Data-Centric Computational Epidemic Forecasting

Alexander Rodríguez

Harshavardhan Kamarthi

B. Aditya Prakash

College of Computing

Georgia Institute of Technology

August 14, 2022

Outline

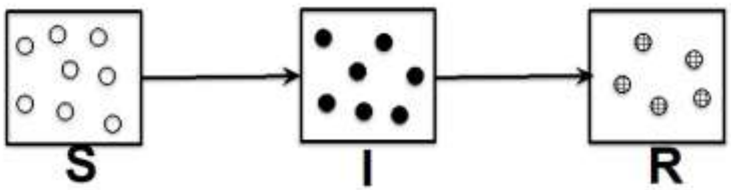
1. Epidemic forecasting (30 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (55 min)
5. **Hybrid models (45 min)**
6. Epidemic forecasting in practice (20 min)
7. Open challenges (20 min)

- 30 min break after Part 4
  - Feel free to catch us for coffee

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
  - …

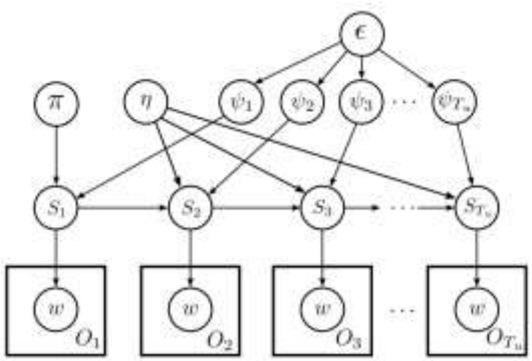
Stat/ML/AI models: overview

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)

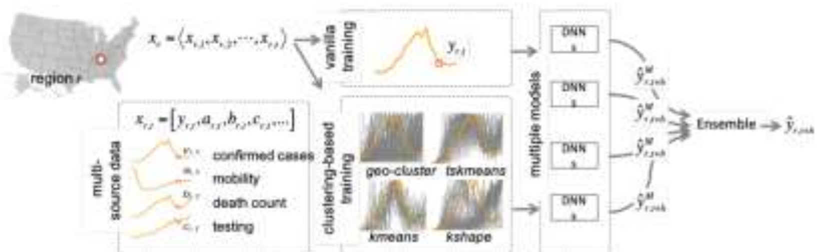
Language and vision



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

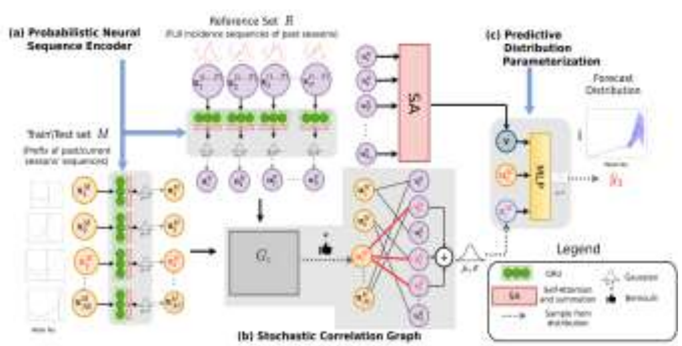
Stat/ML/AI models: overview

Deep learning



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

Density estimation



- 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	Statistical/ML/AI models

Mechanistic models

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

Statistical/ML/AI models

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

Hybrid Models

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

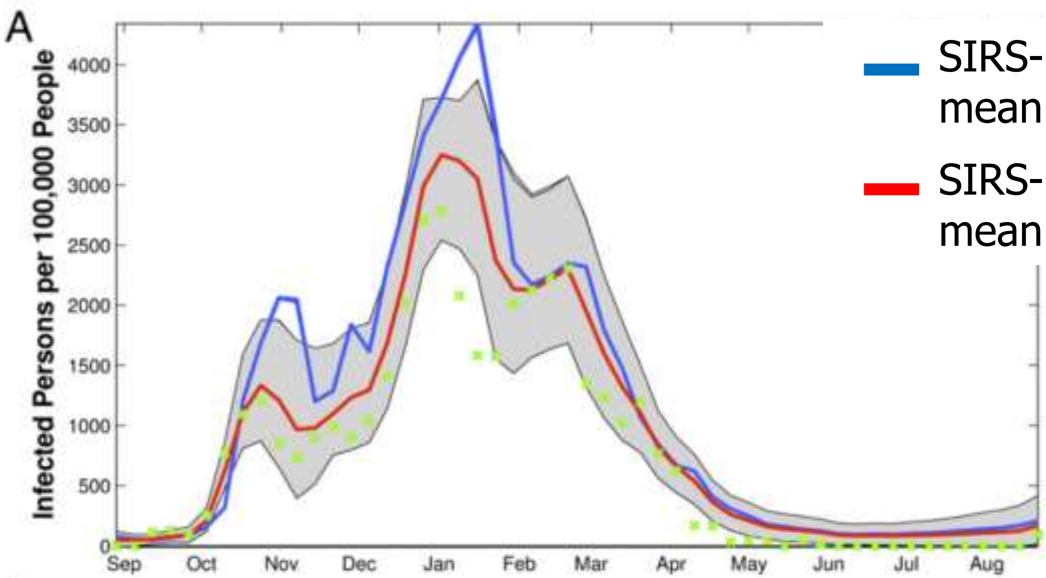
[H1] Mechanistic Model with Statistical Components

- 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)



GFT ILI estimates

SIRS-EAKF ensemble mean prior

SIRS-EAKF ensemble mean posterior

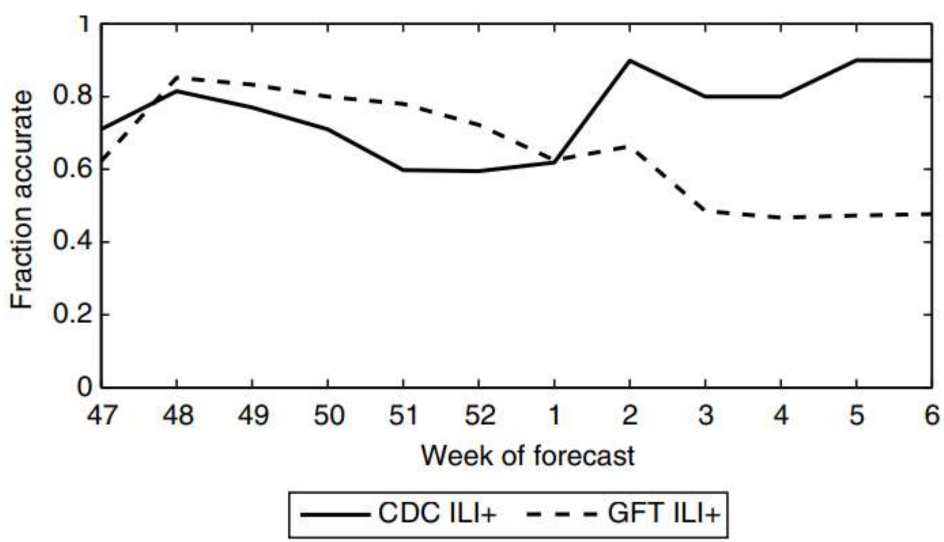
The SIRS equations are given by:

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

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

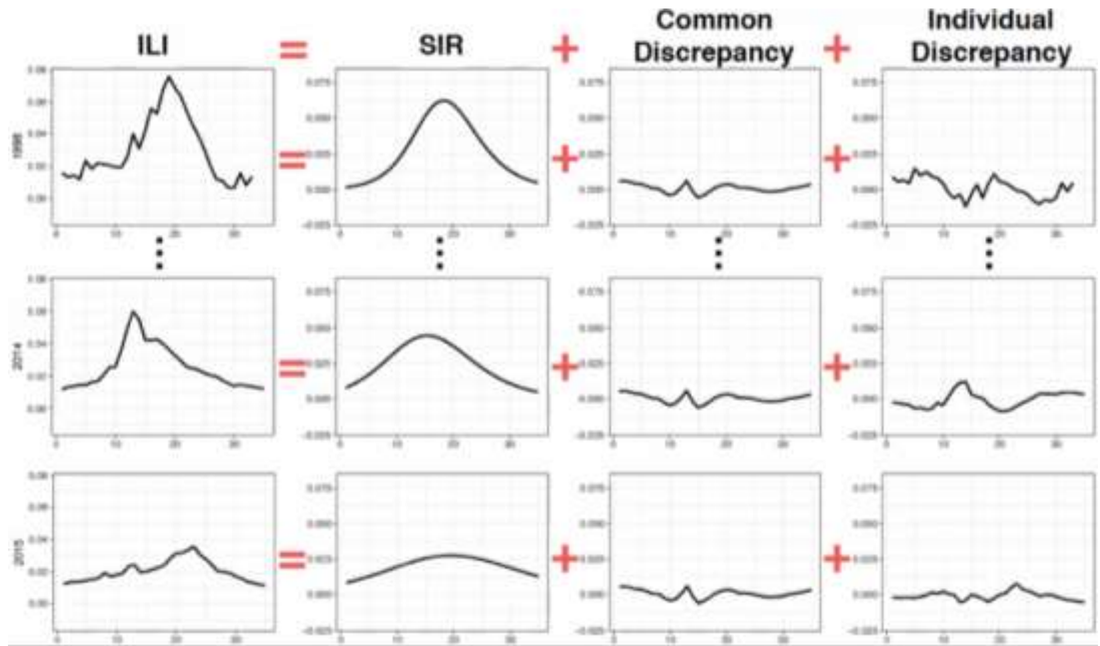


Figure credit: Sara Del Valle, LANL

Bayesian modeling of ILI

- Observable ILI:

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

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

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

SIR model

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

Across all seasons

$$\mu_T \sim \mathrm{N}(0, \sigma^2_{\mu_T}),$$
$$\mu_t | \mu_{t+1} \sim \mathrm{N}(\mu_{t+1}, \sigma^2_\mu).$$

Reverse random walk

Season-specific

$$\delta_{j,T} = -\mathrm{logit}(I_{j,T})$$
$$\delta_{j,t} | \delta_{j,t+1} \sim \mathrm{N}(\alpha_j \delta_{j,t+1}, \sigma^2_{\delta,j})$$

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

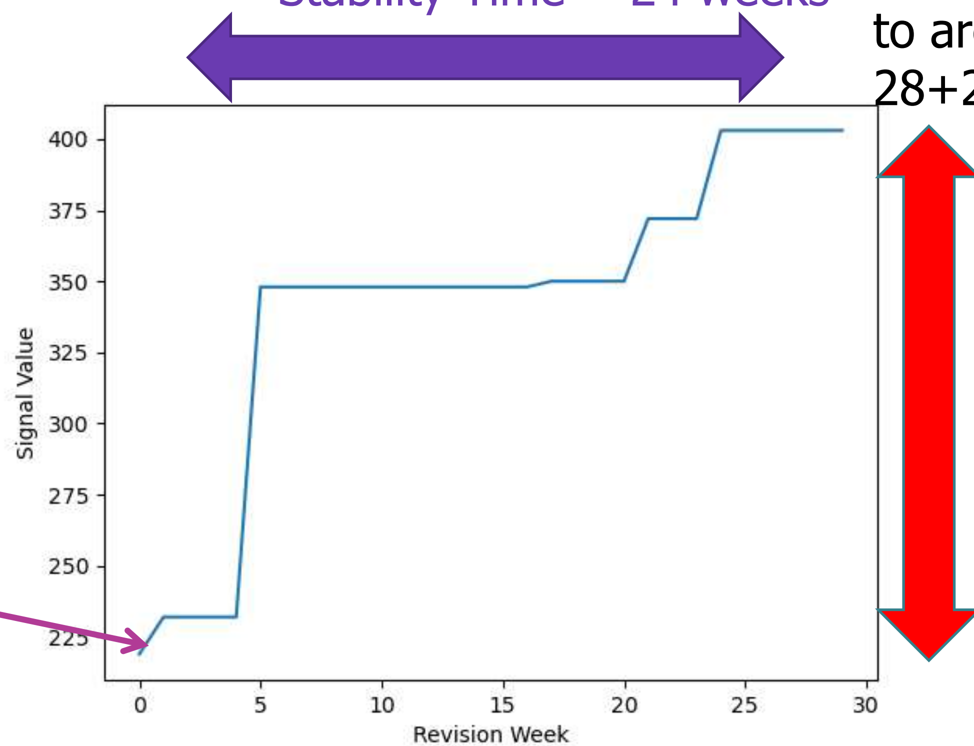
Ex. 2: Model-agnostic correction via data revision representations

[Kamarthi+, ICLR 2022]

- Data revisions (aka backfill)

Stability Time = 24 weeks

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



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

Initial/real-time value = 223

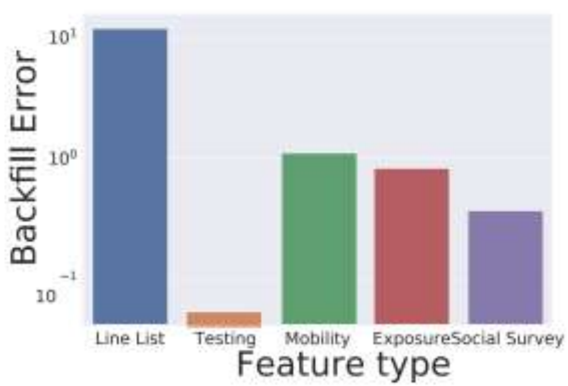
Revision of mortality for TX released on week 28

Rodríguez, Kamarthi, and Prakash 2022

Georgia Tech

Data revisions are ubiquitous

- Causes:
  - Human error, data instability, delays, disasters
- Over half the signals show backfill error over 32%
  - Clinical + digital surveillance, also behavioral data
- Stability time average around 3-4 weeks



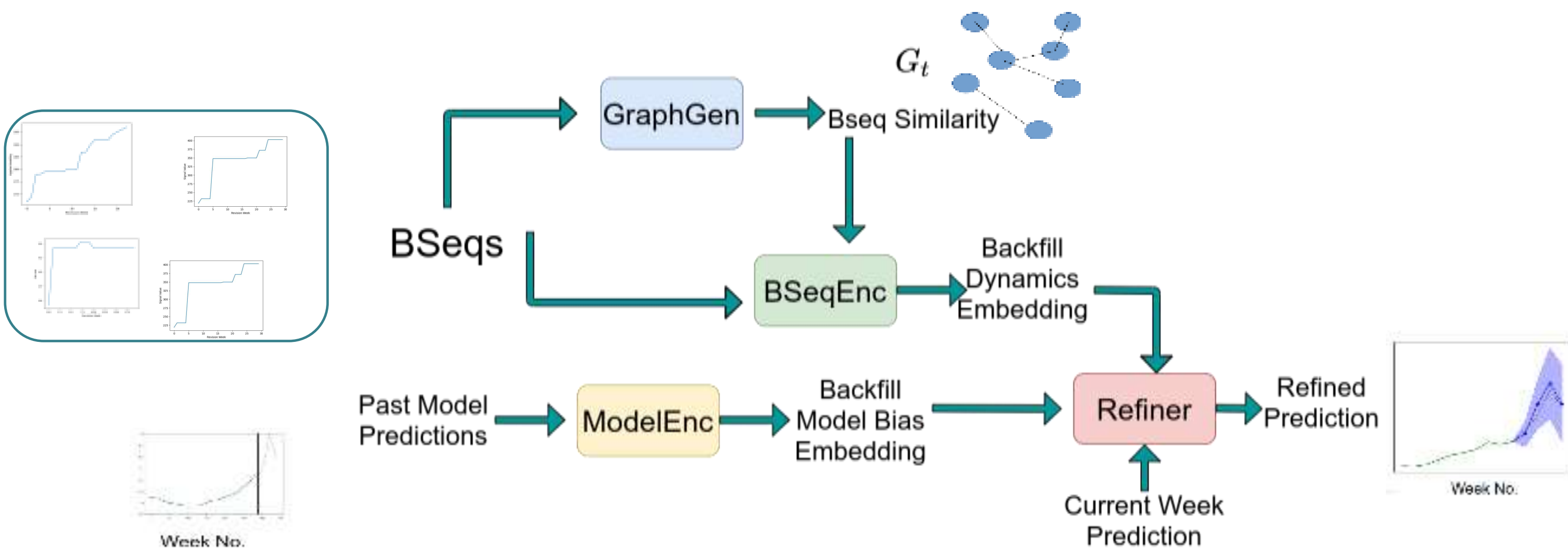
Average Backfill Error across feature types

Georgia Tech

Rodríguez, Kamarthi, and Prakash 2022

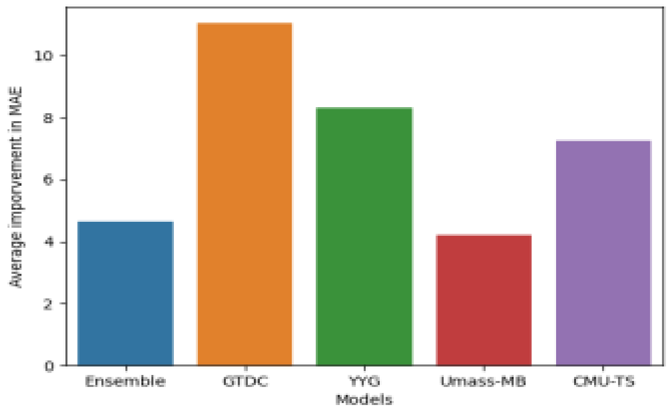
Back2Future

- (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

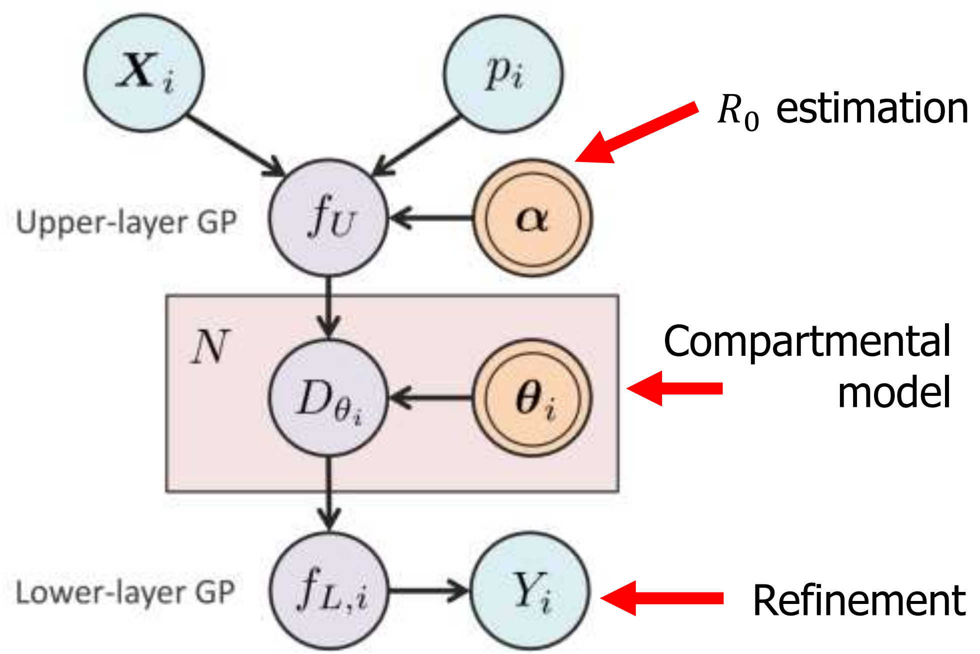
$$R_0(t) = \exp(a \times q(t) + b) + R_{0min}$$

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

Ex. 1: Compartmental Gaussian Processes

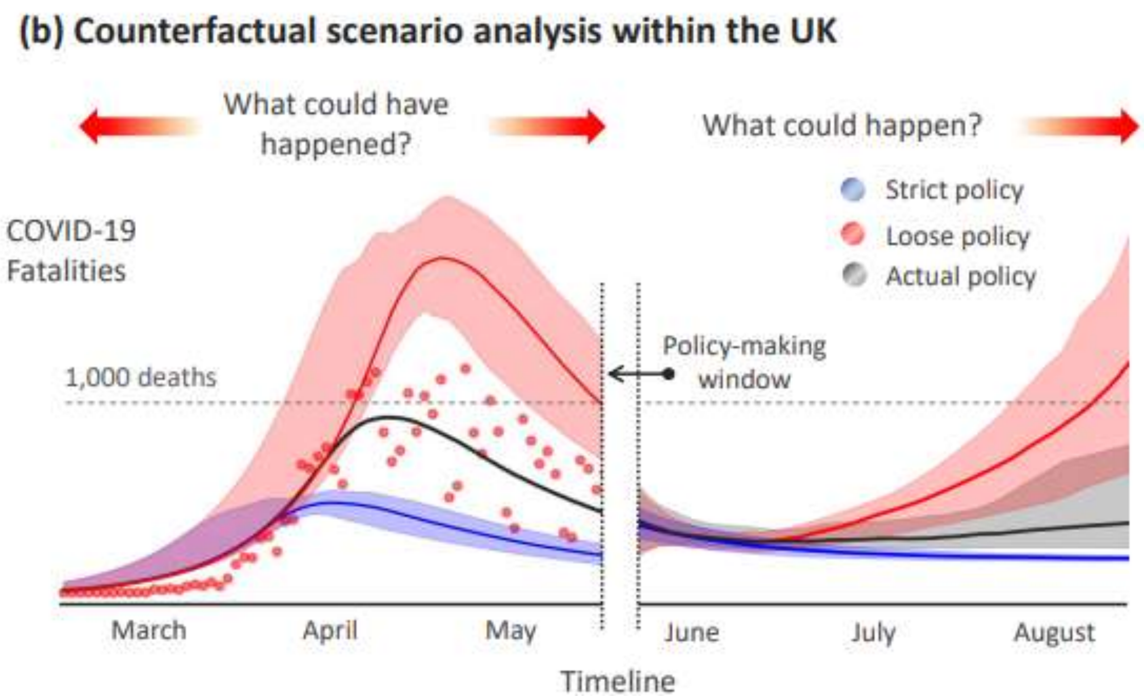
- Hierarchical two-layer Gaussian process (GP).
- Upper-layer GP uses country-specific features + policies in place to estimate R_0
- Lower-layer GP refines predictions

[Qian+ NeurIPS 2020]



Counterfactual based on new set of policies

- What if we have a different input policies p_i?

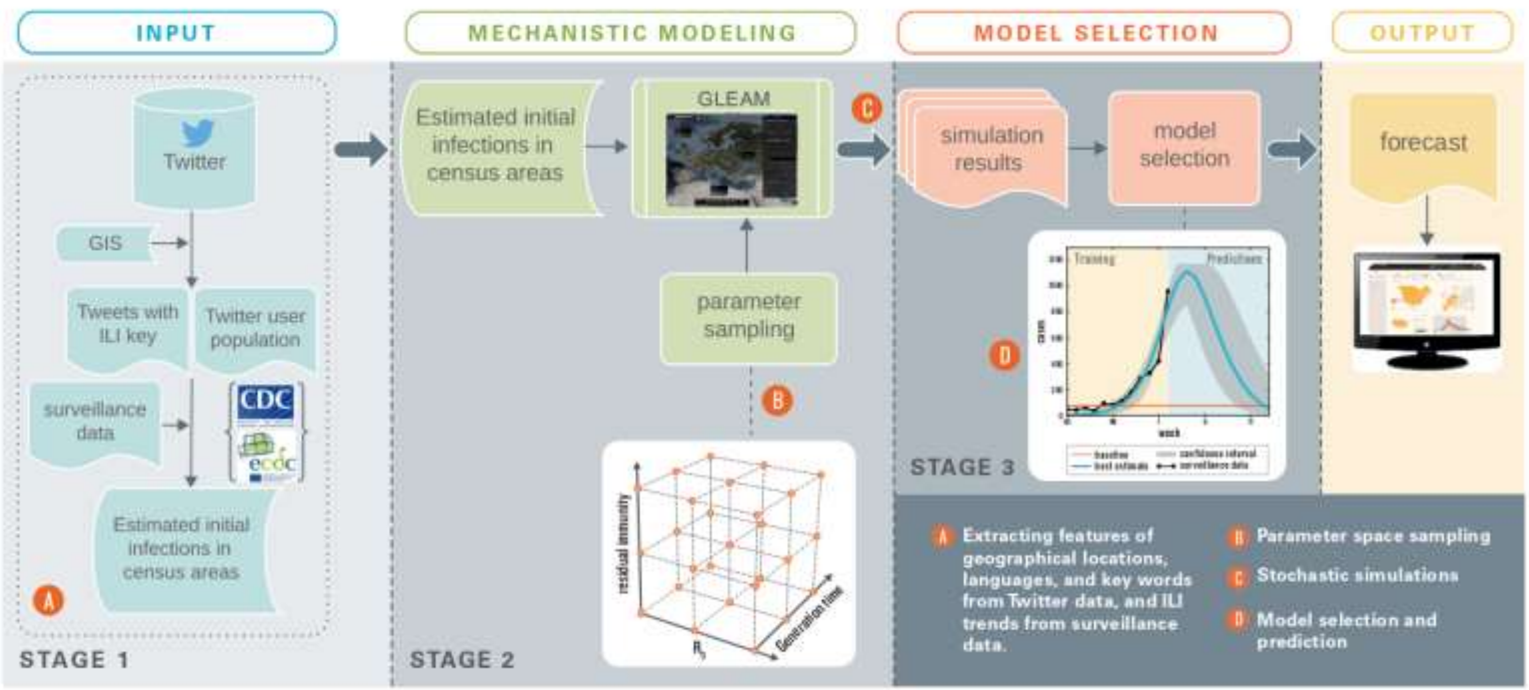


Rodríguez, Kamarthi, and Prakash 2022

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

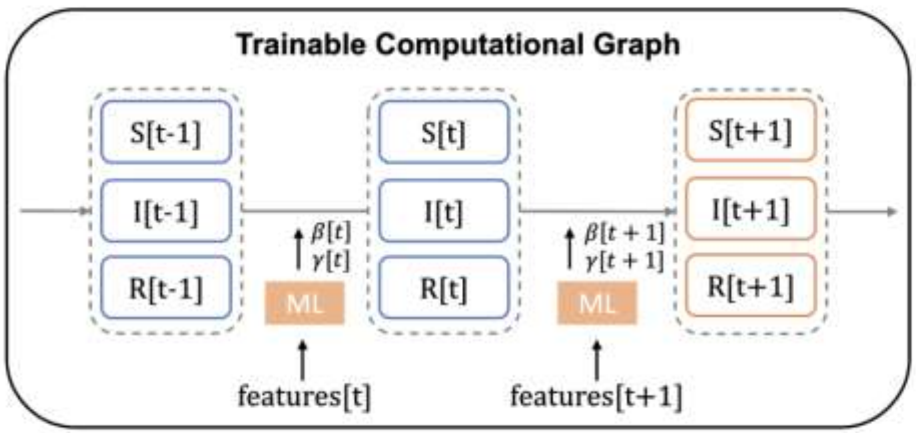
[Arik+, NeurIPS 2020]

- Using additive encoders for time-varying features.

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

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

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



Rate variable	Covariates
$\boldsymbol{\beta}$	Mobility, Interventions, Density, Past Counts
$\boldsymbol{\eta}$	Census, Healthcare supply
$\boldsymbol{\gamma}$	Census, Test count / pos. ratio, Past Counts
h, c, v, $\boldsymbol{\varrho}$, $\boldsymbol{\kappa}$	Census, Econometrics, Healthcare supply

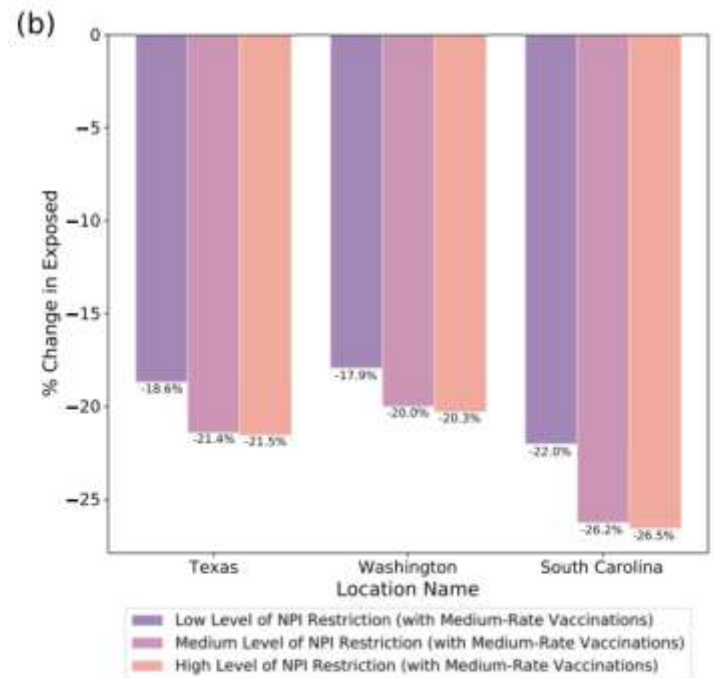
$$v_i[t] = v_{i,L} + (v_{i,U} - v_{i,L}) \cdot \sigma\left(c + b_i + \mathbf{w}^\top \mathrm{cov}(v_i, t)\right)$$

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

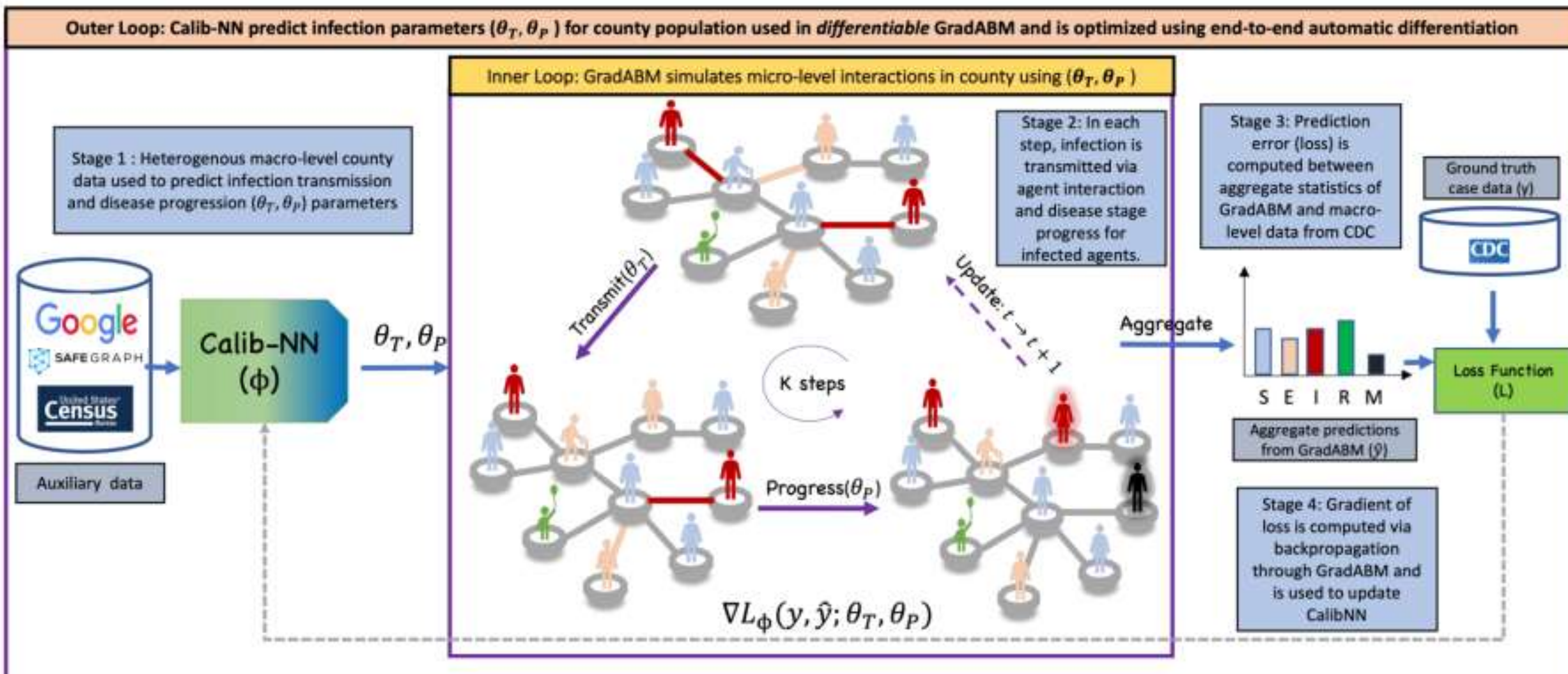
- Directional penalty regularization:
- Ex. 1: If the mobility increases, the average contact rates increase
- Ex. 2: If state of emergency (SoE) is introduced, contact rates decrease

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

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)

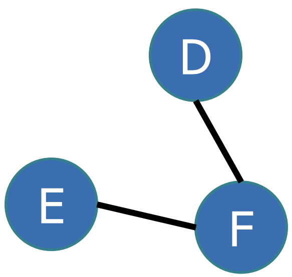
Stage 2: Disease transmission + progression
Message passing in graph neural network (GNN) + reparametrization trick



Georgia Tech

Reformulation of disease transmission and progression

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



$$f(t,s_i,a_s,n) = \frac{R S_{a_s} A_{s_i} B_n}{\bar{I}} \int_{t-1}^{t} G_\Gamma(u;\mu_i,\sigma_i^2)\,du,$$

$$q(t,s_i,a_s,n) = 1 - e^{-\lambda(t,s_i,a_s,n)},$$

$$\boldsymbol{h}_t^n = q_t\left(\boldsymbol{h}_{t-1}^n, \bigcup_{n_j:n_j\to n}^{k} f_t\left(\boldsymbol{h}_{t-1}^n, k, \boldsymbol{h}_{t-1}^{n_j}\right)\right)$$

Reformulation of disease transmission and progression

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

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

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

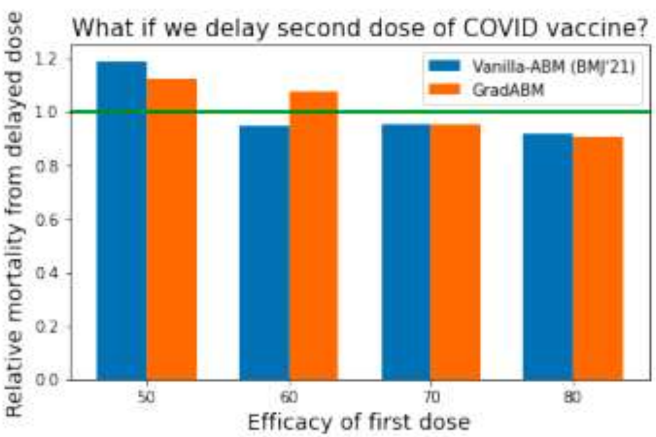
$$= \begin{cases} \texttt{Transmit}(X_i^t, \mathcal{N}_i, (X_j^t)_{j \in \mathcal{N}_i}, \theta_T^t), & \text{if } d_i^t = S, \\ \texttt{Progress}(X_i^t, \theta_P^t), & \text{if } d_i^t \in \{E, I\}. \end{cases} \qquad (2)$$

Benefits

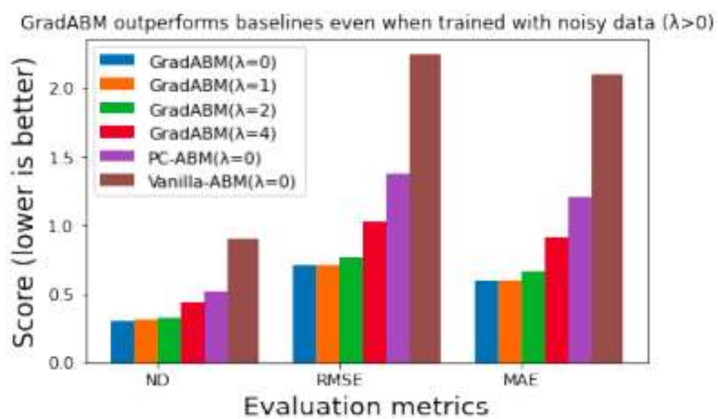
Forecasting

Model	COVID-19 ND	COVID-19 RMSE	COVID-19 MAE	Influenza ND	Influenza RMSE	Influenza 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	$\mathbf{0.97 \pm 0.18}$	$\mathbf{50.99 \pm 12.12}$	$\mathbf{30.02 \pm 5.60}$	$\mathbf{0.41 \pm 0.02}$	$\mathbf{1.47 \pm 0.06}$	$\mathbf{1.22 \pm 0.06}$
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)

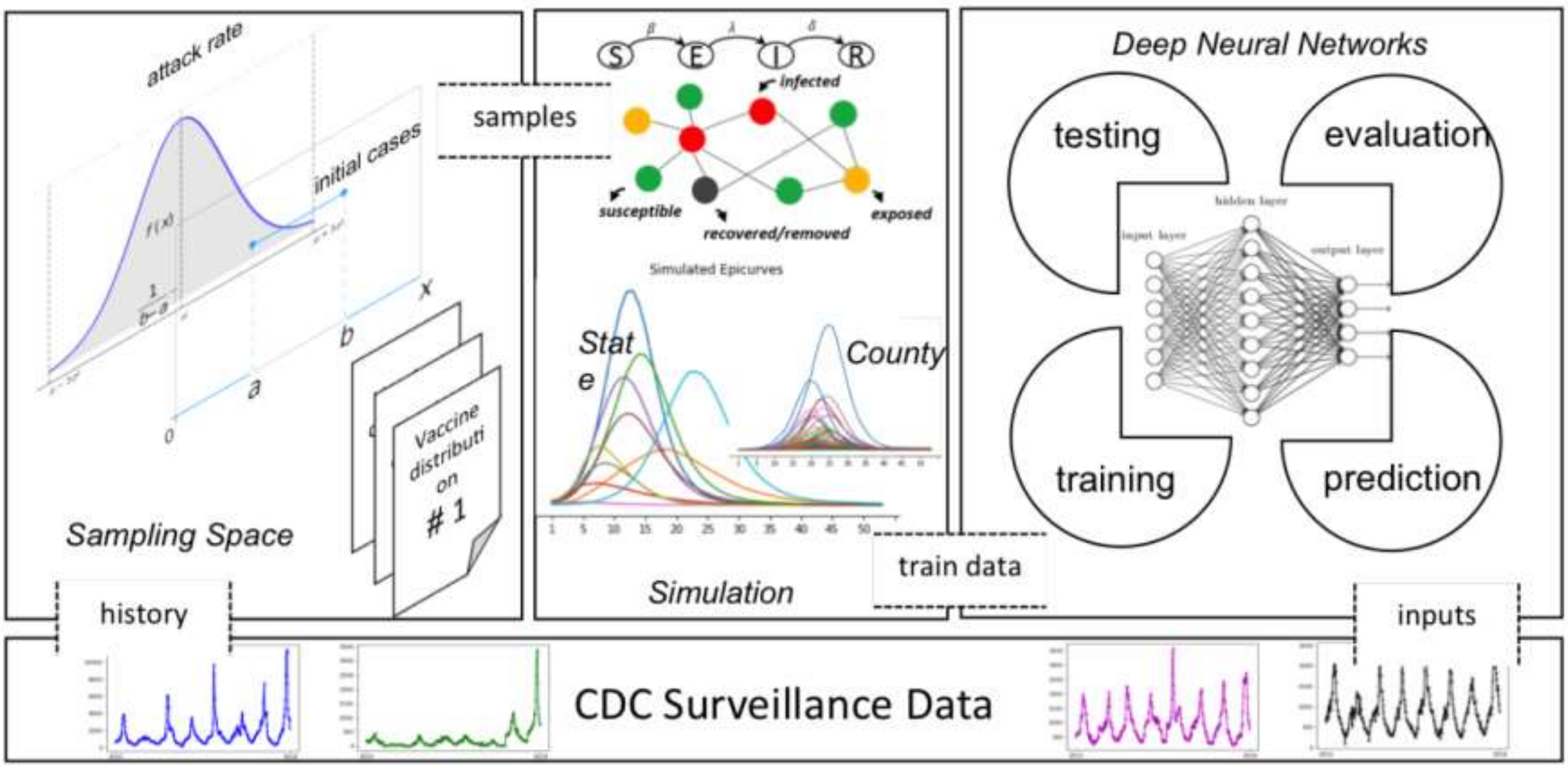
- Approaches:
  1. Mechanistic model with statistical components
  2. **Priors from mechanistic models inform statistical model**
  3. Wisdom of crowds

[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

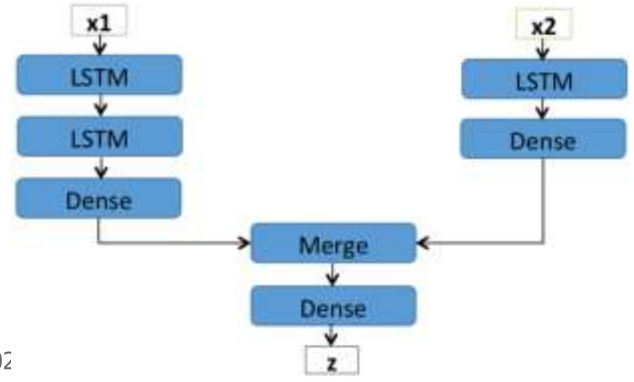
Ex. 1: Deep learning w/ with Synthetic Information (DEFSI)

[Wang+, AAAI 2019]



Major components of DEFSI framework

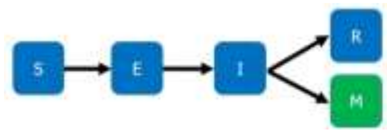
- (1) Disease model parameter space construction
  - Agent-based simulator **EpiFast** w/ SEIR
  - Parameters from literature [Marathe+, PLOS One 2011]
- (2) Synthetic training data generation
  - High-resolution (more granular than reported by CDC)
- (3) Deep neural network training and forecasting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

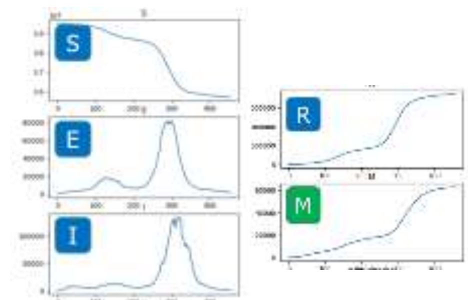


Ex. 2: Epidemiologically-informed Neural Networks (EINNs) [Rodríguez+, arXiv 2022]

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

Time (t) → PINNs → SEIRM(Ω_t)





- EINNs connect features to latent epi dynamics

Features + time (t) → EINNs → SEIRM(Ω_t)

- Gradient approx.: $\frac{d\mathbf{s}_t^F}{dt} = \frac{d\mathbf{s}_t^F}{d\mathbf{e}_t^F}\frac{d\mathbf{e}_t^F}{dt} \approx \frac{d\mathbf{s}_t^F}{d\mathbf{e}_t^F}\frac{d\mathbf{e}_t}{dt}$
- Epi-domain constrains

$$\mathcal{L}^{Mono} = \frac{1}{N+1}\left(\sum_{t=t_0}^{t_N}\left|\frac{dS_t}{dt}\mathrm{ReLu}\left(\frac{dS_t}{dt}\right)\right|^2 + \sum_{t=t_0}^{t_N}\left|-1\frac{dR_t}{dt}\mathrm{ReLu}\left(-\frac{dR_t}{dt}\right)\right|^2\right)$$

- Overcome spectral bias:

$$\Gamma(\mathbf{v}) = [\cos(2\pi\mathbf{B}\mathbf{v}), \sin(2\pi\mathbf{B}\mathbf{v})]^T$$

- ...

Rodríguez, Kamarthi, and Prakash 2022

Results: extend capabilities of both mechanistic and ML models

	Short-term (1-4 wks)			Long-term (5-8 wks)			Trend correlation
Model	NRMSE1	NRMSE2	ND	NRMSE1	NRMSE2	ND	Pearson
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)

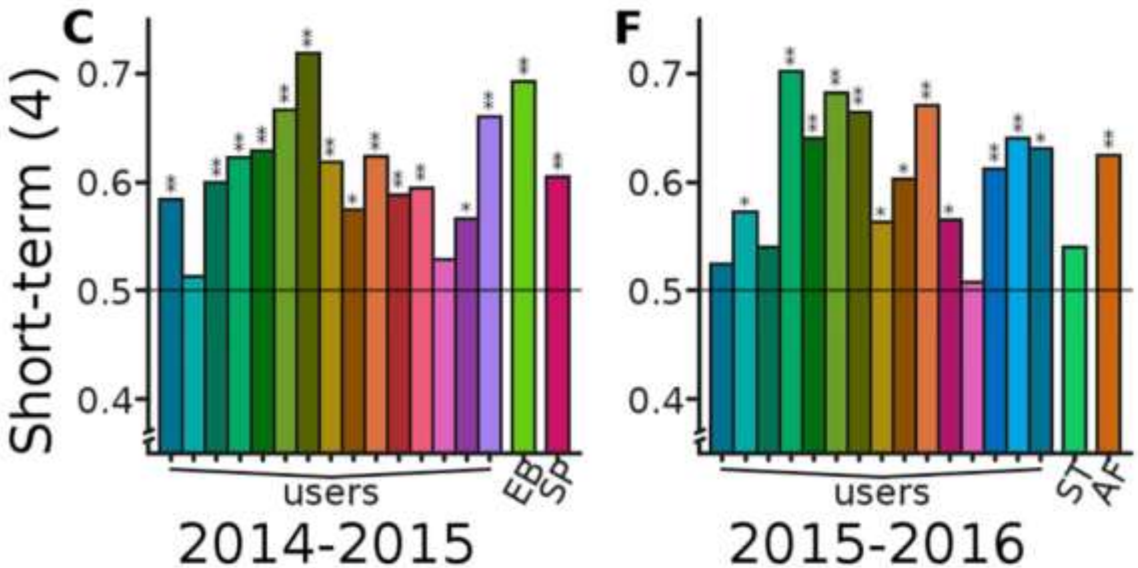
- Leverages multiple predictions from different sources
- Objectives:
  - Account for limitations of a single source of predictions
- Ideas:
  - **Expert predictions**
  - Prediction markets
  - Ensembles

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

Epicast Win Rate



Ex. 2: Experts vs laypeople

[Recchia+, PLOS One 2021]

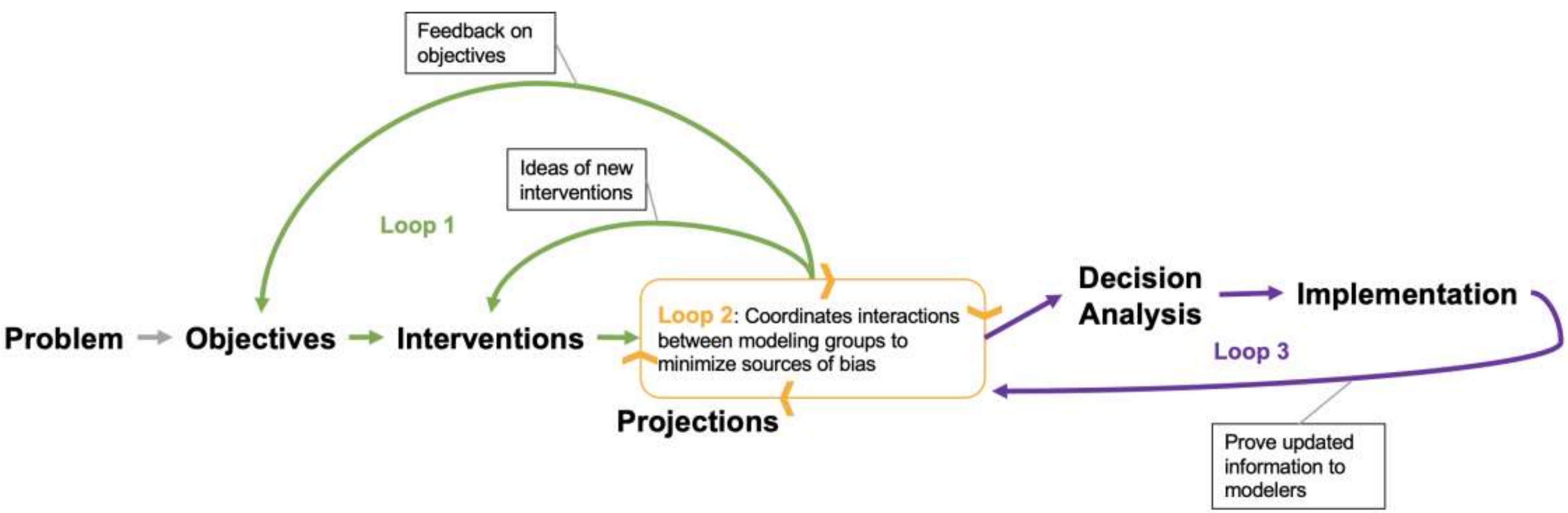
Table 1. Questions asked of participants with corresponding forecast medians, median absolute deviation (MAD), median absolute error (MAE) and median relative error (MRE).

	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

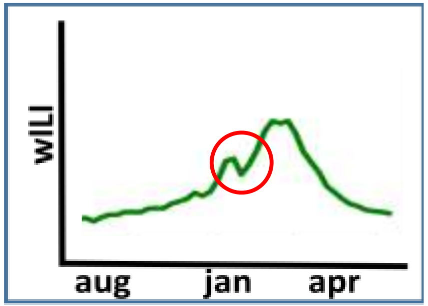
[Shea+, Science 2020]

- 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

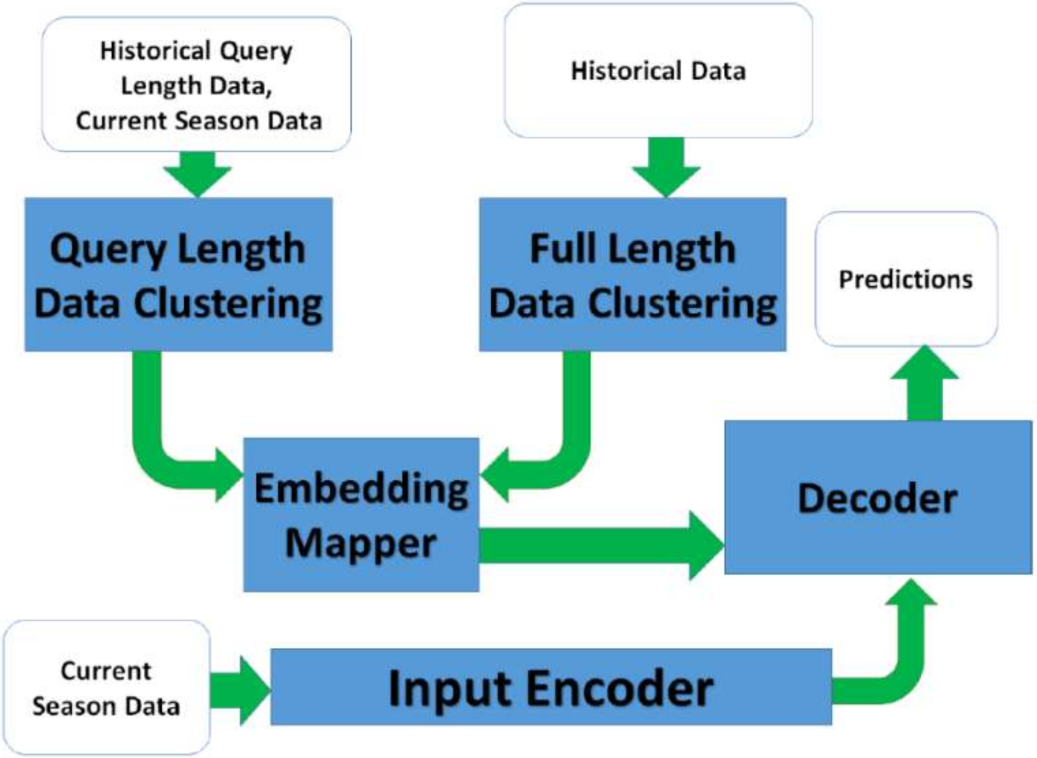
$$g(\theta) = E(|\mu(\theta, t+1, R_1) - \mu(\theta, t+1, R_2)|) - \epsilon \leq 0$$

Squared error in Region 1

Squared error in Region 2

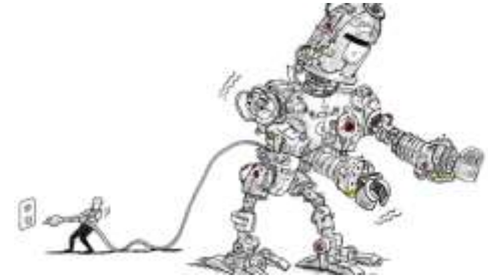
Recall EpiDeep

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



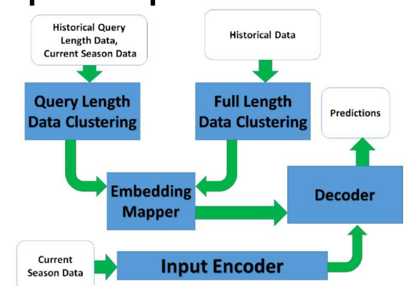
Rodríguez, Kamarthi, and Prakash 2021

Guided-EpiDeep



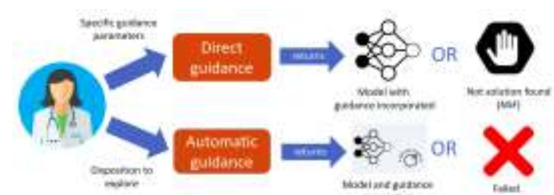
- Seldonian Optimization Framework [Thomas+, 2019]
  - Proposed for AI safety
  - Precludes undesirable behavior of AI model by enforcing behavioral constraints in optimization
  - Has a safety test in unobserved data
- **Main idea:** Enforce Seldonian Framework on Epideep optimization to incorporate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

EpiDeep architecture



\+

Guidance



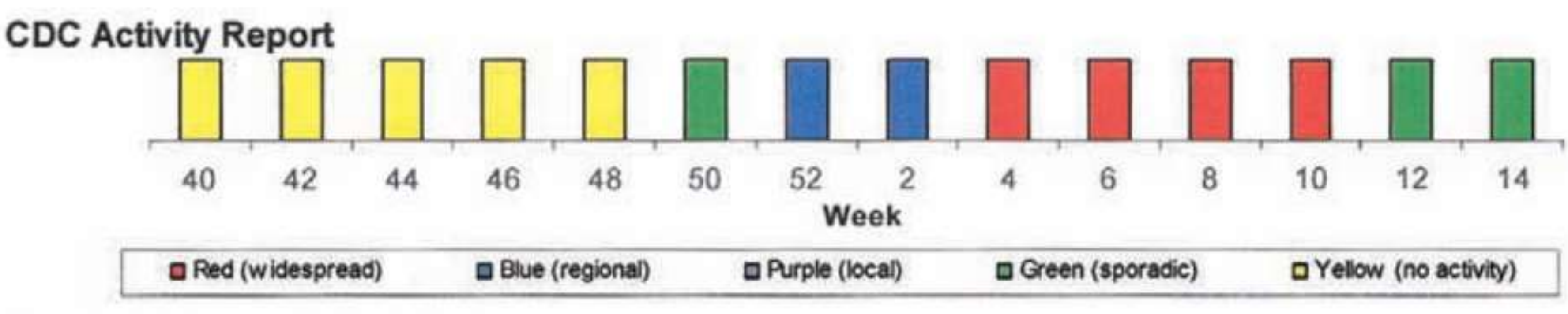
[H3] Wisdom of Crowds (WoC)

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

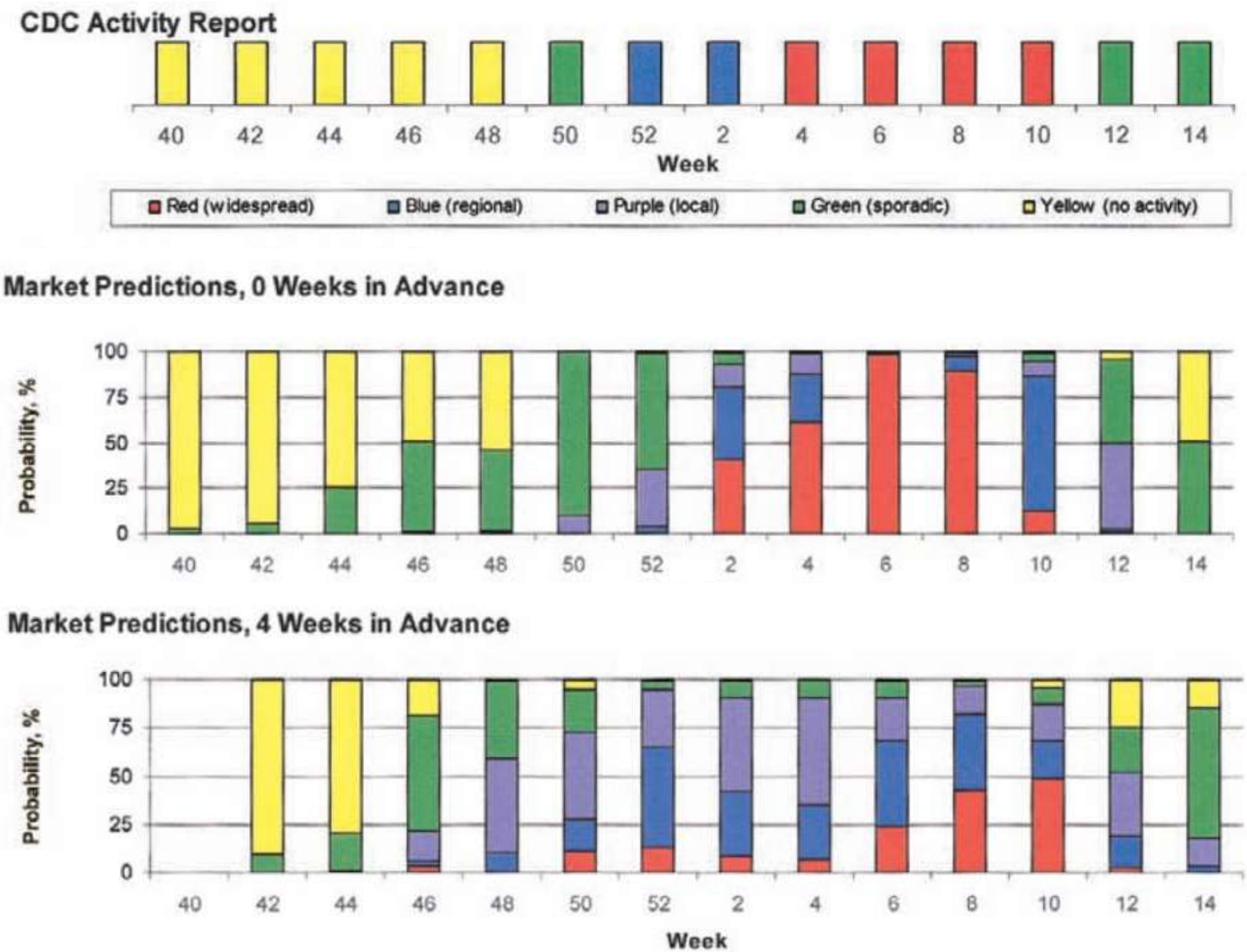
Ex. 1: Prediction markets

[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

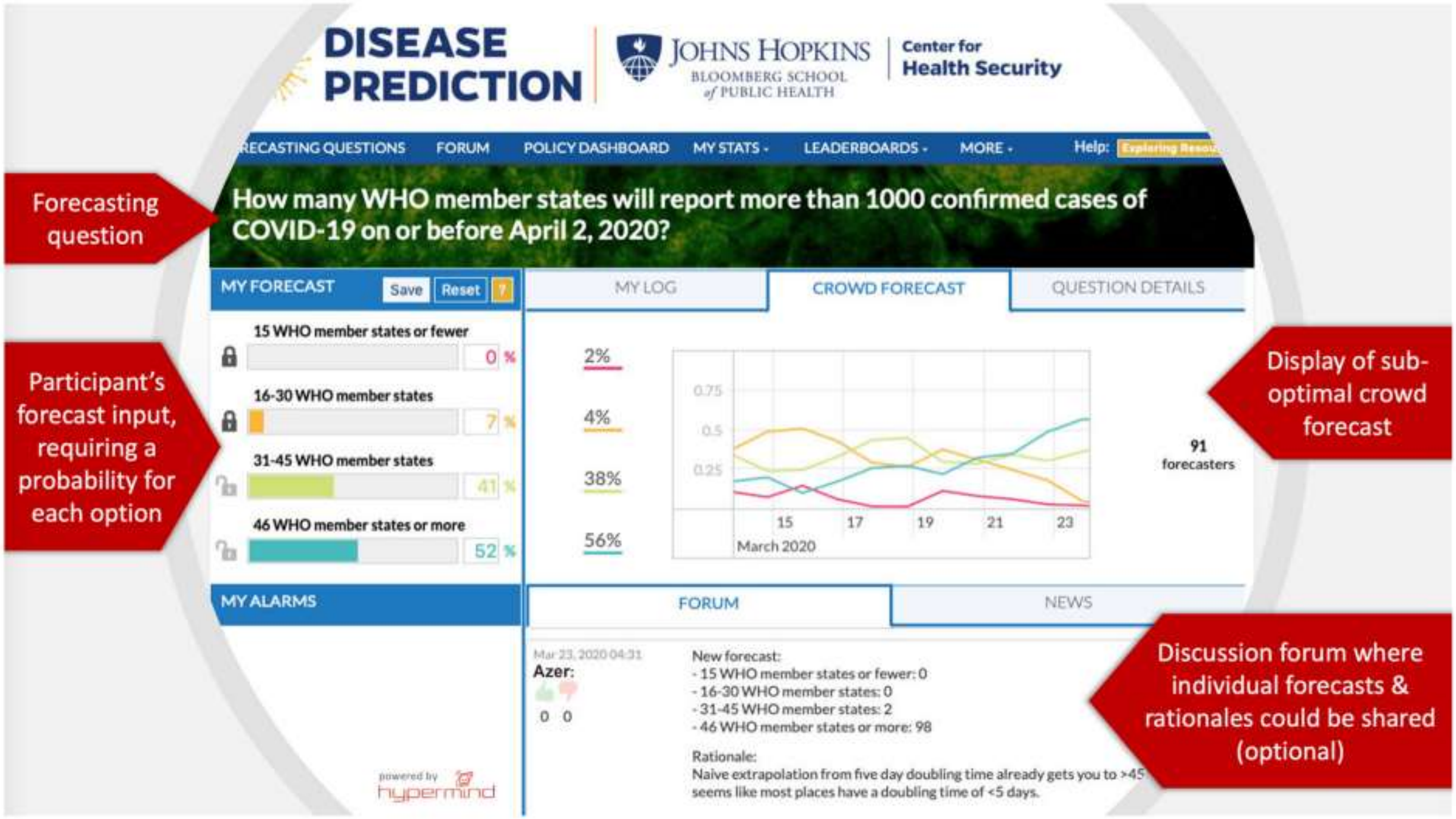


Results


CDC Activity Report
40 42 44 46 48 50 52 2 4 6 8 10 12 14
Week
Red (widespread)
Blue (regional)
Purple (local)
Green (sporadic)
Yellow (no activity)
Market Predictions, 0 Weeks in Advance
Probability, %
100
75
50
25
0
40 42 44 46 48 50 52 2 4 6 8 10 12 14
Market Predictions, 4 Weeks in Advance
Probability, %
100
75
50
25
0
40 42 44 46 48 50 52 2 4 6 8 10 12 14
Week

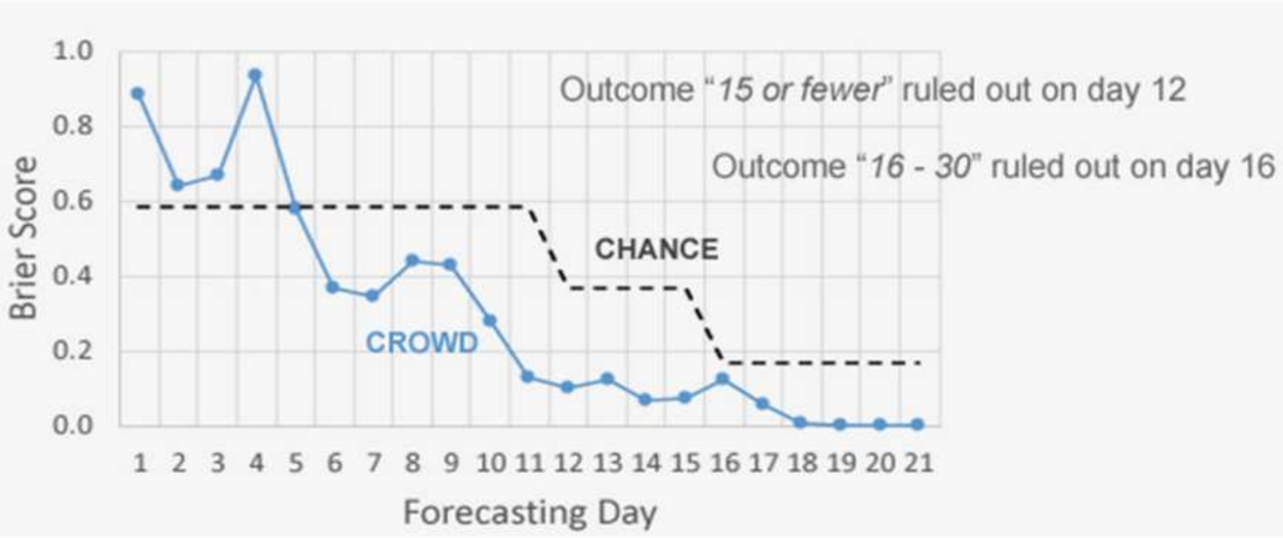
Ex. 2: For pandemic response

[Sell+, BMC Public Health 2021]



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



Rodríguez, Kamarthi, and Prakash 2022

[H3] Wisdom of Crowds (WoC)

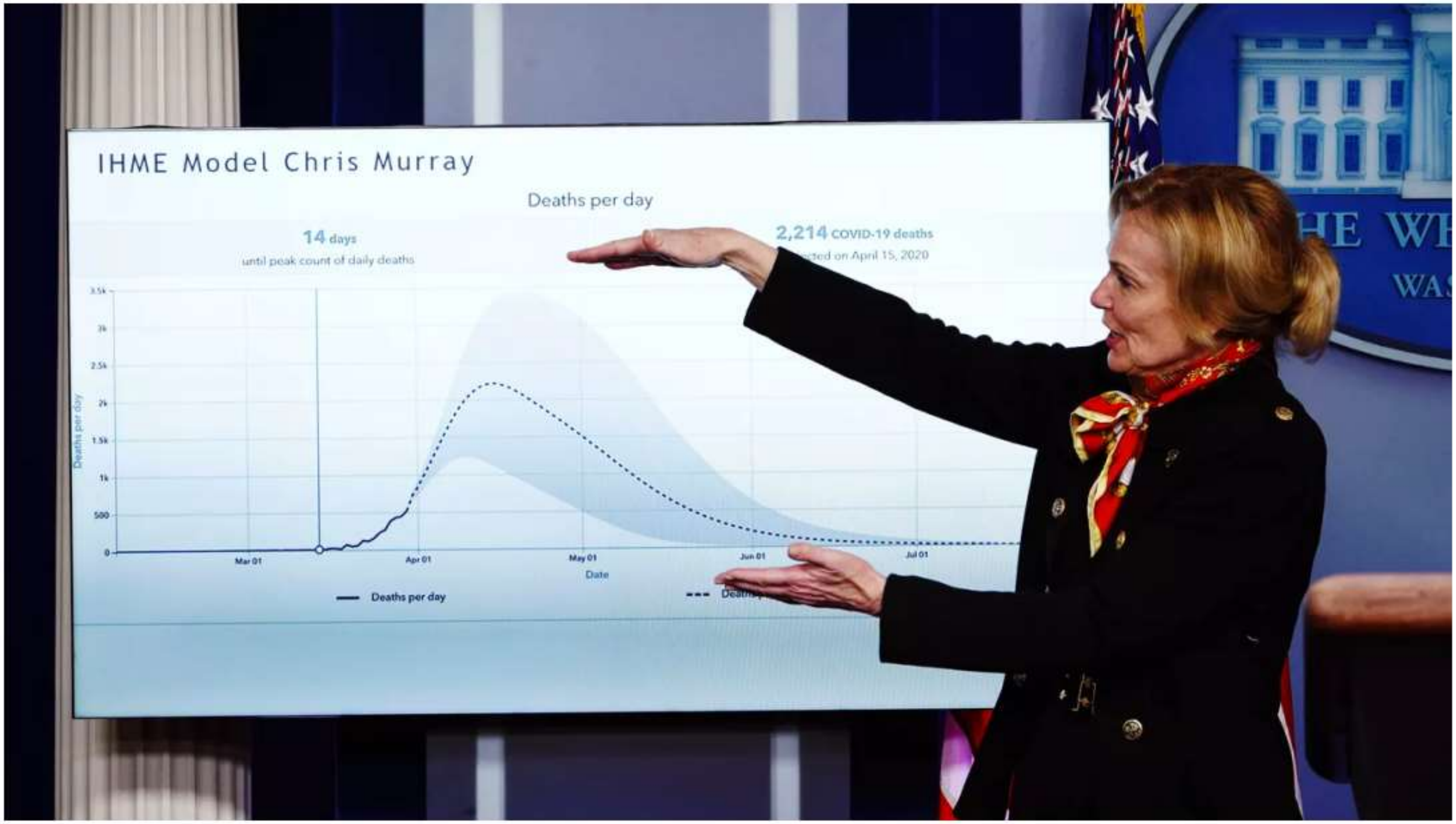
- Leverages multiple predictions from different sources
- Objectives:
  - Account for limitations of a single source of predictions
- Ideas:
  - Expert predictions
  - Prediction markets
  - **Ensembles**

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

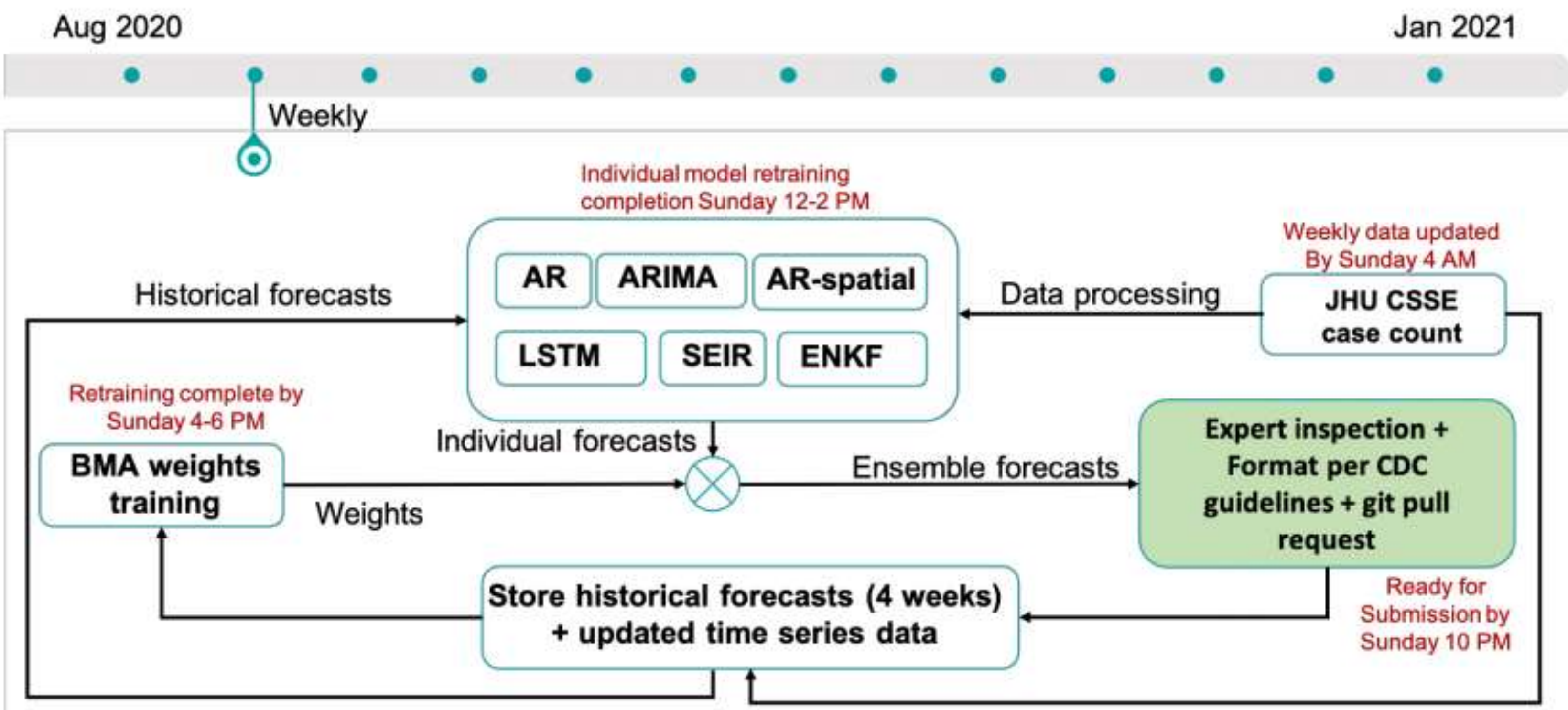


Slide credit: Nicholas Reich, UMass Amherst
covid19forecasthub.org/doc/talks/

Ex. 2: Bayesian ensemble

[Adiga+, KDD 2021]

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



COVID-19 ForecastHub

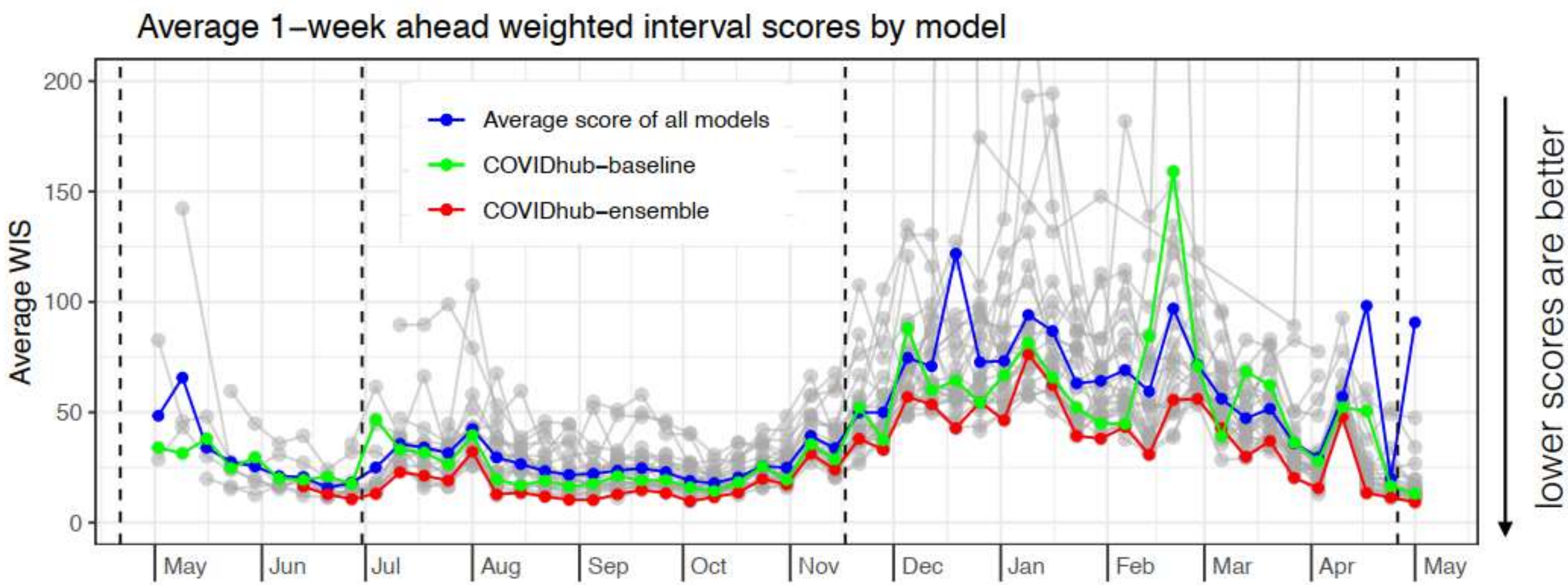
Diversity of COVID-19 models

- 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 (30 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (55 min)
5. Hybrid models (45 min)
6. **Epidemic forecasting in practice (20 min)**
7. Open challenges (20 min)

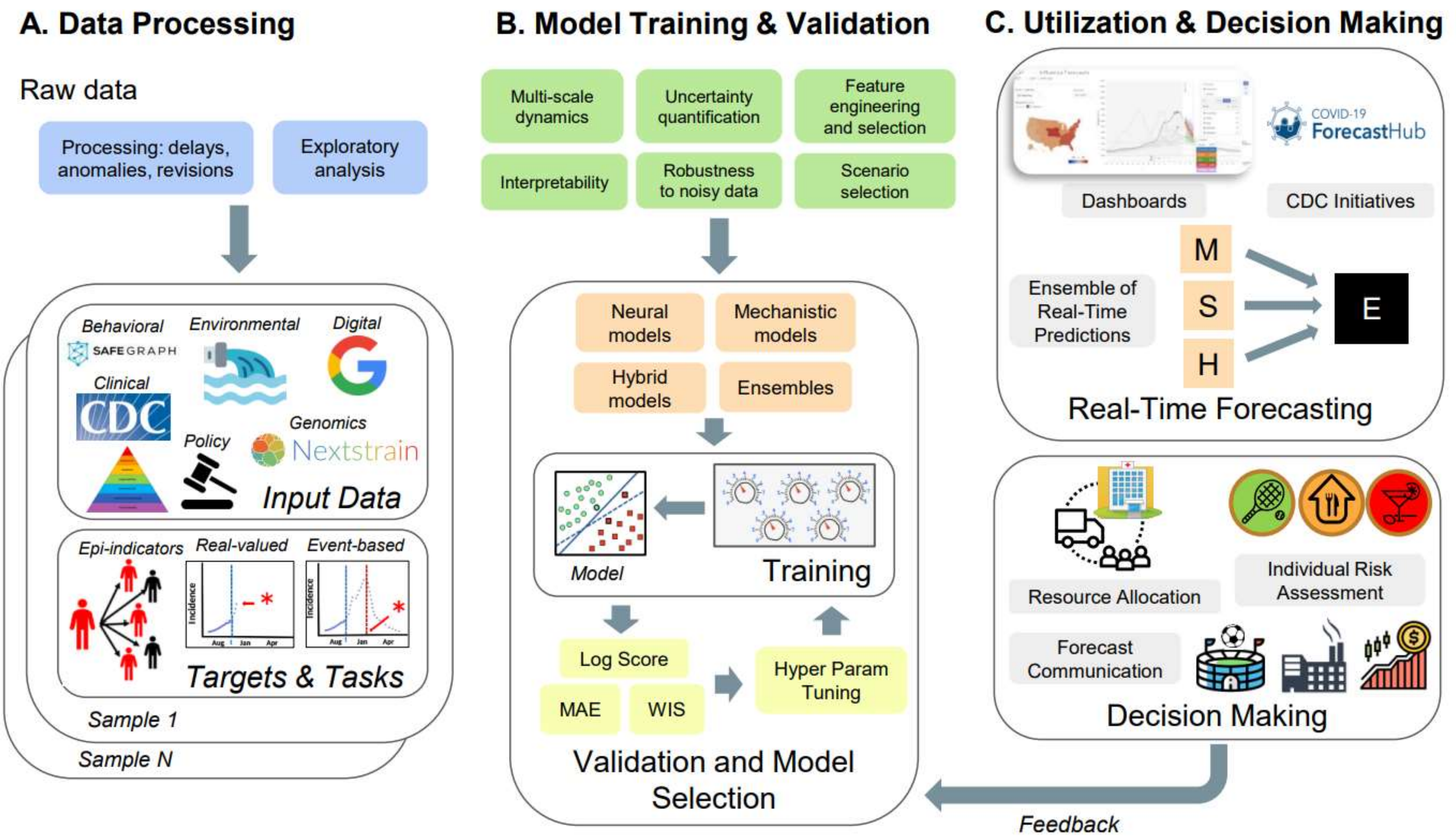
- 30 min break after Part 4
  - Feel free to catch us for coffee

Break 3 min

Part 6: Epidemic Forecasting in Practice

Remind

Epidemic Forecasting Pipeline



Rodríguez, Kamarthi, and Prakash 2022

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

Epidemic Forecasting in Practice (Outline)

- Real-time forecasting on the ground
- Forecasting and decision making

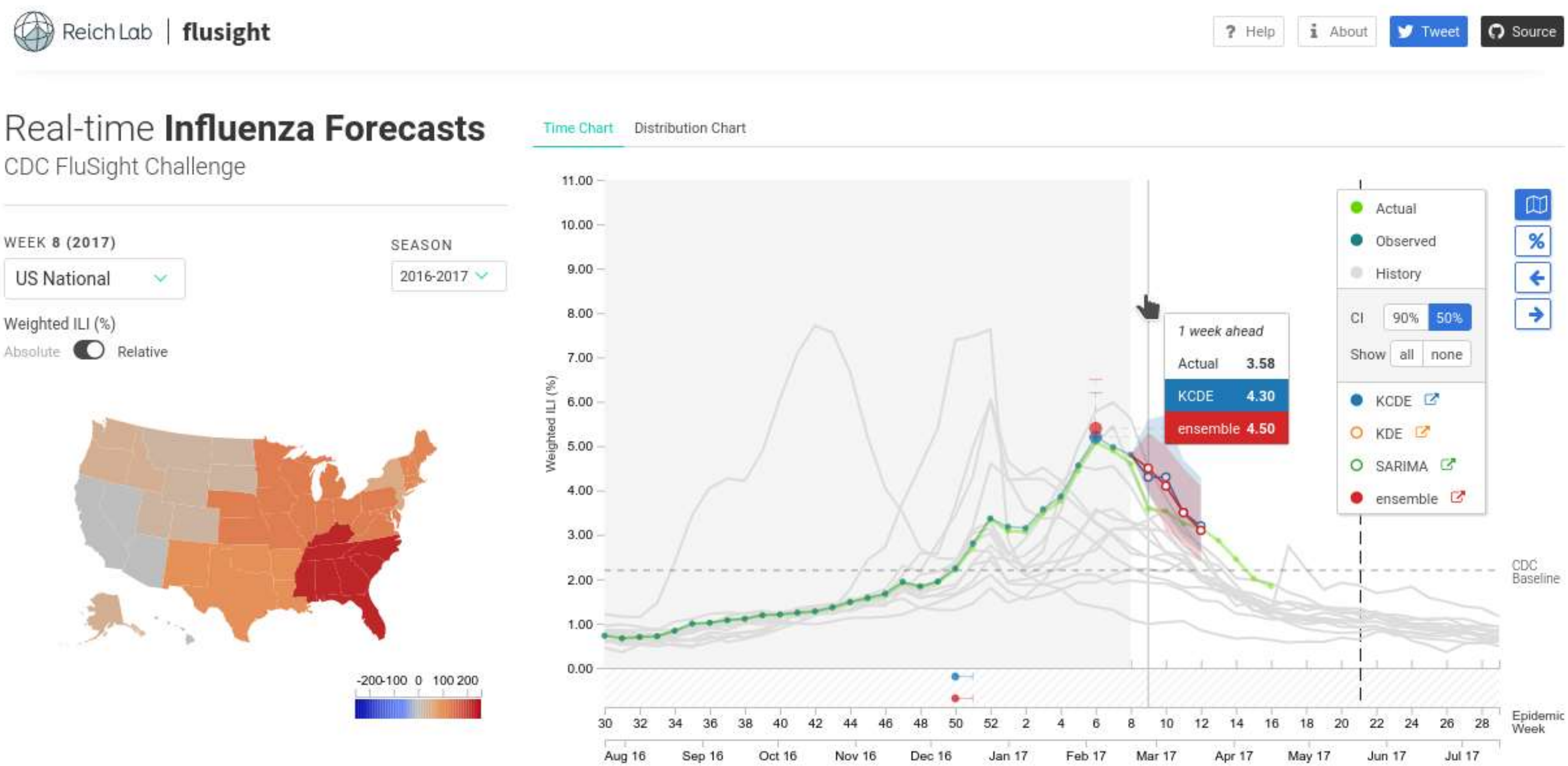
- Topics:
  1. **Collaborative initiatives**
  2. Experiences of individual forecasters
  3. Bridging forecasting with decision making

[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

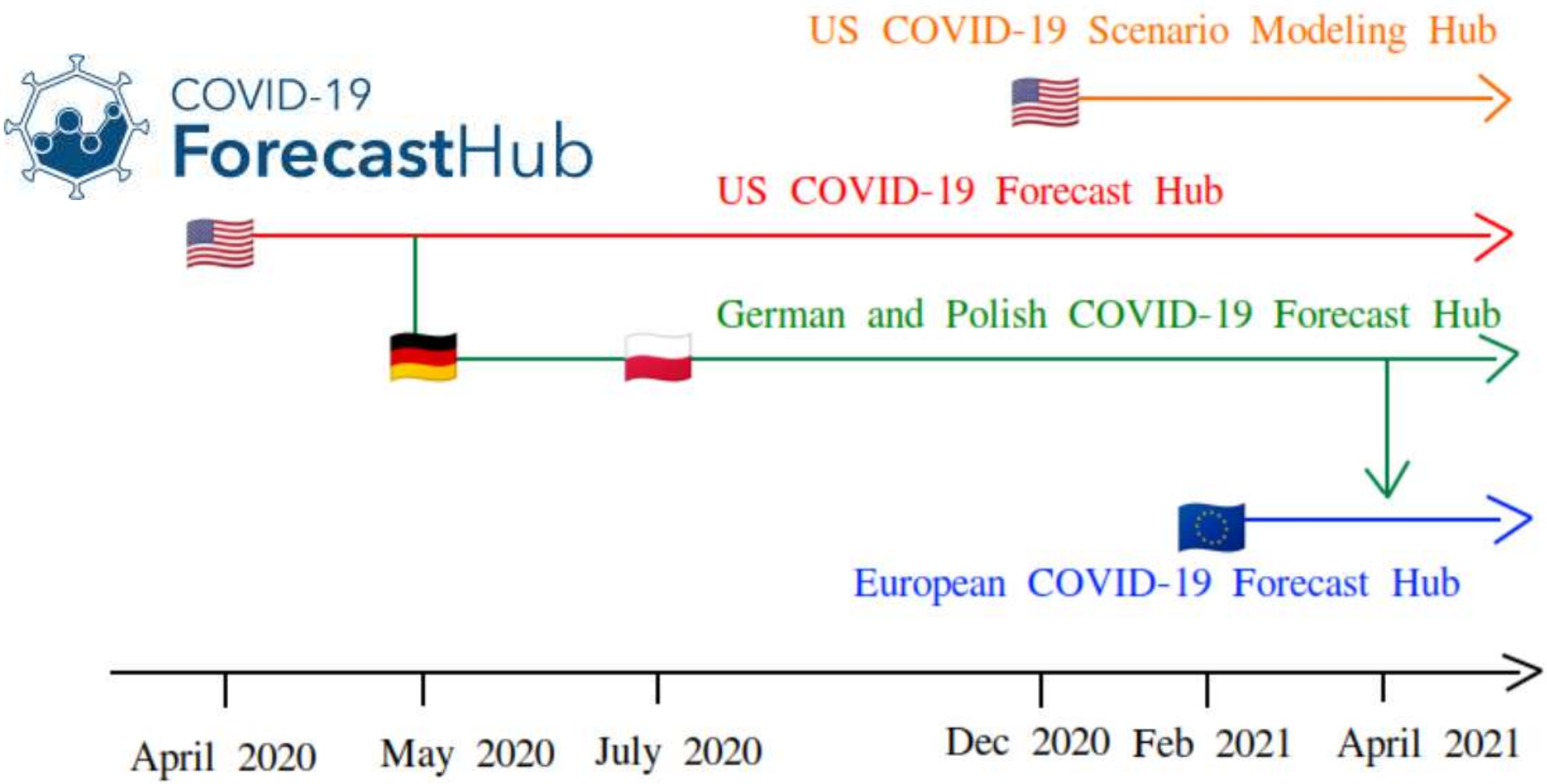
Slide credit: Matt Biggerstaff, US CDC

FluSight Challenge



Source: reichlab.io/flusight/

COVID-19 Forecast Hubs



Source: Johannes Bracher, KIT Karlsruhe and HITS Heidelberg

Standardization efforts of real-time forecast submissions

zoltar About Projects Docs Sign in

A Forecast Archive

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

Learn more »

Project: COVID-19 Forecasts Config

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

COVID-19 ForecastHub

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

[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

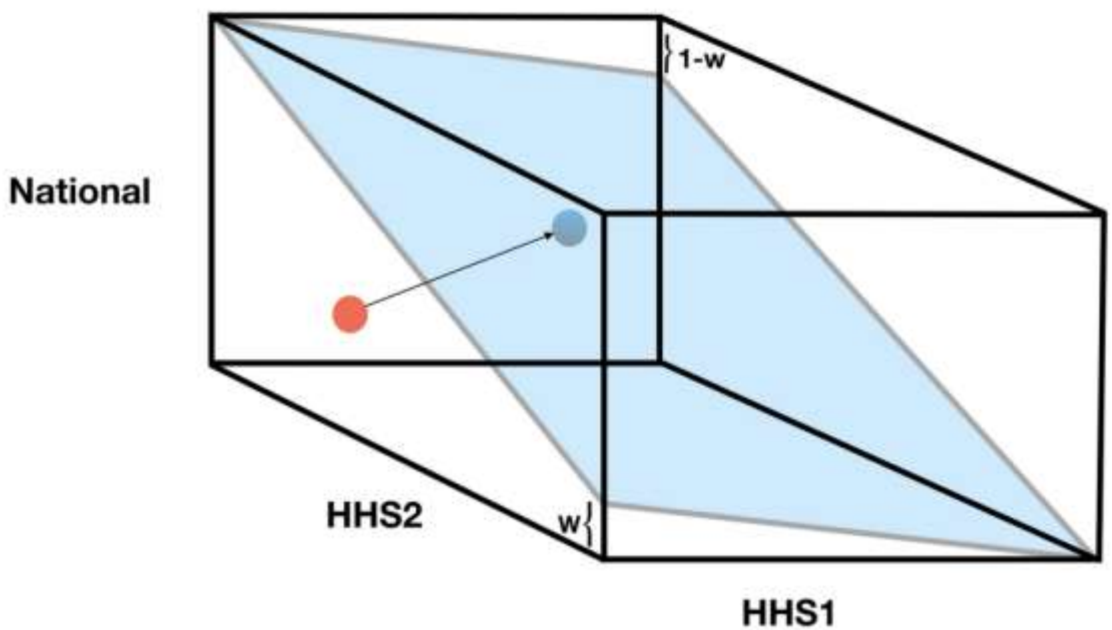
[Kandula+, Royal Society Interface 2018]

- 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

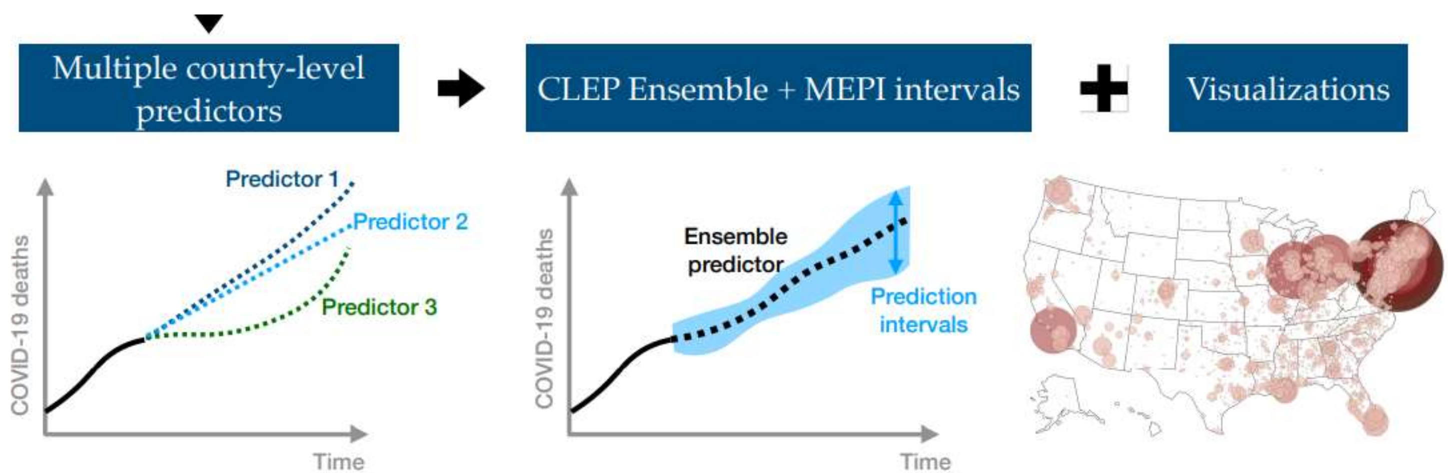
[Gibson+, PLOS Comput. Bio 2021]

- 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

COVID-19 ForecastHub

Top 5 ranked model

[Gibson+, medRxiv 2020]

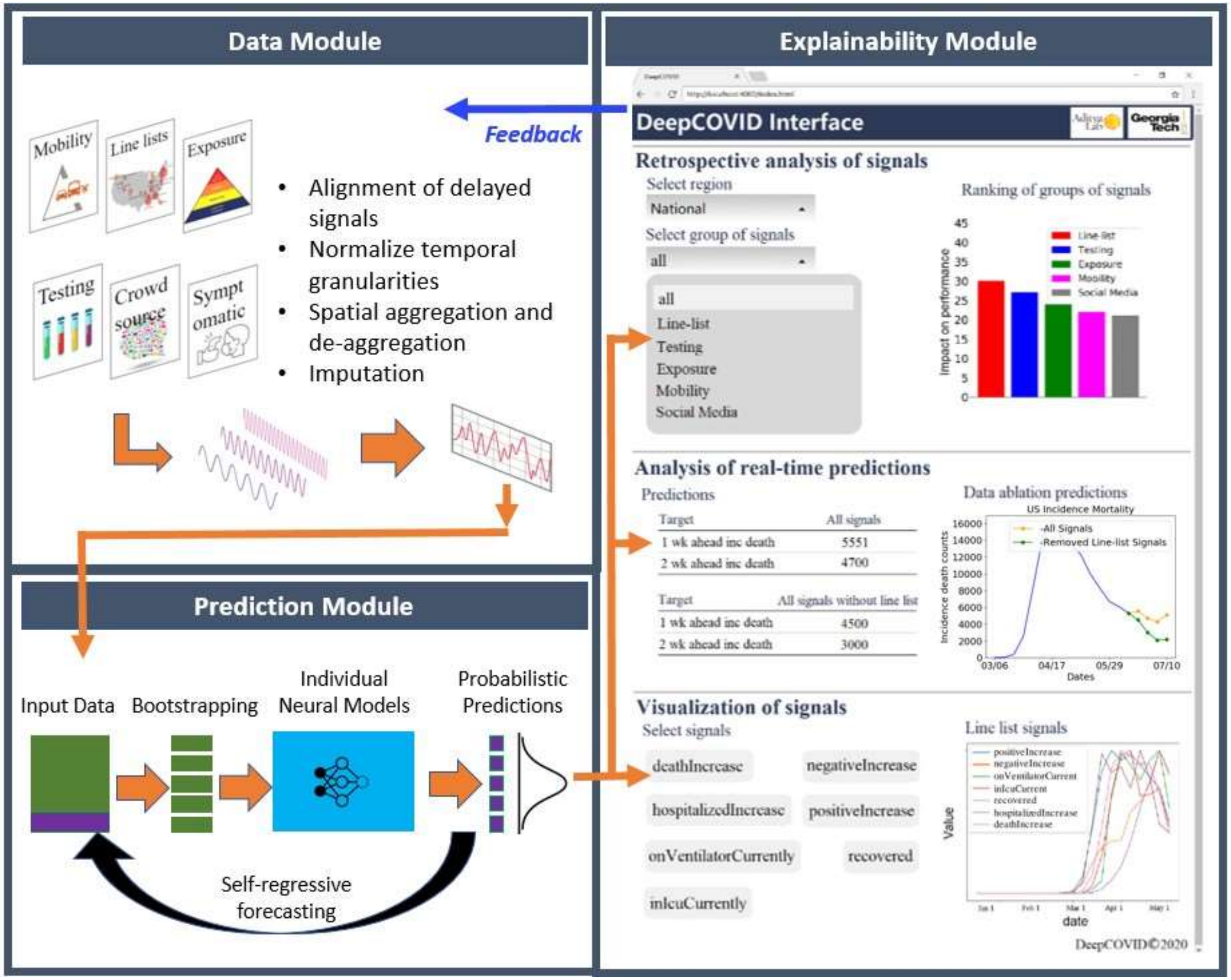
- MechBayes:
  - SEIRD + Bayesian framework w/ informative priors
- Use a quality assurance procedure
- Expert on the loop:
  - Visualize most recent data
- Anomalies:
  - Check notifications from data sources (JHU)
  - Backdistribute when needed

Ex. 5: Operational DL Framework: DeepCOVID

COVID-19 ForecastHub

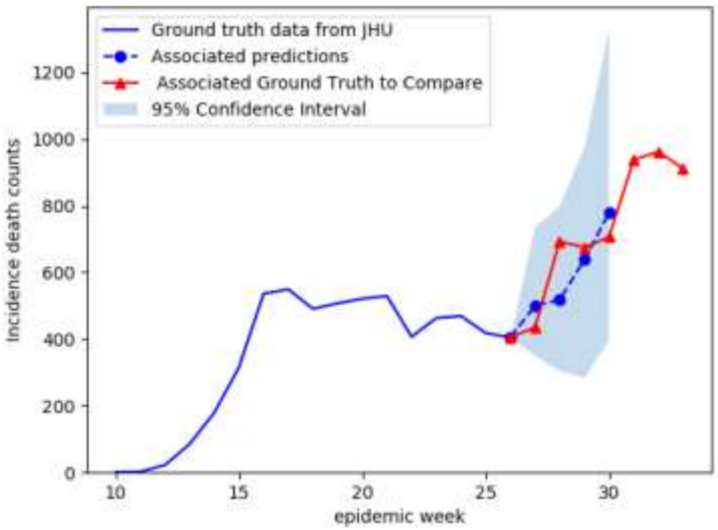
Top 5 ranked model

[Rodríguez+, AAAI 2021]

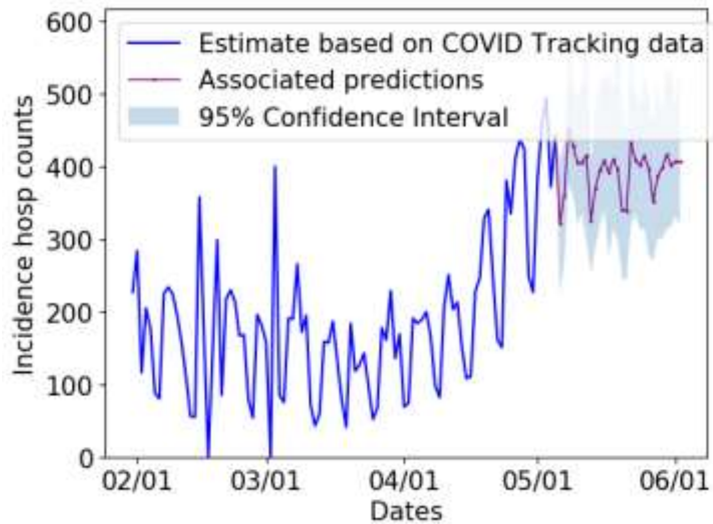


Highlights of results

Anticipate Trend Changes

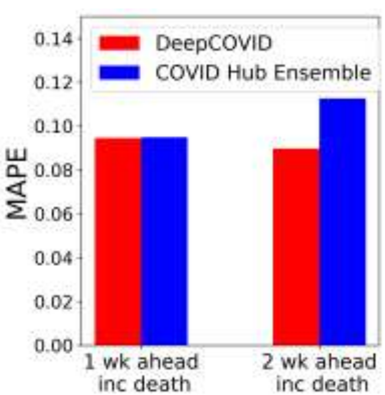


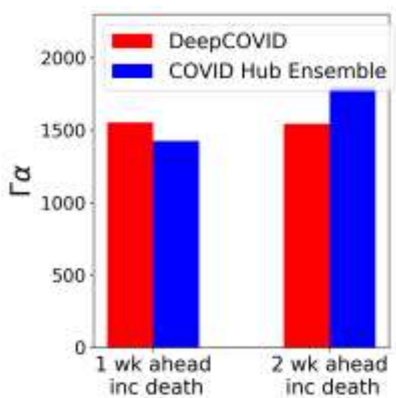
Capture finer-grain patterns



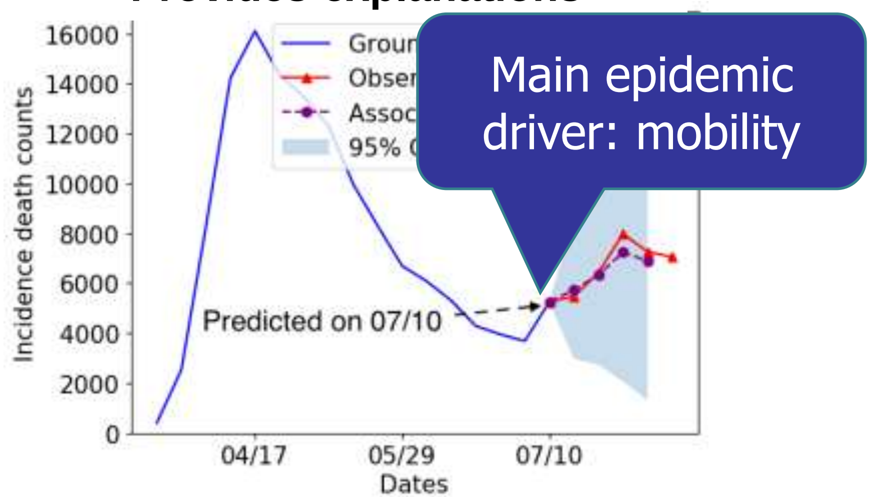
Lower is better

Excels in short-term forecasting





Provides explanations



Rodríguez, Kamarthi, and Prakash 2022

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**

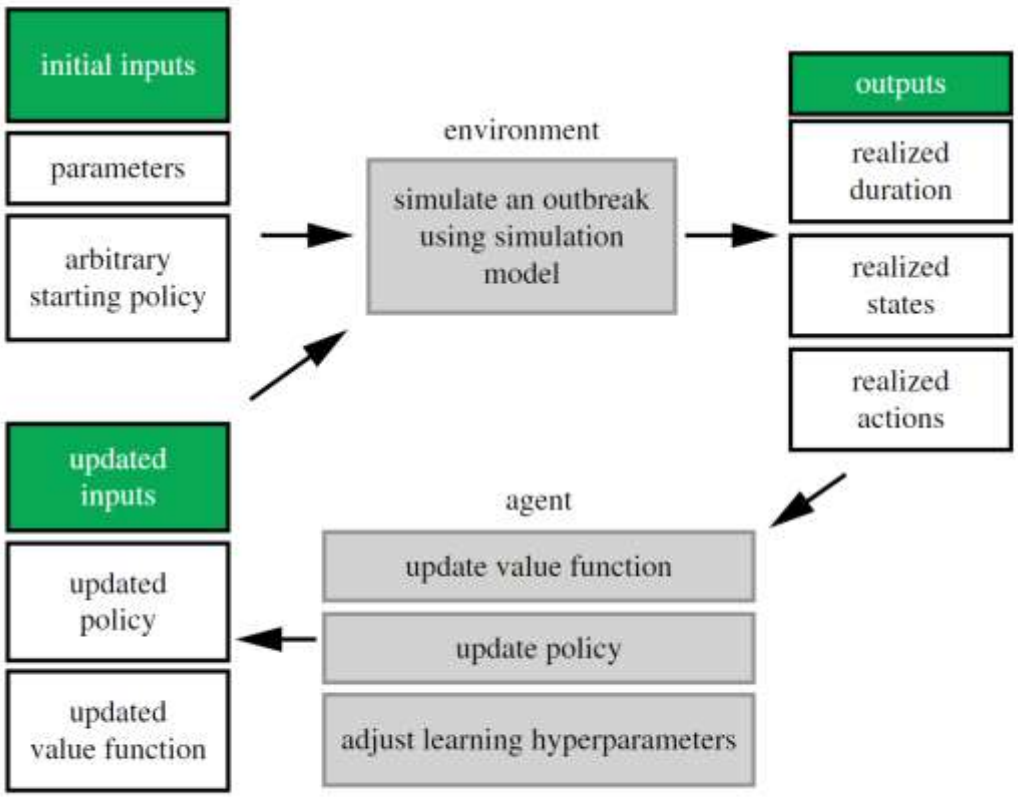
[3] Bridging forecasting with decision making

- 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

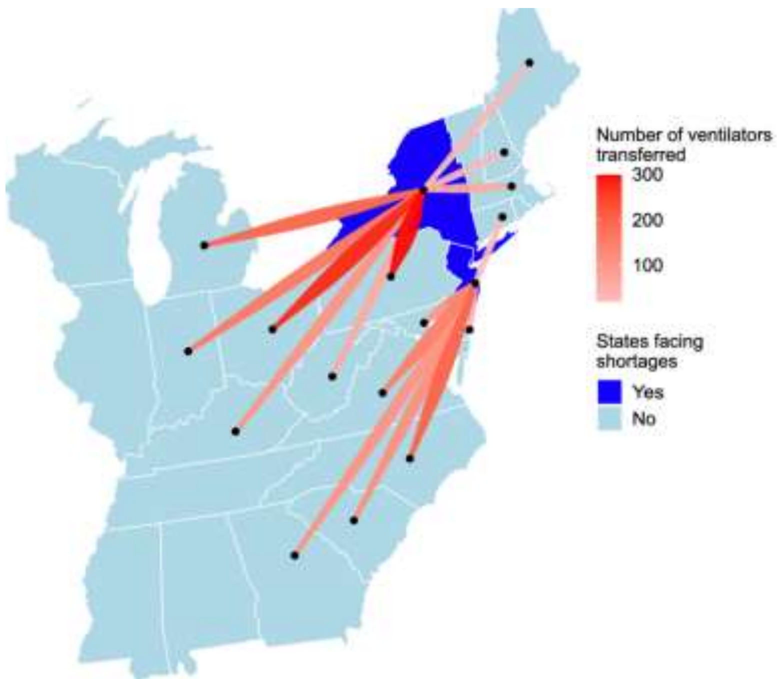
[Probert+ PloS 2018, RS 2019]

- 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



Outline

1. Epidemic forecasting (30 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (55 min)
5. Hybrid models (45 min)
6. Epidemic forecasting in practice (20 min)
7. **Open challenges (20 min)**

- 30 min break after Part 4
  - Feel free to catch us for coffee

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

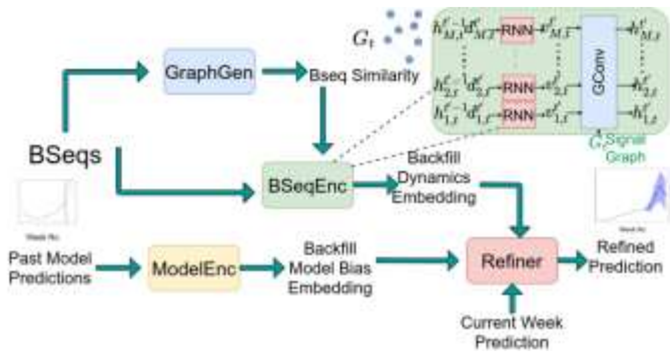
Health data breaches slowing from 2021's record high, report suggests

Published July 19, 2022

Hailey Mensik

[C1] Data quality (Opportunities)

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

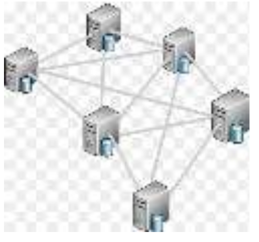


ML correction of data revisions

- 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)

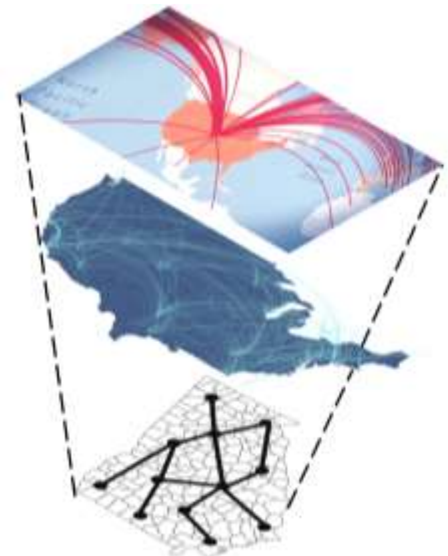
- **Causal ML and reinforcement learning (RL)**
  - Discovering causal relations among mutltivariate data and interventions
  - RL for policy analysis
- Examples:
  - Causal feature selection in time series [Mastakouri and Schölkopf, NeurIPS 2021]
  - COVID-19 testing analysis via RL [Bastani+ Nature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

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

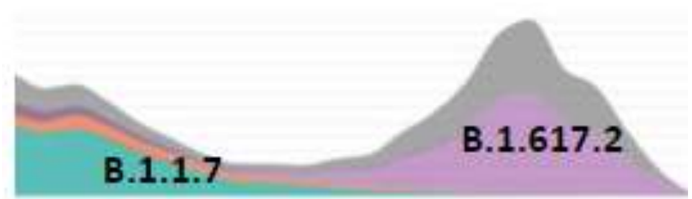
- **Testbed** for data-centered models for scenario analysis

[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

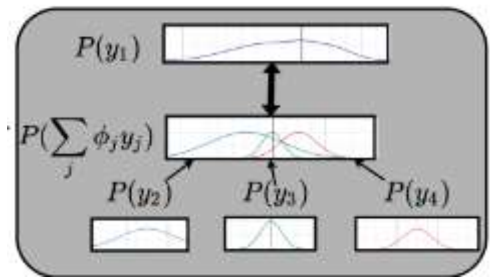


- Incorporate pathogen and behavioral multi-scale dynamics



[C3] Modeling multi-scale dynamics (Opportunities)

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



Coherent forecasts

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

[C3] Modeling multi-scale dynamics (Opportunities)

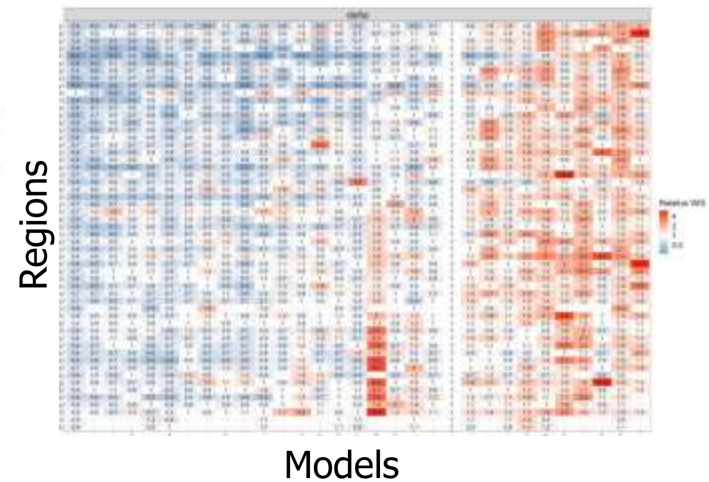
- **Multi-scale modeling**
  - Pathogen dynamics
  - Behavioral models

- Examples:
  - Evolution of pathogens (phylodynamics) [Kraemer+ Science 2021]

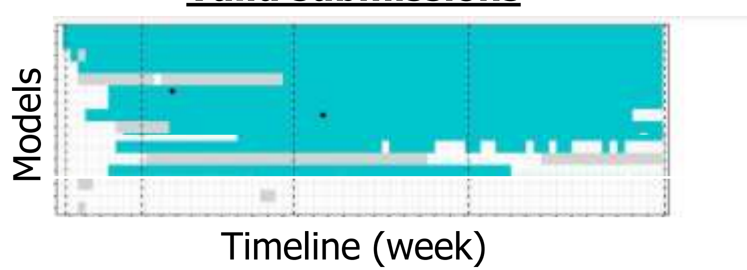
[C4] Improving the ensembles and WoC predictions (Challenges)

- 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



Valid submissions



[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)

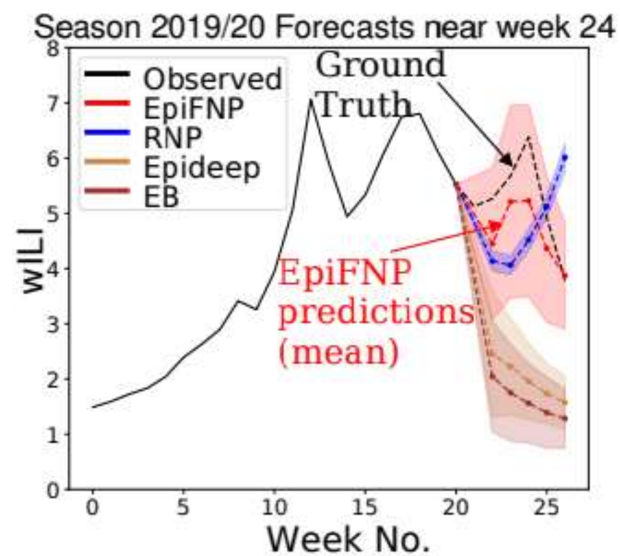
- **HCI for Data Gathering**
  - Capture confidence of prediction, uncertainty bounds
  - Data on multimodal and conditional distributions of beliefs

- Examples:
  - Efficient elicitation for collective crowd answers [Joon+ CSCW 2019]

[C5] Well-calibrated Forecasts (Challenges)

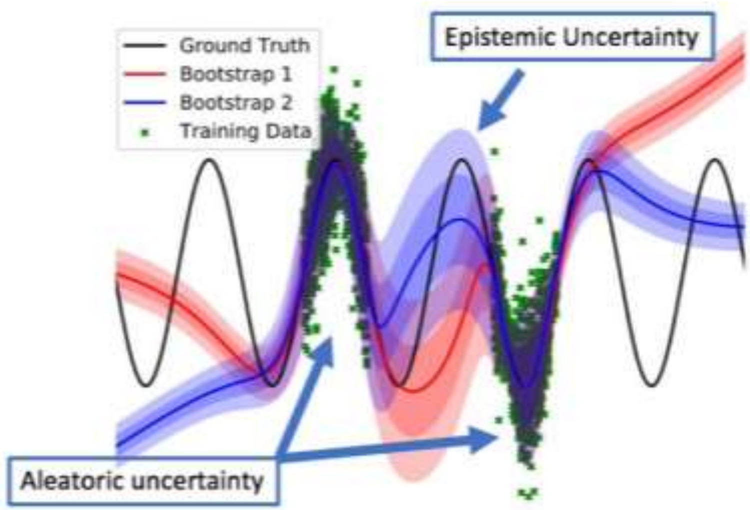
- Uncertainty quantification of forecasts
  - Help at and high-stake decision making

- Helps in explainability during novel/unseen scenarios



[5] Well-calibrated Forecasts (Opportunities)

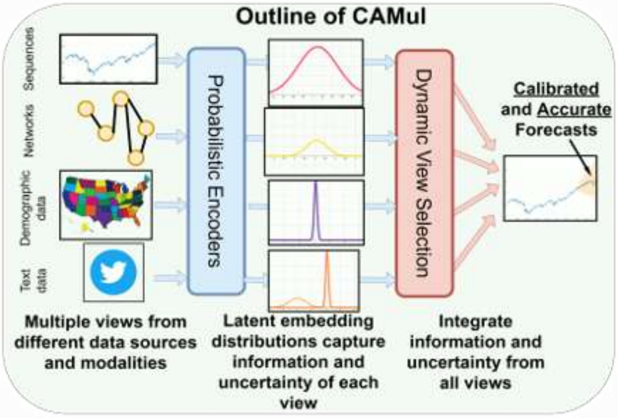
- **Explain sources of uncertainty:**
  - Differentiate aleatoric (data) and epistemic (model) uncertainty



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

[C5] Well-calibrated Forecasts (Opportunities)

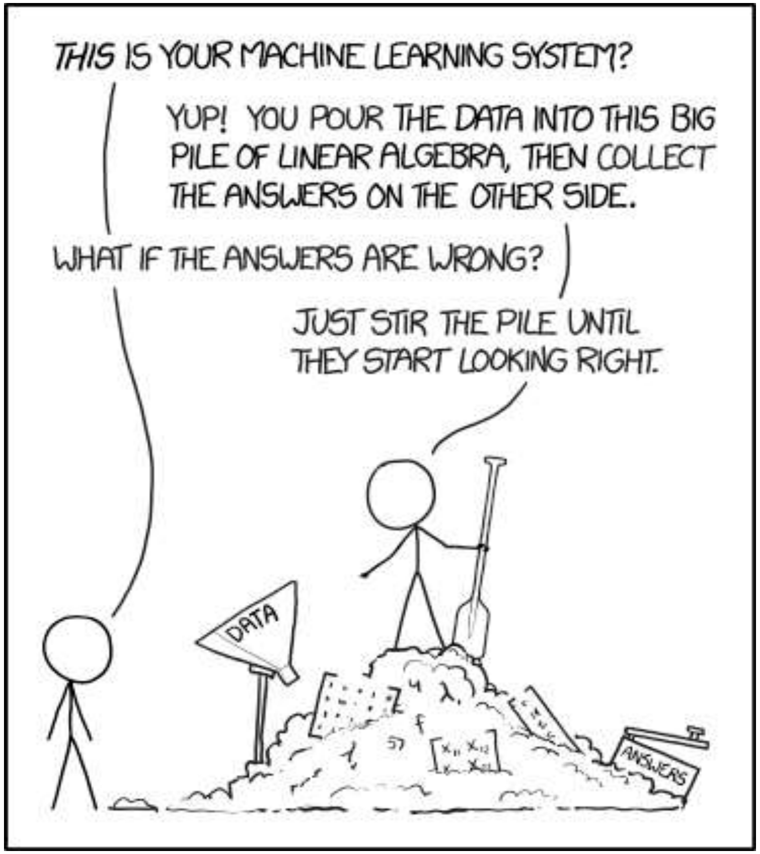
- **Modeling multiple sources of uncertainty**
  - From multiple modalities
  - Useful also in anomaly detection, eval. of data quality
- Examples
  - Neural Gaussian Process for multiple sources of uncertainty [Kamarthi+ WWW 2022]



[C6] Explainable Forecasts (Challenges)

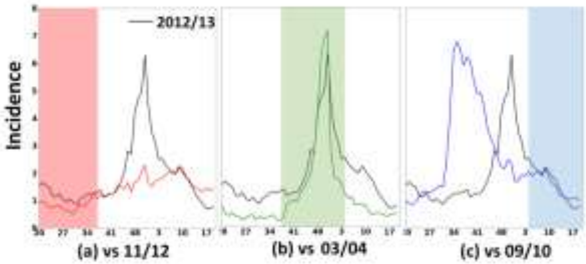
- Key to bridge decision making with forecasting

- 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

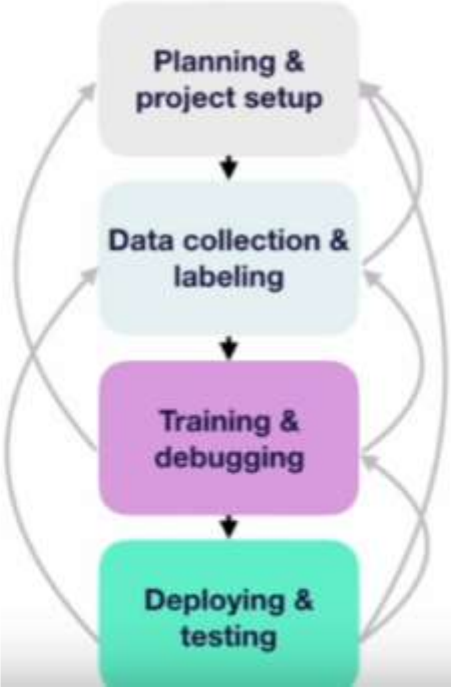


- Examples:
  - Explainable neural representations, Saliency Maps [Molnar 2020]
  - Similarity with past data points [Adhikari+ KDD 2019, Kamarthi+ NeurIPS 2021]
  - Feature-level importance [Rodriguez+ AAAI 2021]

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

Courtesy [Full Stack Deep Learning (FSDL) 2022]

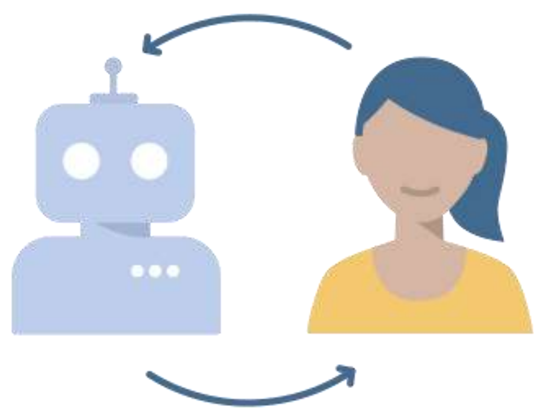


[C7] Technical debt of real-world deployment (Opportunities)

- **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]

Outline

1. Epidemic forecasting (30 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (55 min)
5. Hybrid models (45 min)
6. Epidemic forecasting in practice (20 min)
7. Open challenges (20 min)

- 30 min break after Part 4
  - Feel free to catch us for coffee

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?

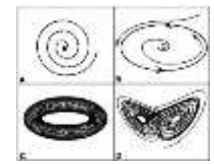
- What measures should the government and people take?
  - Pharmaceutical interventions: vaccinations, anti-virals, other therapeutics
  - Non-pharmaceutical interventions: social distancing, closing schools and workplace, using masks, hand hygiene in hospitals
  - Allocate and distribute medical equipment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

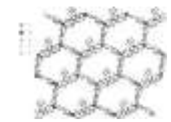
Data + Models + Efficient Algorithms + Simulations + Optimization tools

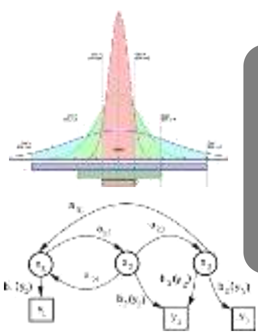
Why data science?

- 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.

facebook

Reminder on Tutorial Webpage

- github.com/AdityaLab/kdd-22-epi-tutorial
- All Slides posted there.
- Talk video as well (later).

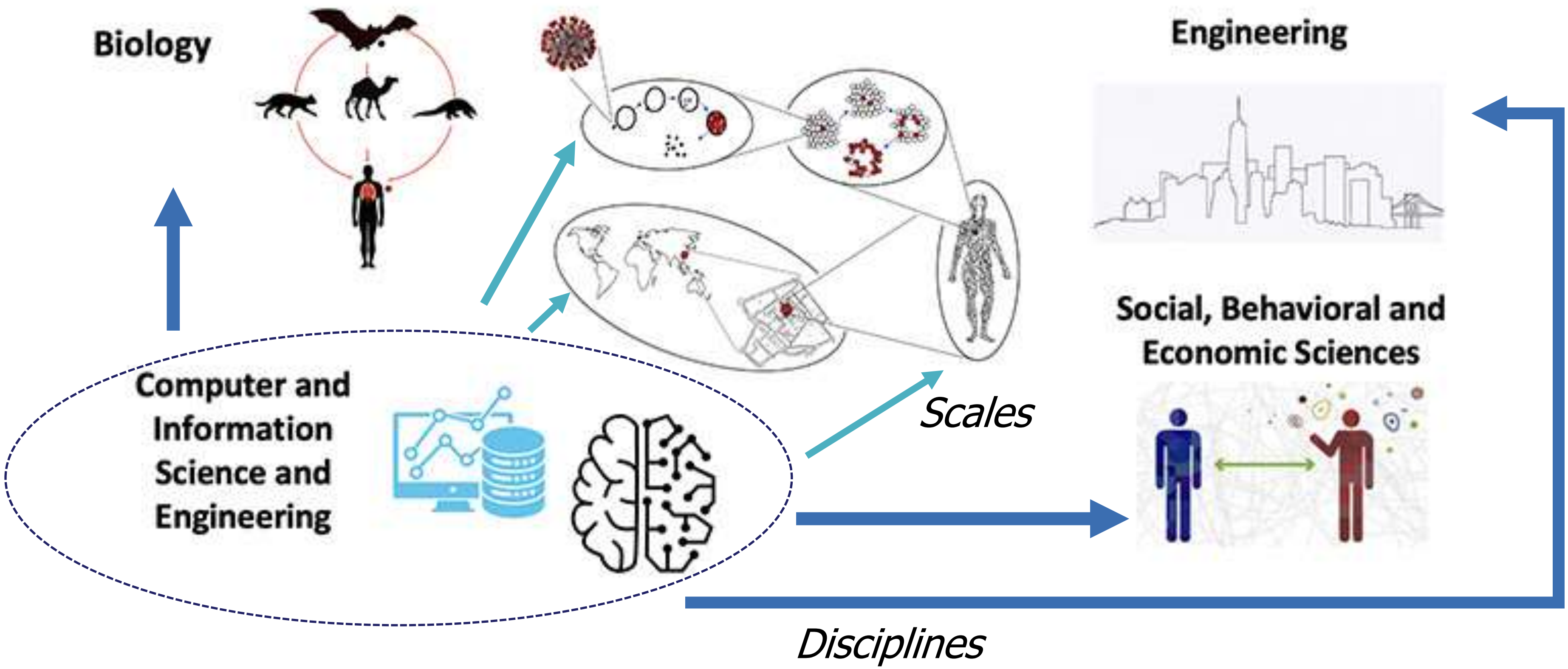
- **License**: for education and research, you are welcome to use parts of this presentation, for free, with standard academic attribution. For-profit usage requires written permission by the authors.

Stay tuned

- Epidemiology meets Data Science Workshop
  - https://epidamik.github.io/
  - Hosted **tomorrow** at KDD 2022
  - Keynotes:
    - Rachel Slayton (CDC)
    - Bryan Wilder (CMU)
    - Cecile Viboud (NIH)

epiDAMIK @ KDD 2022

- 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)

Fill survey: forms.gle/HbLP8Y8QMC27Yg8EA

Also available in github.com/AdityaLab/kdd-22-epi-tutorial

Thanks!

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