Data-Centric Computational Epidemic Forecasting

Alexander Rodríguez
Harshavardhan Kamarthi
B. Aditya Prakash

College of Computing
Georgia Institute of Technology

August 14, 2022

About us

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

About us

- Alexander Rodríguez
  - 5th year PhD student, graduating July 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: cc.gatech.edu/~acastillo41/

About us

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

Tutorial Webpage

KDD22 Tutorial: Data-Centric Epidemic Forecasting

Survey paper companion: PDF

Slides PART 1: PDF

Slides PART 2: PDF

Website: adityalab.cc.gatech.edu/talks/22-kdd-epi-tutorial.html

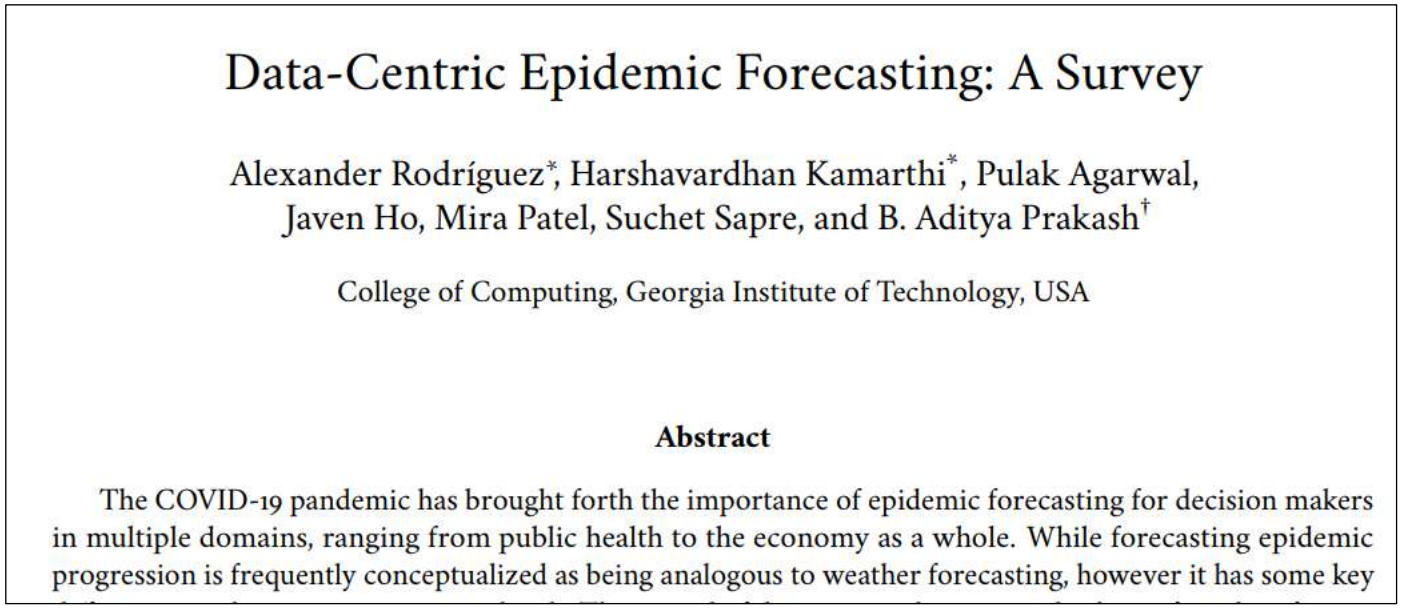
Tutorial abstract

The recent COVID-19 pandemic has brought forth the importance of epidemic forecasting to equip decision makers

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

Survey companion arxiv.org/abs/2207.09370

- Tutorial largely based on recent survey paper



Data-Centric Epidemic Forecasting: A Survey

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

College of Computing, Georgia Institute of Technology, USA

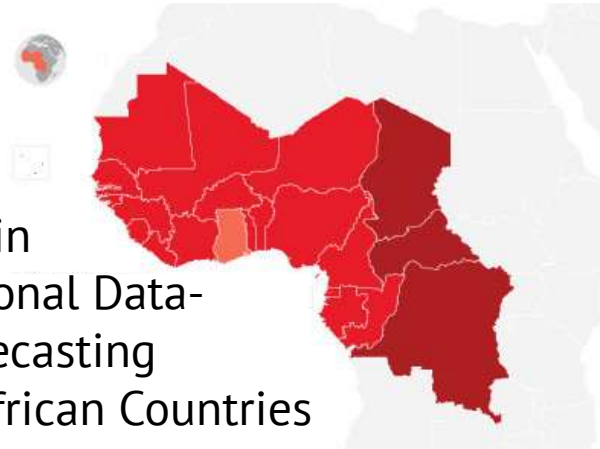
Abstract

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

All citations in this tutorial can be found there

Data-centric epidemic forecasting for practitioners

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



Workshop in Computational Data-driven Forecasting for West-African Countries

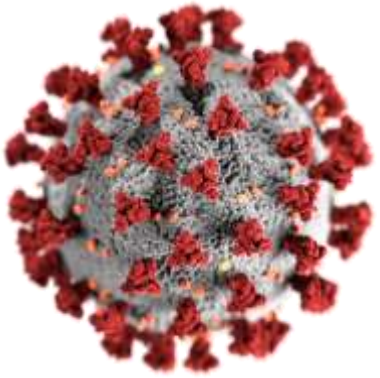
Outline

1. Epidemic forecasting (30 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
   - 30 min break, feel free to catch us for coffee
4. Statistical/ML/AI models (55 min)
5. Hybrid models (45 min)
6. Epidemic forecasting in practice (20 min)
7. Open challenges (20 min)

Plan for the Tutorial

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

COVID-19 pandemic



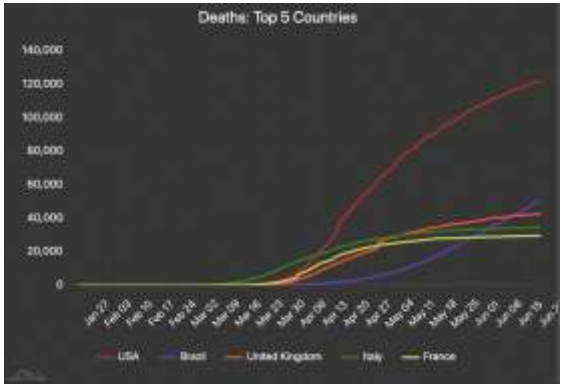
- **Global pandemic**
  - 500+ million cases
  - 6+ million deaths
- Hard to imagine any aspect of life not been affected
- Never before have epi. concepts captured public attention and imagination so vividly! Examples:
  - Reproduction number
  - Non-pharmaceutical interventions
  - Social distancing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

COVID-19 trajectories

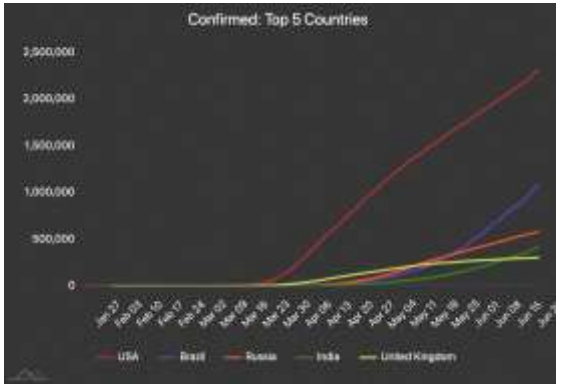
Global spatial incidence distribution

Spatial incidence distribution in USA

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



Cumulative Mortality



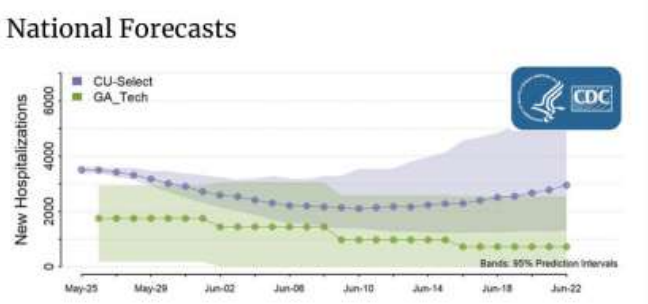
Confirmed cases Cumulative

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

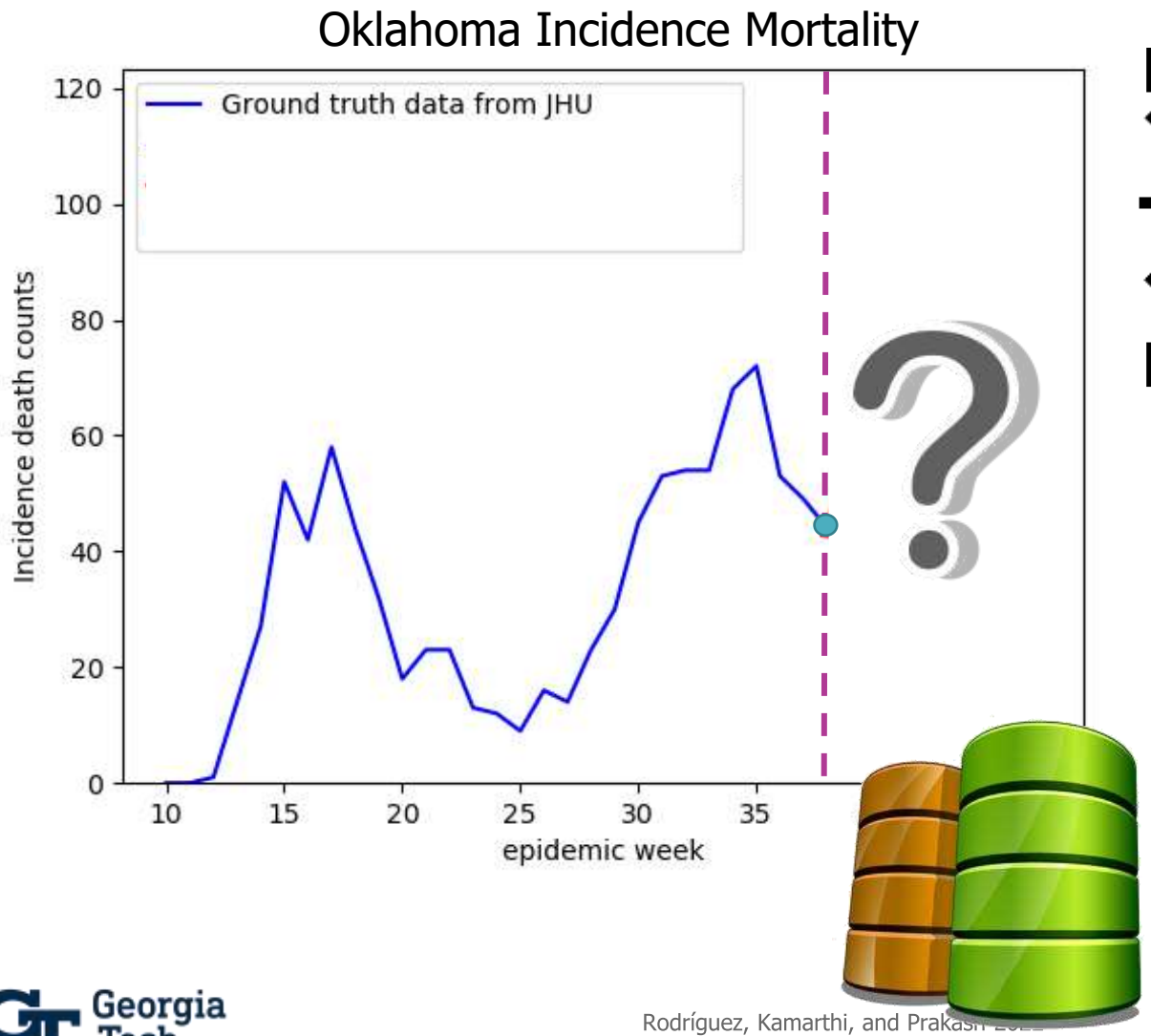
Why Forecasting?

An outlook to the future allow communities to

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



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

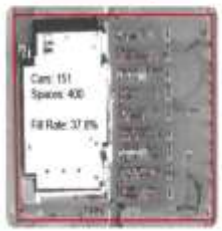
Increasing data collection

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

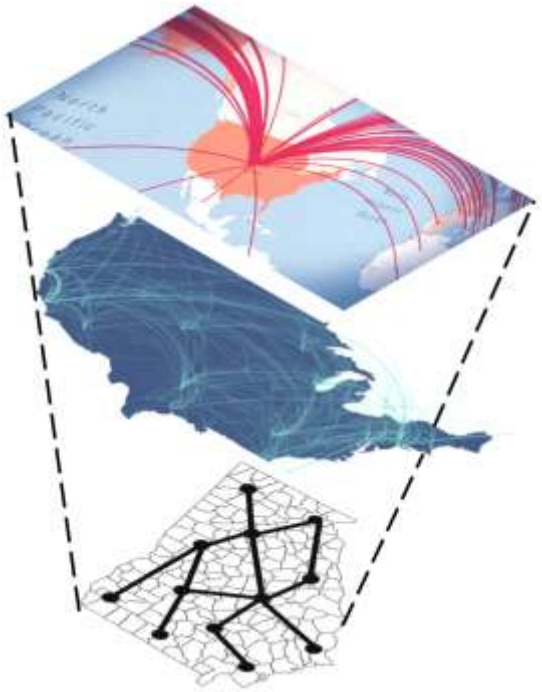
Medical record

GAATTCATACCAGATCACCGGATTCCCGACTCCAAATGTGTCCCCCTCACAC
TCCCCCGATTACCGTCTTCTGCTCTTAGACCACTCTACCCTATTCCCCACACT
CACCGGAGCCAAAGCCGCGGCCCTTCCGTTCCGATTACCGAAAAGACCCCA
CCCGTAGGTGGCAAGCTAGCTTAAGTAACGCCACTTCGATTAACGAGGAAA
AATACATAACTGACCTATTATCGAGTTCAGATCAAGGTCAGGAACAAAGAA
ACACCGATTACCGTAACCGTAAGATARTGGTATCGATACGTAGACAGTTTA

Estimating "fill rate" in parking lots



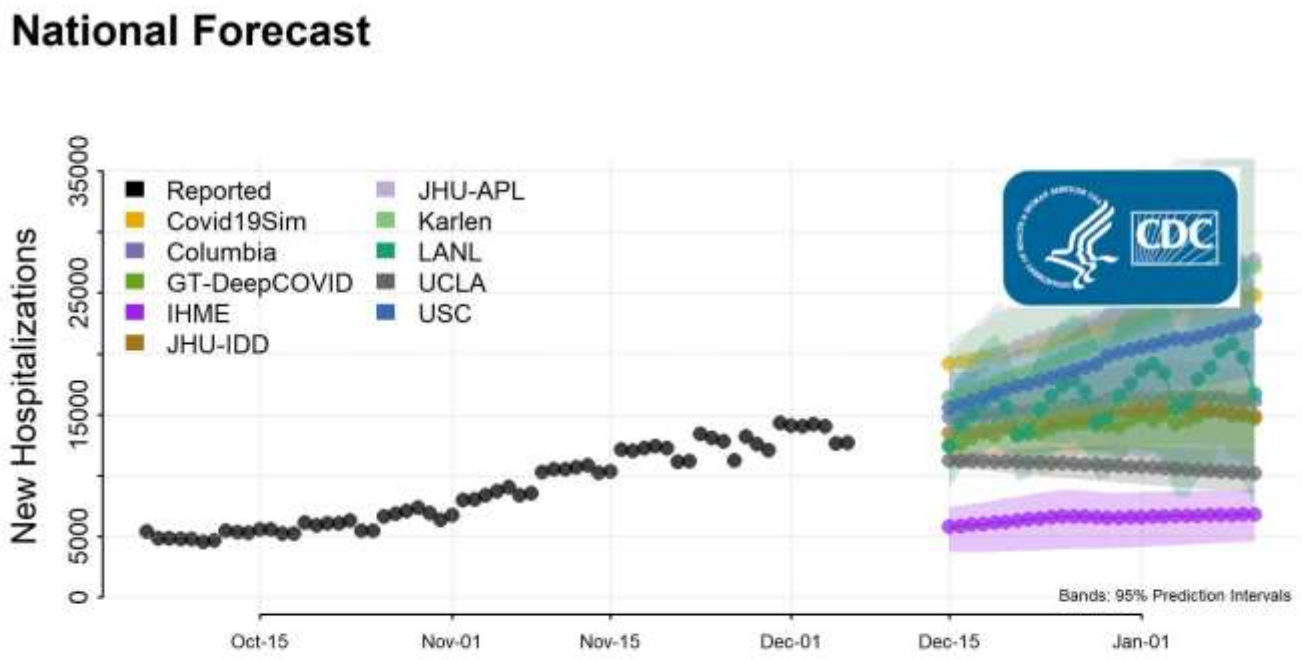
Why Computational Data-centered Forecasting?



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

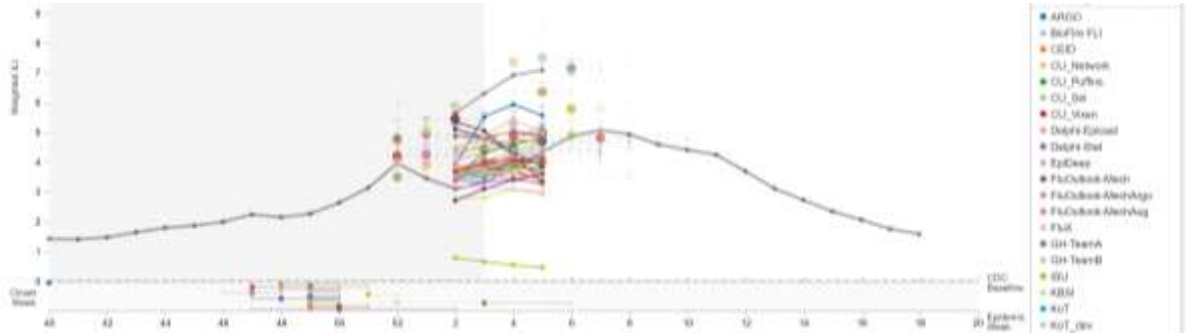
Our approach

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



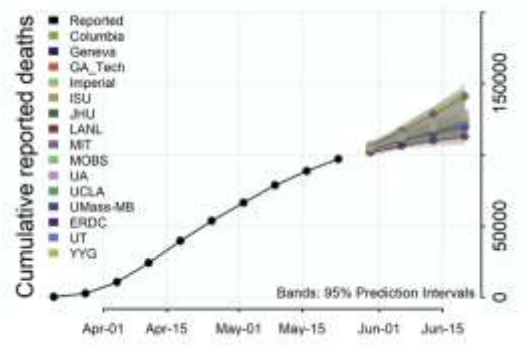
Our Participation in CDC Forecasting Initiatives

Target 1: Influenza like illness per week



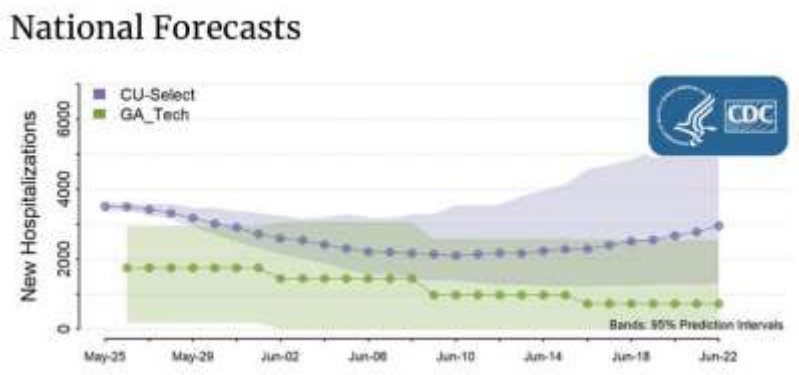
Last few years
Also in COVID-ILI (March 2020)

Target 2: Weekly Covid Mortality



Since April End 2020

Target 3: Daily Covid Hospitalizations



Our Impact

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

1 of 11 shown on their page

1st Prize

facebook

Carnegie Mellon University

Out of 115 global participants

2nd Prize

C3.ai COVID-19 Grand Challenge

43
Countries

777
Participants

AdityaLab @ Georgia Tech

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

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

Recent Publications

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

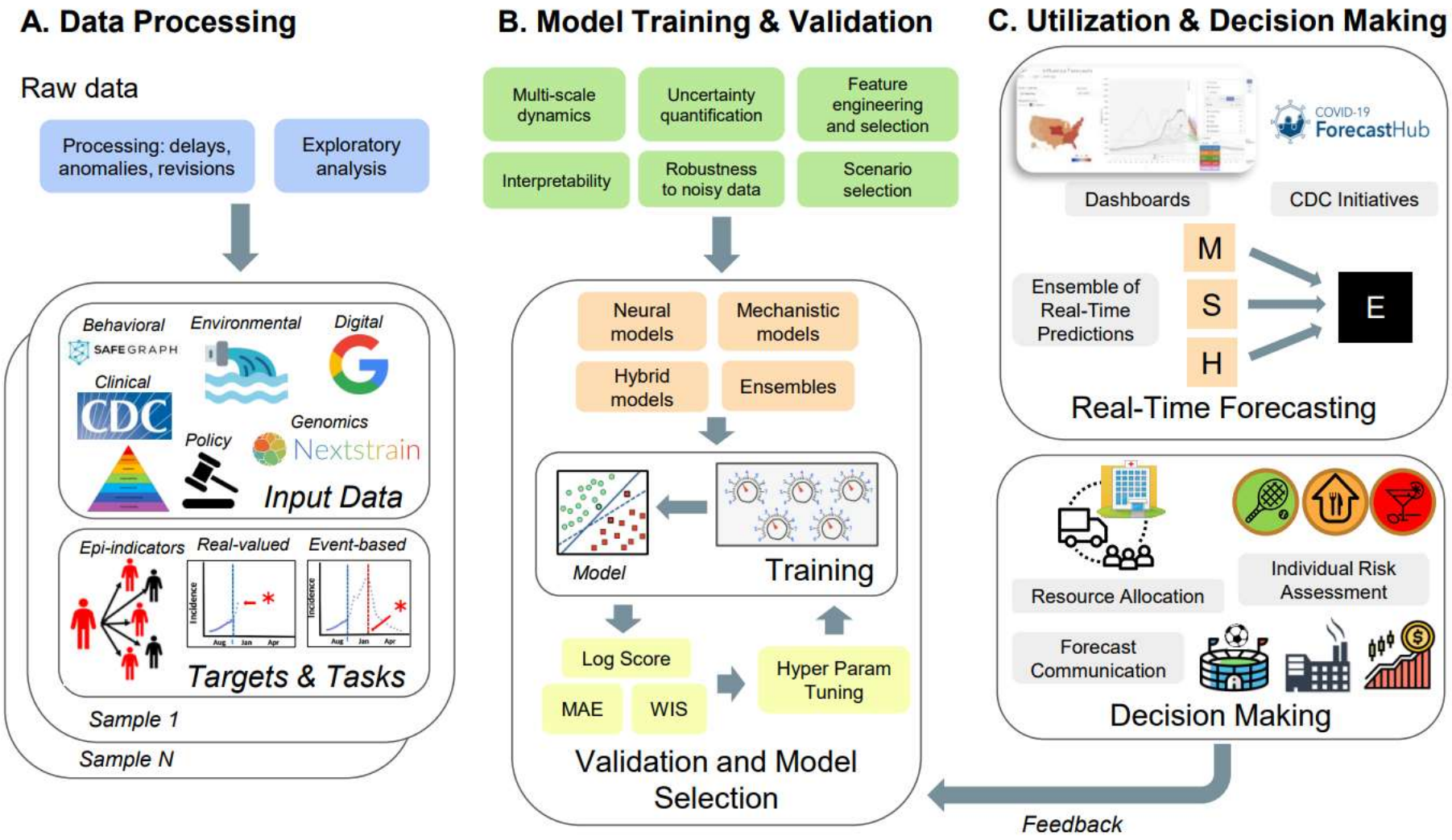
Outline

1. **Epidemic forecasting (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 1: Epidemic Forecasting

Epidemic Forecasting Pipeline

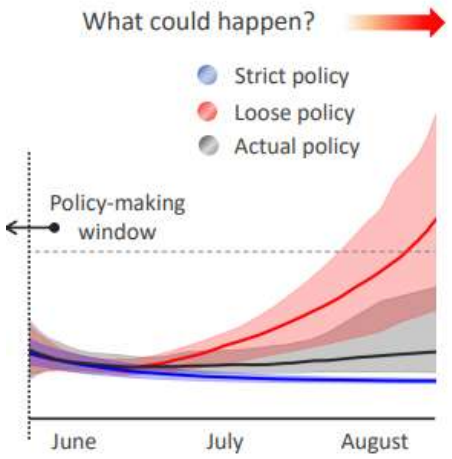


Epidemic Forecasting Setting

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

Preliminaries: Projections vs Predictions

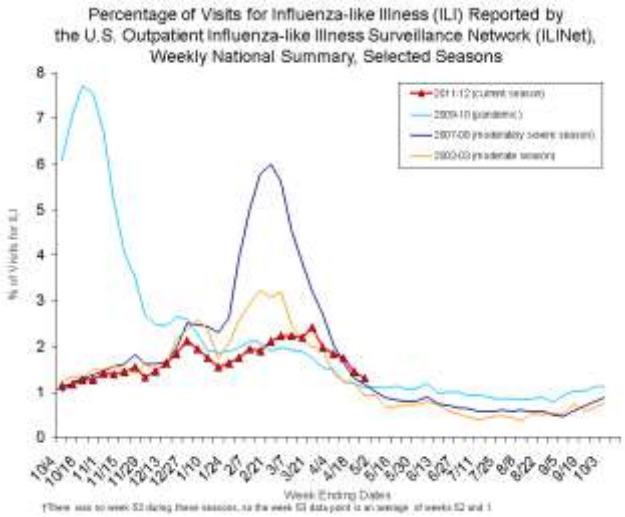
- Projections:
  - Outcomes for **specific scenario**. E.g.: Mask Mandates, Lockdowns
  - Require mechanistic assumptions/domain knowledge
- Predictions:
  - **Most likely outcome** based on past data.
- This tutorial: Mostly focus on predictions
  - But models from projections can be extended to predictions



Courtesy of [Qiann+ NeurIPS 2020]

Preliminaries: Prediction Seasons

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



Epidemic Forecasting Setting

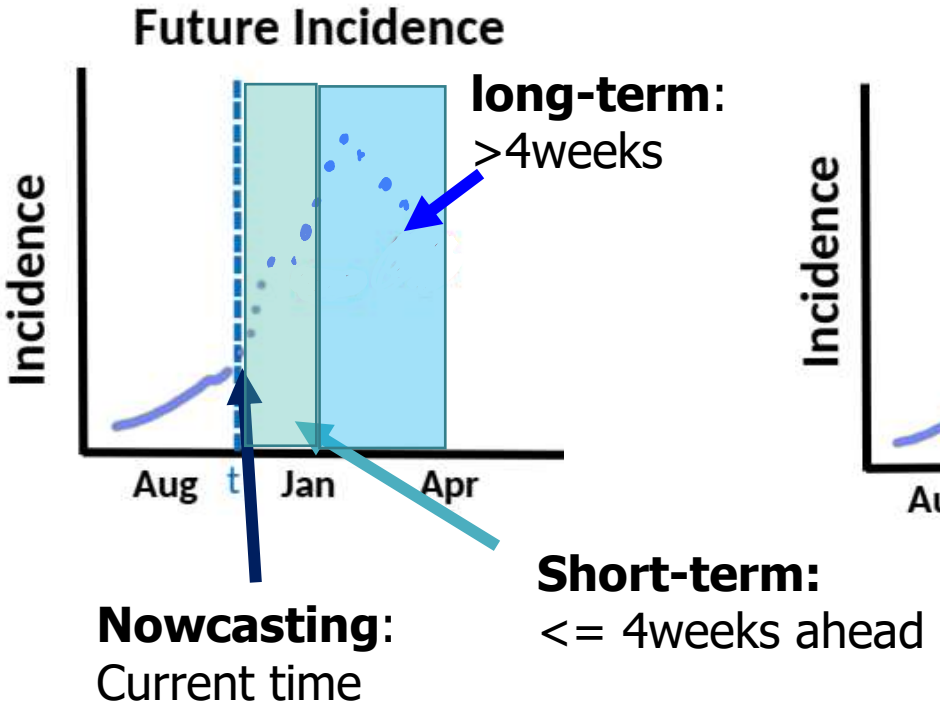
1. **Forecasting Tasks**
2. Targets of interest
3. Spatial and temporal scales
4. Datasets
5. Model evaluation

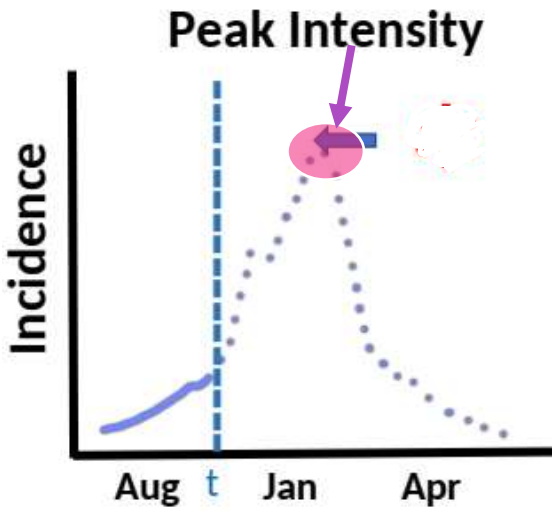
Different forecasting tasks

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

[1.1] Real-valued predictions

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



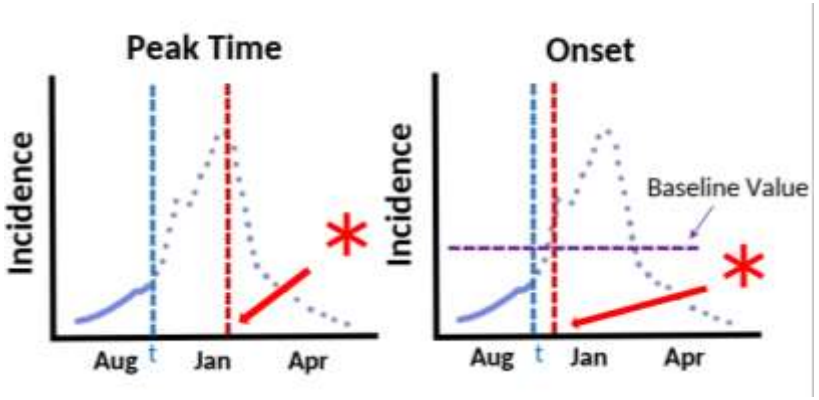


Why predicting the current value (nowcasting)?

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

[1.2] Event-based predictions

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



[1.3] Epidemiological Indicators

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

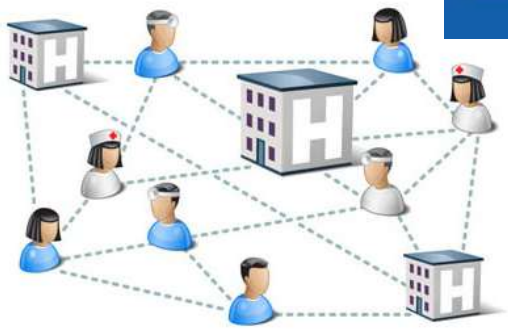
Epidemic Forecasting Setting

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

[2] Targets of Interest

Details of target data sources discussed later

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



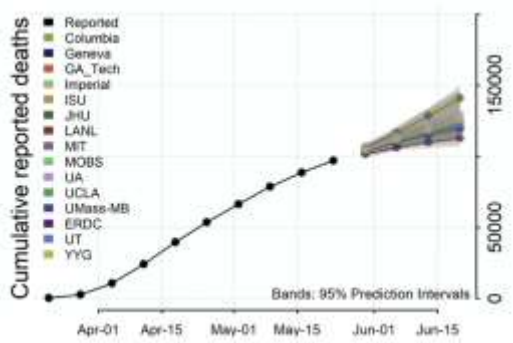
ILINet surveillance network

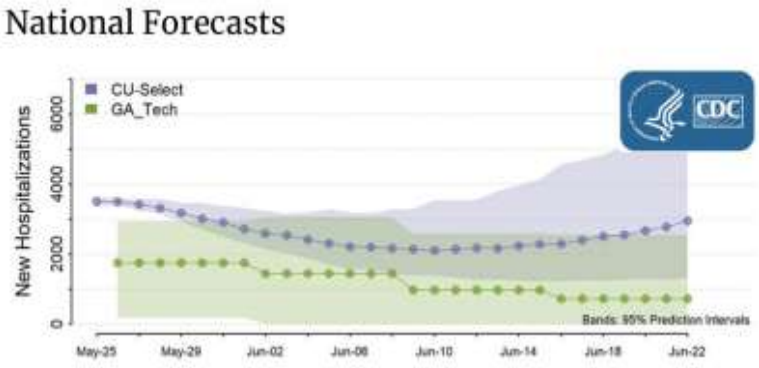
Epidemic Forecasting Setting

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

[3] Spatial and Temporal Scales

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





Epidemic Forecasting Setting

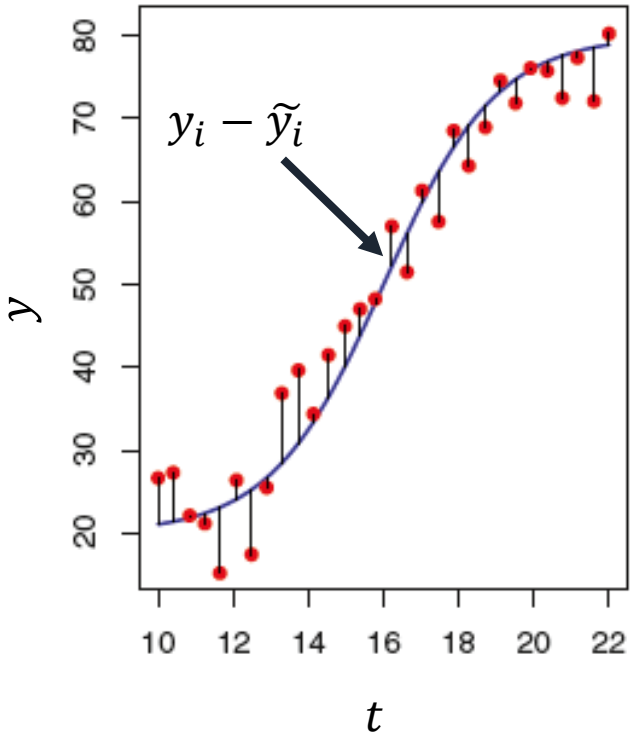
1. Forecasting Tasks
2. Targets of interest
3. Spatial and temporal scales
4. **Model evaluation**
5. Datasets

[5] Success metrics

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

Evaluation of point forecasts

- RMSE: $\sqrt{\frac{\sum_{i=1..T}(y_i - \widetilde{y_i})^2}{T}}$
- 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,…



Eval. of probabilistic forecasts (1)

- Consider both accuracy and uncertainty of distributions/confidence intervals
- Log Score:
  - Log probability of ground truth outcome (binned)

$$\frac{1}{T}\sum_{i=1}^{T}\ln(p_i(y_i))$$

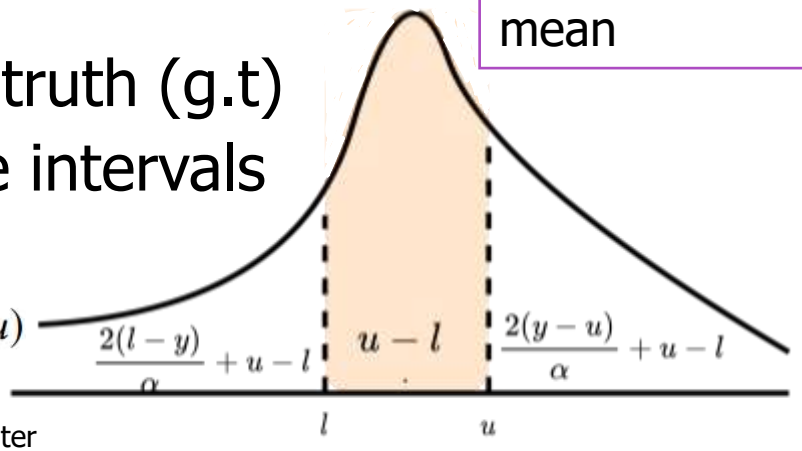
  - Each term clipped at -10 for stability and interpretability

Eval. of probabilistic forecasts (2)

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

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

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



Penalize if distribution is too flat (low confidence)

Penalize if g.t is less than lower bound

Penalize if g.t is greater than upper bound

- Weighted interval Score [Bracher+ 2021]
  - Aggregates for multiple α

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

Eval. of probabilistic forecasts (3)

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

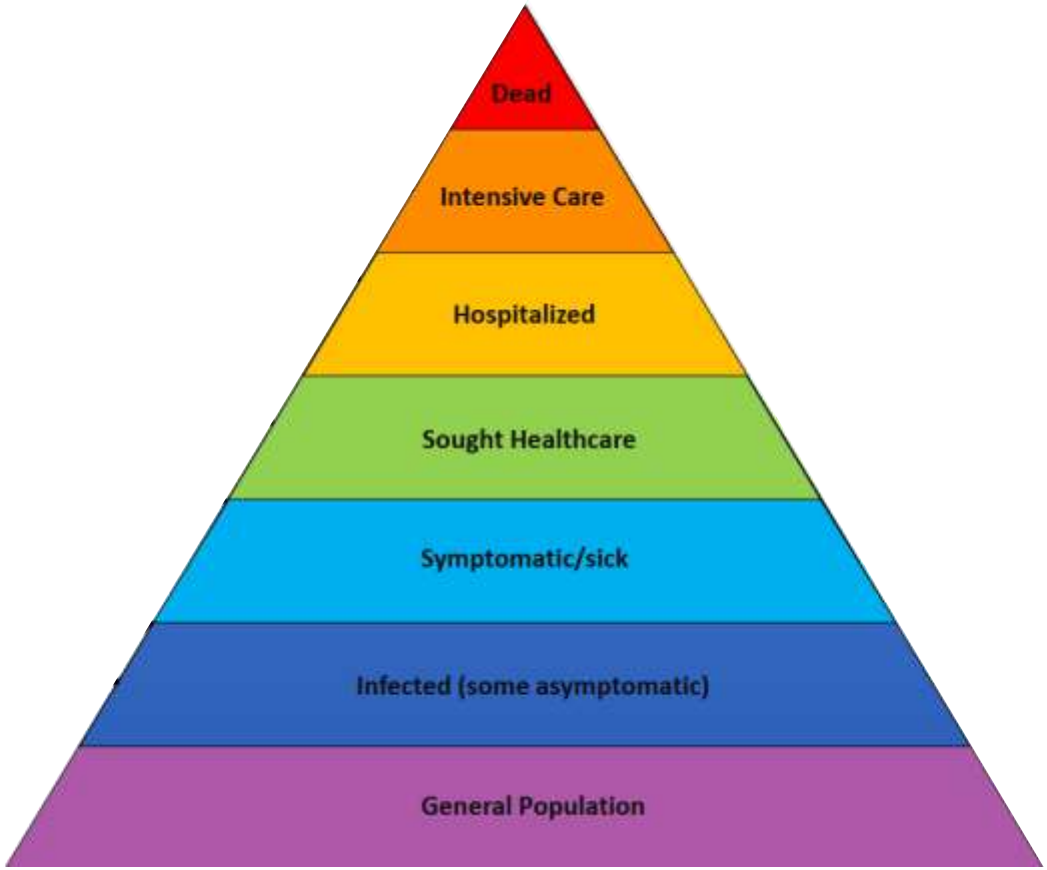
How to choose eval. metrics?

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

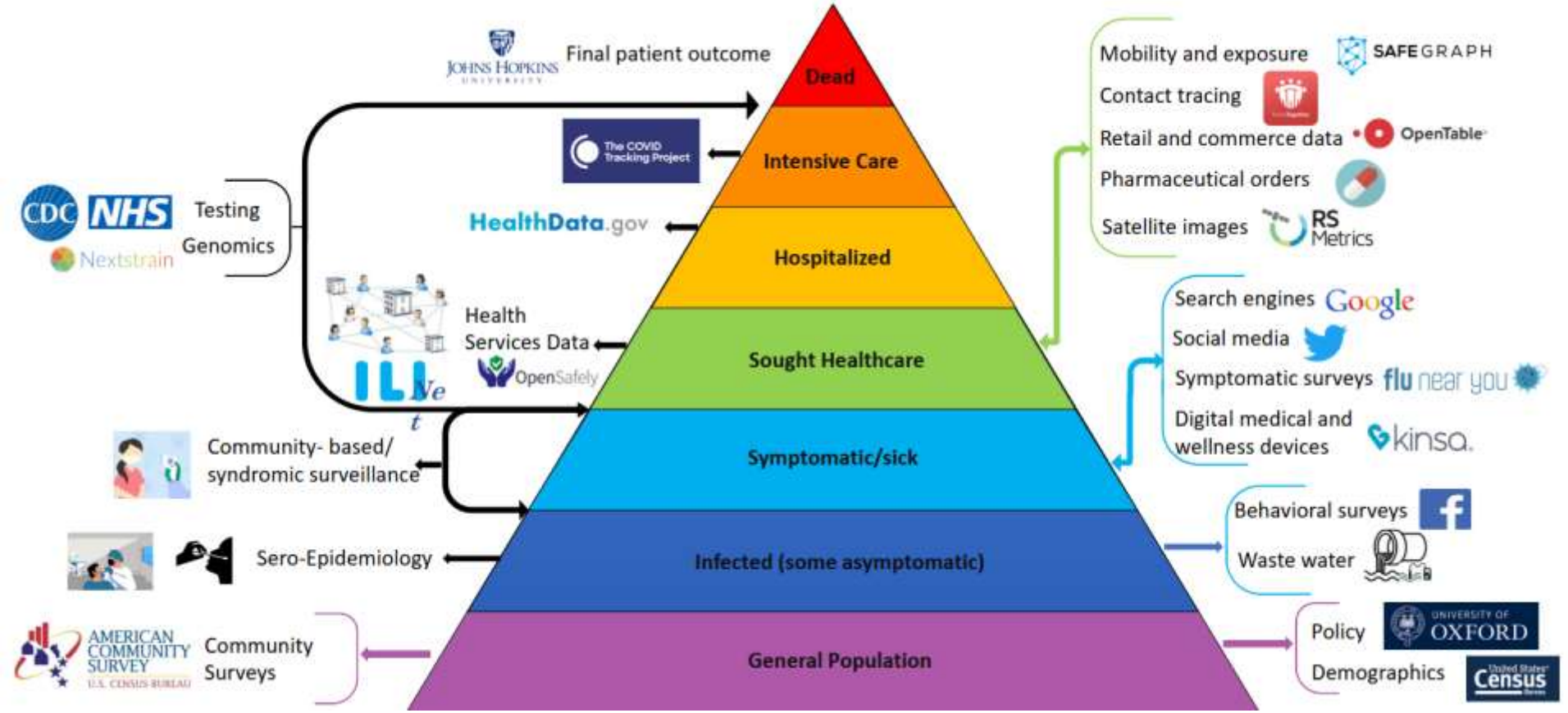
Epidemic Forecasting Setting

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

[4] Datasets: surveillance pyramid



Surveillance pyramid and datasets



Sources of Data

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

(1) Line-list data

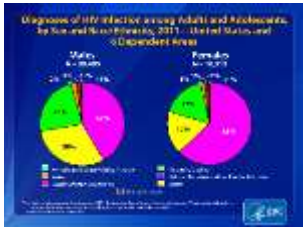
- Who, when and where a person was infected

Hospital records

Lab surveys

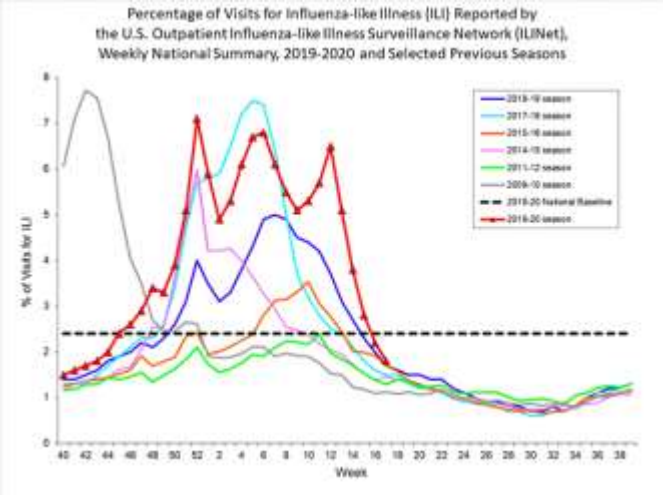
Population surveys

Surveillance Reports



(2) Health services records

- Aggregate records collected from health-service providers
- Include inpatient and outpatient records
- E.g.: ILINet
  - Influenza-Like Illnesses (ILI) from out-patients
  - Collected by CDC and aggregated along 8 HHS regions



Traditional data sources

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

(3) Electronic Health Records (EHR)

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

New data sources: Sky is the limit

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

Digital epidemiology

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Review

Digital Epidemiology

Marcel Salathé[1,2]*, Linus 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

(4) Digital Surveillance: Search Engines

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

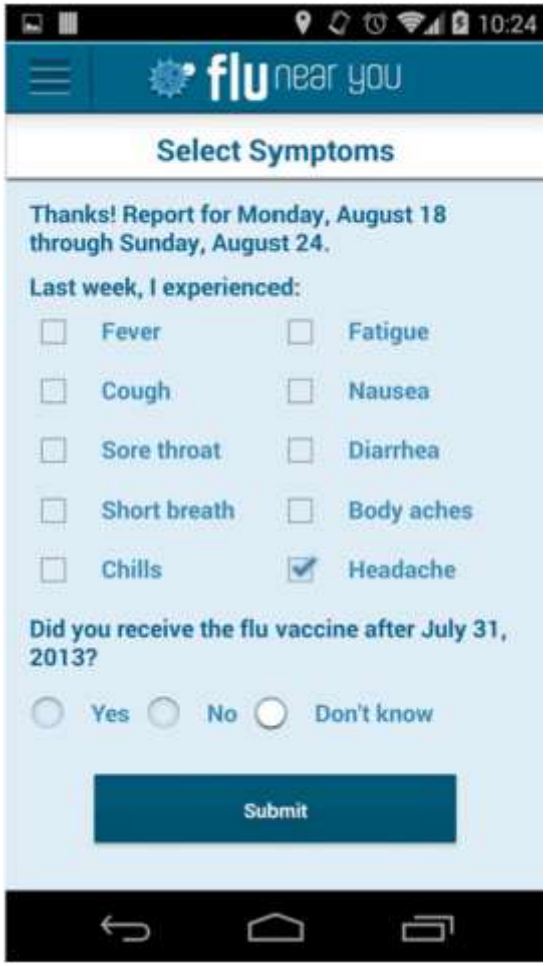
WIKIPEDIA
The Free Encyclopedia

(5) Digital Surveillance: Social Media

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

(6) Digital Surveillance: Online Surveys

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

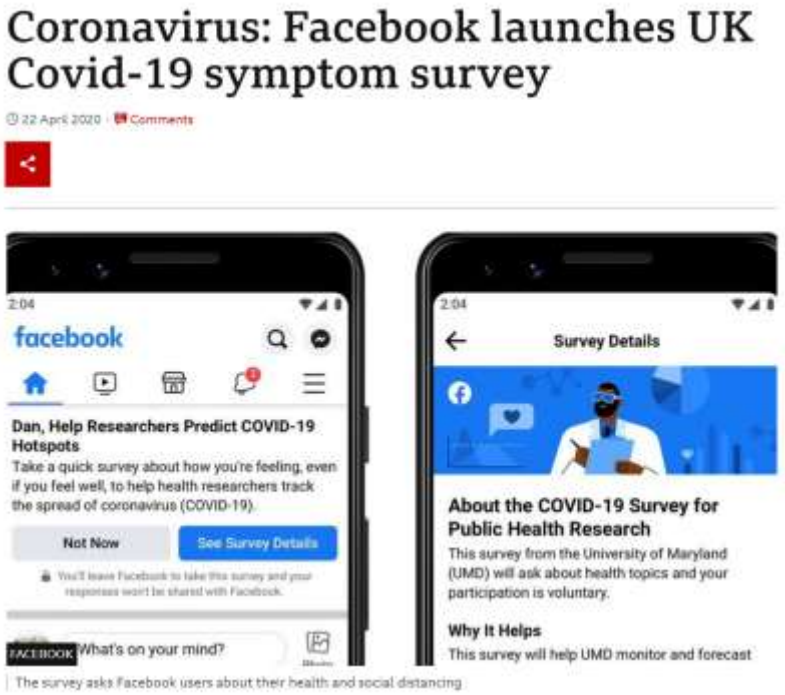


Pros/cons of digital surveillance

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

(7) Behavioral Data: Digital Surveys

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



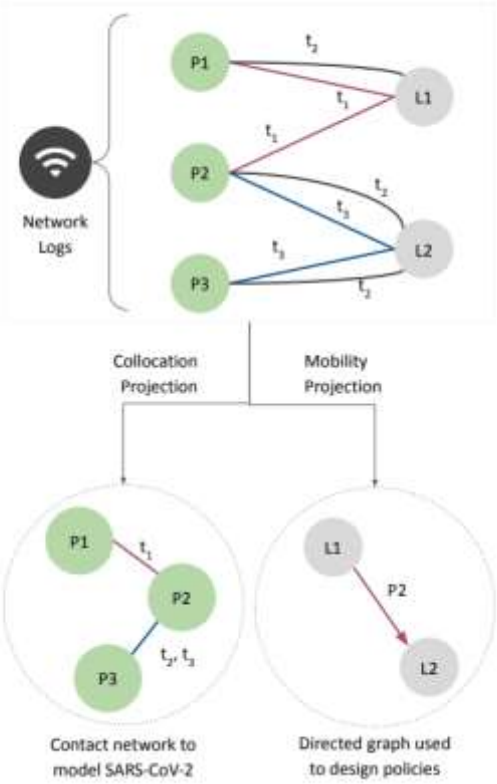
(8) Behavioral Data: Mobility

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

SAFEGRAPH

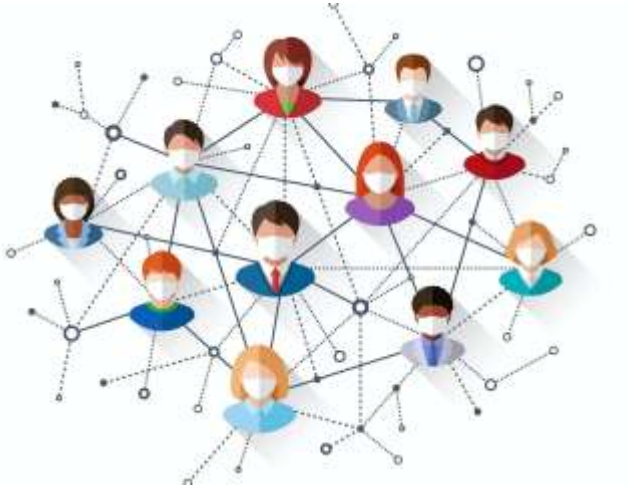
(9) Behavioral Data: Contact Tracing

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



Mobility and Contact Tracing: Pros & Cons

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



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

Researchers find way to fix privacy flaw

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

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

By Herlyn Kaur
Posted Wed 18 May 2022 at 7:04pm

(10) Satellite Images

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

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

(11) Genomics

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

(12) Environmental Sources

- Meteorological
  - Temperature, humidity, etc. Influence transmission
- Zoonotic
  - Track diseases born in animals and transmitted to humans (E.g.: Bats for Covid-19)
  - E.g.: Microsoft Premonition project (for mosquitoes)
- Wastewater
  - Study genetic remnants (RNA) from wastewater sludge
  - Useful for early detection of outbreaks [Peccia+ 2020 Nature]
- Pro: Doesn't require extensive human involvement, cost-effective in long-run
- Cons: Require cutting-edge infrastructure

Epidemic Forecasting Setting

1. Forecasting Tasks
2. Targets of interest
3. Spatial and temporal scales
4. Datasets
5. **Success metrics**

Outline

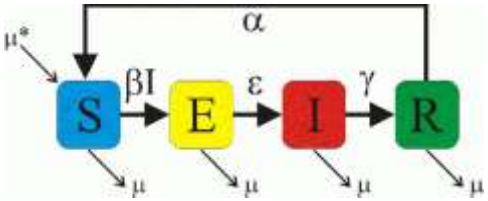
1. Epidemic forecasting (30 min)
2. **Modeling paradigms - Overview**
3. Mechanistic models (15 min)
4. Statistical/ML/AI models (55 min)
5. Hybrid models (45 min)
6. Epidemic forecasting in practice (20 min)
7. Open challenges (20 min)

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

Part 2: Modeling Paradigms

Mechanistic Models

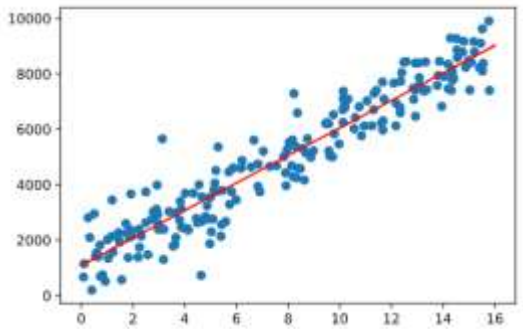
- Encode mechanism of epidemic spread
  - Based on domain-based constraints
- Intuition:
  - People move from compartments based on the disease progression
  - Differential equations describe movement

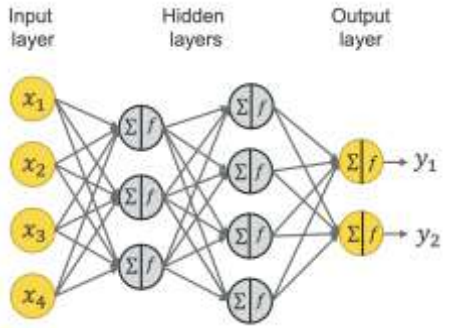


- Modeling approaches:
  1. Mass-action models (ODE models)
  2. Metapopulation models
  3. Agent-based networked models

Statistical/ML/AI Models

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





Rodríguez, Kamarthi, and Prakash 2022

Hybrid Models

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

Model vs Data sources

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

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

Model vs Data sources

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

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

Georgia Tech

Rodríguez, Kamarthi, and Prakash 2022

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

Detailed table in survey

102 methods

250+ references

Dating back to 2000

Georgia Tech

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 3: Mechanistic Models

Mechanistic Models

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

Mechanistic Models (Outline)

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

Note: 1 and 2 are also known as compartmental models

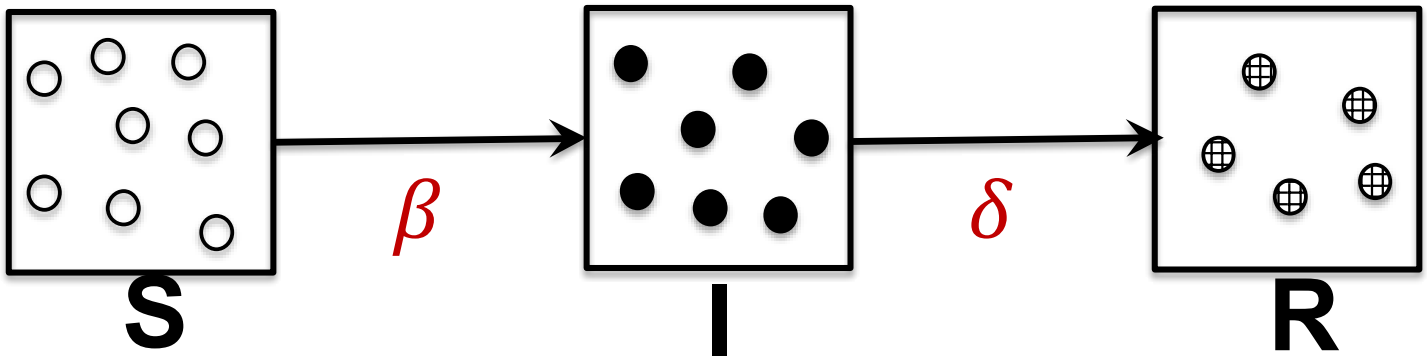
Mechanistic Models (Outline)

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

[M1] Mass-action models

[Hethcote, SIAM Review 2000]

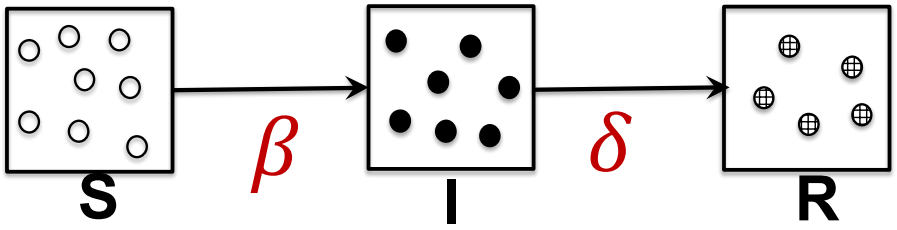
- One of the simplest models
  - Susceptible: healthy, can get infected
  - Infected: can infect others through contact
  - Recovered: cannot infect others



Assumptions

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

SIR Model



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

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

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

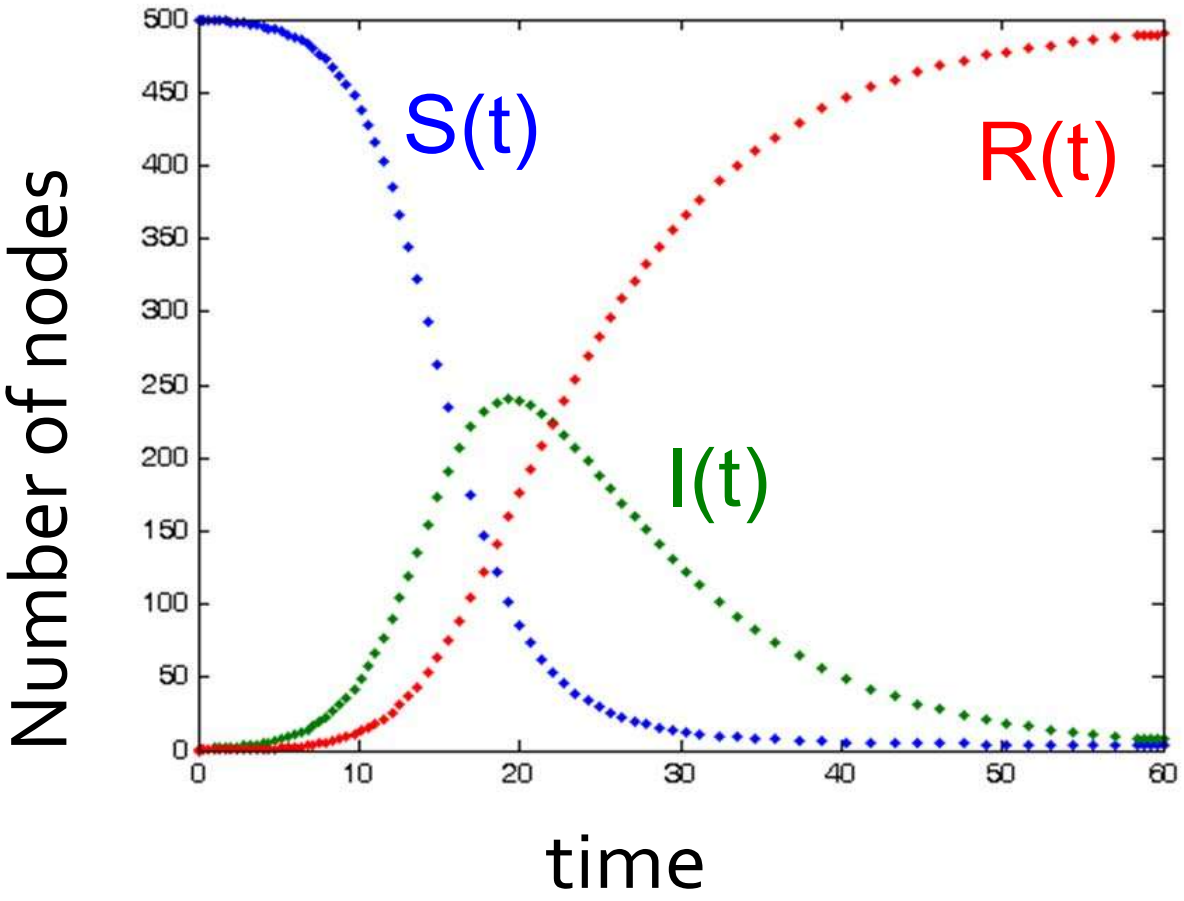
Number of new infections =
\beta * # infection attempts

Number of infected nodes curing

Solving SIR

- No closed form solution!

SIR: numerical output



Many many extensions

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

Threshold Phenomenon: R0

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

- This implies

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

Threshold Phenomenon

- So, $R_0 = \beta/\delta$
  - Basic Reproductive number: average number of secondary cases caused by one individual
- If $S(0) < \delta/\beta = 1/R_0$
  - Epidemic dies out
  - Large epidemic if and only if R0 > 1
  - Hence estimating R0 very important!
    - Why?
    - Immunization: reduce S(0) to below 1/R0

Mechanistic Models (Outline)

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

[M2] Metapopulation Models

[Dimitrov and Meyers, INFORMS 2010]

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

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

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

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

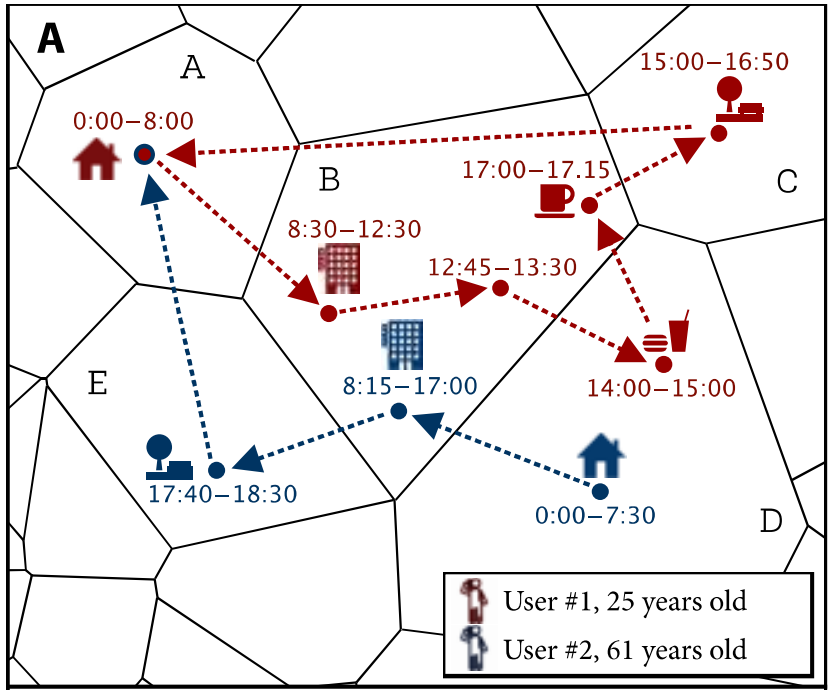
But… Human contact patterns are not random

Source: Mi Jin Lee at petterhol.me

How to Capture Them? Example: Using Call Data Records

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

#raw data



#trips

B

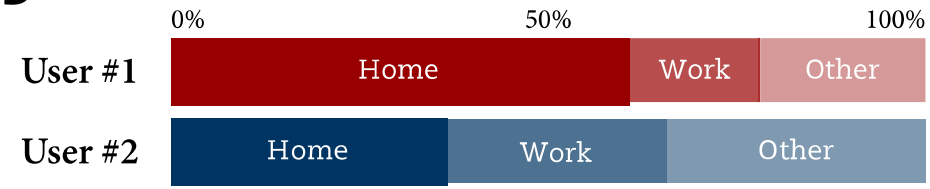
	A	B	C	D	E
A	0	1	0	0	0
B	0	1	1	1	1
C	1	0	0	0	0
D	0	2	0	0	0
E	1	0	0	0	0

#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



Numerous COVID-19 examples

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

Mechanistic Models (Outline)

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

[M3] Agent-based networked models

[Marathe and Vullikanti, CACM 2013]

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

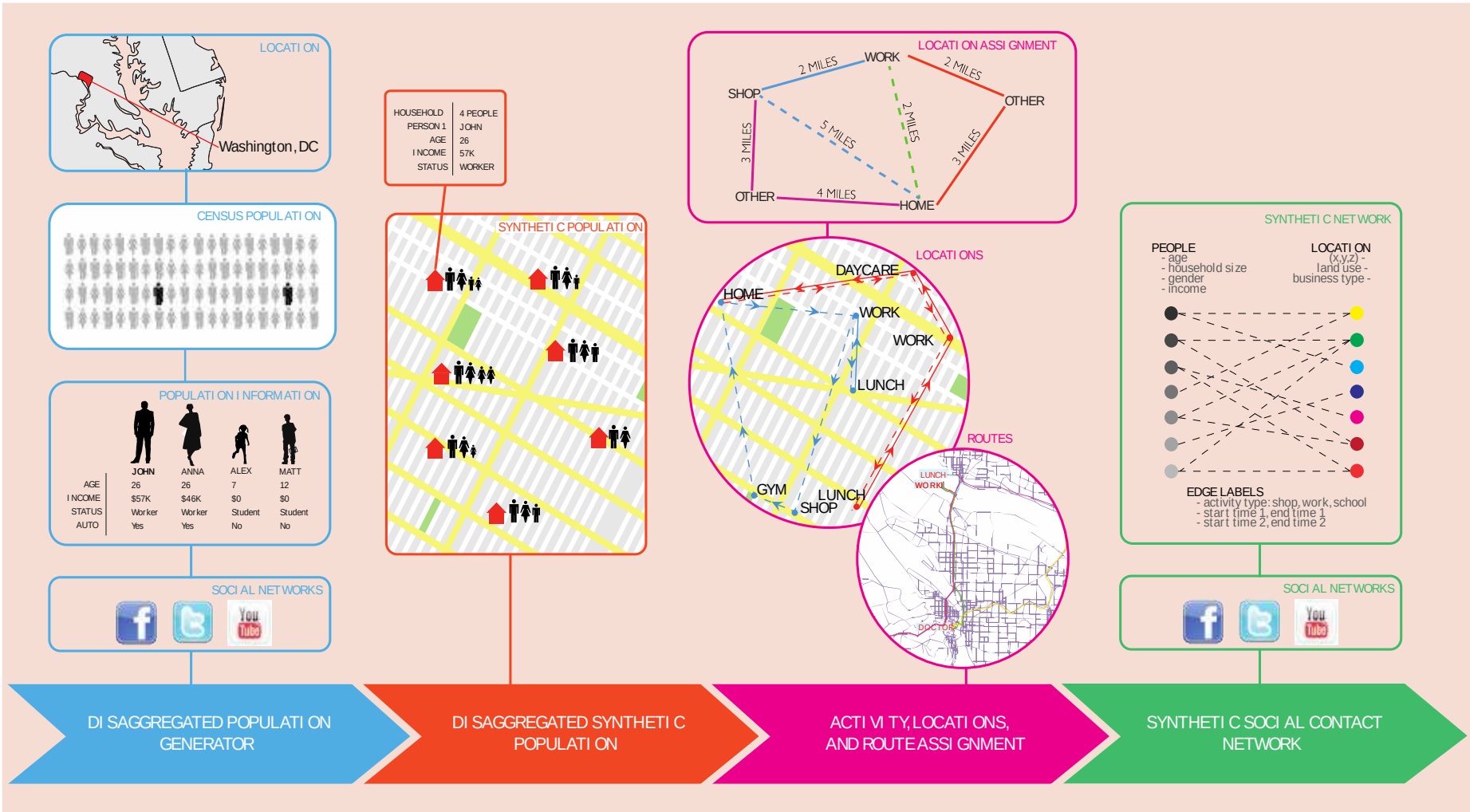
First principles Approach for Constructing Social Contact Networks

- For individuals in a population [Marathe and Vullikanti, CACM 2013]
  - Demographics (who)
  - Sequences of their activities (what)
  - Times of their activities (When)
  - Places/locations of their activities (where)
  - Reasons for their activities (Why)

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

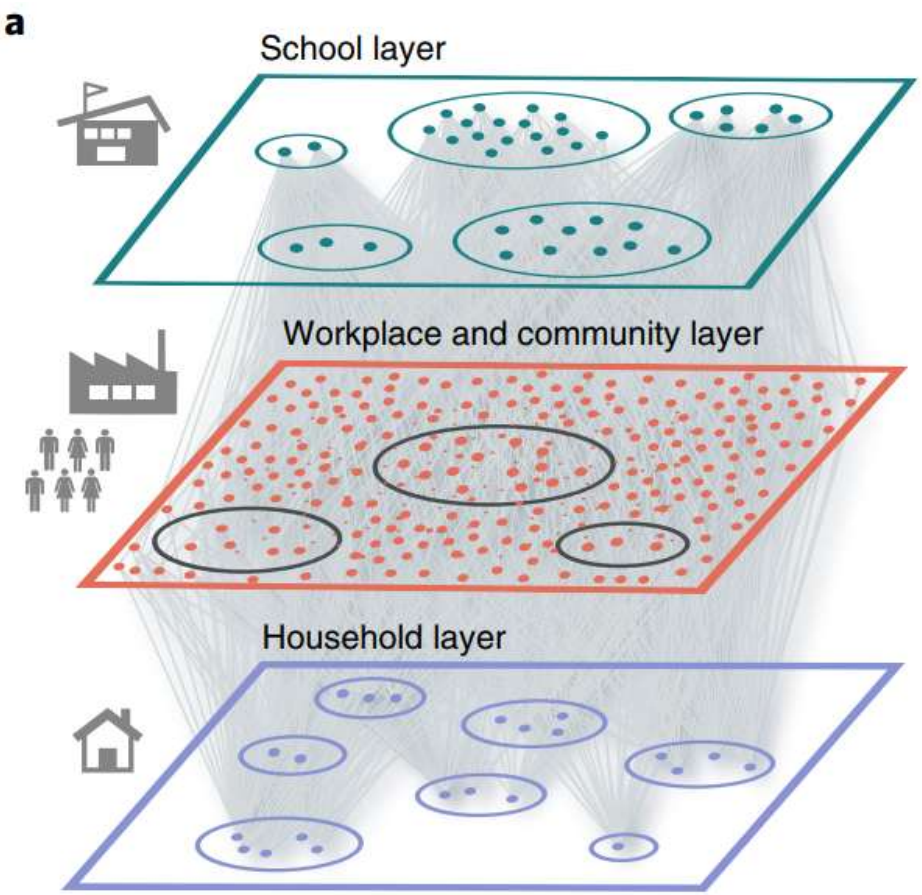
First principles Approach for Constructing Social Contact Networks

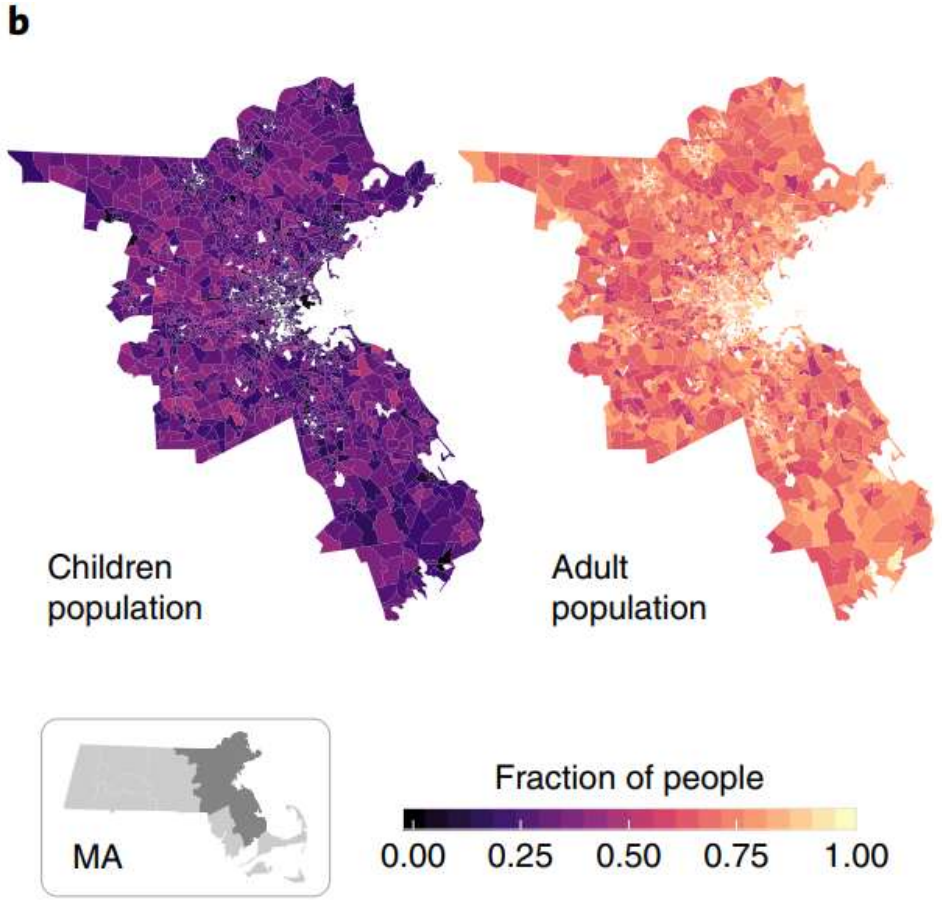
[Marathe and Vullikanti, CACM 2013]



Example: COVID-19 in MA

[Aleta et al, Nature Human Behavior 2020]





Calibration of Mechanistic Models

- Estimate parameters
  - Beta, delta, initial conditions

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

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

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

Pros/Cons Mechanistic Models

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

Pros/Cons contd.

- Sometimes does not reflect reality
  - SARS example
    - High 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.

Break 3 min

Outline

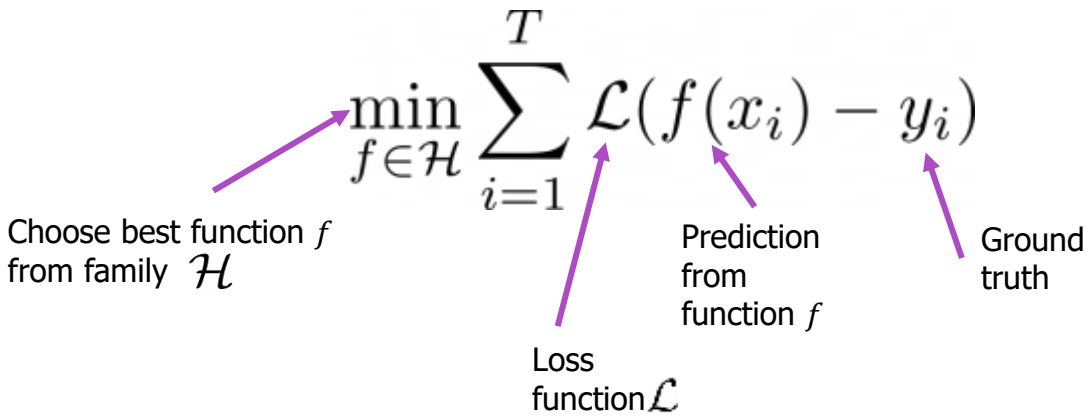
1. Epidemic forecasting (30 min)
2. Modeling paradigms - Overview
3. Mechanistic models (15 min)
4. **Statistical/ML/AI models (55 min)**
5. Hybrid models (45 min)
6. Epidemic forecasting in practice (20 min)
7. Open challenges (20 min)

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

Part 4: Statistical, Machine-learning/AI models

Statistical/Machine Learning Models

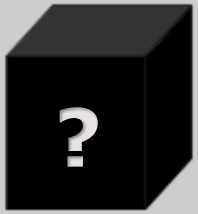
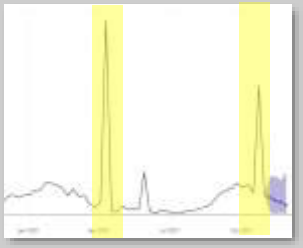
- Intuition:
  - Find the best function from a family of functions that approximate forecast target given input data.
  - Best approximate is found using past training data.



Why Stat./ML models?

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

Overview of challenges for Stat./ML models

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

Statistical, ML/AI Models (Outline)

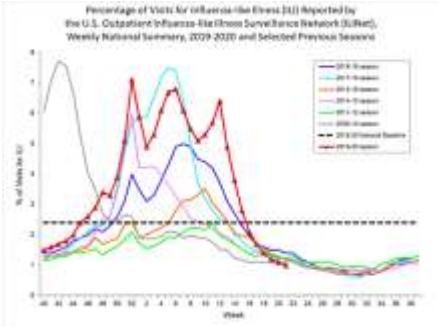
- Approaches:
  1. Regression Models
  2. Language and Vision Models
  3. Neural Models
  4. Density Estimation

Statistical, ML/AI Models (Outline)

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

[S1] Regression Models

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



Idea 1: AutoRegressive Models

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

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

Future target to predict

Past values of epi curve

Noise

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

Ex. 1: Google Flu Trends

[Ginsberg+ Nature 2009]

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

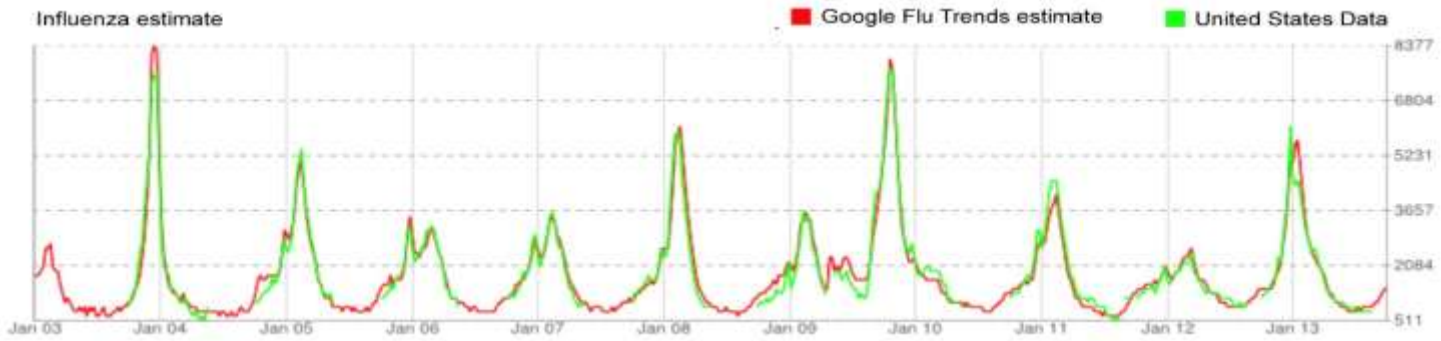
$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$

  - P = ILI (physician visits)
  - Q = Fraction of search queries that are ILI-related

CNN 2008

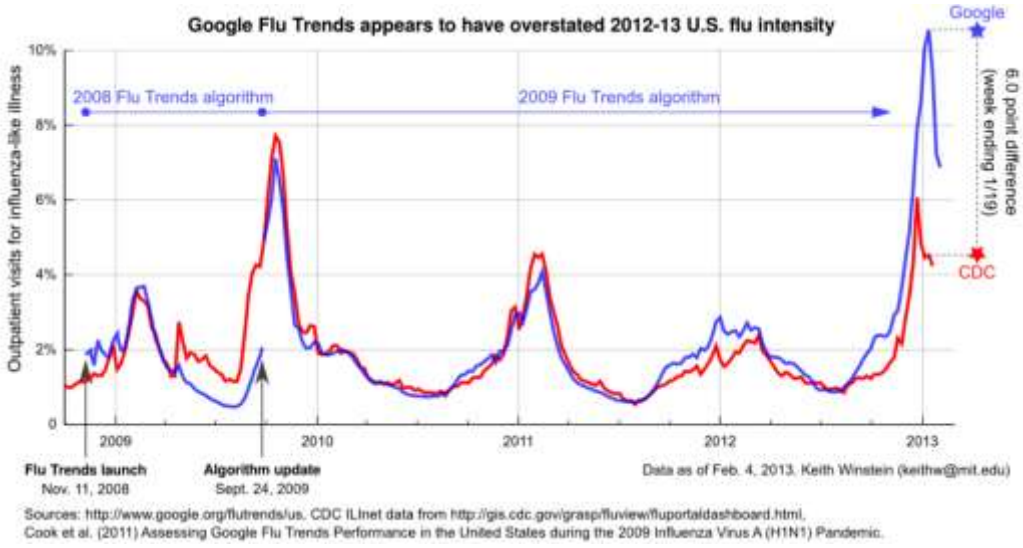
Google Flu Trends (Contd.)

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



However,…

- Didn’t capture changing trends in keyword correlates, i.e. didn’t handle data drift
  - Failed to capture H1N1 pandemic, overestimate 2012-13 season [Olson+ PLoS Comp. Bio 2013]



Ex. 2: ARGO

[Yang+ Sci. Reports 2017]

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

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

Past ILI

Query volume

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

Examples of search terms used

ARGO2 (Extension)

[Ning+ Sci. Reports 2019]

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

Regression models: Extensions

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

Statistical, ML/AI Models (Outline)

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

[S2.1] Language models

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

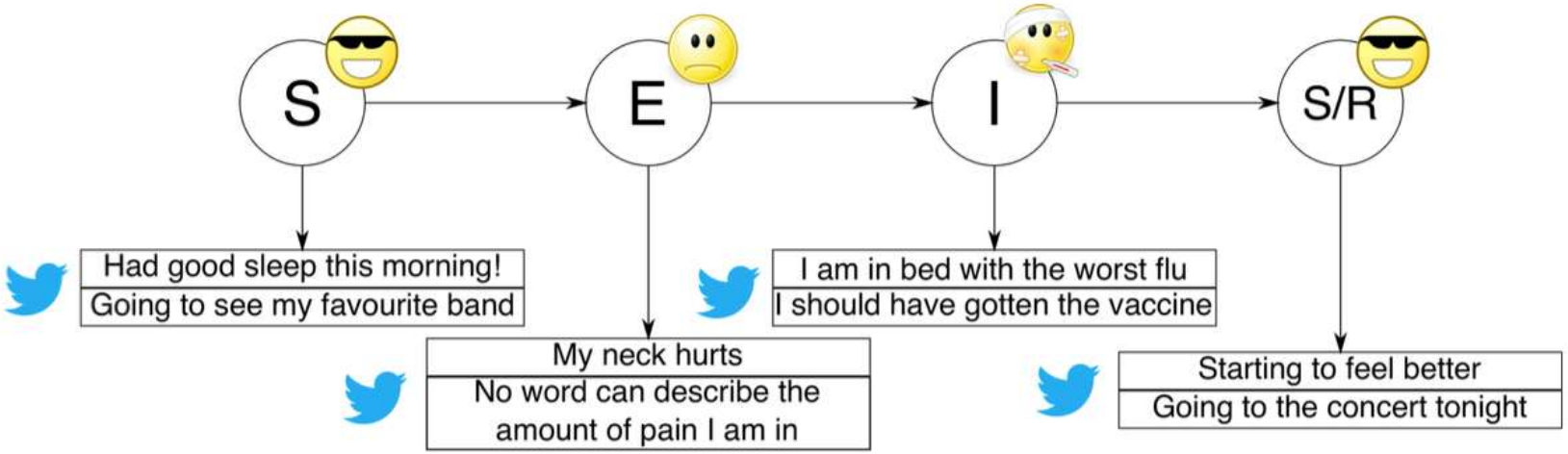
Idea 1: Using Tweets to forecast H1N1 pandemic

[Chen+ ICDM 2017]

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

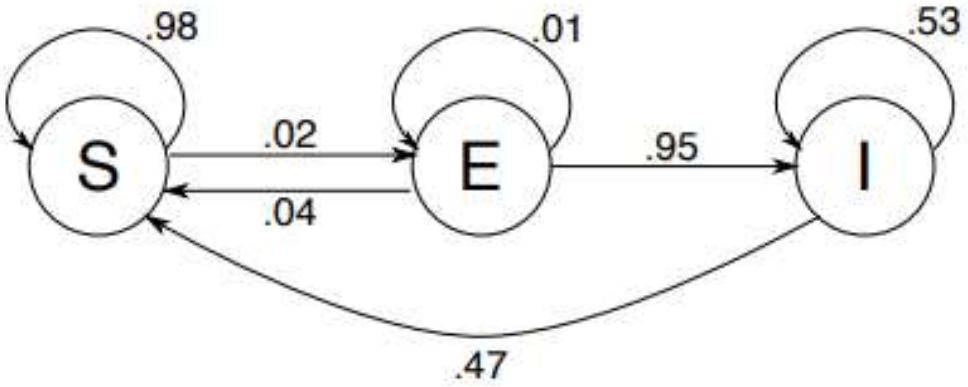
States of infection cycle

- Model states of infection cycle using tweets



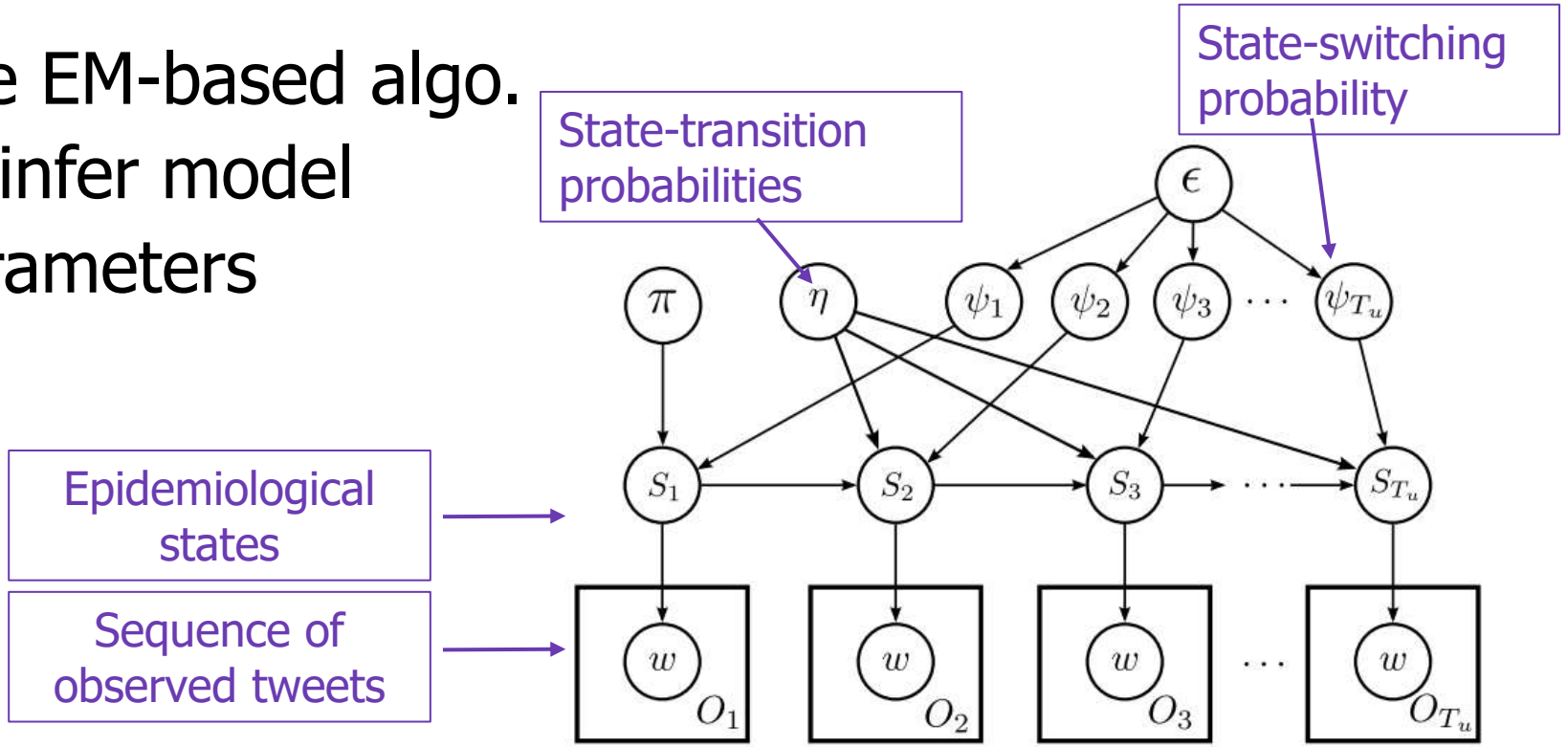
Model transition across states

- Transition probabilities across epidemiological states
- Automatically learned by HFSTM



HFSTM model (Contd.)

- Use EM-based algo. To infer model parameters



[S2.2] Vision models

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

Ex 1: Satellite images to detect Flu outbreaks

[Butler+ IEEE Annals of Comp. 2014]

- RS Metric satellite imagery dataset
- Vision based automated algorithms to estimate no. of vehicles
  - Parking lots, streets, etc Of hospitals

- Linear regression on occupancy rate to model weekly ILI case counts.

Ex 2: Covid-19 Example [Nsoesis+ arXiv 2020]

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

Statistical, ML/AI Models (Outline)

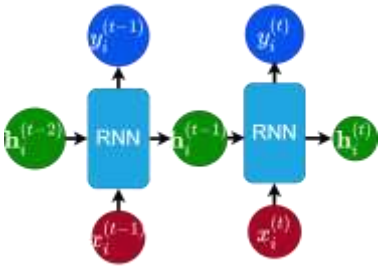
- Approaches:
  1. Regression Models
  2. Language and Vision Models
  3. **Neural Models**
  4. Density Estimation

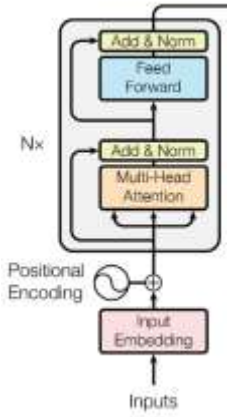
[S3] Neural Models

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

Idea 1: Off the shelf sequential models

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





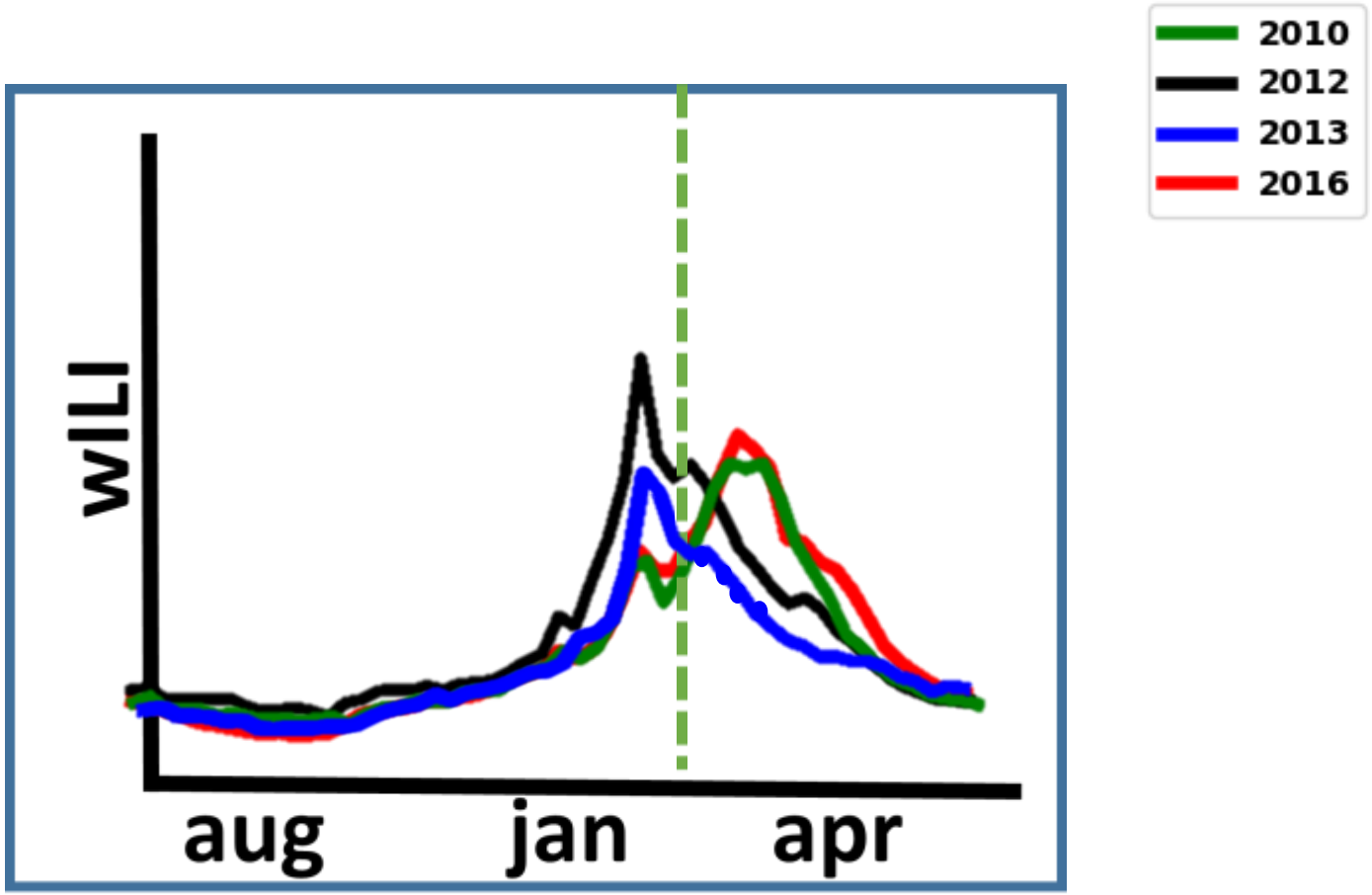
Adaptating to epidemiology... some ideas

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

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

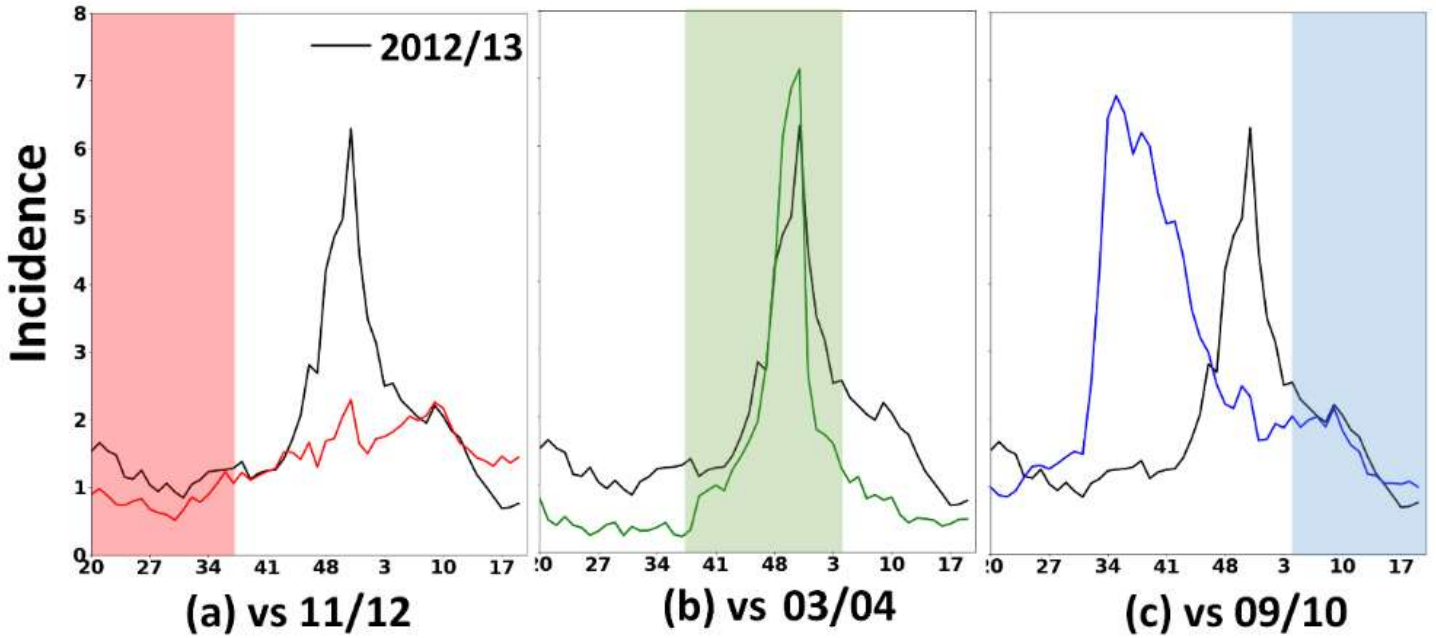
[Adhikari+, KDD 2019]

- Idea: **clustering for prediction**



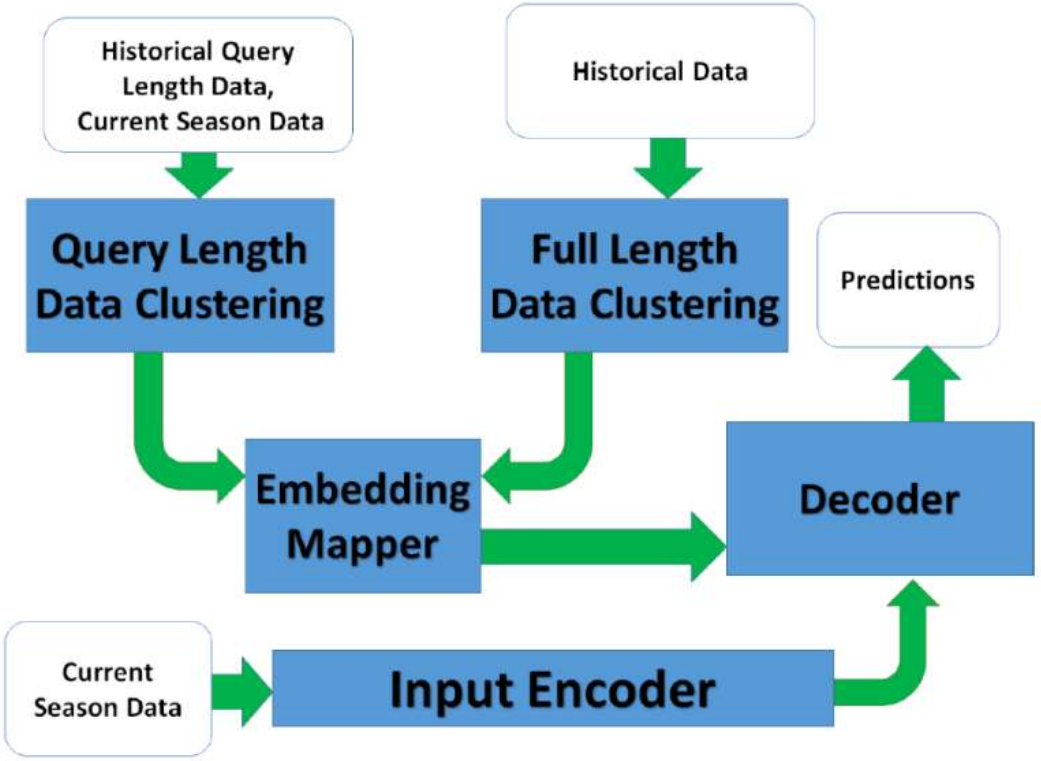
Model temporal dynamics via similarity CONTD.

- Idea: **Dynamic clustering for prediction**



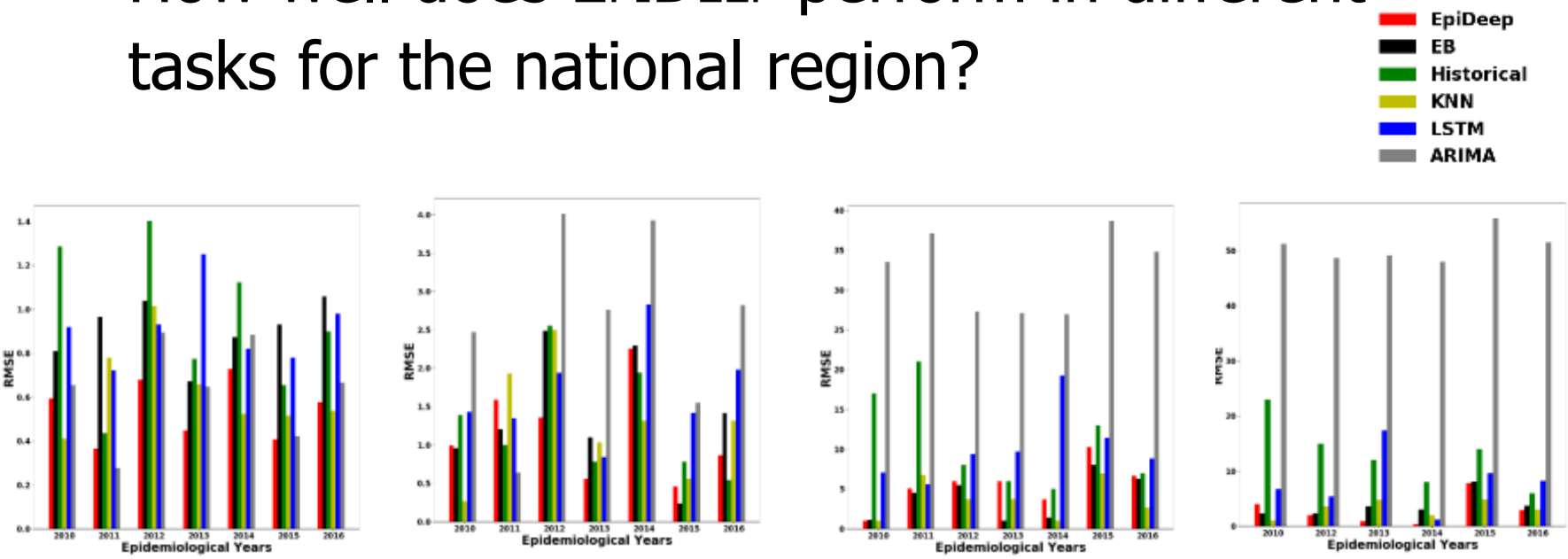
Model temporal dynamics via similarity CONTD.

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



Performance: National region

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



Future Incidence | Peak Intensity | Peak Week | Onset

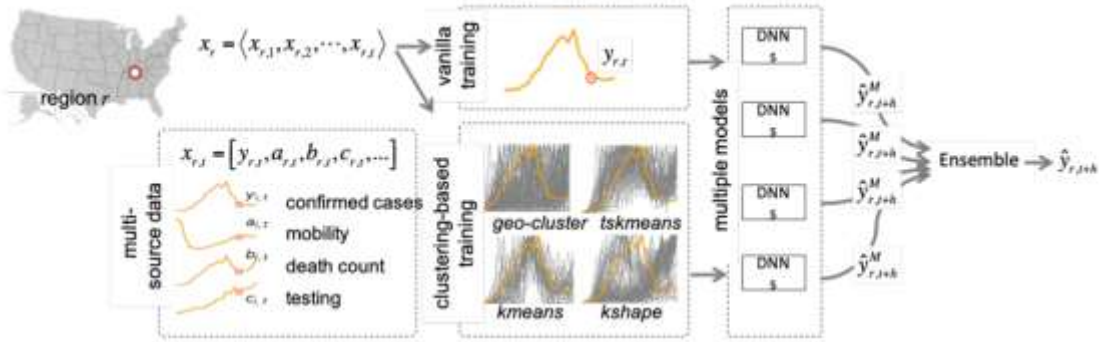
Lower is better

EpiDeep outperforms baselines in most settings.

Ex. 2: Using multiple clustering methods

[Wang+, BigData 2020]

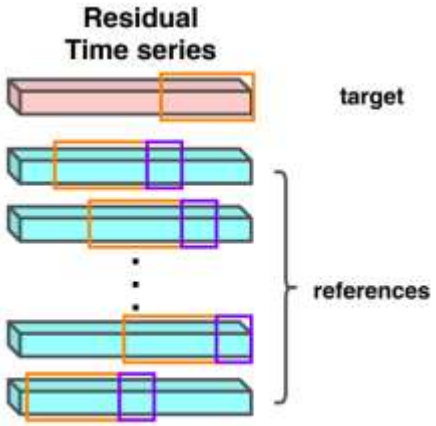
- Temporal and geo. similarity (adjacent regions)
  - Train one model for set of regions
  - Regions clustered by geographical similarity or using clustering algorithms
  - Multiple models with different clustering strategies combined using ensembling (more on ensembles later)



Ex. 3: Inter-series attention

[Jin et al., SDM 2021]

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

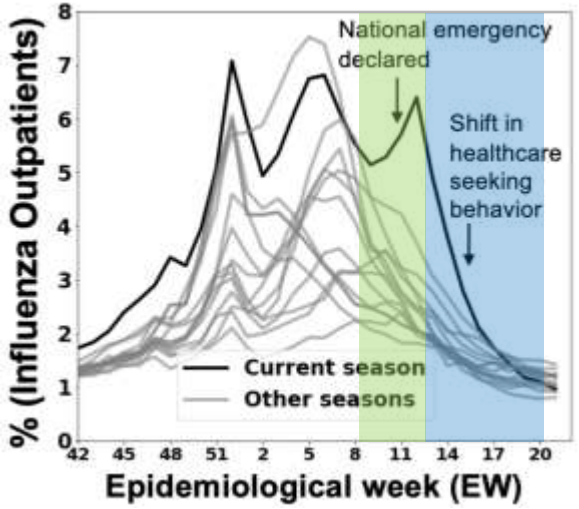


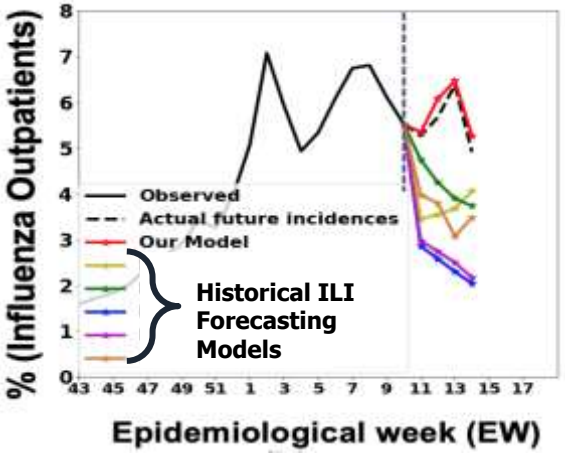
Idea 2: Transfer knowledge representations

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

A Novel Forecasting Setting

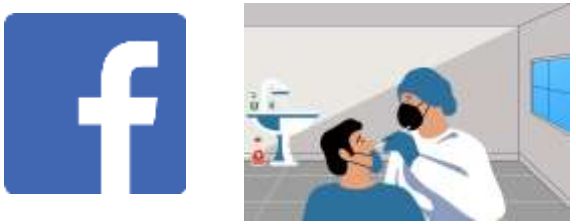
- Influenza counts may be affected by
  - COVID "contamination"
  - Shift in healthcare seeking behavior
- This new scenario lead us a <u>novel</u> forecasting problem
- Historical flu models unable to adapt to new trends

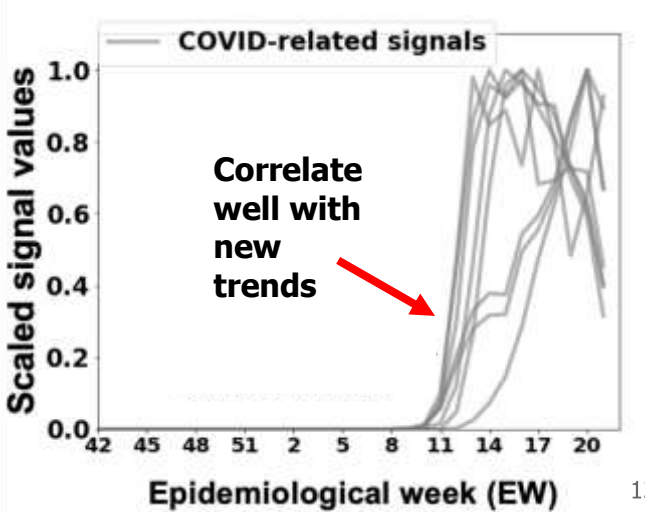




New COVID-related signals correlate with new trends

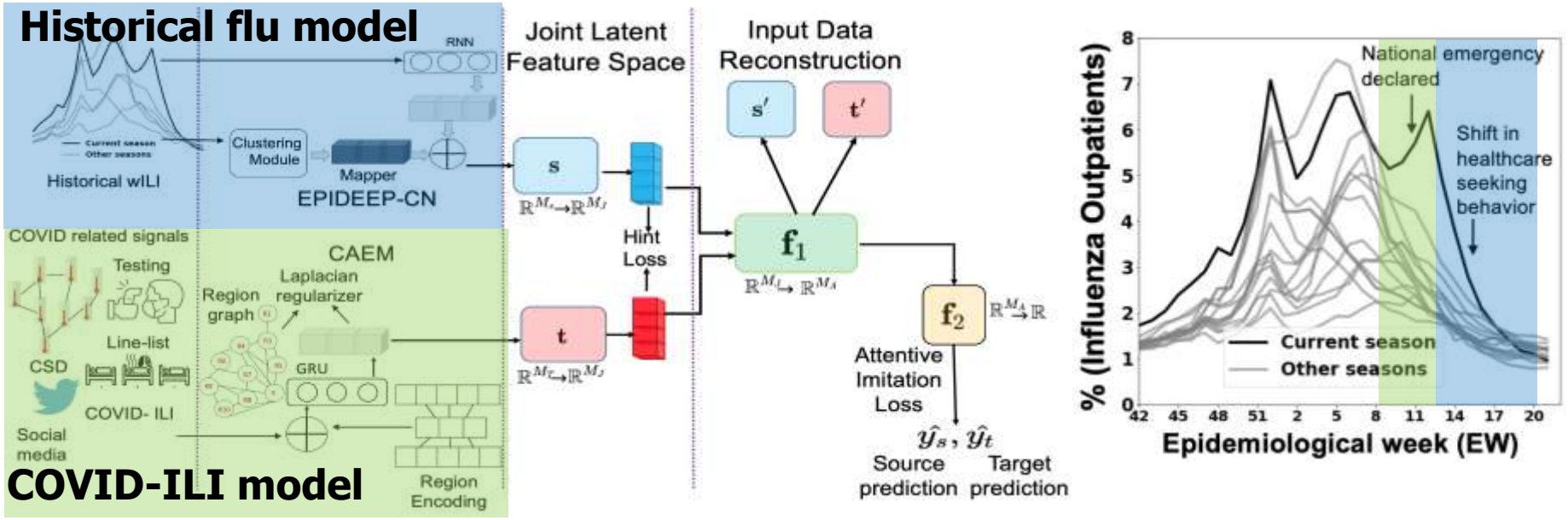
- Line-list based
- Testing
- Crowdsourced
- Mobility
- Exposure
- Social Media surveys





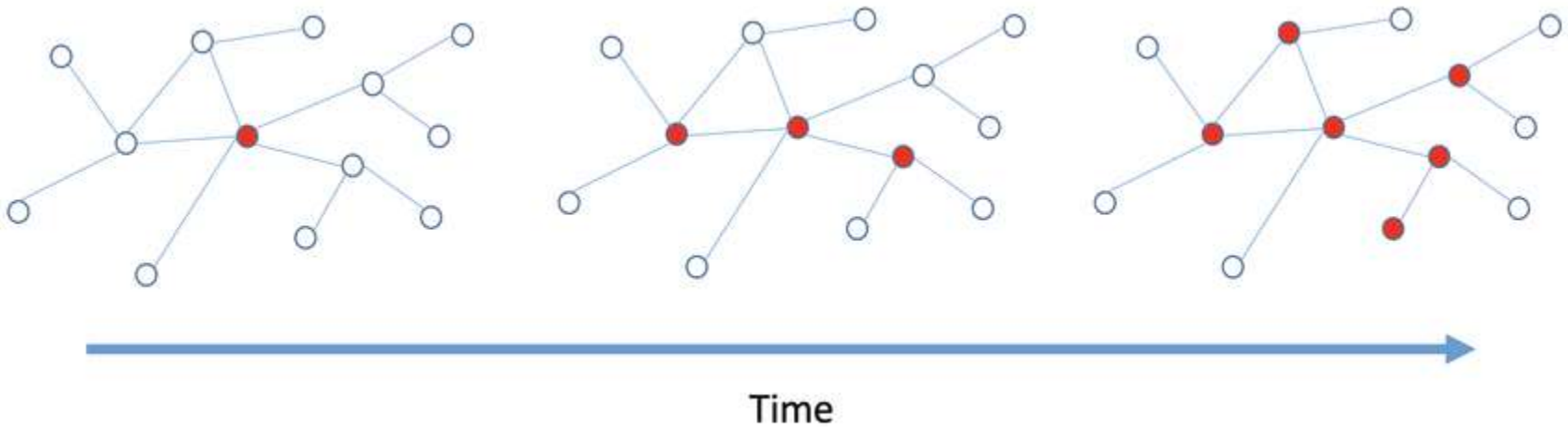
Attentive transfer learning for heterogeneous domains

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



Idea 3: Incorporate spatial structure

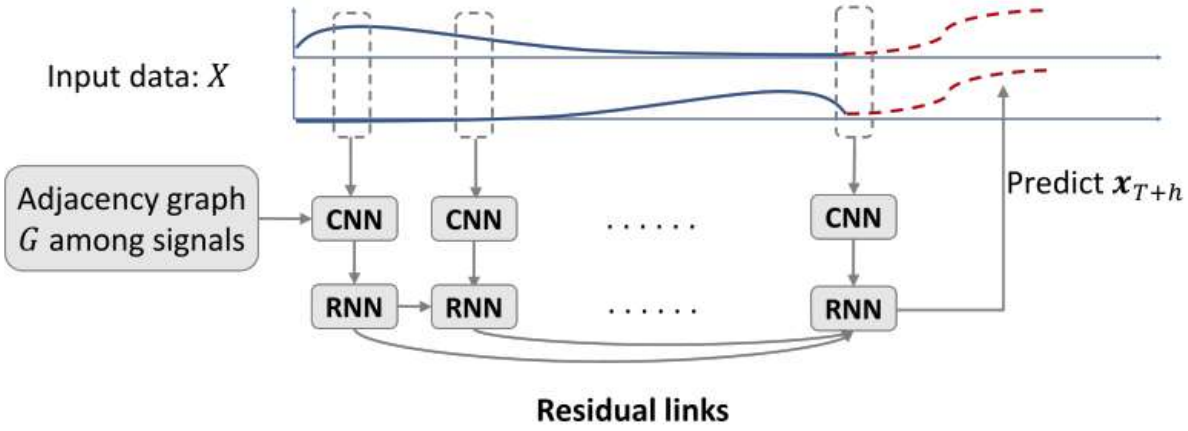
- Pathogens propagate to adjacent regions
  - And then to new adjacent regions
- Propagation over spatial graphs



Ex 1: Convolutional modules to aggregate related regions

[Wu+, SIGIR 2020]

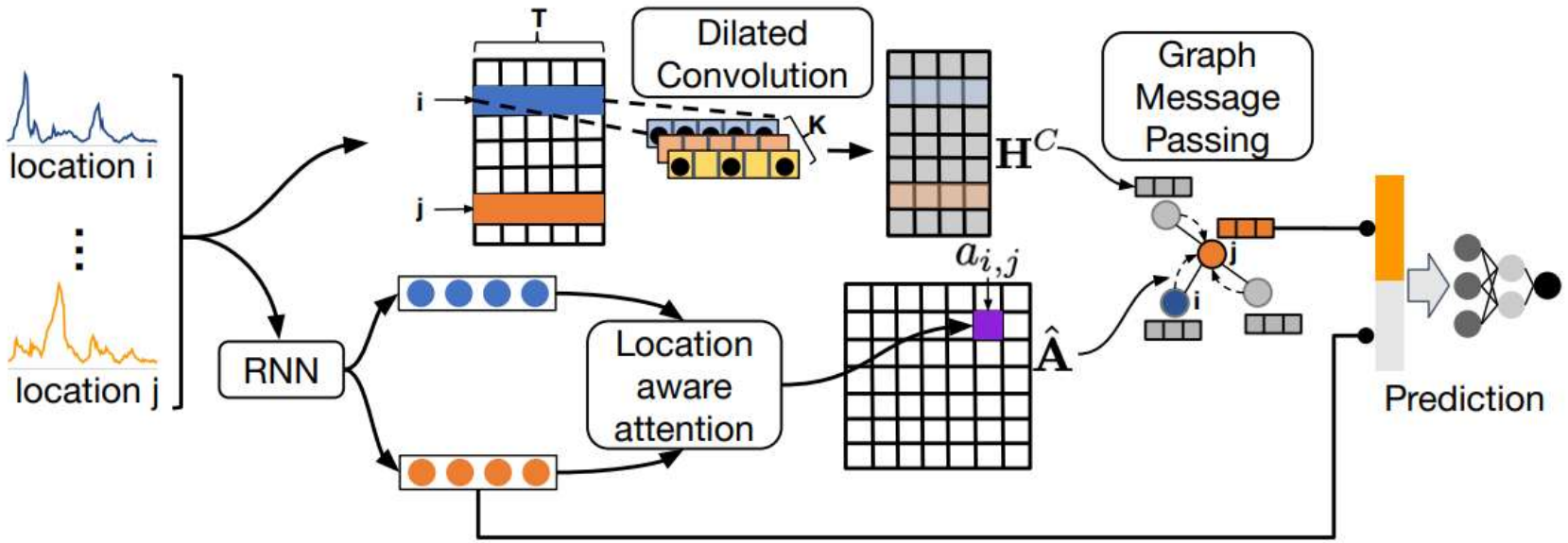
- Combine CNN and RNN
  - CNN: Model regional proximity
  - RNN: Temporal dynamics
  - Residual connections: Better generalization



Ex 2: Graph message passing for spatial propagation

[Deng+, CIKM 2020]

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



ColaGNN in long-term forecasting

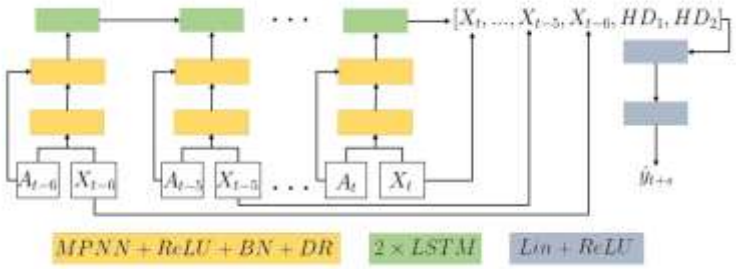
- Dilated Convolution observed to improve long-term predictions as well

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

Ex 3: Transfer Learning via Graph Neural Networks

[Panagopoulos+, AAAI 2021]

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



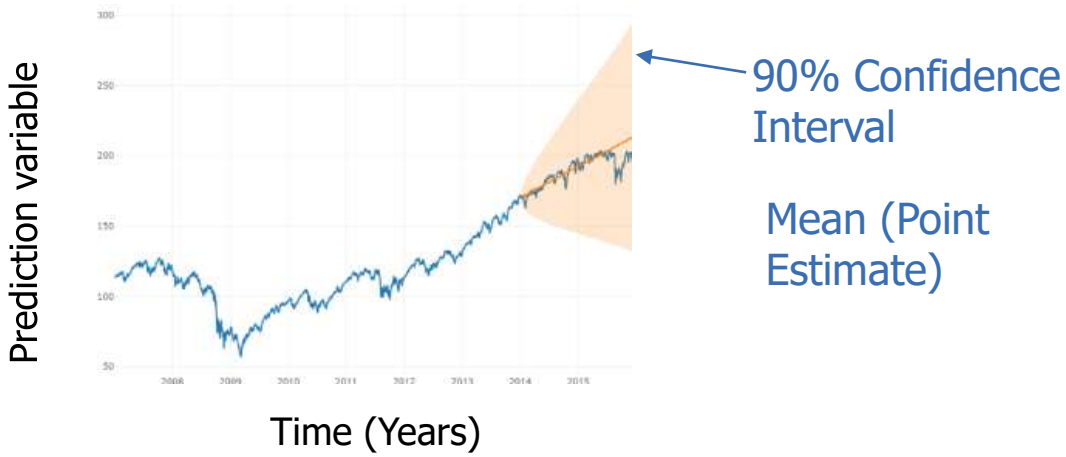
Statistical, ML/AI Models (Outline)

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

[S4] Density Estimation Models

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

$$P\left(y_i^{(t+\tau)} | Y_i^{(1:t)}, X_i^{(1:t)}\right)$$



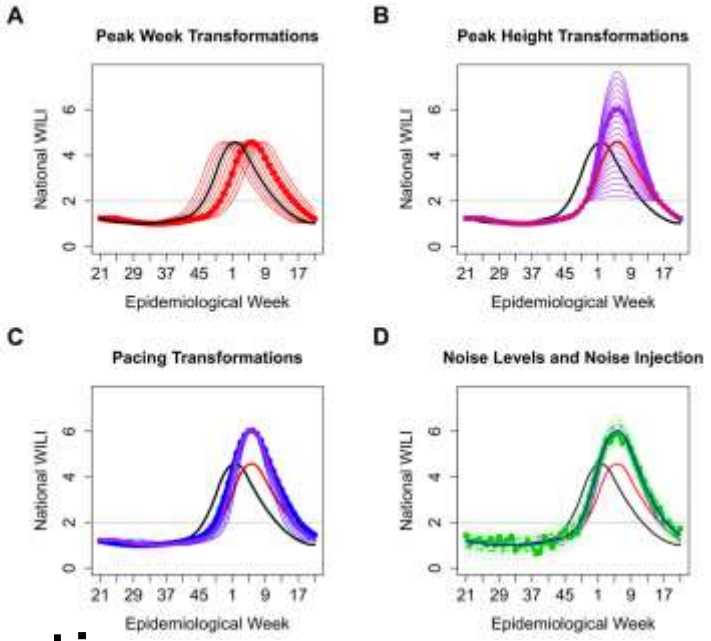
Types of Density Estimation Models

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

Ex 1: Empirical Bayes (Parametric)

[Brooks+ PLoS 2015]

- Current season's epidemic curve is a probabilistic distribution of features
- Epi-curve function parameters:
  - Similarity in shape to past sequences
  - Peak height, week
  - Scaling factor of the curve

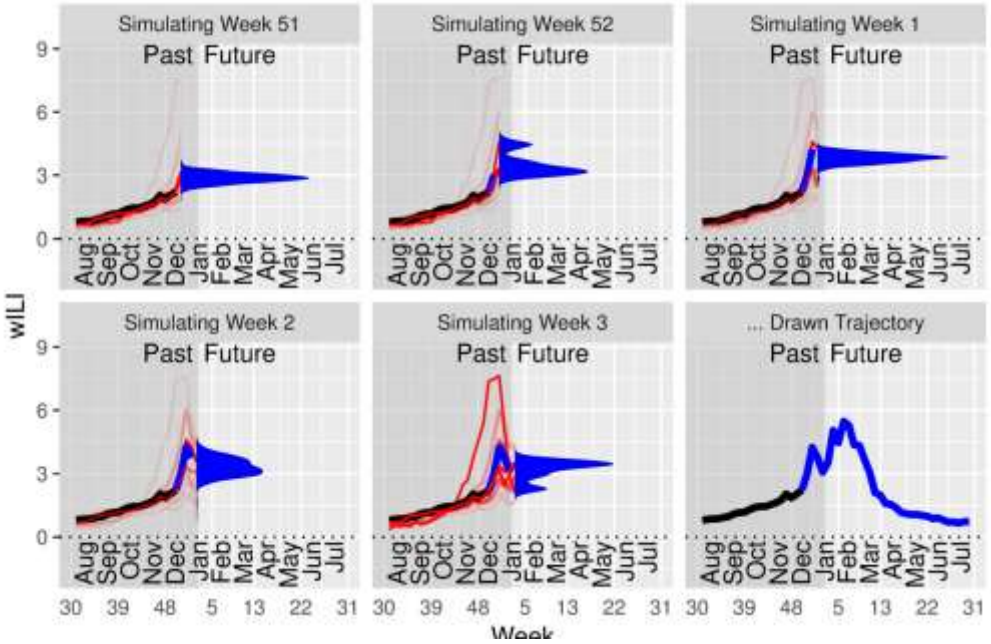


- All modelled into forecast distribution
- Use Bayesian Inference to calibrate for current season

Ex 2: Delta Density (Non-parametric)

[Brooks+ PLoS 2017]

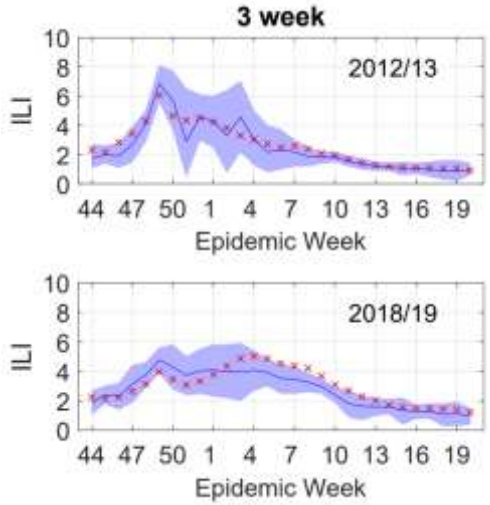
- Use kernel density estimation to leverage similarity with historical seasons
- One of the top models in Flusight 2017 challenge

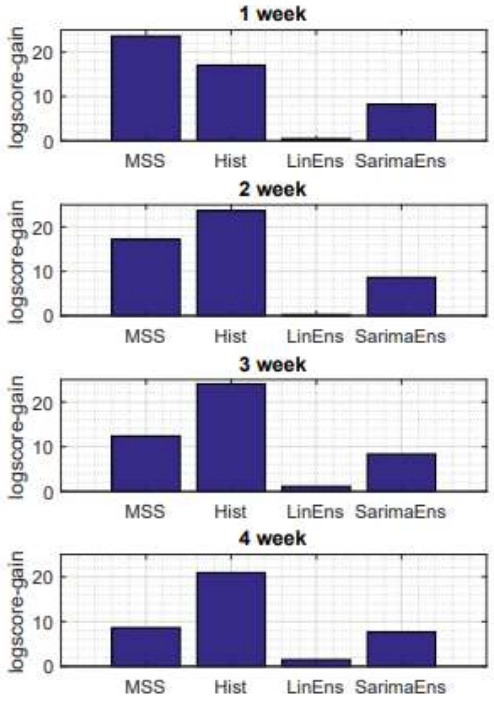


Ex 3: Gaussian Process (Non-Parametric)

[Zimmer+ ICML 2019]

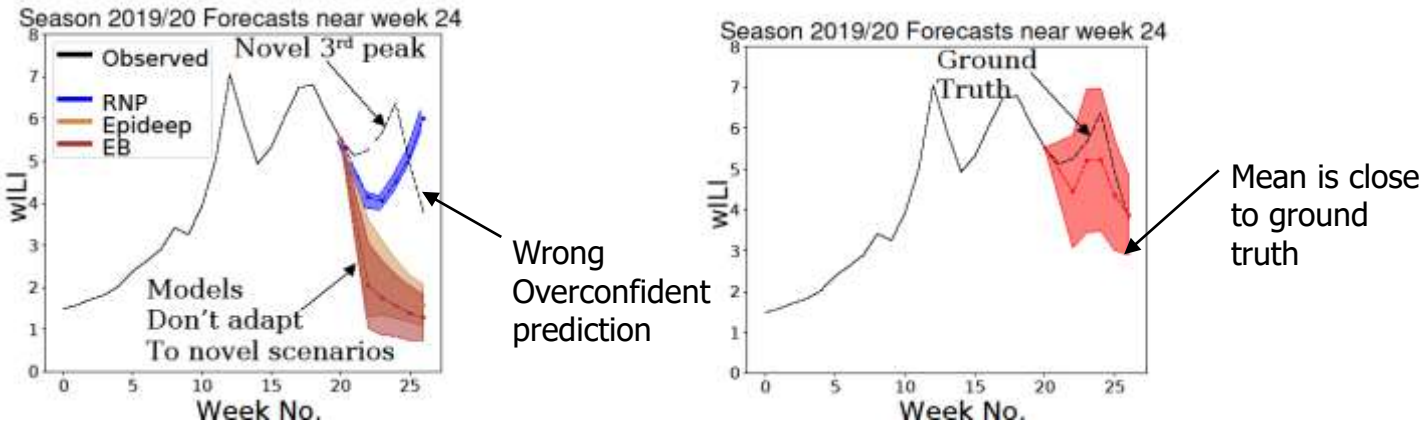
- Used Gaussian Process over incidence values of previous seasons
- Showed reasonable confidence intervals and state-of-art log score over past models





Neural models for calibrated forecasts

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

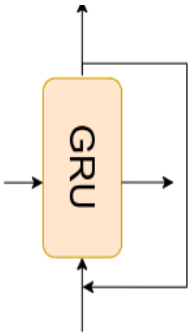


Rodríguez, Kamarthi, and Prakash 2022

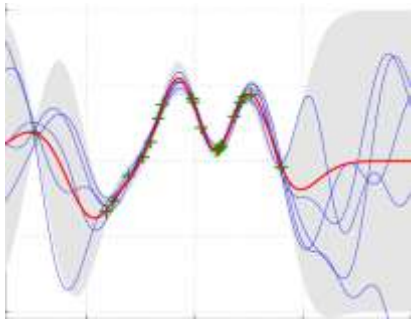
Ex 4: EpiFNP: Neural non-parametric model for better calibration

[Kamarthi+, NeurIPS 2021]

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



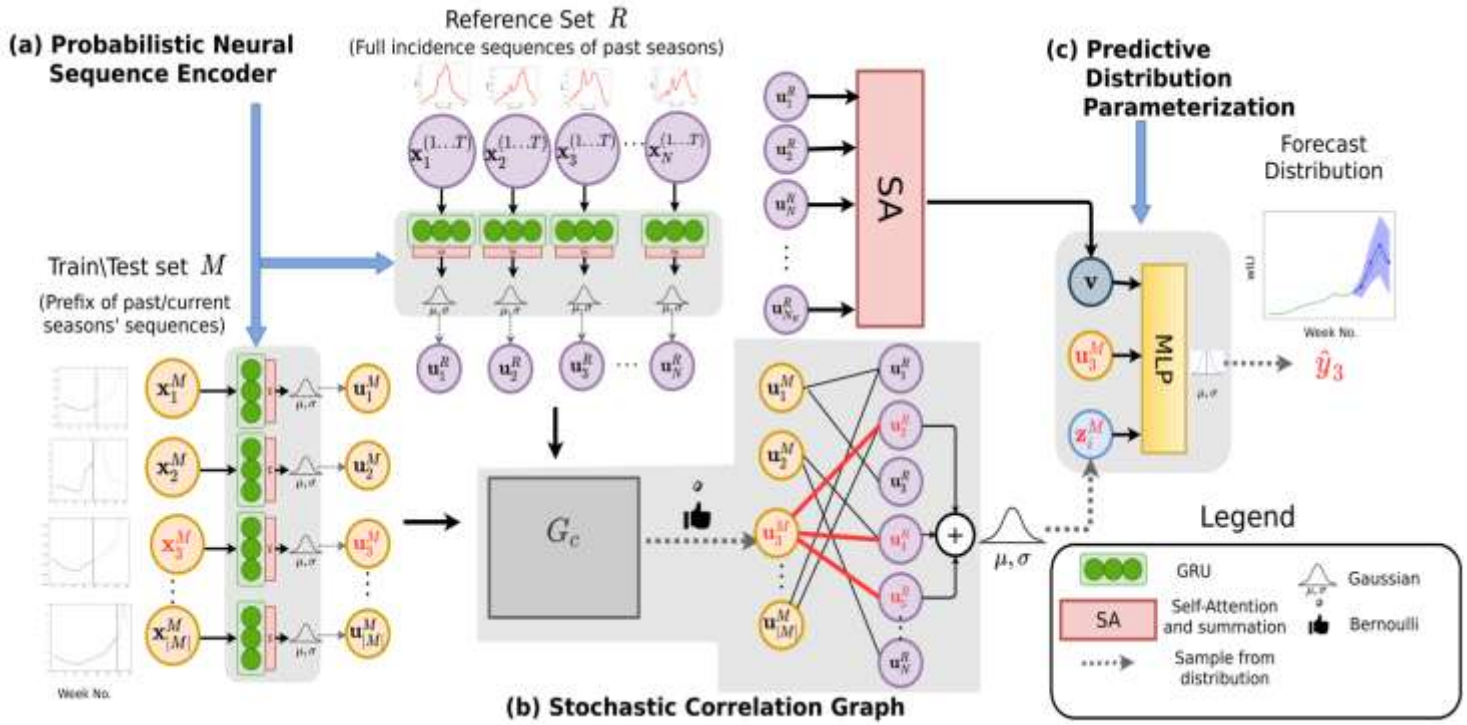
Deep Sequential Models



Accurate and Well-calibrated neural forecasting model

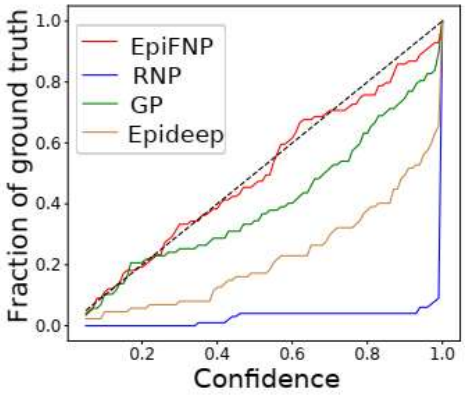
EpiFNP: Architecture

Sequential representations + neural Gaussian processes

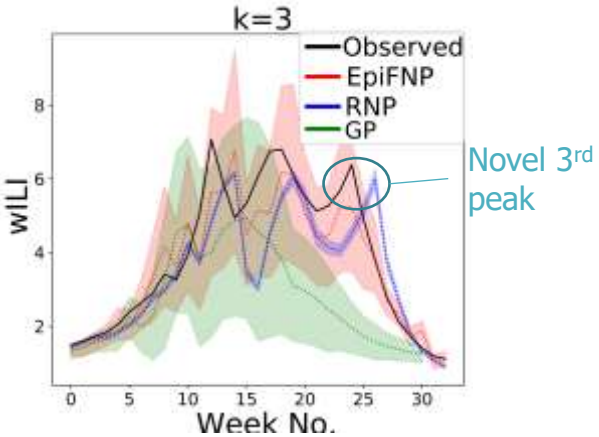


Rodríguez, Kamarthi, and Prakash 2022

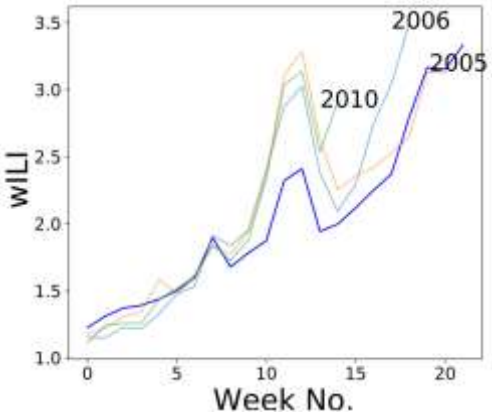
Results



Well calibrated predictions



Adapt to novel patterns



Explaining predictions

Most similar seasons chosen by EpiFNP

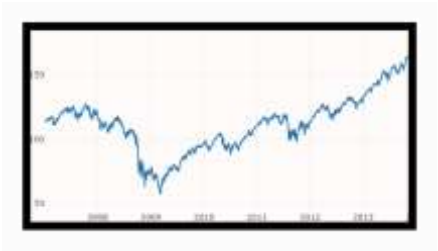
Georgia Tech

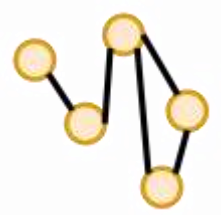
Extending to modeling multiple views/modalities

Sequences

Static Features

Graphs

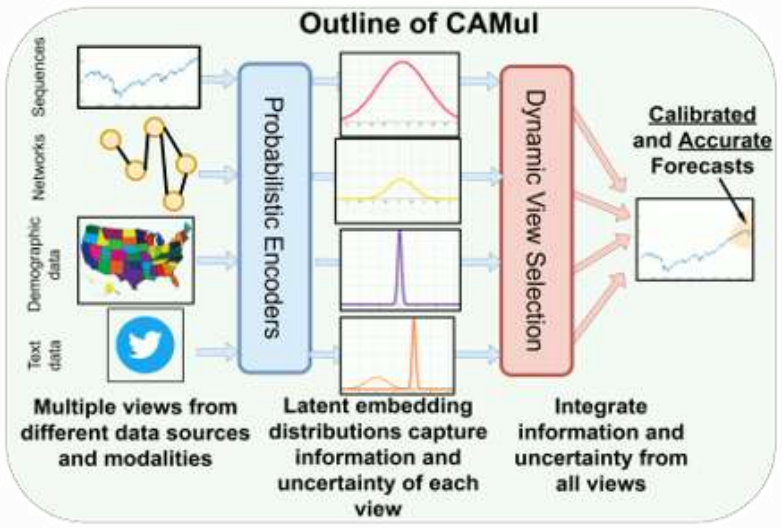




CaMuL: multi-view time series forecasting

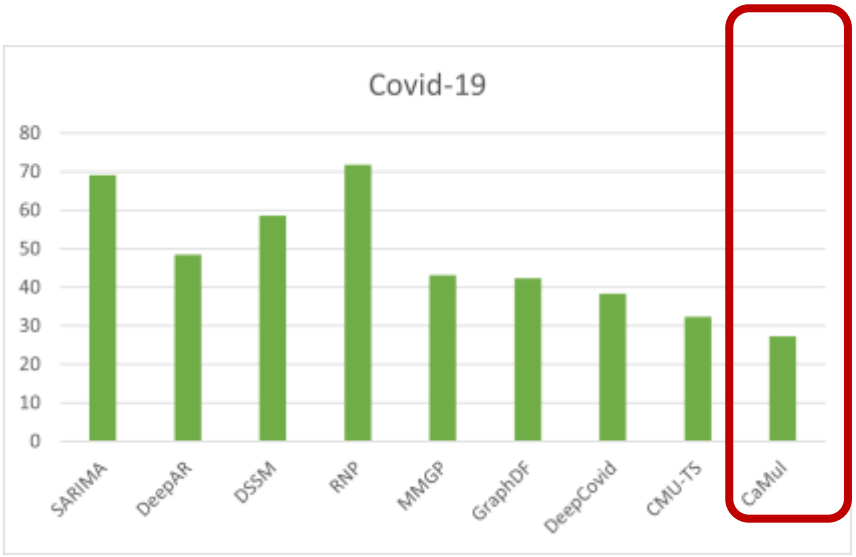
[Kamarthi+, WWW 2022]

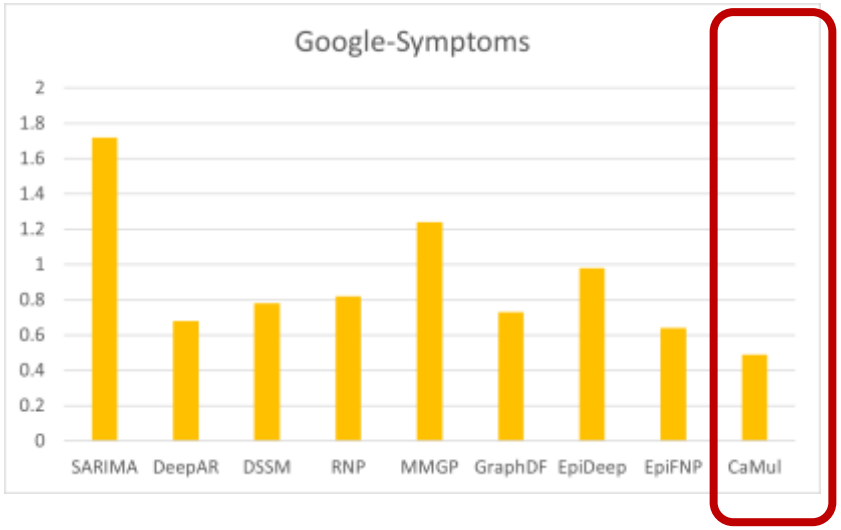
- Encode representation for each data source
- Select important views adaptively
- Combine important views to provide calibrated forecasts



CaMuL: Results

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





Pros/Cons Statistical Models

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

Pros/Cons Contd.

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

Part 5: Hybrid Models

Will continue after break