Machine Learning To Track The Spread Of COVID-19

Viktor Prasanna

prasanna@usc.edu

ceng.usc.edu/~prasanna

dslab.usc.edu

University of Southern California

Disclaimer

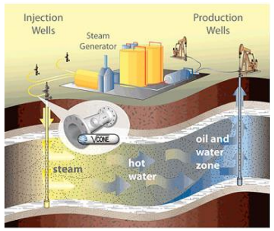
- This presentation targets a broad audience
- Most of the low level technical details have been skipped to simplify the flow
- For all the technical details, please refer to the paper below
    - A. Srivastava and V. K. Prasanna, “Learning to Forecast and Forecasting to Learn from the COVID-19 Pandemic,” arXiv, April 2020
    - https://arxiv.org/abs/2004.11372
- The results from COVID-19 spread analysis should be interpreted with caution from people with the appropriate technical background

Outline

- Data Driven Research At USC's Data Science Lab (DSLab)
- Machine Learning (ML) 101
- Basics of Prediction
- Examples Of Prior Work At DSLab
  - Smart Grid
  - Smart Oil Field
  - Internet Traffic Prediction
  - Chikungunya Epidemic Prediction
- Covid-19 Epidemic Prediction
  - Models
  - Our approach
  - Results
  - Role of ML
- Covid-19 Interactive Visualization Tool
- Ongoing work
- Concluding Remarks
- Q & A

Data Driven Research At Data Science Lab (DSLab)

Oil Fields



Steam Job Benefit Prediction
Slippage Detection

Load/Generation Prediction
Energy Resource Scheduling
Voltage Regulation

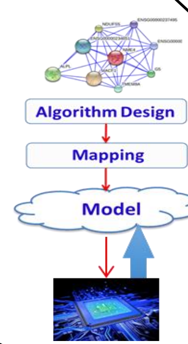
Power Grids

Machine Learning

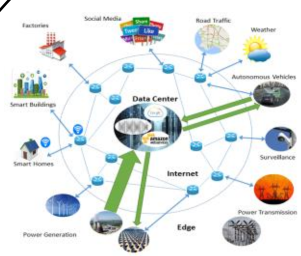
Optimization

dslab.usc.edu

Runtime Performance Prediction
Building Block Selection
Algorithm to HW Mapping



System Design



Internet Design

Network Traffic Prediction
Scalable Traffic Measurement
Traffic Routing

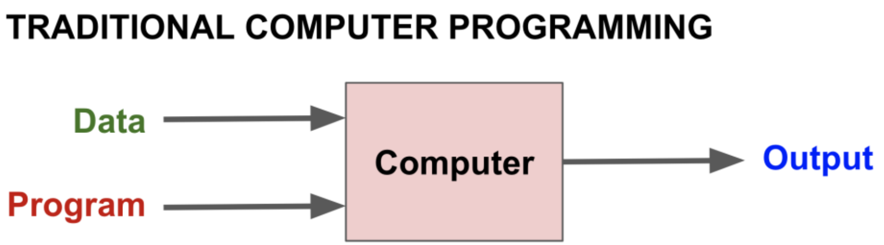
Outline

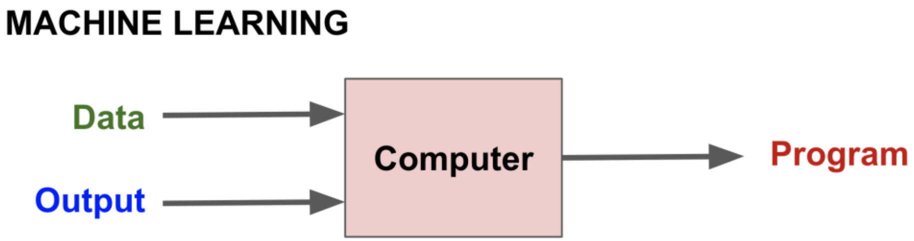
- Data Driven Research At USC's Data Science Lab (DSLab)
- **Machine Learning (ML) 101**
- Basics of Prediction
- Examples Of Prior Work At DSLab
  - Smart Grid
  - Smart Oil Field
  - Internet Traffic Prediction
  - Chikungunya Epidemic Prediction
- Covid-19 Epidemic Prediction
  - Models
  - Our approach
  - Results
  - Role of ML
- Covid-19 Interactive Visualization Tool
- Concluding Remarks
- Q & A

What Is Machine Learning?

Definition 1: Field of study that gives computers the ability to learn without being explicitly programmed [A. Samuel ‘59]

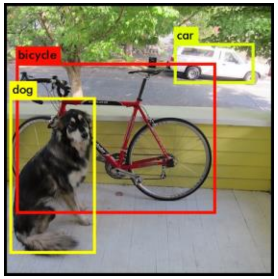
Definition 2: A computer program that improves its *performance* at some task through *experience* [T. Mitchel ‘97]

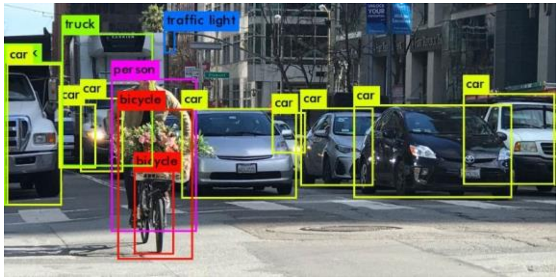




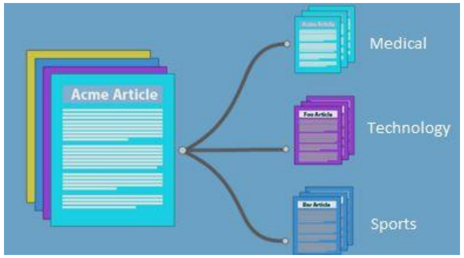
Examples Of Machine Learning Applications

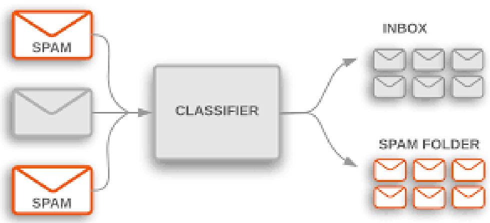
- Object recognition
  - Recognize and tag objects in images



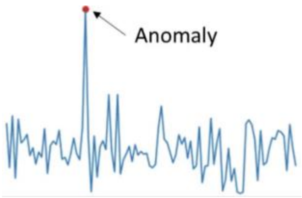


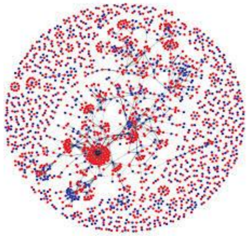
- Document classification
  - Assign a topic to a document
  - Spam Email Detection





- Other applications
  - Anomaly detection
  - Fraud Detection
  - Playing games (e.g. Go)

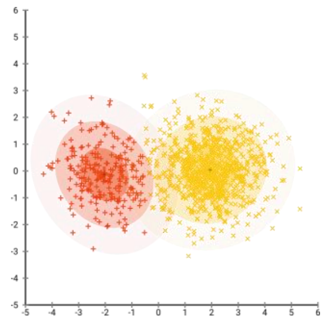




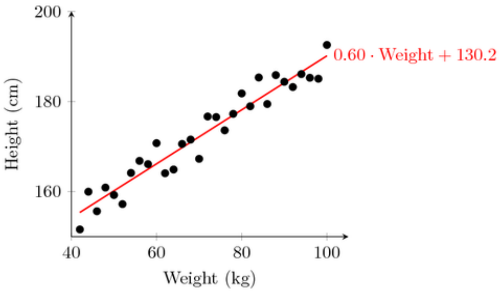
Main Categories Of Machine Learning Algorithms

- Classification
  - Assign a category to each input item (e.g. assign documents to categories)
    - For each input we get its output label

- Clustering
  - Partition a list of input items into homogeneous groups
    - For each input, we only get the group ID it belongs to, without any other label attached

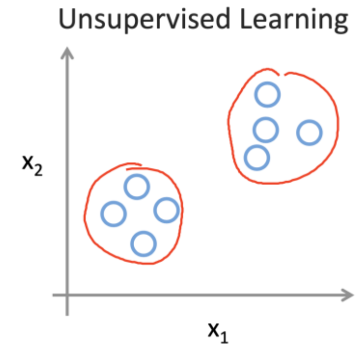


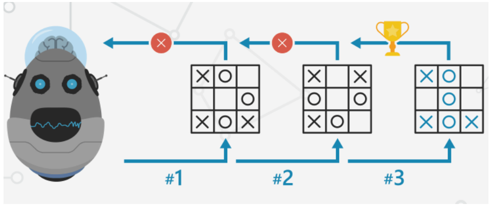
- Regression
  - Predict a real value for each input item (e.g. predict height based on weight)
    - We get a curve that fits the data



Main Types Of Learning

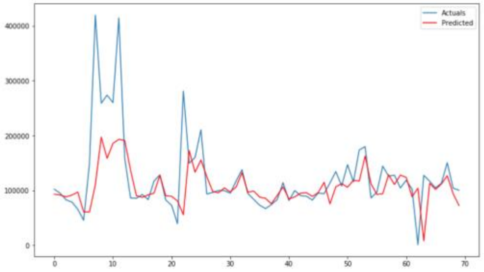
- Supervised Learning
  - The learner receives a set of "labelled" data that represent the correct answers (i.e. outputs) for each input
  - data in the form: (input, correct output)
    - E.g. classification and regression algorithms
- Unsupervised Learning
  - The learner does not receive "labelled" data, and makes predictions for all unseen points
  - data in the form: (input, ?)
    - E.g. clustering algorithms
- Reinforcement Learning
  - The learner actively interacts with (and potentially affects) the environment and receives an immediate reward for each action
  - The learner maximizes the long-term reward received
  - data in the form = (input, some output, reward for this output)
    - e.g. an algorithm to play Tic-tac-toe





Important Machine Learning Terminology

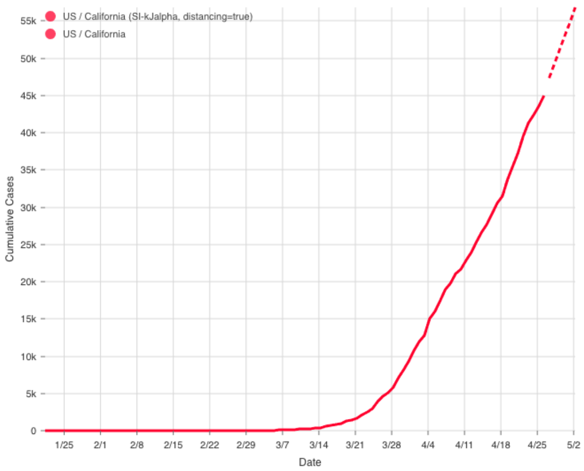
- Training
  - The process of creating a machine learning algorithm
- Inference
  - The process of using a trained machine learning algorithm to draw conclusions
    - e.g. make a classification decision, or a real value prediction etc.
- Hyperparameters
  - Parameters that are not determined by the learning algorithm, but rather specified as inputs to the learning algorithm, before learning begins
- Accuracy
  - The ratio of correct predictions over all the predictions (e.g. in classification tasks)
- Mean Absolute Percentage Error (MAPE)
  - The average percentage difference of the predicted values from the correct ones (e.g. to assess regression quality)
- Root Mean Square Error (RMSE)
  - Another common error metric like MAPE



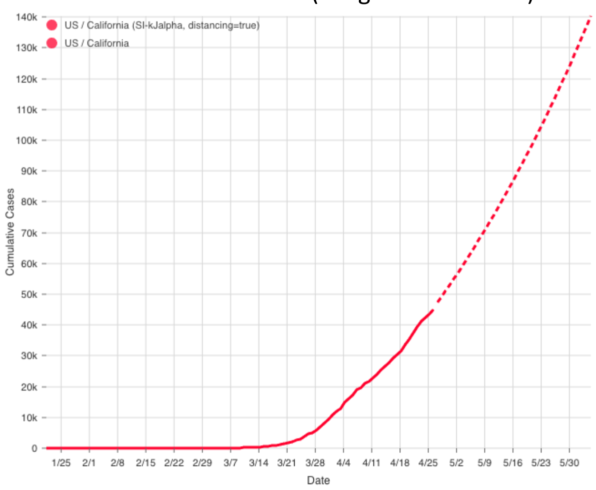
Basics of Prediction

- Prediction
  - a statement about a future event
    - e.g. predict the next value of a time-series
- Predictions can be classified as
  - Short-term vs. Long-term, depending on the length of the prediction window
    - e.g. next time epoch vs the next multiple epochs
  - What is considered "short" or "long" also depends on the application and the data availability/aggregation process
    - e.g. COVID-19 data are published once per day, so short-term can be few days, and long-term can be > 1 week
- Uncertainty increases with the prediction horizon
  - Nevertheless, both short and long term predictions can be very **important** for
    - planning, resource allocation, policy making, and general decision making under uncertainty

COVID-19 Cases (Short-Term Horizon)



COVID-19 Cases (Long-Term Horizon)

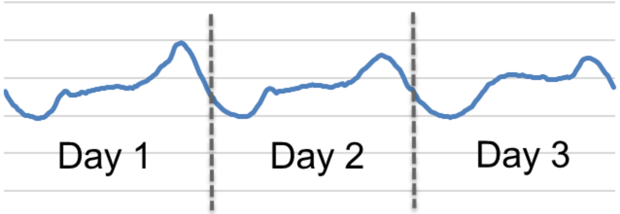


Outline

- Data Driven Research At USC's Data Science Lab (DSLab)
- Machine Learning (ML) 101
- Basics of Prediction
- **Examples Of Prior Work At DSLab**
  - Smart Grid
  - Smart Oil Field
  - Internet Traffic Prediction
  - Chikungunya Epidemic Prediction
- Covid-19 Epidemic Prediction
  - Models
  - Our approach
  - Results
  - Role of ML
- Covid-19 Interactive Visualization Tool
- Ongoing Work
- Concluding Remarks
- Q & A

Temporal Ensemble Learning for Load Prediction in Smart Power Grids

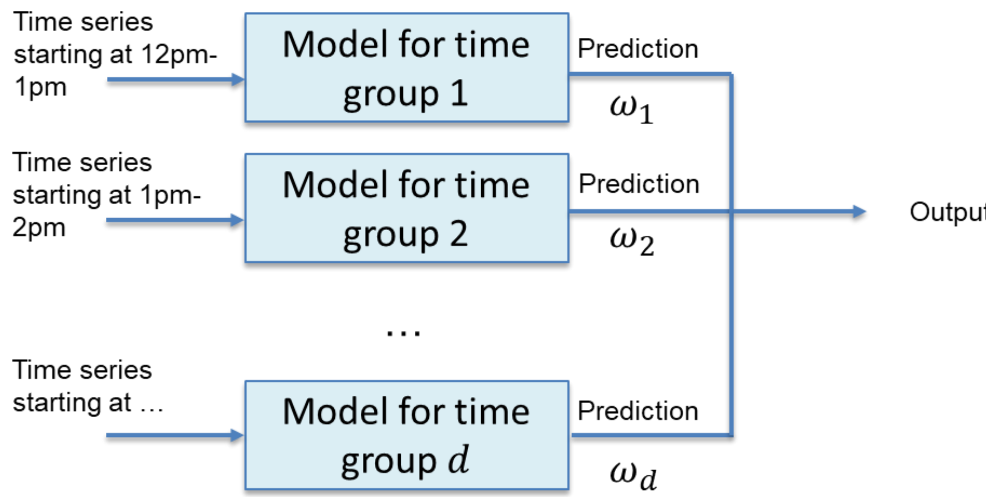
Problem: Given historical power consumption data, predict future power consumption



Key Observation: Daily Periodicity in Consumption Data

Methodology - Temporal Ensemble Learning

- Train Specialized models for specific time of day – **Temporal Features**
- Output a weighted sum of temporal features as predicted value - **Ensemble**
- **ML Models –** Kernel Regression (KR), Support Vector Regression (SVR)



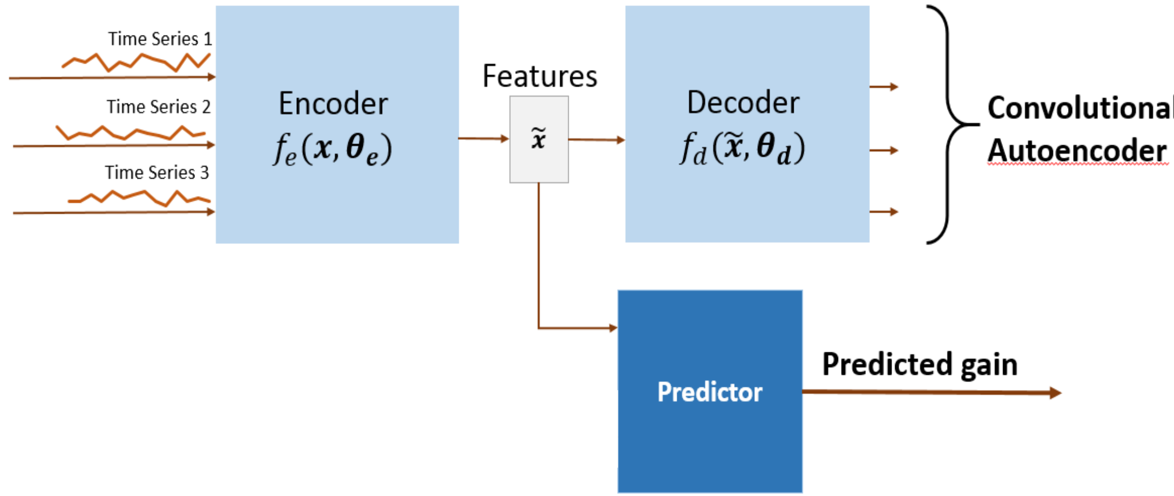
	Ensemble		Single Model	
	MAPE(%)	RMSE	MAPE(%)	RMSE
KR	1.03	124.41	1.16	158.13
SVR	1.05	126.41	1.50	186.97

Temporal Ensemble Models achieve high accuracy (1-2% error rate) compared with traditional prediction methods (ARIMA, NYISO, etc. which achieve ~5% error rate)

OReONet: Deep Convolutional Network for Oil Reservoir Optimization

Steam job Candidate Selection

- Injecting steam increases well temperature → increases production
- Predict the benefit of steam jobs on wells
- “benefit” = the gain obtainable from performing a steam job



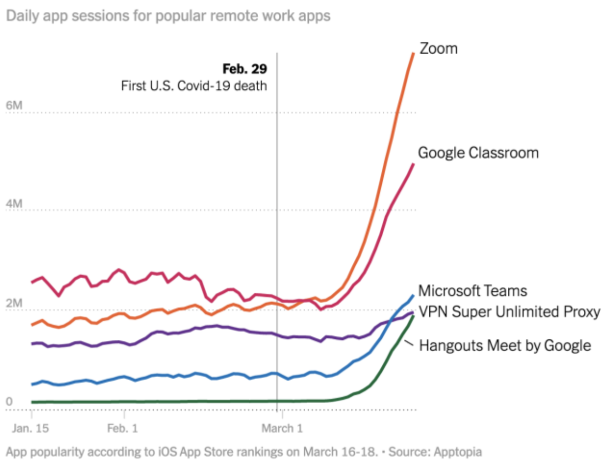
	Linear Regression		Support Vector Regression		Kernel Regression	
	MF	AF	MF	AF	MF	AF
Mean Squared Error	361.4	351.2	15.47	2.61	10.4	0.59
Overlap coefficient(50)	0.06	0.18	0.44	0.68	0.36	0.74
Precision@50	0.14	0.26	0.6	0.98	0.56	0.98

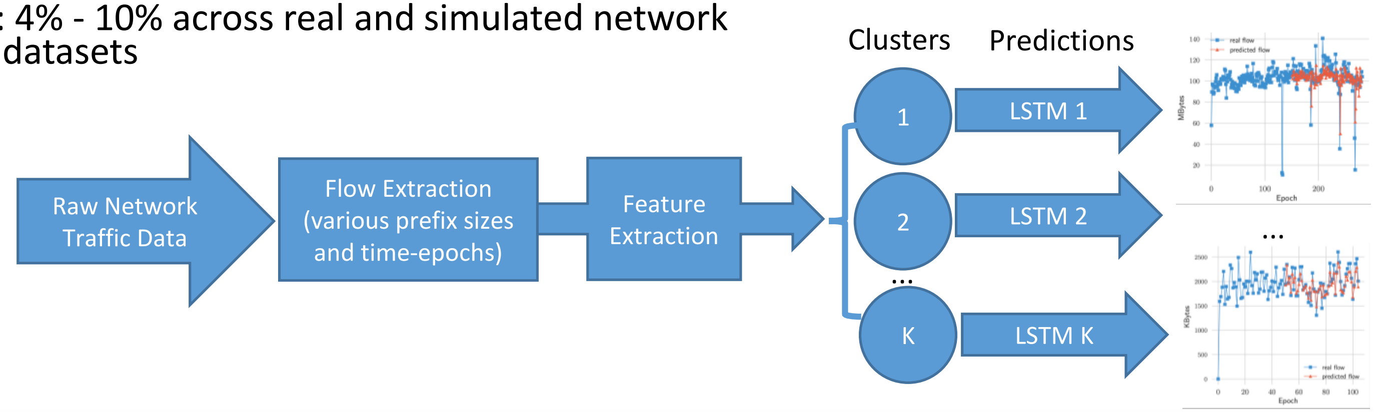
Average gain of 176% compared to 124% achieved by on field operators

Internet Traffic Prediction

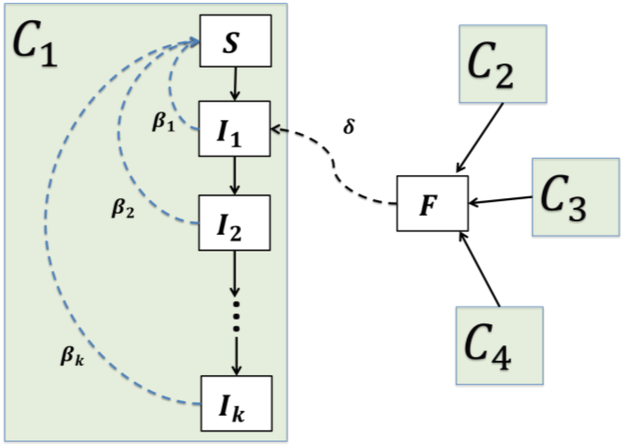
- Predicting network traffic in short time scales is very important for
  - Traffic engineering
  - Power savings in Data Center or backbone ISP networks
  - Improved Quality-of-Experience (QoE) at the end-user
- Particularly useful during Covid-19 due to rapid multimedia growth (Zoom, Netflix, etc.)
- Developed clustering-based LSTM prediction models
  - Groups network time-series into similarity groups and then model them with a specialized model for each group
  - MAPE: 4% - 10% across real and simulated network traffic datasets

Internet Traffic Growth During Covid19





DARPA Grand Challenge – CHIKV (2014-2015)



Heterogeneous infection rate model with human mobility

CHIKV epidemic: Country-level predictions. Weekly over 8 months, 55 countries

One of 10 winners of DARPA Grand Challenge 2015 for predicting CHIKV epidemic

Ajitesh Srivastava, "*Computing Cascades: How to Spread Rumors, Win Campaigns, Stop Violence and Predict Epidemics*", PhD Thesis, USC August 2018

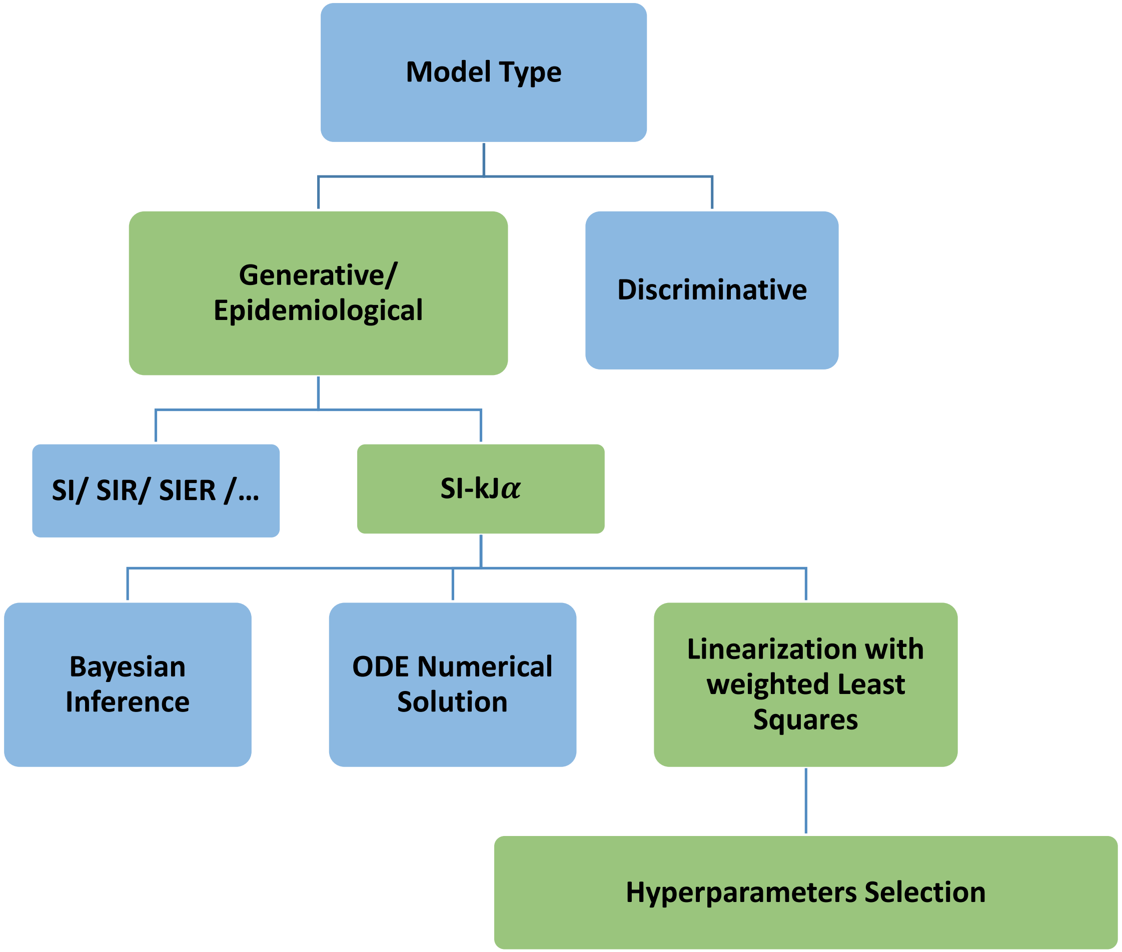
Outline

- Data Driven Research At USC's Data Science Lab (DSLab)
- Machine Learning (ML) 101
- Basics of Prediction
- Examples Of Prior Work At DSLab
  - Smart Grid
  - Smart Oil Field
  - Internet Traffic Prediction
  - Chikungunya Epidemic Prediction
- **Covid-19 Epidemic Prediction**
  - Models
  - Our approach
  - Results
  - Role of ML
- Covid-19 Interactive Visualization Tool
- Ongoing Work
- Concluding Remarks
- Q & A

Why Forecast?

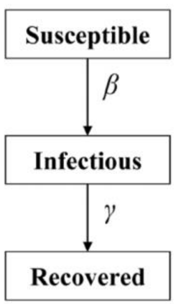
- Preparedness and resource management needs **state/county/city level predictions**:
  - How many masks, testing kits, beds are needed tomorrow/next week at a given hospital
  - How to distribute state/country resources across all the hospitals in a state/country
- How do we come out of “stay-at-home” order?
  - Should some venues remain closed and some open, initially?
- Need accurate forecasts for simulation of future scenarios

Modeling Choices



SIR Model

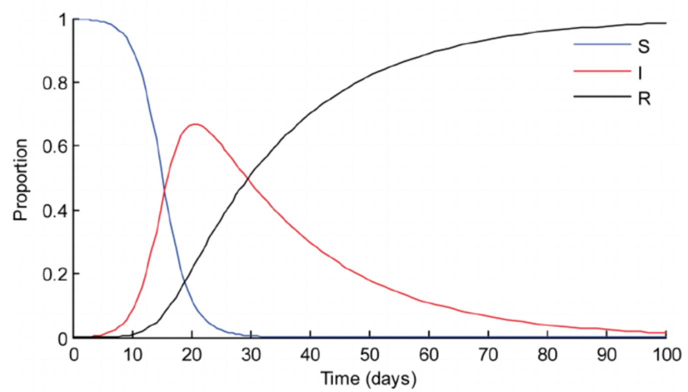
- SIR model has been used to study the spread of various infectious diseases such as measles, mumps, and rubella
  - S: the number of susceptible
  - I: the number of infectious
  - R: the number of recovered or deceased (or immune) individuals

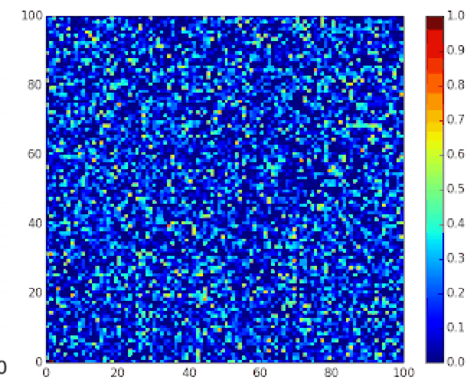


$$\frac{dS}{dt} = -\beta SI$$

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

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





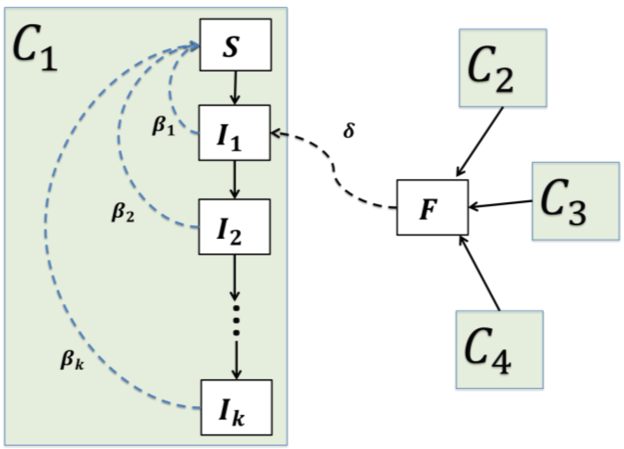
https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology

Heterogeneous Infection Rate with Human Mobility

$$\Delta I_t^p = \frac{S_{t-1}^p}{N^p}\sum_{i=1}^{k} \beta_i^p \left(I_{t-iJ}^p - I_{t-(i-1)J}^p\right) + \delta \sum_q F(q,p)\frac{\sum_{i=1}^{k} \beta_i^q \left(I_{t-iJ}^q - I_{t-(i-1)J}^q\right)}{N^q}$$

Community spread

Travel spread



$$\beta^p = \begin{bmatrix}\beta_1^p & \dots & \beta_k^p & \delta^p\end{bmatrix}$$

$$\text{And, } \mathbf{X}_t^p = \begin{bmatrix} S_t(I_t^p - I_{t-J}^p) \\ \vdots \\ S_{t-(k-1)J}(I_{t-(k-1)J}^p - I_{t-kJ}^p) \\ \sum_q \frac{F(q,p)}{N^q}(I_t^q - I_{t-kJ}^q) \end{bmatrix}^T.$$

$$\Delta I_t^p = \beta^p \mathbf{X}_t^p$$

Learning with weighted least square minimization

$$\sum_{t=1}^{T}\left(\alpha^{\frac{T-t}{2}}\Delta \hat{I}_t^p - \alpha^{\frac{T-t}{2}}\beta_p \mathbf{X}_t^p\right)^2$$

Decaying weights on past data

Results: Short-term Predictions (1)

- Using data by April 10th (not including travel)

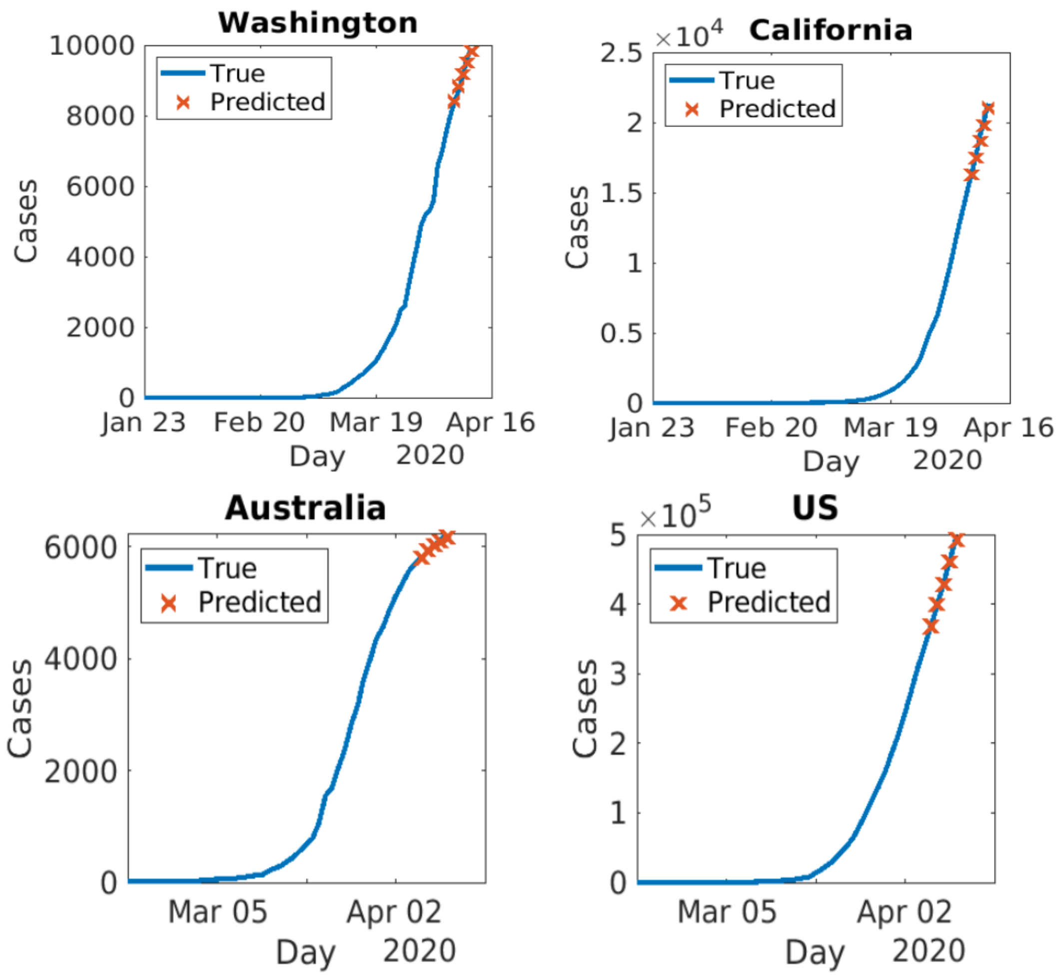
	Method	RMSE (US)	MAPE (US)	RMSE (Global)	MAPE (Global)
Adaptive	SI-kJα (variable)	333.3	6.82%	462.6	13.64%
Adaptive	SI-kJα (fixed)	342.05	6.58%	456.0	**11.22%**
Adaptive	SI-kJα (ensemble)	**316.3**	**5.93%**	**355.9**	11.37%
Single curve fitting	Gen-SEIR	2106.4	14.31%	7471.2*	41.06%*

- Using data by March 21st including travel data

Travel data improved the models

Method	US RMSE	US MAPE	Global RMSE	Global MAPE
travel, variable	**147.3**	19.93%	248.4	21.353%
without travel, variable	166.7	**18.51%**	348.2	23.15%
travel, fixed	207.0	25.08%	**242.6**	**19.50%**
without travel, fixed	186.6	19.52%	286.8	21.42%

Results: Short-term Predictions (2)



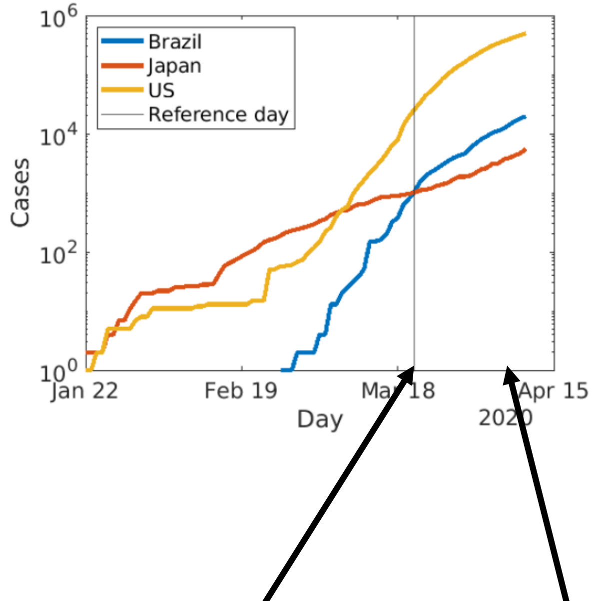
Measuring the Present, using the Past, through Predictions

Objective: Asses the effects of a region's effort to battle COVID-19, for example, contact reduction

Approach Idea: Model is adaptive, captures the changes → compare models

- Define $\tau^p \propto \sum \beta_i^p$ for region p with rate of infection defined using model parameters β_i^p
- Calculate τ_0^p for a reference date t_0 and τ_1^p for t_1 ($t_1 > t_0$)
- **Contact Reduction Score (CRS)** defined as:

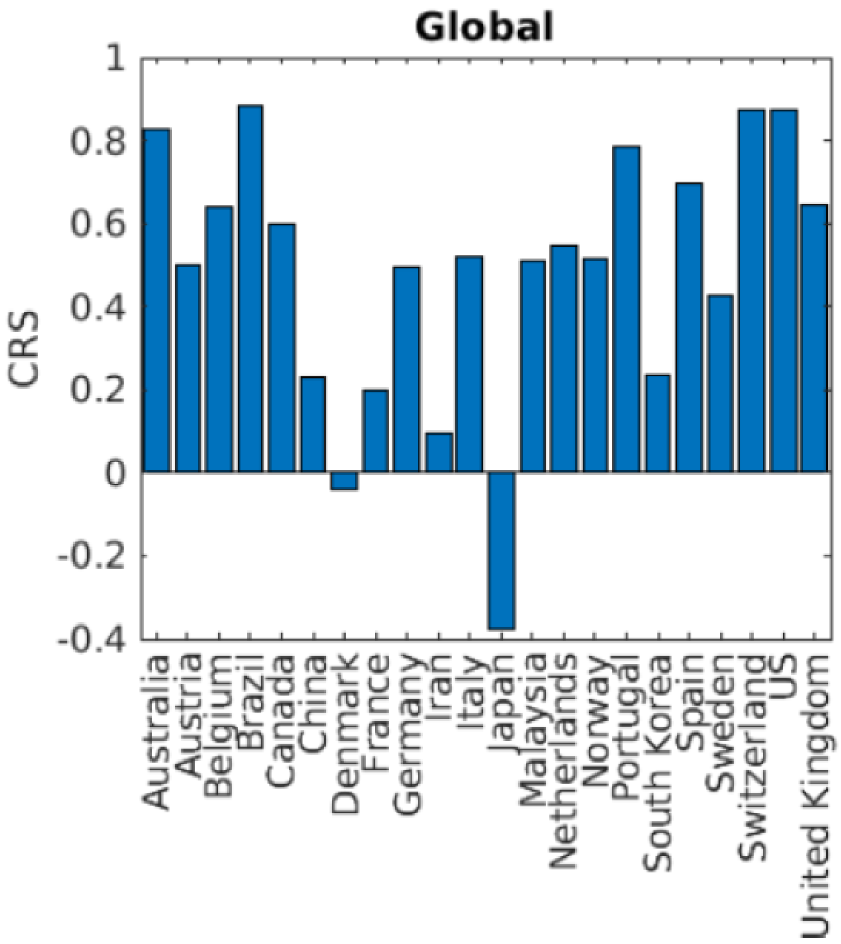
$$\frac{\tau_0^p - \tau_1^p}{\tau_0^p}$$



t_0: τ_0^p

t_1: τ_1^p

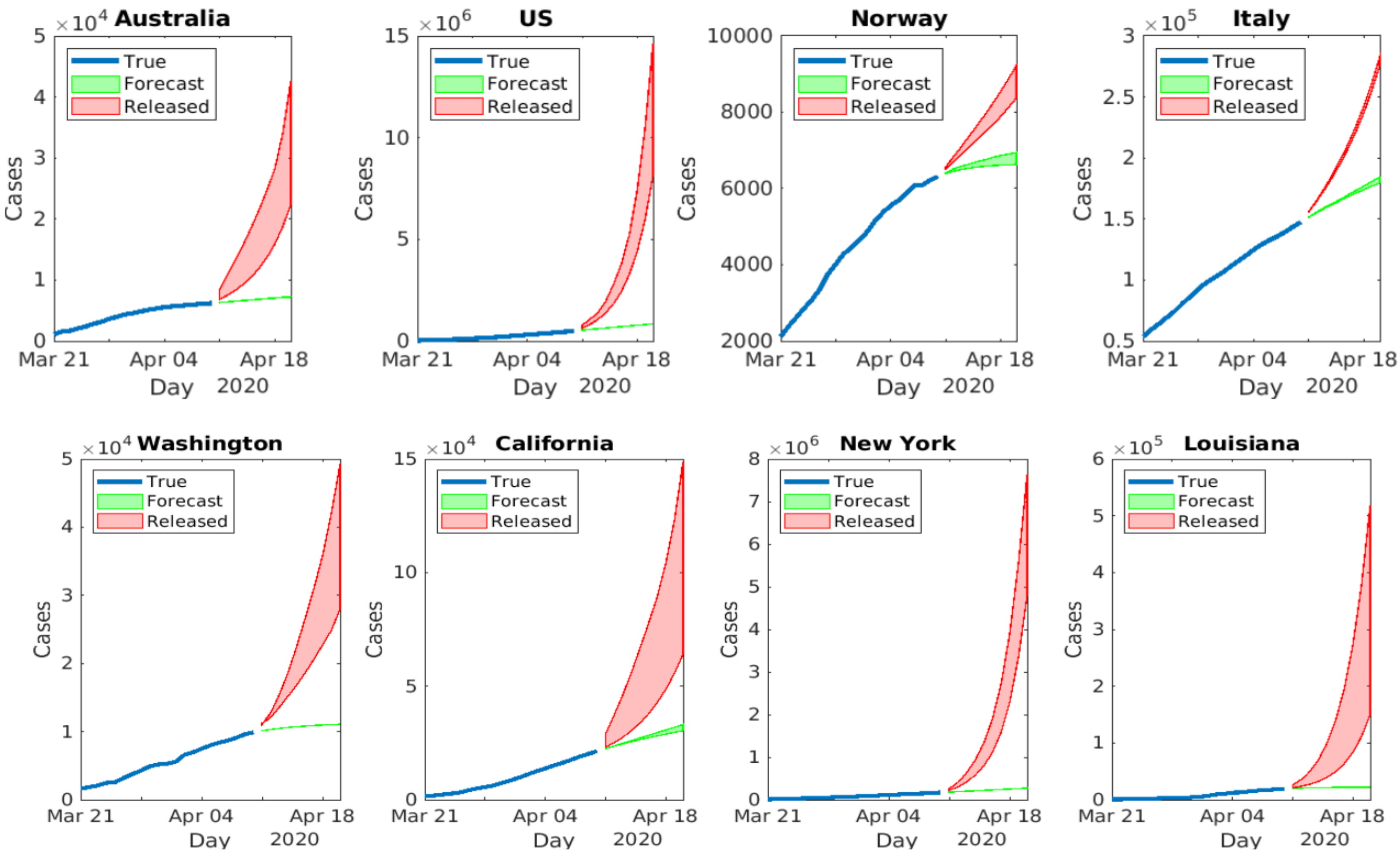
CRS for Global (March 21st-April 10th)



CRS – higher implies better

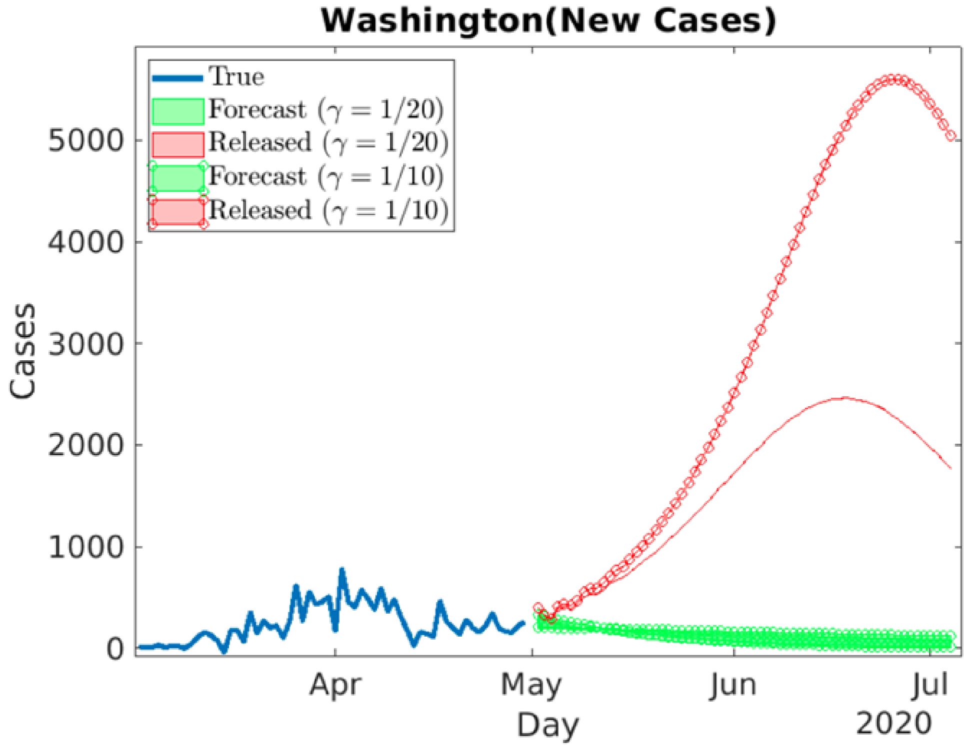
Best CRS: Brazil, Worst CRS: Japan

Forecasts and “What-if”



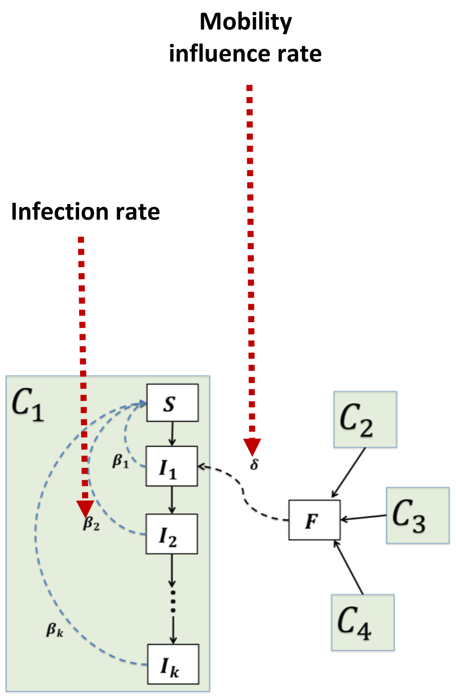
Scenarios Accounting for Unreported Cases

- Model can capture unreported cases as an input from antibody studies
  - With probability γ a COVID case is reported
  - $\gamma = \frac{\text{Reported Cases}}{\text{Estimated Total Cases}}$



The Role Of ML In Covid-19 Predictions

- Traditional SIR models rely on simplified assumptions (e.g. no mobility) and numerical solutions to differential equations
  - Cannot capture the complex mobility patterns and evolving trends
- Our approach is ML based
  - Supports different infection rates depending on how many days one has been infected
  - Learn optimal parameters $\beta_1, \beta_2, \ldots, \beta_K$, and δ using Weighted Least Squares
  - Introduced smoothing to avoid overfitting
  - Adapt to rapid Covid-19 related policy changes that affect future data by using a forgetting factor $\alpha < 1$ during training
    - Give more weight to more recent data
  - Leverage real mobility datasets (flight data)
- Thus, more accurate predictions can be achieved, outperforming the baselines

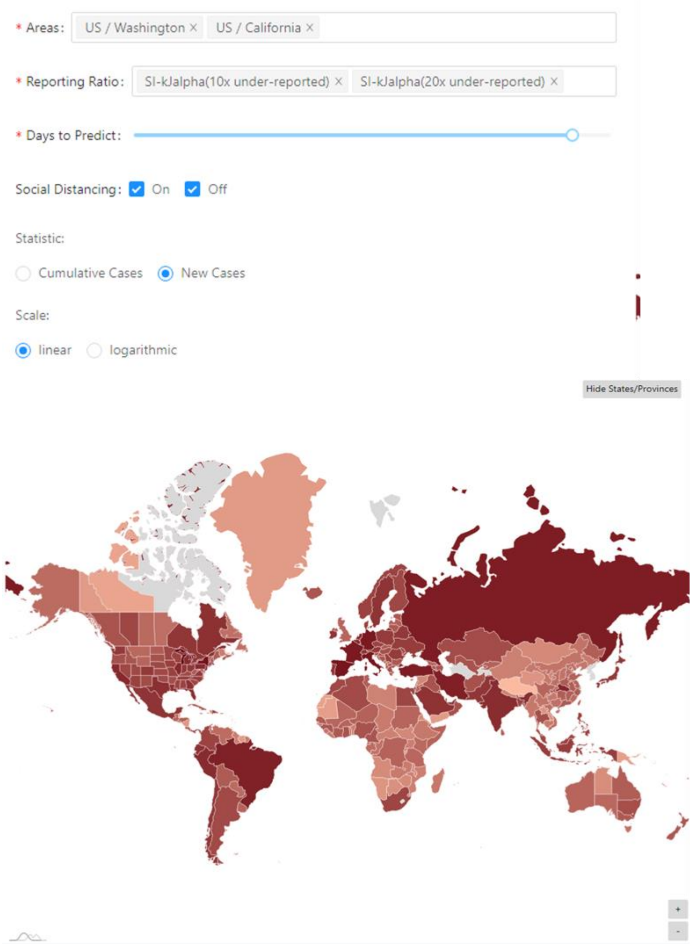


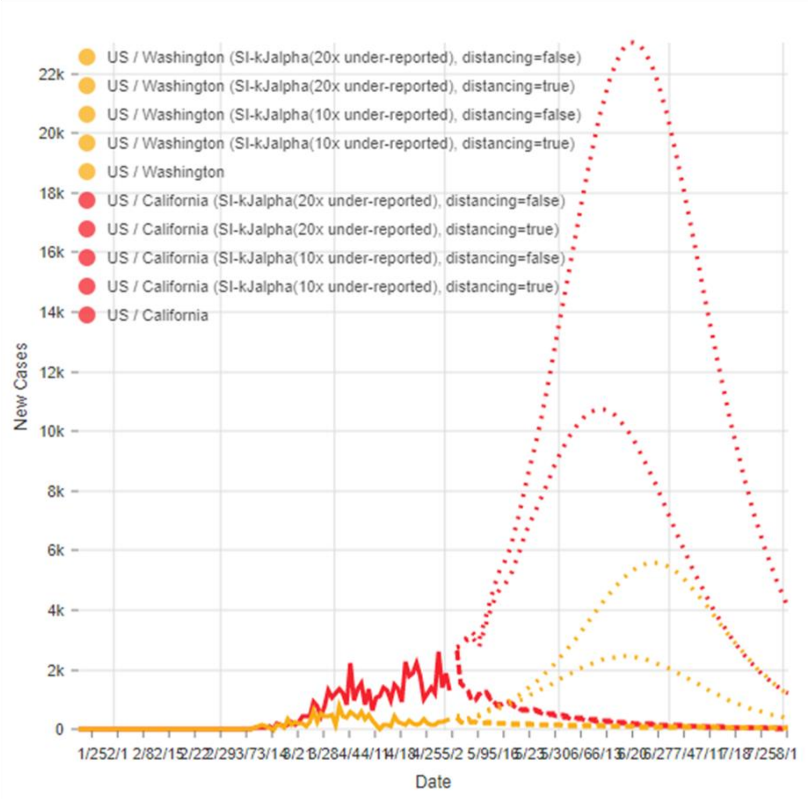
Outline

- Data Driven Research At USC's Data Science Lab (DSLab)
- Machine Learning (ML) 101
- Basics of Prediction
- Examples Of Prior Work At DSLab
  - Smart Grid
  - Smart Oil Field
  - Internet Traffic Prediction
  - Chikungunya Epidemic Prediction
- Covid-19 Epidemic Prediction
  - Models
  - Our approach
  - Results
  - Role of ML
- **Covid-19 Interactive Visualization Tool**
- Ongoing Work
- Concluding Remarks
- Q & A

Forecasting Web Interface

https://jaminche.github.io/COVID-19/





Predictions: Next Steps

- County/city/neighborhood level predictions

- Hybrid hyperparameter/parameter learning scheme
  - Current approach: Each has its own or everyone uses the same hyperparameters
  - Clusters of regions share hyperparameters and even parameters: Consider similar regions when data for given region is not enough

- Incorporating Unreported Cases

Beyond Predictions

- Resource allocation
  - Optimal distribution of test and protection resources
  - Under continued lockdown

- Network diffusion/immunization
  - How to limit mobility so the epidemic is contained
  - …
- Lessons learned for the future
  - Generalized models
  - …

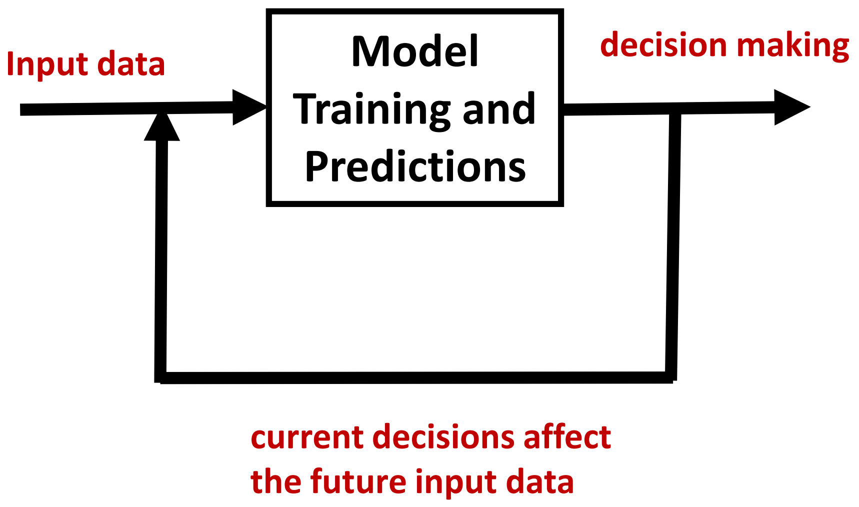
Acknowledgments

- **NSF RAPID:** ReCOVER: Accurate Predictions and Resource Allocation for COVID-19 Epidemic Response (Prasanna and Srivastava)

- Initial Sprint
  - Frost Tianjian Xu (Sophomore, CS): Dataset preparation
  - Jamin Chen (Senior, CS): Integrating our methods into a web-based visualization
  - Prathik Rao (Junior, CE) and Kangmin Tan (Junior, CS): Implementing and evaluating various ML training approaches

Concluding Remarks

- Good hyperparameter selection is critical
- Models should evolve with data
  - Current decisions can affect future input data
    - Feedback loop
  - Need to retrain the models and adapt to changes
- Ensemble approach likely to be the best approach
  - Combine the results from several models instead of one



Thanks

Be Safe

prasanna@usc.edu

dslab.usc.edu