple computer program, forward using a mathematical model to predict the number of Google users who had ILI symptoms (5). This model significantly outperformed GFT in terms of accuracy and reliability.

126

The Final Straw

[Lazer+ Science 2014]

- Search algorithm continually being modified
- Additional search term suggestions Lack of transparency
- Big data ‘hubris’
 - *For the two years ending Sep 2013, Google’s estimates were high in 100 out of 108 weeks. After Oct 2013 update, Google’s estimates are over by 30% for 2013–2014 season*

POLICYFORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,^{1,*} Ryan Kennedy,^{1,4} Gary King,² Alessandro Vespignani^{1,4,3}

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracker should have been hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories and medical facilities (1, 2). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (3, 4), what lessons can we draw from this episode?

The problems we identify are not limited to GFT. Research on whether search or social media can predict x has become commonplace (5–7) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these instruments from a perspective where they can supplant more traditional methods or theories (8). We explore two issues that contributed to GFT’s mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.



Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

run ever since, with a few changes announced in October 2013 (10, 15).

Although not widely reported until 2013, how GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are avoidable (16–18). For example, last week’s errors predict this week’s errors (temporal auto-correlation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (4). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data

Ex. 2: ARGO

[Yang+ Sci. Reports 2017]

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

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

Past ILI Query volume

influenza.type.a	painful.cough
flu.incubation	fever.flu
bronchitis	over.the.counter.flu
influenza.contagious	pneumonia
flu.fever	how.long.is.the.flu
influenza.a	flu.how.long
influenza.incubation	treatment.for.flu
flu.contagious	fever.cough
treating.the.flu	flu.medicine
type.a.influenza	dangerous.fever
symptoms.of.the.flu	high.fever
influenza.symptoms	is.flu.contagious
flu.duration	normal.body
flu.report	normal.body.temperature

Examples of search terms used

ARGO2 (Extension)

[Ning+ Sci. Reports 2019]

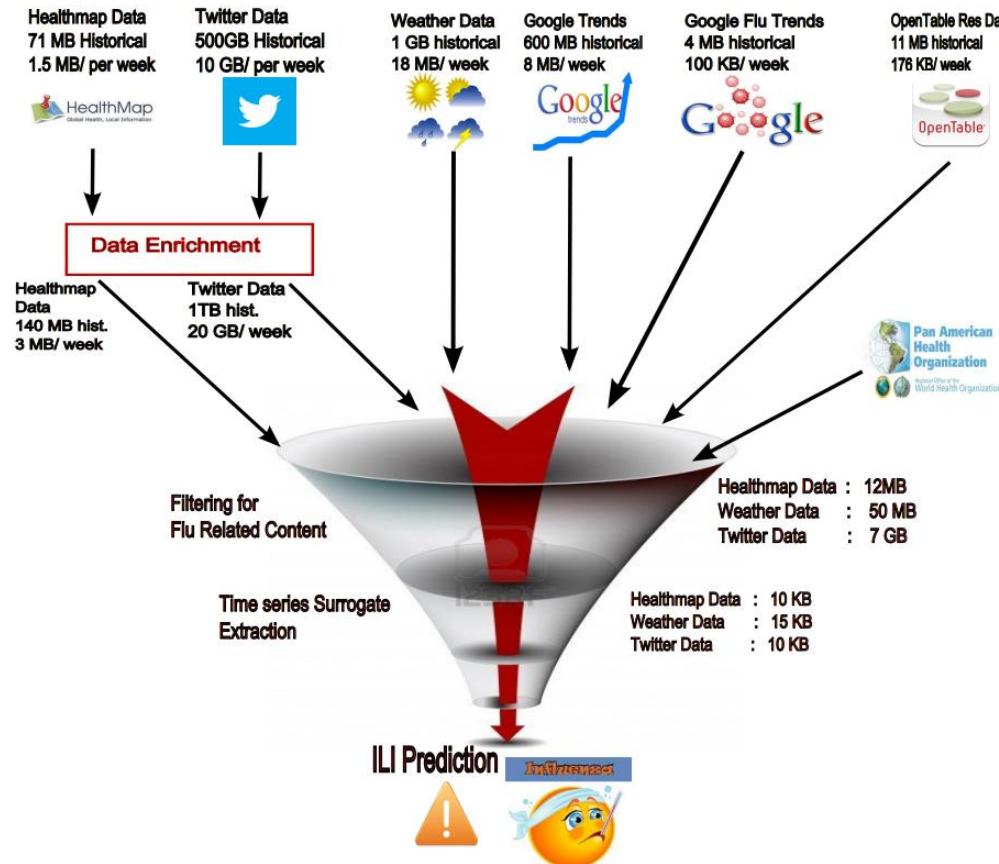
- Simultaneously predict HHS and national level ILI
- Capture interdependencies across regions
- Step 1: Region-level independent prediction
- Step 2: Refining prediction using increments modelled as multi-variate Gaussian with inter-region covariates



Ex. 3: Flu Forecasting based on Surrogate Data

[Chakraborty+ SDM 2014]

- ILI prediction based on heterogeneous data sources



Ex 4: Ensemble of ML models

- Use multiple ML methods for ILI prediction
 - Matrix Factorization Based Regression (MF)
 - Nearest Neighbor Based Regression (NN)
 - Matrix Factorization Regression using Nearest Neighbor embedding (MFN)

Table 1: Comparing forecasting accuracy of models using individual sources. Scores in this and other tables are normalized to [0,4] so that 4 is the most accurate.

Model	Sources	AR	BO	CL	CR	CO	EC	GF	GT	HN	MX	NI	PA	PY	PE	SV	All
MF	\mathcal{W}	2.78	2.46	2.39	2.14	2.70	2.22	2.12	2.63	2.52	2.73	2.31	2.21	2.49	2.77	2.61	2.47
	\mathcal{H}	2.81	2.31	2.22	1.92	2.43	2.04	2.11	2.57	2.33	2.48	2.39	2.15	2.18	2.47	2.33	2.32
	\mathcal{T}	2.37	2.35	2.18	2.03	2.21	2.12	1.83	2.12	2.29	2.03	1.89	2.06	1.96	2.20	2.21	2.12
	\mathcal{F}	2.34	2.11	2.29	N/A	N/A	N/A	N/A	N/A	N/A	2.71	N/A	N/A	2.31	2.24	N/A	2.33
	\mathcal{S}	2.48	2.21	2.33	2.04	2.31	2.21	1.93	2.03	2.15	2.51	2.42	2.52	2.33	1.93	2.30	2.24
NN	\mathcal{W}	2.92	2.93	2.63	2.52	2.66	2.51	2.71	2.82	2.59	2.62	2.55	2.59	2.61	2.80	2.52	2.66
	\mathcal{H}	2.73	3.10	2.42	2.27	2.83	2.64	2.43	2.25	2.71	2.31	2.61	2.35	2.43	2.39	2.52	2.53
	\mathcal{T}	2.72	2.86	2.31	2.62	2.77	2.52	2.71	2.66	2.51	2.44	2.13	2.01	1.77	2.51	2.20	2.45
	\mathcal{F}	2.11	2.21	2.33	N/A	N/A	N/A	N/A	N/A	N/A	2.19	N/A	N/A	2.41	2.32	N/A	2.26
	\mathcal{S}	2.51	2.31	2.41	1.81	2.52	2.41	2.12	2.29	2.51	2.13	2.61	2.14	2.51	1.87	2.12	2.28
MFN	\mathcal{W}	2.99	3.01	2.88	2.53	2.78	2.81	2.77	2.83	2.61	2.70	2.56	2.66	2.82	2.79	2.51	2.75
	\mathcal{H}	2.81	3.13	2.63	2.58	2.91	2.77	2.57	2.63	2.73	2.50	2.61	2.54	2.51	2.69	2.61	2.68
	\mathcal{T}	2.74	3.03	2.51	2.64	2.83	2.51	2.81	2.71	2.60	2.48	2.13	2.55	2.19	2.57	2.31	2.57
	\mathcal{F}	2.33	2.41	2.34	N/A	N/A	N/A	N/A	N/A	N/A	2.69	N/A	N/A	2.54	2.48	N/A	2.46
	\mathcal{S}	2.61	2.44	2.55	2.22	2.61	2.52	2.71	2.31	2.62	2.48	2.61	2.31	2.53	2.23	2.13	2.46

Regression models: Extensions

- Alternate search queries
 - Wikipedia [[McIver+ PLoS Comp Bio 2014](#)], UpToDate [[Santillana+ CID 2014](#)]
- Non-linear regression methods
 - GLM [[Wang+ KDD 2015](#)]
- Hierarchical Models
 - Elastic Nets [[Zou+ WWW 2018](#)], MDL [[Matsubara+ KDD 2014](#)]
 - Multi-Task Gaussian process [[Williams+ NIPS 2007](#)]

Statistical, ML/AI Models (Outline)

- Approaches:
 1. Regression Models
 2. **Language and Vision Models**
 3. Neural Models
 4. Density Estimation

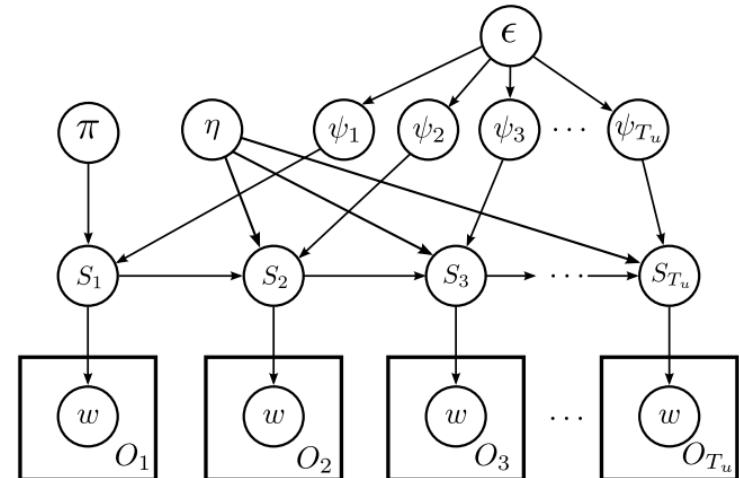
[S2.1] Language models

- Large sources of online text data
 - Social Media, Blogs, Search queries
- Incorporating textual data
 - Regression models using hand-designed linguistic features [[Lampos+ ECML 2010, Culotta+ 2010](#)]
 - Leveraging pre-trained word embeddings [[Zou+ WWW 2019](#)]
 - Topic Models [[Paul+ AAAI 2011, Chen+ ICDM 2017](#)]

Idea 1: Using Tweets to forecast H1N1 pandemic

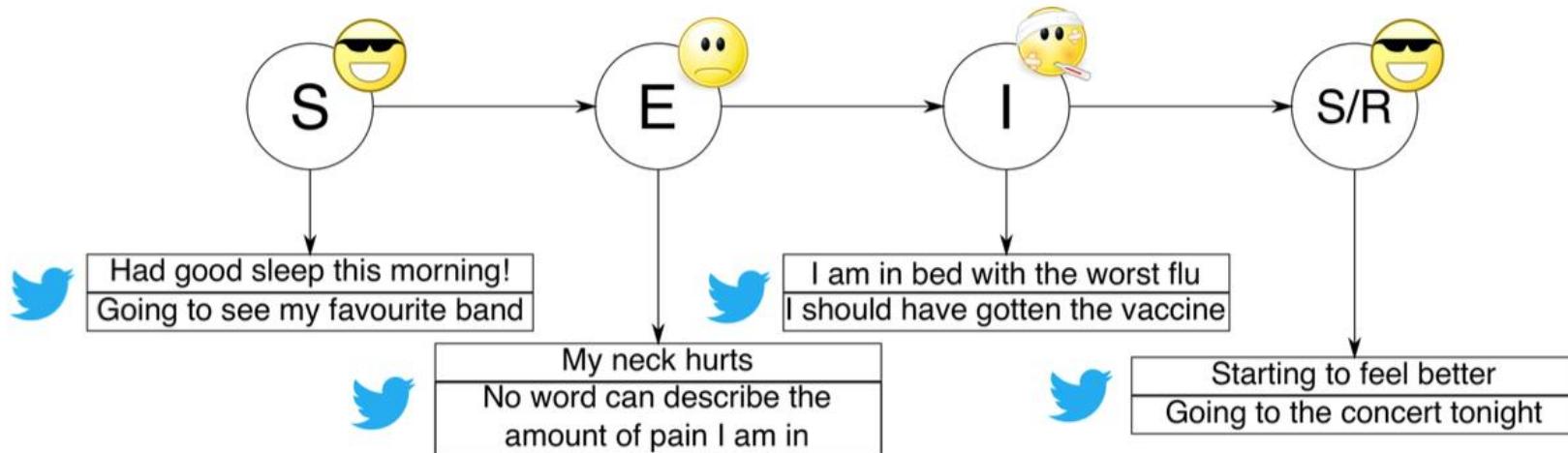
[Chen+ ICDM 2017]

- Temporal Topic modelling HFSTM
 - Use words of tweets to infer latent epidemiological states of users
- Combines
 - Information propagation on Twitter
 - Epidemiological model



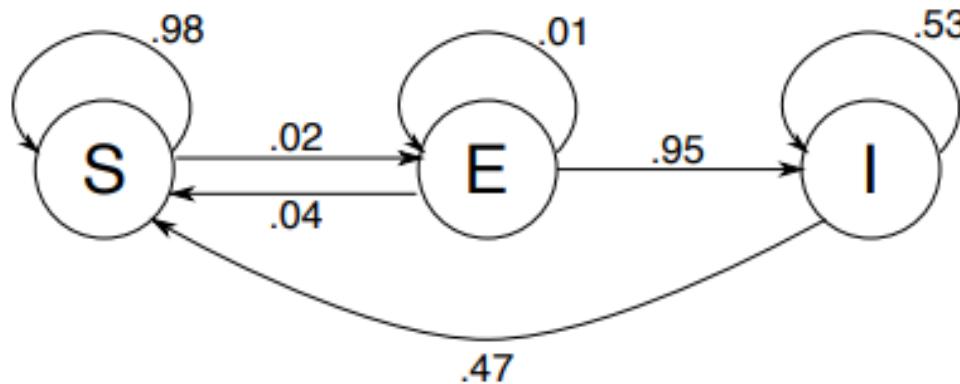
States of infection cycle

- Model states of infection cycle using tweets



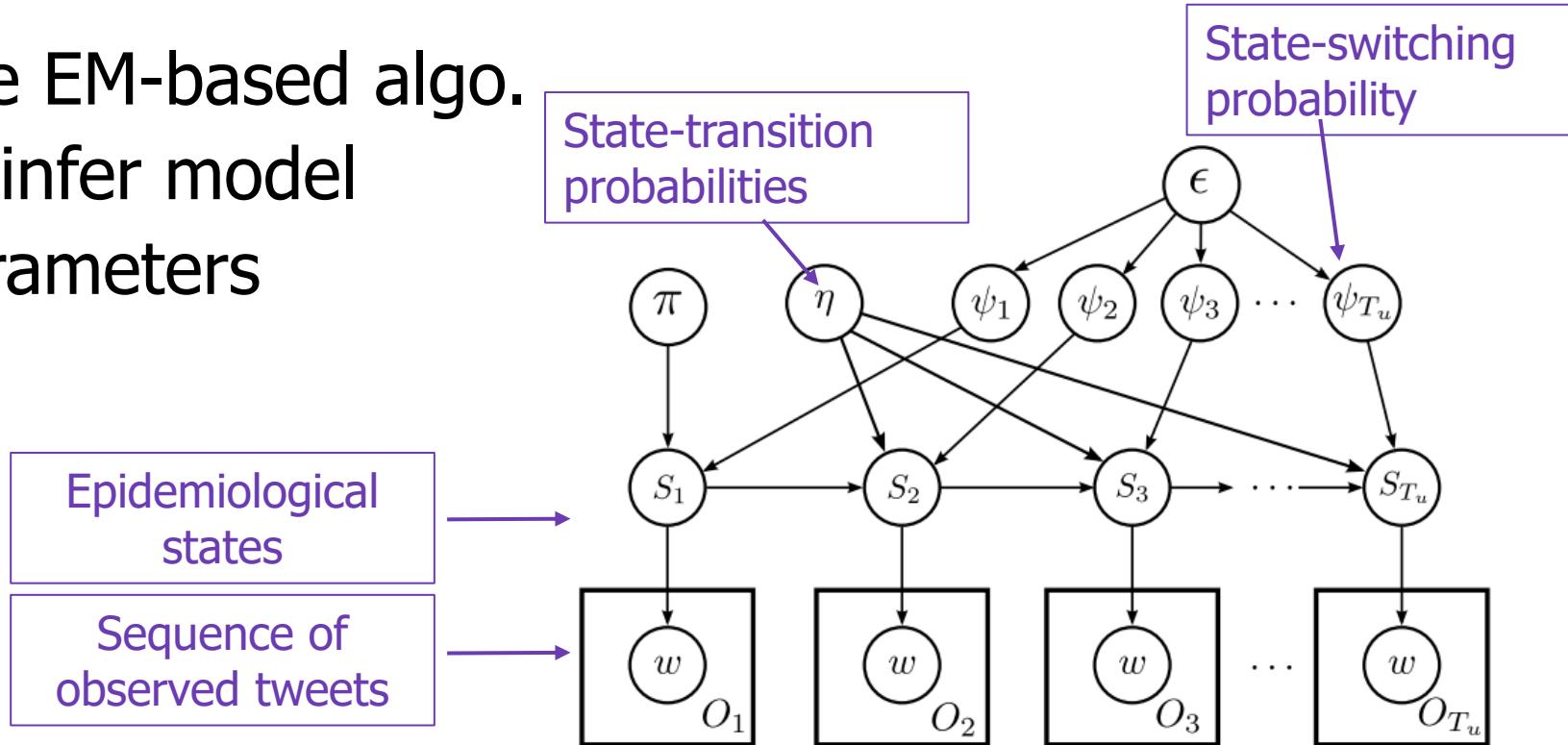
Model transition across states

- Transition probabilities across epidemiological states
- Automatically learned by HFSTM



HFSTM model (Contd.)

- Use EM-based algo.
To infer model
parameters



Idea 2: Cross-lingual word embeddings for transfer learning

[Zou+ WWW 2019]

- They chose the candidate queries based on word embedding of search queries.
 - Cross-lingual word embeddings
 - Trained on Wikipedia corpus (aligned text)
- Transfer learning: use the weights learned from a region to another region
 - US → France; US → Spain; US → Australia

Table 5: Top-5 target queries (with source mappings) in terms of mean ILI estimate impact (%) in the 10 weeks with the lowest and greatest MAE (all test periods), for all target countries (TC), based on their respective optimal transfer learning models.

TC	Mappings during accurate estimates	Mappings during inaccurate estimates
FR	flu incubation period → grippe durée (10.9), cough fever → la toux (6.3), how to treat flu → comment soigner une grippe (6), fever flu → fièvre de la grippe (5.47), flu treatment → traitement de la grippe (4.95)	24 hour flu → grippe intestinale (13.24), influenza a treatment → grippe traitement (8.07), remedies for colds → rhume de cerveau (6.75), child temperature → température du corps (6.37), child fever → fièvre adulte (6.04)
ES	symptoms of flu → symptômes grippe (9.04), fever flu → con gripe (7.49), cough fever → la tos (6.34), flu incubation period → cuanto dura una gripe (5.19), how to treat a fever → para bajar la fiebre (5.03)	mucinez for kids → tratamiento de la gripe (20.76), child fever → sinusitis (7.76), influenza a treatment → con gripe (7.02), symptoms pneumonia → bronquitis (6.04), child temperature → temperatura corporal (5.62)
AU	treatment for the flu → flu treatment (9.85), cough fever → cough and fever (8.05), flu type → influenza type (5.37), symptoms of flu → symptoms of flu (5.11), flu incubation period → flu incubation period (5.03)	24 hour flu → flu duration (11.51), child temperature → warmer (9.77), how to treat a fever → have a fever (6.94), tamiflu and breastfeeding → flu while pregnant (6.81), robitussin cf → colds (5.18)

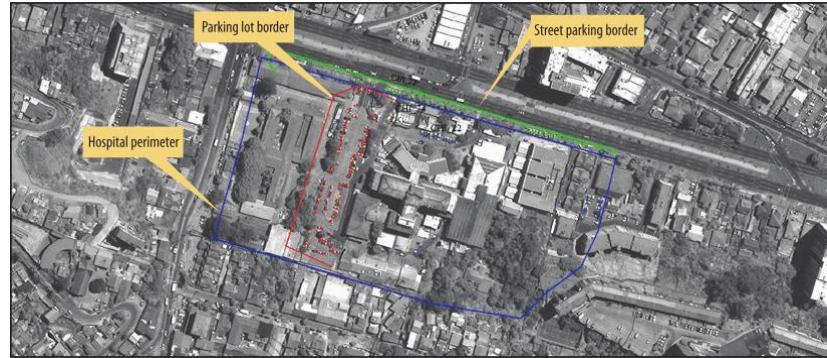
[S2.2] Vision models

- Recent works using satellite data
 - Flu [Butler+ IEEE Annals of Comp. 2014]
 - Covid [Nsoesie+ arXiv 2020]
- Images of places sensitive to outbreak
 - Parking lots near hospitals
- Still a nascent area of research

Ex 1: Satellite images to detect Flu outbreaks

[Butler+ IEEE Annals of Comp. 2014]

- RS Metric satellite imagery dataset
- Vision based automated algorithms to estimate no. of vehicles
 - Parking lots, streets, etc Of hospitals
- Linear regression on occupancy rate to model weekly ILI case counts.



Ex 2: Covid-19 Example

[Nsoesis+ arXiv 2020]

- Modeled early outbreaks (Dec '19 – May '20) in Wuhan, China
- Collected High-res satellite images from RS Metrics
 - Six hospitals, Seafood markets, two railway stations
- Automated segmentation of parking spots and high-traffic areas
 - Followed by manual counting of vehicles
- Also used search queries (Baidu) and traditional ILI counts as additional features

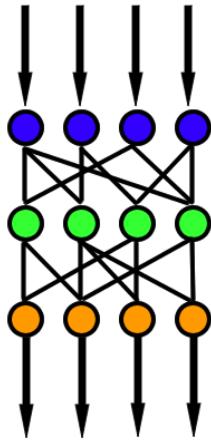
Statistical, ML/AI Models (Outline)

- Approaches:
 1. Regression Models
 2. Language and Vision Models
 - 3. Neural Models**
 4. Density Estimation

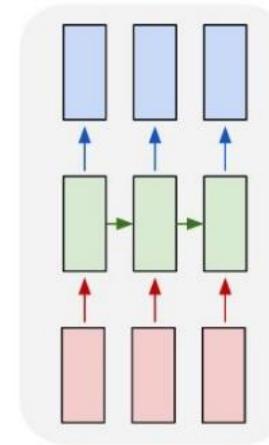
[S3] Neural Models

- Why deep learning?
 - Capture non-linear patterns in high-dimensional data with minor assumptions
 - Flexible learning of rich representations that generalize to complex domains
 - Leverage multiple sources of data of variety of modalities

Neural models for different modalities

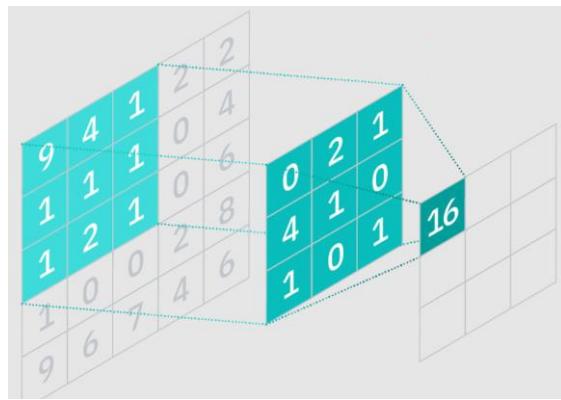
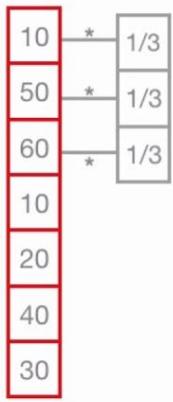


Feed-forward:
Static features data

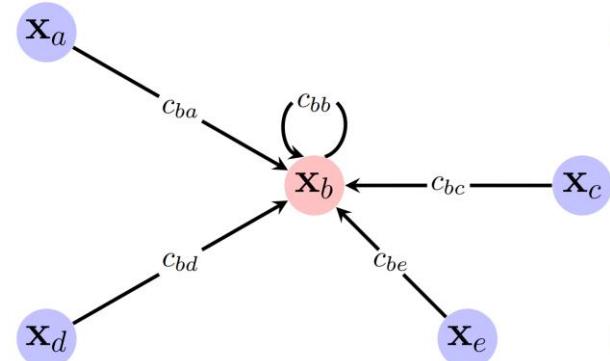


Recurrent Networks,
Transformers:
Sequential data

Neural models for different modalities (Contd.)



Convolutional Networks:
Sequence, Images, Videos

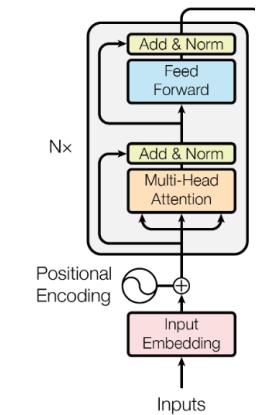
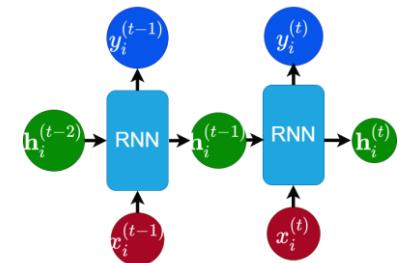


Graph Neural Networks:
Networks and Relational
data

[Bronstein+2021]

Idea 1: Off the shelf sequential models

- Captures complex, long-range patterns
- Capable of using high-dimensional features
- Popular models
 - Recurrent neural models [Rumelhart+ 1985]
 - Transformers [Vaswani+ NIPS 2017]
- Examples:
 - LSTM [Venna+ IEEE Access 2018] and transformers [Wu+ 2019] for flu forecasting
 - LSTM [Ayyoubzadeh+ JMIR 2020] for Covid-19 pandemic



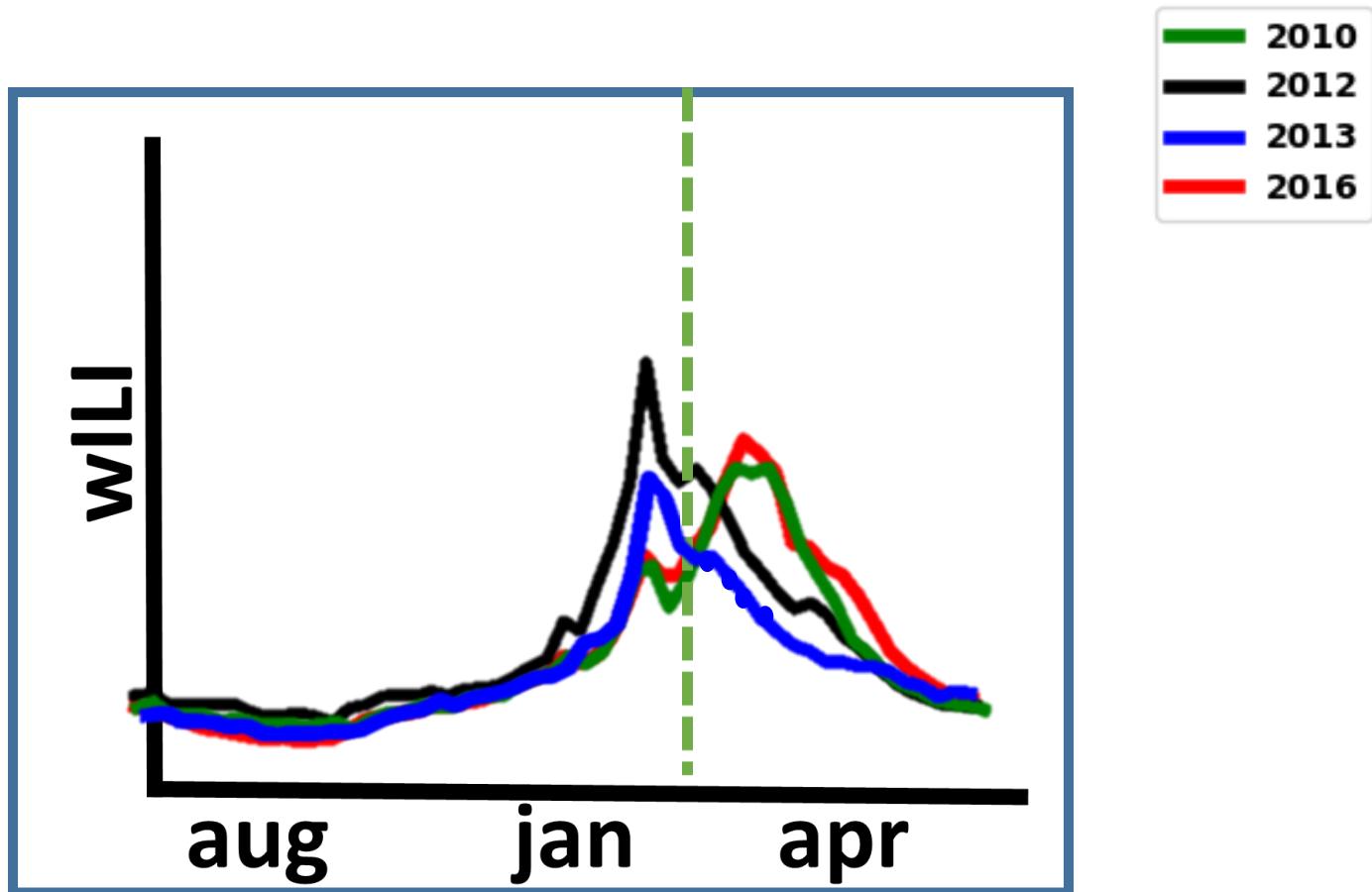
Adapting to epidemiology... some ideas

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

Idea 2: Model temporal dynamics via similarity (ex. 1)

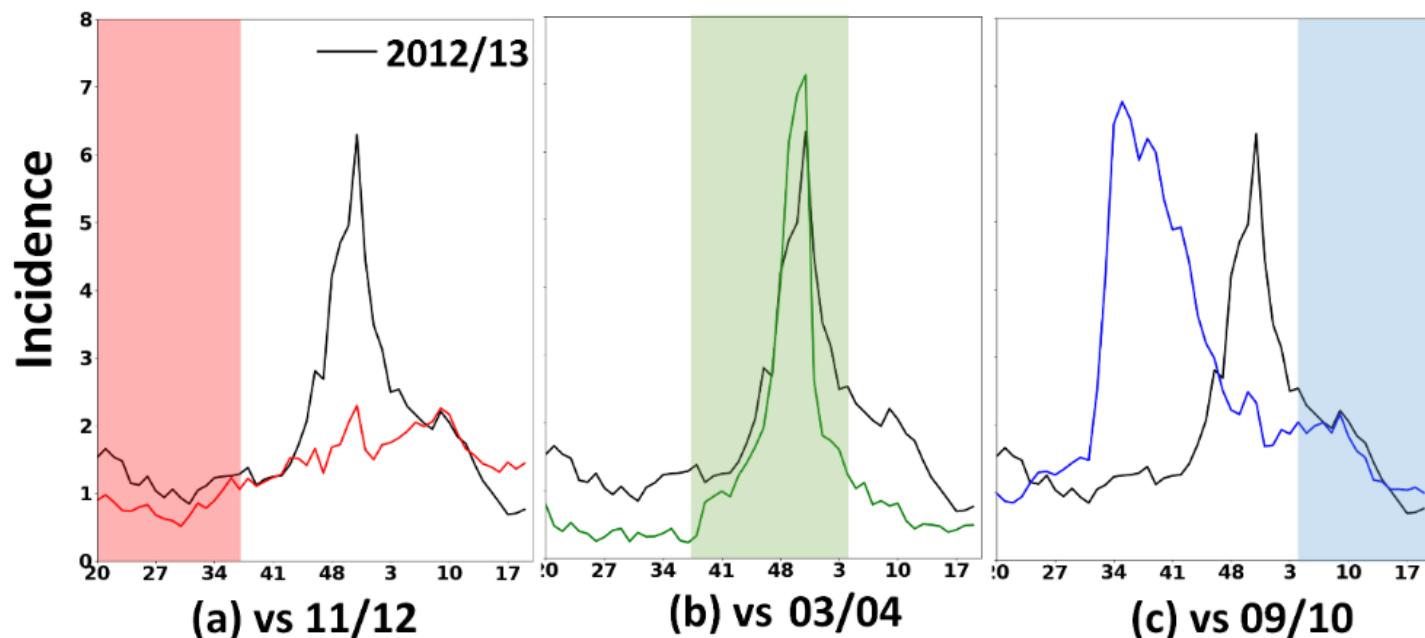
[Adhikari+, KDD 2019]

- Idea: clustering for prediction



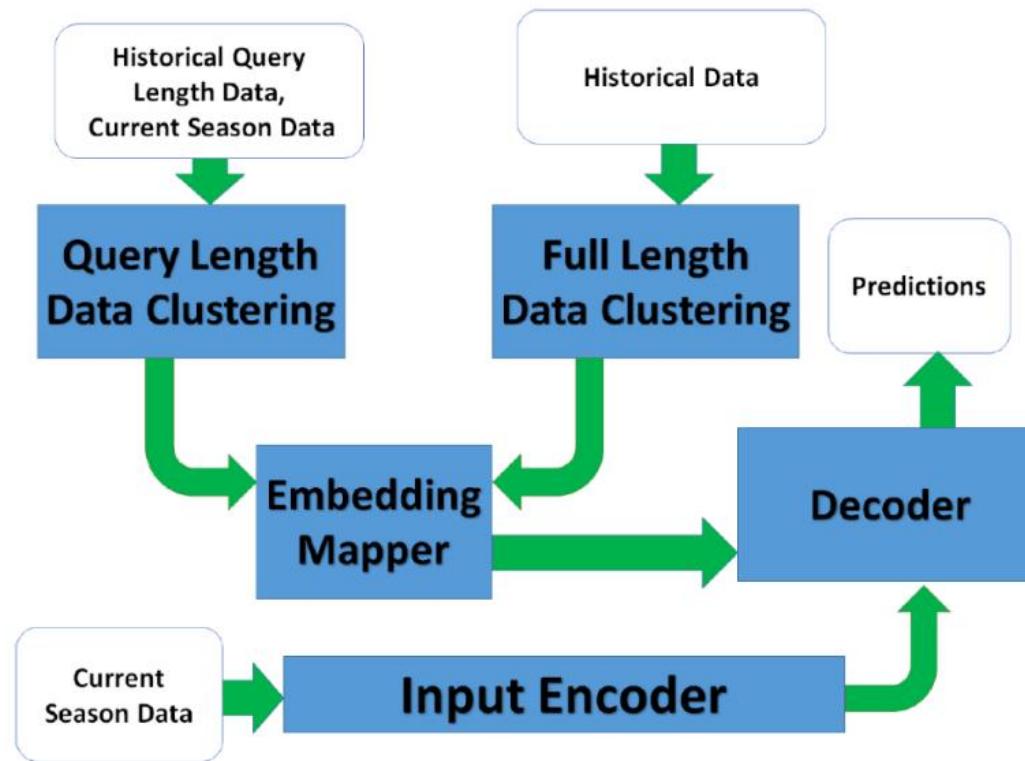
Model temporal dynamics via similarity CONTD.

- Idea: Dynamic clustering for prediction



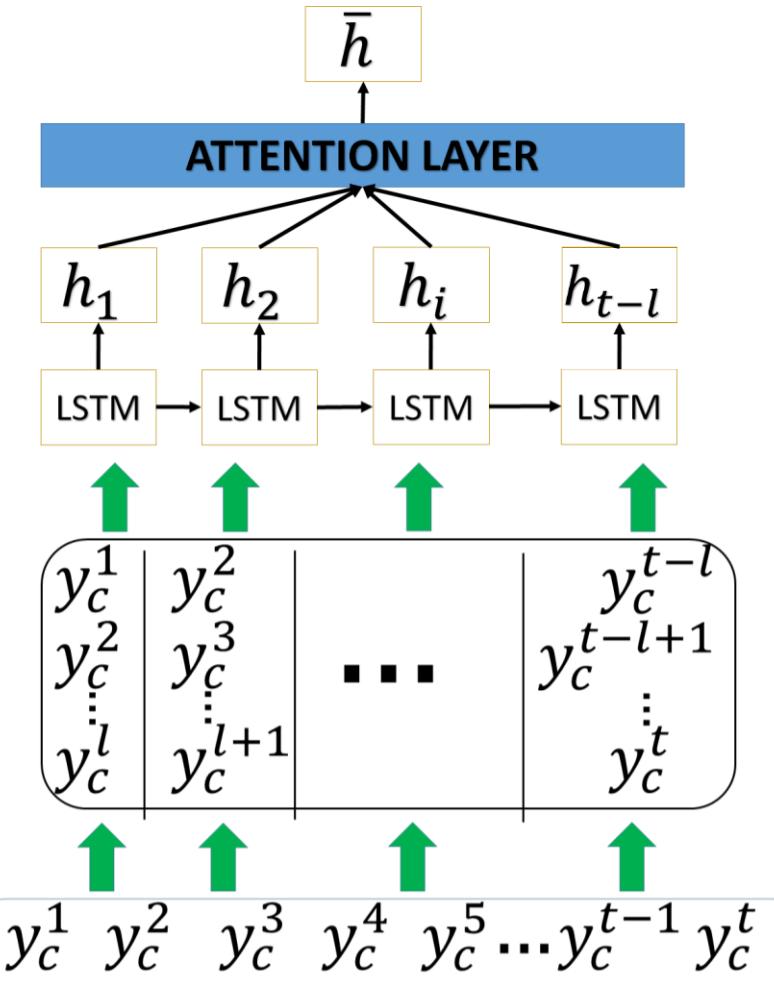
Model temporal dynamics via similarity CONTD.

- Idea: Dynamic deep clustering for prediction with limited data



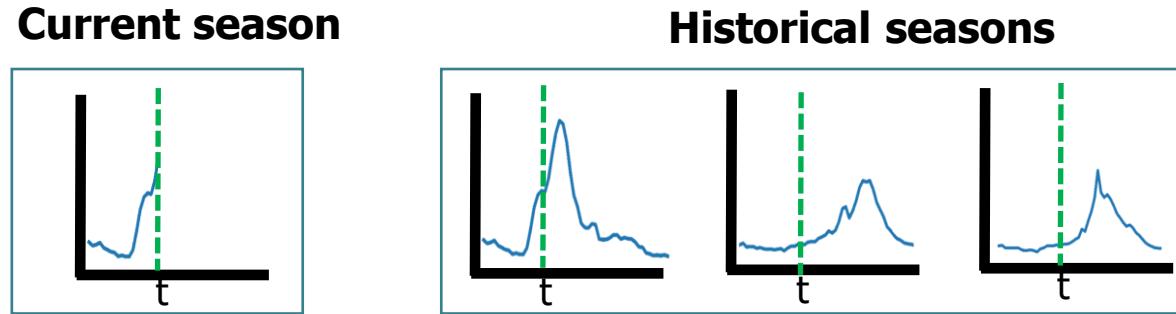
Component 1: Input Encoder

- Encodes partially observed season $Y_c = \{y_c^1, y_c^2, \dots, y_c^t\}$
 - Uses LSTM to learn a representation of Y_c
- Not enough
 - CDC **revises** and **updates** data
 - Unreliability of the latest data
 - Solution: assign different weight part of the data
 - Learn attention weights over the LSTM representation



Embed Historical Seasons

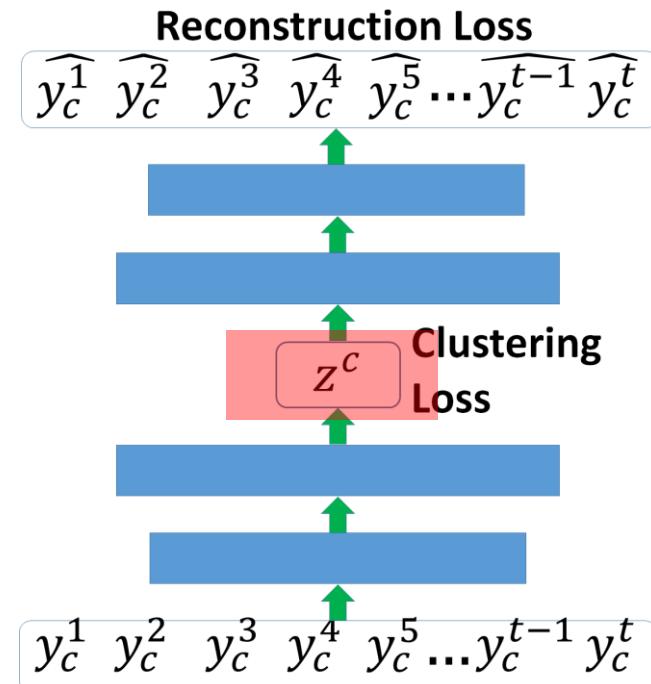
- Embed the historical seasons to capture the similarity with the current season
- Current season is observed only **till week t**



- Use snippets of historical seasons till week t to learn embedding

Component 2: Clustering Query length Data

- Learn meaningful embeddings of the historical data **till week t**
- Improved Deep Embedded Clustering [Guo+, 2017]
 - Auto-encoder with clustering loss



Component 2: Clustering Query length Data (Contd.)

- Soft-Assignment Distribution

$$q_{ij} = \frac{(1 + \|z_i^t - m_j\|^2)^{-1}}{\sum_j (1 + \|z_i^t - m_j\|^2)^{-1}}$$

- Target Distribution

- Encourage points to be assigned to nearest cluster

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})}$$

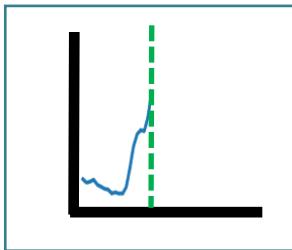
- Minimize KL Divergence between the two to refine clusters

$$L_c^t = \sum_i \sum_j p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right)$$

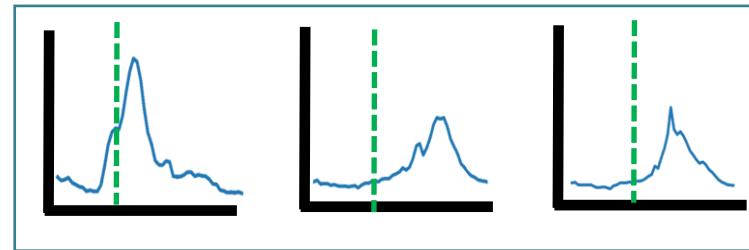
Embedding Component

- Not all data is utilized to learn the embeddings

Current seasons



Historical seasons



- Want to utilize the entire data
- Our Idea
 - Embed the full length historical seasons separately
 - Learn a mapper between two embedding space

Component 4: Decoder

- Different architectures for different tasks
- Task 1

- Future incidence for next four observations

Architecture	Loss function
$y^* = f_{next}(\bar{h}_j, z_k^T)$ Feed-forward	L2

- Task 2

- The peak intensity $\max_i y_c^i \quad i=1 \dots T$

Similar to above

- Task 3

- The peak time $\arg \max_i y_c^i \quad i=1 \dots T$

$$x_t = W_p f(\bar{h}_j, z_k^T)$$
$$P(t | x_t) = \frac{\exp(x_t)}{\sum_i \exp(x_i)}$$

Cross-Entropy

- Task 4

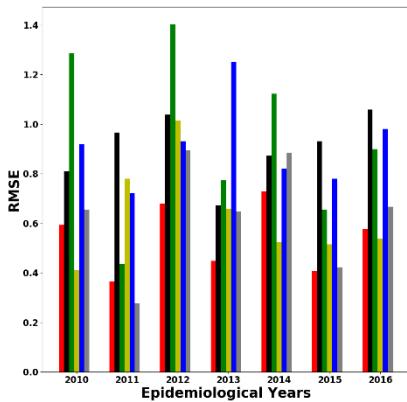
- The onset $Week j$ such that $\forall_{i=j}^{j+3}, y_c^i \geq b_c$

Similar to above

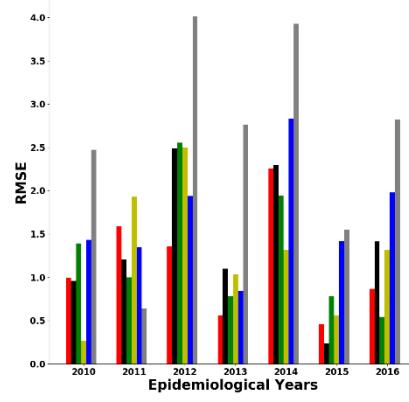
Performance: National region

- How well does EPIDEEP perform in different tasks for the national region?

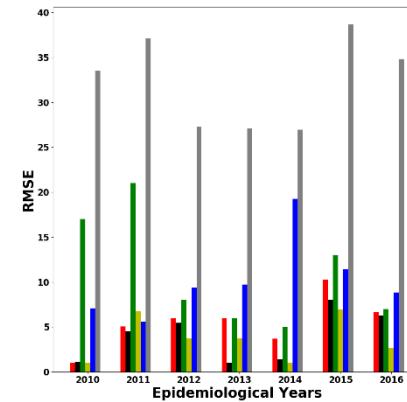
EpiDeep
EB
Historical
KNN
LSTM
ARIMA



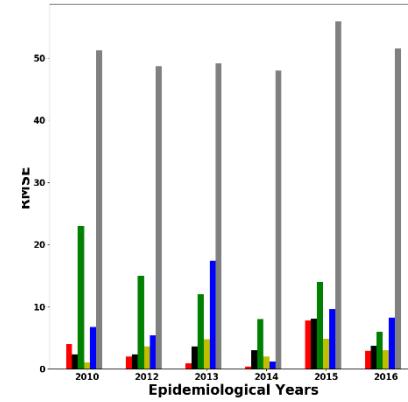
Future Incidence



Peak Intensity



Peak Week



Onset

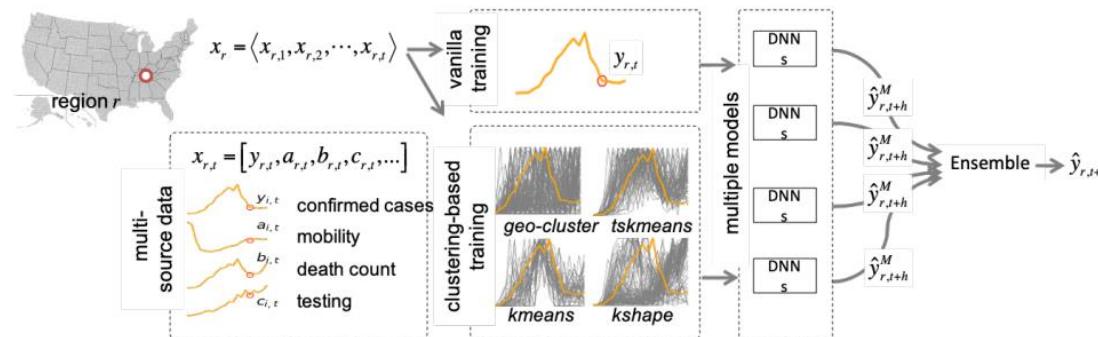
Lower is better

EpiDeep outperforms baselines in most settings.

Ex. 2: Using multiple clustering methods

[Wang+, BigData 2020]

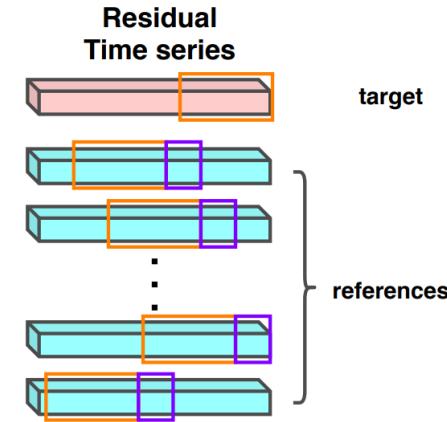
- Temporal and geo. similarity (adjacent regions)
 - Train one model for set of regions
 - Regions clustered by geographical similarity or using clustering algorithms
 - Multiple models with different clustering strategies combined using ensembling (more on ensembles later)



Ex. 3: Inter-series attention

[Jin et al., SDM 2021]

- Idea: Use attention-based similarity to learn from time-series of all regions
- Model:
 - Segment past time-series
 - Transform the segments into fixed embeddings via Convolutional layers
 - Use similarity between target (input) and segments using attention to predict target output
 - Joint training of model across all regions

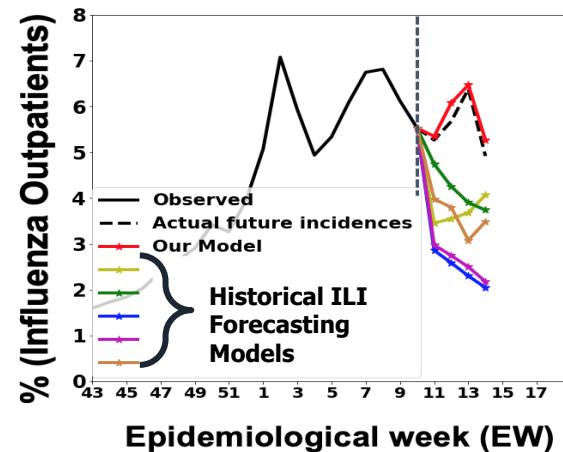
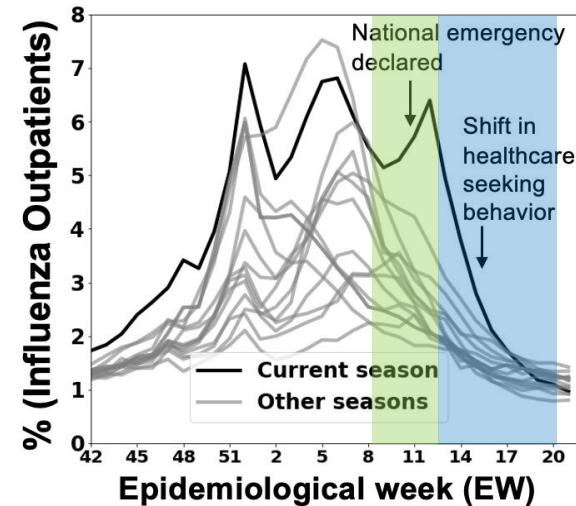


Idea 2: Transfer knowledge representations

- Transfer learning
 - Leverage implicit knowledge from large data/models to scarce data scenarios
 - Reduce compute cost
- Examples:
 - From one country to another country
 - Even in different continents [Panagopoulos+ AAAI 2020]
 - From a historical scenario to a novel scenario
 - From pre-COVID flu to COVID-contaminated flu counts [Rodríguez et al., AAAI 2020]

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

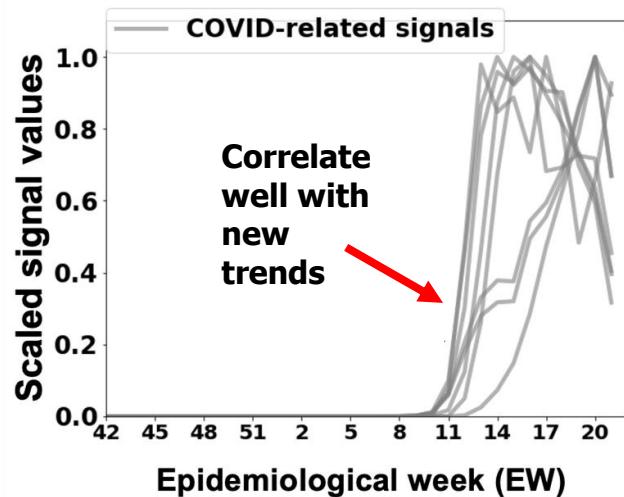


New COVID-related signals correlate with new trends

- Line-list based
- Testing
- Crowdsourced
- Mobility
- Exposure
- Social Media surveys

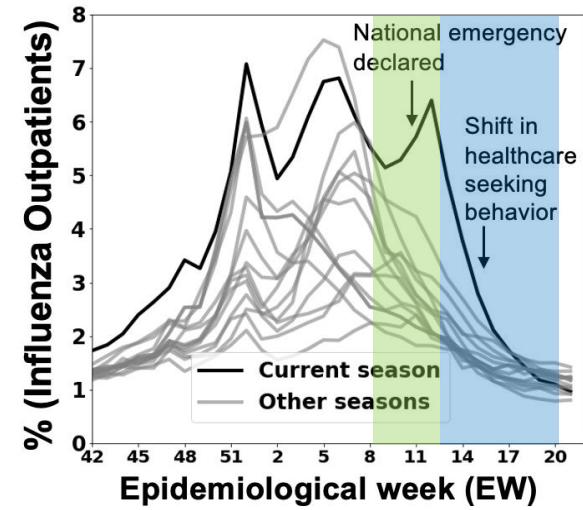
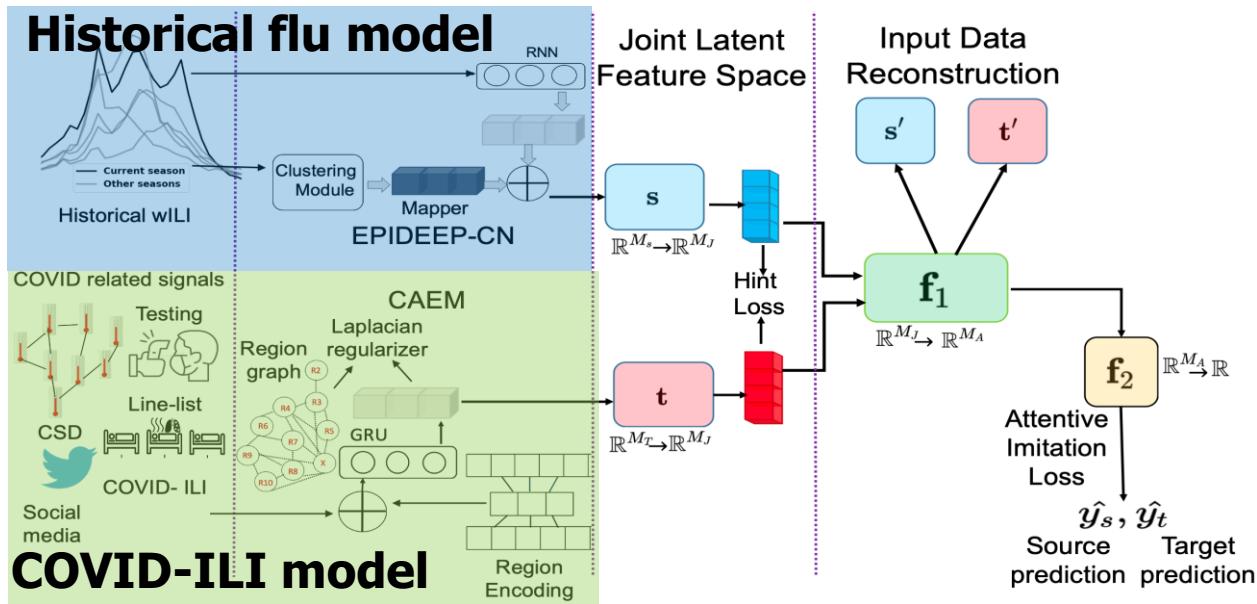


Center for Systems Science
and Engineering



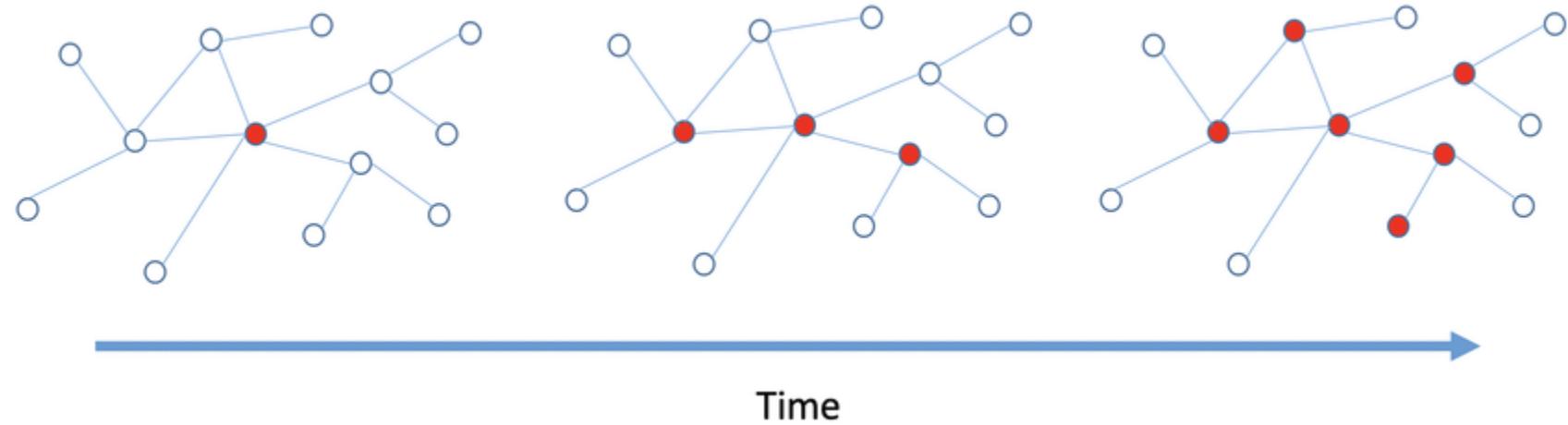
Attentive transfer learning for heterogeneous domains

- CALI-Net: steer a historical flu model (EpiDeep, KDD 2019) with new COVID-related signals



Idea 3: Incorporate spatial structure

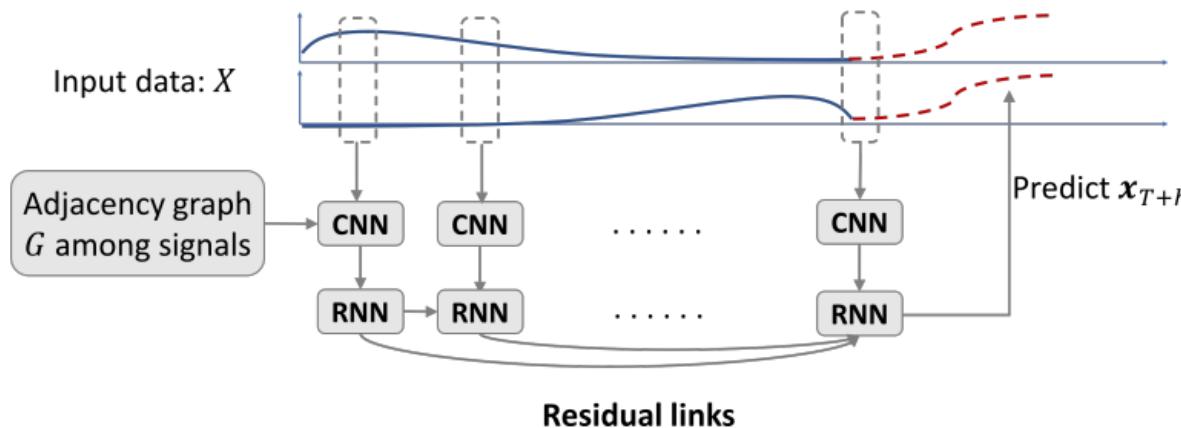
- Pathogens propagate to adjacent regions
 - And then to new adjacent regions
- Propagation over spatial graphs



Ex 1: Convolutional modules to aggregate related regions

- Combine CNN and RNN
 - CNN: Model regional proximity
 - RNN: Temporal dynamics
 - Residual connections: Better generalization

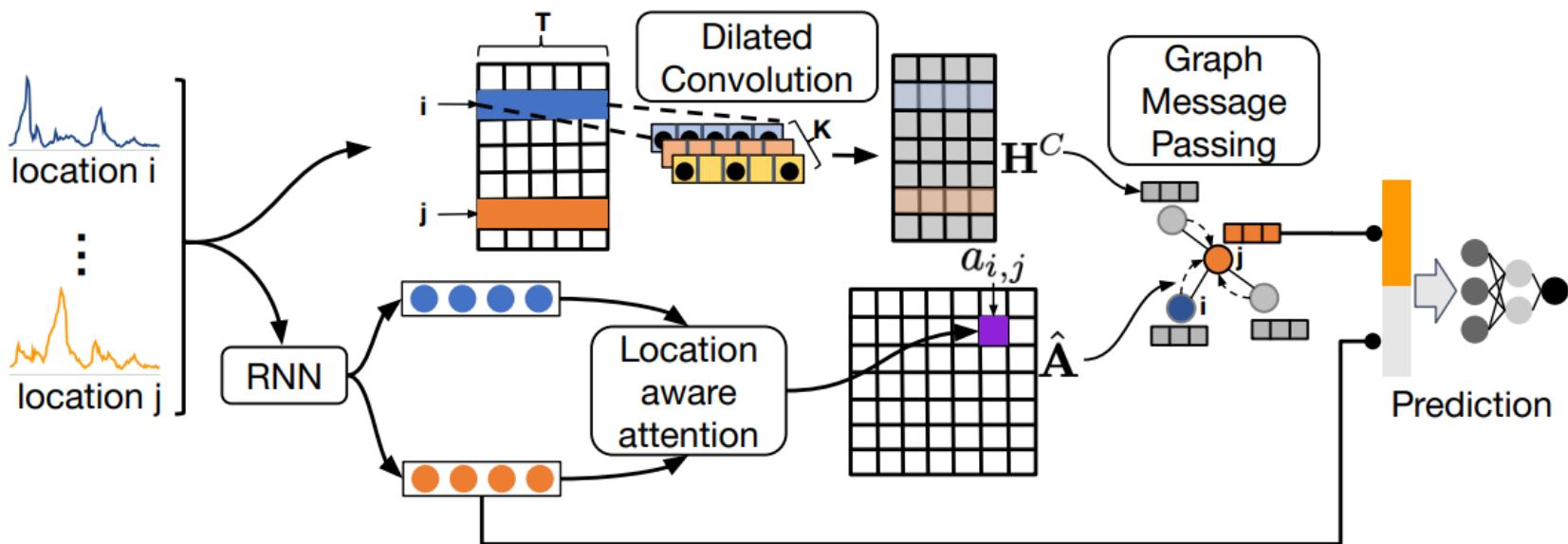
[Wu+, SIGIR 2020]



Ex 2: Graph message passing for spatial propagation

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

[Deng+, CIKM 2020]



ColaGNN in long-term forecasting

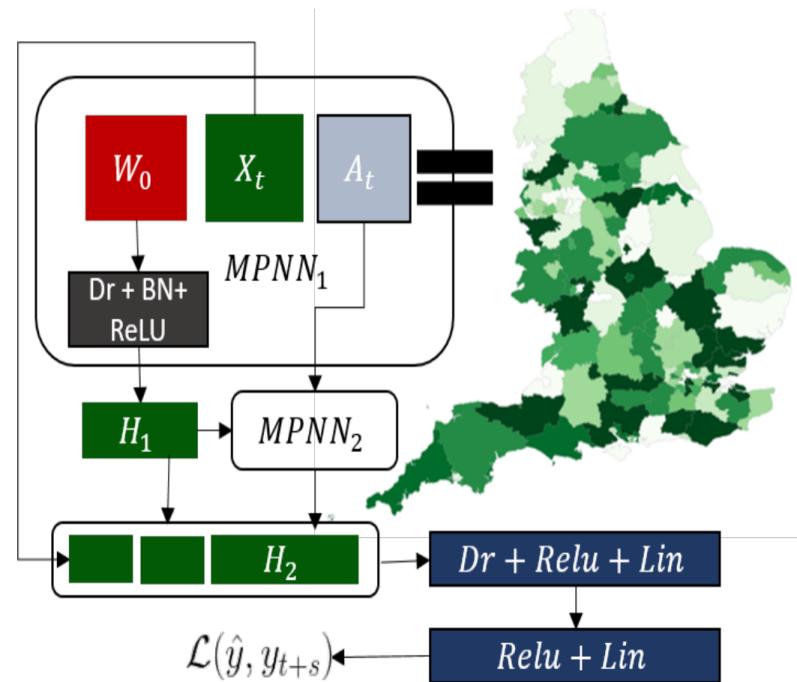
- Dilated Convolution observed to improve long-term predictions as well

RMSE(↓)	Japan-Prefectures					US-Regions					US-States				
	2	3	5	10	15	2	3	5	10	15	2	3	5	10	15
GAR	1232	1628	1988	2065	2016	536	715	991	1377	1465	150	187	236	314	340
AR	1377	1705	2013	2107	2042	570	757	997	1330	1404	161	204	251	306	327
VAR	1361	1711	2025	1942	1899	741	870	1059	1270	1299	290	276	295	324	352
ARMA	1371	1703	2013	2105	2041	560	742	989	1322	1400	161	200	250	306	326
RNN	1001	1259	1376	1696	1629	513	689	896	1328	1434	149	181	217	274	315
LSTM	1052	1246	1335	1622	1649	507	688	975	1351	1477	150	180	213	276	307
RNN+Attn	1166	1572	1746	1612	1823	613	753	1065	1367	1368	152	186	234	315	334
DCRNN	1502	1769	2024	2019	1992	711	874	1127	1411	1434	165	209	244	299	298
CNNRNN-Res	1133	1550	1942	1865	1862	571	738	936	1233	1285	205	239	267	260	250
LSTNet	1133	1459	1883	1811	1884	554	801	998	1157	1231	199	249	299	292	292
ST-GCN	996	1115	1129	1541	1527	697	807	1038	1290	1286	189	209	256	289	292
Cola-GNN	929	1051	1117	1372	1475	480	636	855	1134	1203	136	167	202	241	237
% relative gain	6.7%	5.7%	1.1%	11.0%	3.4%	5.3%	7.6%	4.6%	2.0%	2.3%	8.7%	7.2%	5.2%	7.3%	5.2%

Ex 3: Transfer Learning via Graph Neural Networks

[Panagopoulos+, AAAI 2021]

- Construct graphs from mobility data across regions of country
- Combine GNN (MPNN) and LSTM to capture spatial and temporal relations
- Use Meta-Learning [Finn+ ICML '17] to train over all regions (to further adapt to low data regions)



Recap of Ideas for Neural Forecasting

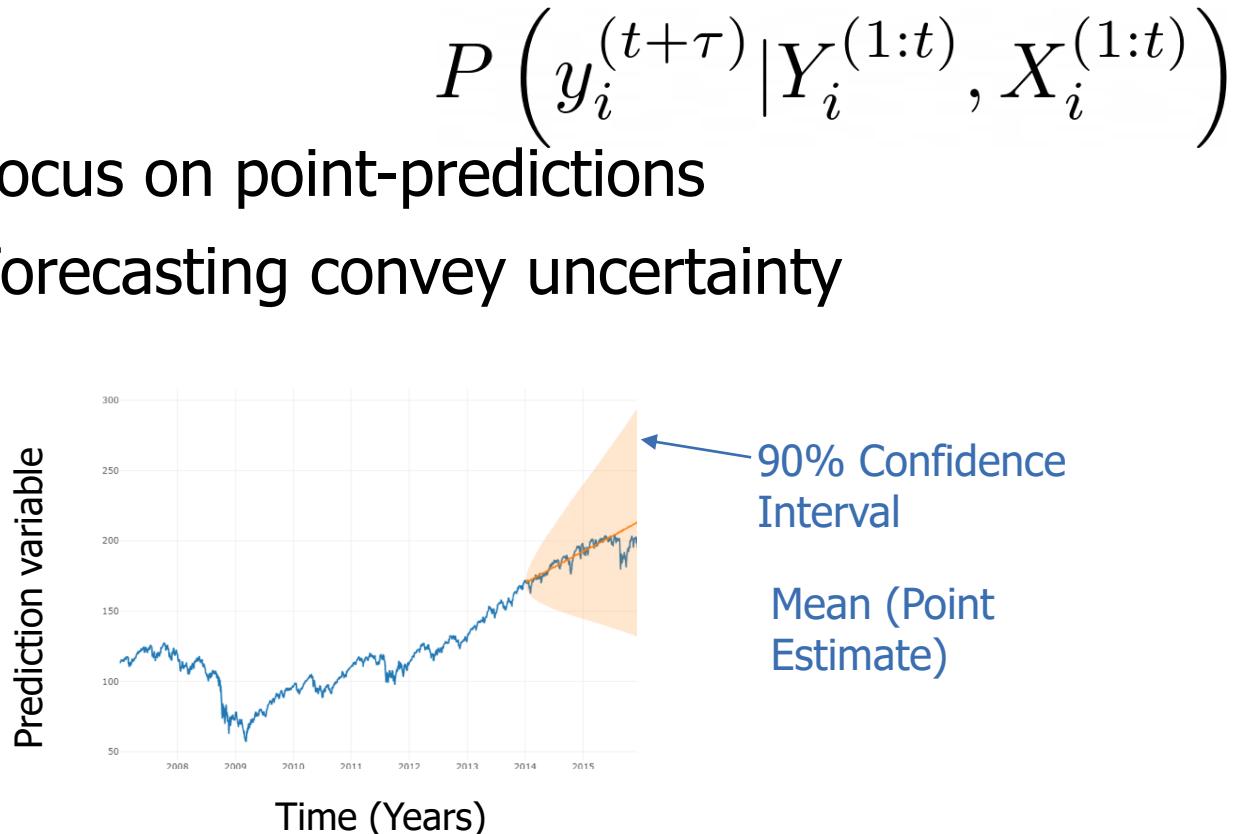
- Idea 1: Leverage similarity
 - Temporal: seasonal, segments
 - Geographical
- Idea 2: Transfer learning
 - Heterogeneous domain transfer learning
 - Knowledge distillation
 - Meta-learning
- Idea 3: Spatial structure
 - Temporal and spatial convolution
 - Graph neural networks

Statistical, ML/AI Models (Outline)

- Approaches:
 1. Regression Models
 2. Language and Vision Models
 3. Neural Models
 - 4. Density Estimation**

[S4] Density Estimation Models

- Directly model the forecast distribution
- Why?
 - Most works: focus on point-predictions
 - Probabilistic Forecasting convey uncertainty



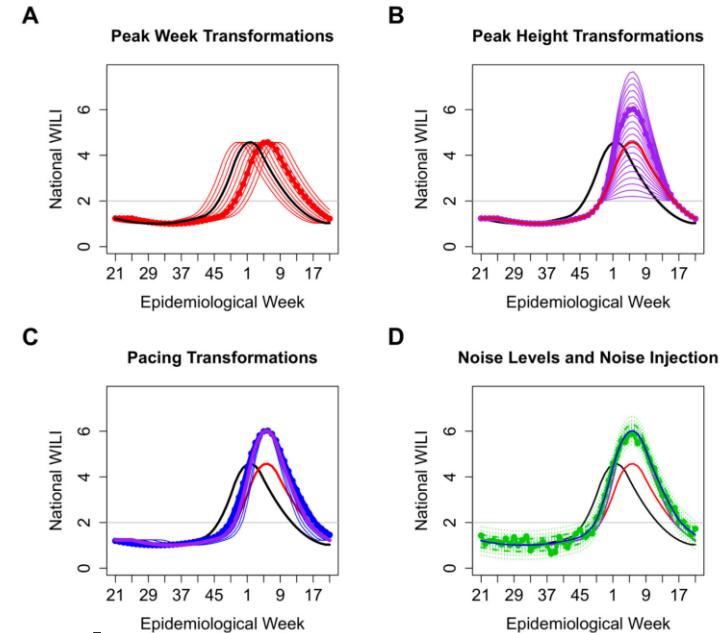
Types of Density Estimation Models

- *Parametric*: parameters of distribution as function of features
- *Non-parametric*: Function of training datapoints leveraging similarity
- *Neural probabilistic models*: Deep learning to capture complex patterns for improved calibration

Ex 1: Empirical Bayes (Parametric)

[Brooks+ PLoS 2015]

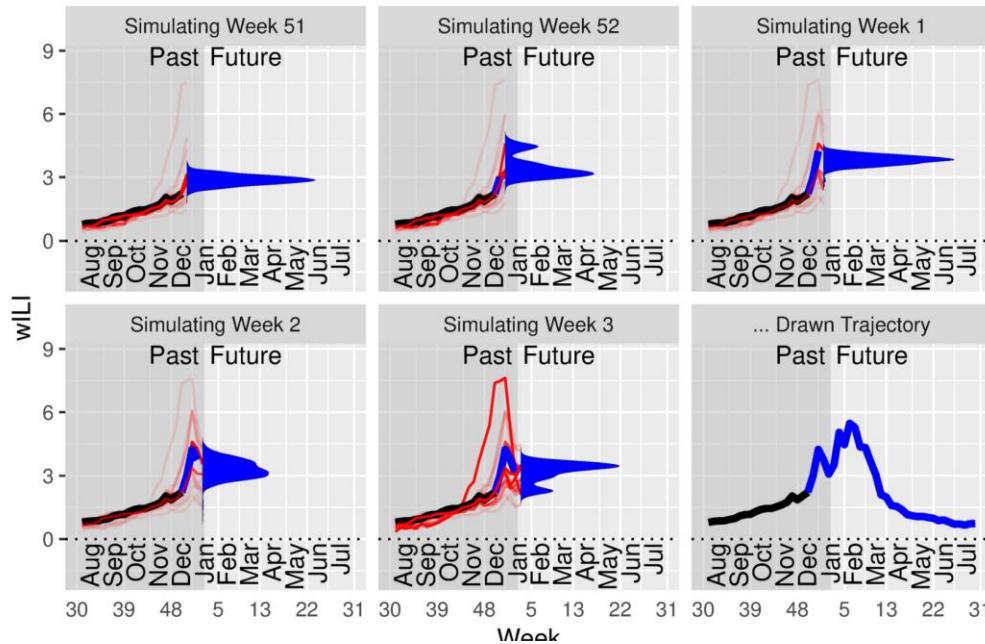
- Current season's epidemic curve is a probabilistic distribution of features
- Epi-curve function parameters:
 - Similarity in shape to past sequences
 - Peak height, week
 - Scaling factor of the curve
- All modelled into forecast distribution
 - Use Bayesian Inference to calibrate for current season



Ex 2: Delta Density (Non-parametric)

[Brooks+ PLoS 2017]

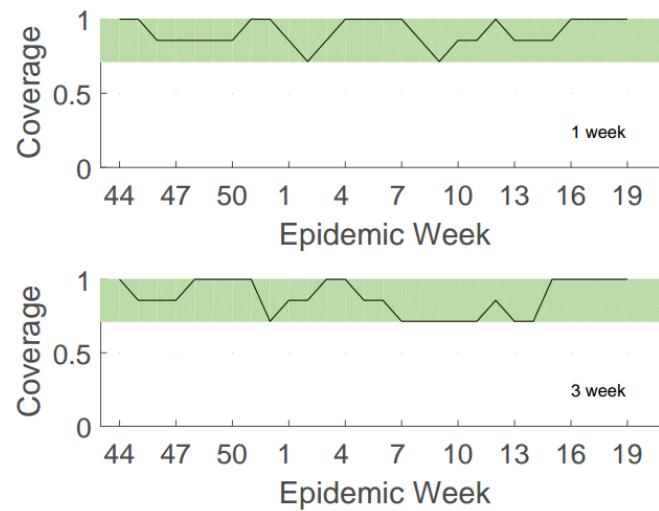
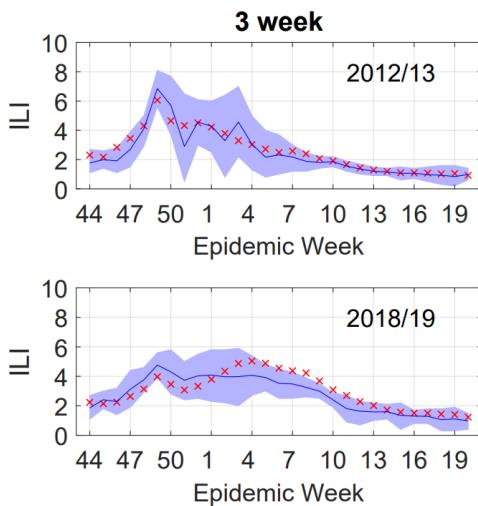
- Use kernel density estimation to leverage similarity with historical seasons
- One of the top models in Flusight 2017 challenge



Ex 3: Gaussian Process (Non-Parametric)

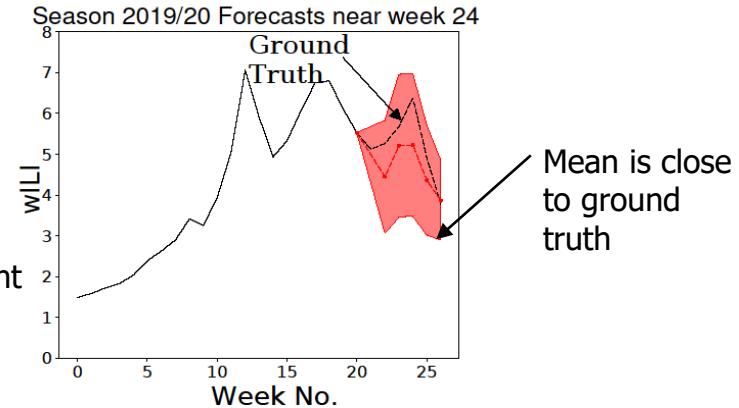
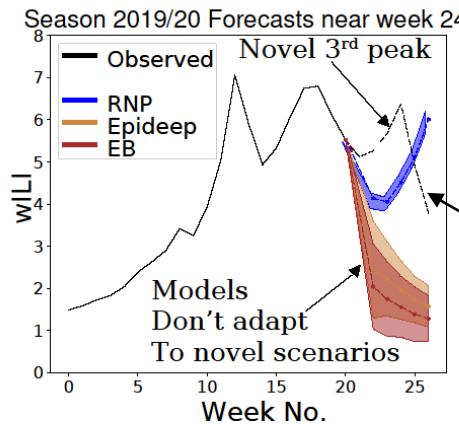
[Zimmer+ ICML 2019]

- 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

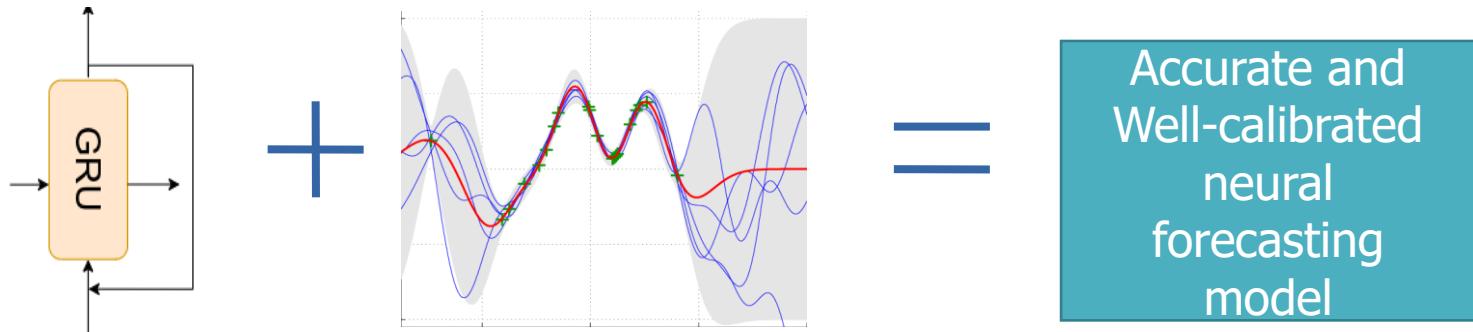
- Most statistical methods don't leverage complex patterns to learn well-calibrated forecasts
 - Can't adapt to provide reliable forecast uncertainty on novel patterns



Ex 4: EpiFNP: Neural non-parametric model for better calibration

[Kamarthi+, NeurIPS 2021]

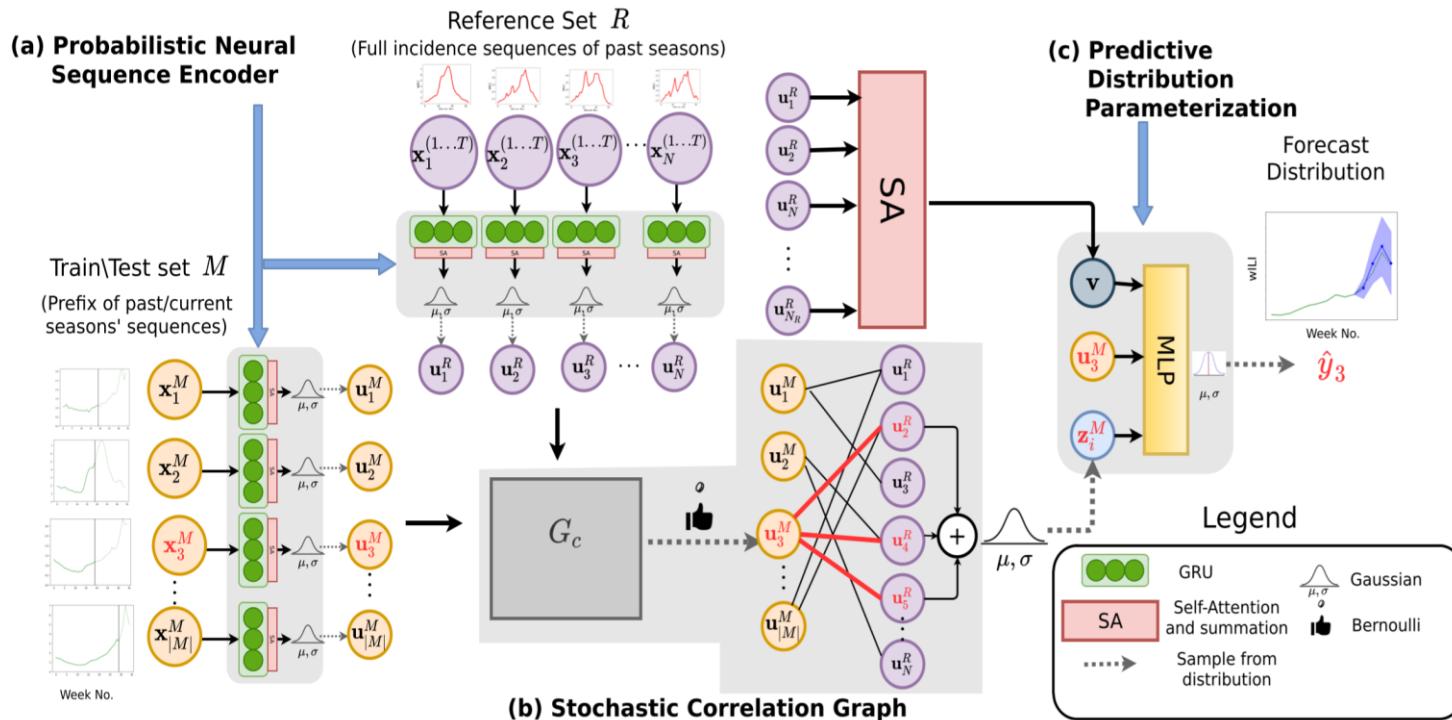
- Neural Sequential models to capture long term sequential patterns
- Non-parametric Gaussian Process
 - Flexibly model forecast distribution
 - Leveraging similarities with past historical sequences



Deep Sequential
Models

EpiFNP: Architecture

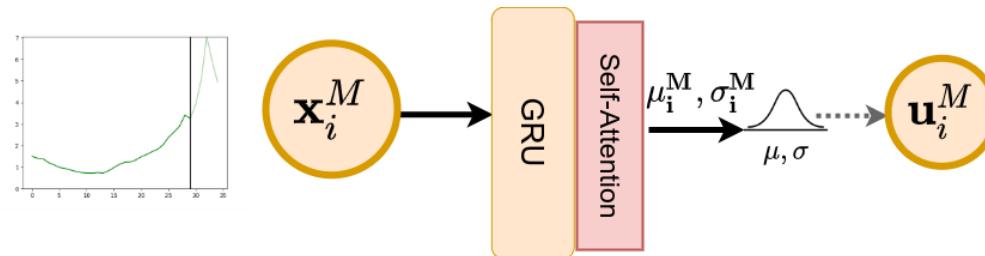
Sequential representations +
neural Gaussian processes



Rodríguez, Kamarthi, and Prakash 2023

Component 1: Probabilistic Neural Encoder

- Encodes current week and historical sequences into a latent vector distribution – Multi-variate Gaussian
- Quantify **uncertainty of input sequence**



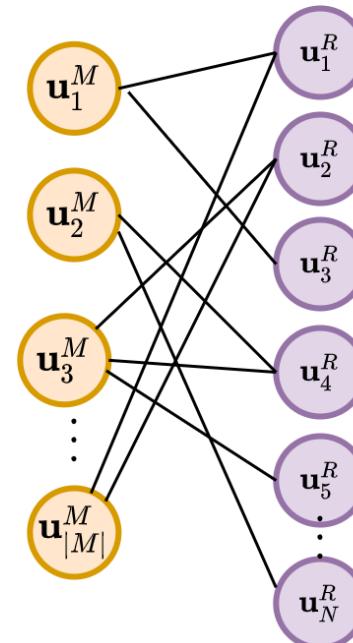
Component 2: Stochastic Correlation Graph

- Leverages similarity between current (training/test set) and historical sequences (reference set) in latent space

$$d_{i,j} = \kappa(\mathbf{u}_i^M, \mathbf{u}_j^R)$$

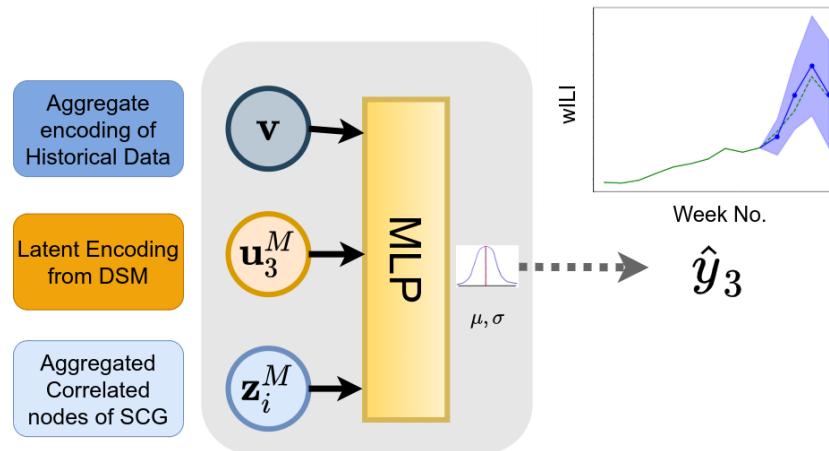


Connect an edge with probability

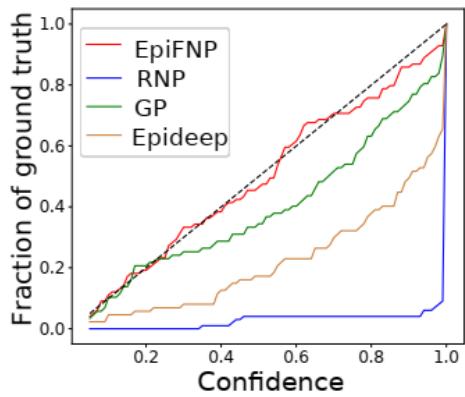


Component 3: Predictive Distribution Decoder

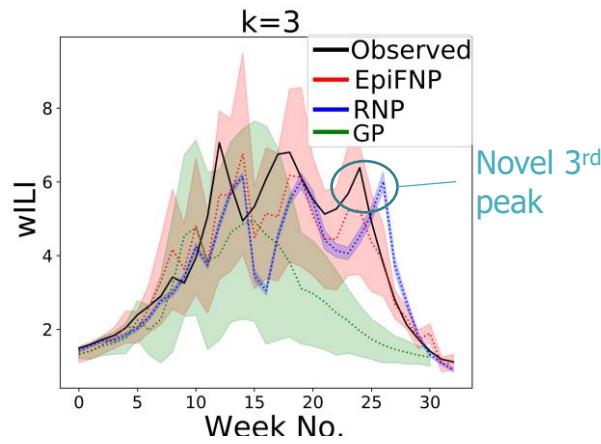
- Combines uncertainty from different perspectives to parameterize predictive distribution:



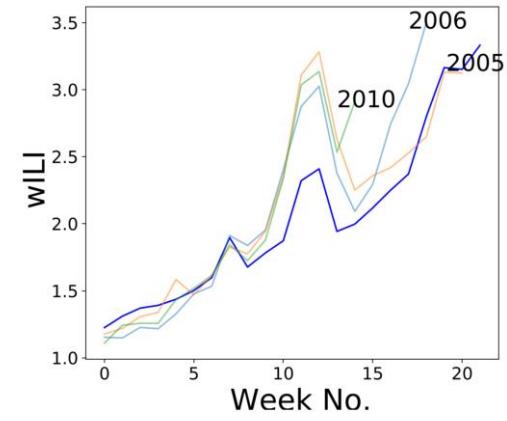
Results



Well calibrated predictions



Adapt to novel patterns



Explaining predictions

Most similar seasons chosen by EpiFNP

Extending to modeling multiple views/modalities



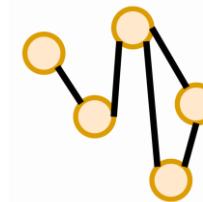
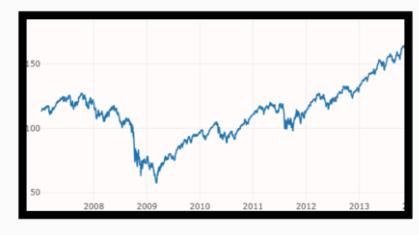
Sequences



Static Features



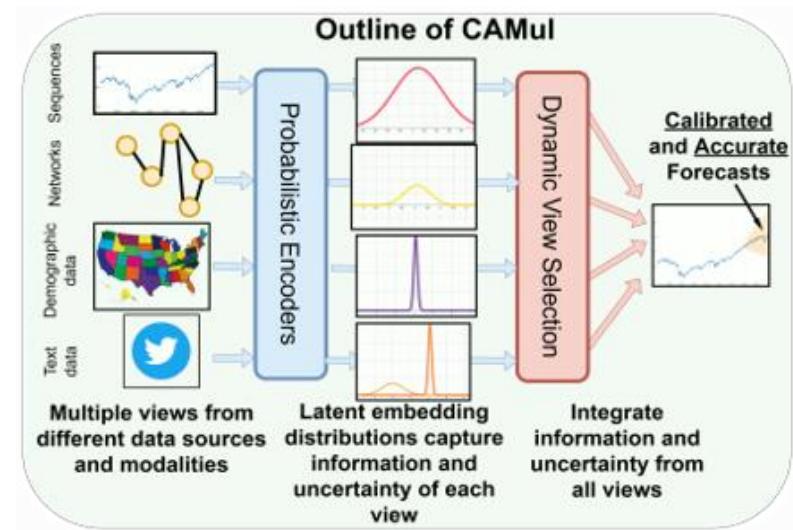
Graphs



CaMuL: multi-view time series forecasting

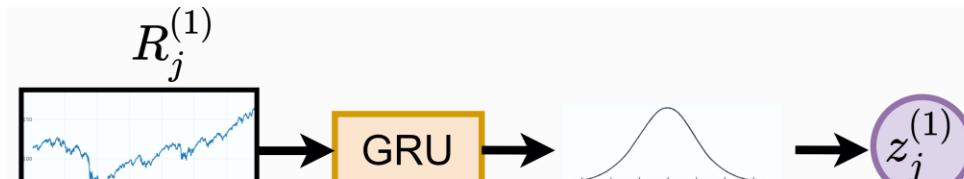
[Kamarthi+, WWW 2022]

- Encode representation for each data source
- Select important views adaptively
- Combine important views to provide calibrated forecasts

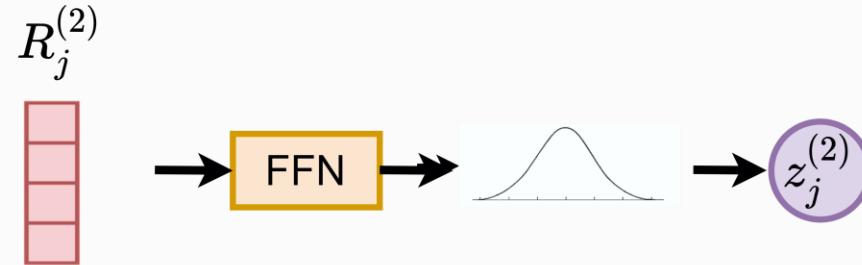


Component 1: Multi-view Latent Encoders

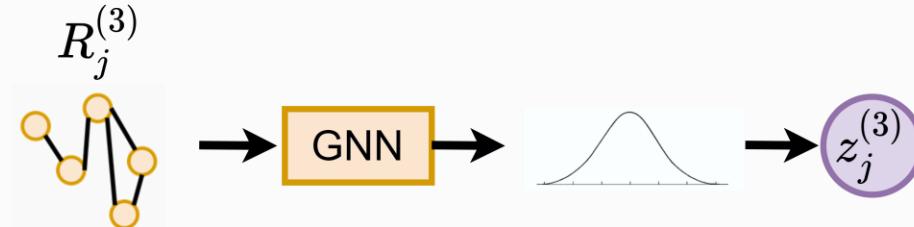
Sequence View



Static View

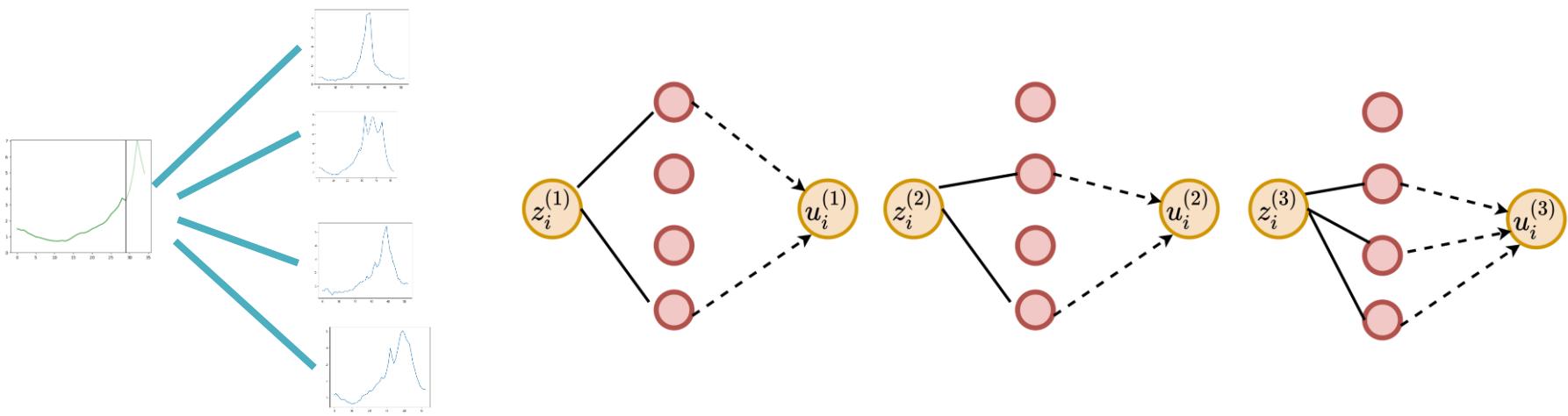


Graph View



Component 2: View Specific Correlation Graph

- Extracts probabilistic similarity relation between training and reference points for each view

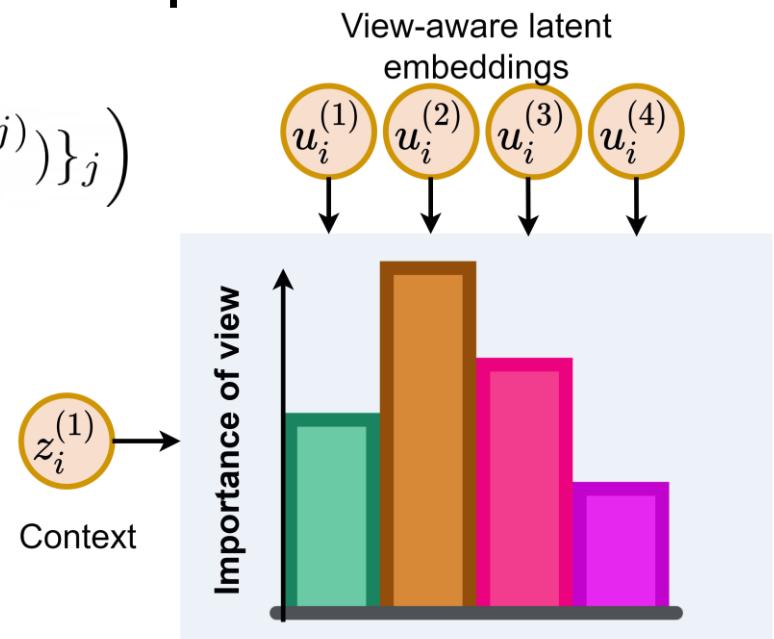


Component 3: Context Specific Dynamic View Selection

- Derive importance of each view based on context
- Context: Latent embedding of input time-series

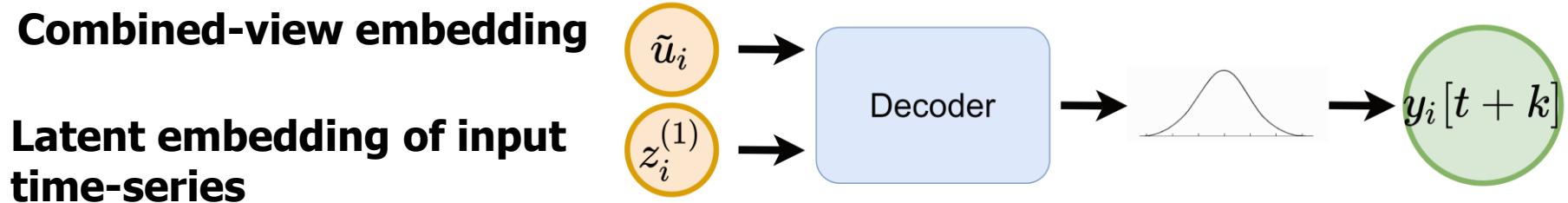
$$\{\alpha_i^{(j)}\}_j = \text{Softmax} \left(\{NN_1(z_i^{(1)})^T NN_2(h_i^{(j)})\}_j \right)$$

Cross-Attention



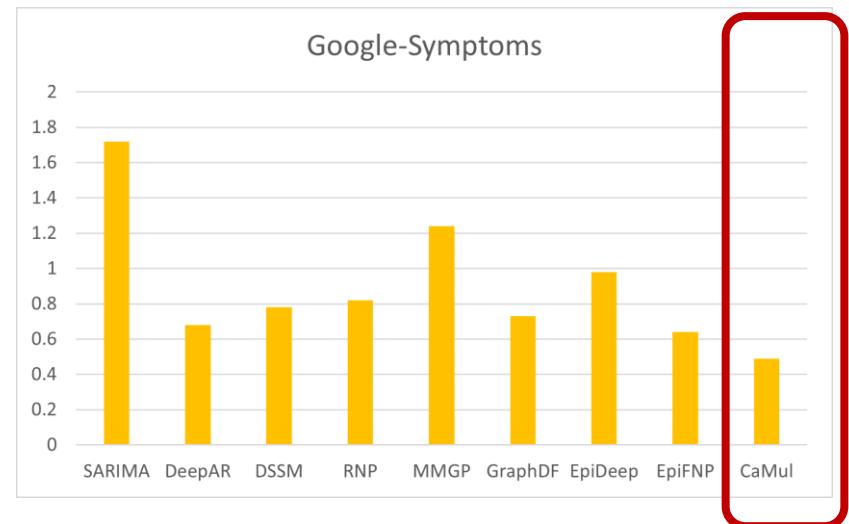
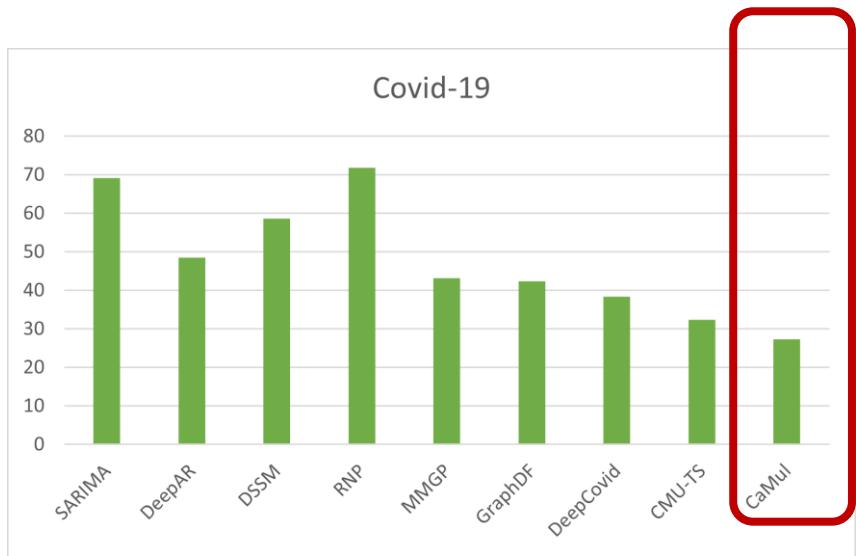
Component 4: Decoder

- Combine information and uncertainty from all views and time-series to get output distribution

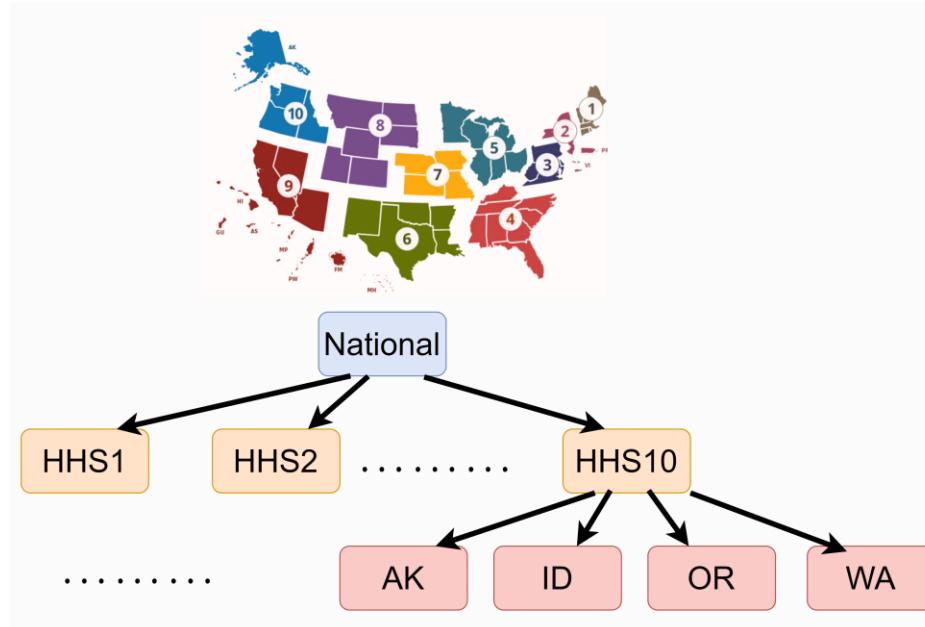


CaMuL: Results

- 18-30% accuracy and calibration for Covid-19 and Flu tasks (CRPS scores)



ProfHiT: Extension to Hierarchical Datasets

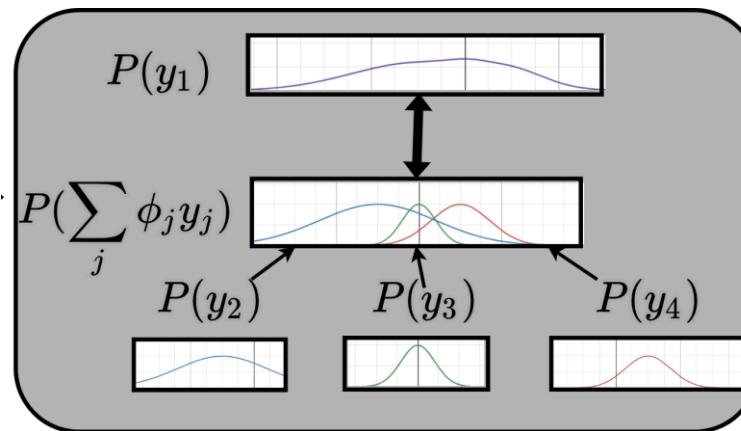


$$H_{\mathcal{T}} = \{\mathbf{y}_i = \sum_{j \in \mathcal{C}_i} \phi_{ij} \mathbf{y}_j : \forall i \in \{1, 2, \dots, N\}\}$$

ϕ_{ij} = Fraction of population of i in j

Idea: Probabilistic Consistent Hierarchical Forecasting

- Ensure the **forecast distributions** of each time-series satisfies hierarchical constraints



Results: Better Accuracy and Calibration leveraging hierarchy

- Outperforms consistently at most levels of hierarchy

Models/Data		Tourism-L		Labour		Wiki		Flu-Symptoms		FB-Survey	
		CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS	CRPS	LS
Baselines	DeepVAR	0.17	0.61	0.045	0.75	0.232	0.83	0.610	3.25	7.32	5.32
	TSFNP	0.21	1.19	0.071	1.41	0.287	0.86	0.460	0.93	5.53	7.84
	MINt	0.5	0.58	0.045	4.12	0.243	0.78	0.630	3.18	5.39	6.35
	ERM	0.56	0.53	0.045	3.63	0.221	0.74	0.620	2.75	6.14	4.23
	HIERE2E	0.15	0.38	0.034	0.51	0.211	0.46	0.420	0.81	4.12	1.13
	SHARQ	0.17	0.41	0.054	0.47	0.241	0.52	0.470	1.42	3.12	0.81
	PROFHIT (Ours)	0.12	0.33	0.026	0.21	0.184	0.35	0.250	0.28	1.43	0.45

Pros/Cons Statistical Models

- Leverage large variety of data directly
 - Handle high-dimensional data structures
 - Data may not directly relate to mechanics of epidemic curve
 - Doesn't need domain-specific information on epidemic dynamics
- State of the art in multiple forecasting tasks
 - Short-term forecasting
 - Staple in modern forecasting initiatives and challenges

Pros/Cons Contd.

- Unaware of epidemic spread mechanisms
 - Poor performance in long-term
 - Due to lack of knowledge on epidemic dynamics
- Unable of evaluating what-if scenarios
 - Not easily adapted to predict counterfactual scenarios
- Need constant monitoring and fine-tuning for real word deployment (more on this later!)
 - Loss in performance/adaptability to drift in data distributions
(Eg: sudden outbreaks, errors/change in data collection)

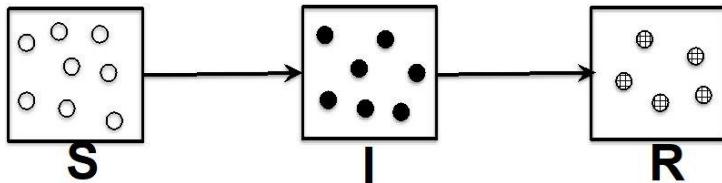
Outline

1. Epidemic forecasting: data and setup (40 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (60 min)
 - 30 min break at 3:15 PM, feel free to catch us for coffee
5. **Hybrid models** (45 min)
 - 5 min break
6. Epidemic forecasting in practice (25 min)
7. Open challenges and final remarks (20 min)

Part 5: Hybrid Models

Mechanistic models: overview

Mass-action models



- Based on ordinary differential equations (ODEs)
- Assume homogeneity of population and interactions

Metapopulation models

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

- Breaks down population into sub-populations to model heterogeneity
- Model spread within and across sub-populations

Mechanistic models: overview

Agent-based models (ABMs)



- Every person in population is represented
- Interactions (disease spread) are based on interaction networks at multiple settings:
 - workplace
 - school
 - ...

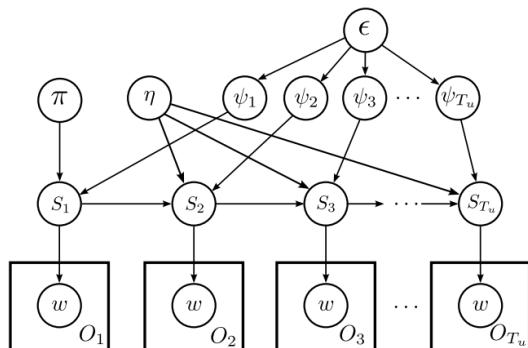
Stat/ML/AI models: overview

Regression

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

- Sparse and autoregressive models
- Google Flu Trends (GFT)

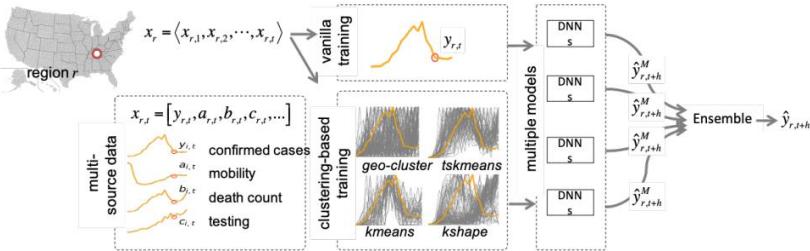
Language and vision



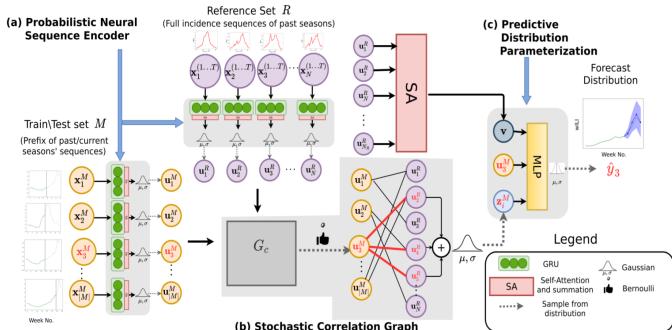
- Topic modeling + disease progression
- Combines
 - Information propagation on Twitter
 - Epidemiological model

Stat/ML/AI models: overview

Deep learning



Density estimation



- Capture non-linear patterns in high-dimensional data with minor assumptions
- Leverage multiple sources of data of variety of modalities

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

Summary of work until now

Mechanistic models

- Explicit causal mechanisms of epidemic spread
- Excel in understanding and what-if analysis

Statistical/ML/AI models

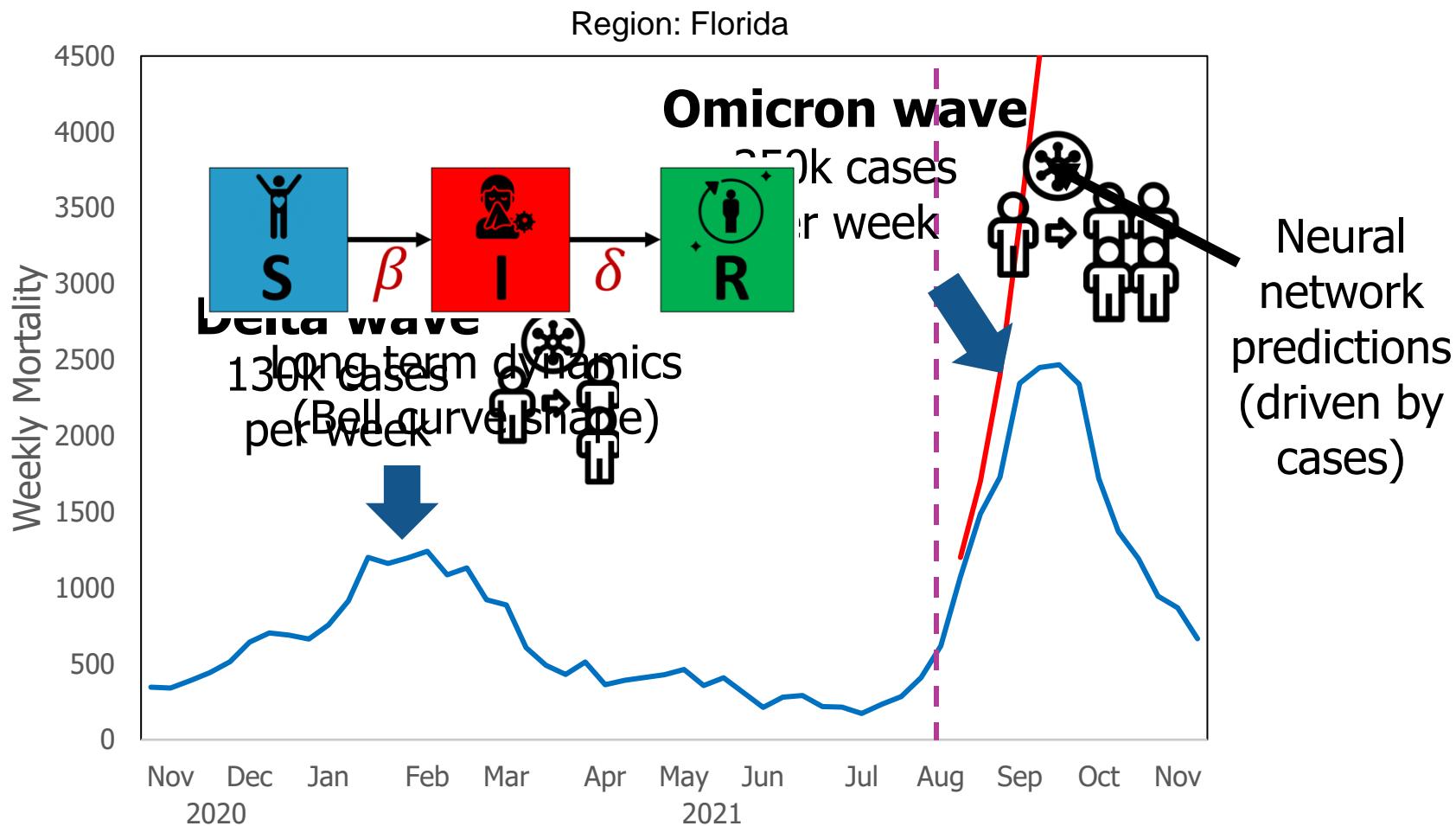
- Learn from data with little constraints
- Excel in short-term forecasting

Hybrid Models

- Marry the expressivity of statistical models with theory-based mechanisms of mechanistic models

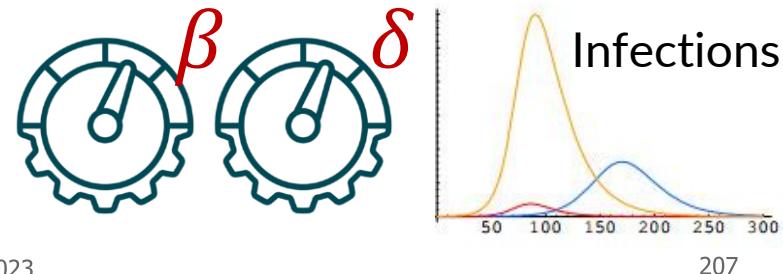
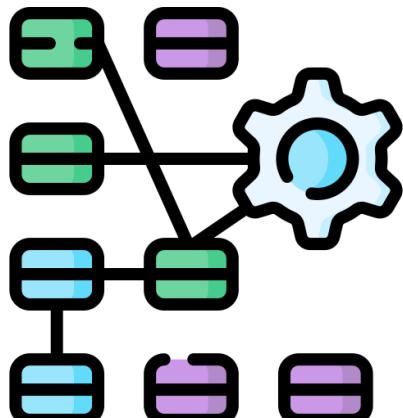


Problems with Current AI Methods



Problems with Mechanistic Models

- Non-trivial to add novel data sources
 - Need to explicitly model the relationships
 - Often laborious & adds overhead complexity
- Calibration complicated and cost-prohibitive
 - A small error can result in very different prediction
 - Takes too many resources to calibrate large models



Hybrid Models (Outline)

- Approaches:
 1. Mechanistic model with statistical components
 2. Priors from mechanistic models inform statistical model
 3. Wisdom of crowds

Hybrid Models (Outline)

- Approaches:
 1. **Mechanistic model with statistical components**
 2. Priors from mechanistic models inform statistical model
 3. Wisdom of crowds

[H1] Mechanistic Model with Statistical Components

- Mechanistic aided by stat/ML components
- Objectives:
 - Incorporate data into mech. calibration
 - Account for modeling limitations
- Ideas:
 - Data assimilation
 - Discrepancy modeling
 - Estimate parameters of a mechanistic model from features

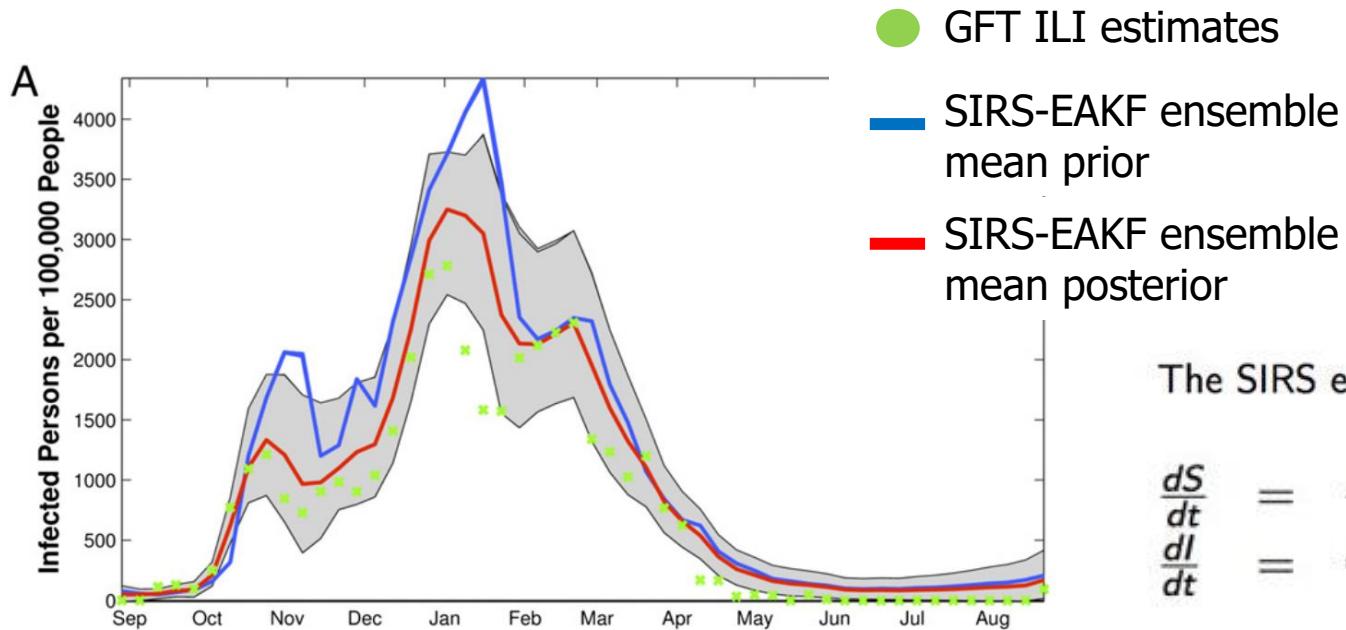
[H1] Mechanistic Model with Statistical Components

- Mechanistic aided by stat/ML components
- Objectives:
 - Incorporate data into mech. calibration
 - Account for modeling limitations
- Ideas:
 - **Data assimilation**
 - Discrepancy modeling
 - Estimate parameters of a mechanistic model from features

Idea 1: Data Assimilation

[Shaman and Karspeck, PNAS 2012]

- Incorporates Google Flu Trends ILI estimates via ensemble adjustment Kalman filter (EAKF)



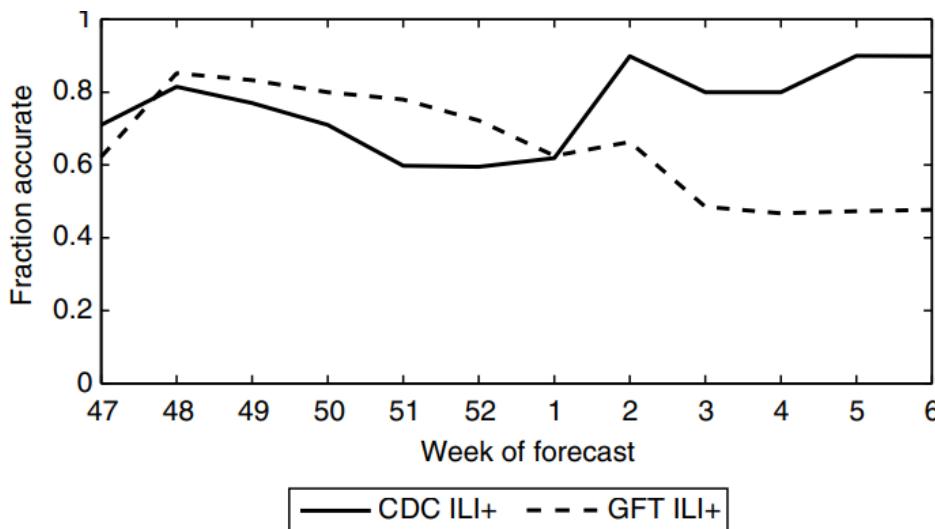
The SIRS equations are given by:

$$\begin{aligned}\frac{dS}{dt} &= \frac{N-S-I}{L} - \frac{\beta(t)SI}{N} - \alpha \\ \frac{dI}{dt} &= \frac{\beta(t)SI}{N} - \frac{I}{D} + \alpha\end{aligned}$$

Real-time forecasting results

[Shaman+, Nat. Comm. 2013]

- First example of real-time forecasting
- Evaluated peak timing and peak value prediction
- By week 52, prior to peak for majority of cities, 63% of forecasts were accurate



[H1] Mechanistic Model with Statistical Components

- Mechanistic aided by stat/ML components
- Objectives:
 - Incorporate data into mech. calibration
 - Account for modeling limitations
- Ideas:
 - Data assimilation
 - **Discrepancy modeling**
 - Estimate parameters of a mechanistic model from features

Idea 2: Discrepancy modeling

Ex. 1: Mech. correction via Bayesian

[Osthus+, Bay. Analysis 2019]

- Refines/corrects mechanistic predictions with a hierarchical Bayesian model.
- Refinement components:
 - Common discrepancy: across all flu seasons
 - Individual discrepancy: season-specific

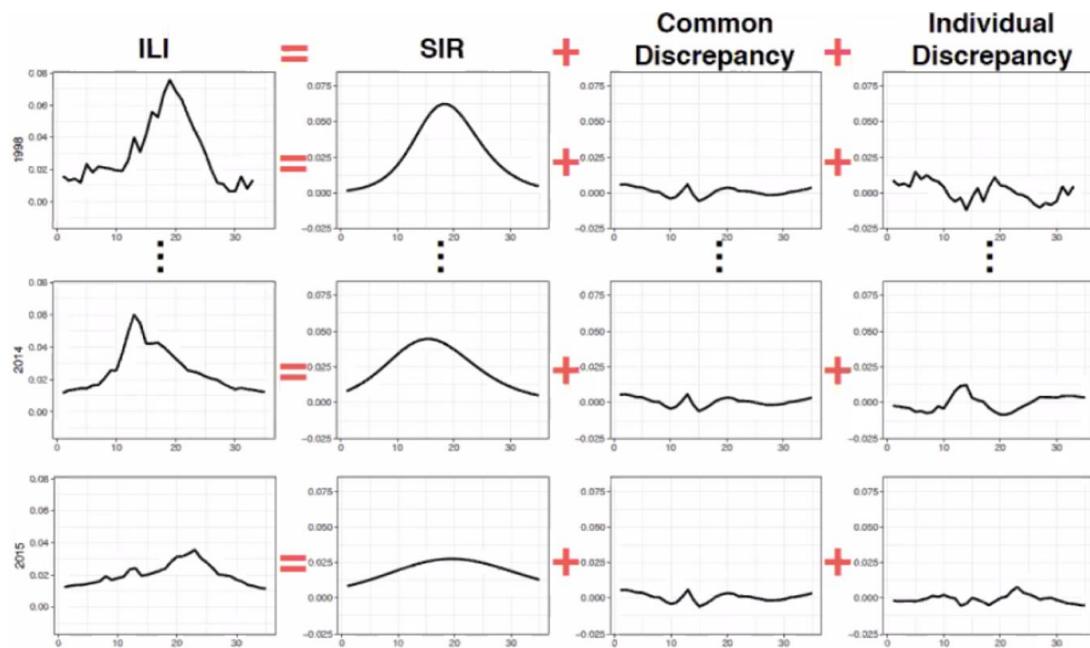


Figure credit: Sara Del Valle, LANL

Bayesian modeling of ILI

- Observable ILI:

$$y_{j,t} \sim \text{Beta}(\lambda\pi_{j,t}, \lambda(1 - \pi_{j,t})), \quad E(y_{j,t}) = \pi_{j,t}, \quad \text{SD}(y_{j,t}) = \left(\frac{\pi_{j,t}(1 - \pi_{j,t})}{1 + \lambda} \right)^{0.5}.$$

- True proportion of population w/ ILI, $\pi_{j,t}$:

$$\text{logit}(\pi_{j,t}) = \text{logit}(I_{j,t}) + \mu_t + \delta_{j,t}.$$

SIR model

$$\frac{dS}{dt} = -\beta SI,$$

$$\frac{dI}{dt} = \beta SI - \gamma I,$$

$$\frac{dR}{dt} = \gamma I,$$

Across all seasons

$$\mu_T \sim N(0, \sigma_{\mu_T}^2),$$

$$\mu_t | \mu_{t+1} \sim N(\mu_{t+1}, \sigma_{\mu}^2).$$

Reverse random walk

Season-specific

$$\delta_{j,T} = -\text{logit}(I_{j,T})$$

$$\delta_{j,t} | \delta_{j,t+1} \sim N(\alpha_j \delta_{j,t+1}, \sigma_{\delta,j}^2)$$

Reverse random walk

Extensions

- Ex.2: Dante [Osthus and Moran, Nat. Comm. 2021]
 - Top model in 2018/19 CDC FluSight Challenge
 - Team: Los Alamos National Lab (LANL)
 - Multi-scale (hierarchical) consistency:
 - States -> HHS Regions
 - Joint modeling of all regions/seasons:
 - Shares information across all of them
- Ex. 3: Inferno [Osthus, PLOS Comput. Bio 2022]
 - Accelerates Dante via dropping joint modeling
 - Enables parallelization

Remember Real-time Forecasting

Input: Prediction weeks W , forecasting horizon K in weeks, time series of features X_t until time t , time series of target Y_t until time t .

FOR w in W : // for each prediction week

 FOR k in K : // for each week ahead

1. Pre-process data X_t and Y_t
2. Train model M with gradient-based optimization
3. Forecast target with M

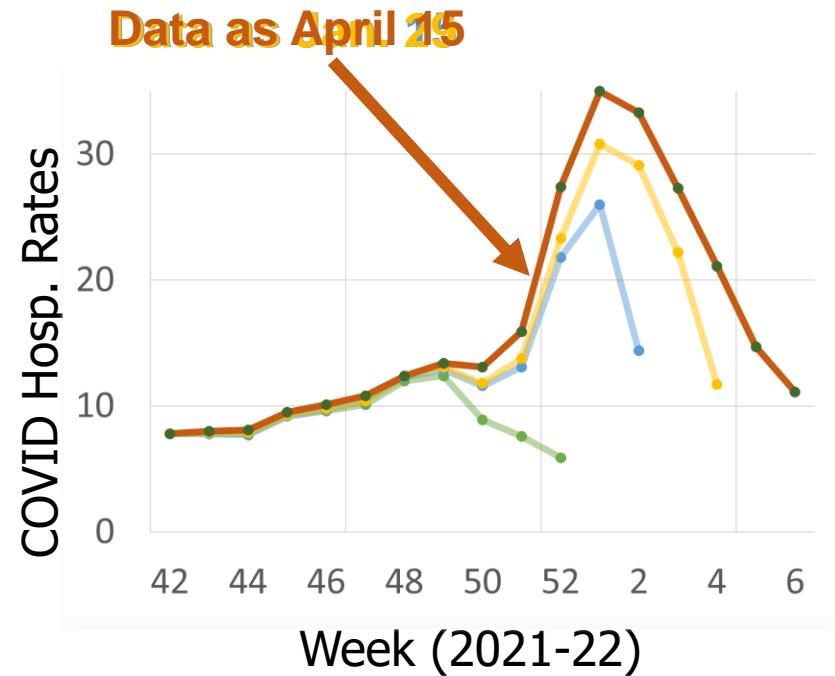
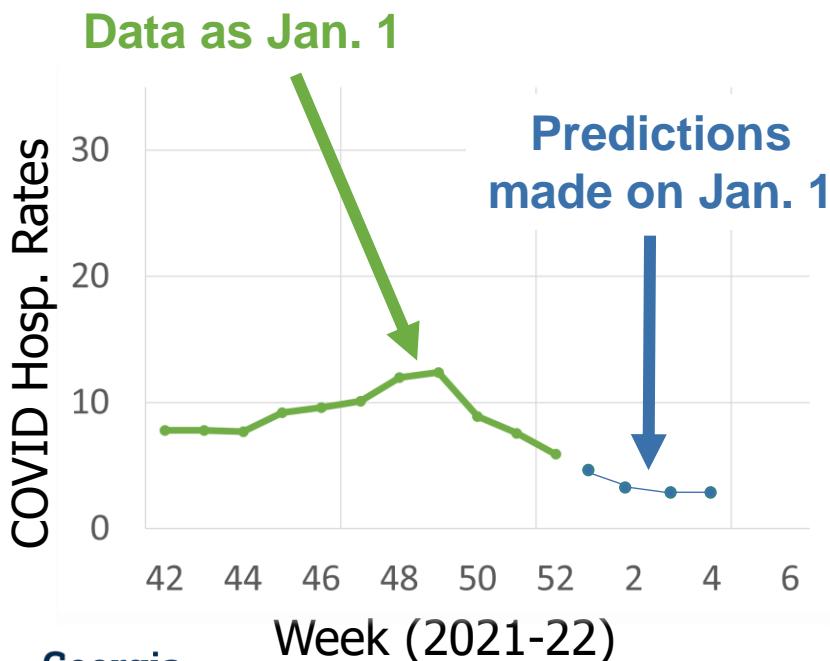
 ENDFOR

ENDFOR

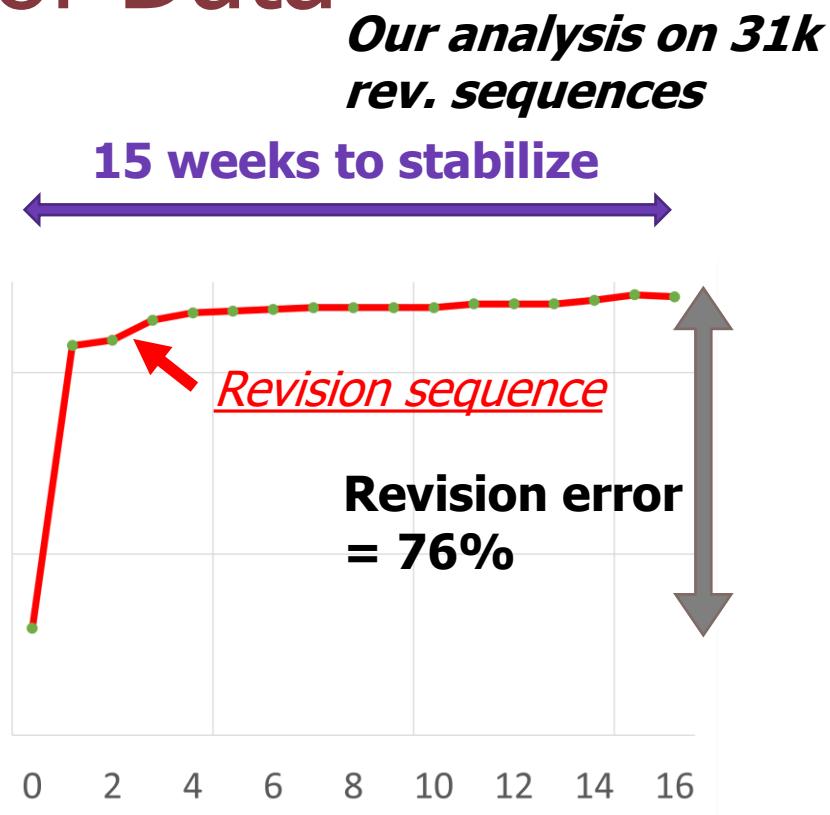
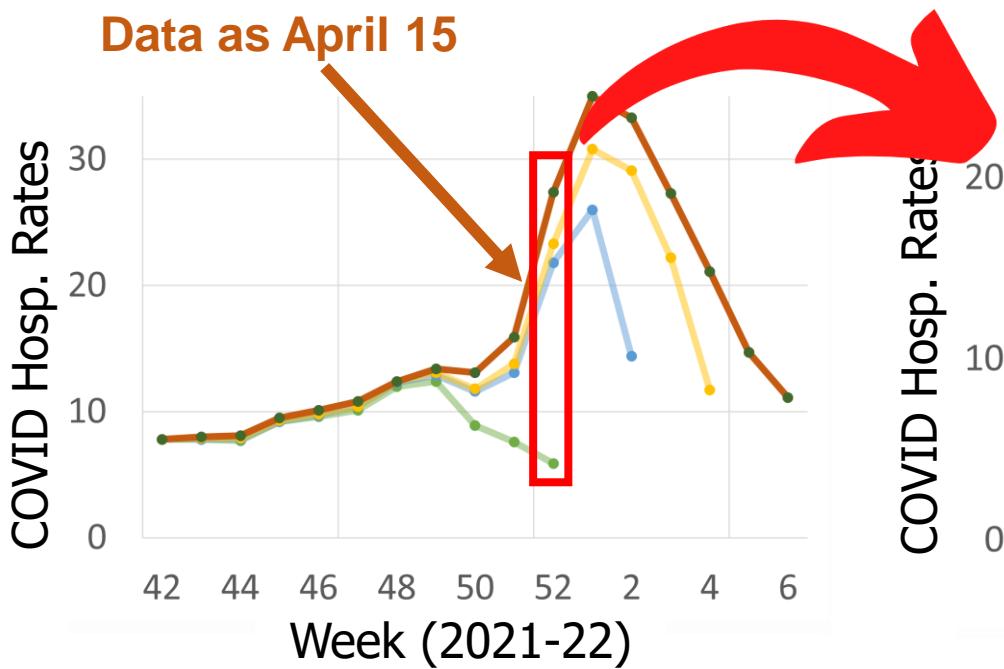
Issues Arising from Real-time Settings

- These data quality issues are not present in carefully controlled environments.

Example: Data revisions, reporting delays, missing values



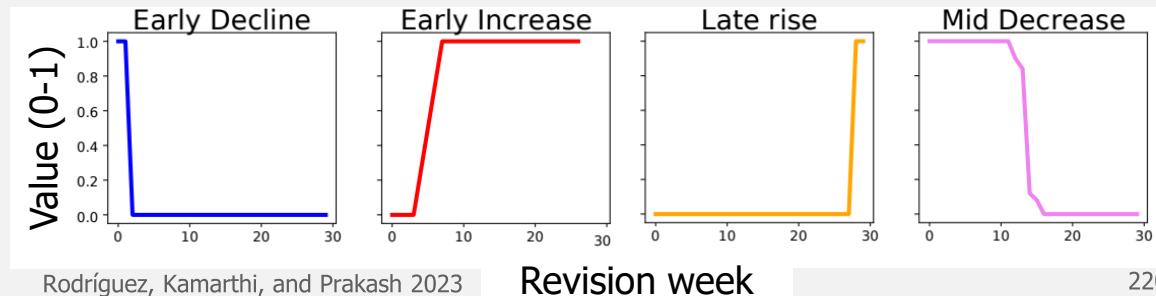
Analyzing Dynamics of Data Revisions



Significance

- Half signals with 30+% of revision error
- 4 weeks to stabilize on average

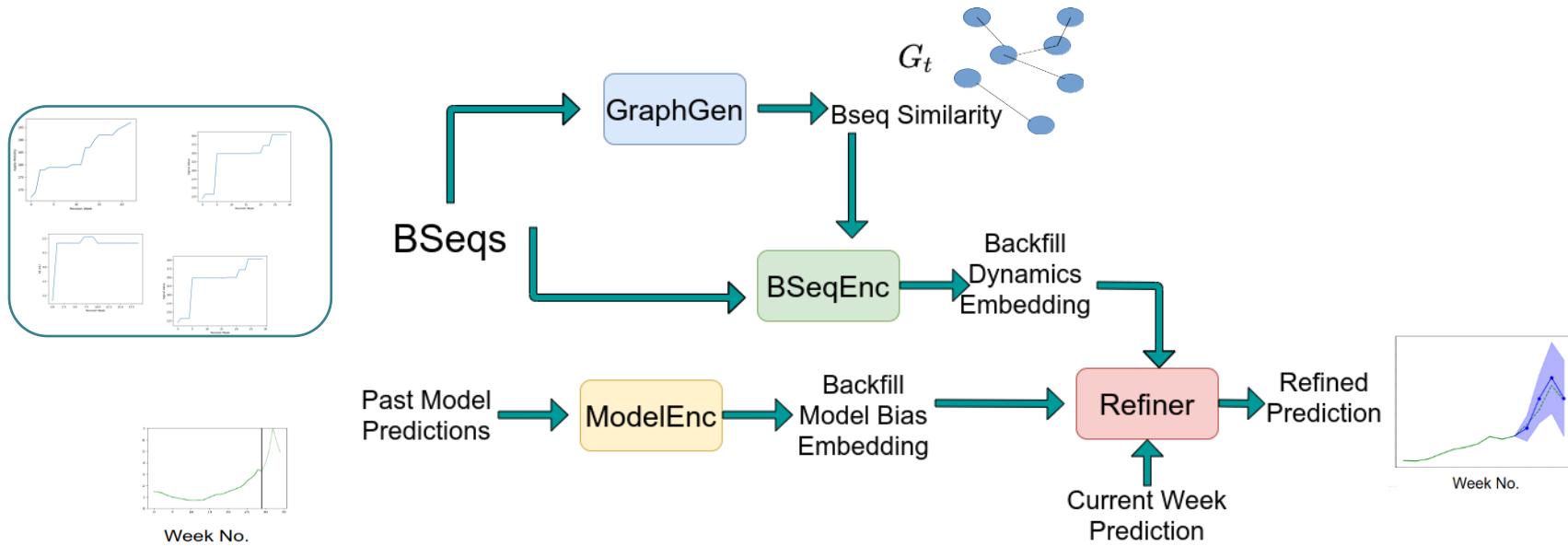
Diversity of revision dynamics



Ex. 2: Model-agnostic correction via data revision representations

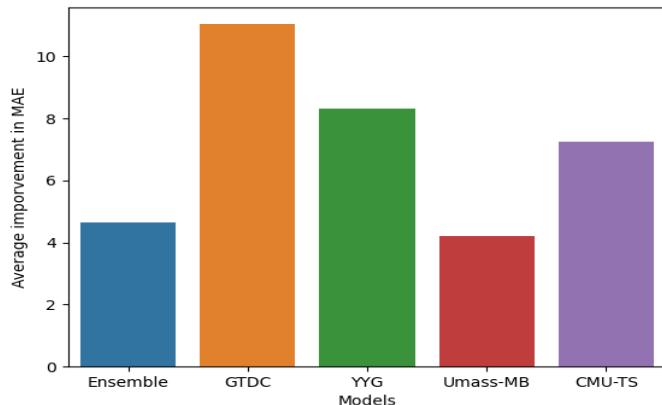
[Kamarthi+, ICLR 2022]

- (1) Learns backfill patterns
- (2) 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
- Wrapper for any model (mech. or stat.)



Takeaway: data quality issues can be helped with statistical correction

Try it out!
github.com/AdityaLab/Back2Future

[H1] Mechanistic Model with Statistical Components

- Mechanistic aided by stat/ML components
- Objectives:
 - Incorporate data into mech. calibration
 - Account for modeling limitations
- Ideas:
 - Data assimilation
 - Discrepancy modeling
 - **Estimate parameters of a mechanistic model from features**

Typical way: Based on laboratory experiments

[Shaman and Kohn, PNAS 2009]

Humidity-driven SIRS

- Flu reproduction number estimated based on laboratory experiments with humidity

The SIRS equations are given by:

$$\begin{aligned}\frac{dS}{dt} &= \frac{N-S-I}{L} - \frac{\beta(t)SI}{N} - \alpha \\ \frac{dI}{dt} &= \frac{\beta(t)SI}{N} - \frac{I}{D} + \alpha\end{aligned}$$

where the AH modulated reproductive number is given by

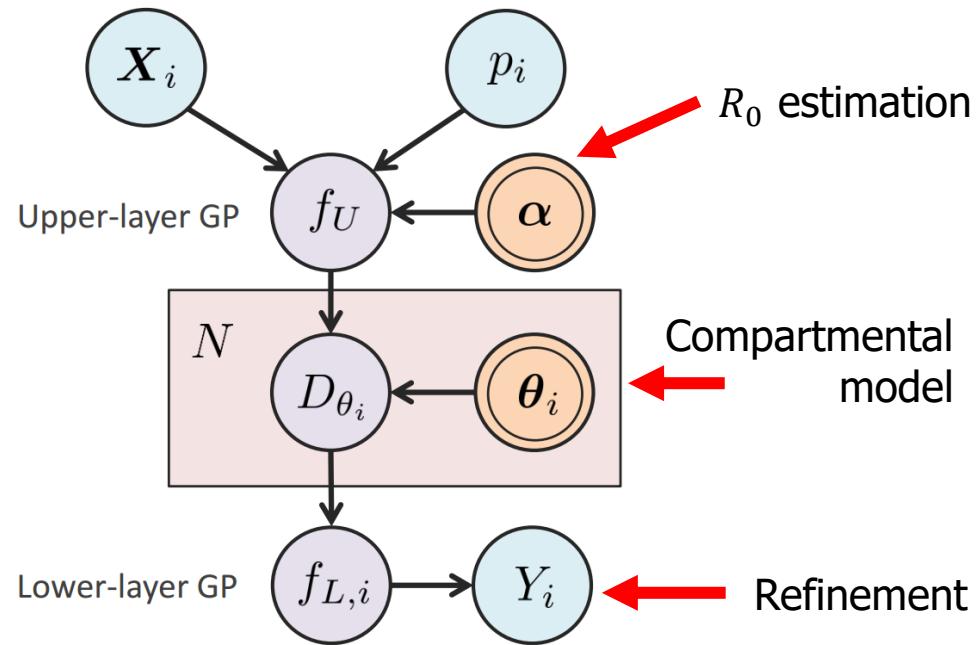
$$R_0(t) = \exp(a \times q(t) + b) + R_{0min}$$

where, $a = -180$ and $b = \log(R_{0max} - R_{0min})$. $q(t)$ is the time varying specific humidity.

Ex. 1: Compartmental Gaussian Processes

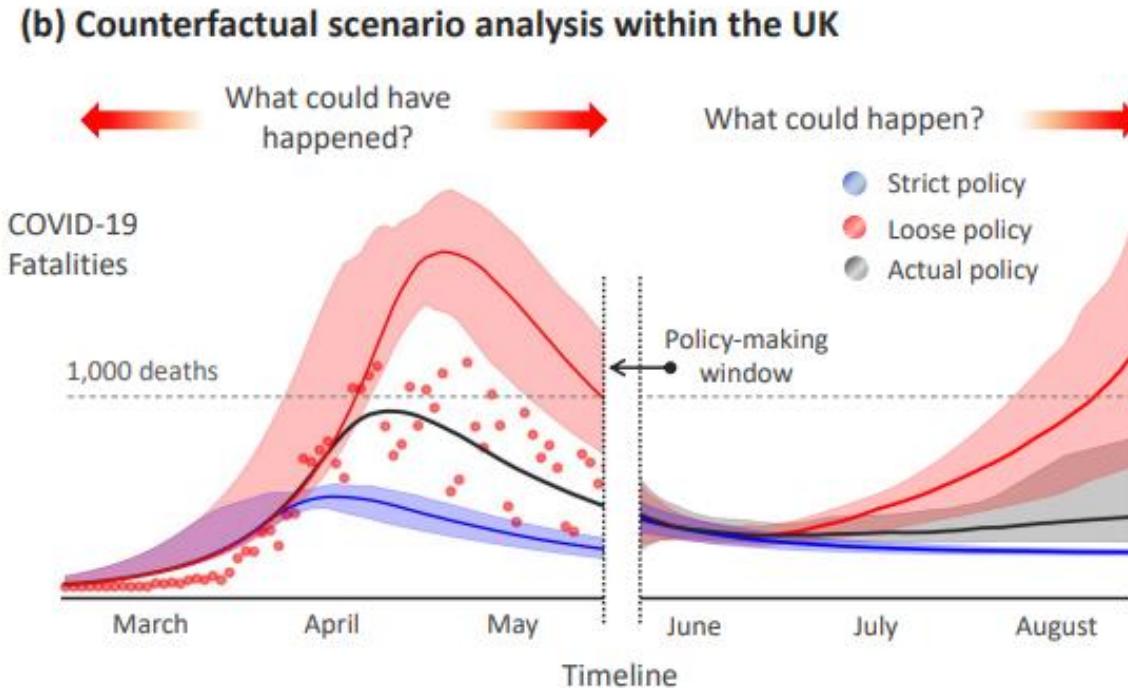
- Hierarchical two-layer Gaussian process (GP).
- Upper-layer GP uses country-specific features + policies in place to estimate R_0
- Lower-layer GP refines predictions

[Qian+ NeurIPS 2020]



Counterfactual based on new set of policies

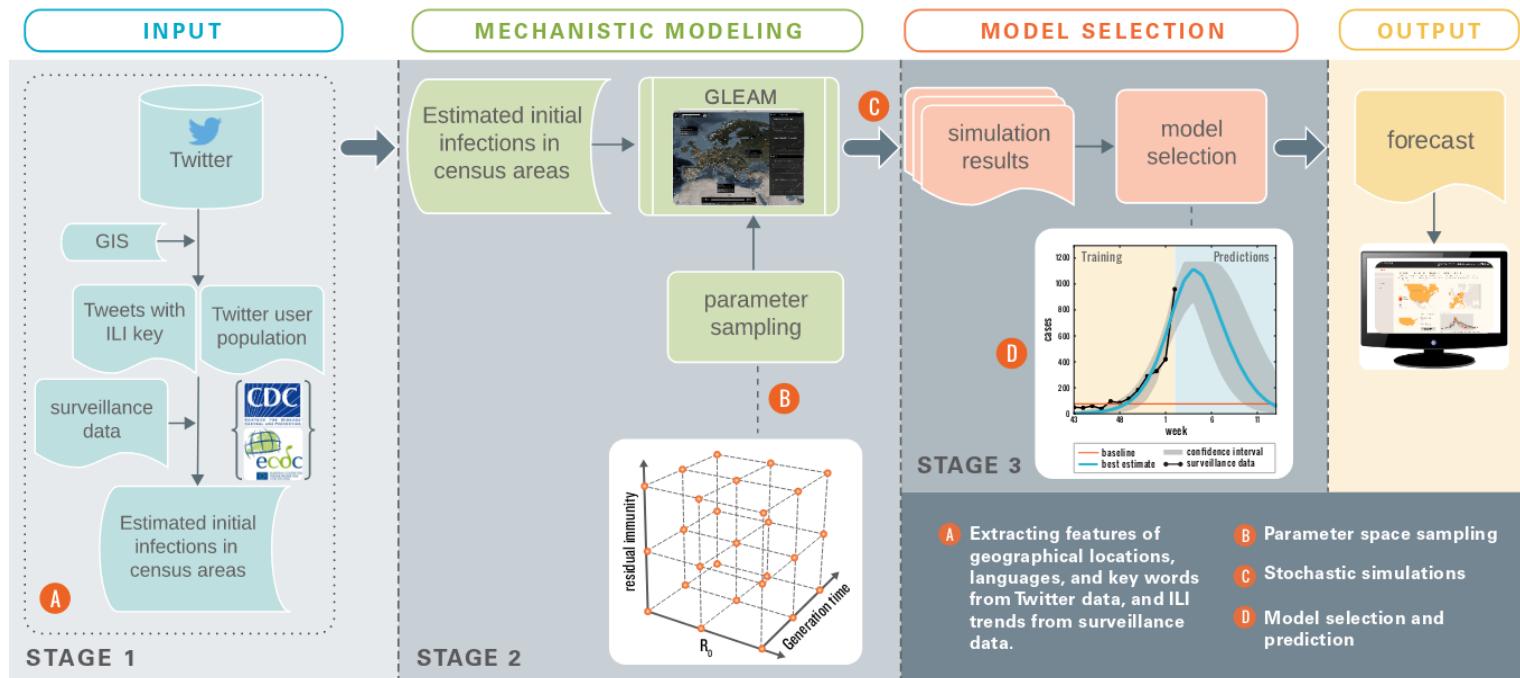
- What if we have a different input policies p_i ?



Ex. 2: Also with metapopulation models

[Zhang+, WWW 2017]

- Extends metapopulation model GLEAM by learning mechanistic initial conditions from geo-localized



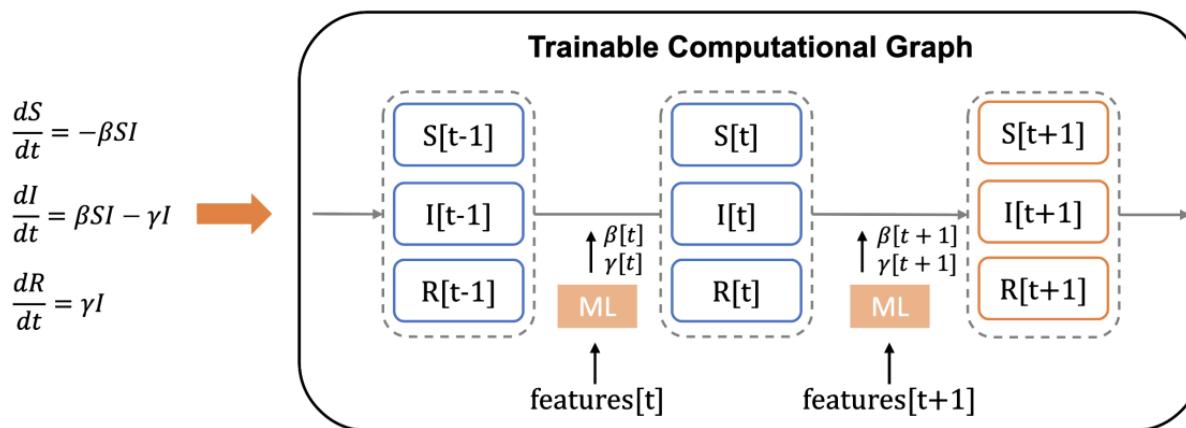
Recent idea: End-to-end learning

- Very popular in deep learning
- Differentiable modules take advantage of gradient-based optimization
- Recent work is exploring how to do this w/ mechanistic models
- Connections to:
 - Physics/biological simulators [[Gaw+](#), Sci. Reports 2019]
 - Systems biology [[Yazdani+](#) PLOS Comp. Bio 2020]
 - ...

Ex. 3: End-to-end Differentiable Learning with ODEs

[Arik+, NeurIPS 2020]

- Using additive encoders for time-varying features.



Rate variable	Covariates
β	Mobility, Interventions, Density, Past Counts
η	Census, Healthcare supply
γ	Census, Test count / pos. ratio, Past Counts
h, c, v, ρ, κ	Census, Econometrics, Healthcare supply

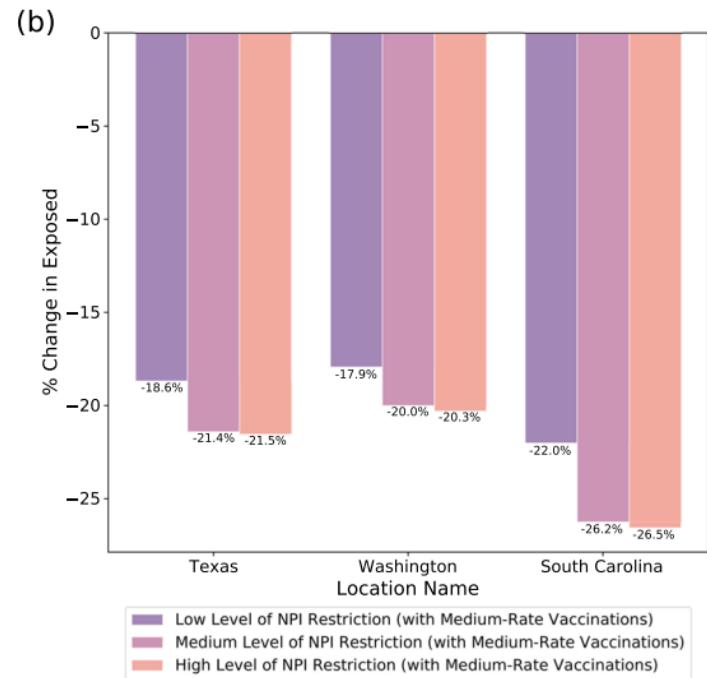
$$v_i[t] = v_{i,L} + (v_{i,U} - v_{i,L}) \cdot \sigma(c + b_i + \mathbf{w}^\top \text{cov}(v_i, t))$$

Update parameters via gradient-based optimization (RMSProp)

Ex. 4: Extension to what-if forecasting

[Arik+, npj Dig. Medicine 2021]

- What-if forecasting via time-varying features as NPI scenario
- NPI scenarios:
 - Mobility increase
 - State of emergency introduced



Incorporating priors from epidemiological knowledge

- Directional penalty regularization:
- Ex. 1: If the mobility increases, the average contact rates increase
- Ex. 2: If state of emergency (SoE) is introduced, contact rates decrease

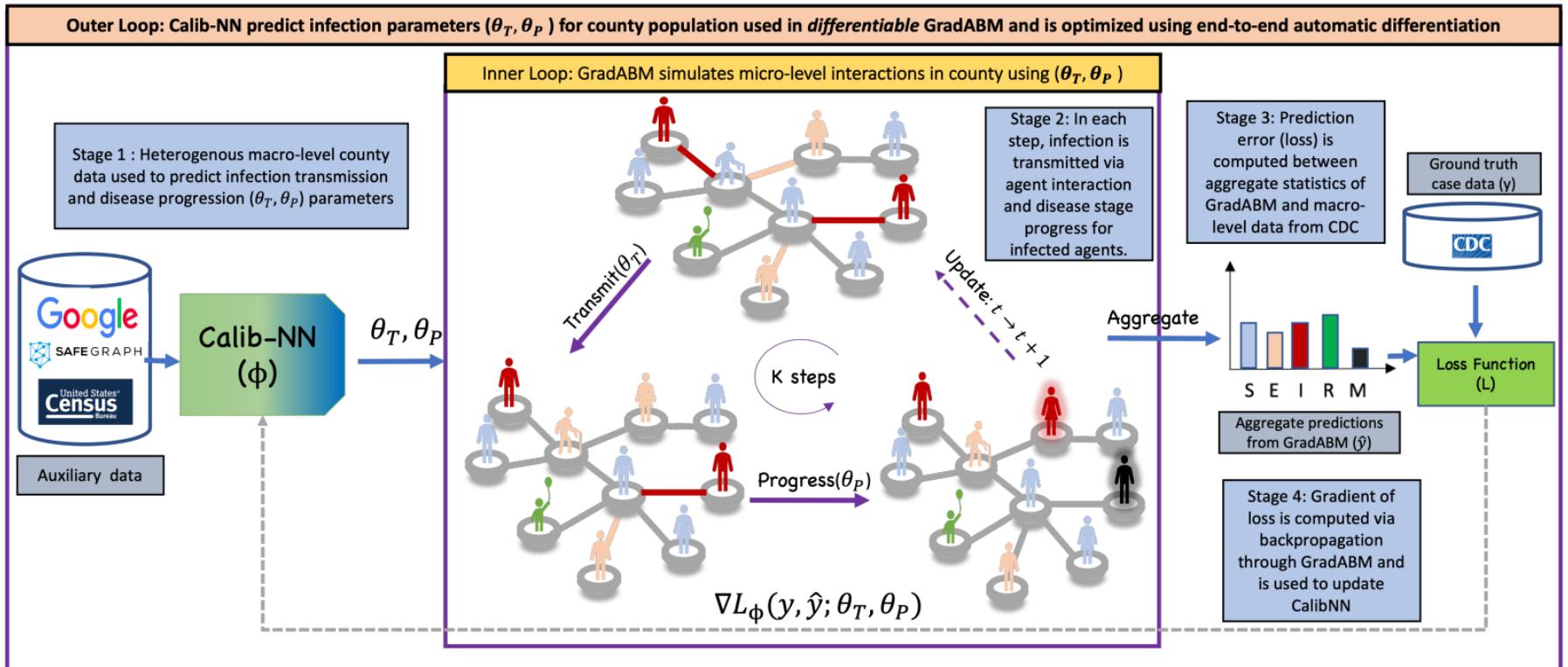
$$L_{dir} = \sum_{i \in \text{Mobility}} \max(-w_i, 0) + \sum_{j \in \text{NPIs or SoE}} \max(w_j, 0)$$

Ex. 5: End-to-end Differentiable Learning with ABMs

[Chopra and Rodríguez+, AI4ABM @ ICML 2022] Best paper award

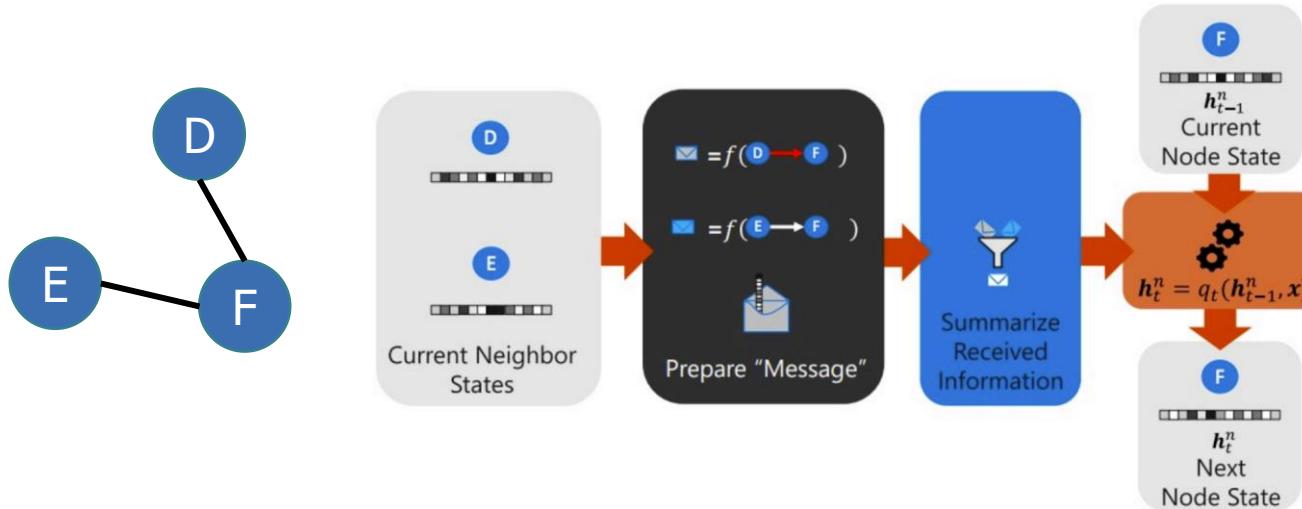
Stage 1: Param. prediction
Encoder-decoder GRU
(Calib-NN)

Stage 2: Disease transmission + progression
Message passing in graph neural network (GNN)
+ reparametrization trick



Reformulation of disease transmission and progression

- Transmission as message passing operation over sparse graphs (permutation invariance).



$$h_t^n = q_t \left(h_{t-1}^n, \bigcup_{\substack{k \\ n_j : n_j \rightarrow n}} f_t \left(h_{t-1}^n, k, h_{t-1}^{n_j} \right) \right)$$

Reformulation of disease transmission and progression

- Sample from infection probability via Gumble-softmax (reparameterization).
- Disease progression is a linear operation with stochastic steps

$$d_i^{t+1} = \text{Update}(X_i^t, \mathcal{N}_i, (X_j^t)_{j \in \mathcal{N}_i}, \theta_T^t, \theta_P^t), \quad (1)$$

where $\text{Update}(X_i^t, \mathcal{N}_i, (X_j^t)_{j \in \mathcal{N}_i}, \theta_T^t, \theta_P^t)$ is

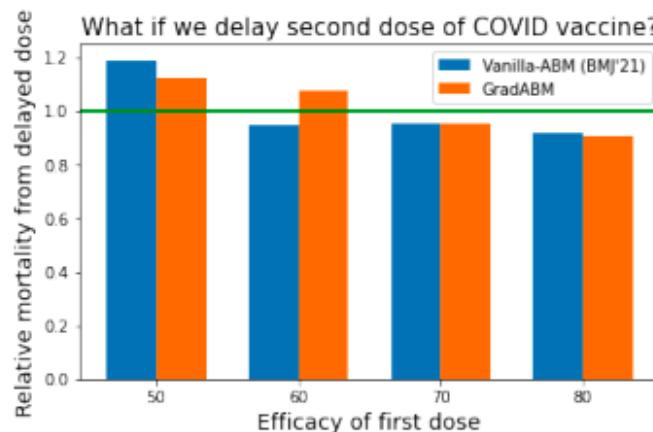
$$= \begin{cases} \text{Transmit}(X_i^t, \mathcal{N}_i, (X_j^t)_{j \in \mathcal{N}_i}, \theta_T^t), & \text{if } d_i^t = S, \\ \text{Progress}(X_i^t, \theta_P^t), & \text{if } d_i^t \in \{E, I\}. \end{cases} \quad (2)$$

Benefits

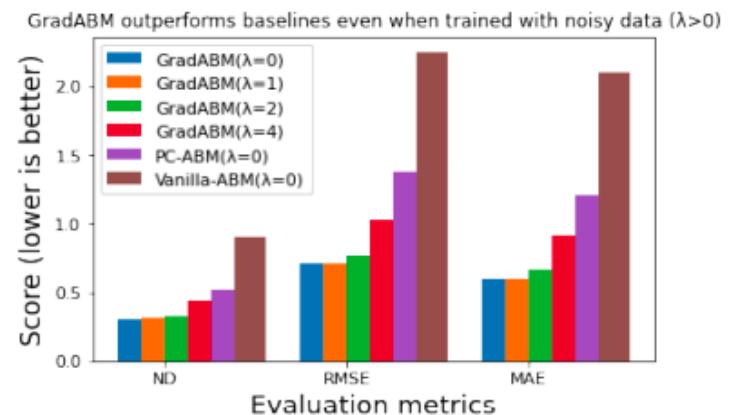
Forecasting

Model	COVID-19			Influenza		
	ND	RMSE	MAE	ND	RMSE	MAE
Vanilla-ABM [48]	8.75	689.92	270.13	0.57	2.03	1.72
PC-ABM [5]	2.21 ± 1.36	121.87 ± 63.97	68.20 ± 41.84	0.59 ± 0.02	2.17 ± 0.05	1.77 ± 0.05
GRADABM	0.97 ± 0.18	50.99 ± 12.12	30.02 ± 5.60	0.41 ± 0.02	1.47 ± 0.06	1.22 ± 0.06
GRADABM (w/o TL)	1.26 ± 0.43	78.22 ± 78.22	38.74 ± 13.35	0.41 ± 0.02	1.47 ± 0.06	1.22 ± 0.06
GRADABM (w/o TL, w/o CALIBNN)	2.39 ± 0.35	205.14 ± 42.56	73.66 ± 10.88	0.88 ± 0.14	2.97 ± 0.44	2.64 ± 0.43

What-if analysis



Robustness



Hybrid Models (Outline)

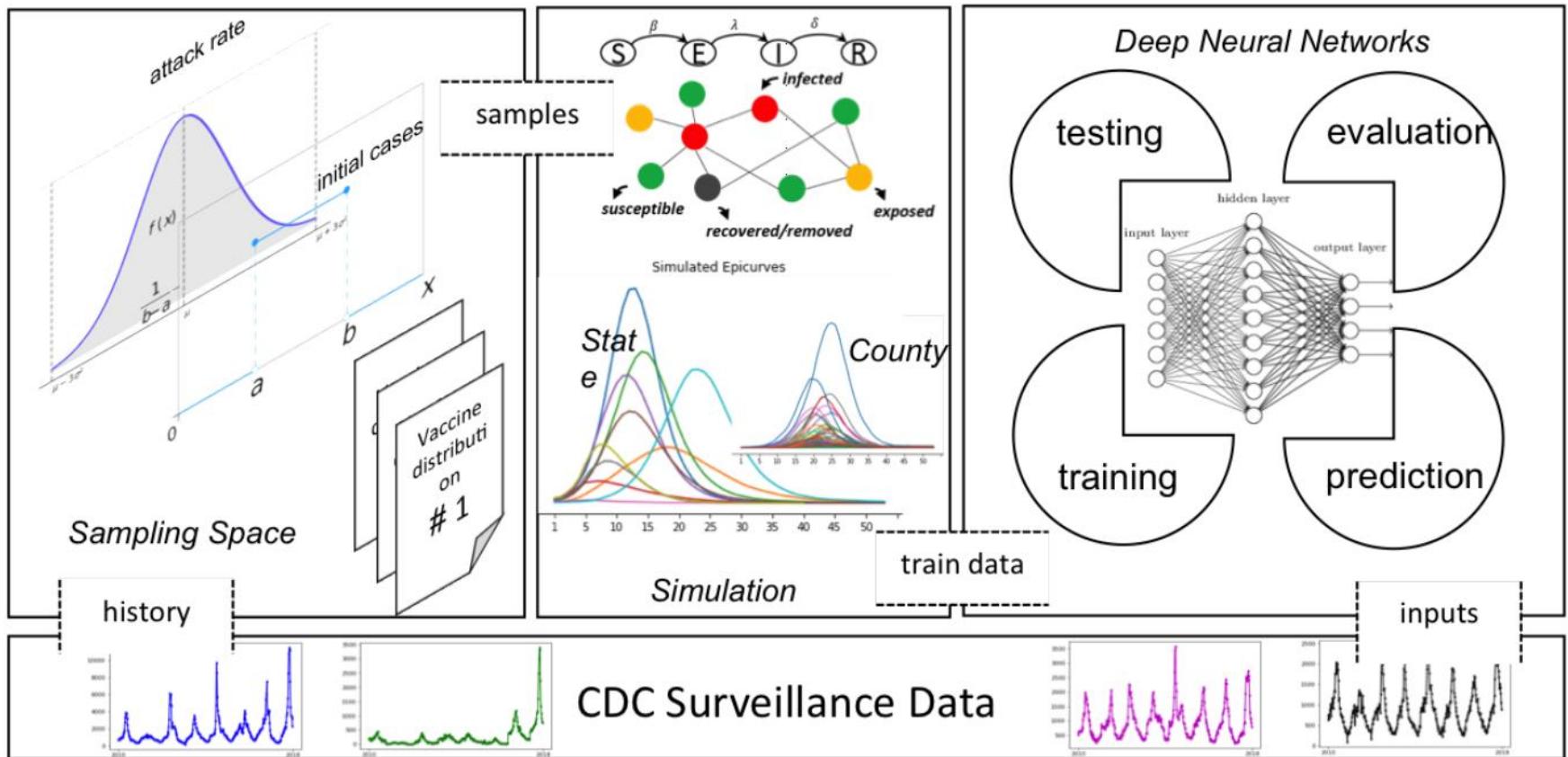
- Approaches:
 1. Mechanistic model with statistical components
 2. **Priors from mechanistic models inform statistical model**
 3. Wisdom of crowds

[H2] Priors from mechanistic models inform statistical model

- Statistical model has some prior knowledge coming from mechanistic model
- Objectives:
 - Address data scarcity
 - Include knowledge on mechanisms of disease spread

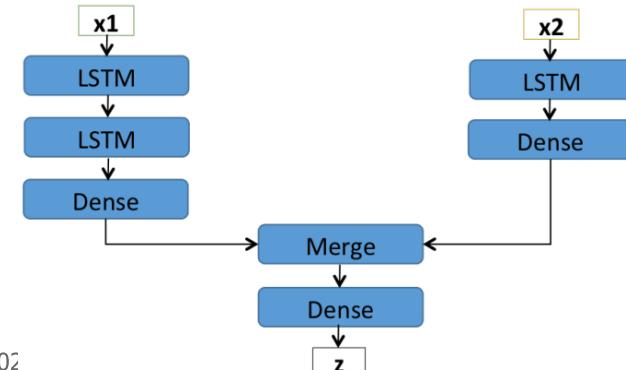
Ex. 1: Deep learning w/ with Synthetic Information (DEFSI)

[Wang+, AAAI 2019]



Major components of DEFSI framework

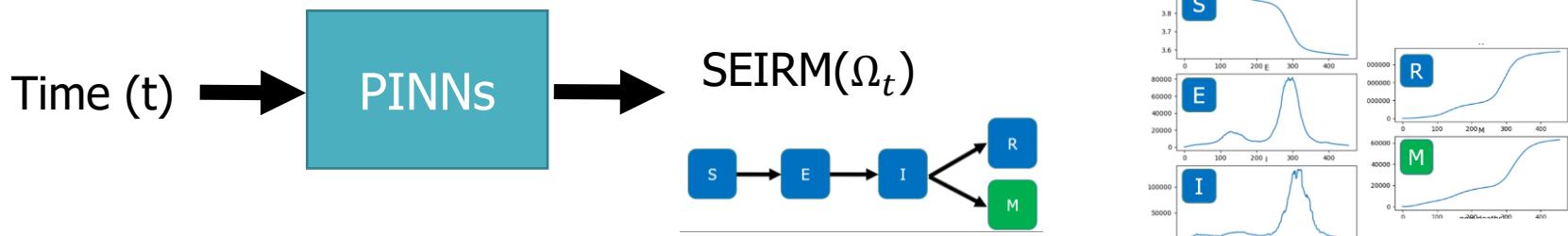
- (1) Disease model parameter space construction
 - Agent-based simulator **EpiFast** w/ SEIR
 - Parameters from literature [[Marathe+](#), PLOS One 2011]
- (2) Synthetic training data generation
 - High-resolution (more granular than reported by CDC)
- (3) Deep neural network training and forecasting
 - RNN 1: Within-season
 - RNN 2: Between-season



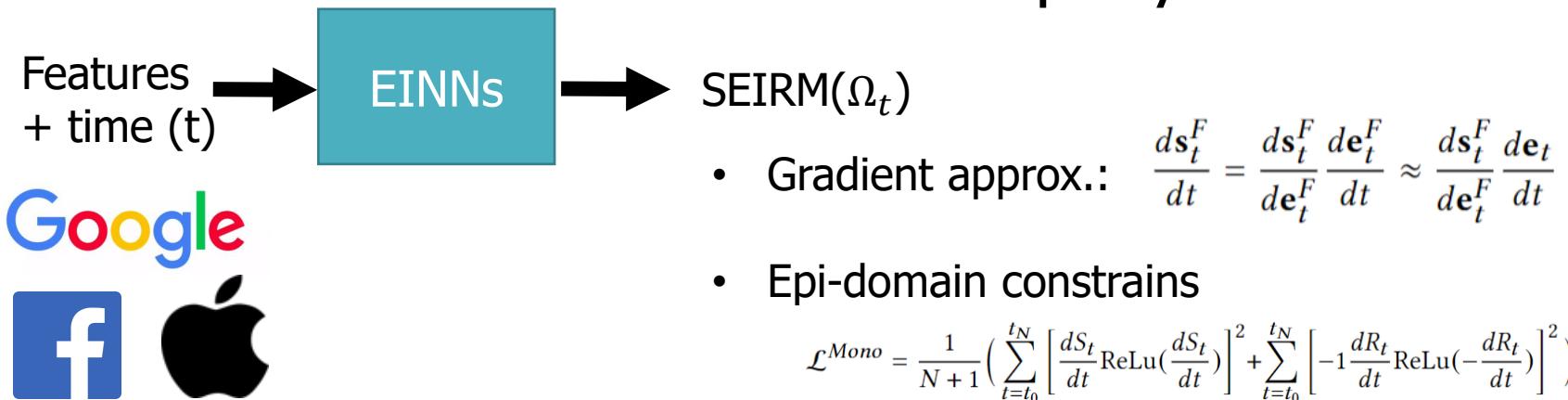
Ex. 2: Epidemiologically-informed Neural Networks (EINNs)

[Rodríguez+, arXiv 2022]

- Physics-informed neural networks: [Raissi+, Comp Phys 2019]



- EINNs connect features to latent epi dynamics



- Overcome spectral bias:
 $\Gamma(v) = [\cos(2\pi\mathbf{B}v), \sin(2\pi\mathbf{B}v)]^{T^{242}}$
- ...



Results: extend capabilities of both mechanistic and ML models

Model	Short-term (1-4 wks)			Long-term (5-8 wks)			Trend correlation
	NRMSE1	NRMSE2	ND	NRMSE1	NRMSE2	ND	
RNN	1.09	0.50	0.86	1.19	0.53	0.96	0.08
SEIRM	2.35	1.13	1.36	7.14	2.99	3.11	0.53
GENERATION	0.79	0.35	0.60	0.93	0.40	0.74	-0.01
REGULARIZATION	1.05	0.48	0.81	1.19	0.53	0.97	0.09
ENSEMBLING	0.91	0.41	0.68	0.93	0.40	0.69	-0.01
PINN (time module standalone)	0.84	0.38	0.64	0.93	0.40	0.72	0.24
EINN-NoGradMatching	0.82	0.36	0.61	0.89	0.38	0.68	0.04
EINN	0.54	0.24	0.38	0.85	0.37	0.66	0.46

Summary of results

	1-4 week accuracy	5-8 week accuracy	Trend correlation
RNN	✓	✗	✗
SEIRM	✗	✗	✓
EINN	✓	✓	✓

Hybrid Models (Outline)

- Approaches:
 1. Mechanistic model with statistical components
 2. Priors from mechanistic models inform statistical model
 - 3. Wisdom of crowds**

[H3] Wisdom of Crowds (WoC)

- Leverages multiple predictions from different sources
- Objectives:
 - Account for limitations of a single source of predictions
- Ideas:
 - Expert predictions
 - Prediction markets
 - Ensembles

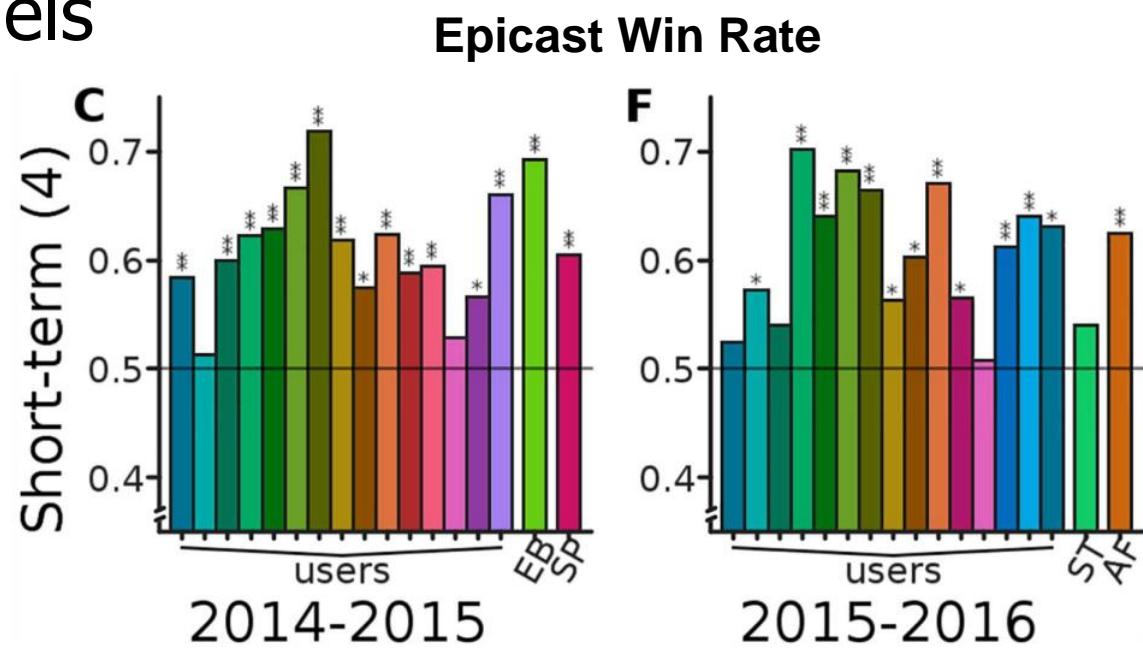
[H3] Wisdom of Crowds (WoC)

- Leverages multiple predictions from different sources
- Objectives:
 - Account for limitations of a single source of predictions
- Ideas:
 - **Expert predictions**
 - Prediction markets
 - Ensembles

Ex. 1: Expert predictions

[Farrow+, PLoS Comput Biol 2017]

- Epicast system to collect crowd-sourced predictions for influenza and chikungunya
- Aggregated predictions > top-performing statistical models



Ex. 2: Experts vs laypeople

[Recchia+, PLOS One 2021]

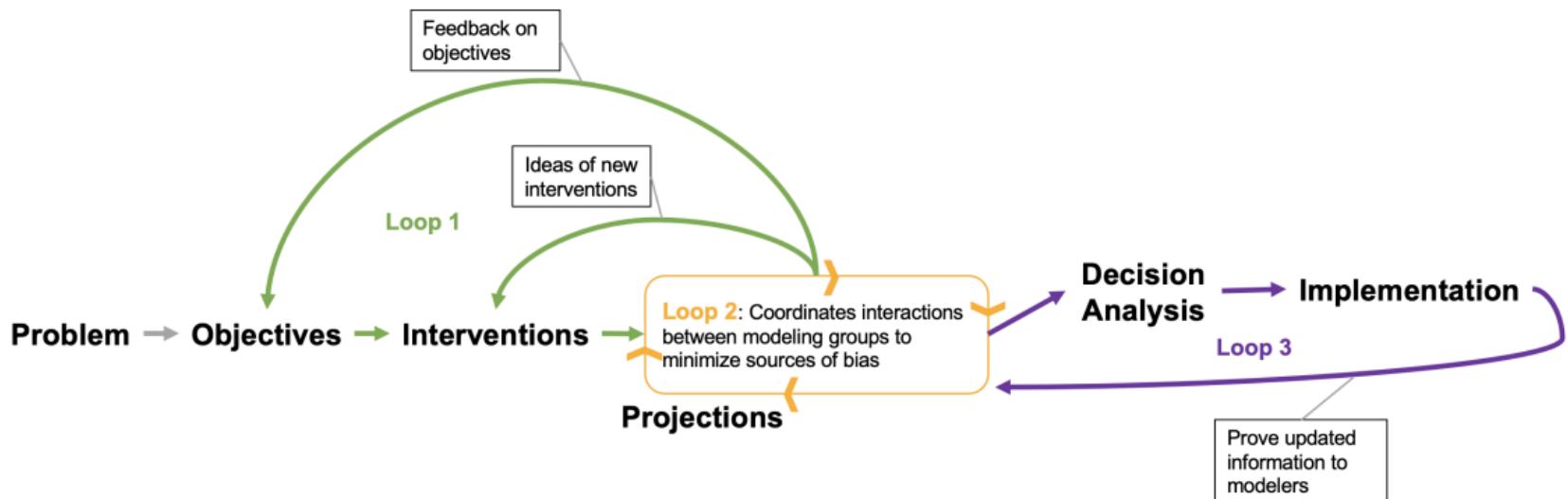
Table 1. Questions asked of participants with corresponding forecast medians, median absolute deviation (MAD), median absolute error (MAE) and median relative error (MRE).

	Question 1	Question 2	Question 3	Question 4
Question	How many people in the country you're living in do you think will have died from COVID-19 by December 31st 2020?	How many people in the country you're living in do you think will have been infected by COVID-19 by December 31st 2020?	Out of every 1000 people who will have been infected by the virus worldwide, how many do you think will have died by December 31st 2020 as a result?	Out of every 1000 people who will have been infected by the virus in the country you're living in, how many do you think will have died by December 31st 2020 as a result?
How true outcome estimate was derived	Total number of “deaths within 28 days of positive test” having a date of death earlier than 1 Jan 2021	Number of infections implied by dividing the total number of COVID-19 deaths in the UK (left) by the UK infection fatality rate estimated by Imperial College COVID-19 response team in Oct 2020	1000 multiplied by the age-specific infection fatality rates estimated by the Imperial College COVID-19 response team in Oct 2020, weighted by worldwide age distribution	1000 multiplied by the UK infection fatality rate estimated by the Imperial College COVID-19 response team in Oct 2020
True outcome estimate	75,346	6,385,254	4.55	11.8
Experts, median (MAD)	30,000 (15,000)	4,000,000 (3,687,500)	10 (5)	9.5 (4.5)
High-numeracy nonexperts, median (MAD)	25,000 (10,000)	800,000 (700,000)	30 (20)	30 (22)
All nonexperts, median (MAD)	20,000 (10,000)	250,000 (247,000)	50 (45)	40 (35)
Expert MAE	45,346	5,585,254	5.45	6.80
High-numeracy nonexpert MAE	55,346	6,085,254	25.45	18.20
Nonexpert MAE	55,346	6,235,254	45.45	28.20

Ex. 3: Expert consensus

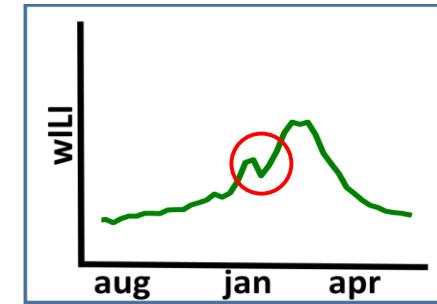
[Shea+, Science 2020]

- Each expert has their own models
- Multi-round framework
 - Exchange modeling assumptions + predictions
 - Designed to alleviate biases from group dominance



Ex. 4: Expert guidance

[Rodriguez+ epiDAMIK @ KDD 2020]



- Smoothness: Difference ϵ between the predicted value and its predecessor should be small

$$g(\theta) = E(|\hat{Y}_{t+1} - Y_t|) - \epsilon \leq 0$$

forecasted value predecessor

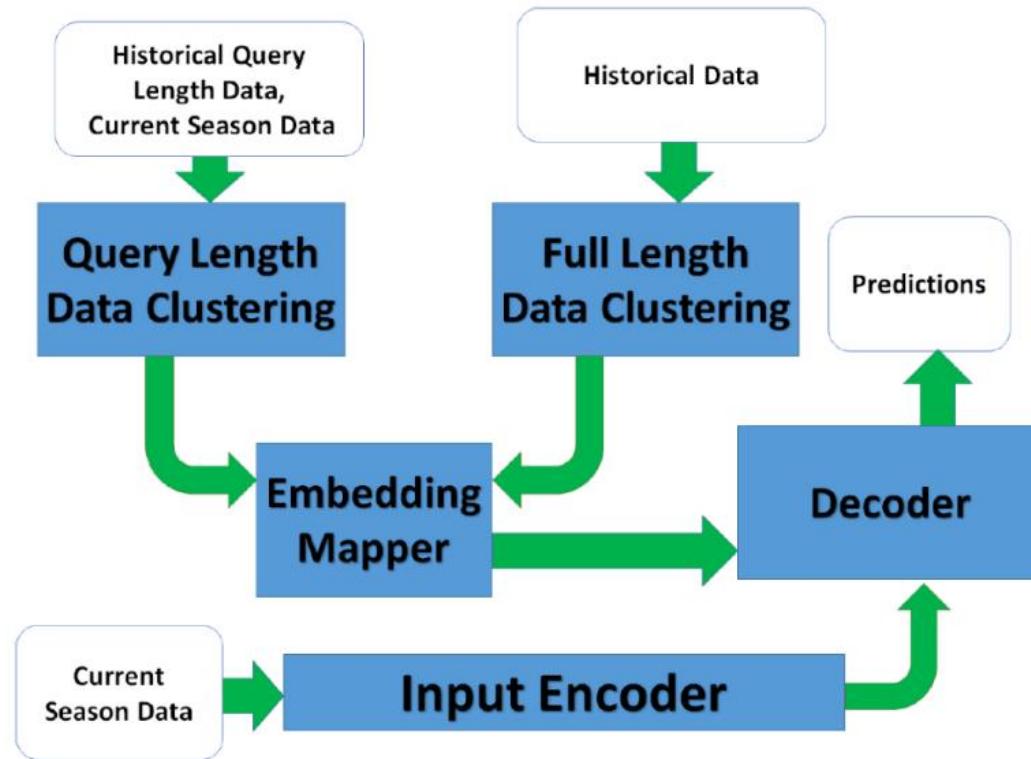
- Regional Equity: Quality of forecast μ between any two regions should be similar

$$g(\theta) = E(|\mu(\theta, t + 1, R_1) - \mu(\theta, t + 1, R_2)|) - \epsilon \leq 0$$

Squared error in Region 1 Squared error in Region 2

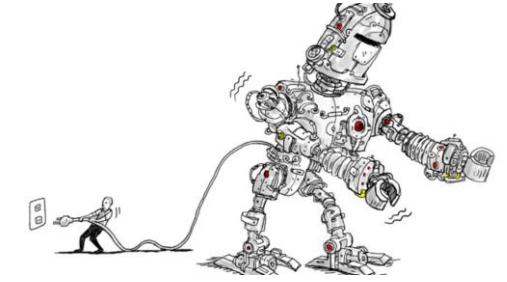
Recall EpiDeep

- Idea: Dynamic deep clustering for prediction with limited data

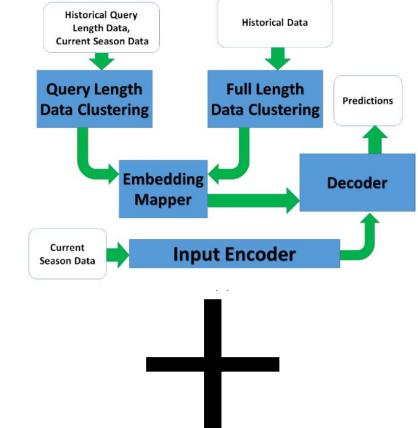


Guided-EpiDeep

- Seldonian Optimization Framework [Thomas+, 2019]
 - Proposed for AI safety
 - Precludes undesirable behavior of AI model by enforcing behavioral constraints in optimization
 - Has a safety test in unobserved data
- **Main idea:** Enforce Seldonian Framework on Epideep optimization to incorporate expert's guidance

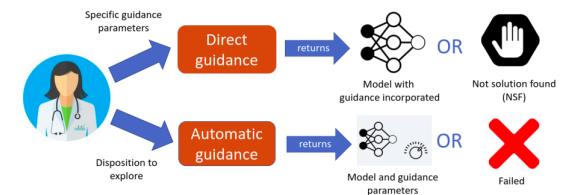


EpiDeep architecture



+

Guidance



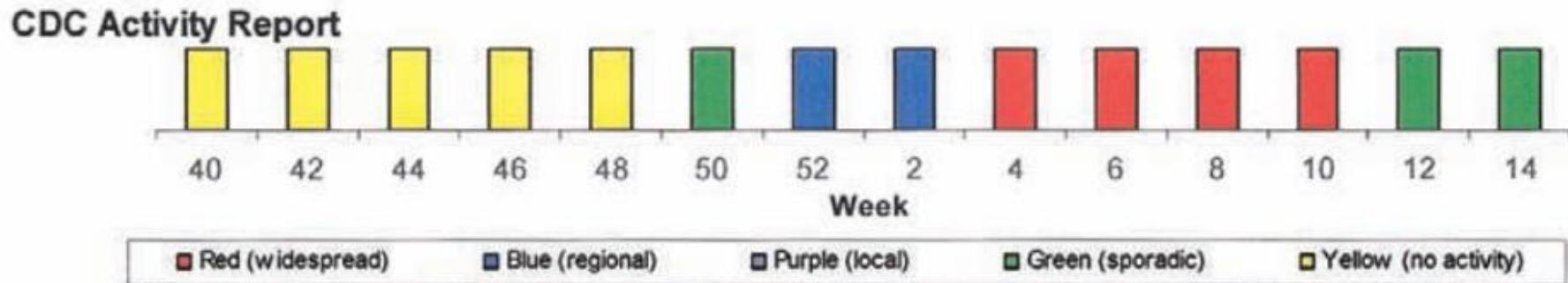
[H3] Wisdom of Crowds (WoC)

- Leverages multiple predictions from different sources
- Objectives:
 - Account for limitations of a single source of predictions
- Ideas:
 - Expert predictions
 - **Prediction markets**
 - Ensembles

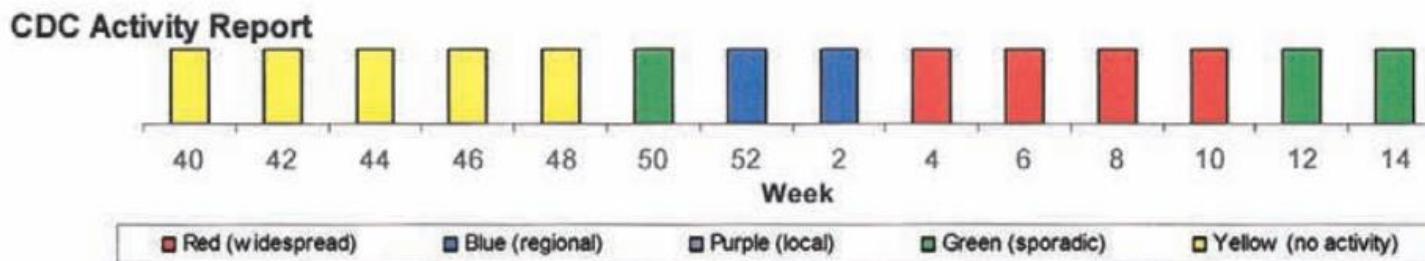
Ex. 1: Prediction markets

[Polgreen+, Clinical Infect. Diseases 2007]

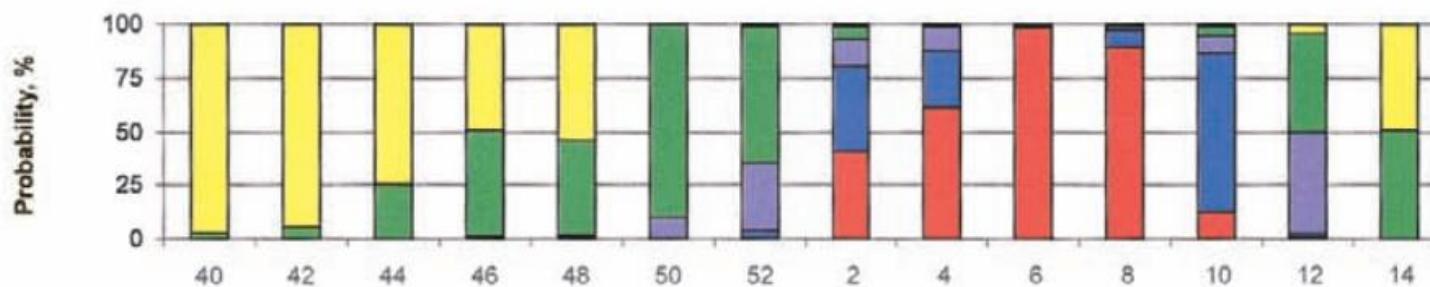
- Participants are healthcare worker or experts
 - 47 “traders” with monetary prizes of \$44.70 to \$213.19
- Up to 8 weeks ahead of ILI.
- Bet into 5 bins (colors) representing epidemic activity



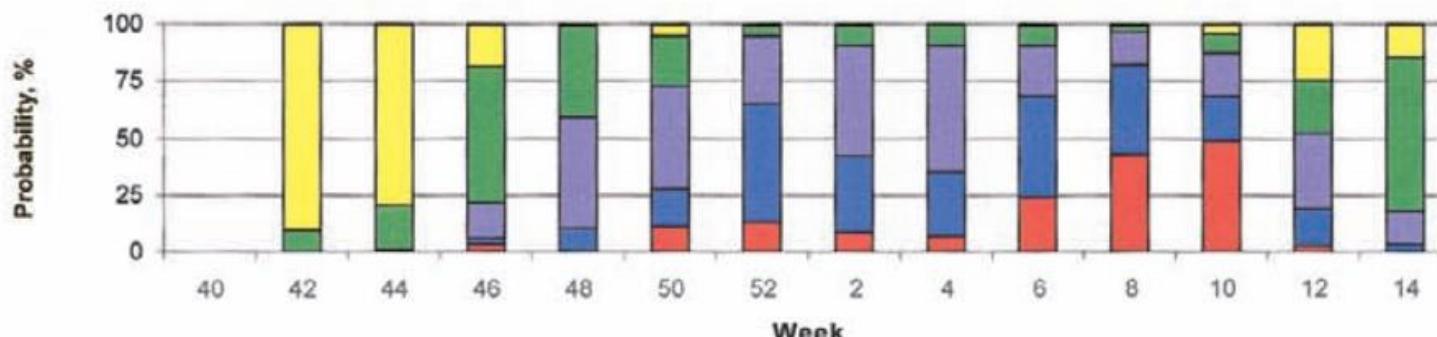
Results



Market Predictions, 0 Weeks in Advance



Market Predictions, 4 Weeks in Advance



Ex. 2: For pandemic response

[Sell+, BMC Public Health 2021]

The screenshot shows the Disease Prediction platform interface. At the top, it features the Johns Hopkins Bloomberg School of Public Health logo and the Center for Health Security. The main header reads "How many WHO member states will report more than 1000 confirmed cases of COVID-19 on or before April 2, 2020?".

Forecasting question: A red arrow points to the question itself.

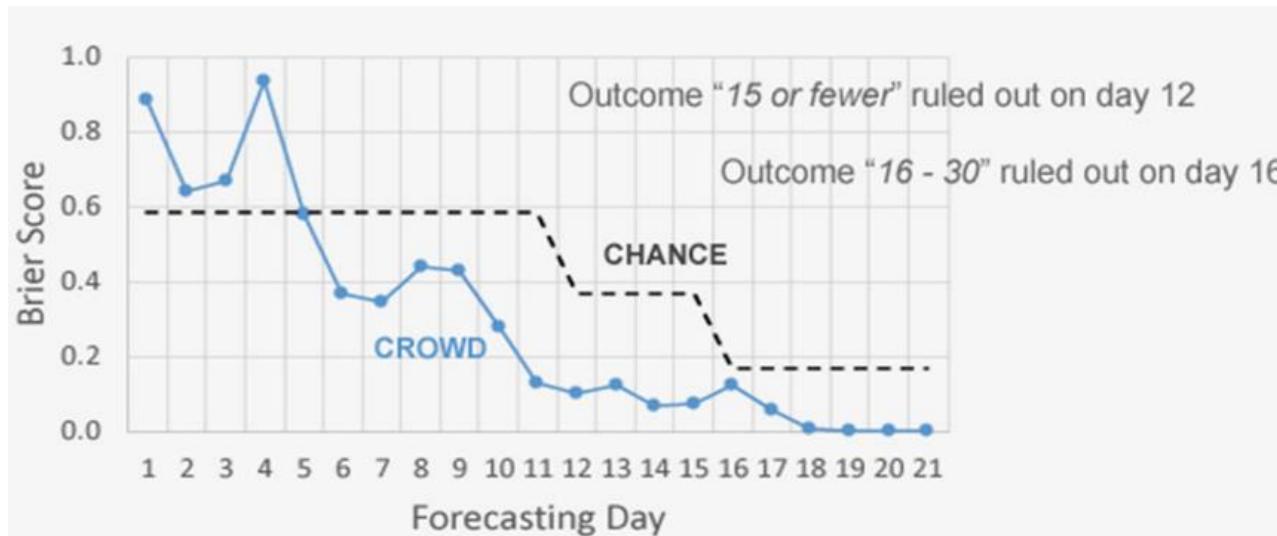
Participant's forecast input: A red arrow points to the "MY FORECAST" section where users can input probabilities for different ranges of WHO member states. The options are: "15 WHO member states or fewer" (0%), "16-30 WHO member states" (7%), "31-45 WHO member states" (41%), and "46 WHO member states or more" (52%).

Display of sub-optimal crowd forecast: A red arrow points to the "CROWD FORECAST" chart, which shows the distribution of forecasts from 91 forecaster. The chart includes a legend for colors: 2% (red), 4% (orange), 38% (green), and 56% (cyan). The x-axis shows dates from March 15 to March 23, 2020. The chart shows a general upward trend in the number of expected cases over time.

Discussion forum: A red arrow points to the "FORUM" section at the bottom, which contains a post by user "Azer" dated Mar 23, 2020 04:31. The post includes a new forecast and a rationale about naive extrapolation.

Study design and results

- Large scale study:
 - 562 participants, +15 months, 61 questions, 19 diseases including, Ebola, flu and COVID-19
- Aggregated predictions > any individual forecast



[H3] Wisdom of Crowds (WoC)

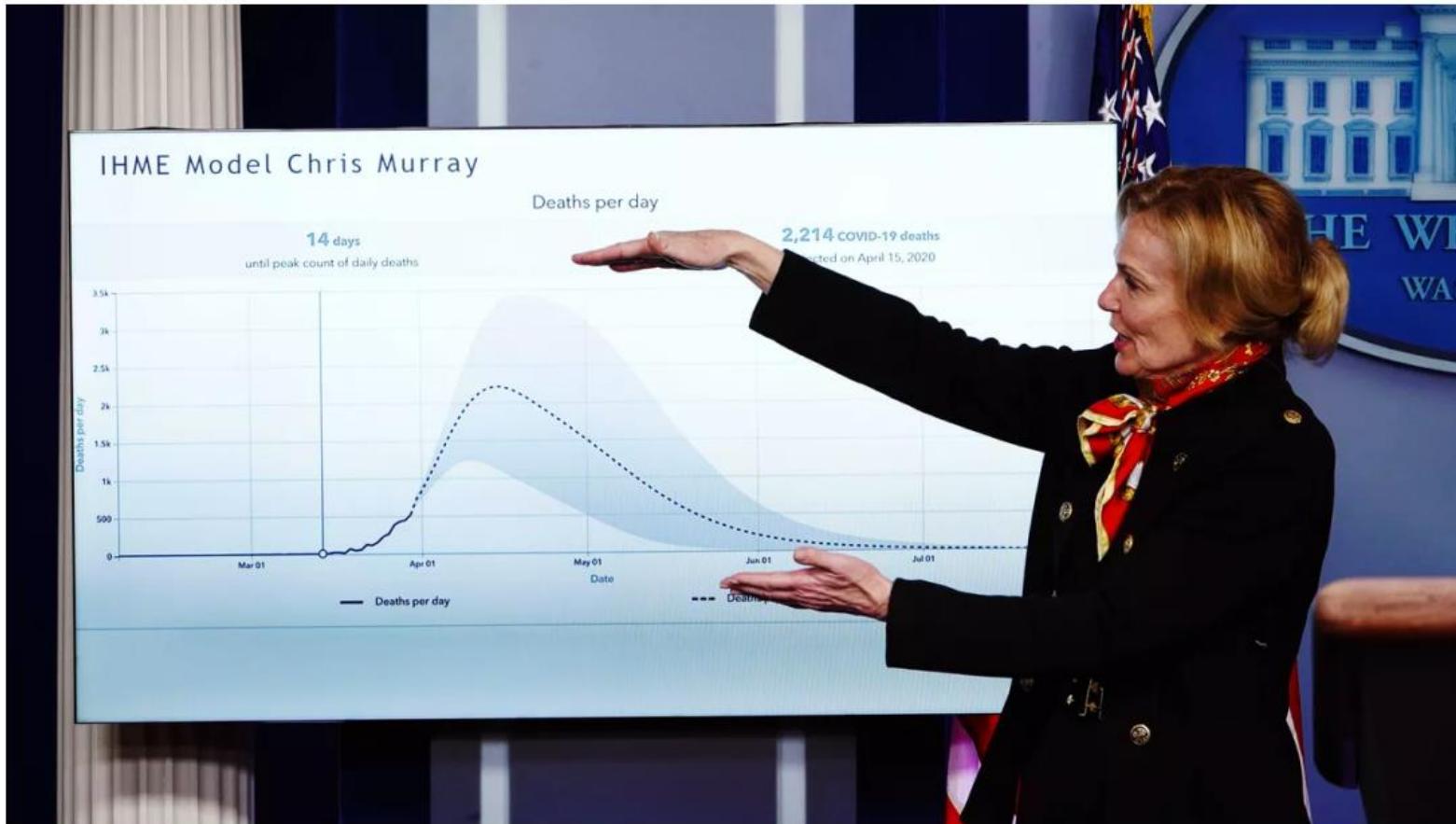
- Leverages multiple predictions from different sources
- Objectives:
 - Account for limitations of a single source of predictions
- Ideas:
 - Expert predictions
 - Prediction markets
 - **Ensembles**

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

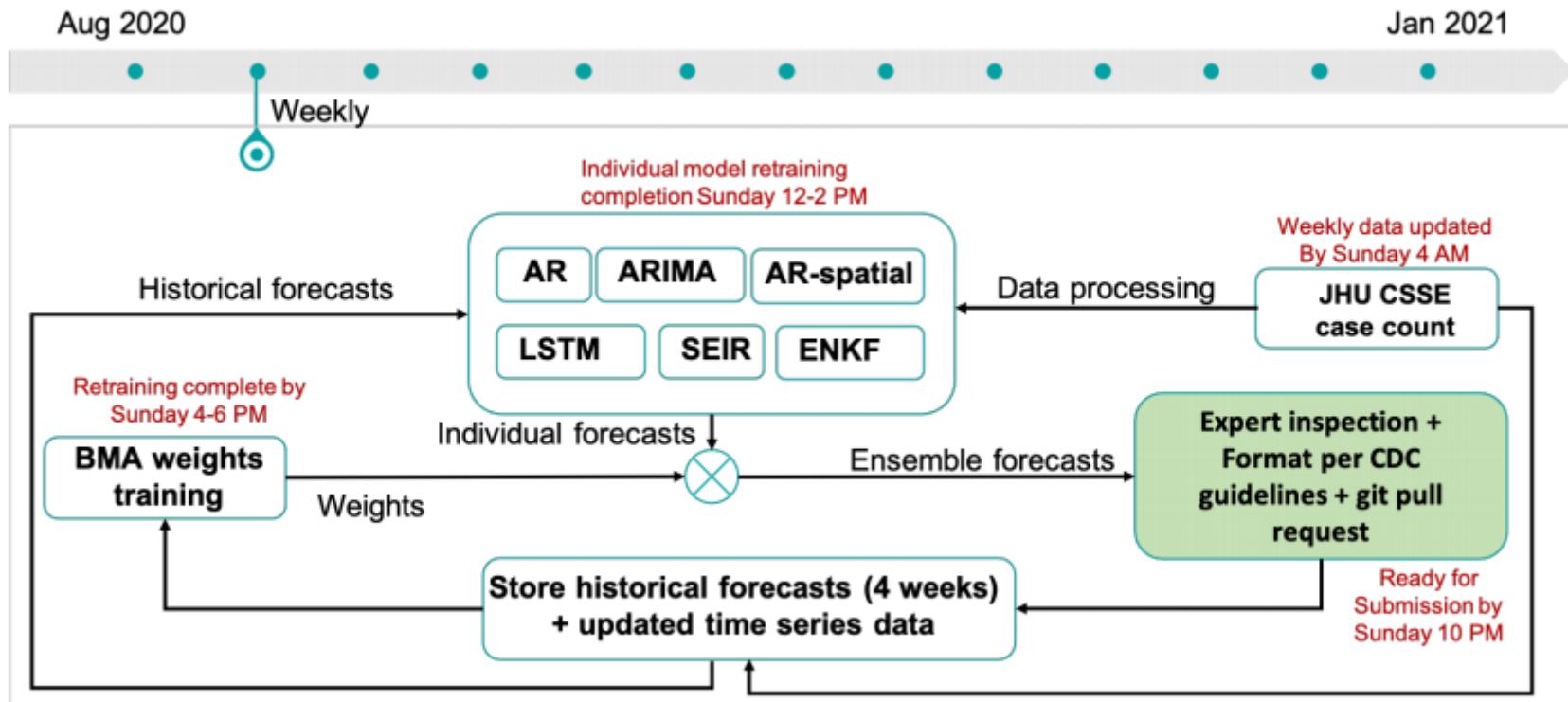


Slide credit: Nicholas Reich, UMass Amherst
covid19forecasthub.org/doc/talks/

Ex. 2: Bayesian ensemble

[Adiga+, KDD 2021]

- Diverse set of models + expert on the loop
 - All models trained by a single team





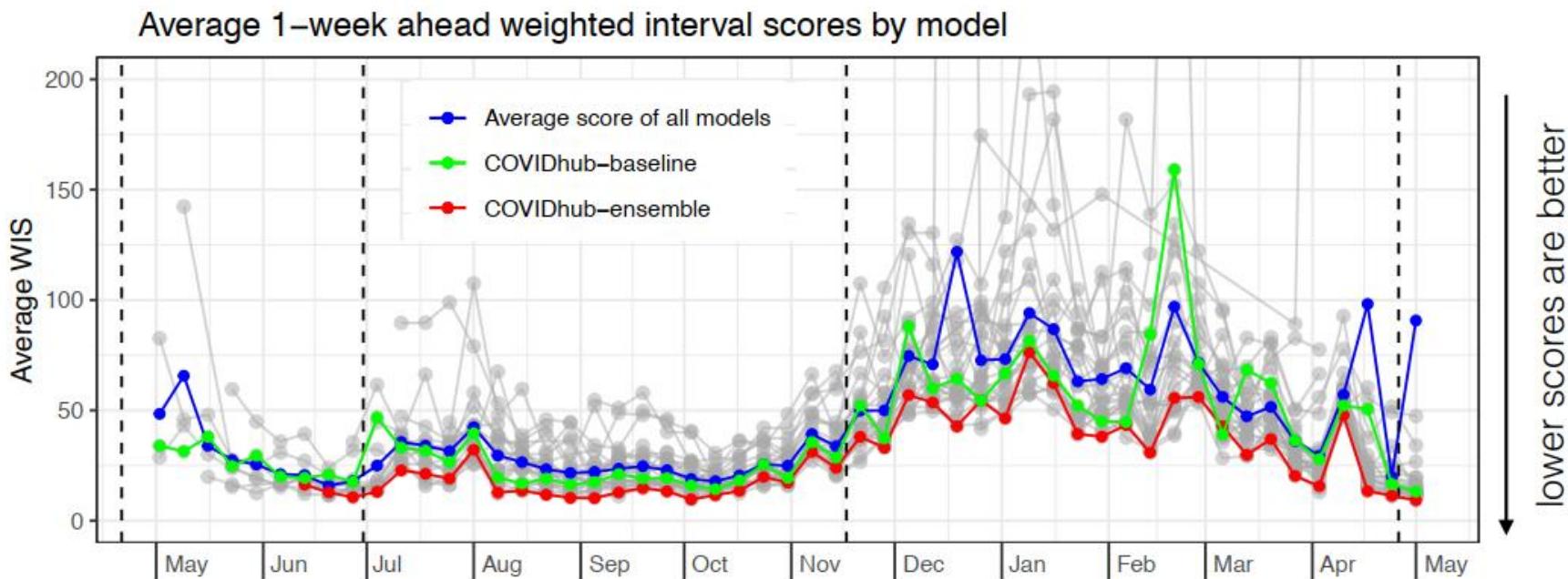
Diversity of COVID-19 models

- 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
covid19forecasthub.org/doc/talks/

Ex. 2: COVID Forecasting Hub ensemble

[Craemer+, PNAS 2022]



What is the optimal ensemble?

		"Trained" (i.e. component forecasts are weighted)	
		No	Yes
"Robust" (i.e. ensemble does not "blow up")	No	Equal-weighted mean	Variations on a weighted mean
	Yes	Median	Variations on a weighted median

- 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
covid19forecasthub.org/doc/talks/

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

Pros/cons hybrid models

- Pros:
 - Extend the capabilities of modeling paradigms
 - Seamlessly incorporation of multimodal data
- Cons:
 - Expert knowledge in mechanistic models and/or predictions can be very wrong
 - What-if forecasting from features may be misleading
 - Can't ensure parameters will change in the right direction

Outline

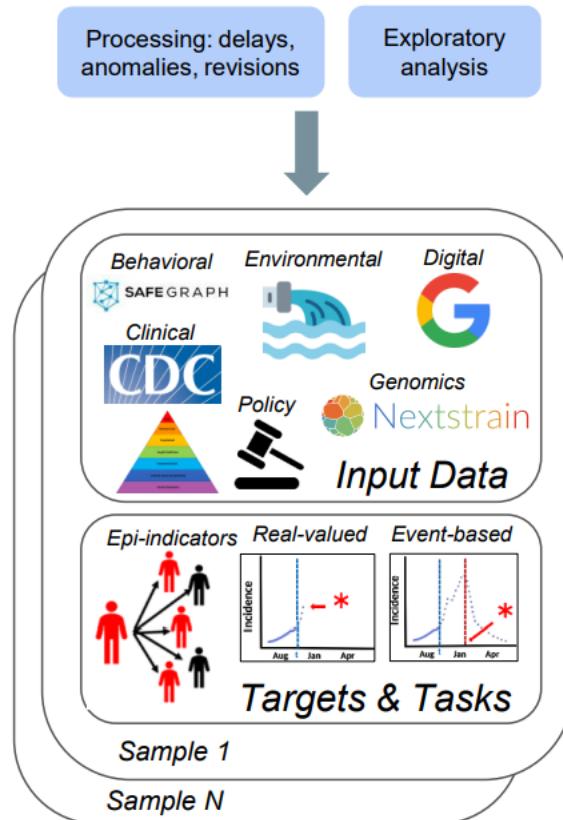
1. Epidemic forecasting: data and setup (40 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (60 min)
 - 30 min break at 3:15 PM, feel free to catch us for coffee
5. Hybrid models (45 min)
 - 5 min break
6. **Epidemic forecasting in practice** (25 min)
7. Open challenges and final remarks (20 min)

Part 6: Epidemic Forecasting in Practice

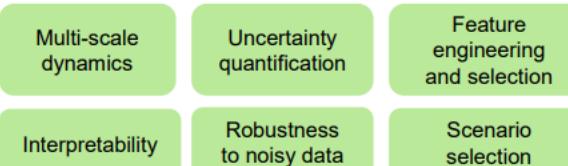
Epidemic Forecasting Pipeline

A. Data Processing

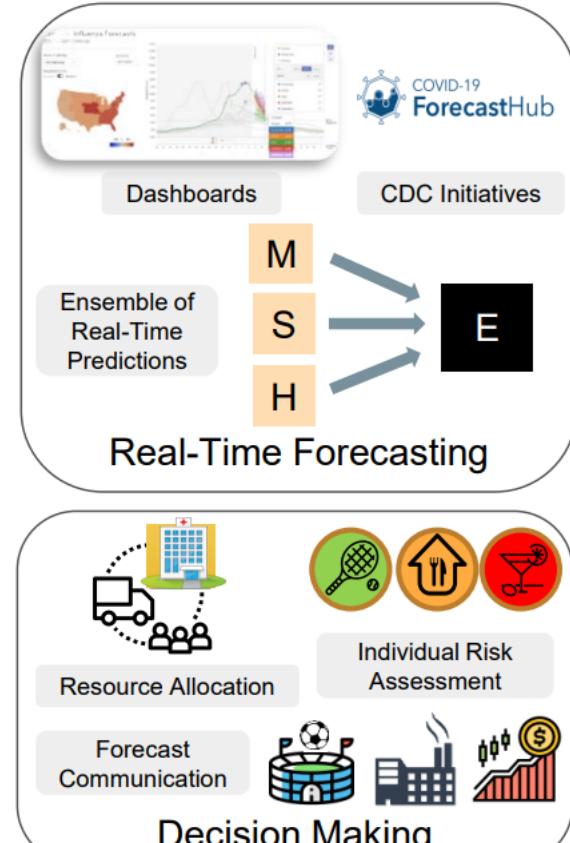
Raw data



B. Model Training & Validation



C. Utilization & Decision Making



Epidemic Forecasting in Practice (Outline)

- Real-time forecasting on the ground
- Forecasting and decision making
- Topics:
 1. Collaborative initiatives
 2. Experiences of individual forecasters
 3. Bridging forecasting with decision making
 4. Ethics and fairness

Epidemic Forecasting in Practice (Outline)

- Real-time forecasting on the ground
- Forecasting and decision making
- Topics:
 1. **Collaborative initiatives**
 2. Experiences of individual forecasters
 3. Bridging forecasting with decision making
 4. Ethics and fairness

[1] Collaborative 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
 - 2017-2019 Influenza hospitalizations – US National
 - 2017-2020 Influenza – US States
 - 2019-2020 Ae. aegypti & Ae. Albopictus mosquitoes (Arboviral) – US counties
 - 2019-2020 Department of Defense Influenza – US military facilities
 - 2020 West Nile neuroinvasive disease – US counties

Slide credit: Matt Biggerstaff, US CDC

FluSight Challenge

Reich Lab | [flusight](#)

? Help [About](#) [Tweet](#) [Source](#)

Real-time Influenza Forecasts

CDC FluSight Challenge

WEEK 8 (2017)

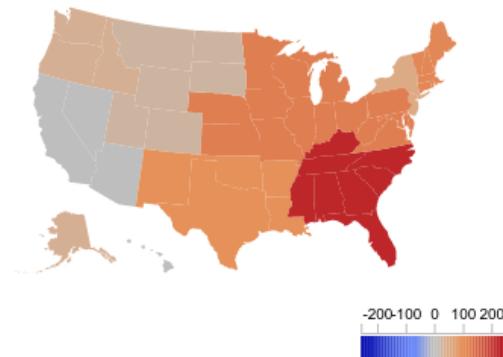
US National

Weighted ILI (%)

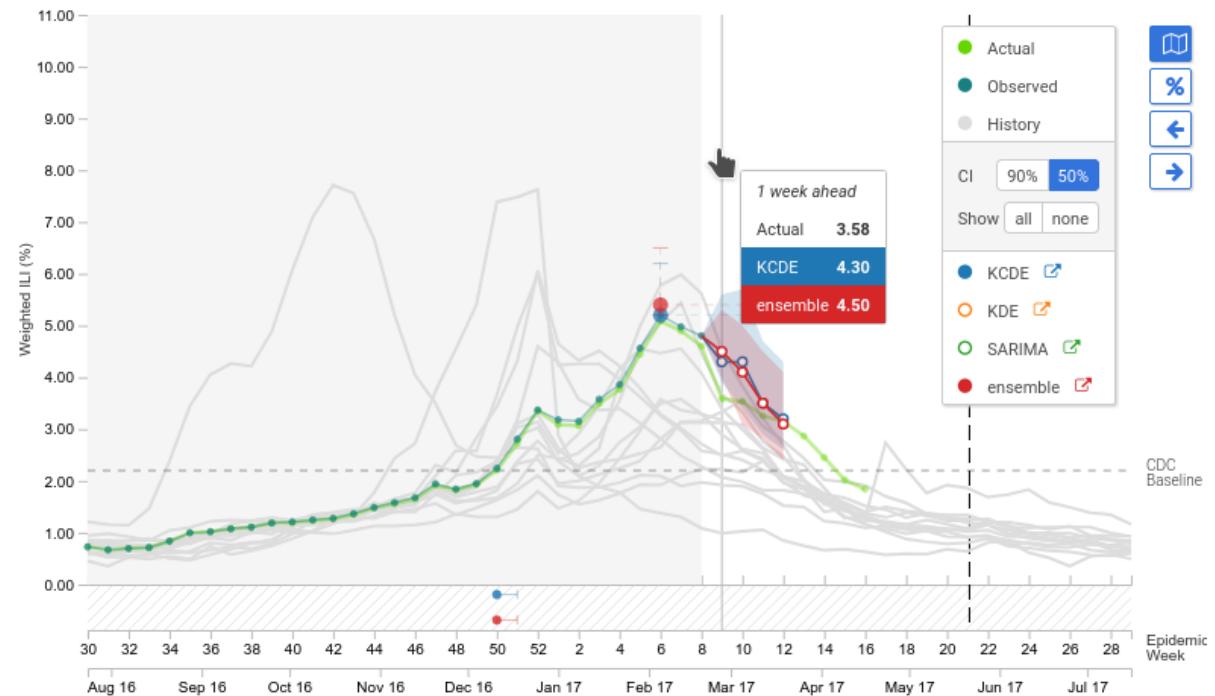
Absolute Relative

SEASON

2016-2017

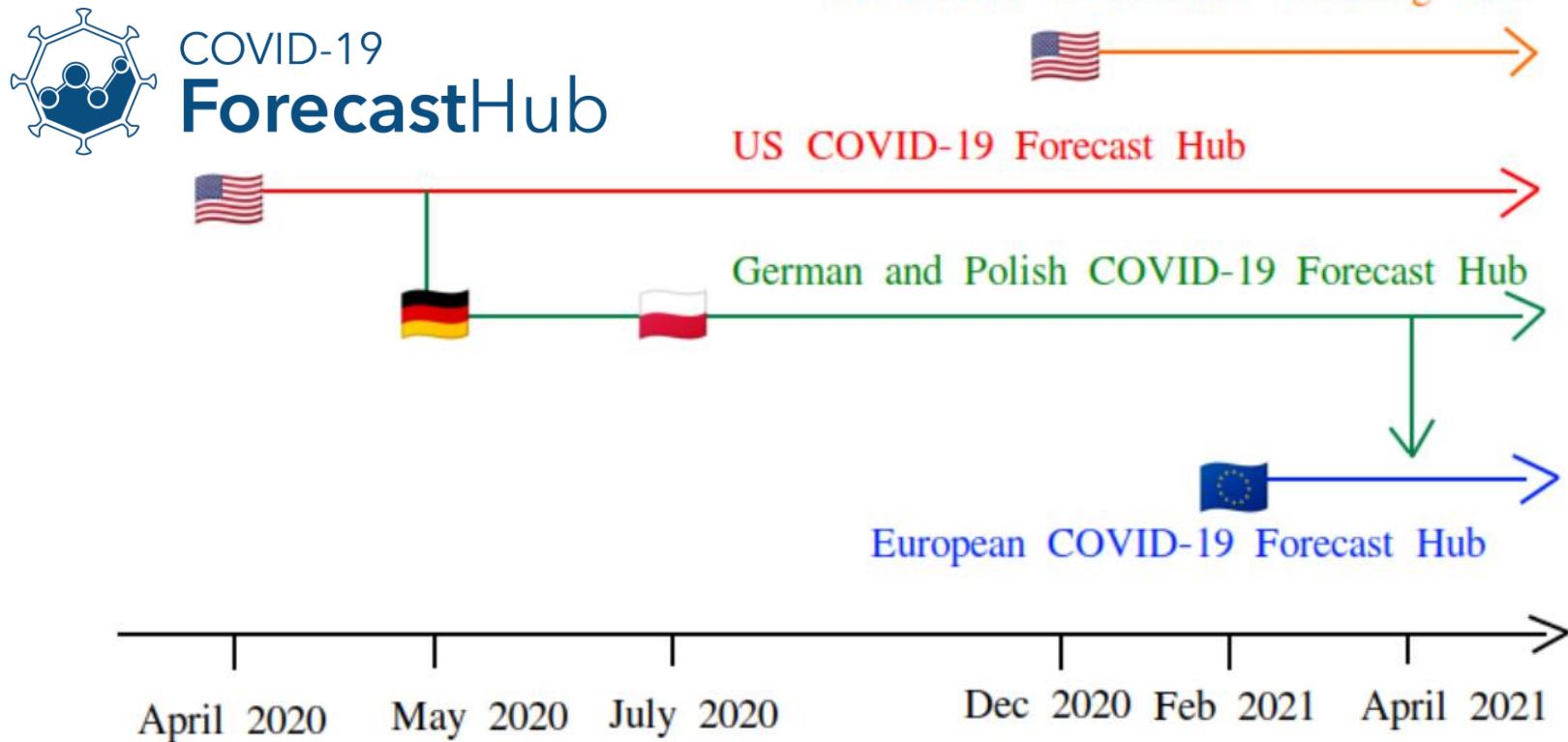


Time Chart Distribution Chart



Source: reichlab.io/flusight/

COVID-19 Forecast Hubs



Source: Johannes Bracher, KIT Karlsruhe and HITS Heidelberg

Standardization efforts of real-time forecast submissions

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 >

0 1 2 3 4 5

Project: COVID-19 Forecasts Config

Summary:	111 models, 5424 forecasts, 79,213,246 predictions
Owner:	covid19hub
Model Owners:	ydh28, vrushti-mody
Time Interval Type	Week



Epidemic Forecasting in Practice (Outline)

- Real-time forecasting on the ground
- Forecasting and decision making
- Topics:
 1. Collaborative initiatives
 2. **Experiences of individual forecasters**
 3. Bridging forecasting with decision making
 4. Ethics and fairness

[2] Real-time Experiences of Forecasters: FluSight Challenge

[Reich+, PNAS 2019]

- Study of 22 different models across 5+ years of submitted real-time forecasts
- Observations:
 - Top 5: First 4 are stat/ML models, 5th is mechanistic
 - Hardships in incorporating novel data sources
 - Effects of large data revisions

Ex. 1: Post-processing step: goal-oriented adjustments for competitions

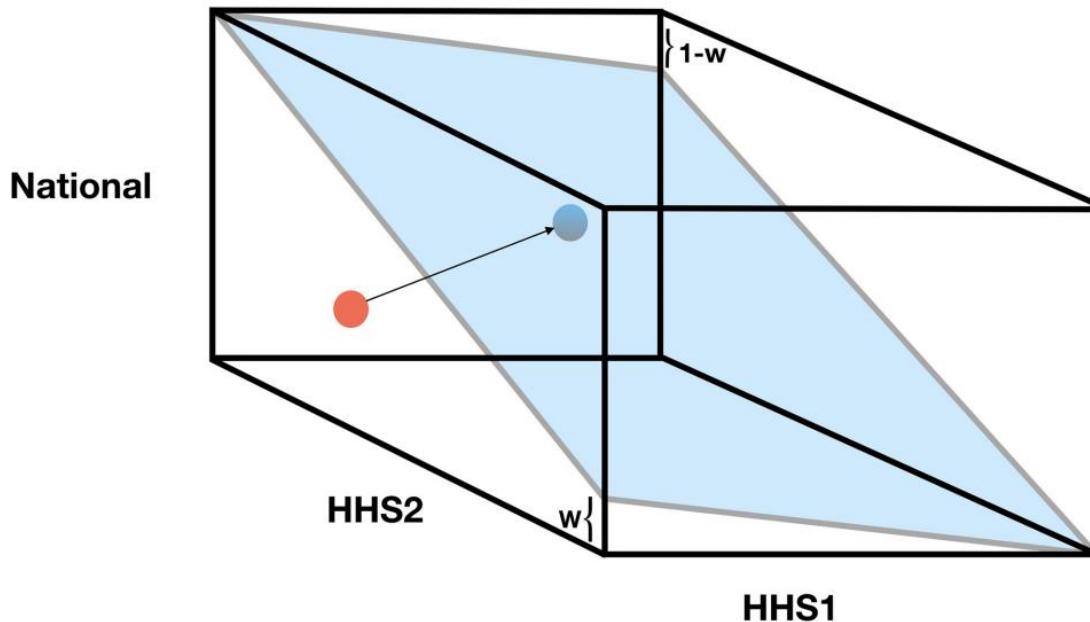
[Kandula+, Royal Society Interface 2018]

- Study on a diverse set of mech. & stat. models
- (1) Adjust for data revisions
 - Given time of year
- (2) Unrealistically wide distribution
 - Post-processing uncertainty quantification
 - Reduce probabilities for unlikely events
- (3) Avoid scoring penalty
 - Add small prob value to avoid log score harsh penalty

Ex. 2: Post-processing step: hierarchical coherence

[Gibson+, PLOS Comput. Bio 2021]

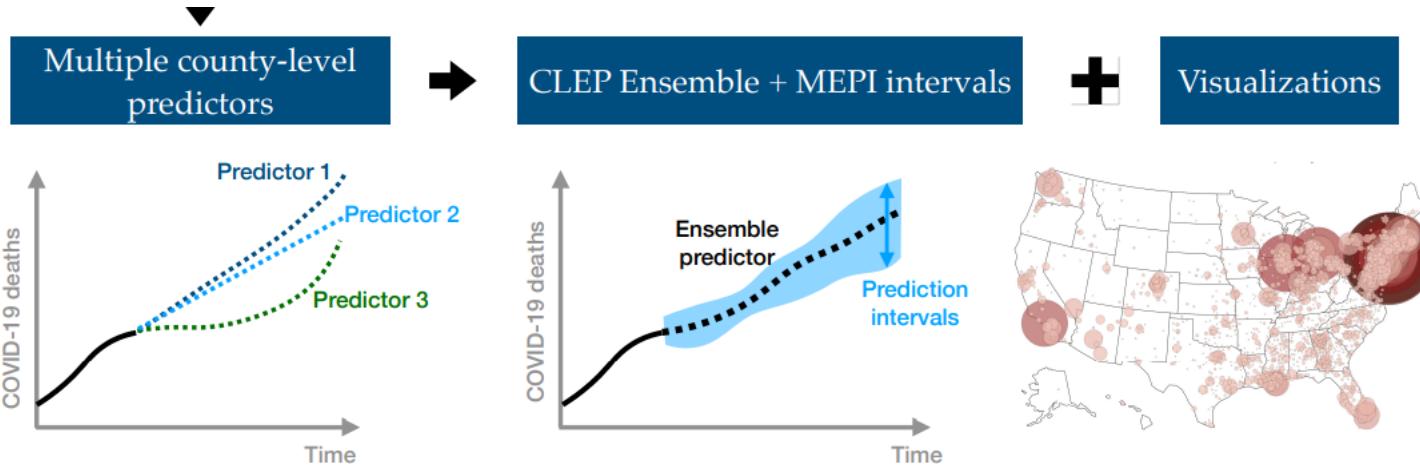
- Regional predictions should aggregate to National
- 79% increase in forecast skill



Ex. 3: Real-time Experiences of Forecasters: COVID-19

[Altieri+, Harvard Data Sci. Review 2021]

- Data quality issues:
 - Mismatch in reports from different sources
 - Inconsistencies data definitions across geographies
- CLEP ensemble:
 - Weighted comb. of linear and exponential predictors



Ex. 4: Bayesian Mechanistic Model

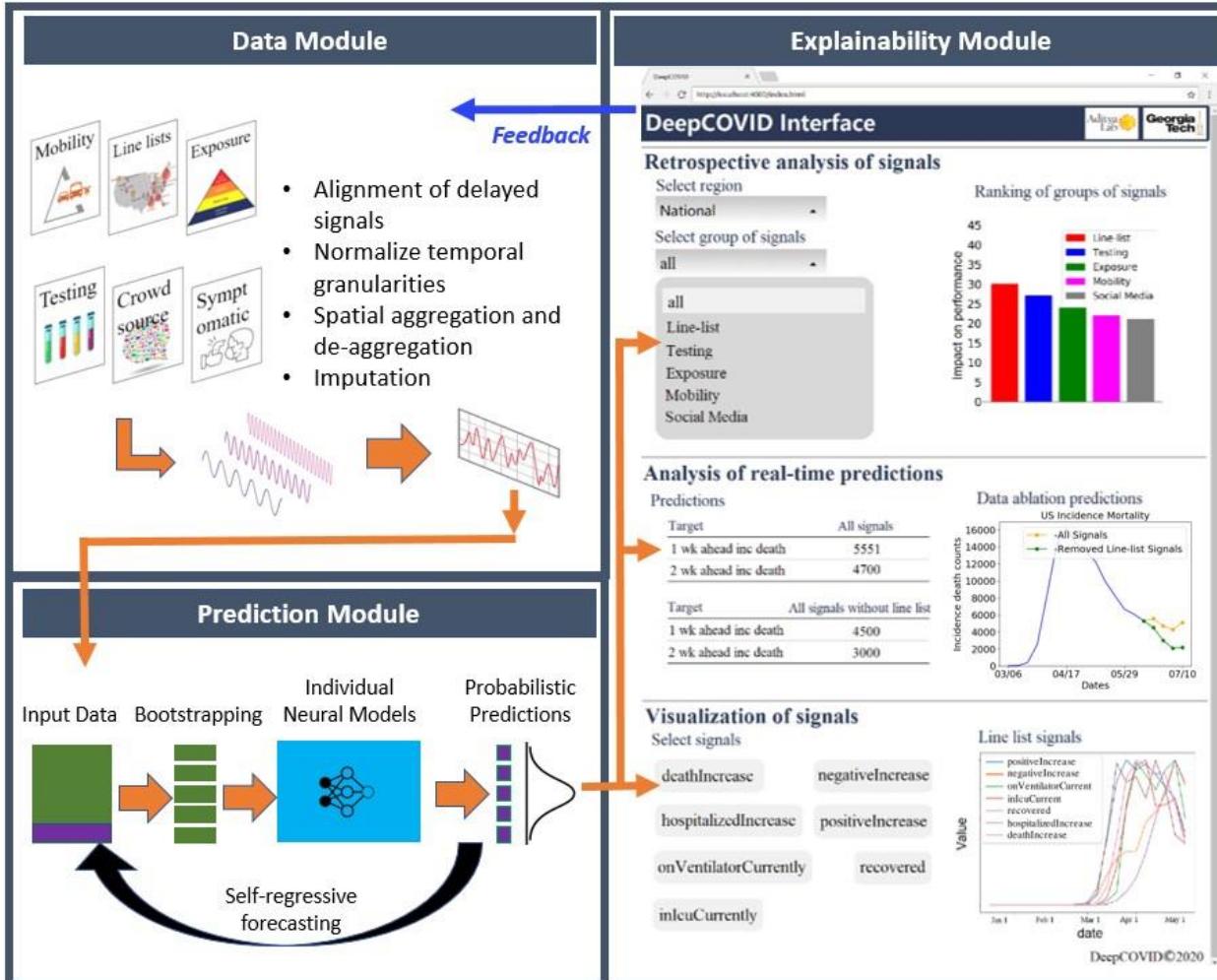
[Gibson+, medRxiv 2020]

- MechBayes:
 - SEIRD + Bayesian framework w/ informative priors
- Use a quality assurance procedure
- Expert on the loop:
 - Visualize most recent data
- Anomalies:
 - Check notifications from data sources (JHU)
 - Backdistribute when needed

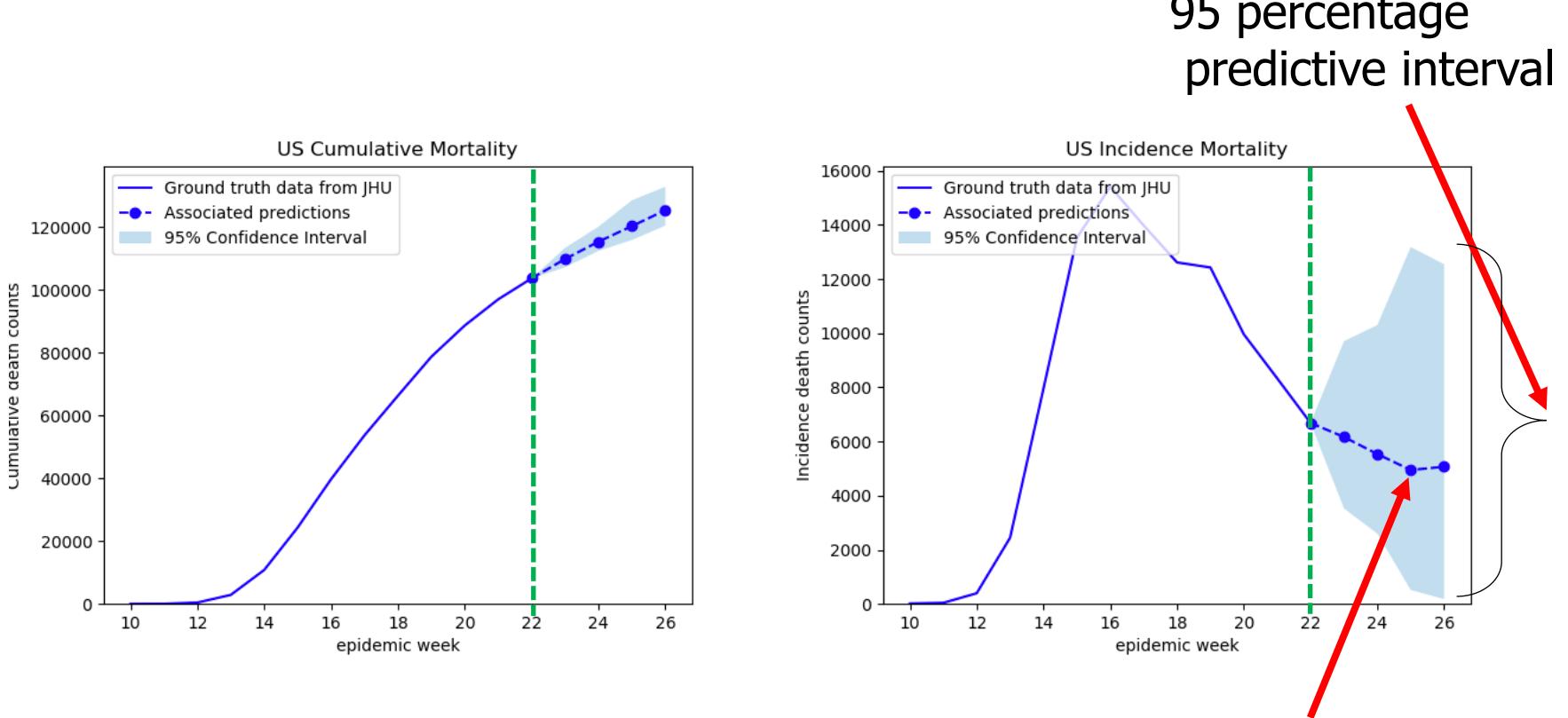
Ex. 5: Operational DL

Framework: DeepCOVID

[Rodríguez+, AAAI 2021]



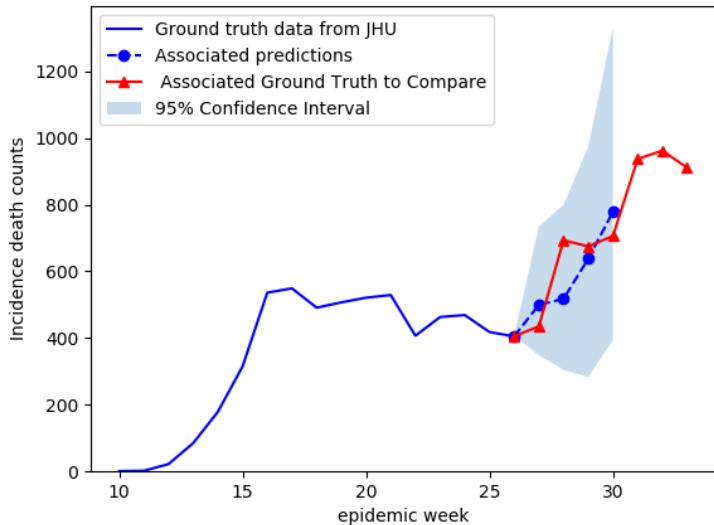
Example Forecast Visualization



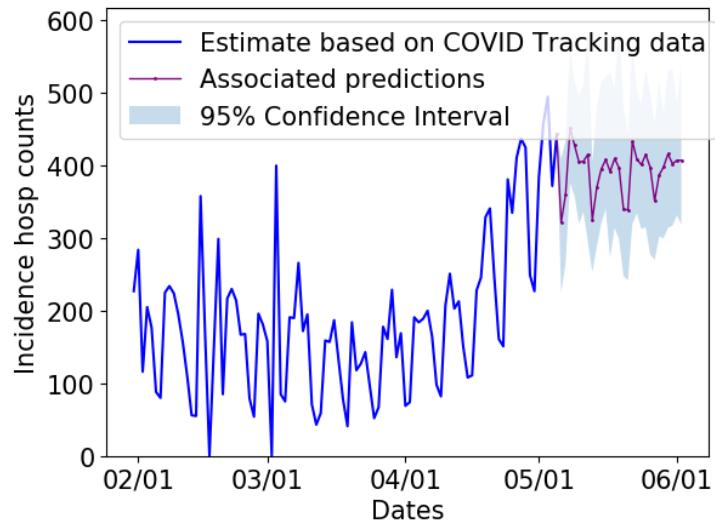
**Predictions on EW 22
(= May 25, 2020)**

Highlights of results

Anticipate Trend Changes

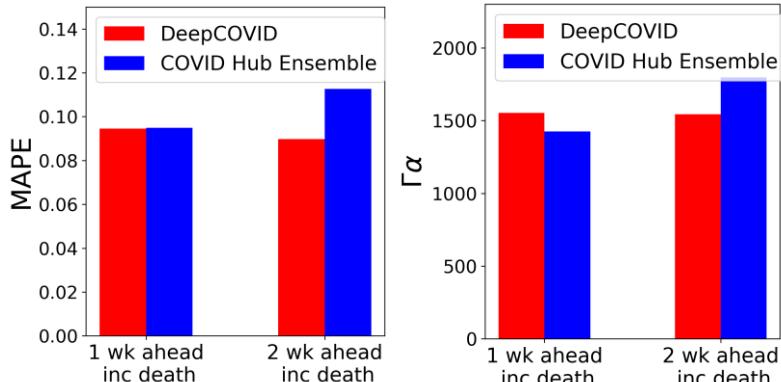


Capture finer-grain patterns

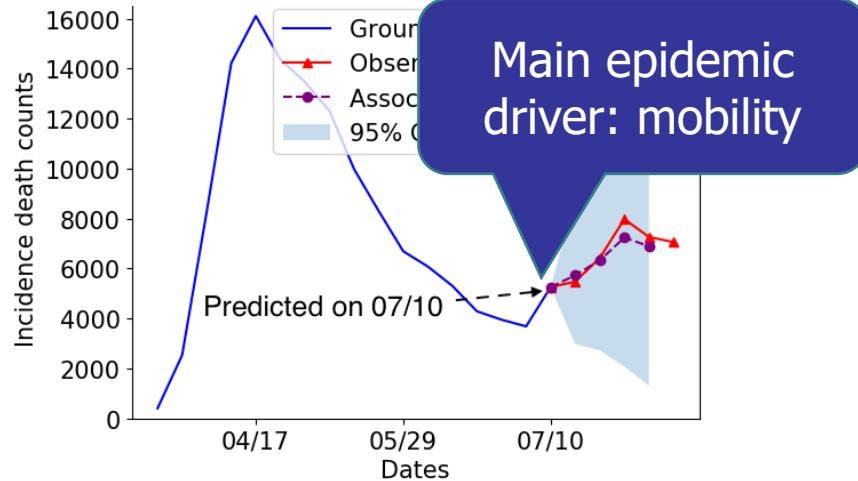


Lower is better
better

Excels in short-term forecasting



Provides explanations



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

Epidemic Forecasting in Practice (Outline)

- Real-time forecasting on the ground
- Forecasting and decision making
- Topics:
 1. Collaborative initiatives
 2. Experiences of individual forecasters
 - 3. Bridging forecasting with decision making**
 4. Ethics and fairness

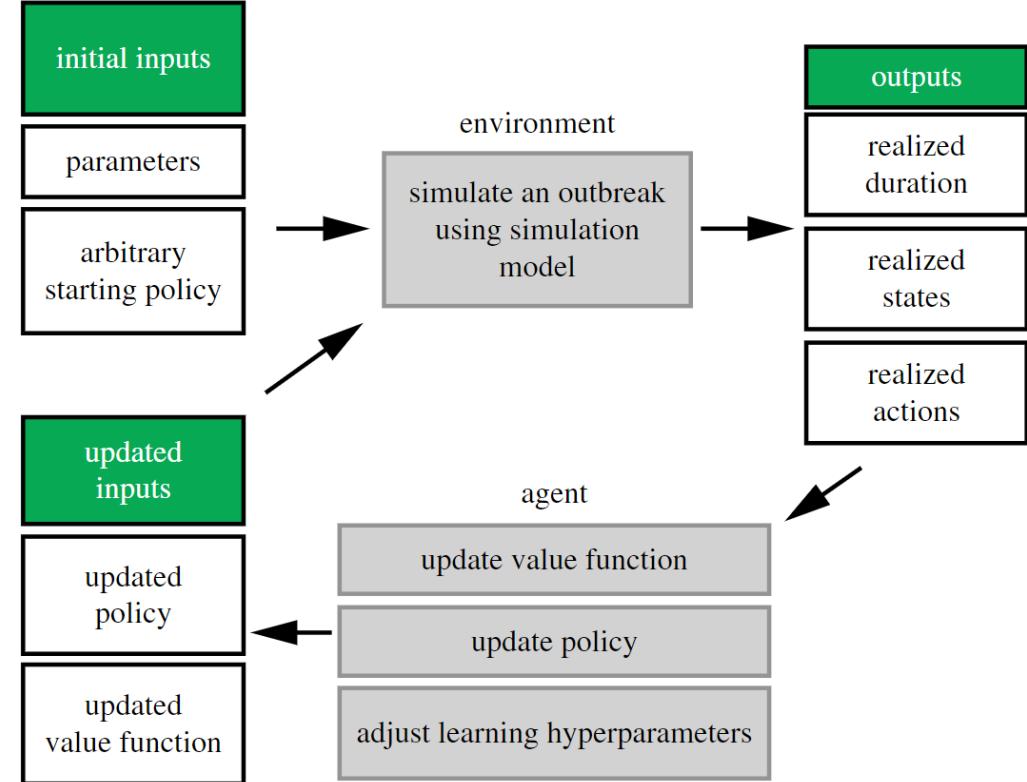
[3] Bridging forecasting with decision making

- Leverage predictions to inform decision making for policymakers, public health workers, supply chains, etc.
- Types:
 - Strategic: Large-scale policies (lockdowns, mandates...)
 - Tactical: Small-scale, high density action space, to accomplish a narrow goal (logistics, distribution of vaccines...)

Ex 1: Strategic Interventions for mitigating foot and mouth disease

[Probert+ PloS 2018, RS 2019]

- Simulations based on past outbreak data.
- Can be solved as Sequential Decision making problem (leverage Reinforcement Learning)



Ex 1: Strategic Interventions for mitigating foot and mouth disease

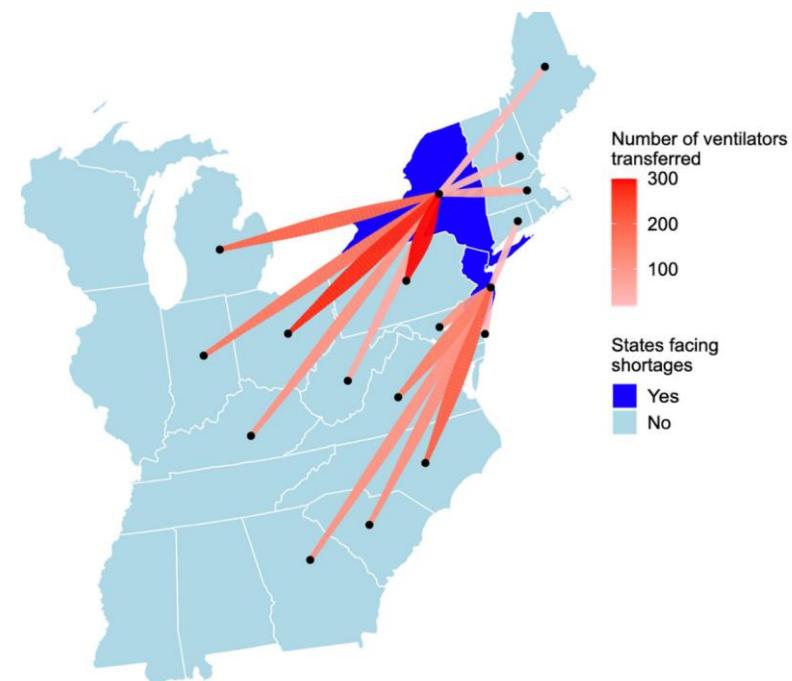
[Probert+ PloS 2018, RS 2019]

- Control measures (possible interventions):
 - Vaccinate animals: Costly but preserves cattle
 - Cull farm animals: Cheap to stop spread but loss of cattle (long-term costly)
- Set rewards: no. of cattle saved and cost on vaccination
- Solved using a Deep RL algo. (DQN) [Minh+ Nature 2013]

Ex 2: Tactical Interventions for ventilator allocation

[Bertsimas+, HCMS 2021]

- Leverage future case forecasts to model optimal resource-allocation
- Tradeoff:
 - Satisfy future demand for ventilators
 - Reduce inter-state transport cost



Epidemic Forecasting in Practice (Outline)

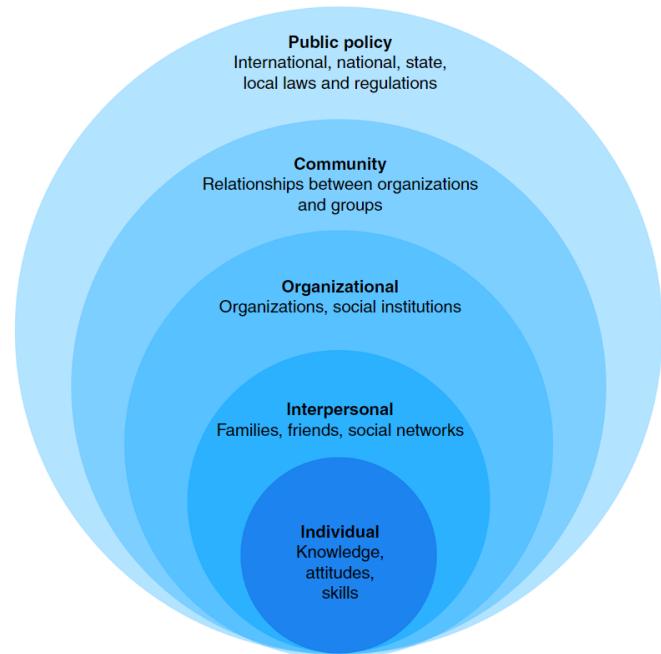
- Real-time forecasting on the ground
- Forecasting and decision making
- Topics:
 1. Collaborative initiatives
 2. Experiences of individual forecasters
 3. Bridging forecasting with decision making
 - 4. Ethics and fairness**

ML and Algorithmic Fairness in Public Health

- Main components:

- Data in public health
 - Data from surveys and government reports
 - Person-generated data
- Algorithms
 - Identification of factors for health outcomes
 - Intervention design
 - Prediction of outcomes
 - Allocation of resources

[Mhasawade+ Nat. Mach. Intell. 2021]

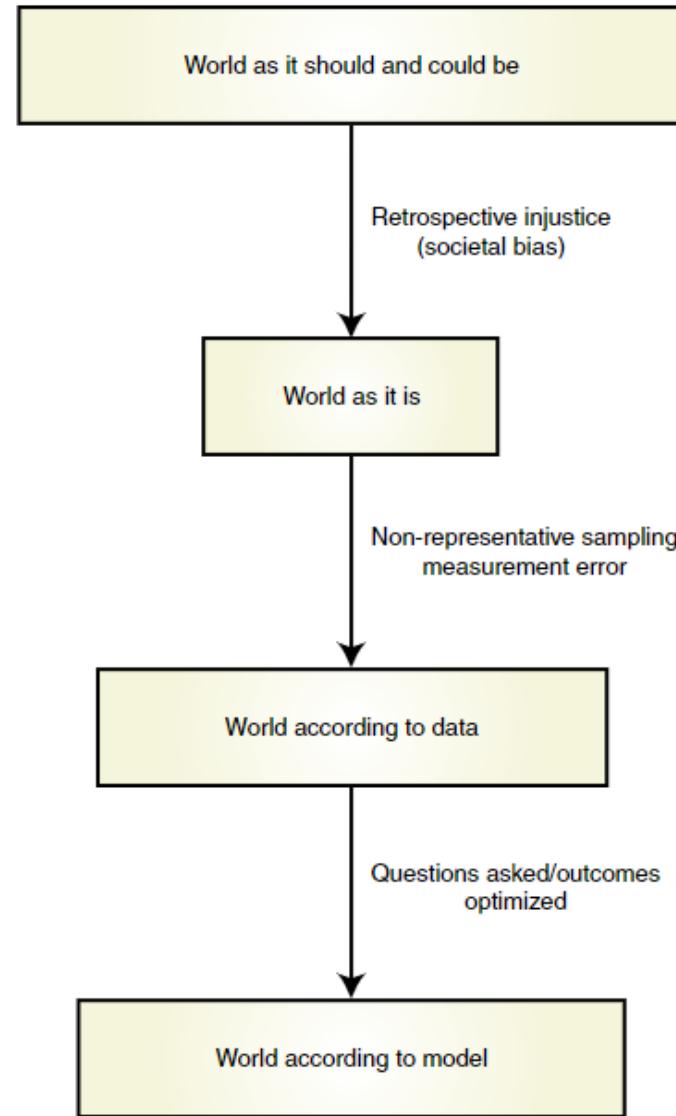


Socio-ecological model of health:

- Multi-level individual and environmental factors that determine health
- Macro-level properties to understanding inequities

Sources of bias and algorithmic fairness

- Disparities can manifest in generating and using data
- Applying ML can impact disparities and equity
- Algorithmic fairness comes into play at the end of this pipeline
 - via questions asked/outcomes optimized



Data in public health (1)

- **Data from surveys and government reports**
 - Aggregated individual-level information. Examples:
 - National Health and Nutrition Examination Survey (NHANES)
 - Demographic and Health Surveys programme (DHS)
 - Different indicators of health, designed via specific constructs
 - Housing quality can be measured via rental status, sanitation status, crowding, indoor air quality, ...
 - Disparity issues:
 - Constructs are misleading
 - Privacy concerns

Data in public health (2)

- **Person-generated data**

- Data from digital sources which are high-resolution and geo-linked in near-real time. Examples:
 - Twitter and Instagram
 - Internet surveys
- Provide opportunities to better measure the social determinants of health in a targeted way, by person, location and/or time
- Disparity issues:
 - Opt-in nature
 - Cost associated with access

Examples of ML algorithms applications

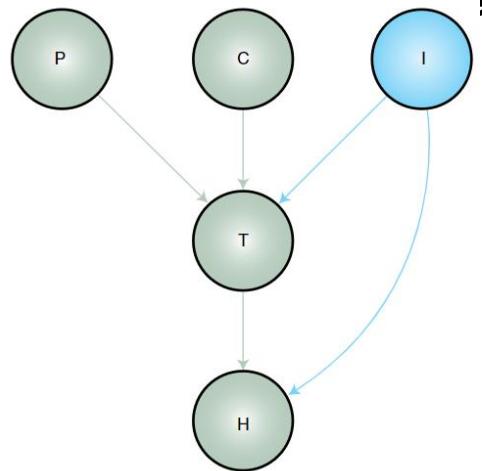
- **Identification of factors for health outcomes**
 - Establishing the multivariate empirical relation between the probability of disease outbreak and environmental condition
- **Intervention design**
 - Targeting the individual towards depression management, self-efficacy for weight loss, and smoking cessation.
- **Prediction of outcomes**
 - Mortality risk, hospital readmission and disease prognoses from imaging or other clinical data.
 - Population-level predictions
- **Allocation of resources**
 - Deciding what laboratory measurements should be taken, when and on whom, trading off the value of information against the cost of acquisition

Key challenges

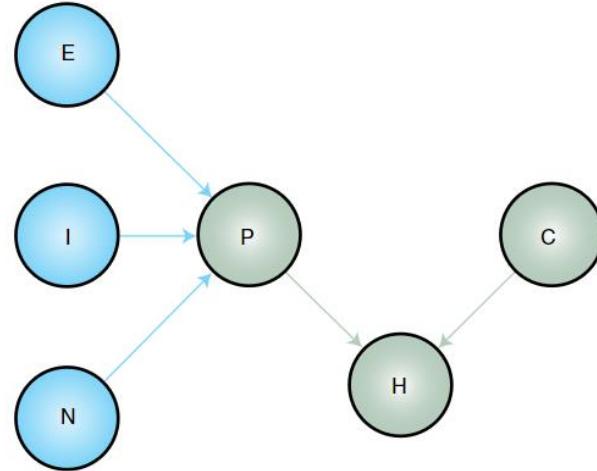
- **Privacy and health data**
 - People may wish to avoid being stigmatized via certain assessments
 - Lack of privacy-centered standards and regulations
- **Assessing external validity**
 - Applying the conclusions of a scientific study outside the context of that study
 - opt-in nature of data collection leads to selection bias
- **Measuring and integrating social determinants of health**
 - There is a need for better measures of upstream factors that are important social determinants.
 - This spans environmental, policy and other social factors such as racism

Challenges in algorithmic fairness

Owning variables vs. Unaccounted variables



Owning variables vs. Unaccounted variables



- If high-risk populations such as the **uninsured** are not included (shaded blue), then inferred relationships and treatment effects would **not be relevant to those most vulnerable**
- Fairness is ensured w.r.t. protected attribute P but this is composed of several factors (shaded blue) that are not accounted, for which the model is unfair.
 - E: education; I: income levels

Challenges in fairness of epidemic forecasting models

[Tsai+ Nat. Dig. Med. 2022]

- **Recognizing biases when selecting outputs**
 - Case data are highly dependent on the number of tests performed in a community
 - Failure to account for this skewed data could then lead to a cycle in which models underestimate the healthcare needs of a low-resourced community
- **Including social determinants**
 - Many demographic and socioeconomic variables are highly interrelated and may be correlated with confounding effects.
 - Recommendation:
 - Include in predictive models only if their effects on outputs are well-characterized by the scientific community and can be decoupled from the rest of data-driven learning
 - Training models with and without each potentially problematic input variable and observing the effect on performance (and understand the significance of their effects via F-tests)

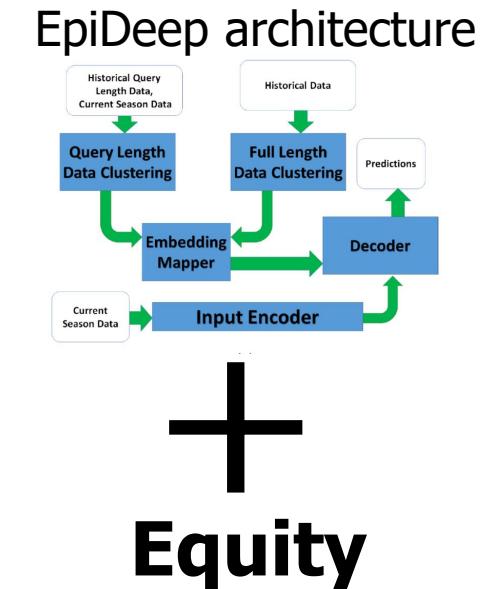
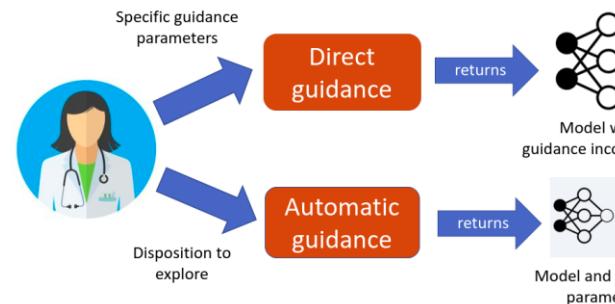
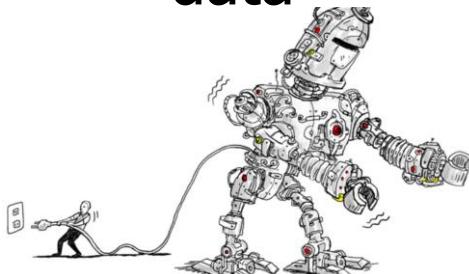
Challenges in fairness of epidemic forecasting models (2)

- **Evaluating models through an equality lens**
 - Not only accuracy
 - Add equality analyses
 - Forecasting models undergo subgroup analysis to verify that results are comparably accurate for each group and that there are no systematic biases
- **Choosing the appropriate geographic unit of analysis**
 - Your zip code is more important than your genetic code: access to healthcare resources remains largely tied to geography
 - Building models at overly-aggregated geographies may conceal the critical heterogeneity that can guide policy action and reveal inequities

Regional Equity for Neural Forecasting Models

[Rodriguez+ epiDAMIK @ KDD 2020]

- Leverage Seldonian optimization framework [Thomas+, 2019]
 - Proposed for AI safety
 - Precludes undesirable behavior of AI model by enforcing behavioral constraints in optimization
 - Has a safety test in unobserved data



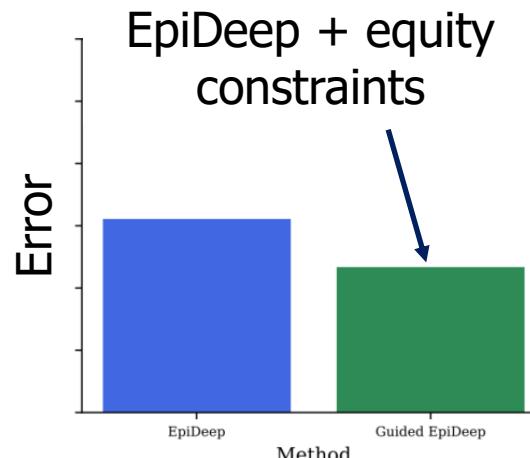
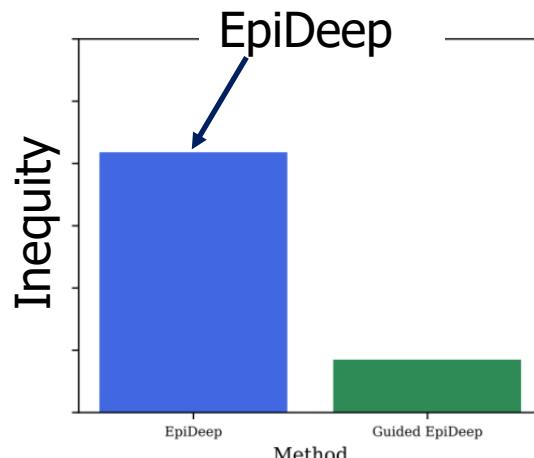
Constraining Regional Equity

- Regional Equity: Quality of forecast μ between any two regions ~~should be similar~~

$$|\mathbb{E}[\mu(\theta, t+1, R_1)] - \mu(\theta, t+1, R_2)| - \epsilon \leq 0$$

Squared error in Region 1

Squared error in Region 2



Goal: Reduce inequity and maintain (or even improve) forecasting error

Outline

1. Epidemic forecasting: data and setup (40 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (60 min)
 - 30 min break at 3:15 PM, feel free to catch us for coffee
5. Hybrid models (45 min)
 - 5 min break
6. Epidemic forecasting in practice (25 min)
7. **Open challenges and final remarks** (20 min)

Part 7: Open Challenges and Opportunities

[C1] Data-related challenges (Challenges)

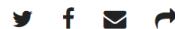
- Dealing with collection/reporting errors
- Adapt to revisions and anomalies
- Data privacy, anonymity and security
 - Many datasets can contain sensitive personal data (EHR, Mobility)



MultiCare announces breach that could impact over 18,000 patients' health data and records

BY ALLEN SIEGLER

AUGUST 11, 2022 5:00 AM



DIVE BRIEF

Health data breaches slowing from 2021's record high, report suggests

Published July 19, 2022

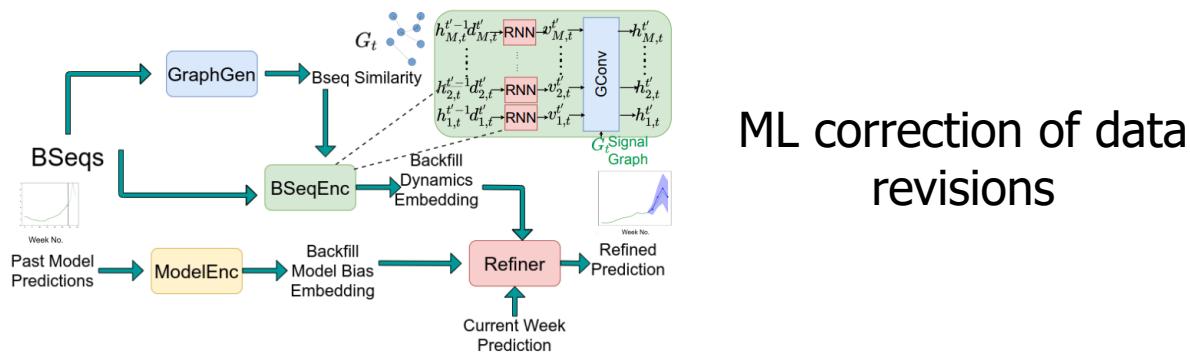


Hailey Mensik



[C1] Data quality (Opportunities)

- **Statistical data correction:**
 - Correct revision and reporting errors to improve data quality

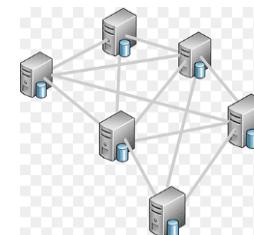


ML correction of data revisions

- Examples:
 - Modelling data revision dynamics [[Kamarthi+ ICLR 2022](#)]
 - Data inequity and bias [[Rodriguez+ epiDAMIK@KDD 2020](#)]

[C1] Data collection and privacy (Opportunities)

- **Privacy-preserving techniques:**
 - Abide privacy, security laws (Eg: GDPR, HIPAA)
 - Using Differential Privacy + Federated learning to learn from deanonymized sensitive data [Zhang+ KBS 2021]
- **Building accessible data infrastructures:**
 - Researchers safely submit and access data at scale
 - Access multiple data versions when subject to revisions



[C2] Moving beyond short-term forecasting (Challenges)

- Unrealistic longer-term and what-if predictions of ML models
- Mechanistic models do this but face difficulties including data
 - What does data tell us about long-term patterns?
 - How can past data in interventions can inform new interventions?

[C2] Moving beyond short-term forecasting (Opportunities)

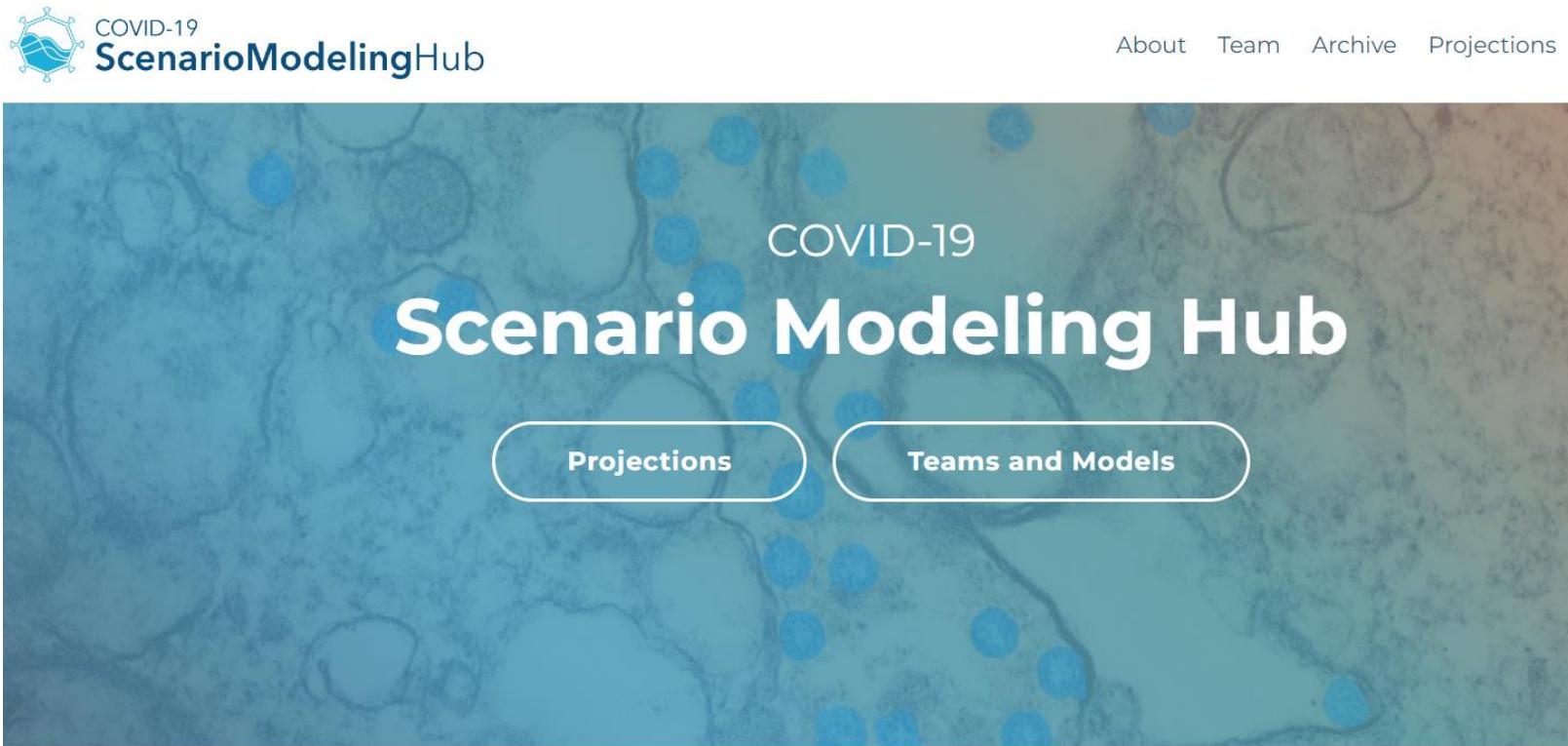
- **Scientific AI:** End-to-end integration of mechanistic and ML models
 - Neural networks interacting w/ epidemiological models
 - AI for scientific discovery
- Examples:
 - DEFSI [Arik+ NeurIPS 2020]
 - EINNs [Rodriguez+ arXiv 2022]
 - Differentiable ABMs [Chopra and Rodriguez+ AI4ABM @ ICML 2022]

[C2] Moving beyond short-term forecasting (Opportunities)

- **Causal ML and reinforcement learning (RL)**
 - Discovering causal relations among multivariate data and interventions
 - RL for policy analysis
- Examples:
 - Causal feature selection in time series [Mastakouri and Schölkopf, NeurIPS 2021]
 - COVID-19 testing analysis via RL [Bastani+ Nature 2021]

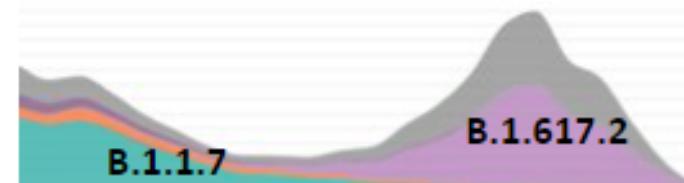
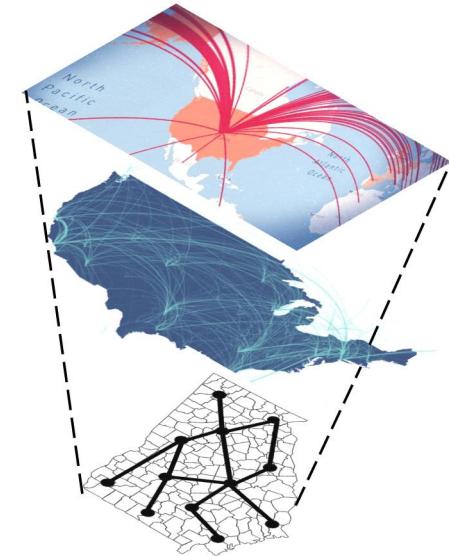
[C2] Moving beyond short-term forecasting (Opportunities)

- **Testbed** for data-centered models for scenario analysis



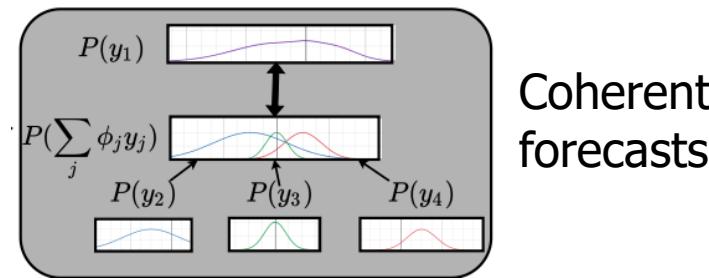
[C3] Modeling multi-scale dynamics (Challenges)

- Temporal and spatial multi-scales
 - Coherent probabilistic forecasts
 - city vs county vs state vs HHS
 - Robustness to noise and missing data
- Incorporate pathogen and behavioral multi-scale dynamics



[C3] Modeling multi-scale dynamics (Opportunities)

- **Hierarchical modeling**
 - Probabilistic coherency across hierarchies



- Examples:
 - Spatially coherent probabilistic forecasts [\[Kamarthi+ arXiv 2022\]](#)
 - Post-processing spatial consistency [\[Gibson+, PLOS Comput. Bio 2021\]](#)

[C3] Modeling multi-scale dynamics (Opportunities)

- **Multi-scale modeling**

- Pathogen dynamics
- Behavioral models



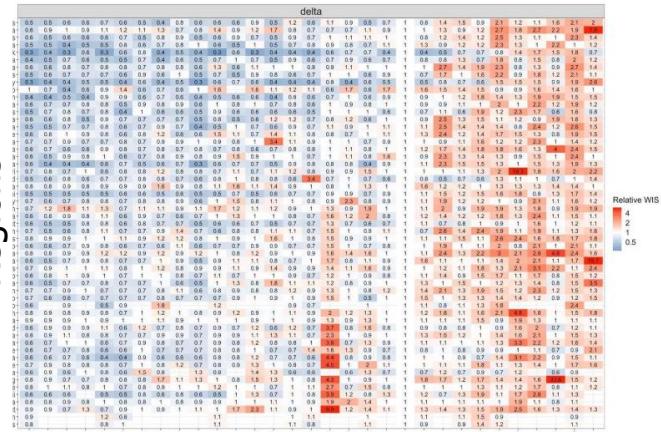
- Examples:

- Evolution of pathogens (phylodynamics) [\[Kraemer+
Science 2021\]](#)

[C4] Improving the ensembles and WoC predictions (Challenges)

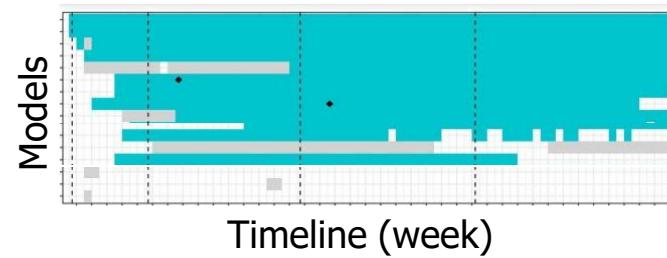
- Models change:
 - Performance (across regions and time)
 - Methodology (often not reported)
- WoC:
 - Information inefficiencies (e.g., misinformation)
 - Confidence varies across source

Performance on Delta wave



Models

Valid submissions



Timeline (week)

[C4] Improving the ensembles and WoC predictions (Opportunities)

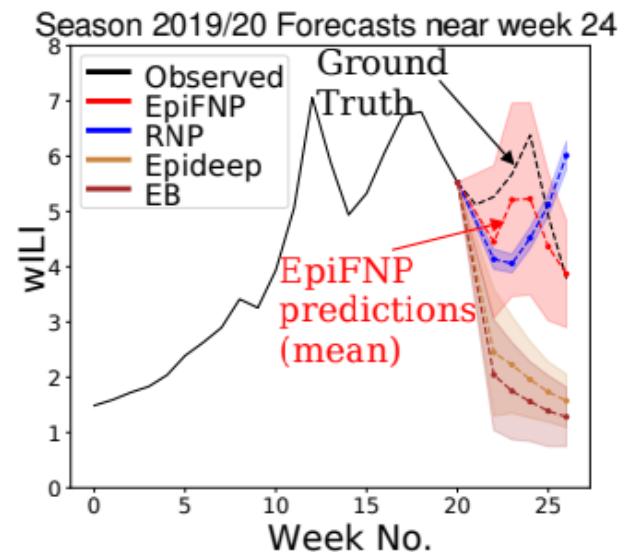
- **Novel weighting schemes:**
 - Spatio-temporal model weighting/selection
 - Multiple pooling or aggregation mechanisms
- Examples:
 - Mixture of experts [Riquelme+, NeurIPS 2021]
 - Optimal ensemble weighting [Shahhosseini+, ML with Applications 2022]

[C4] Improving the ensembles and WoC predictions (Opportunities)

- **HCI for Data Gathering**
 - Capture confidence of prediction, uncertainty bounds
 - Data on multimodal and conditional distributions of beliefs
- Examples:
 - Efficient elicitation for collective crowd answers [[Joon + CSCW 2019](#)]

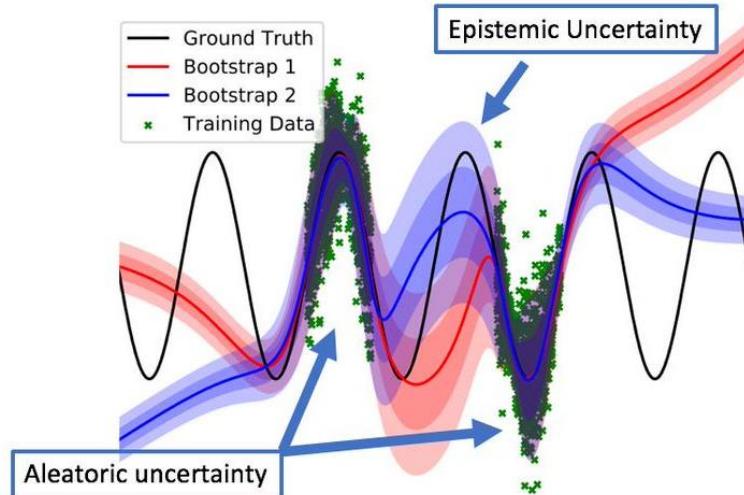
[C5] Well-calibrated Forecasts (Challenges)

- Uncertainty quantification of forecasts
 - Help at and high-stake decision making
- Helps in explainability during novel/unseen scenarios



[C5] Well-calibrated Forecasts (Opportunities)

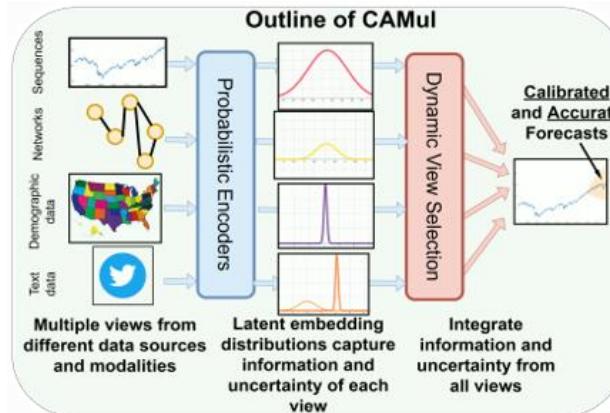
- **Explain sources of uncertainty:**
 - Differentiate aleatoric (data) and epistemic (model) uncertainty



- Examples
 - Neural SDEs [Kong+ ICML 2020]
 - Probabilistic ensemble [Chua+ NeurIPS 2018]

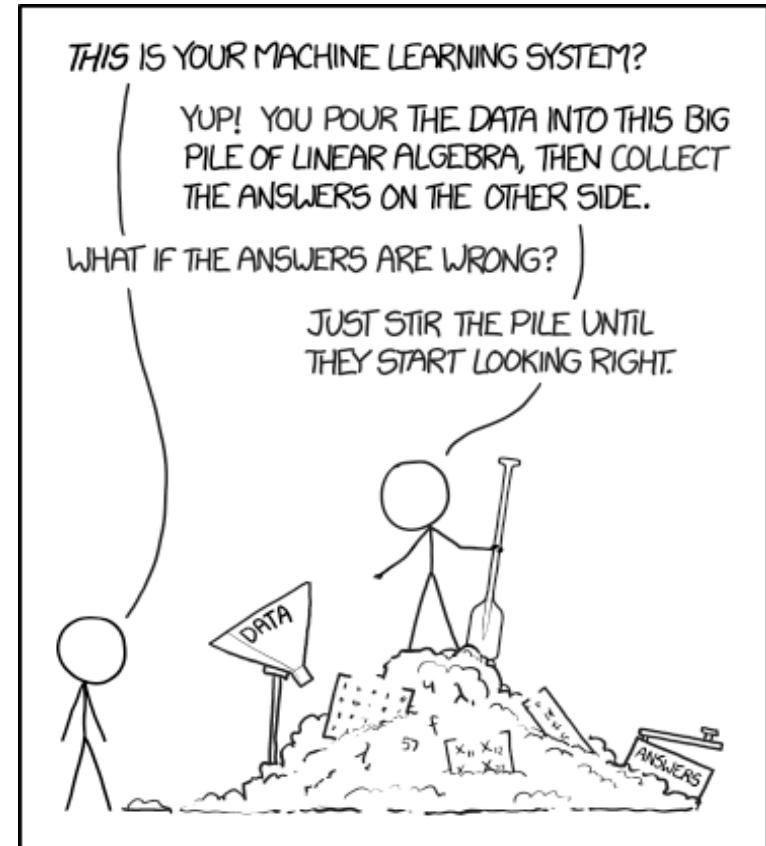
[C5] Well-calibrated Forecasts (Opportunities)

- **Modeling multiple sources of uncertainty**
 - From multiple modalities
 - Useful also in anomaly detection, eval. of data quality
- Examples
 - Neural Gaussian Process for multiple sources of uncertainty
[\[Kamarthi+ WWW 2022\]](#)



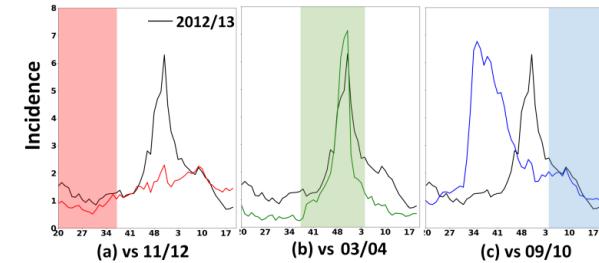
[C6] Explainable Forecasts (Challenges)

- Key to bridge decision making with forecasting
- However:
 - ML predictions not explainable
 - Important problem in stat. and neural models



[C6] Explainable Forecasts (Opportunities)

- **Explainable AI:** interpret predictions from the models
 - Feature-level importance measures
 - Similarity with historical data-points
- Examples:
 - Explainable neural representations, Saliency Maps [Molnar 2020]
 - Similarity with past data points [Adhikari+ KDD 2019, Kamarthi+ NeurIPS 2021]
 - Feature-level importance [Rodriguez+ AAAI 2021]

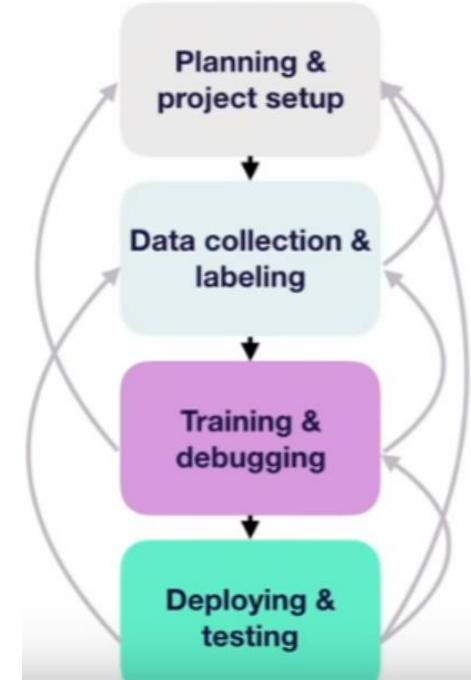


[C7] Technical debt of real-world deployment (Challenges)

- Model deployment involves lot of human involvement (even for Stat/ML models)
 - Require continuous monitoring and testing

Courtesy [Full Stack Deep Learning (FSDL) 2022]

- Called *technical debt* (borrowed from traditional software engineering)

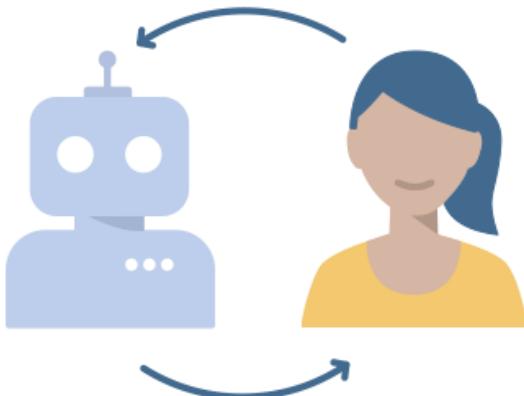


[C7] Technical debt of real-world deployment (Opportunities)

- **Systematic handling of modeling issues:**
 - Detecting data drift, adding new/correcting features
 - Updating/recalibrating model parameters to changing data distribution
- Examples:
 - AutoML solutions [He+ KBS 2021, Real+ ICML 2020]

[C7] Technical debt of real-world deployment (Opportunities)

- **Synergy of humans and models:**
 - Easily adapt expert knowledge



- Examples
 - Human-in-loop learning [Budd+ MIA 2021, Wilder+ IJCAI 2021]

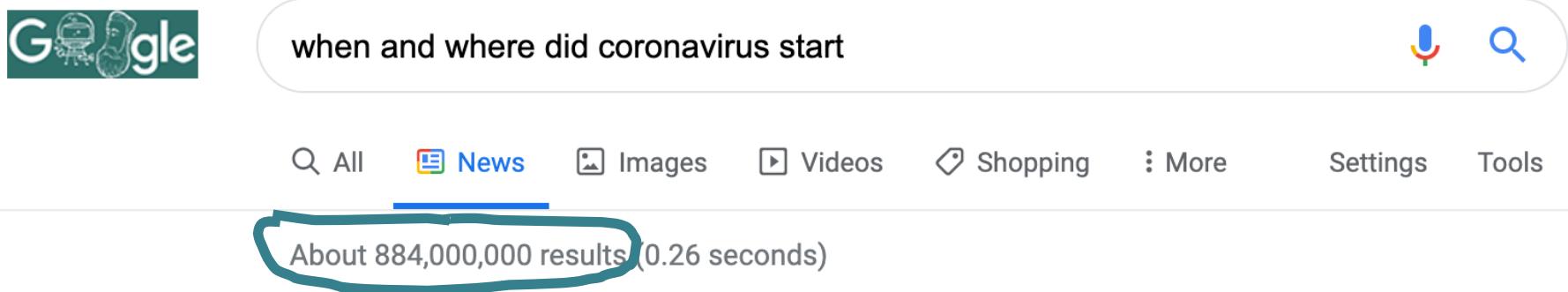
Final Remarks

[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!

[3] Asking What, When?

- 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**

[4] How to control?

- What measures should the government and people take?
 - Pharmaceutical interventions: vaccinations, anti-virals, other therapeutics
 - Non-pharmaceutical interventions: social distancing, closing schools and workplace, using masks, hand hygiene in hospitals
 - Allocate and distribute medical equipment and staff
- Typically resources are limited
 - Not enough vaccinations, hospital beds, ventilators
 - Need to take contact patterns into account!

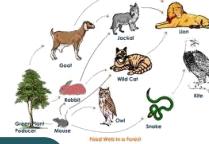
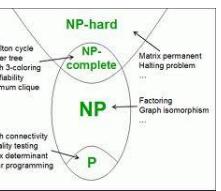
**Data + Models + Efficient
Algorithms + Simulations
+ Optimization tools**

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

Big Picture

Data Science for Epidemiology



Theory
&
Algo.

Biology

Physics



Comp.
Systems

Social
Science



ML &
Stats.

Econ.



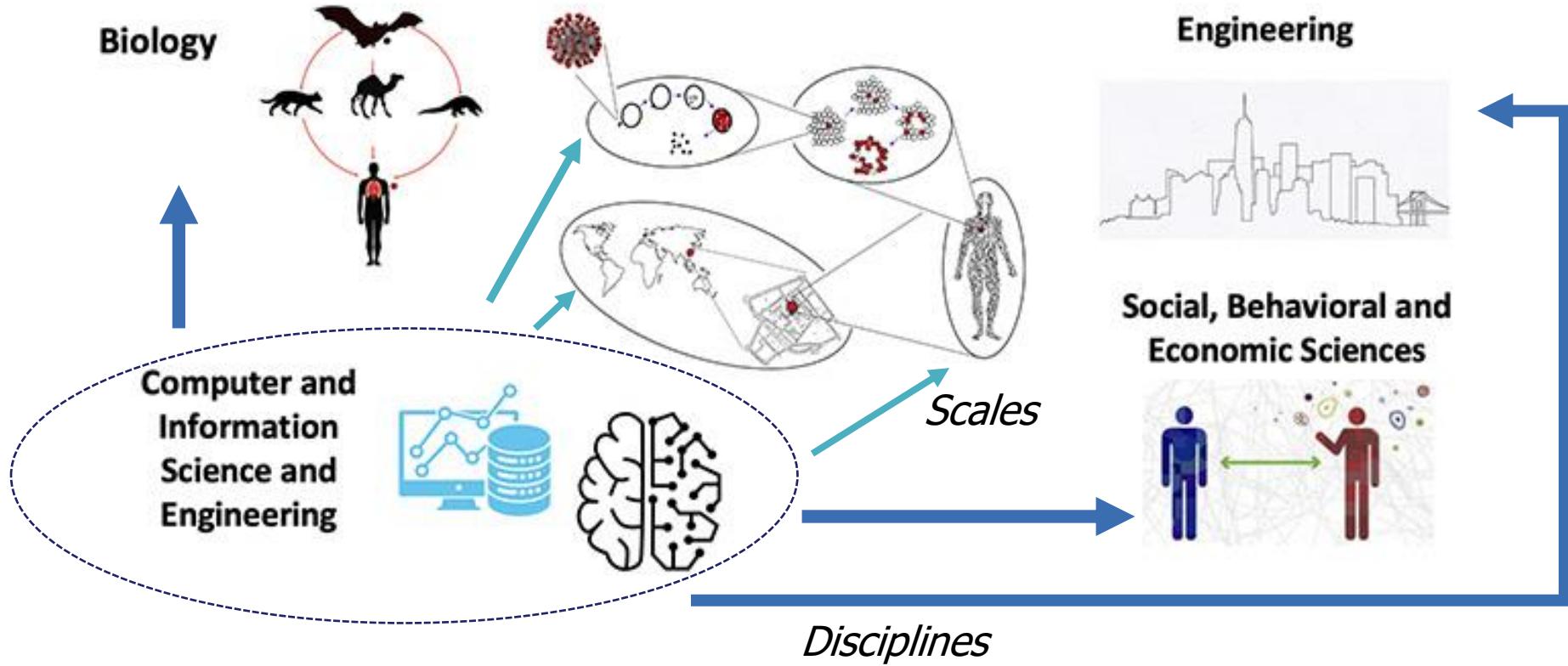
Reminder on Tutorial Webpage

- github.com/AdityaLab/kdd-22-epi-tutorial
- All Slides 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

- Epidemiology meets Data Science Workshop
 - <https://epidamik.github.io/>
 - Tentative venue: KDD 2023 in Long Beach, CA
 - Keynotes 2022:
 - Rachel Slayton (CDC)
 - Bryan Wilder (CMU)
 - Cecile Viboud (NIH)
- And more exciting research and tools!





We organized the **National PREVENT symposium**: Cross-cutting disciplines and scales for pandemic prevention and prediction

Videos and report: prevent-symposium.org



Acknowledgments

At Georgia Tech:

- Pulak Agarwal
- Javen Ho
- Mira Patel
- Suchet Sapre
- Jiaming Cui
- Jiajia Xie
- Lingkai Kong

- Bijaya Adhikari (Iowa)
- Anil Vullikanti (UVA)
- Naren Ramakrishnan (VT)
- Chao Zhang (GaTech)
- Sriram Pemmaraju (Iowa)
- Ayush Chopra (MIT)
- Ramesh Raskar (MIT)

Thanks!

- CDC COVID-19 Forecasting Hub
- Data collection volunteers
- Funding agencies



Fill survey: forms.gle/bvB8K1KwTo9knQUt5

Also available in
github.com/AdityaLab/aaai-23-ai4epi-tutorial

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