Data-driven Computational Epidemic Forecasting

Alexander Rodríguez

Harshavardhan Kamarthi

B. Aditya Prakash

College of Computing

Georgia Institute of Technology

December 1, 2021

AdityaLab @ Georgia Tech

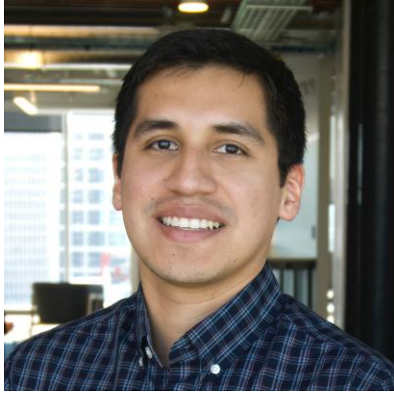
- One of our lab’s focus: explore performance of data-driven methods in epidemiology/public health (surveillance, interventions, vaccination,…)
  - Data from multiple source is often more sensitive to what is happening ‘on the ground’
  - Complementary helpful perspective to other traditional methods

Aditya Lab

About us

- PI: B. Aditya Prakash
  - Assoc. Professor
  - PhD. CMU, 2012.
  - Data Mining, Applied ML
  - Networks and Sequences
  - Applications:
    - Epidemiology and Public Health
    - Urban Computing
    - The web
    - Security
  - Homepage: https://www.cc.gatech.edu/~badityap/

About us



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

About us

- Harshavardhan Kamarthi
  - 2nd year PhD student
  - Research Interests
    - Epidemic forecasting
    - Probabilistic forecasting and uncertainty quantification
    - Deep Probabilistic models
  - Homepage: https://harsha-pk.com/

Workshop Webpage

- https://adityalab.cc.gatech.edu/workshops/21-forecasting-f4sg.html or b.gatech.edu/3cBPfQ7
- All Slides will be posted there. Talk video as well (later).
- **License**: for education and research, you are welcome to use parts of this presentation, for free, with standard academic attribution. For-profit usage requires written permission by the authors.

Outline

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

- 15 min breaks after Part 2 and Part 3
  - We'll be available for questions

Plan for the Workshop

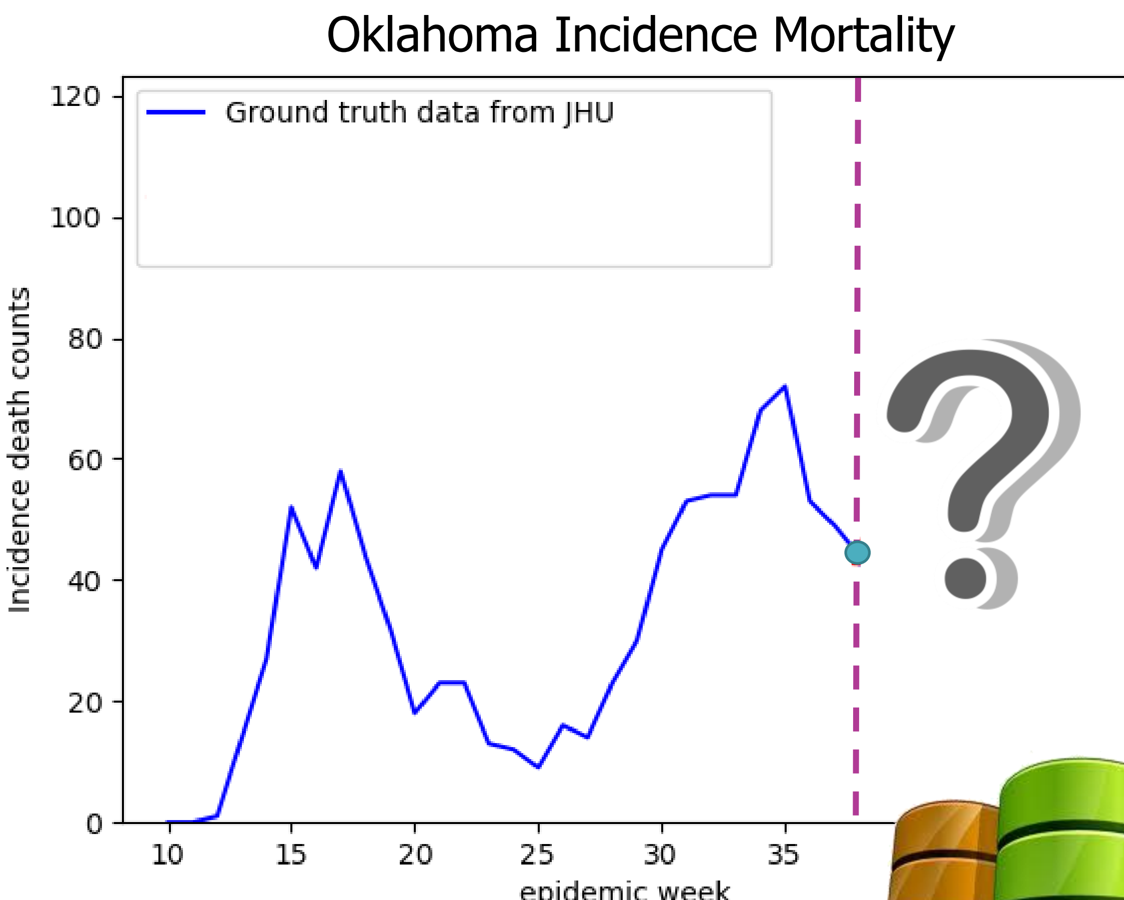
- Theory and research
  - Setting up the epidemic forecasting problem
  - General epidemiology: key concepts and models
  - Statistical modeling and deep learning
    - Research innovations
- Practice
  - US real-time forecasting experiences
  - Coding examples
    - Mechanistic models
    - Statistical models
  - Demo session
    - Statistical correction of forecasts

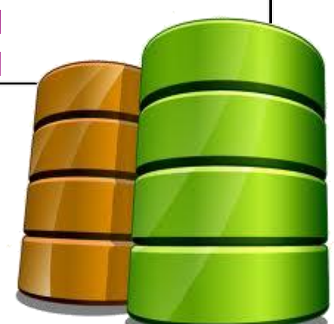
Workshop focus:
- Computational data-driven methods
- Short-term forecasting (up to 4 weeks ahead)

Forecasting Infectious Diseases

- Why? Allocate resources/budget, inform public policy, improve preparedness
- Background:
  - Traditional methods are based on ODEs and agent-based models
  - Data collection has increased
    - Methods have difficulties ingesting these data sources

Real-time Epidemic Forecasting





Possible near future:

- Goes down
- Stays still
- Goes up

Depends on:

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

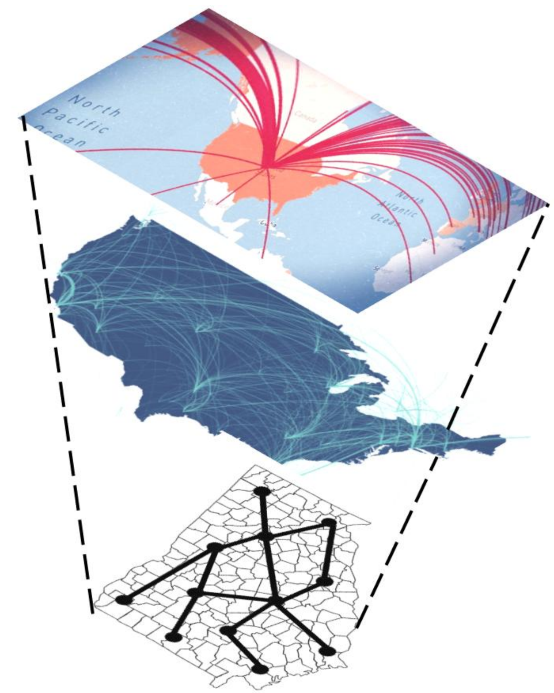
SAFEGRAPH

Google kinsa

Georgia Tech

Why Computational Data-driven Forecasting?

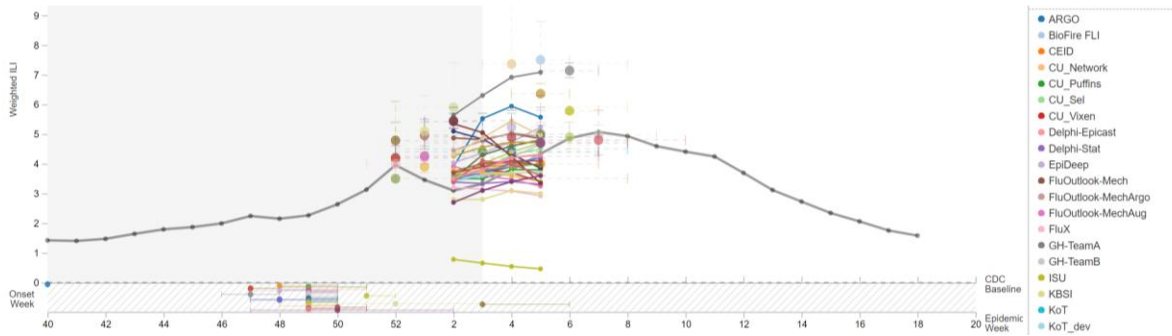
- Epidemic spread is a spatiotemporal phenomena over multi-scale networks



- New end-to-end methods available capable of modeling data with minimal assumptions
- Before and after the COVID-19 pandemic: Explored **performance** and **utility** of data-driven models in short-term forecasting

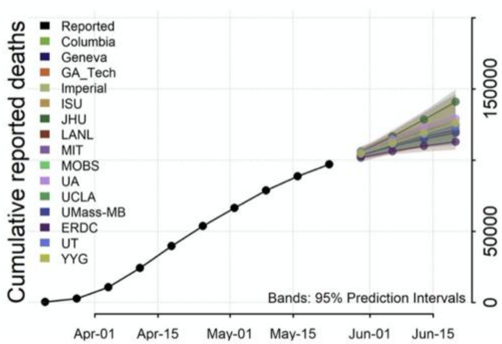
Our Participation on CDC Forecasting Initiatives

Target 1: Influenza like illness per week



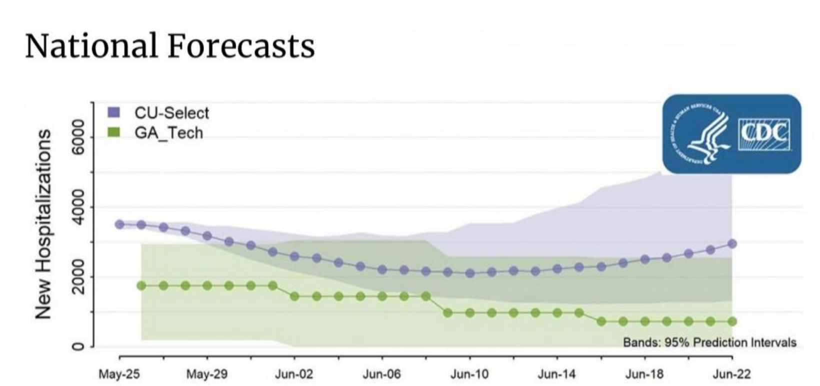
Last few years
Also in COVID-ILI (March 2020)

Target 2: Weekly Covid Mortality



Since April End 2020

Target 3: Daily Covid Hospitalizations



Our Impact

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

CDC

FiveThirtyEight

1 of 11 shown on their page

1st Prize

facebook

Carnegie Mellon University

The COVID-19 Symptom Data Challenge

Out of 115 global participants

2nd Prize

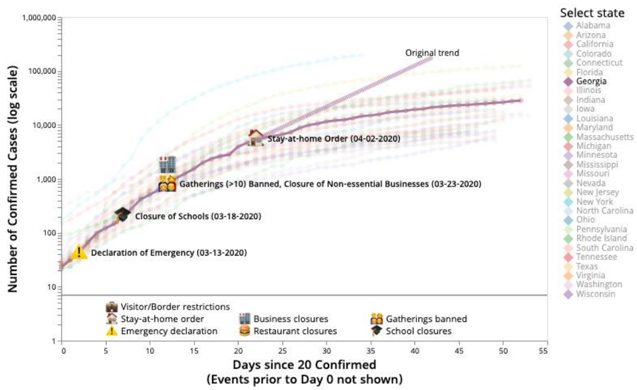
C3.ai COVID-19 Grand Challenge

43 Countries

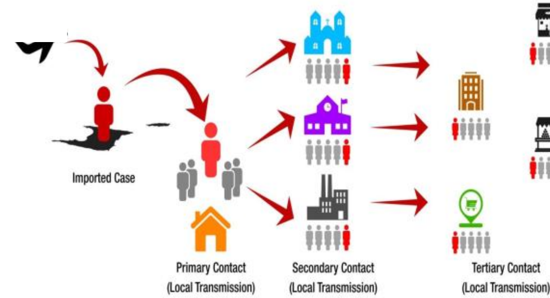
777 Participants

COVID Response

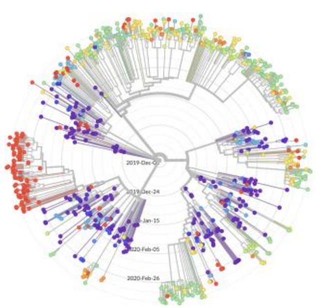
https://www.cc.gatech.edu/~badityap/covid.html



Visualizing impact of nonpharmaceutical interventions

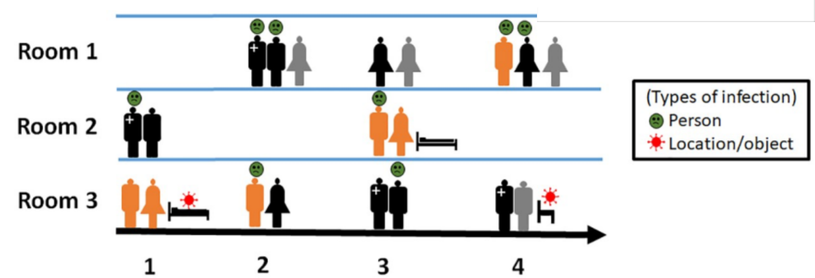


On-campus Mobility and Data-driven Interventions



Anon ID	Infection Time	Infected By	Location	Age	Gender	Strain
1	Dec 15, 2019		98122	27	M	C17747T
2	Jan 4, 2019	1	98112	44	F	C12915T
3		1	98144		M	
4	Jan 17, 2020	2	98105	56	F	
5	Jan 19, 2020			32		C1419T
6	Jan 25, 2020	3	98134		M	
7	Jan 27, 2020	???	98168	22	F	
8	Jan 27, 2020	???			F	
9	Feb 2, 2020	???	98125		M	

Adaptive surveillance



Hospital Acquired Infections

… and others like vaccine allocation algorithms etc.

Rodríguez, Kamarthi, and Prakash 2021

Recent Publications

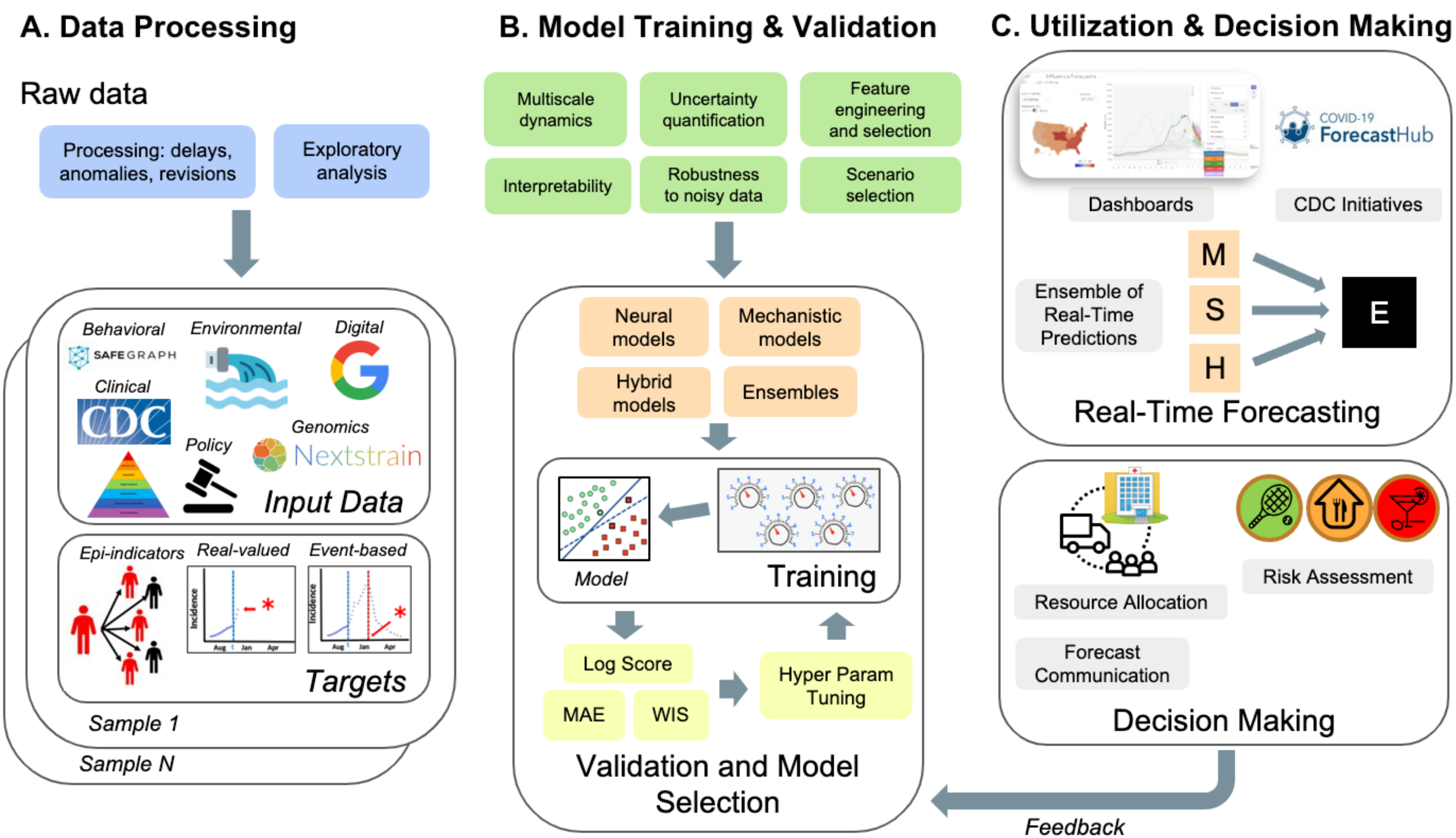
- A. Rodríguez, N. Muralidhar, B. Adhikari, Anika Tabassum, N. Ramakrishnan, B. A.Prakash. Steering a Historical Disease Forecasting Model Under a Pandemic: Case of Flu and COVID-19. In AAAI-21.
- A. Rodríguez, A. Tabassum, J. Cui, J. Xie, J. Ho, P. Agarwal, B. Adhikari, B. A. Prakash. DeepCOVID: An Operational DL-driven Framework for Explainable Real-time COVID-19 Forecasting. In IAAI-21.
- 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.
- H. Kamarthi, A. Rodríguez, B. A. Prakash. Back2Future: Leveraging Backfill Dynamics for Improving Real-time Predictions in Future. In submission (available as arXiv preprint).
- 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).
- E. Cramer et al. Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US In submission (available as medRxiv preprint).

Coming up soon

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

Part 1: Epidemic Forecasting

Epidemic Forecasting Pipeline



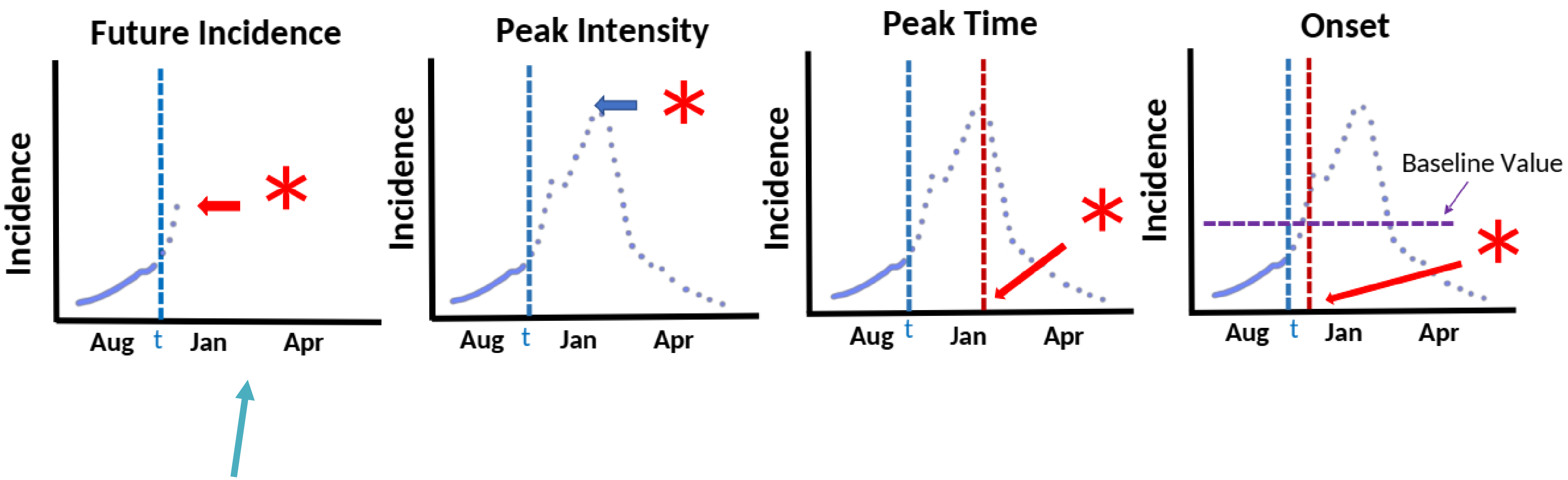
Rodríguez, Kamarthi, and Prakash 2021

Epidemic Forecasting Setting

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

[1] Common Forecasting Tasks

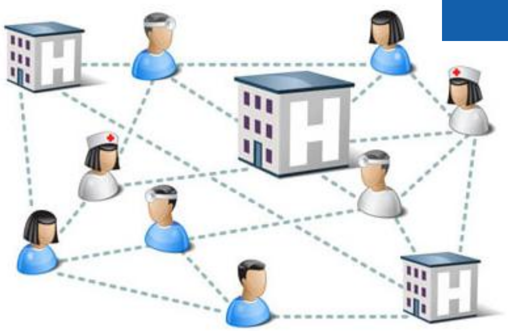
- Used in annual CDC Flu forecasting challenge



Short-term forecasting: Up to 4 weeks ahead

[2] Targets of Interest

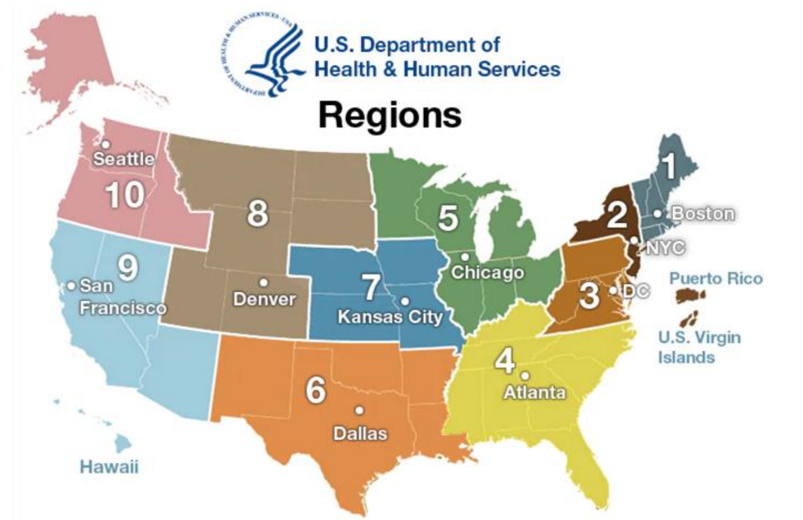
- Influenza
  - %ILI: symptomatic outpatients
    - Syndromic surveillance
  - Lab-tested hospitalizations
- COVID-19
  - Mortality
  - Hospitalizations
  - Cases

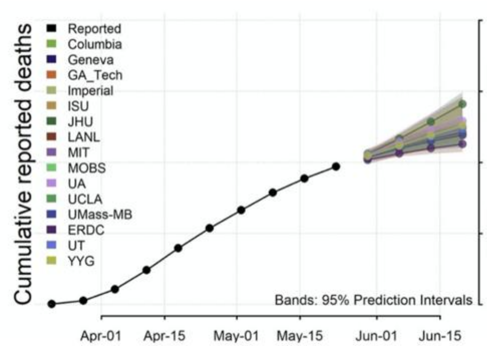


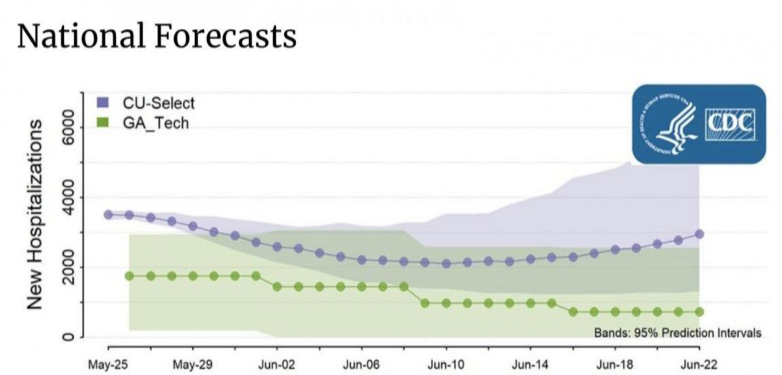
ILINet surveillance network

[3] Spatial and Temporal Scales

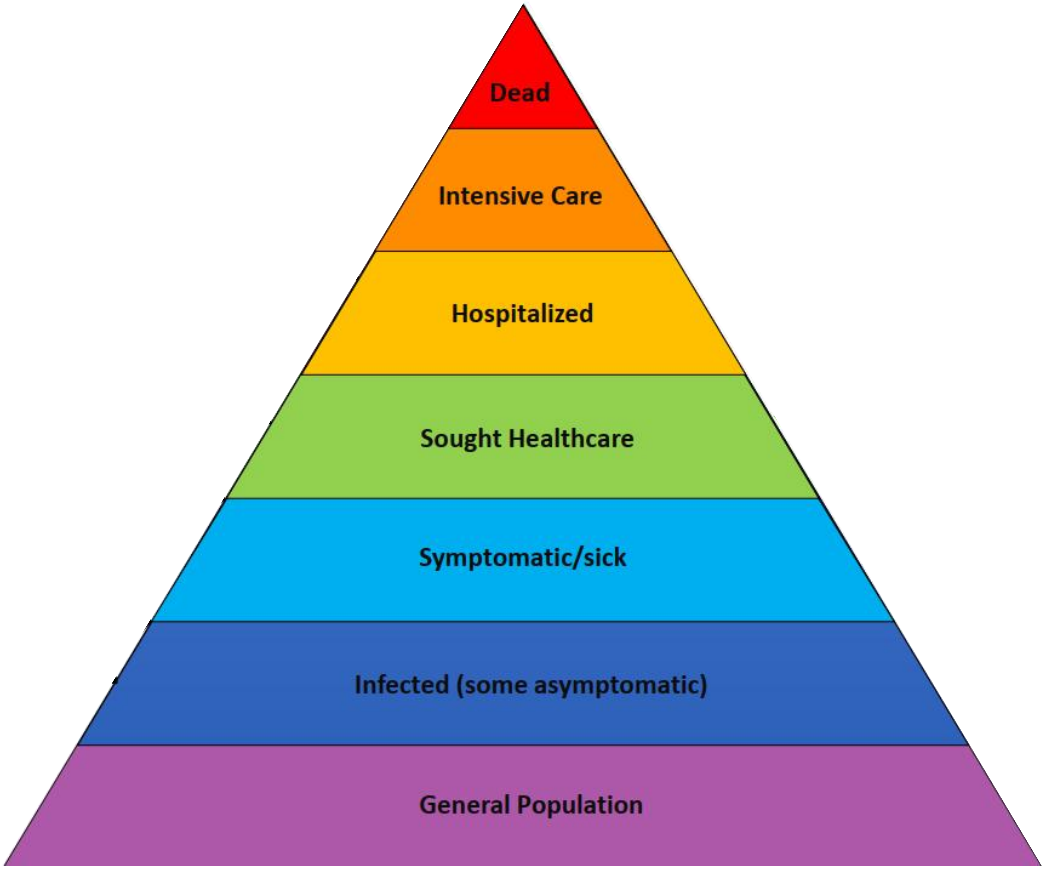
- Spatial scales:
  - National
  - Region/state/province
  - County/city (less common)
- Temporal scales:
  - Weekly
  - Daily







[4] Datasets: surveillance pyramid



Line-list data

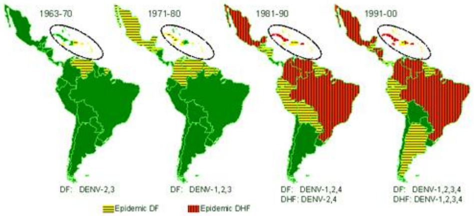
- Who, when and where a person was infected

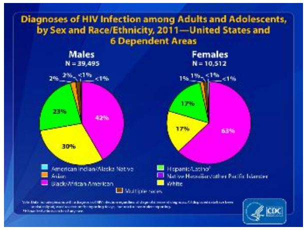
Hospital records

Lab surveys

Population surveys

Surveillance Reports





Digital epidemiology

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

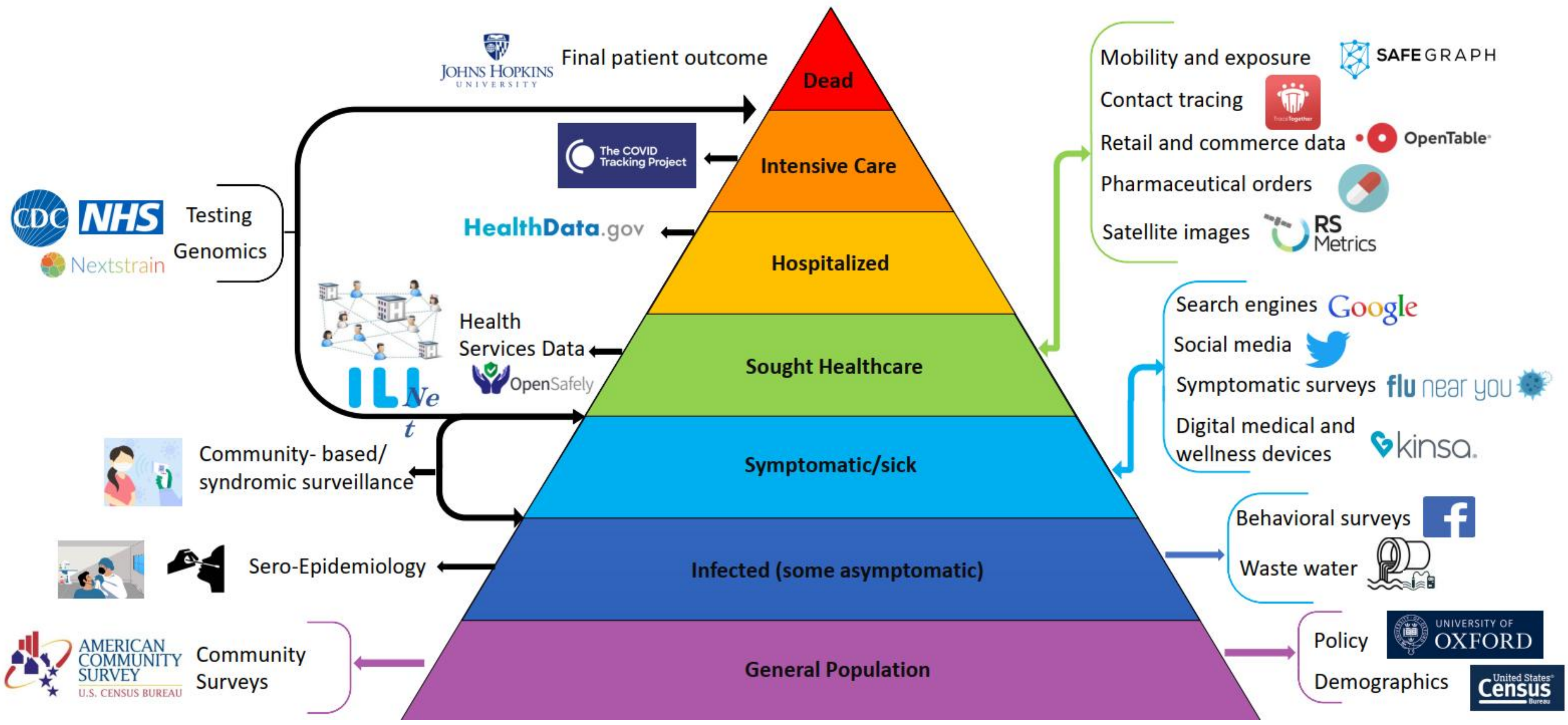
Review

Digital Epidemiology

Marcel Salathé[1,2]*, Linus Bengtsson[3], Todd J. Bodnar[1,2], Devon D. Brewer[4], John S. Brownstein[5], Caroline Buckee[6], Ellsworth M. Campbell[1,2], Ciro Cattuto[7], Shashank Khandelwal[1,2], Patricia L. Mabry[8], Alessandro Vespignani[9]

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

Surveillance pyramid and datasets

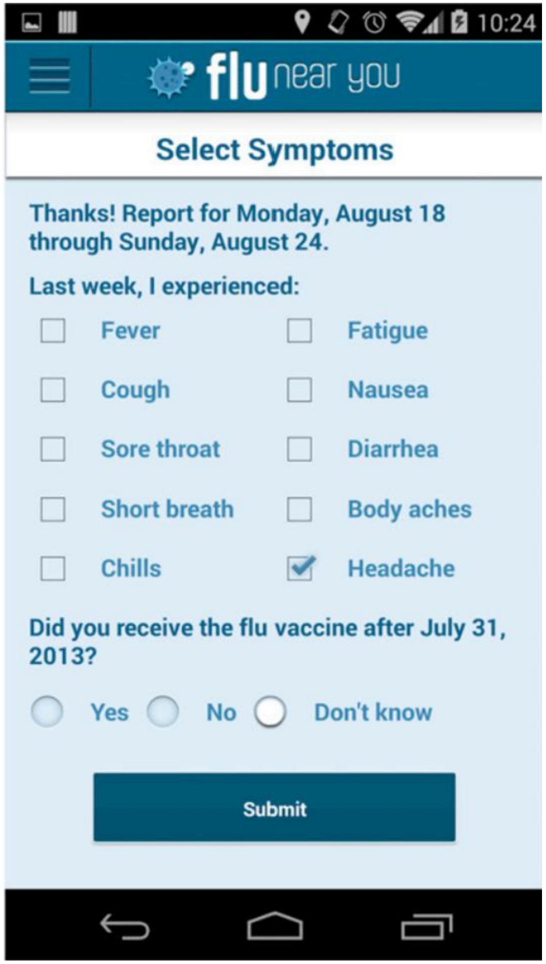


Search Engines and Social Media

- Search activity
  - Ad-hoc search engines
  - Specialized search engines
- Social media
  - Tweets
  - RSS feed

Online Surveys

- Symptomatic surveys
- Behavioral surveys
  - Adoption of public health recommendations
  - Mask wearing
  - Social distance



Mobility

- Quantify contact patterns within and across communities
- Sources:
  - Mobile call records
  - Mobile apps

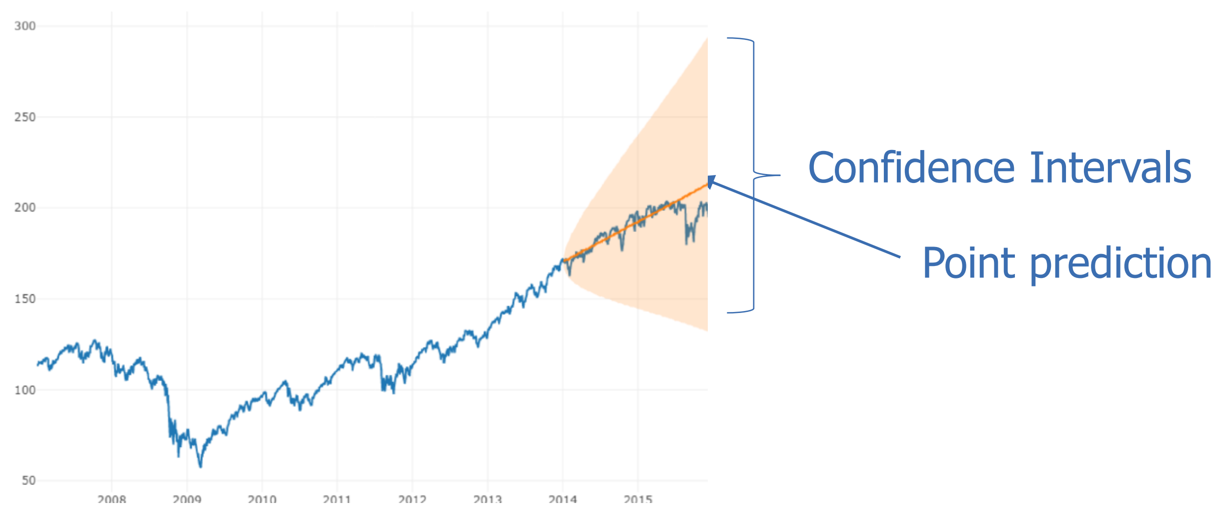
SAFEGRAPH

Satellite Images

[Brownstein+ 2020]

[5] Model evaluation

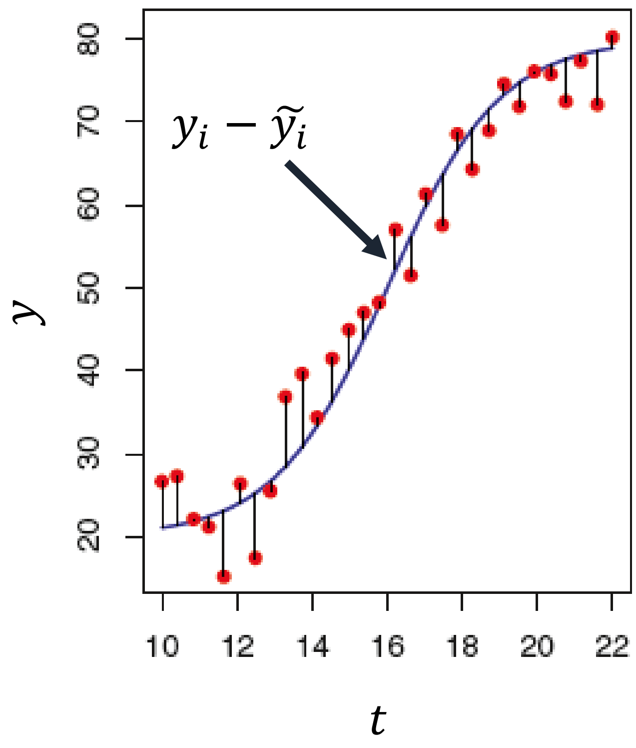
- Point Forecasts: Single value per forecast
- Probabilistic Forecasts: Probability distribution of forecast
  - Captures uncertainty, useful for decision making



Rodríguez, Kamarthi, and Prakash 2021

Evaluation of Point Forecasts

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



Evaluation of Probabilistic Forecasts

- Log Score: $\frac{1}{T}\sum_{i=1}^{T}\ln(p_i(y_i))$
  - Log probability of ground truth outcome (binned)
- Other metrics
  - Coverage score
  - Interval score & Weighted Interval Score (WIS) [Bracher+ 2021]

$$\mathrm{IS}_\alpha(F,y) = (u-l) + \frac{2}{\alpha}(l-y)\mathbb{1}(y<l) + \frac{2}{\alpha}(y-u)\mathbb{1}(y>u)$$

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

How to choose eval. metrics?

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

Part 2: Mechanistic Models

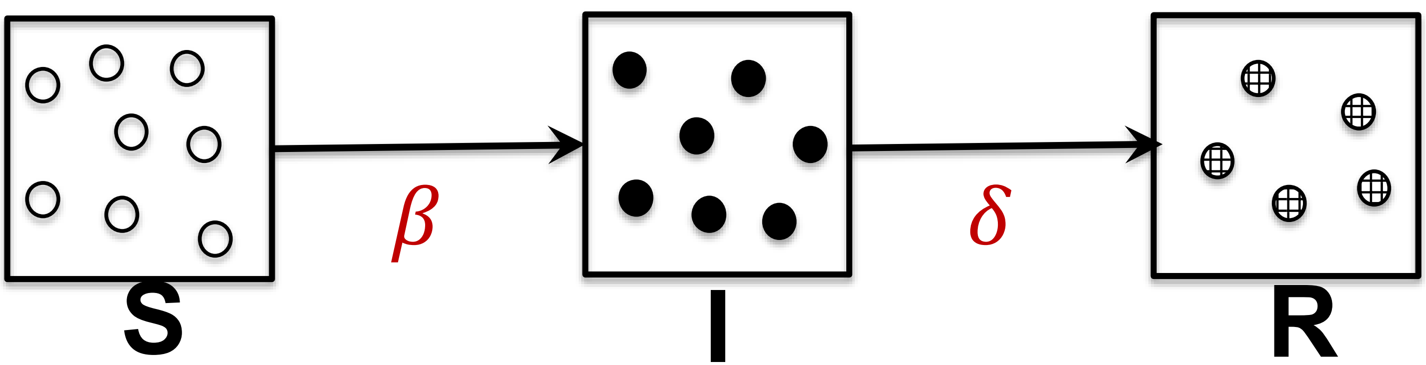
Mechanistic models

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

- Modeling approaches:
  1. Mass-action models (ODE models)
  2. Metapopulation models
  3. Agent-based networked models

[1] ODE Models: SIR

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



Assumptions

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

SIR Model

$$\frac{dS}{dt} = -\beta S I$$

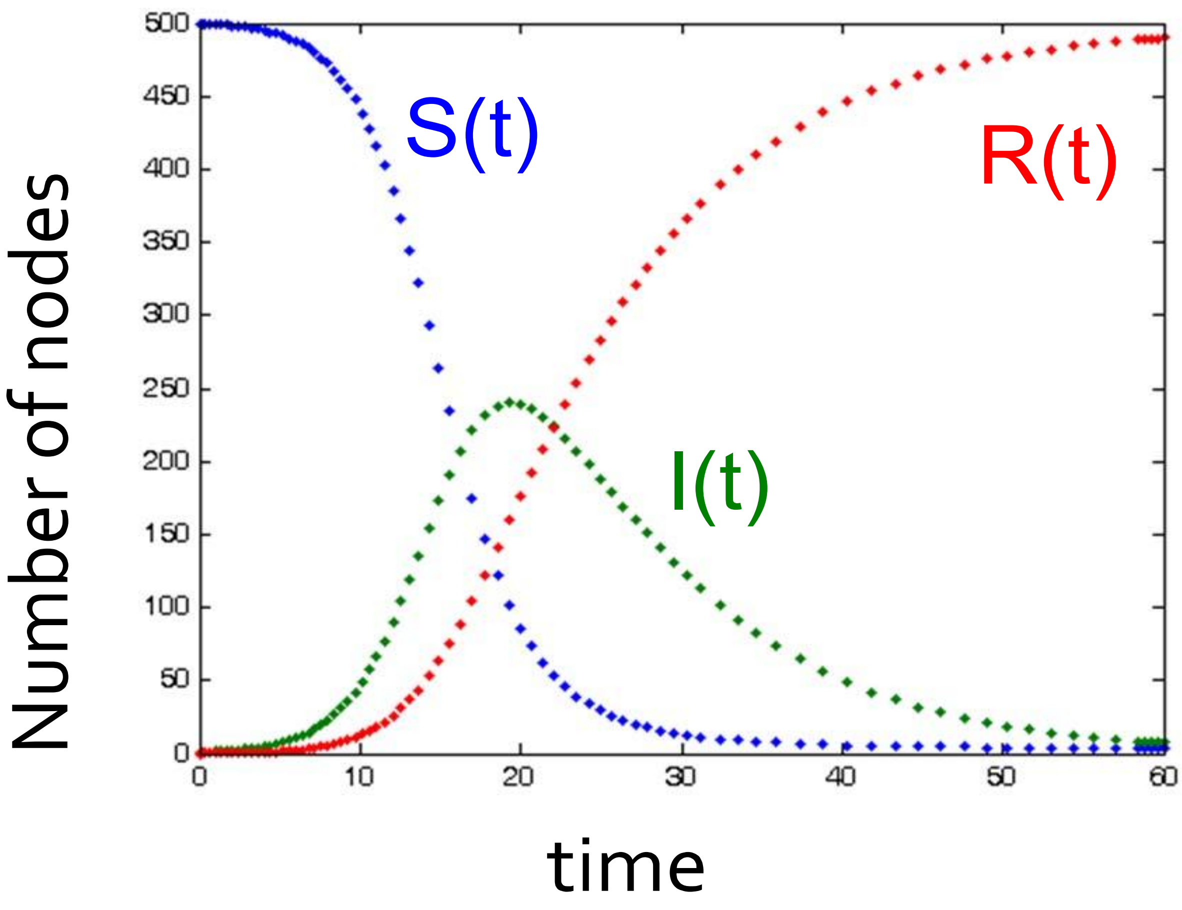
$$\frac{dI}{dt} = \overbrace{\beta S I}^{\text{Number of new infections = \textbackslash beta * \# infection attempts}} - \underbrace{\delta I}_{\text{Number of infected nodes curing}}$$

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

Solving SIR

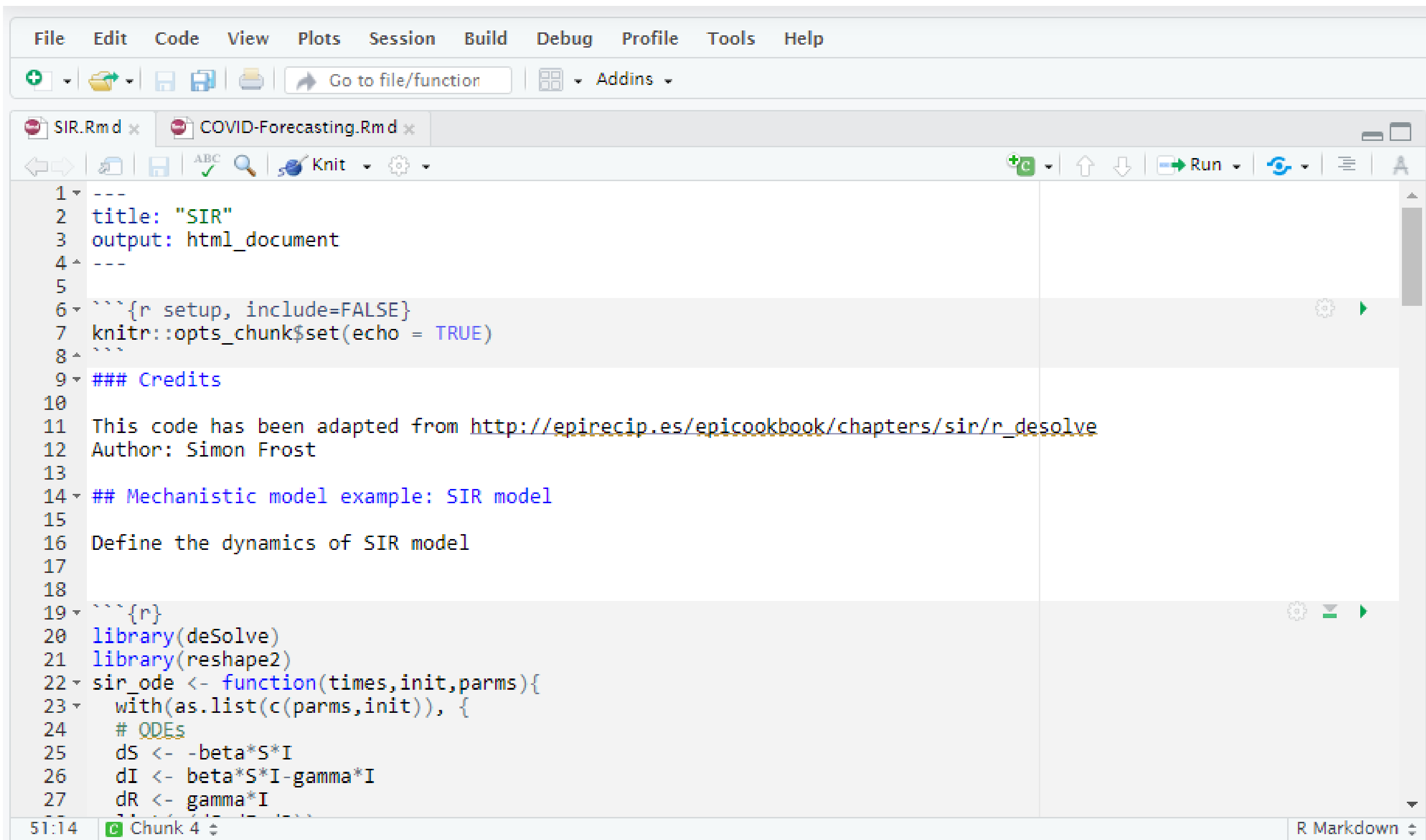
- No closed form solution!

SIR: numerical output



Online interactive example

Data-Driven Computational Epidemic Forecasting / R-Notebook



Many many extensions

- With birth/death rates (‘vital dynamics’)
- Variable contact rates
- Age-structured models
- Make things stochastic
- Multiple viruses/diseases
- ……..
- ……..
- See Hethcote 2000, and the book by May and Anderson 1992

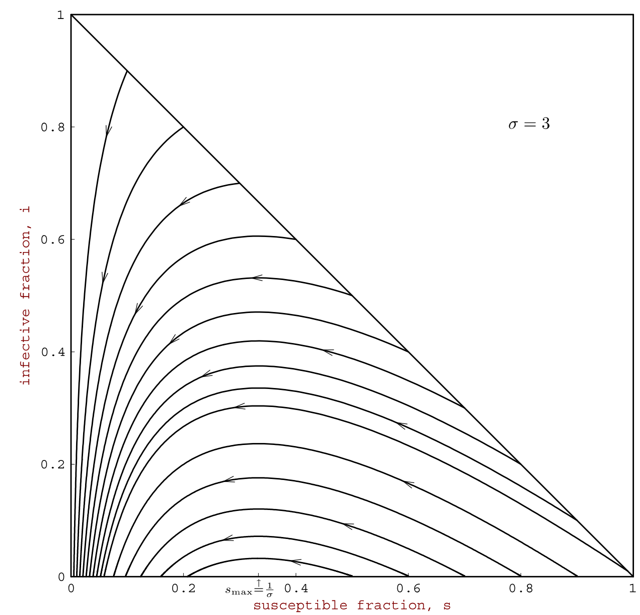
SIR: implicit solution

$$S(t) = S(0)e^{-R_0(R(t)-R(0))}$$

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

$$R_0 = N\beta/\delta$$

Reproductive Number



Threshold Phenomenon: R0

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

- This implies

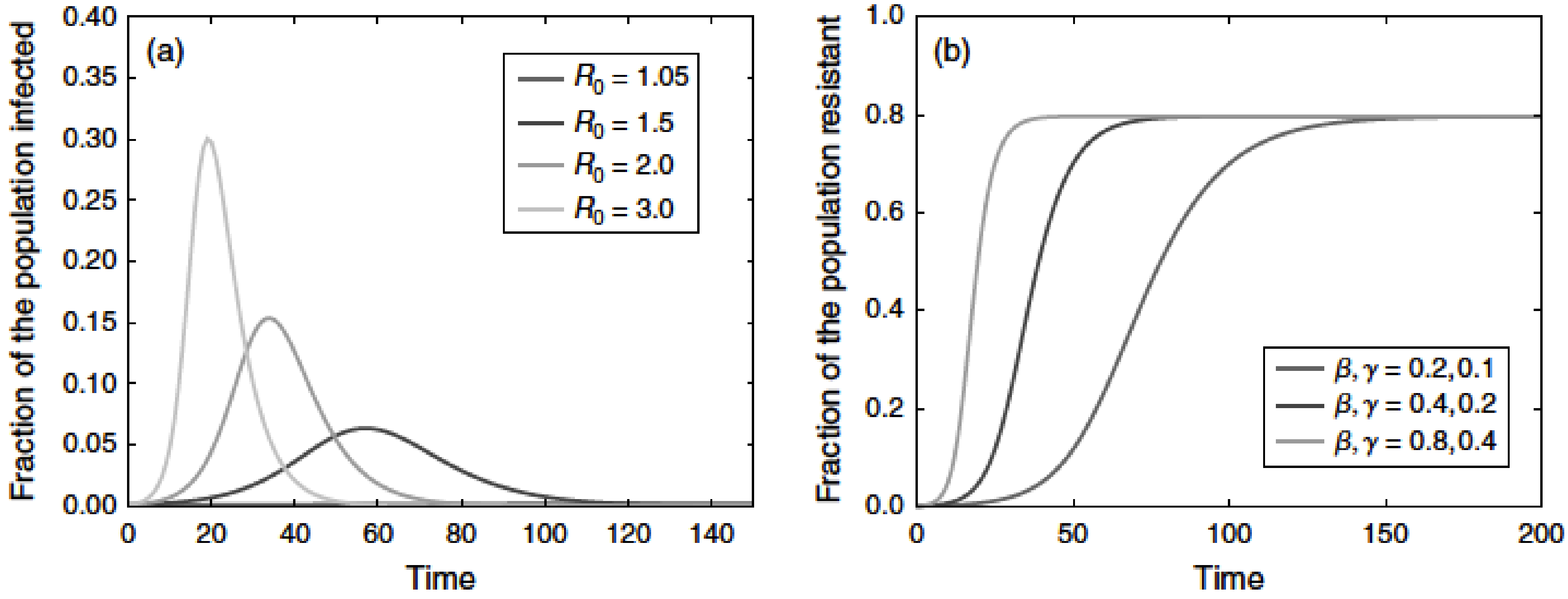
$$\frac{dI}{dt} < 0 \quad \text{if} \quad S(0) < \delta/\beta$$

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

Threshold Phenomenon

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

R0 and disease dynamics



Source: Dimitrov and Meyers, INFORMS 2010

R0 of various diseases

Disease	Transmission	R_0
Measles	Aerosol	12–18[29][30]
Chickenpox (varicella)	Aerosol	10–12[31]
Mumps	Respiratory droplets	10–12[32]
Rubella	Respiratory droplets	6–7[b]
COVID-19 (Delta variant)	Respiratory droplets and aerosol	5–8[37]
Polio	Fecal–oral route	5–7[b]
Pertussis	Respiratory droplets	5.5[38]
Smallpox	Respiratory droplets	3.5–6.0[39]
COVID-19 (Alpha variant)	Respiratory droplets and aerosol	4–5[37]
HIV/AIDS	Body fluids	2–5[40]
COVID-19 (ancestral strain)	Respiratory droplets and aerosol[41]	2.9 (2.4–3.4)[42]
SARS	Respiratory droplets	2–4[43]
Diphtheria	Saliva	2.6 (1.7–4.3)[44]
Common cold	Respiratory droplets	2–3[45]
Ebola (2014 outbreak)	Body fluids	1.8 (1.4–1.8)[46]
Influenza (2009 pandemic strain)	Respiratory droplets	1.6 (1.3–2.0)[2]
Influenza (seasonal strains)	Respiratory droplets	1.3 (1.2–1.4)[47]
Andes hantavirus	Respiratory droplets and body fluids	1.2 (0.8–1.6)[48]
Nipah virus	Body fluids	0.5[49]
MERS	Respiratory droplets	0.5 (0.3–0.8)[50]

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

Source: Wikipedia 2021

[2] Metapopulation Models

- Spatially structured
- For example: modeling COVID-19 and influenza, Zika, Ebola…
- Model heterogeneity by using travel data
  - But assume homogeneity at 'right' granularities

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}

Georgia Tech

Metapopulation Models contd.

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

- Written in terms of X^{eff}, Y^{eff}, Z^{eff}

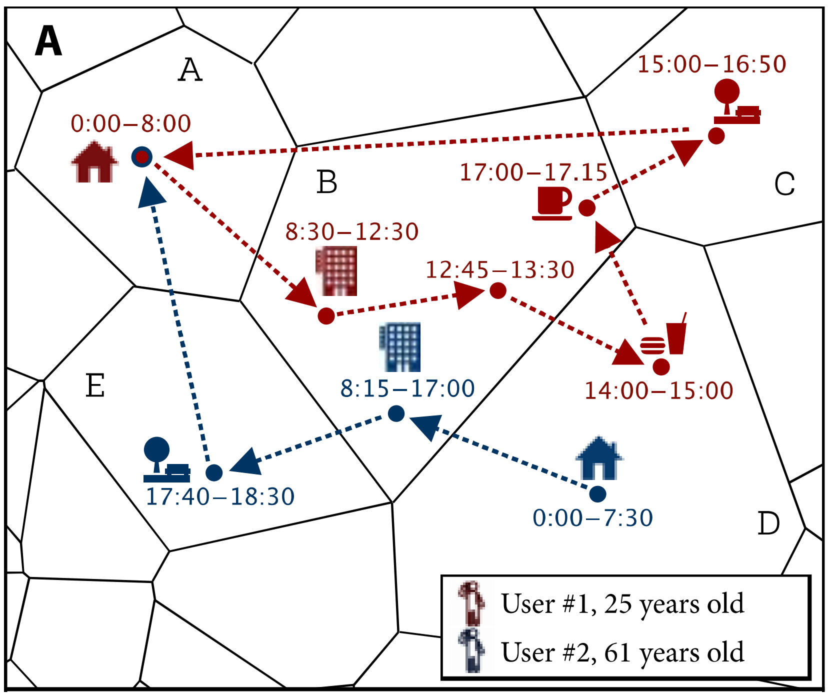
But… Human contact patterns are not random

Source: Mi Jin Lee at petterhol.me

How to Capture Them? Example: Using Call Data Records

- Many recent studies on this topic

#raw data



#trips

B

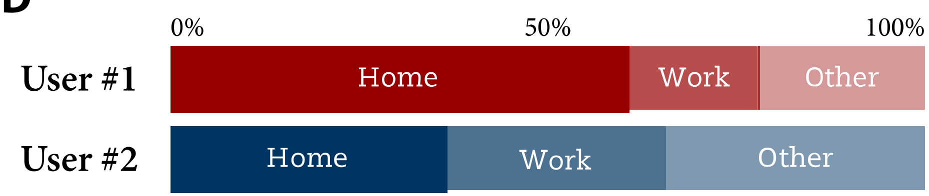
	A	B	C	D	E
A	0	1	0	0	0
B	0	1	1	1	1
C	1	0	0	0	0
D	0	2	0	0	0
E	1	0	0	0	0

#contacts

C

	<20	20–29	30–59	60–79	80+
<20	0	0	0	0	0
20–29	0	0	0	1	0
30–59	0	0	0	0	0
60–79	0	1	0	0	0
80+	0	0	0	0	0

D



[Oliver et al, Sci. Adv. 2020]

Numerous COVID-19 examples

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

[3] Agent-based networked models

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

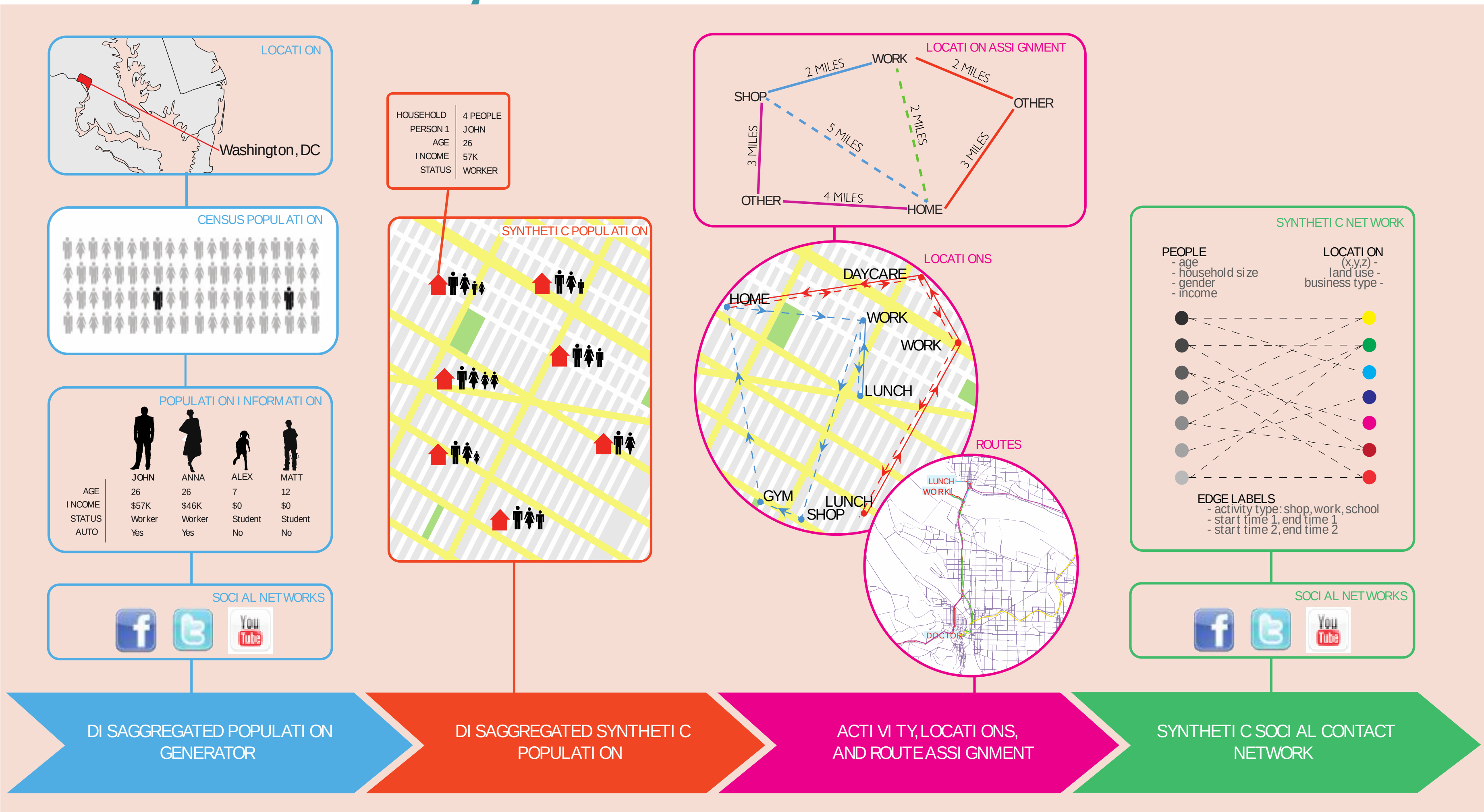
First principles Approach for Constructing Social Contact Networks

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

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

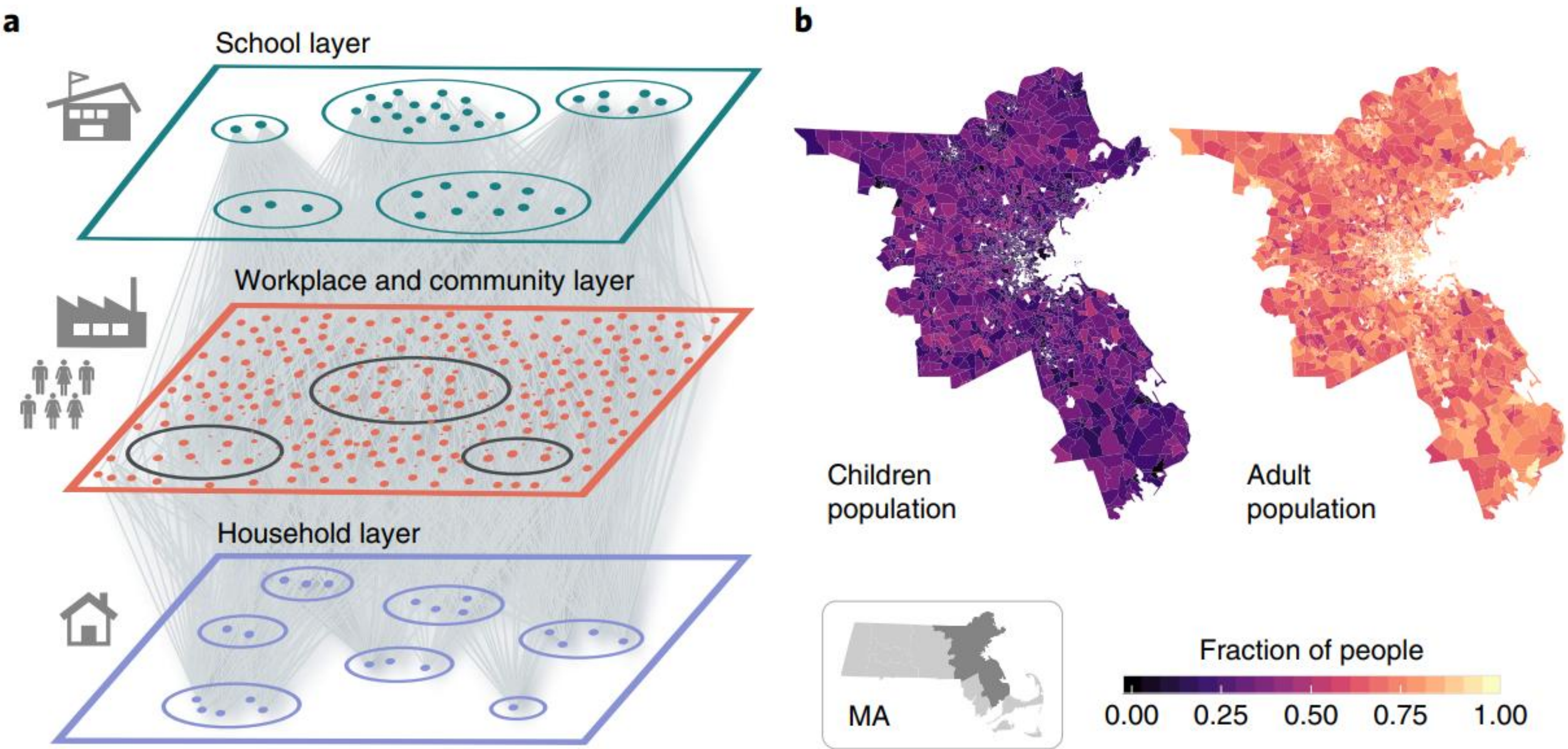
[Marathe and Vullikanti, CACM 2013]

First principles Approach for Constructing Social Contact Networks



[Marathe and Vullikanti, CACM 2013]

Example: COVID-19 in MA



[Aleta et al, Nature Human Behavior 2020]

Calibration of Mechanistic Models

- Estimate parameters
  - Beta, delta, initial conditions

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

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

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

Typically

- Ranges of parameters
  - From epidemiological data

- Try to model uncertainty in the data
  - Multiple stochastic calibrations

Pros/Cons Mechanistic Models

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

- Useful to get intuition and some broad principles
  - More qualitative rather than quantitative

Pros/Cons contd.

- Sometimes does not reflect reality
  - SARS example
    - High R0 (2.2-3.6)
    - Estimates were based on hospital wards, where full mixing was reasonable
- Calibration is challenging
  - Small deviations in parameters can lead to very different results

Remarks

- A lot more to say about mechanistic models
  - Only reviewed some concepts and models
- Other resources:
  - N. Dimitrov and L. Meyers. 2010. Mathematical approaches to infectious disease prediction and control. INFORMS, 1–25
  - 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.

Part 3: Statistical Models

Statistical Models

- Also known as phenomenological models.
- Intuition:
  - Find the best function from a family of functions that approximate forecast target given input data.
  - Best approximate is found using past training data.

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

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

[1] Regression Models

- Assume a linear relationship between input features and future forecast $\tilde{y} = w_0 + \mathbf{w}^{\mathbf{T}}\mathbf{x}$
- The features x can be high-dimensional set of multi-modal features
  - Eg: Past values of epidemic curve (called AutoRegressive models), Search query volumes , word occurrence in text, etc.

AutoRegressive Models

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

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

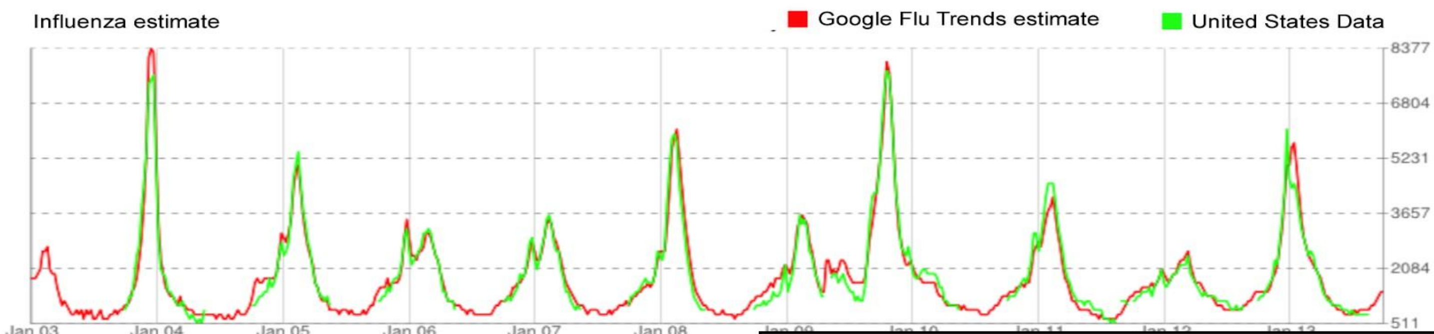
- We can also add difference between values as features (like in ARIMA)

Google Flu Trends

[Ginsberg+ 2009 Nature]

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

$$logit(P) = \beta_0 + \beta_1 \to logit(Q) + \varepsilon$$



High-impact work, media coverage

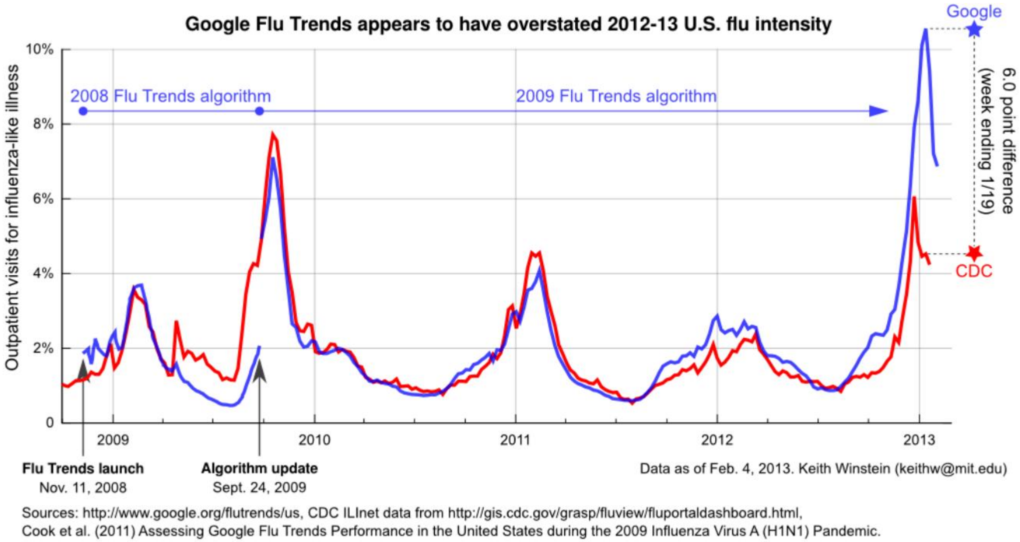

The New York Times

Google Uses Searches to Track Flu's Spread

By Miguel Helft
Nov. 11, 2008

However,…

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



ARGO

[Yang+ 2017 SR]

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

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

ARGO2

[Ning+ 2019 Sci. Reports]

- 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

Georgia Tech

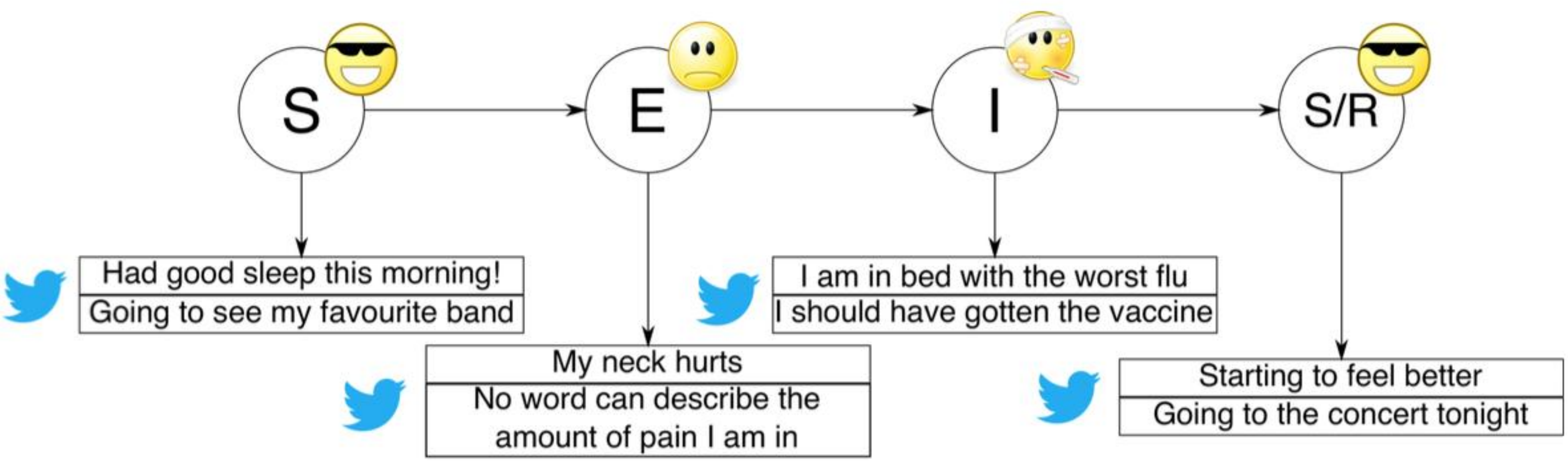
[2] Language models: using Tweets to forecast H1N1 pandemic

[Chen+ ICDM ‘17]

- Topic modelling approach
  - Cluster tweets
- Combines
  - Information propagation on Twitter
  - Epidemiological model

States of infection cycle

- Model states of infection cycle using tweets



Forecasting

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

(a) S state (b) E state (c) I state

HFSTM Model

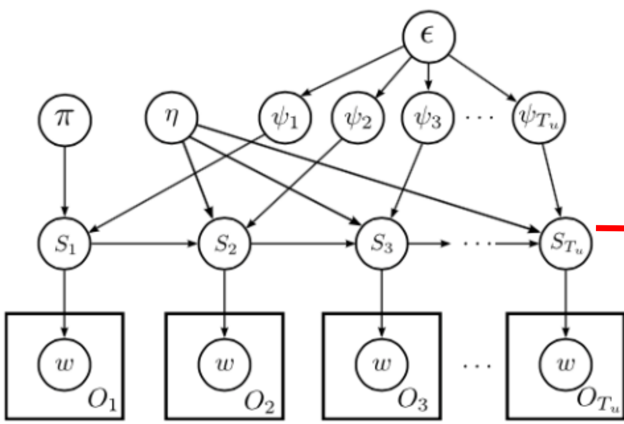
- Generating tweets
  - Generate state for tweet
  - Generate topic for word

S: This restaurant is really good

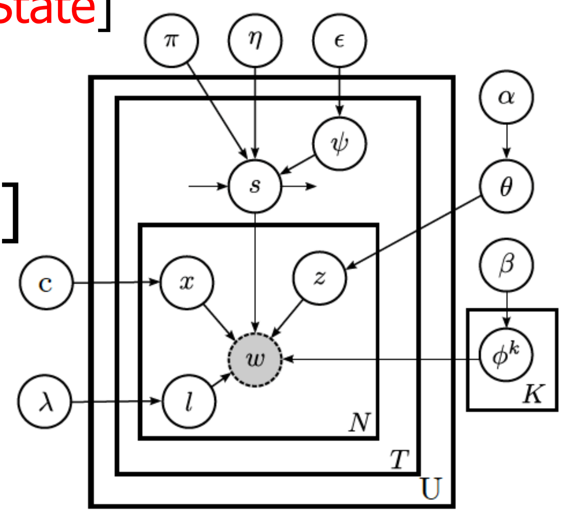
E: The movie was good but it was freezing

I: I think I have flu

Topic: [Background,
Non-flu,
State]

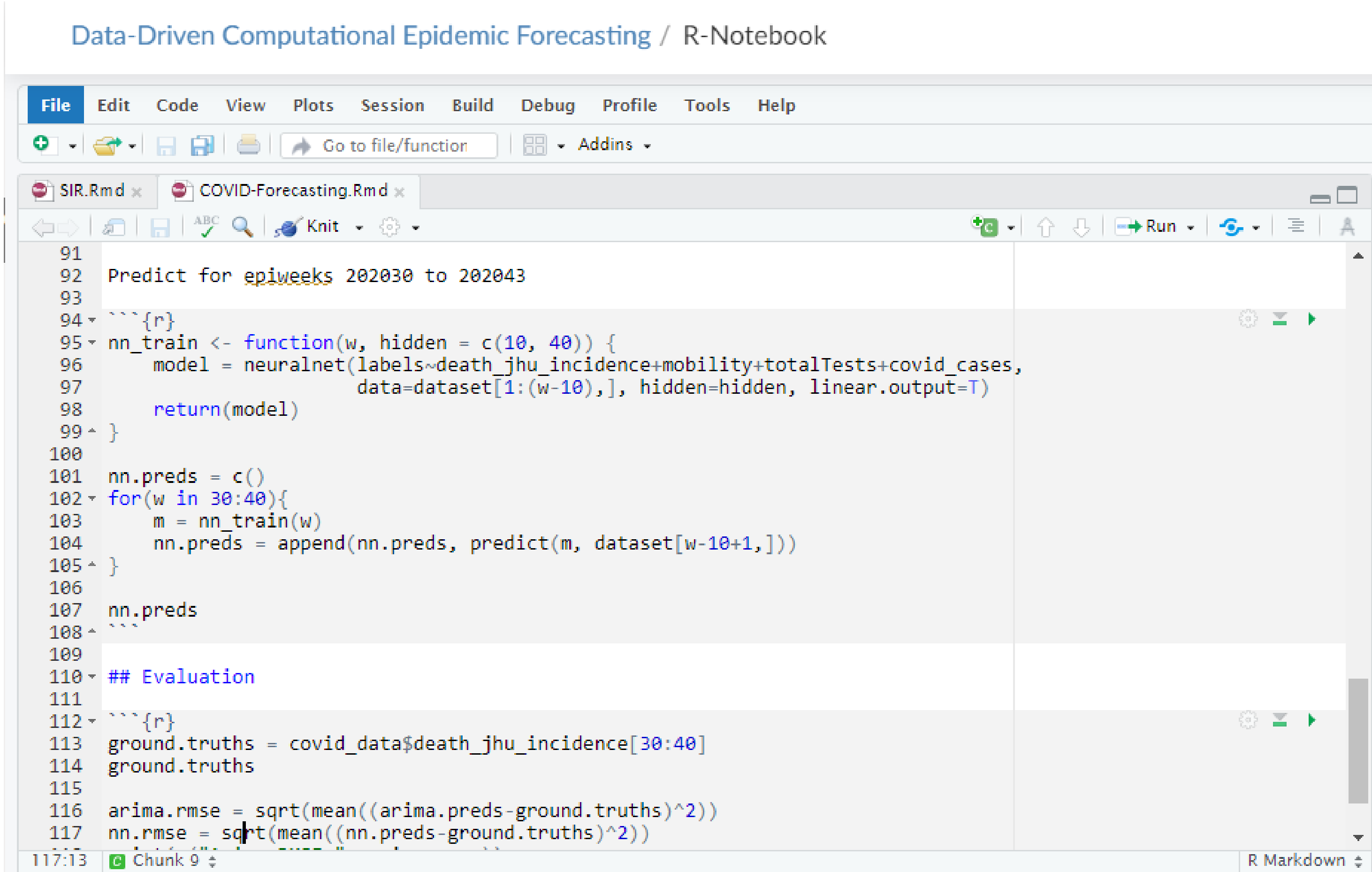


State: [S,E,I]



Use EM Algorithm for learning parameters

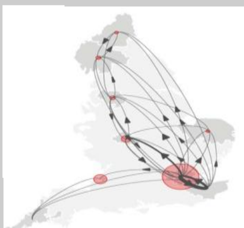
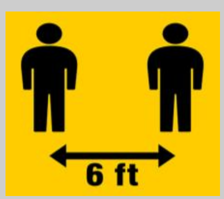
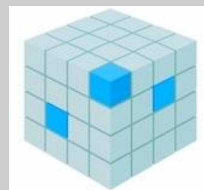
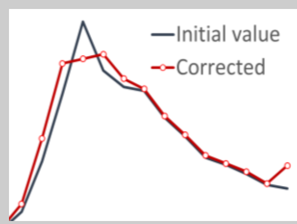
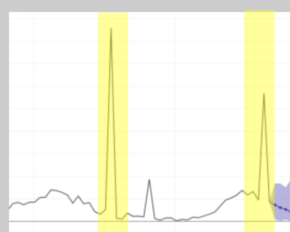
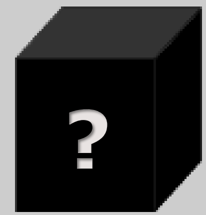
Online interactive example



[3] Neural Models

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

Modeling considerations for neural models

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

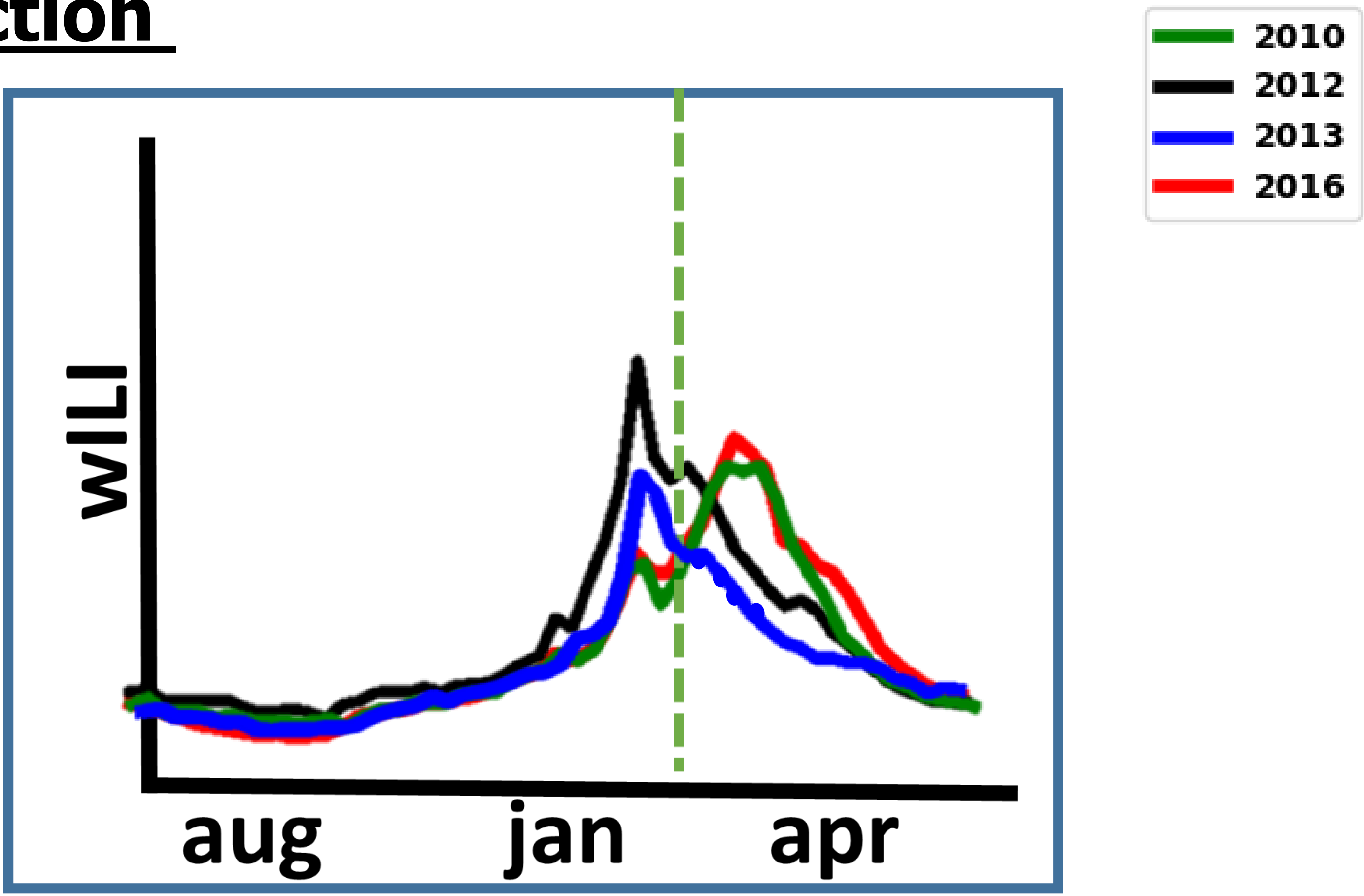
Rodríguez, Kamarthi, and Prakash 2021

Modeling ideas

1. Model temporal dynamics via similarity
    - Overcome data sparsity
    - Enable interpretability
2. Transfer knowledge representations
    - Learn from other relevant domains
3. Incorporate spatial structure
    - Model the spread over adjacent regions
    - Propagation over networks

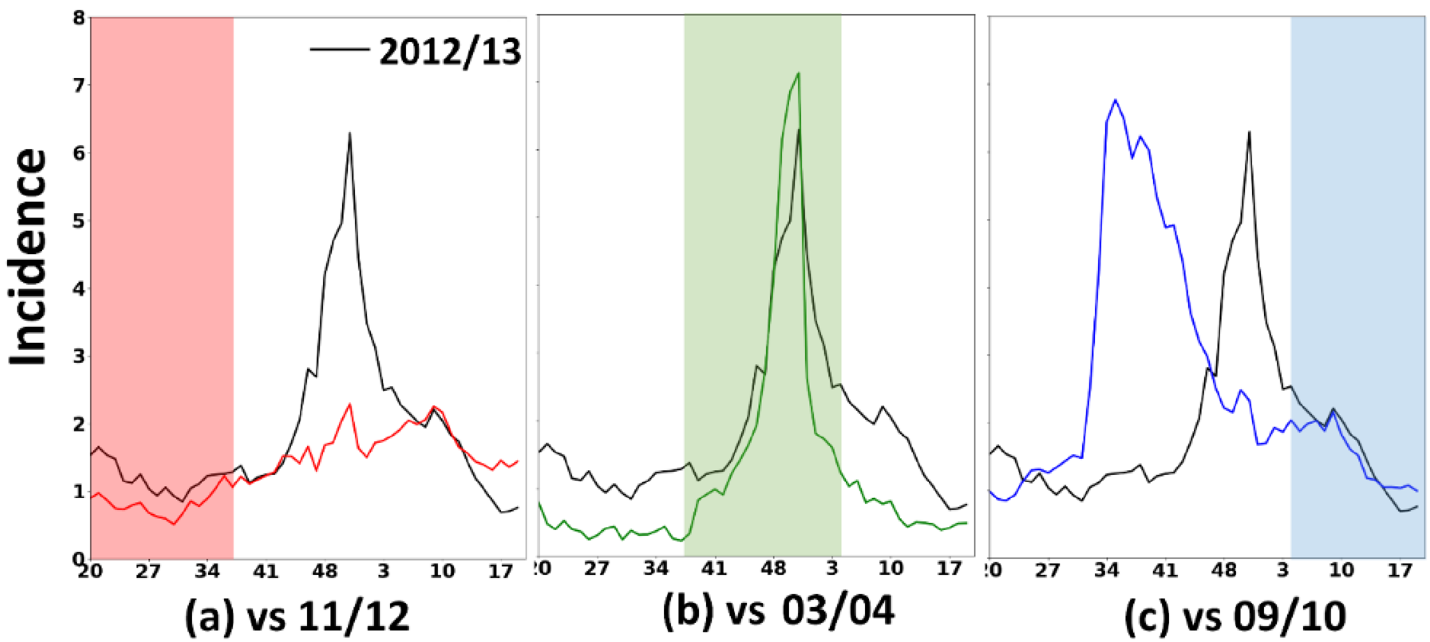
Modeling idea 1: Model temporal dynamics via similarity

- Idea: **clustering for prediction**



Model temporal dynamics via similarity CONTD.

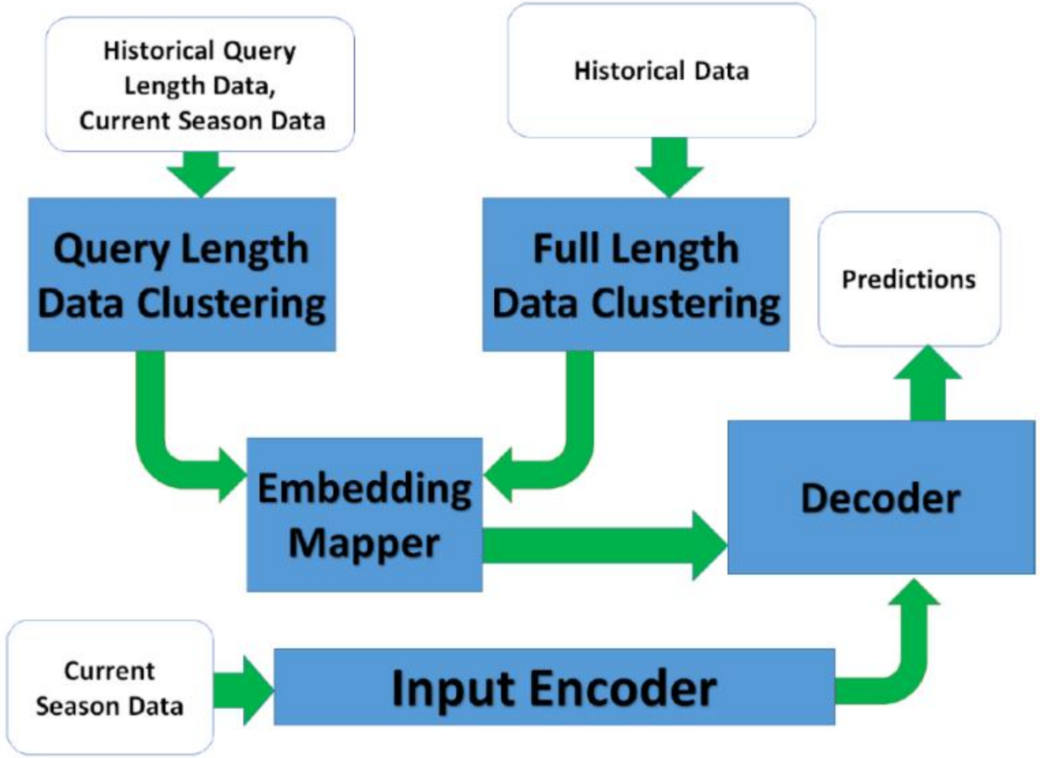
- Idea: **Dynamic clustering for prediction**



Rodríguez, Kamarthi, and Prakash 2021

Model temporal dynamics via similarity CONTD.

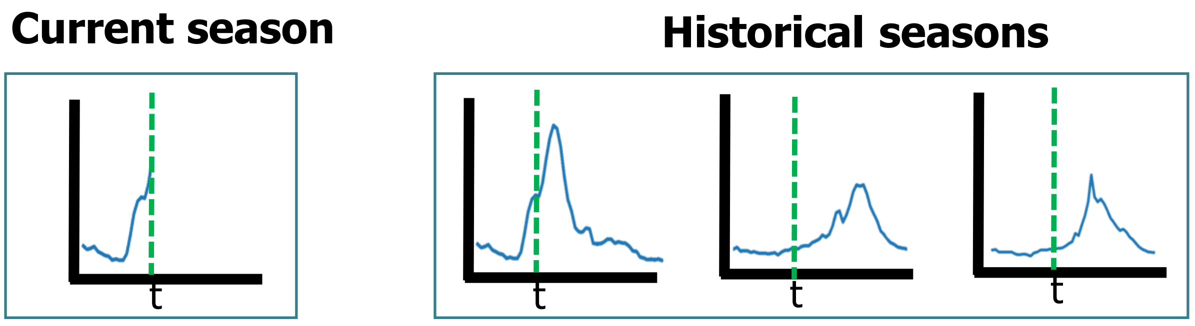
- Idea: **Dynamic deep clustering for prediction with limited data**



Rodríguez, Kamarthi, and Prakash 2021

Find similarity to historical seasons

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

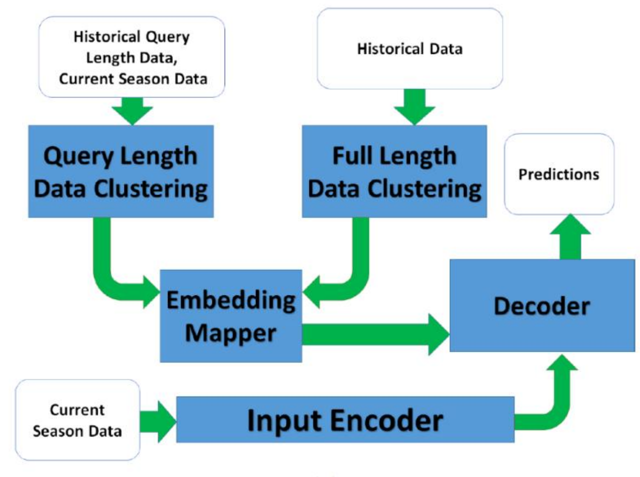


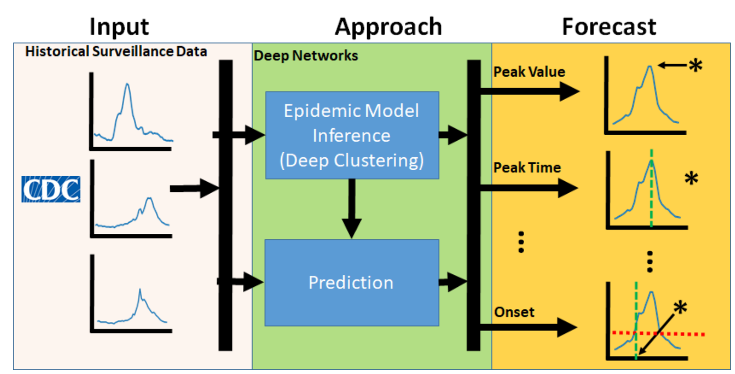
- Use snippets of historical seasons till week t to learn embedding

Data-driven approach: EpiDeep

[Adhikari+, KDD'19]

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





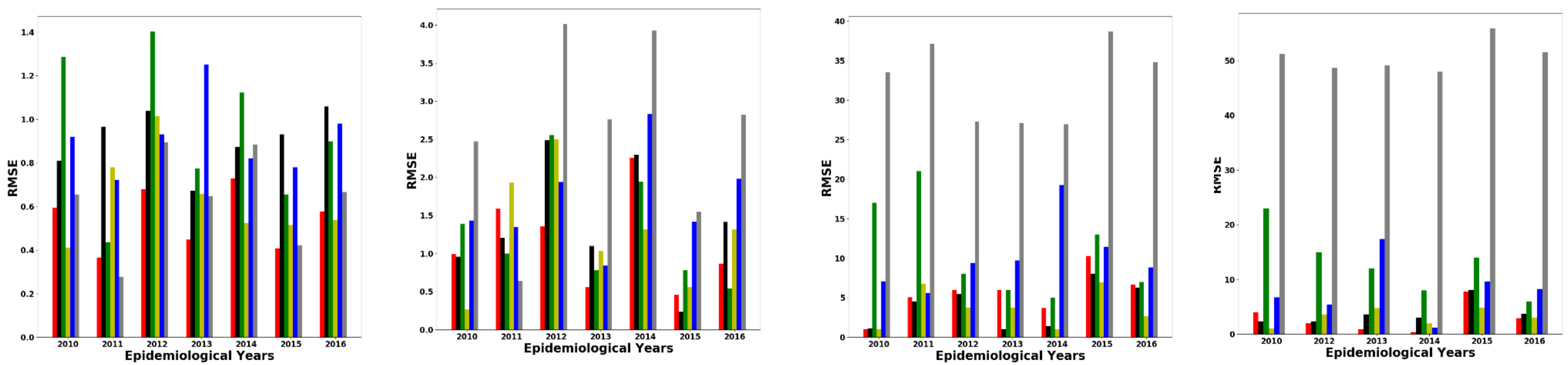
Experiments: Baselines

- **EB**: an empirical Bayesian approach. [Brooks+, PLOS ComBio 2015]
  - published and publicly available version
- **ARIMA**: an auto-regressive method for making predictions on time-series data.
- **HIST**: historical average of all previous seasons.
- **KNN**: selects the top k closest historical seasons to the current season, and make predictions on their average. [Nsoesie+, Stats Com in Infectious Dieases 2011]
- **LSTM**: a version of [Venna+, IEEE Access 2017] without climate and geographical data.

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

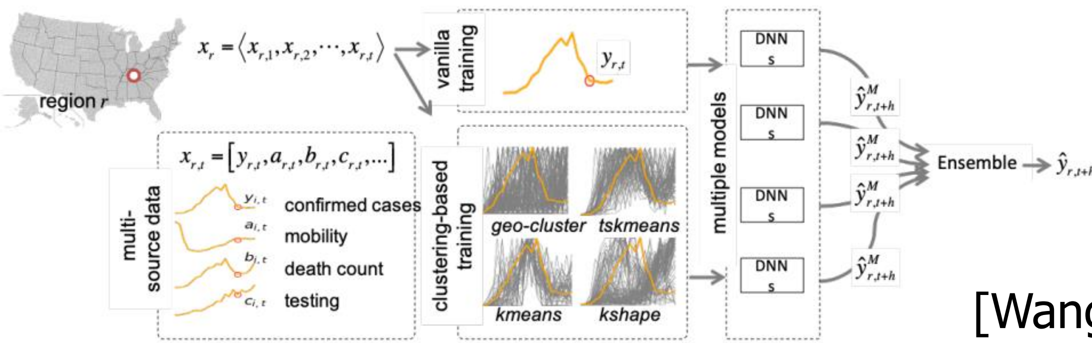
Lower is better

EpiDeep outperforms baselines in most settings.

Georgia Tech

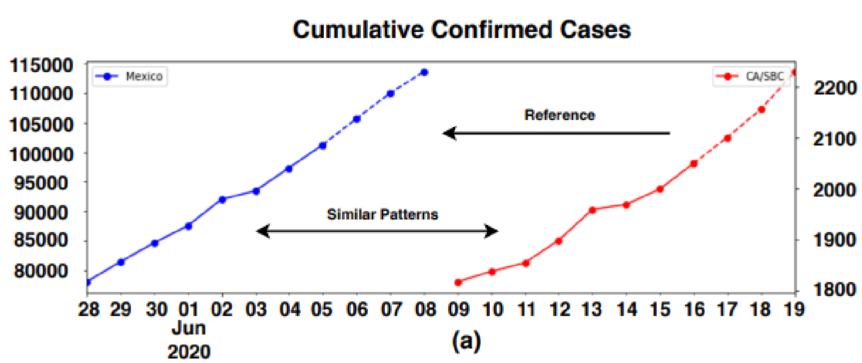
Other examples of modeling temporal similarity

- Temporal and geo. similarity (adjacent regions)



[Wang et al., BigData 2020]

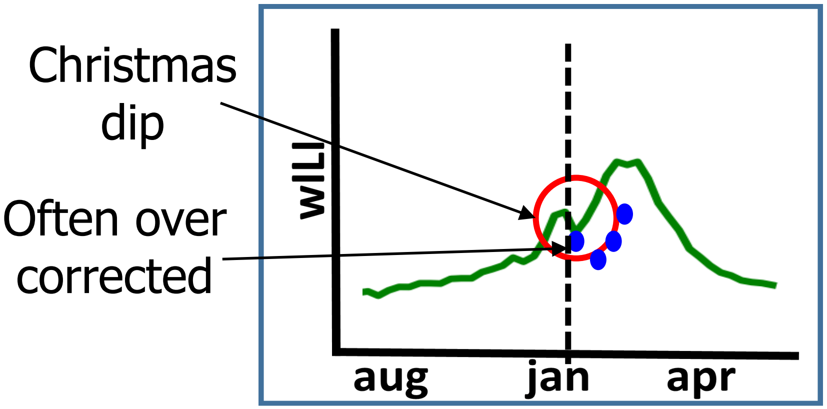
- Inter-series similarity



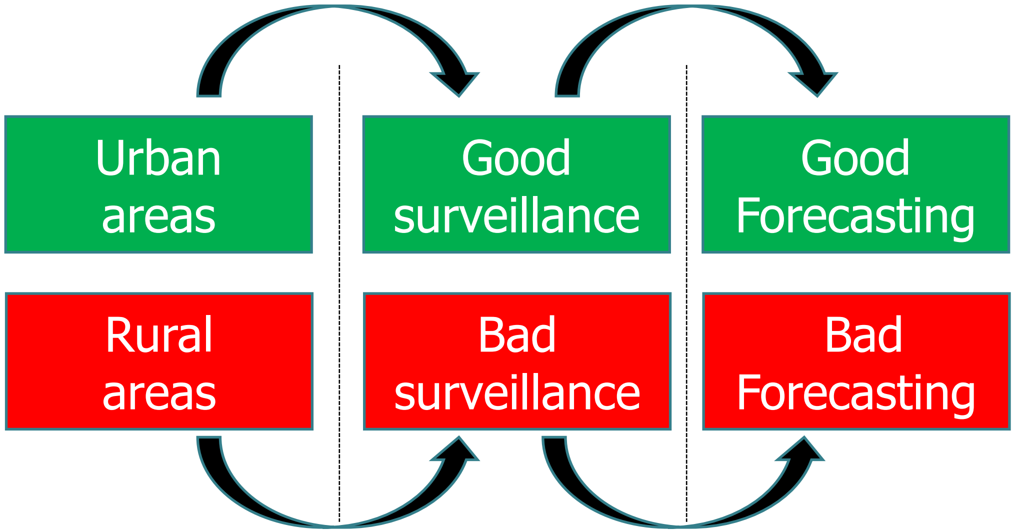
[Jin et al., SDM 2021]

Detour: Incorporating guidance in Epidemic Forecasting

- Epidemiological experts may notice unideal behavior exhibited by statistical approaches



Smoothness



Regional Equity

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

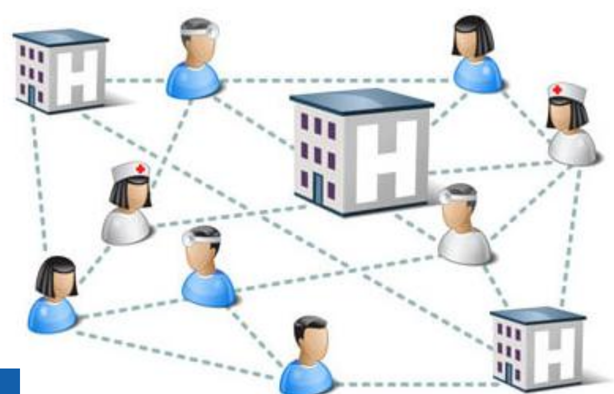
[Rodríguez+, epiDAMIK @ KDD 2020]

Modeling idea 2: Transfer knowledge representations

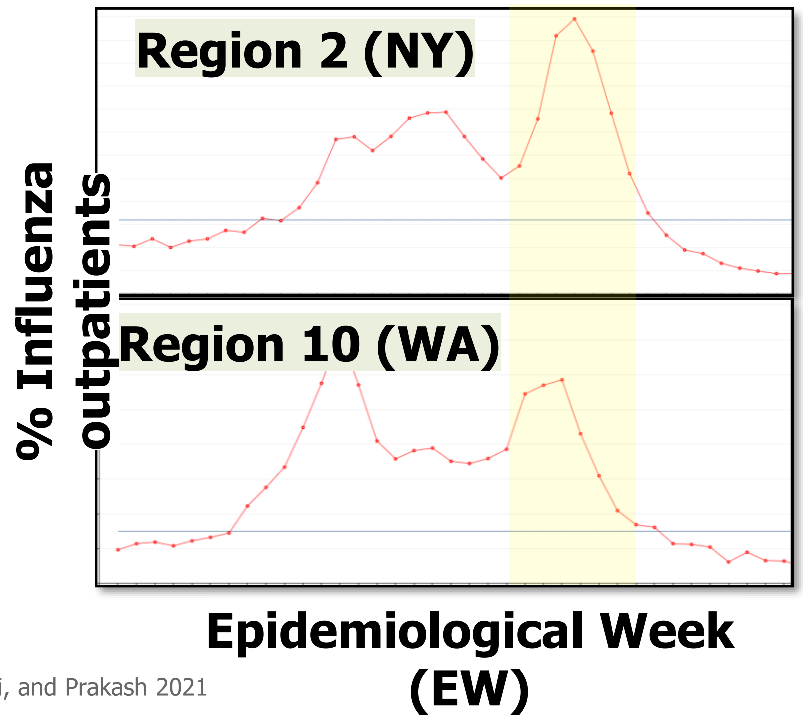
- Neural model automatically learn what to transfer
  - Not everything is relevant! Needs selection
- Examples:
  - From one country to another country
    - Even in different continents
    - In Panagopoulos et al., AAAI 2020
  - From a historical scenario to a novel scenario
    - From pre-COVID flu to COVID-contaminated flu counts
    - In Rodríguez et al., AAAI 2020

Influenza Surveillance in the Early COVID Pandemic

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

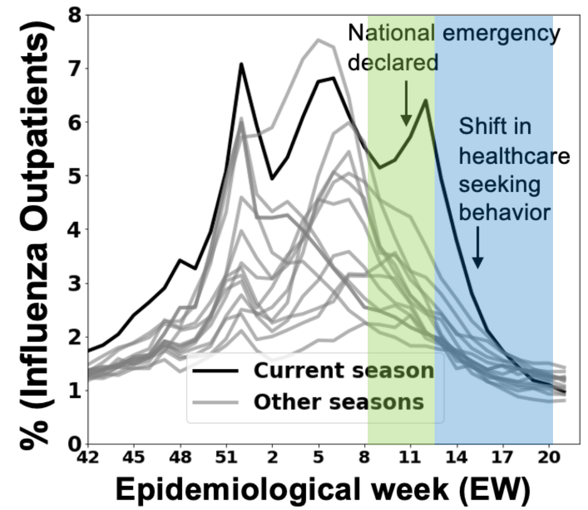


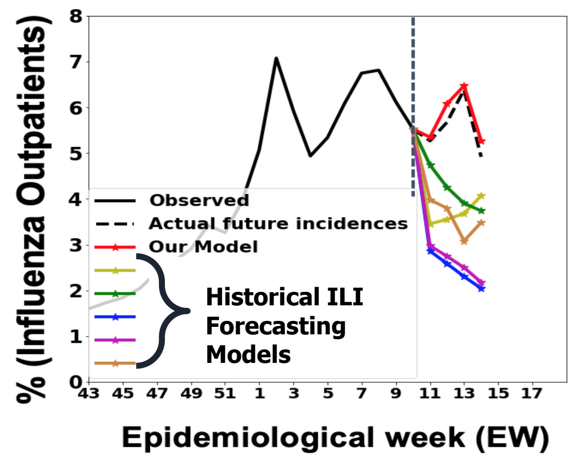
CDC ILINet surveillance network



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





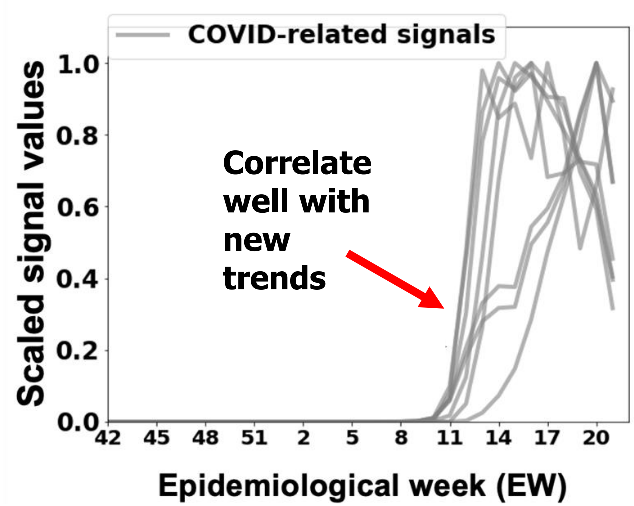
New COVID-related signals correlate with new trends

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

CDC kinsa The COVID Tracking Project

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING
Center for Systems Science and Engineering

Google



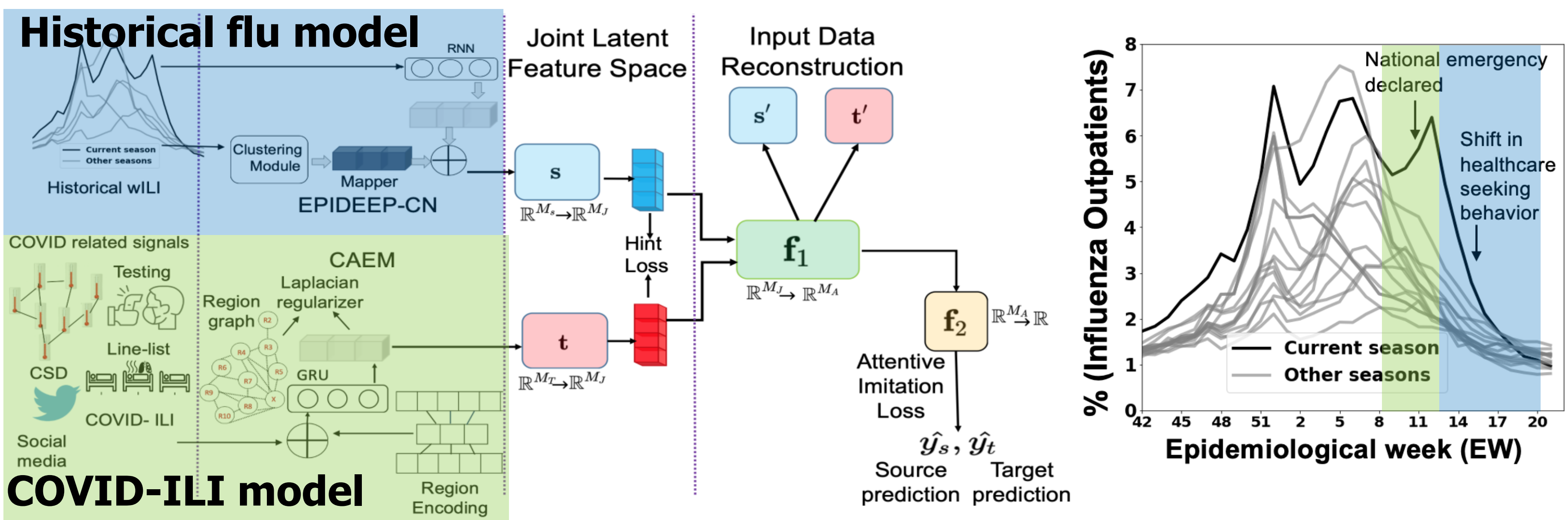
Georgia Tech

Rodríguez, Kamarthi, and Prakash 2021

Attentive transfer learning for heterogeneous domains

[Rodríguez+, AAAI 2021]

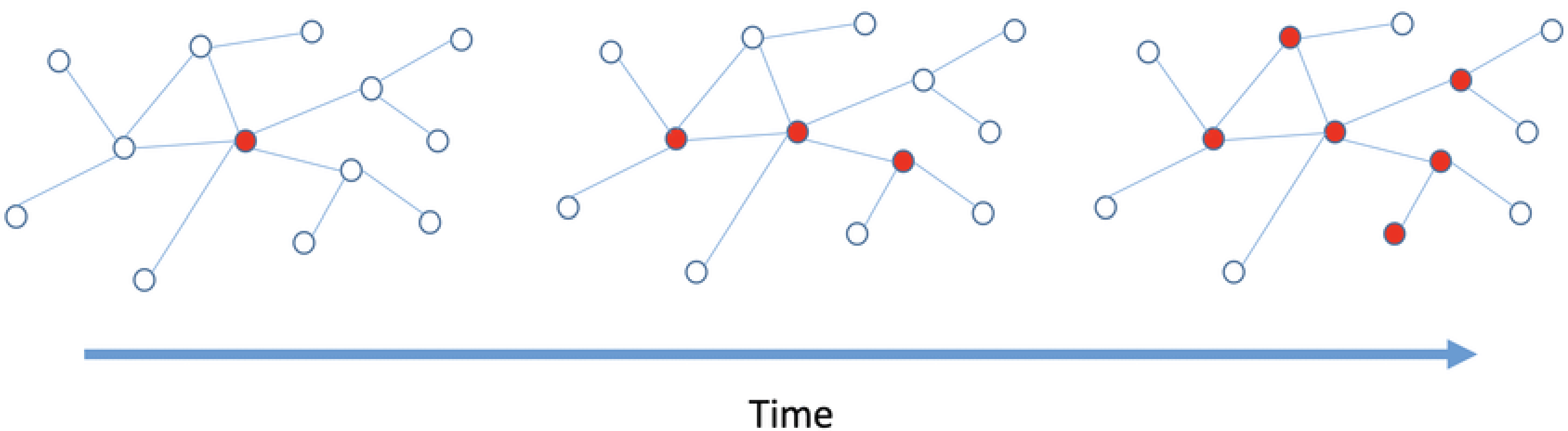
- CALI-Net: steer a historical flu model (EpiDeep, KDD 2019) with new COVID-related signals



Rodríguez, Kamarthi, and Prakash 2021

Modeling idea 3: Incorporate spatial structure

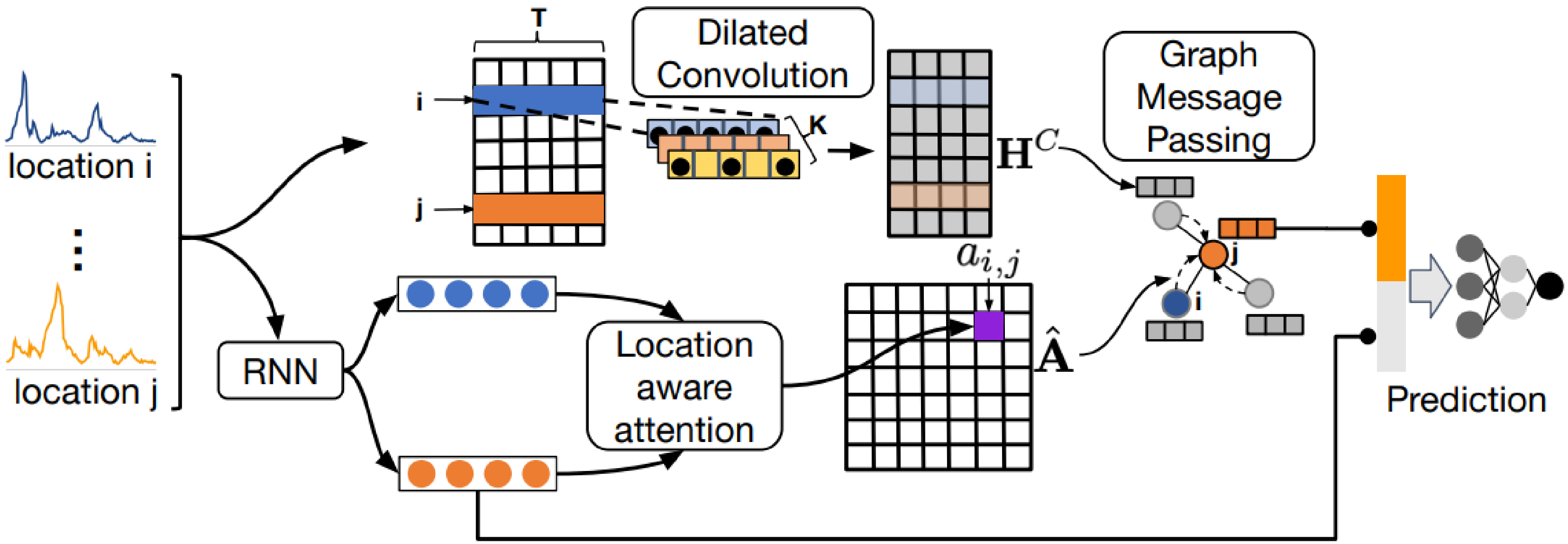
- Pathogens propagate to adjacent regions
  - And then to new adjacent regions
- Propagation over spatial graphs



Graph message passing for spatial propagation

[Deng+, CIKM 2020]

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



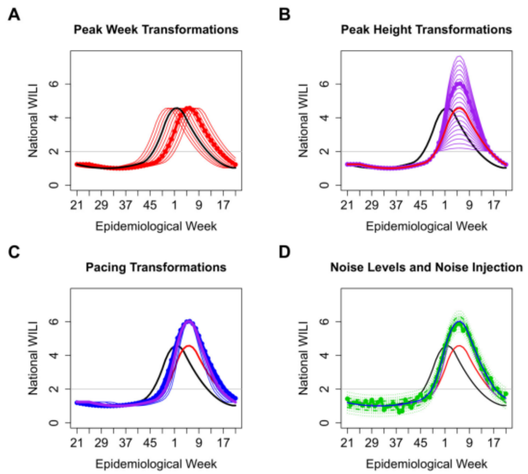
[4] Density Estimation Models

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

Empirical Bayes

[Brooks+ 2015 PLoS]

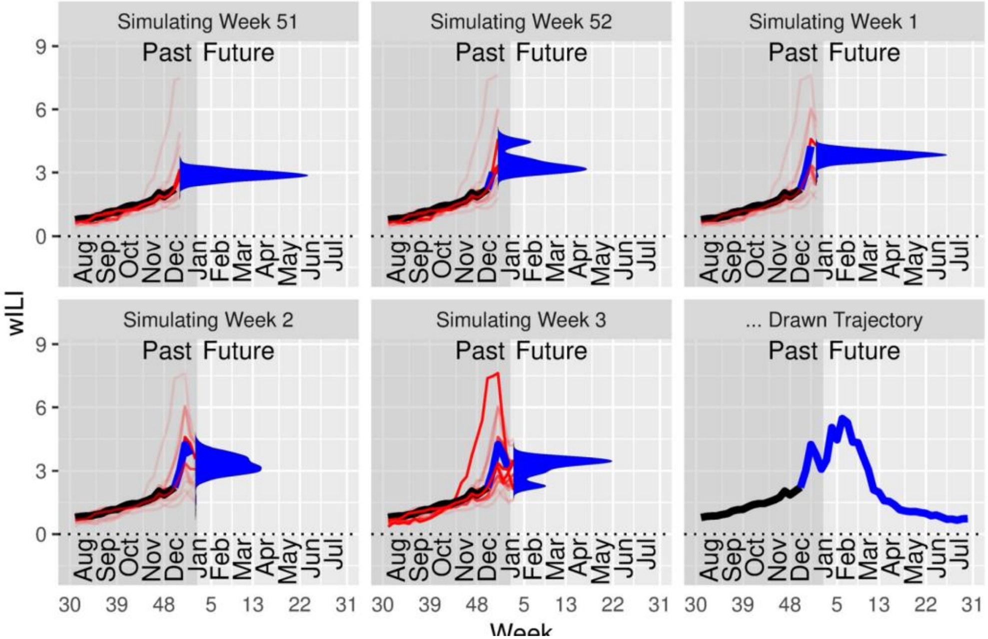
- Idea: Current season's epidemic curve is a probabilistic distribution of features
- Model parameters:
  - Similarity is shape to past sequences
  - Peak height, week
  - Scaling factor of the curve
- All modelled as priors of forecast distribution
- Use Bayesian Inference to calibrate for current season



Delta Density

[Brooks+ 2017 PLoS]

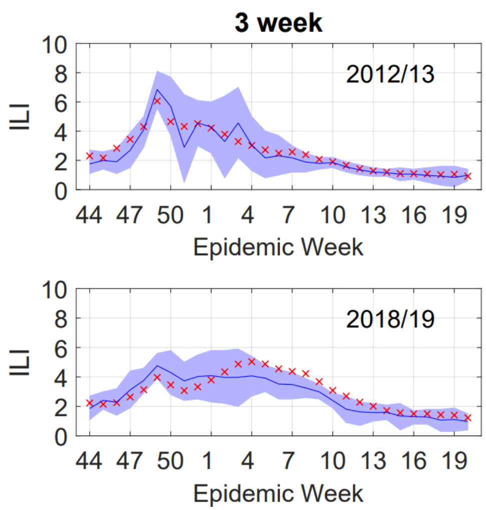
- Use kernel density estimation to leverage similarity with historical seasons
- One of the top models in Flusight 2017 challenge



Gaussian Process

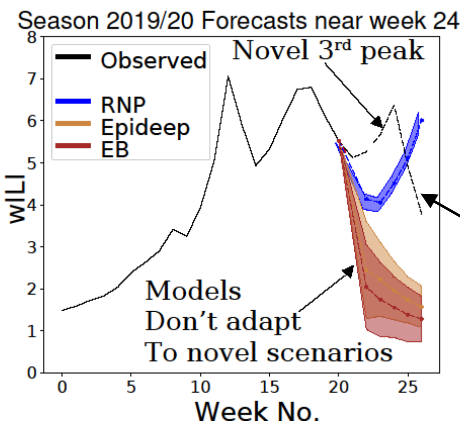
[Zimmer+ 2019 ICML]

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

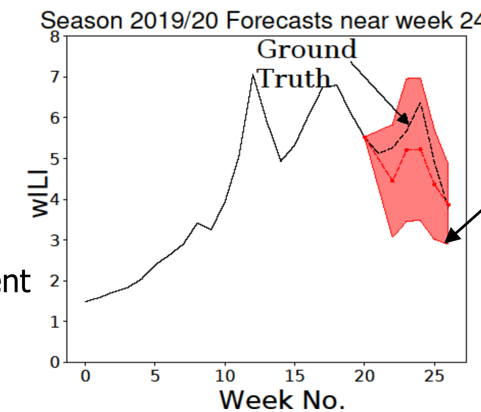


Neural models for calibrated forecasts

- Density Estimation models don't focus on well-calibrated forecasts
  - Can't adapt to provide reliable forecast uncertainty on novel patterns



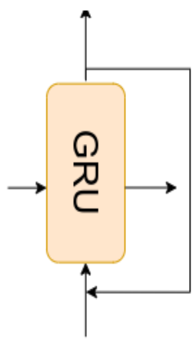
Wrong Overconfident prediction



Mean is close to ground truth

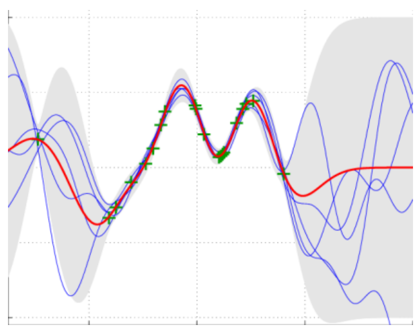
EpiFNP: Neural non-parametric model for better calibration

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



Deep Sequential Models

+



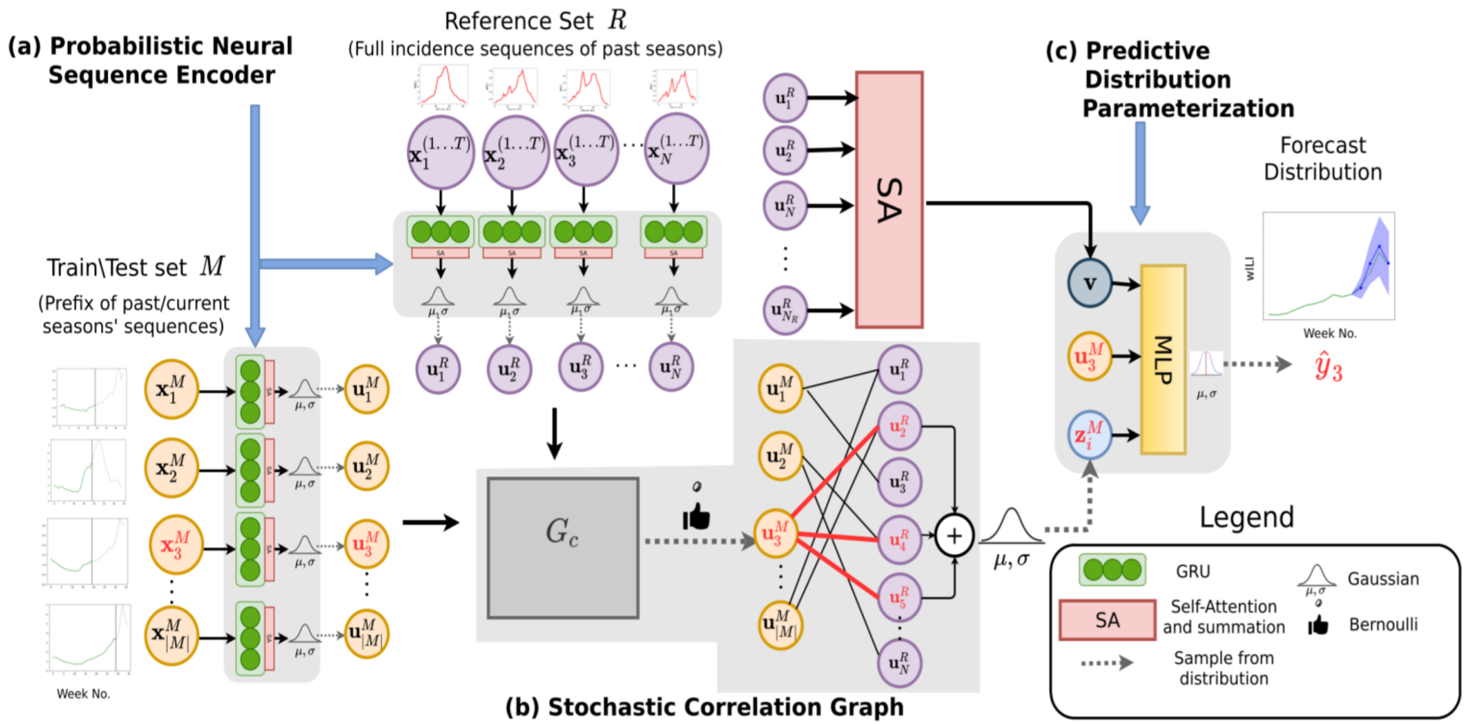
=

Accurate and Well-calibrated neural forecasting model

EpiFNP: Architecture

[Kamarthi+, NeurIPS 2021]

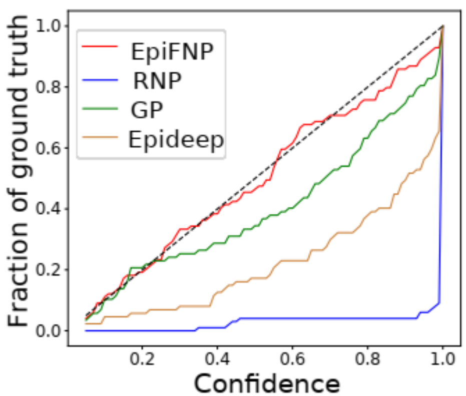
Sequential representations +
neural Gaussian processes



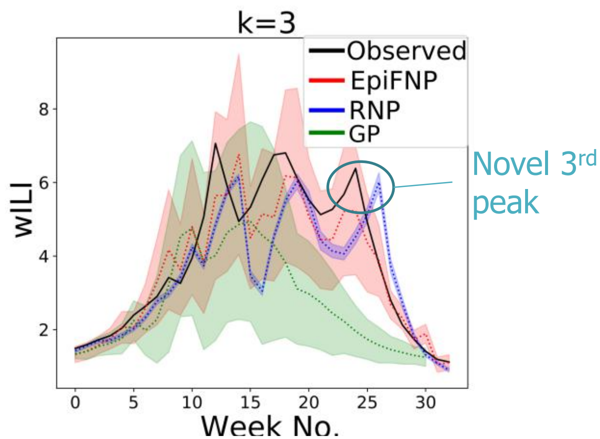
Rodríguez, Kamarthi, and Prakash 2021

Georgia Tech

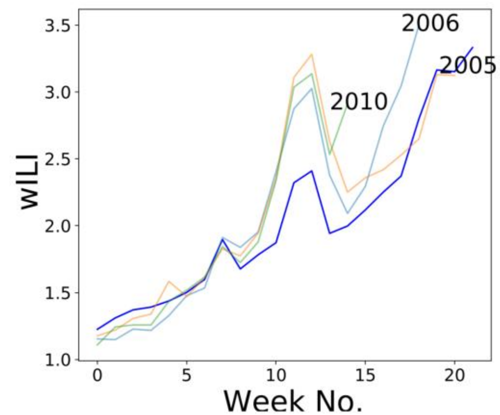
Results



Well calibrated predictions



Adapt to novel patterns



Explaining predictions

Most similar seasons chosen by EpiFNP

Pros/Cons Statistical Models

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

Part 4: Hybrid Models

Hybrid Models

- Use both mechanistic and statistical components as complementary pieces.

- Modeling approaches:
  1. Discrepancy modeling
  2. Parameter estimation

[1] Discrepancy modeling

- Statistical model resolves the discrepancies between a model (often mechanistic) and ground truth data.
- In other words, statistical model **refines/corrects** another model.

Hierarchical Bayesian Model for Mechanistic Discrepancy

[Osthus et al. 2019, Bay. Analysis]

- DBM refines mechanistic predictions with a hierarchical Bayesian model.
- Refinement components:
  - State-specific deviation
  - Season-specific deviation
  - Trends

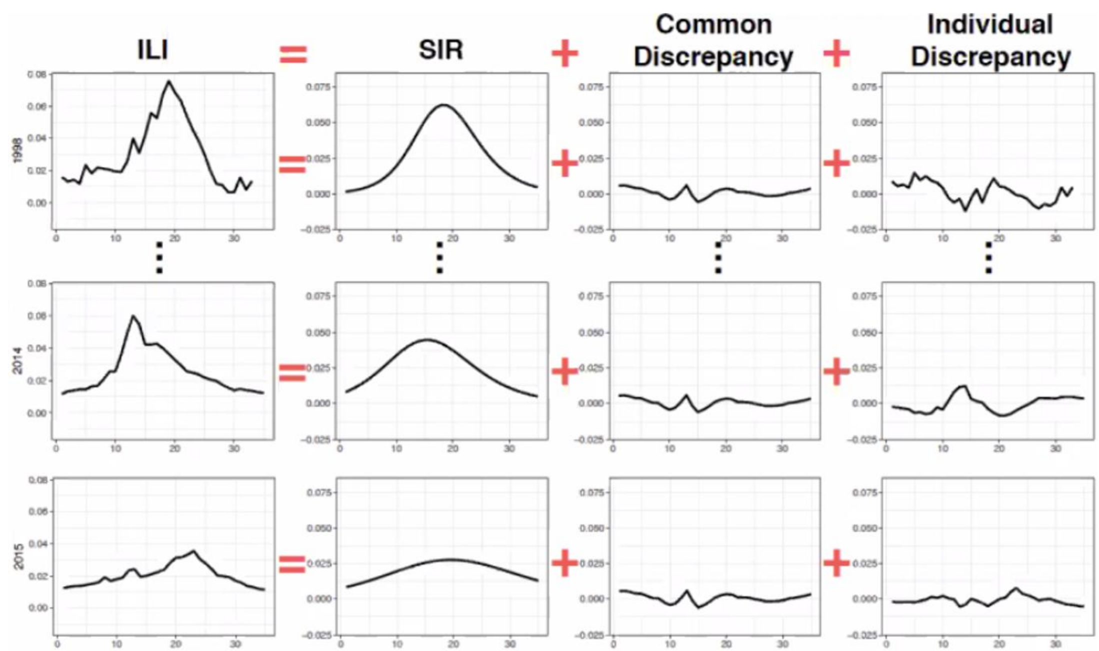
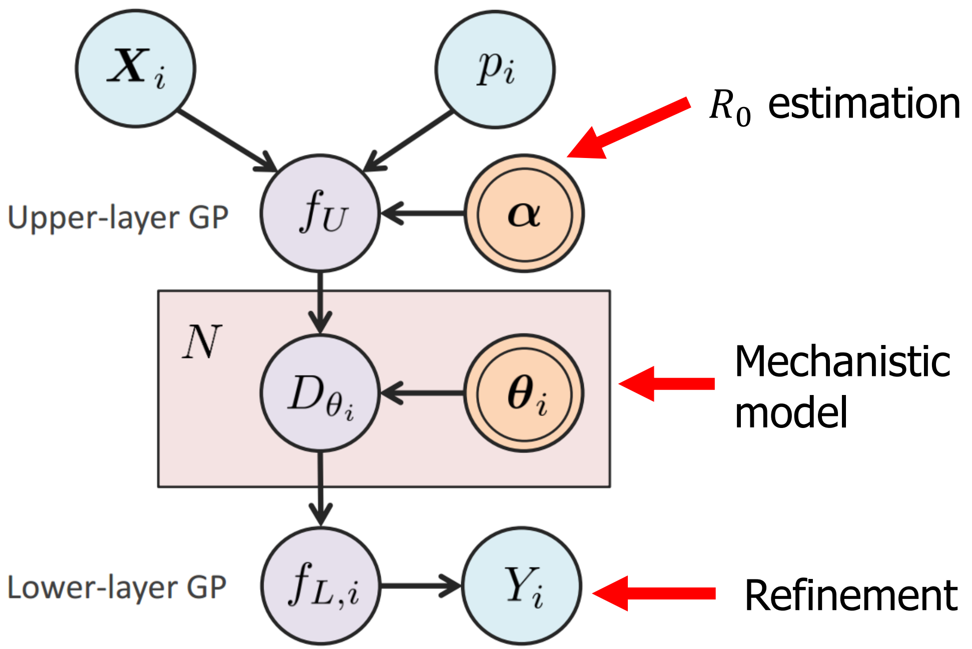


Figure credit: Sara Del Valle, LANL

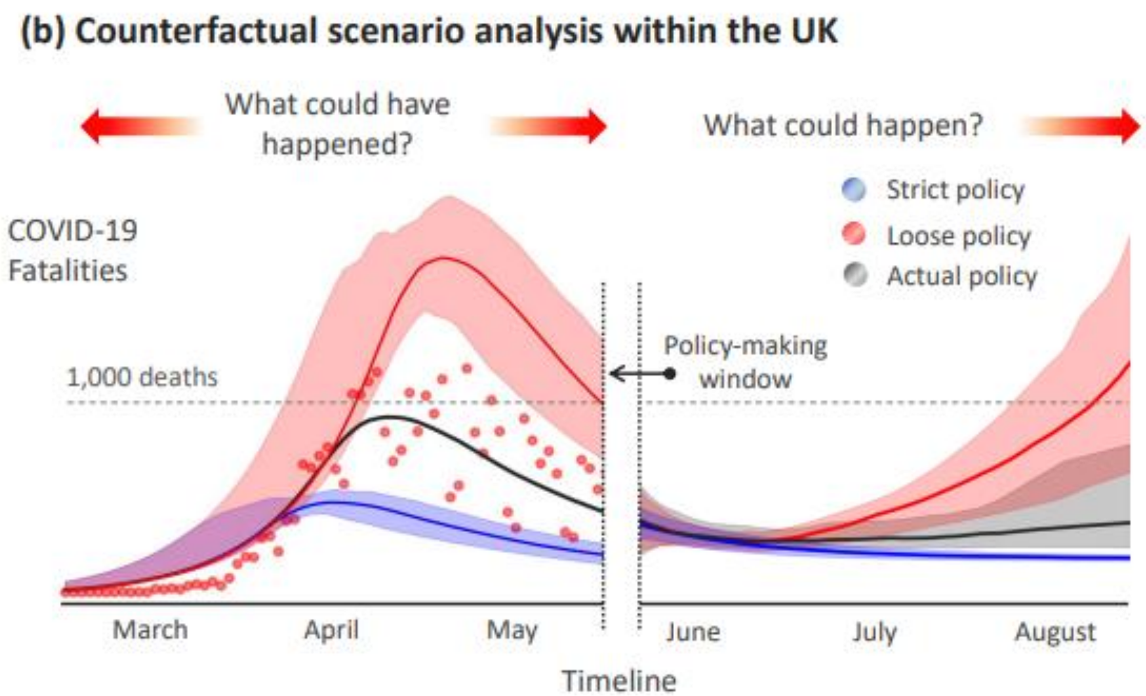
[2] Parameter estimation

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

[Qian+ NeurIPS 2020]



Counterfactual based on new set of policies



Rodríguez, Kamarthi, and Prakash 2021

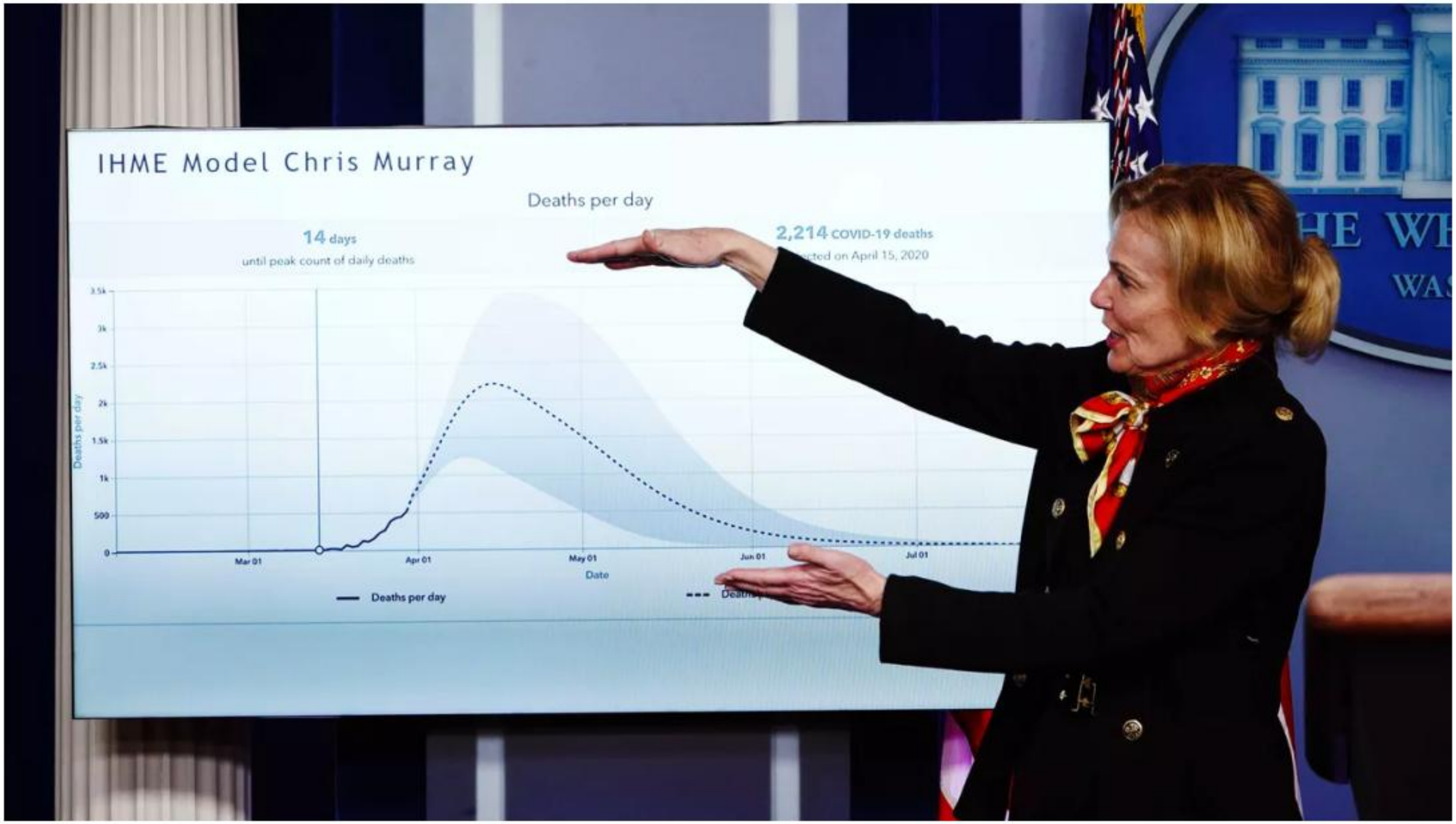
Part 5: Ensembles

Ensembles

- Combining models into an "ensemble" often provides more robust forecasts than any single model
- Consistently found across multiple epidemic forecasting efforts
  - Flu: Reich et al. 2019, PLOS Comp Bio
  - Dengue: Johansson et al. 2019, PNAS
  - Ebola: Viboud et al. 2018, Epidemics

Policy makers needed >1 model

Early April 2020



Slide credit: Nicholas Reich, UMass Amherst

Diversity of COVID-19 models

- IHME-CurveFit: "**hybrid modeling approach** to generate our forecasts, which incorporates elements of statistical and disease transmission models."
- MOBS-GLEAM_COVID: "The GLEAM framework is based on **a metapopulation approach** in which the world is divided into geographical subpopulations. Human **mobility between subpopulations is represented on a network**."
- UMass-MechBayes: "**classical compartmental models from epidemiology**, prior distributions on parameters, models for time-varying dynamics, models for partial/noisy observations of confirmed cases and deaths."
- UT-Mobility: "For each US state, **we use local data from mobile-phone GPS traces** made available by [SafeGraph] to quantify the changing impact of social-distancing measures on 'flattening the curve.' "
- GT-DeepCOVID: "This **data-driven deep learning model** learns the dependence of hospitalization and mortality rate on various detailed syndromic, demographic, mobility and clinical data."
- Google Cloud AI: "a novel approach that integrates **machine learning** into **compartmental disease modeling** to predict the progression of COVID-19"
- Facebook AI: "**recurrent neural networks** with a vector autoregressive model and train the joint model with a specific regularization scheme that increases the **coupling between regions**"
- CMU-TimeSeries: "A **basic AR-type time series model** fit using lagged values of case counts and deaths as features. No assumptions are made regarding reopening or governmental interventions."

Slide credit: Nicholas Reich, UMass Amherst

What is the optimal ensemble?

		"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

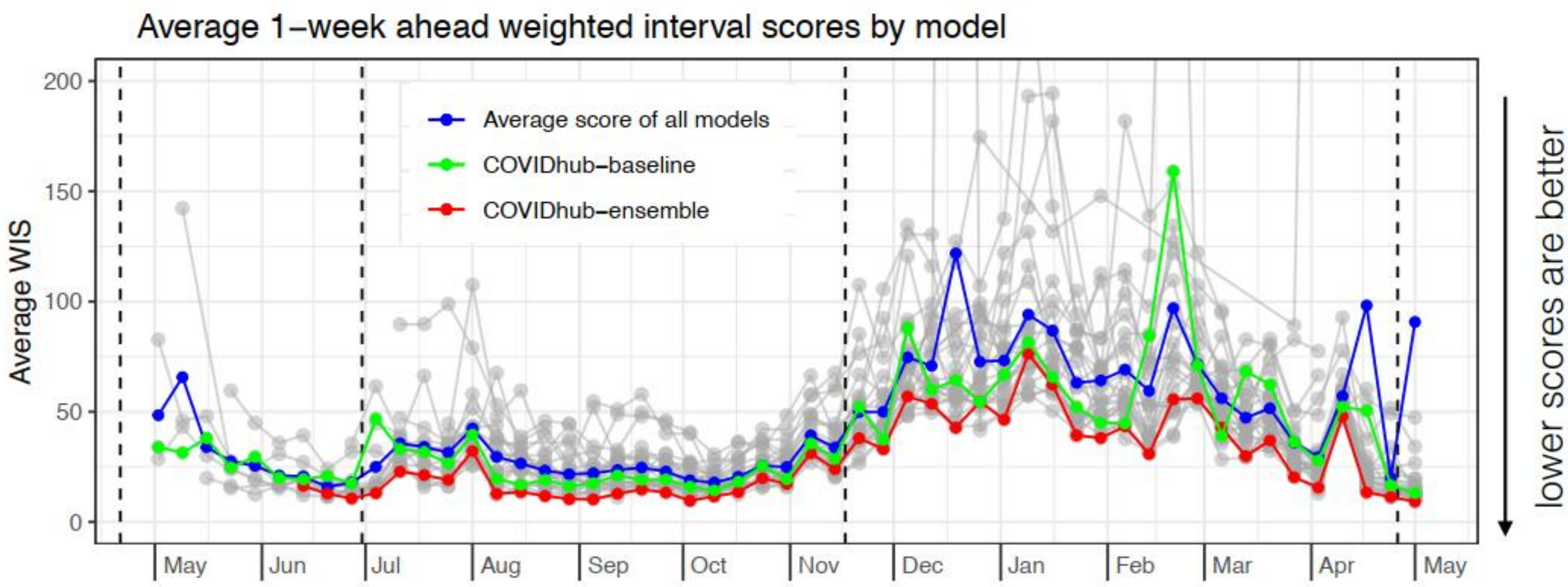
- Median of best 5 or 10 individual models
- Weighted median, weights from a weighted mean ensemble
- Weighted median, weights based on relative WIS

- Takeaway: use a robustly trained ensemble

Slide credit: Nicholas Reich, UMass Amherst

Results in COVID-19

[Craemer+, medRxiv 2021]

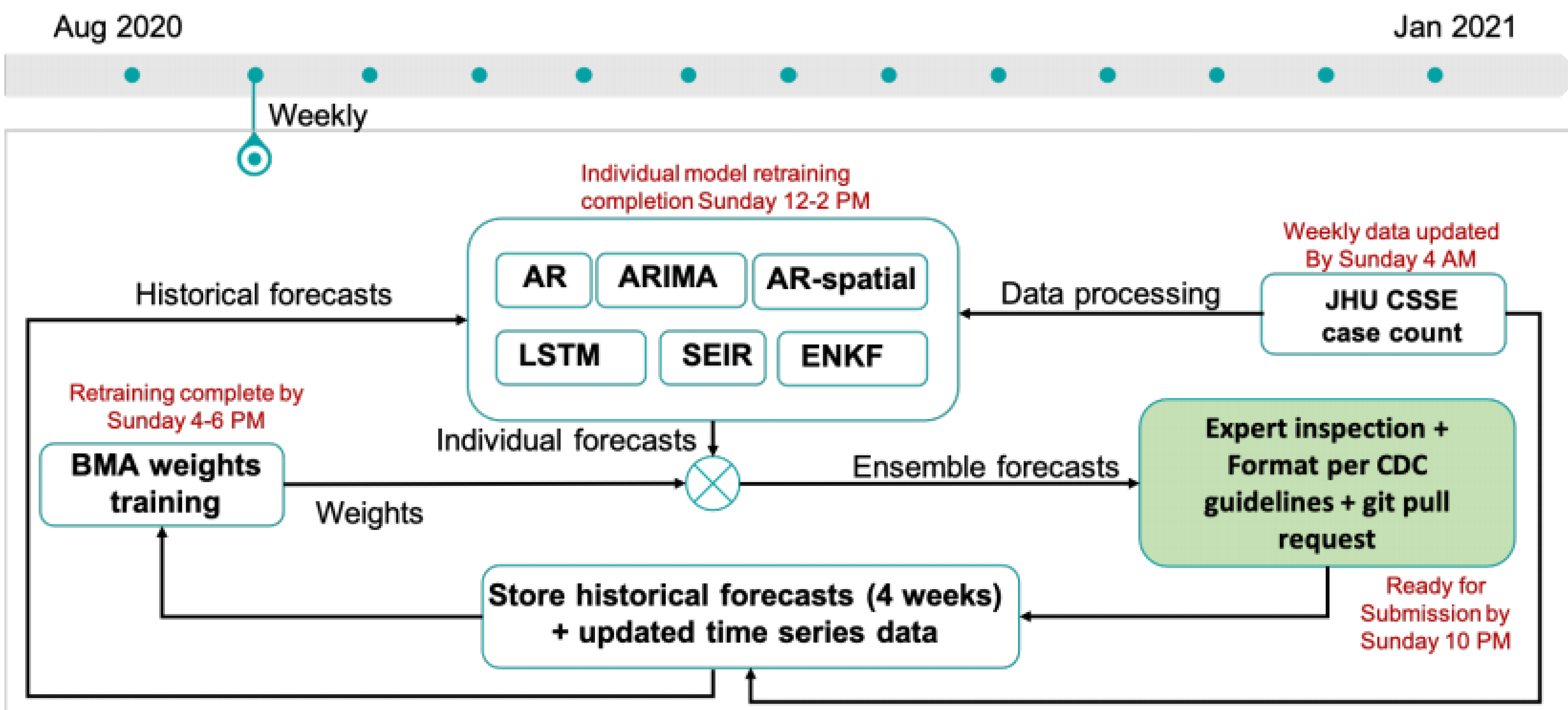


All models are useful

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

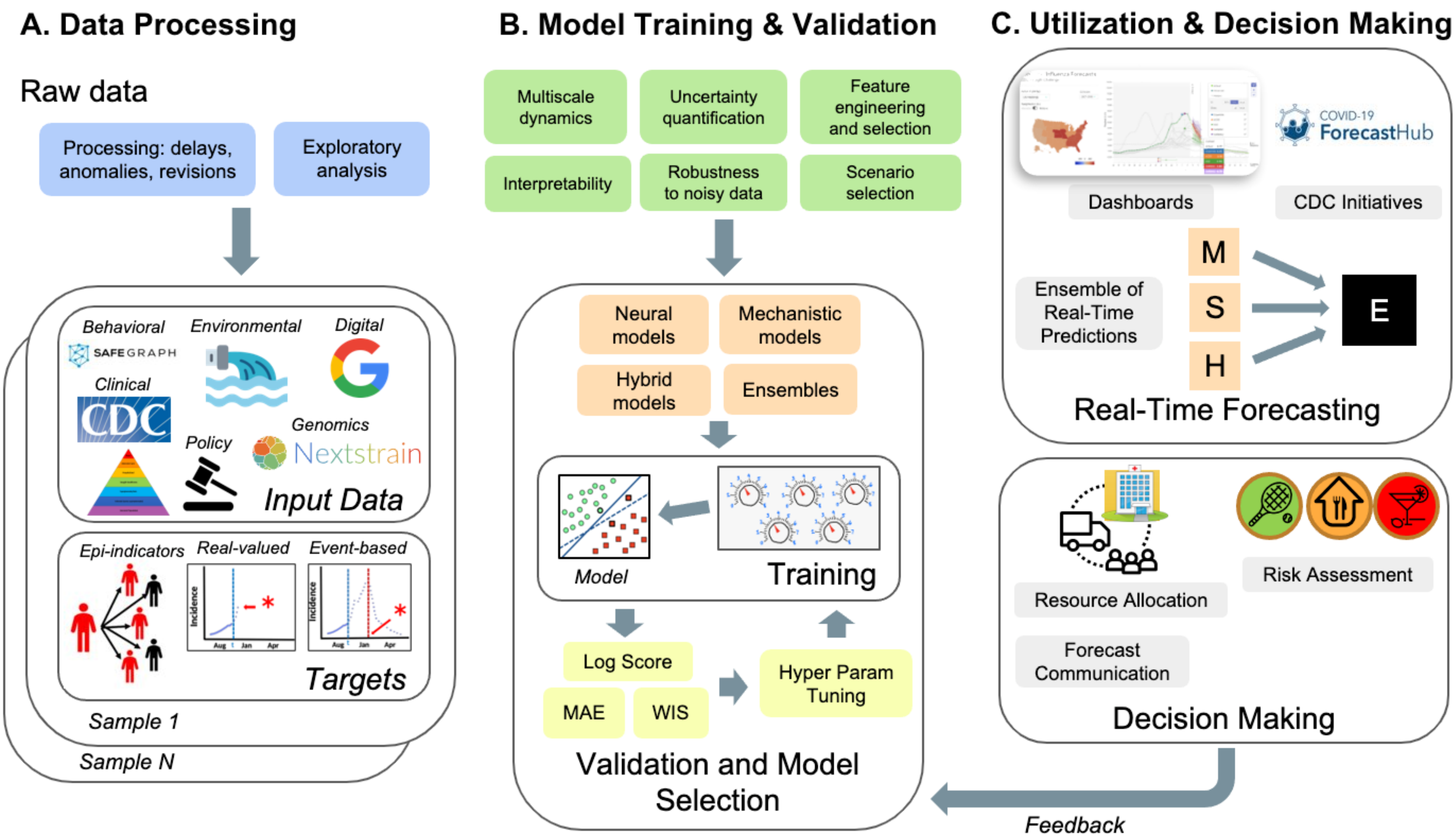
Super-ensembles

[Adiga+, medRxiv 2021]



Part 6: Epidemic Forecasting in Practice

Epidemic Forecasting Pipeline



Rodríguez, Kamarthi, and Prakash 2021

Forecasting in Practice

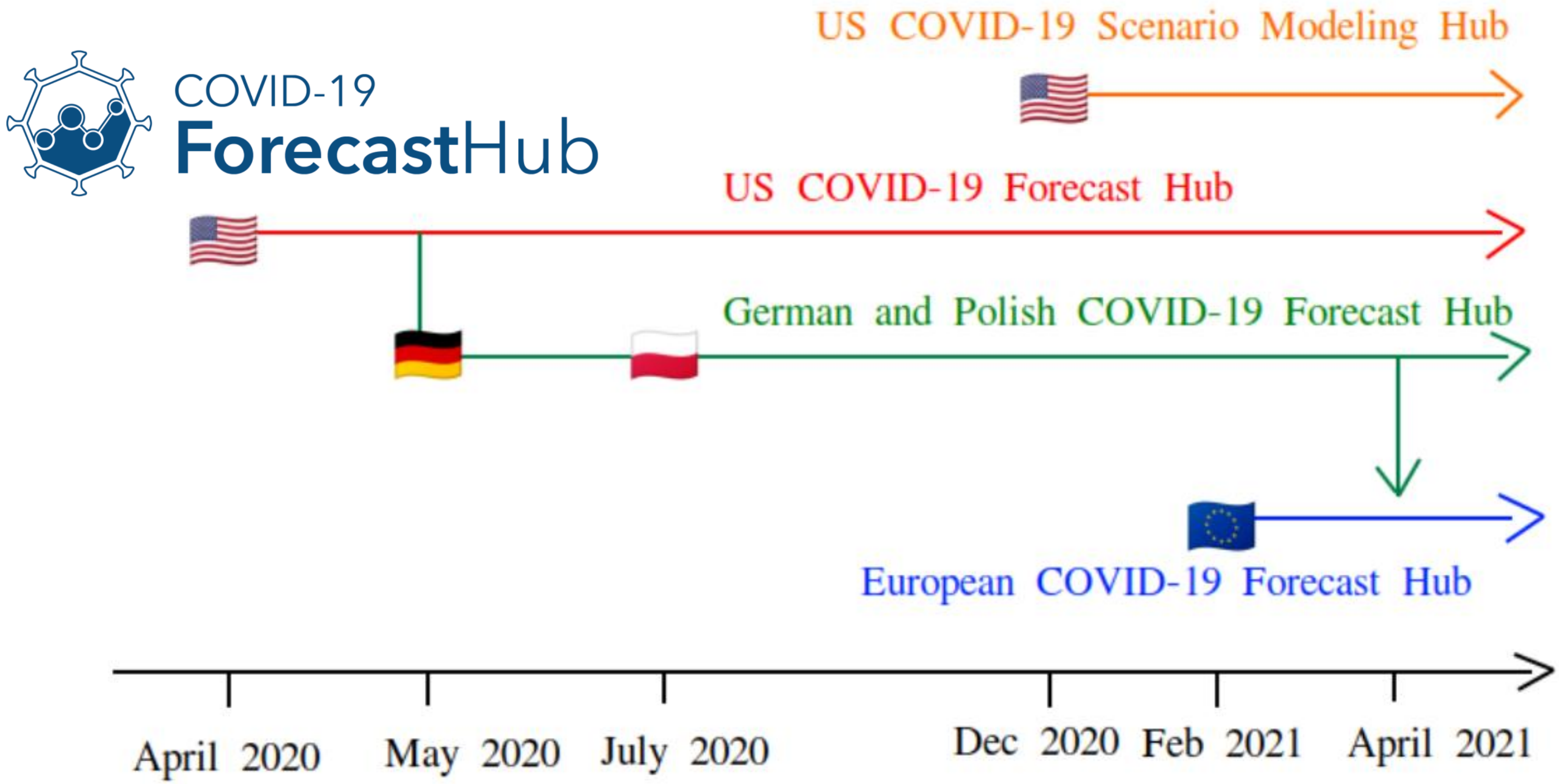
- Topics:
  1. US CDC initiatives
  2. Real time experiences
  3. Decision making

[1] Forecasting Initiatives

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

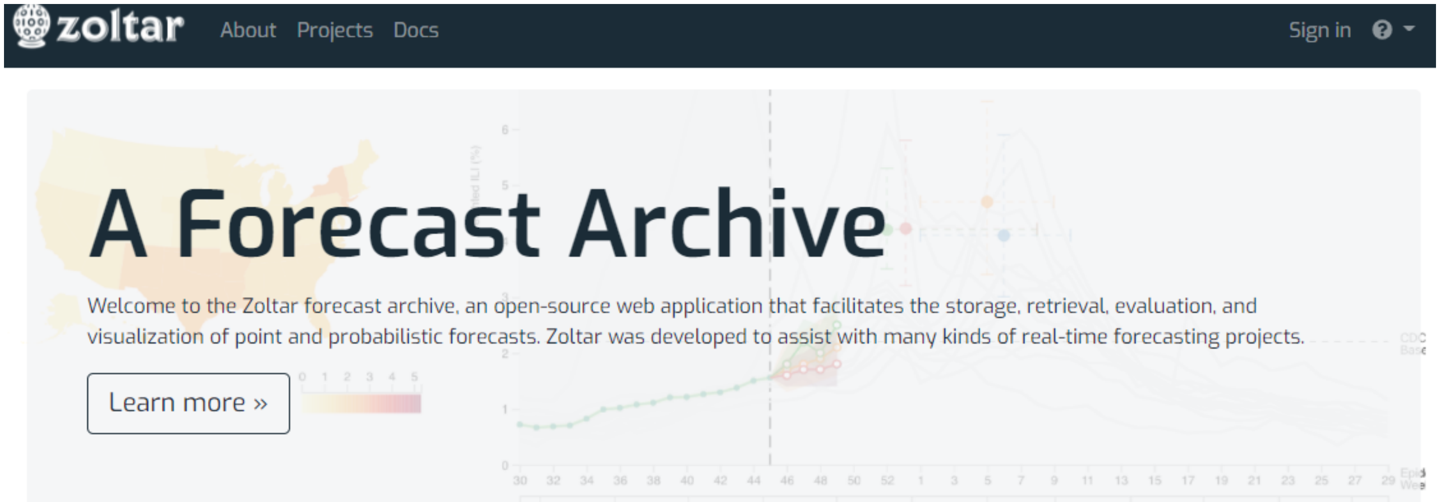
Slide credit: Matt Biggerstaff, US CDC

COVID-19 Forecast Hubs



Source: Johannes Bracher, KIT Karlsruhe and HITS Heidelberg

Standardization efforts of real-time forecast submissions



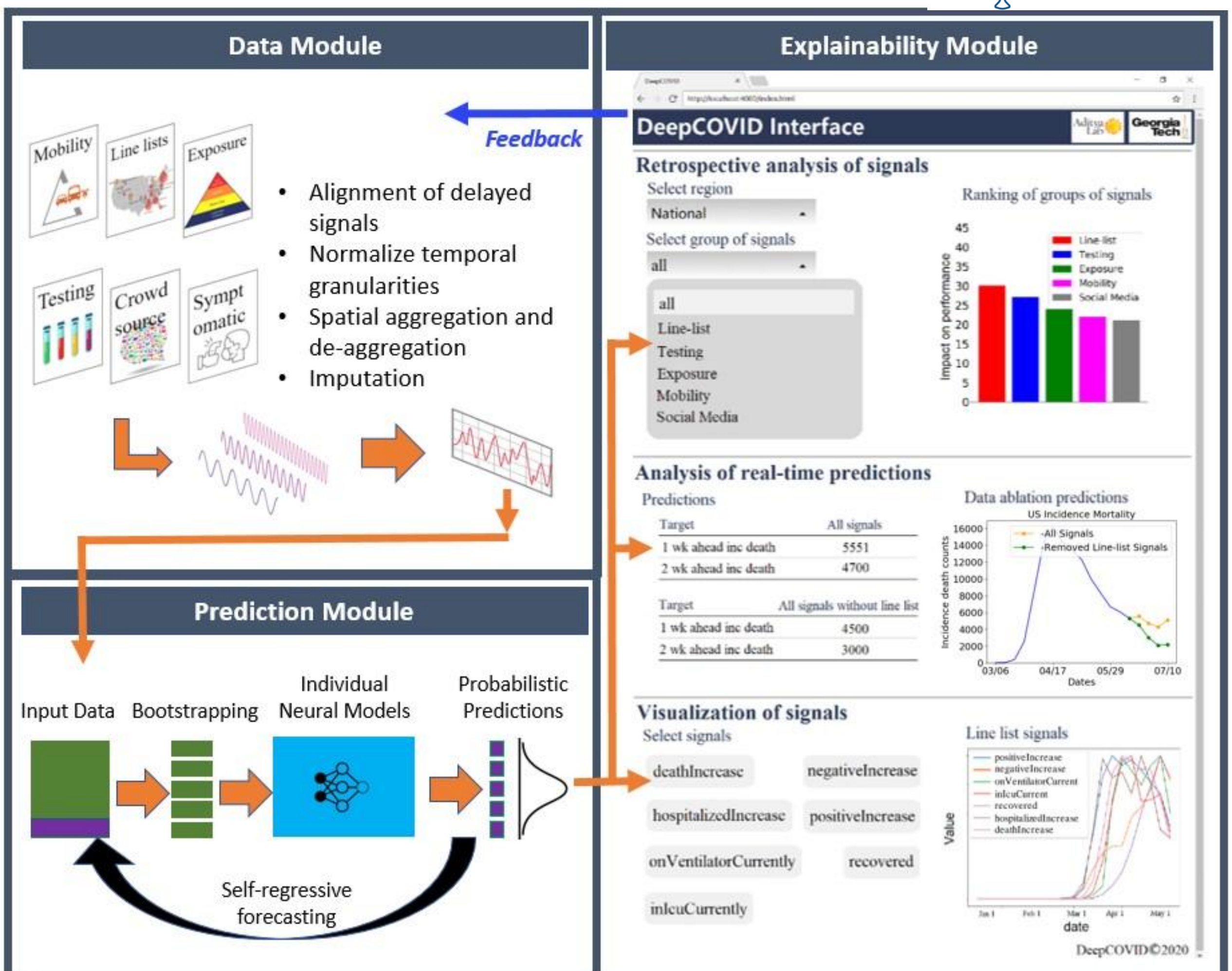
Project: COVID-19 Forecasts Config

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

COVID-19 ForecastHub

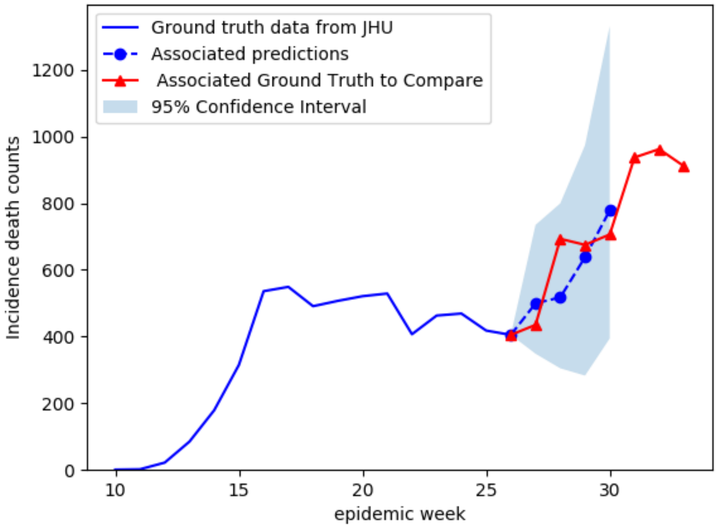
[2] Real-time Experience and Challenges

Operational Deep Learning Framework [Rodríguez+, IAAI 2021]

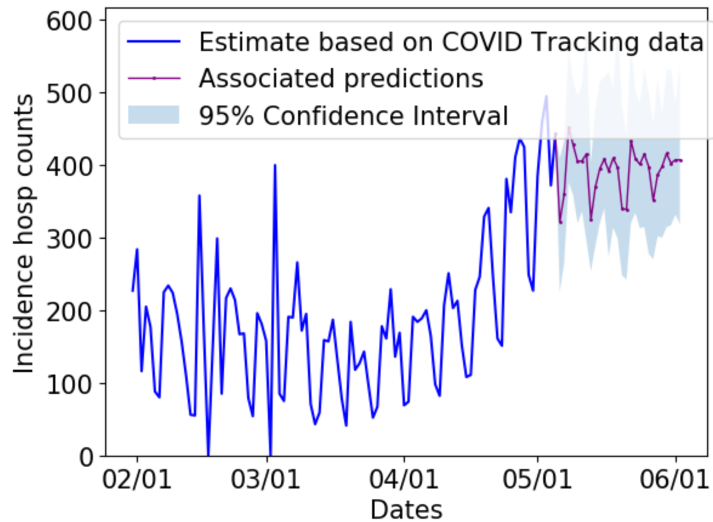


Highlights of results

Anticipate Trend Changes

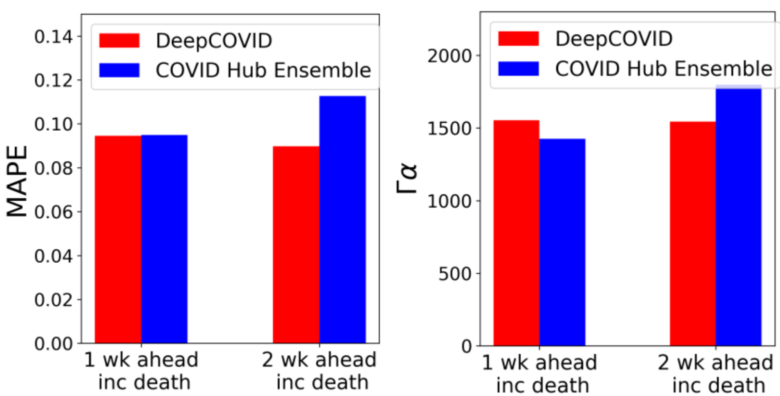


Capture finer-grain patterns

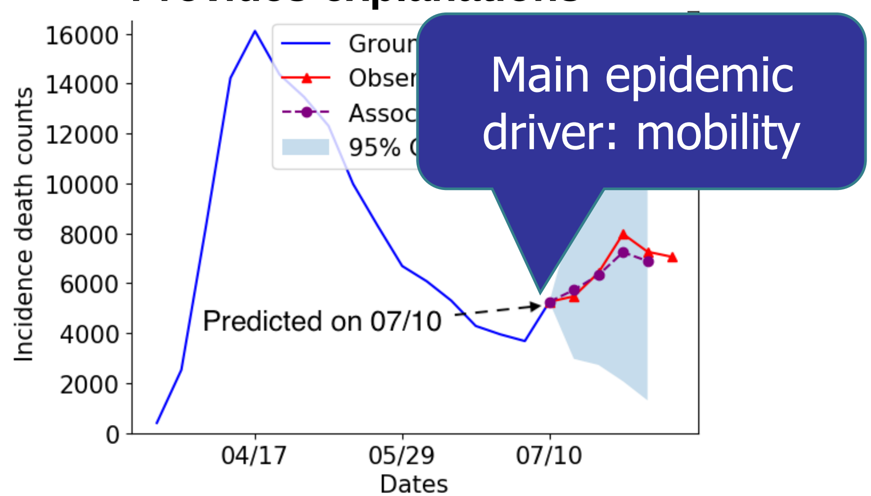


Lower is better

Excels in short-term forecasting



Provides explanations



Top Ranked Model

External evaluation

- Cramer et al. evaluated model predictions submitted to the CDC.
- Evaluation:
  - 1 to 4 week ahead
  - May 2020 - Oct 2021 (1+ year)
  - 51 locations (national + states)
- DeepCOVID ranked **top 5** out of 25 individual models.

Data Challenges: Don't Underestimate!

(C1) Multiple data sources and formats

- Format varies over time

(C2) Select signals with epidemiological significance

(C3) Temporal misalignment

- Delays, pause in reporting, differ in granularity

(C4) Spatial misalignment

- Differ in granularity: county vs state vs national

(C5) Data quality and missing data

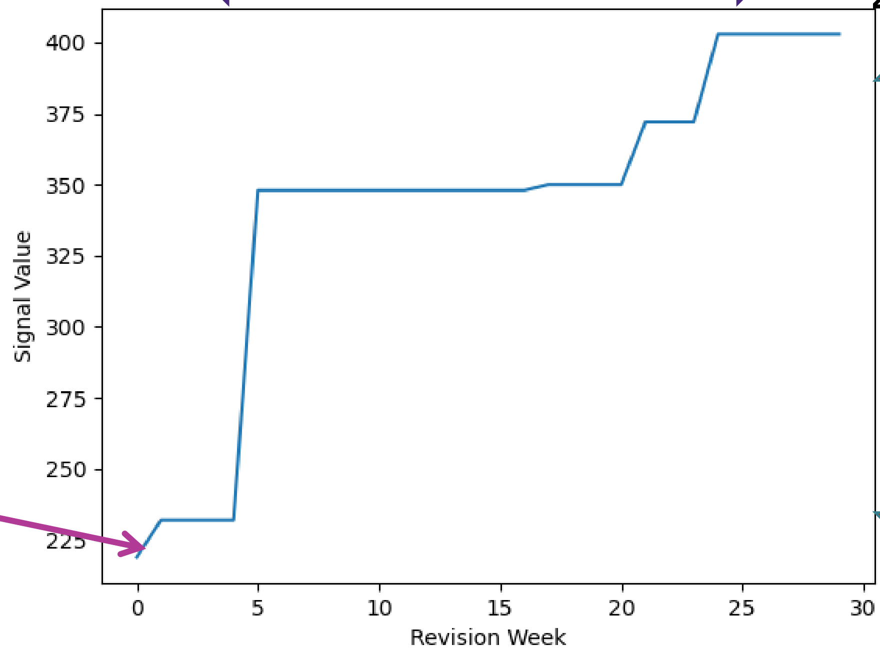
- Noisy and unreliable for some states
- New hospitalizations (target) is not reported by all states

Data Quality issues: Data Revisions

Stability Time = 24 weeks

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

Human error, data instability, delays, disasters



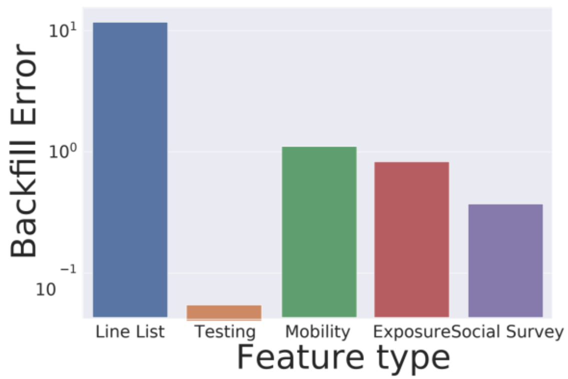
Initial/real-time value = 223

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

Revision of mortality for TX released on week 28

Data revisions are significant

- Over half the signals show backfill error over 32%
- Targets revised by 5%
- Stability time average around 3-4 weeks



Average Backfill Error across feature types

Model performance is affected by data revision

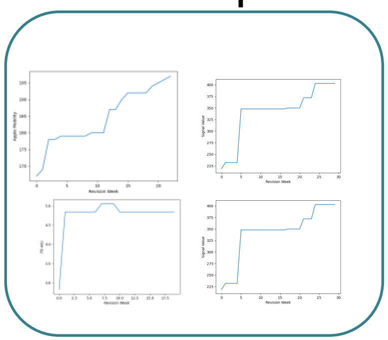
Refining predictions due to backfill

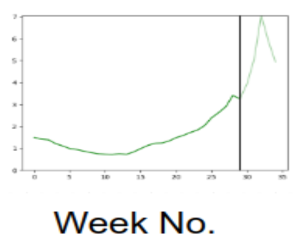
Given

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

Output

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



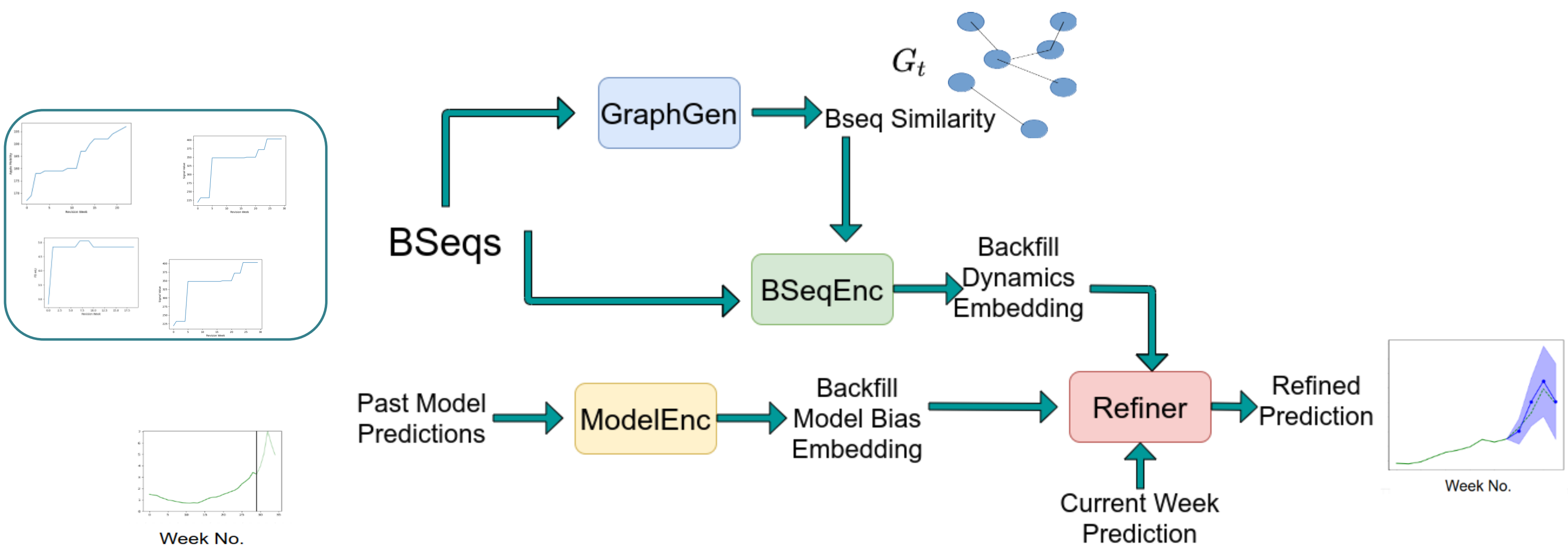


Current week prediction

Refined Prediction

Back2Future

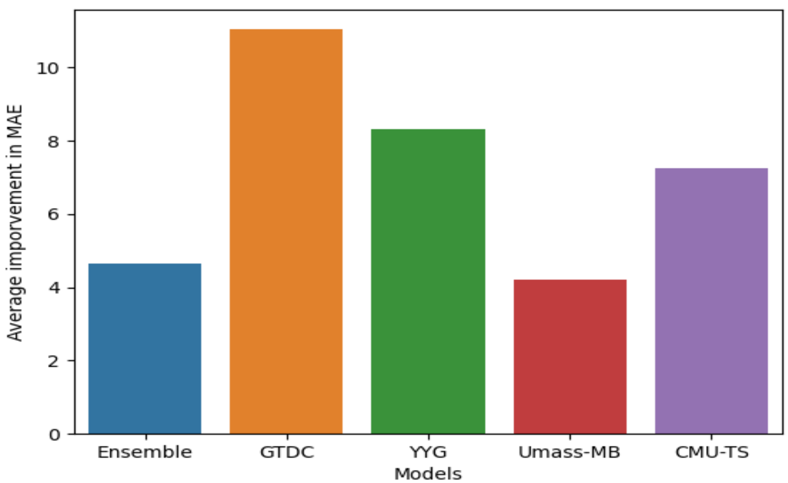
- Learns from past revision patterns of all features
- Refines model predictions of **any** model given prediction history



Rodríguez, Kamarthi, and Prakash 2021

Back2Future: Results

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

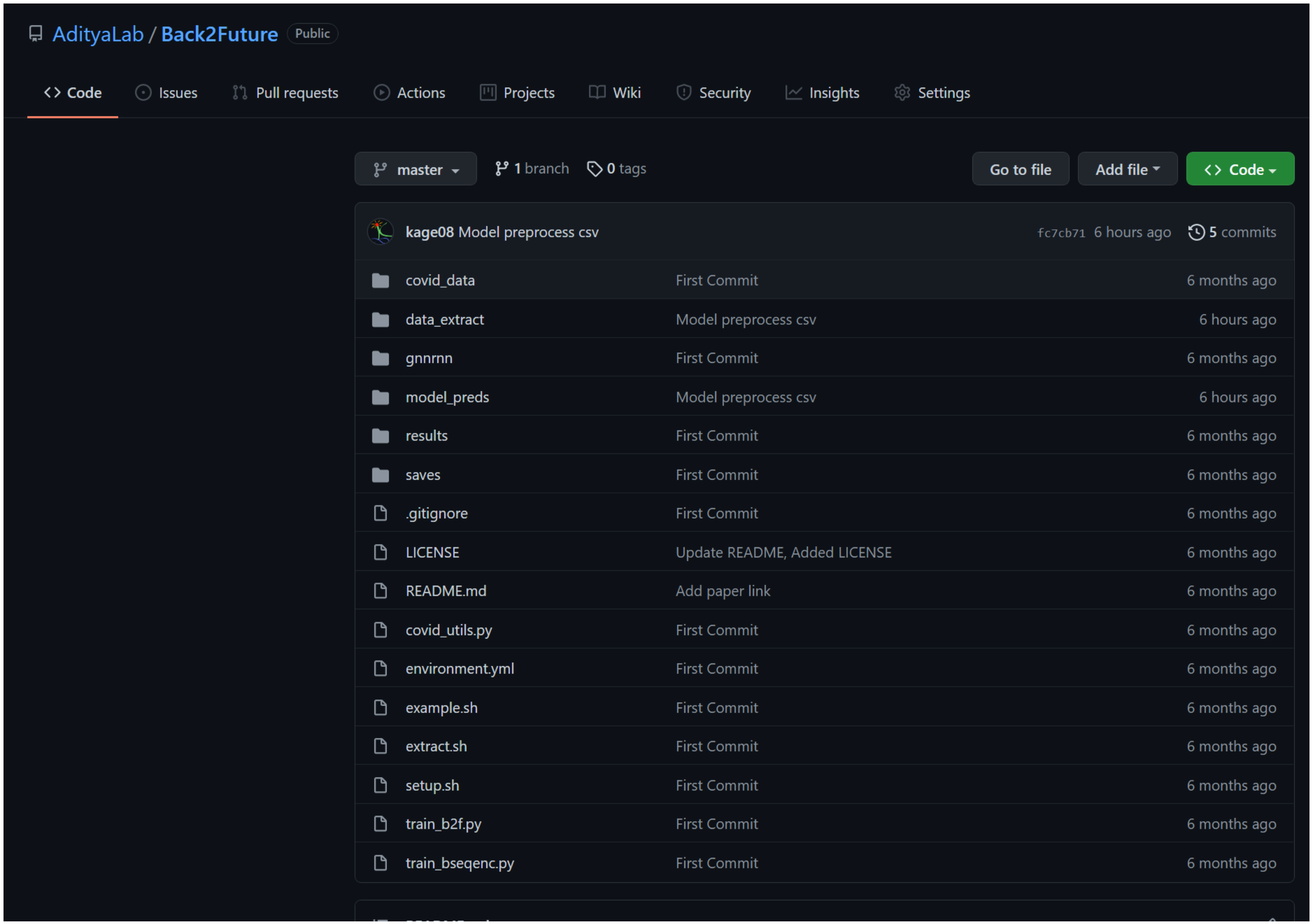


Takeaway: data quality issues can be helped with statistical correction

Rodríguez, Kamarthi, and Prakash 2021

Demo

Link: https://github.com/AdityaLab/Back2Future



[3] Decision making

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

Strategic Interventions for mitigating foot and mouth disease

[Probert+ PloS 2018, RS 2019]

- Use simulations based on past outbreak data.
- Control measures:
  - Vaccinate animals
  - Cull farm animals
- Can be solved as Sequential Decision making problem (leverage Reinforcement Learning)

Tactical Interventions for ventilator allocation

- Bertsimas et al. (2021) leverage future case forecasts to model optimal resource-allocation
- Tradeoff:
  - Satisfy future demand for ventilators
  - Reduce inter-state transport cost

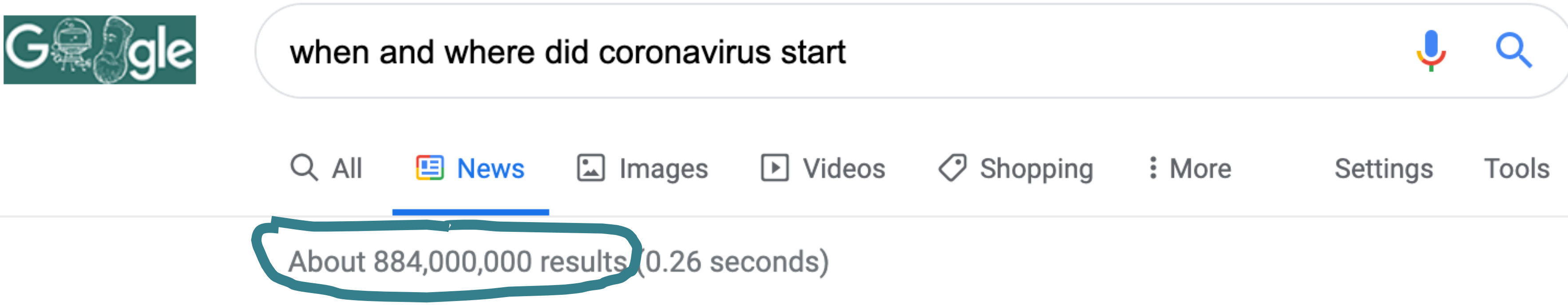
Final Remarks

[1] All models are useful

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

[2] Asking when, where, who

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



  - Very challenging!

[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

Studying epidemics in *real time*

- Editorial, Fineberg and Harvey, Science, May 2009: Epidemics Science in Real-Time
  - Five areas:
    - Pandemic risk,
    - vulnerable populations,
    - available interventions,
    - implementation possibilities
    - pitfalls, and public understanding

EDITORIAL

Harvey V. Fineberg is president of the Institute of Medicine.

Mary Elizabeth Wilson is associate professor of Global Health and Population at the Harvard School of Public Health and associate clinical professor at Harvard Medical School, Boston, MA.

Epidemic Science in Real Time

FEW SITUATIONS MORE DRAMATICALLY ILLUSTRATE THE SALIENCE OF SCIENCE TO POLICY THAN AN epidemic. The relevant science takes place rapidly and continually, in the laboratory, clinic, and community. In facing the current swine flu (H1N1 influenza) outbreak, the world has benefited from research investment over many years, as well as from preparedness exercises and planning in many countries. The global public health enterprise has been tempered by the outbreak of severe acute respiratory syndrome (SARS) in 2002–2003, the ongoing threat of highly pathogenic avian flu, and concerns over bioterrorism. Researchers and other experts are now able to make vital contributions in real time. By conducting the right science and communicating expert judgment, scientists can enable policies to be adjusted appropriately as an epidemic scenario unfolds.

In the past, scientists and policy-makers have often failed to take advantage of the opportunity to learn and adjust policy in real time. In 1976, for example, in response to a swine flu outbreak at Fort Dix, New Jersey, a decision was made to mount a nationwide immunization program against this virus because it was deemed similar to that responsible for the 1918–1919 flu pandemic. Immunizations were initiated months later despite the fact that not a single related case of infection had appeared by that time elsewhere in the United States or the world (www.iom.edu/swinefluaffair). Decision-makers failed to take seriously a key question: What additional information could lead to a different course of action? The answer is precisely what should drive a research agenda in real time today.

In the face of a threatened pandemic, policy-makers will want real-time answers in at least five areas where science can help: pandemic risk, vulnerable populations, available interventions, implementation possibilities and pitfalls, and public understanding. Pandemic risk, for example, entails both spread and severity. In the current H1N1 influenza outbreak, the causative virus and its genetic sequence were identified in a matter of days. Within a couple of weeks, an international consortium of investigators developed preliminary assessments of cases and mortality based on epidemic modeling.*

Specific genetic markers on flu viruses have been associated with more severe outbreaks. But virulence is an incompletely understood function of host-pathogen interaction, and the absence of a known marker in the current H1N1 virus does not mean it will remain relatively benign. It may mutate or acquire new genetic material. Thus, ongoing, refined estimates of its pandemic potential will benefit from tracking epidemiological patterns in the field and viral mutations in the laboratory. If epidemic models suggest that more precise estimates on specific elements such as attack rate, case fatality rate, or duration of viral shedding will be pivotal for projecting pandemic potential, then these measurements deserve special attention. Even when more is learned, a degree of uncertainty will persist, and scientists have the responsibility to accurately convey the extent of and change in scientific uncertainty as new information emerges.

A range of laboratory, epidemiologic, and social science research will similarly be required to provide answers about vulnerable populations; interventions to prevent, treat, and mitigate disease and other consequences of a pandemic; and ways of achieving public understanding that avoid both over- and underreaction. Also, we know from past experience that planning for the implementation of such projects has often been inadequate. For example, if the United States decides to immunize twice the number of people in half the usual time, are the existing channels of vaccine distribution and administration up to the task? On a global scale, making the rapid availability and administration of vaccine possible is an order of magnitude more daunting.

Scientists and other flu experts in the United States and around the world have much to occupy their attention. Time and resources are limited, however, and leaders in government agencies will need to ensure that the most consequential scientific questions are answered. In the meantime, scientists can discourage irrational policies, such as the banning of pork imports, and in the face of a threatened pandemic, energetically pursue science in real time.

– Harvey V. Fineberg and Mary Elizabeth Wilson

10.1126/science.1176297

*C. Fraser *et al.*, *Science* 11 May 2009 (10.1126/science.1176062).

Published by AAAS

Studying epidemics in *real time*

- Modeling **Before** the epidemic

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

- Modeling **After/During** the epidemic

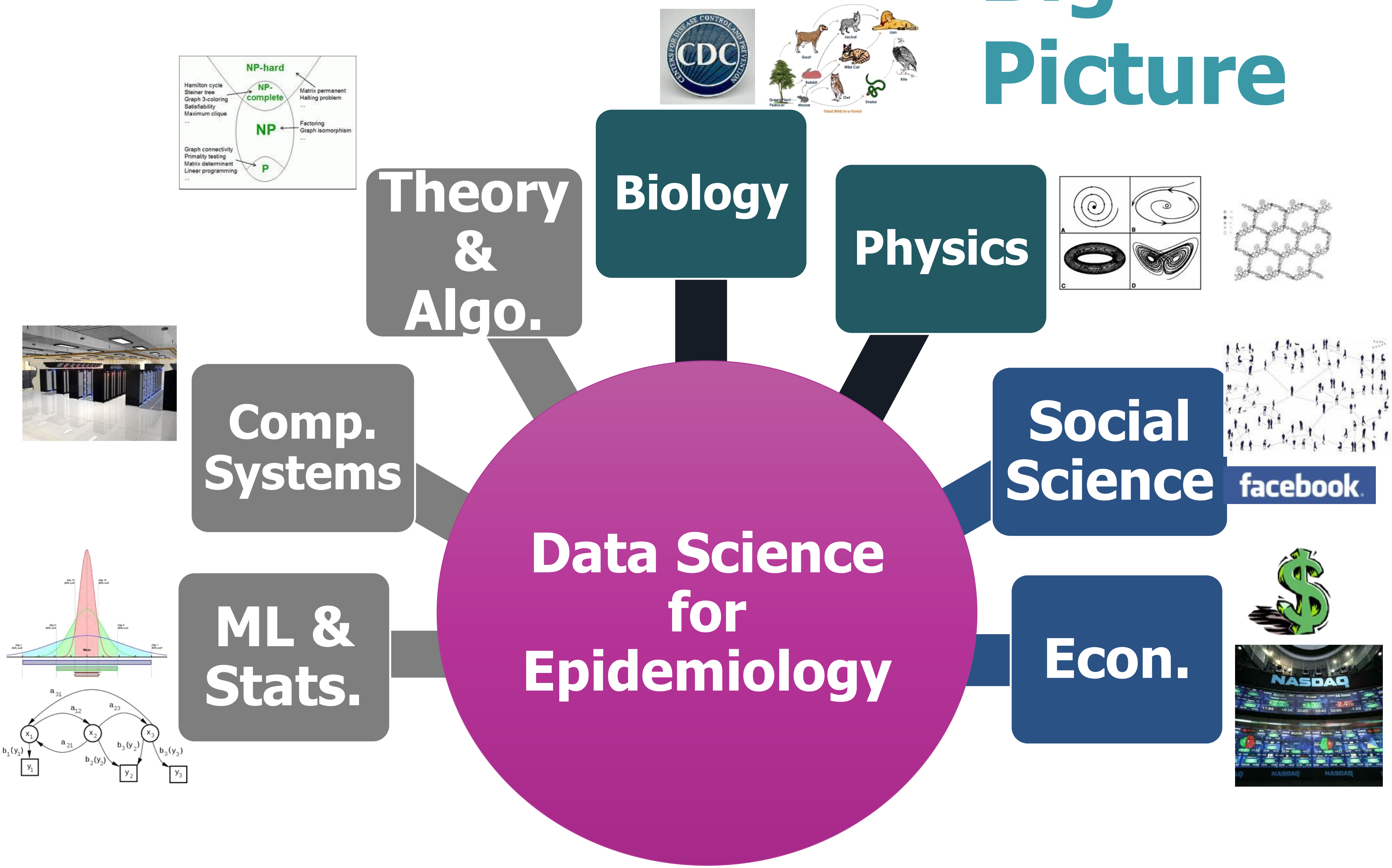
1. Quantifying transmission parameters,
2. Interpreting real-time epidemiological trends,
3. measuring antigenic shift
4. assessing impact of interventions.

Data Science is very important for all of these!

Why data science?

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

Big Picture



NP-hard
NP-complete
NP
P
Hamilton cycle
Steiner tree
Graph 3-coloring
Satisfiability
Maximum clique
Matrix permanent
Halting problem
Factoring
Graph isomorphism
Graph connectivity
Primality testing
Matrix determinant
Linear programming
CDC
Theory & Algo.
Biology
Physics
Comp. Systems
Social Science
facebook
ML & Stats.
Econ.
NASDAQ
Data Science for Epidemiology

Reminder on Workshop Webpage

- https://adityalab.cc.gatech.edu/workshops/21-forecasting-f4sg.html or b.gatech.edu/3cBPfQ7
- All Slides will be posted there.
- Talk video as well (later).

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

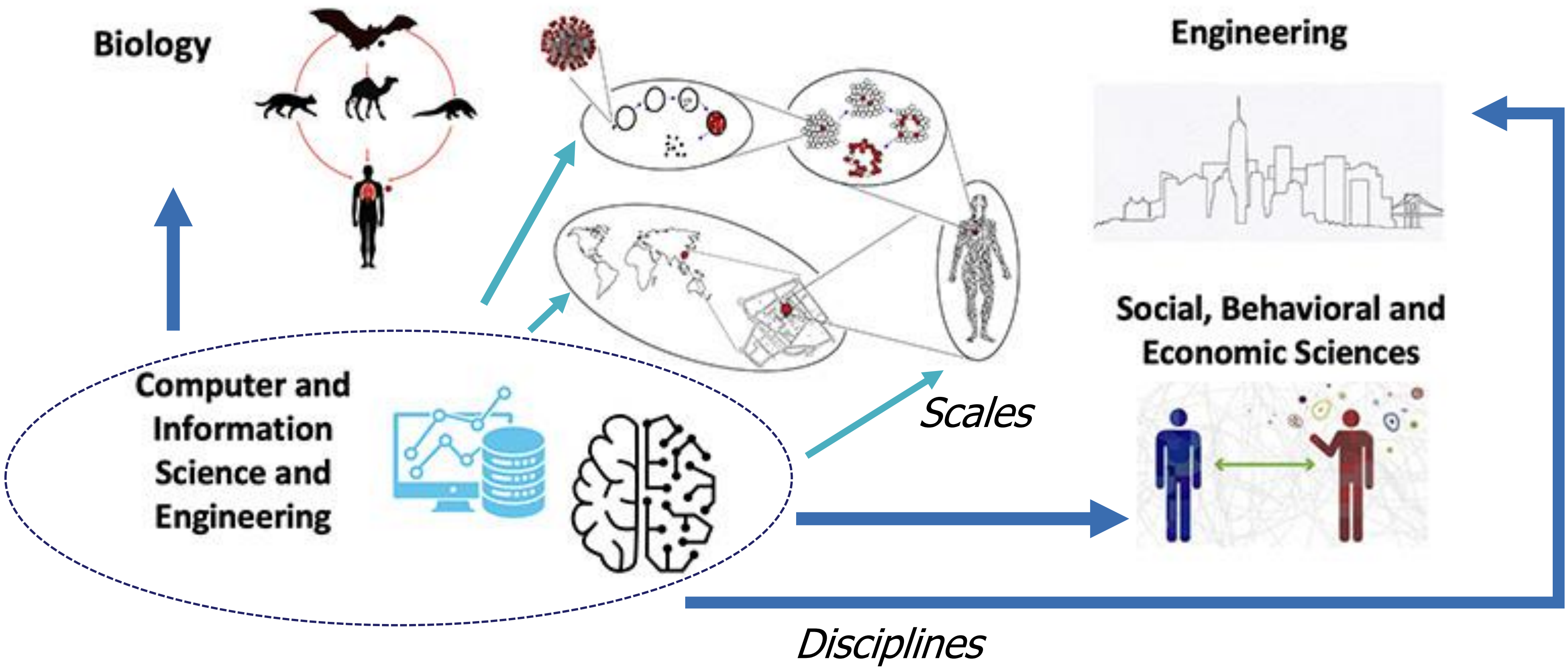
Stay tuned

- Survey paper coming soon

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

epiDAMIK @KDD 2021

- And more exciting research and tools!



We recently organized the **National PREVENT symposium (Feb 22/23):** Cross-cutting disciplines and scales for pandemic prevention and prediction

Videos and handouts: prevent-symposium.org

Thanks!

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

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

Stay in touch!

Alexander Rodríguez
- email: arodriguezc@gatech.edu
- web: cc.gatech.edu/~acastillo41

@arodriguezca

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

@harsha_64

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

@badityap

Georgia Tech