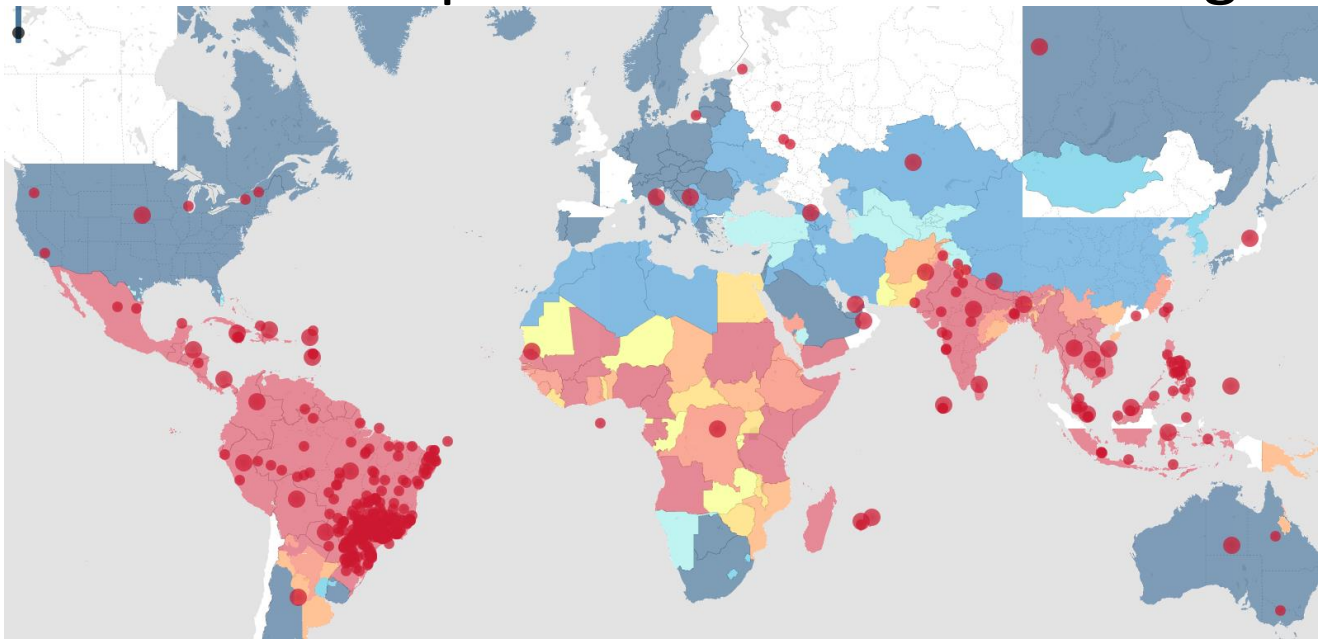
Deng-AI: Predicting Disease Spread

Predicting the spread of Dengue Fever using Time Series Techniques

Greg Condit

Problem Overview

- Dengue Fever is a disease with severity ranging from flu-like symptoms to low blood pressure and death.
- It is not contagious; Dengue Fever can only be spread by mosquitoes
- Typically observed in tropical regions, but cases have increased significantly in recent years, and scientists are warning that climate change is likely to produce shifts that enable mosquitos to cover a much larger region

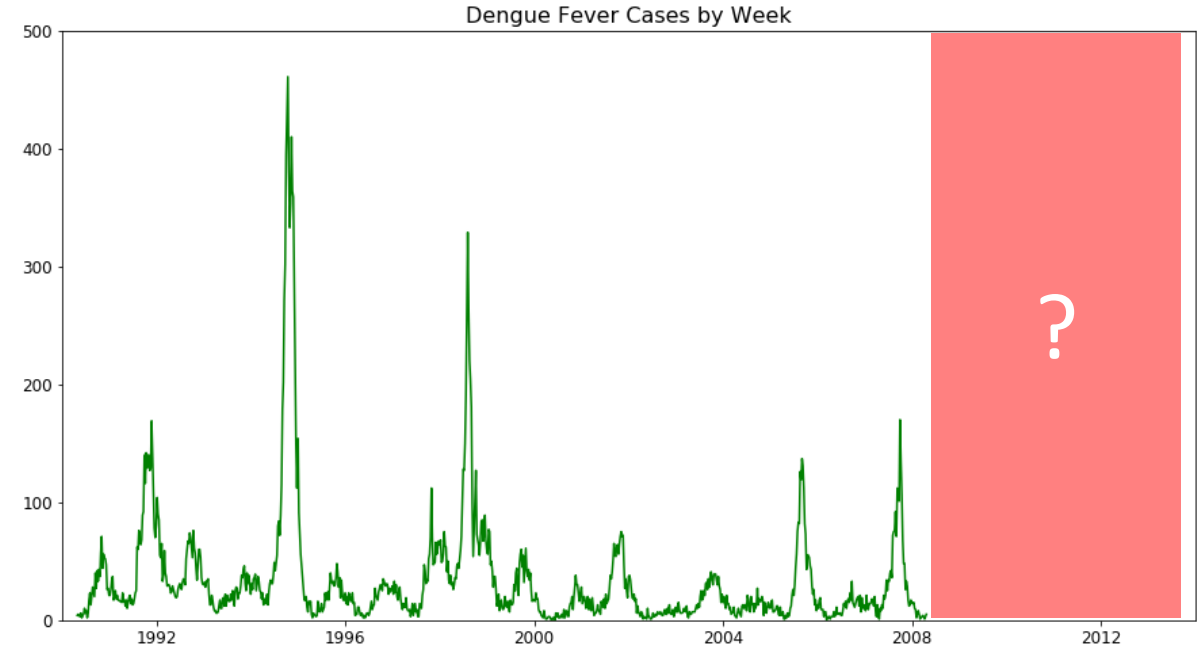


Map of CDC outbreaks

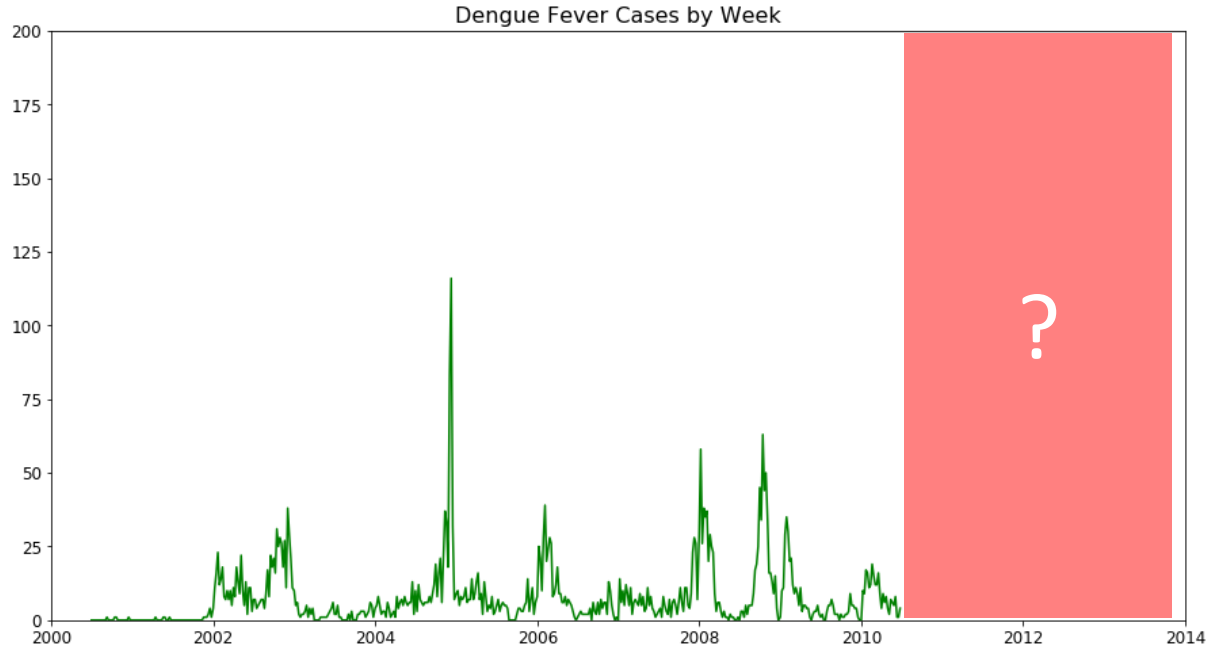
DrivenData.org Competition

- DrivenData.org: “bring cutting-edge practices in data science and crowdsourcing to some of the world's biggest social challenges”
- Deng-AI – Machine Learning Competition to predict outbreaks of Dengue Fever

San Juan, Puerto Rico

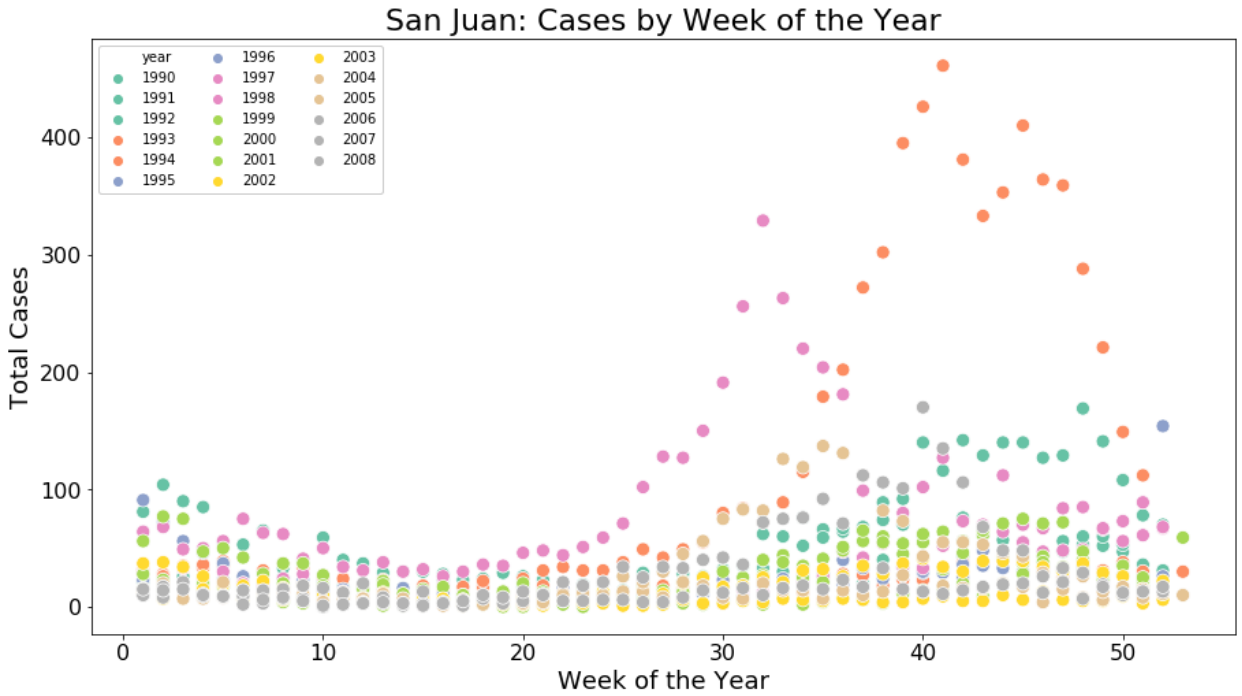


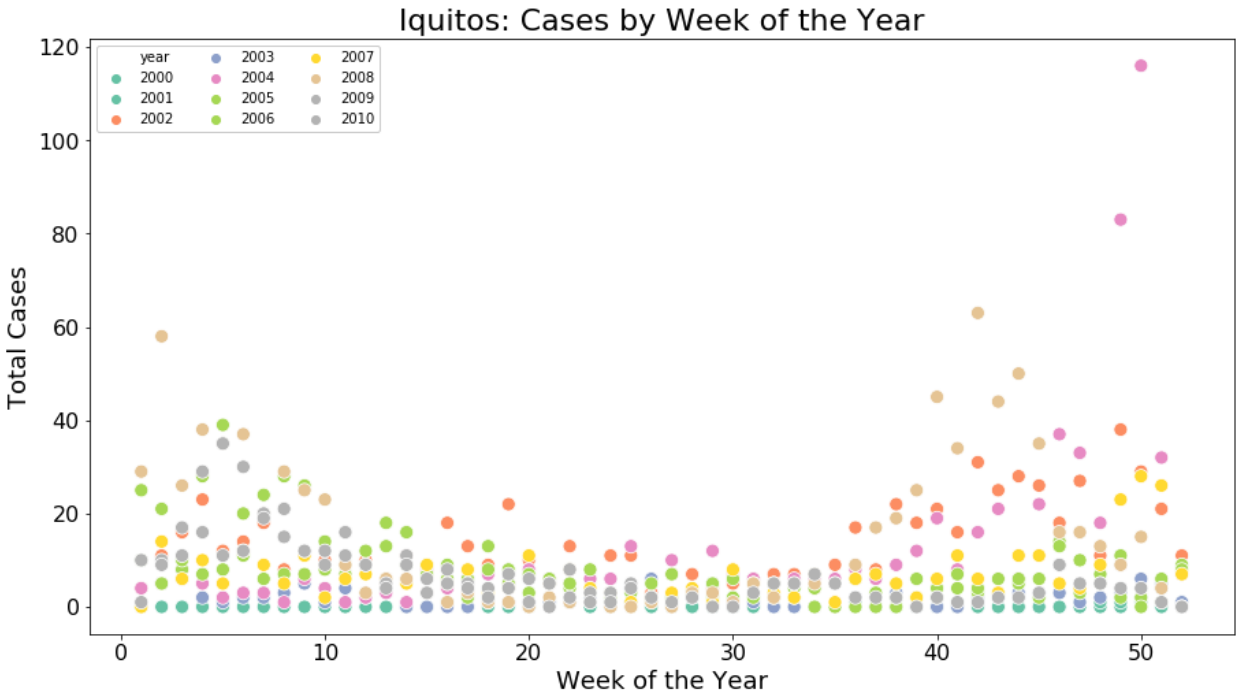
Iquitos, Peru



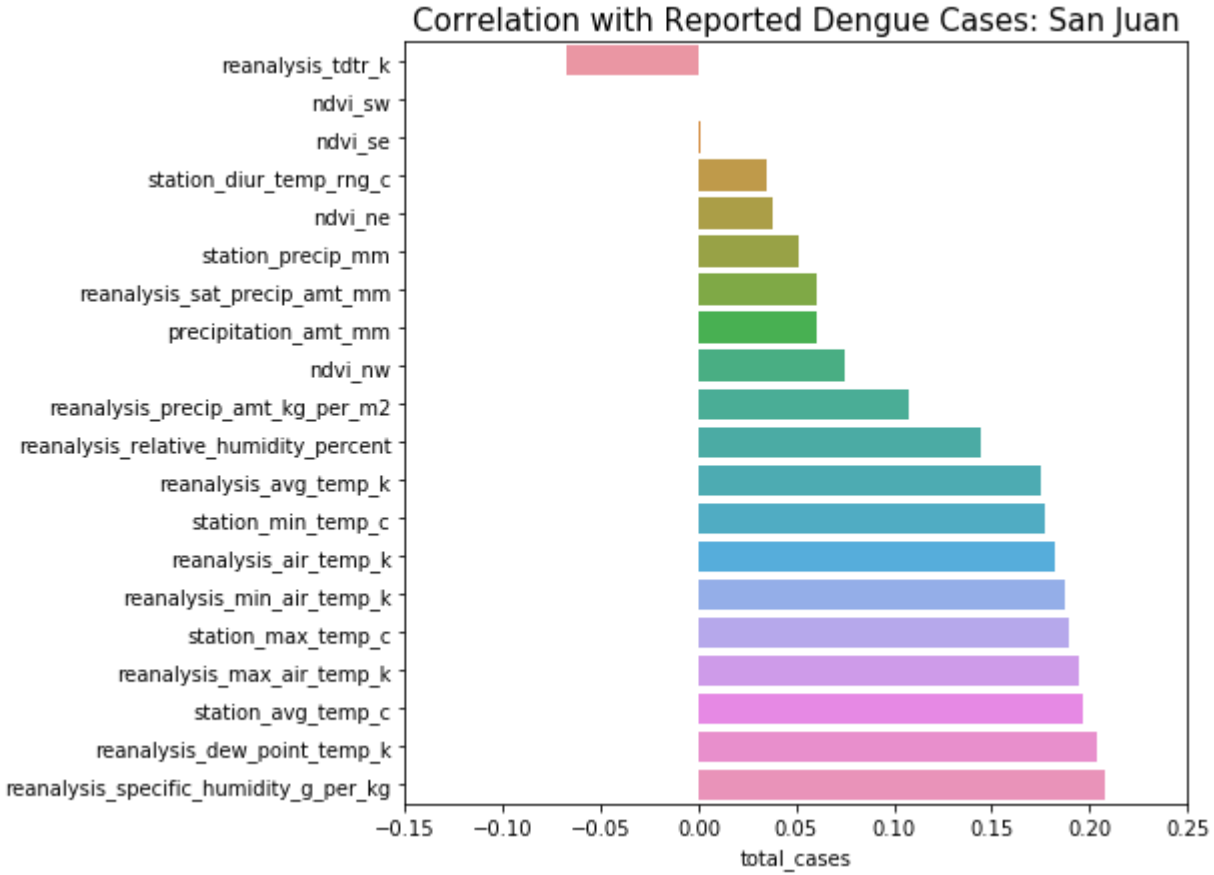
Hypothesis

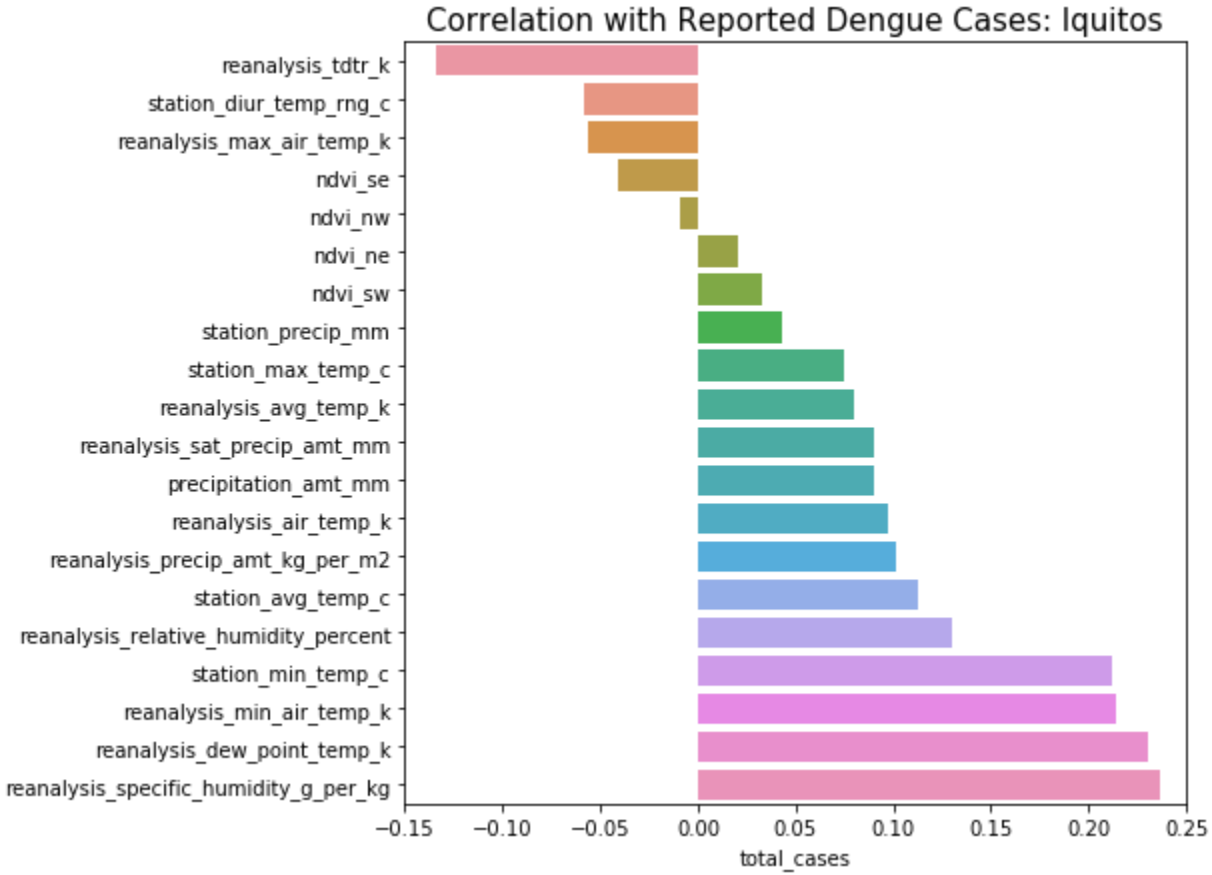
Dengue is spread by Mosquitos, whose breeding patterns are related to weather patterns. Therefore, weather patterns can predict outbreaks.





Features: Lots of weather data

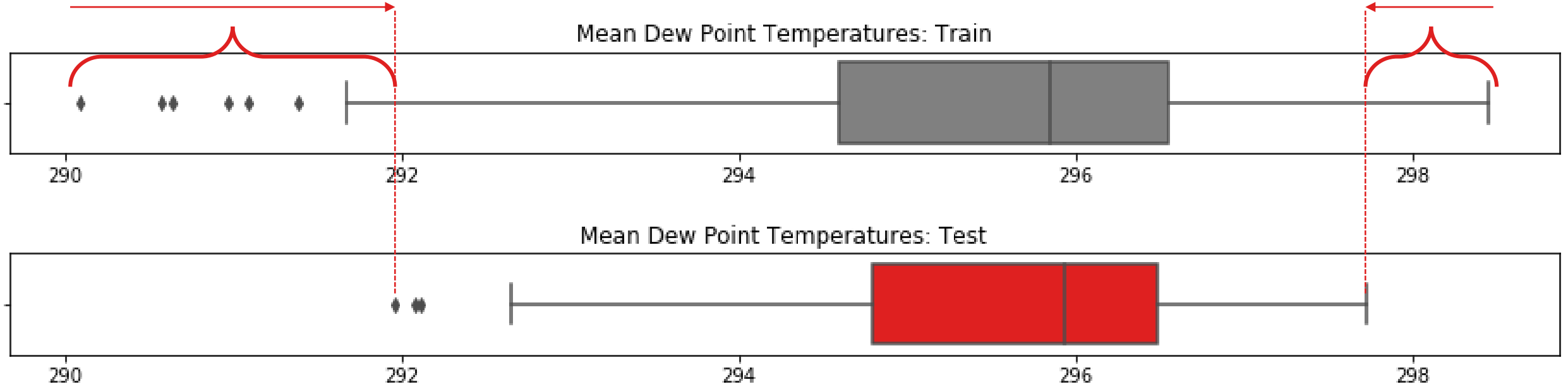




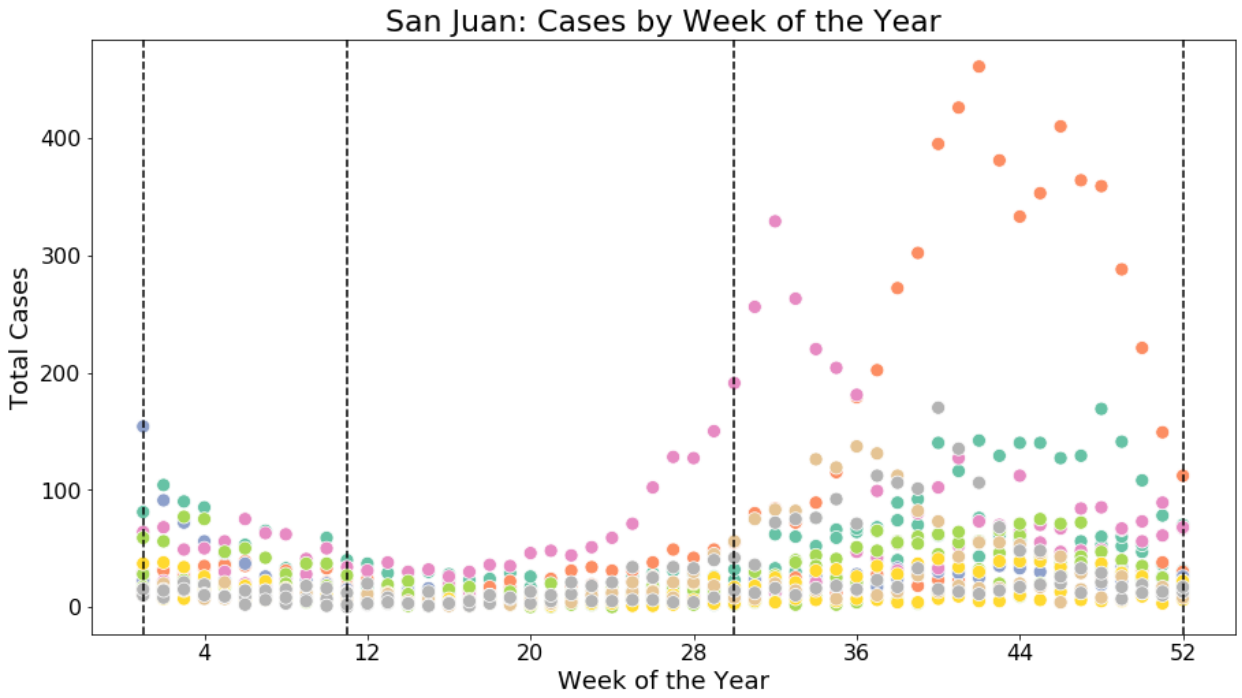
Pearson correlations do not account for sequence

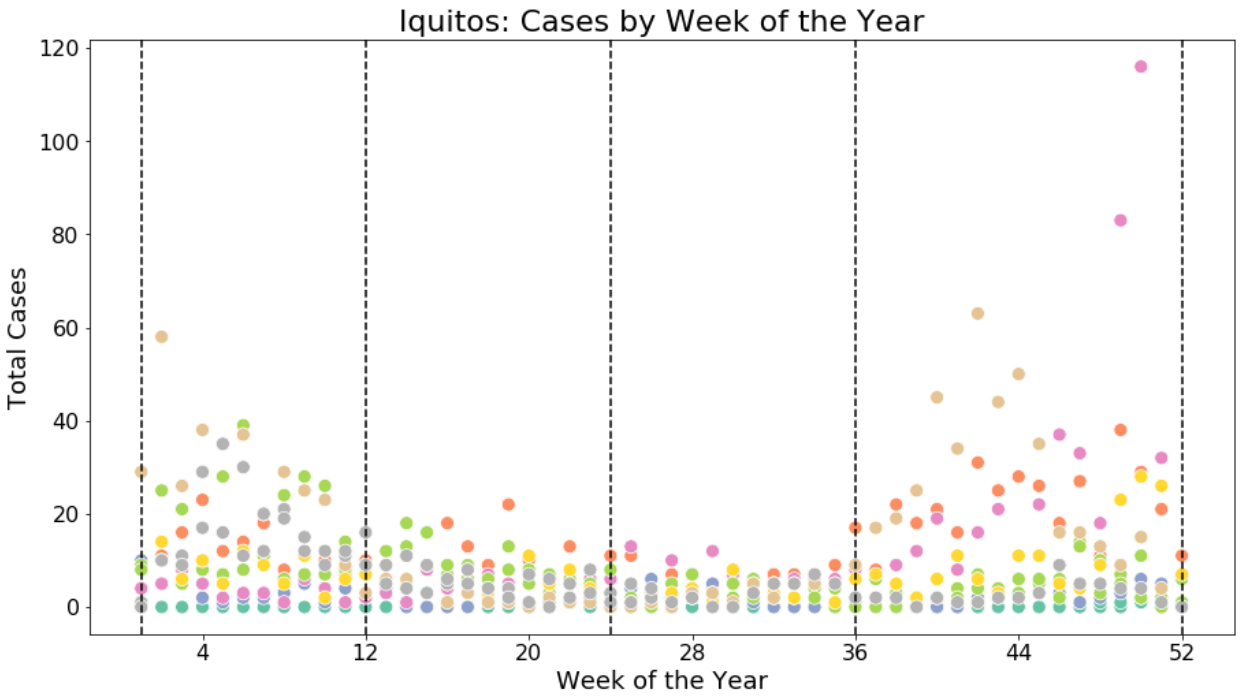
Feature Engineering: Degrading Training Data

- Training minimums raises to Test minimums for each city
- Training maximums raised to Test maximums for each city
- Full scaling was avoided based on research that the actual temperature points matter (see sources)

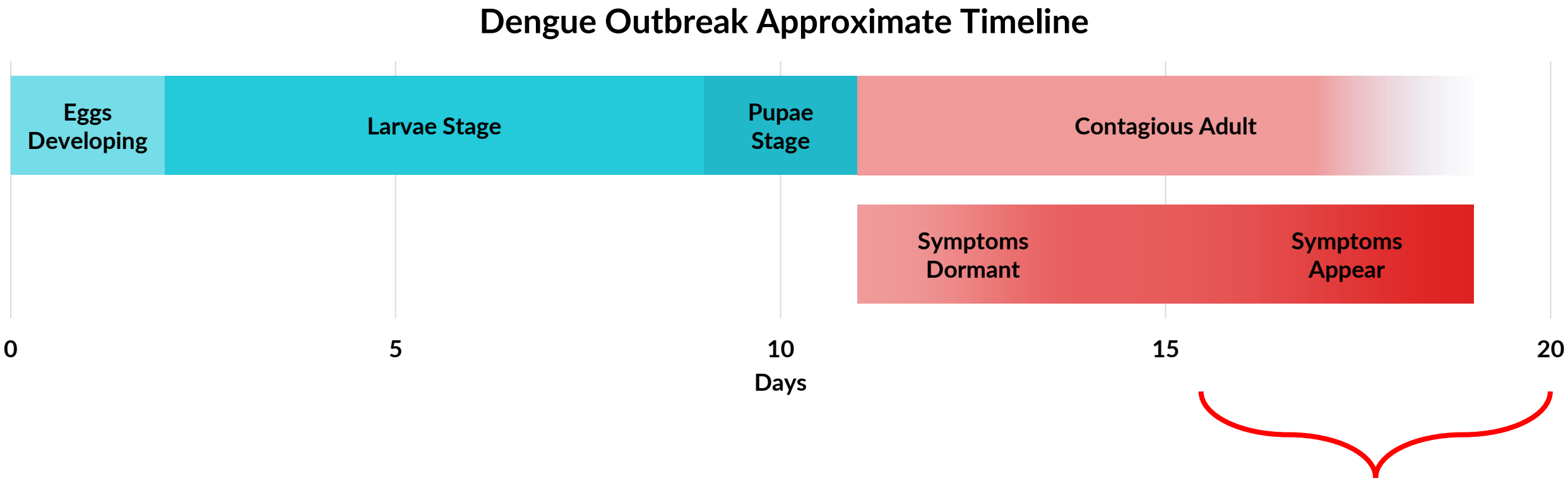


Feature Engineering: One-hot Encoded Seasons





How far back should a time series model “look”?

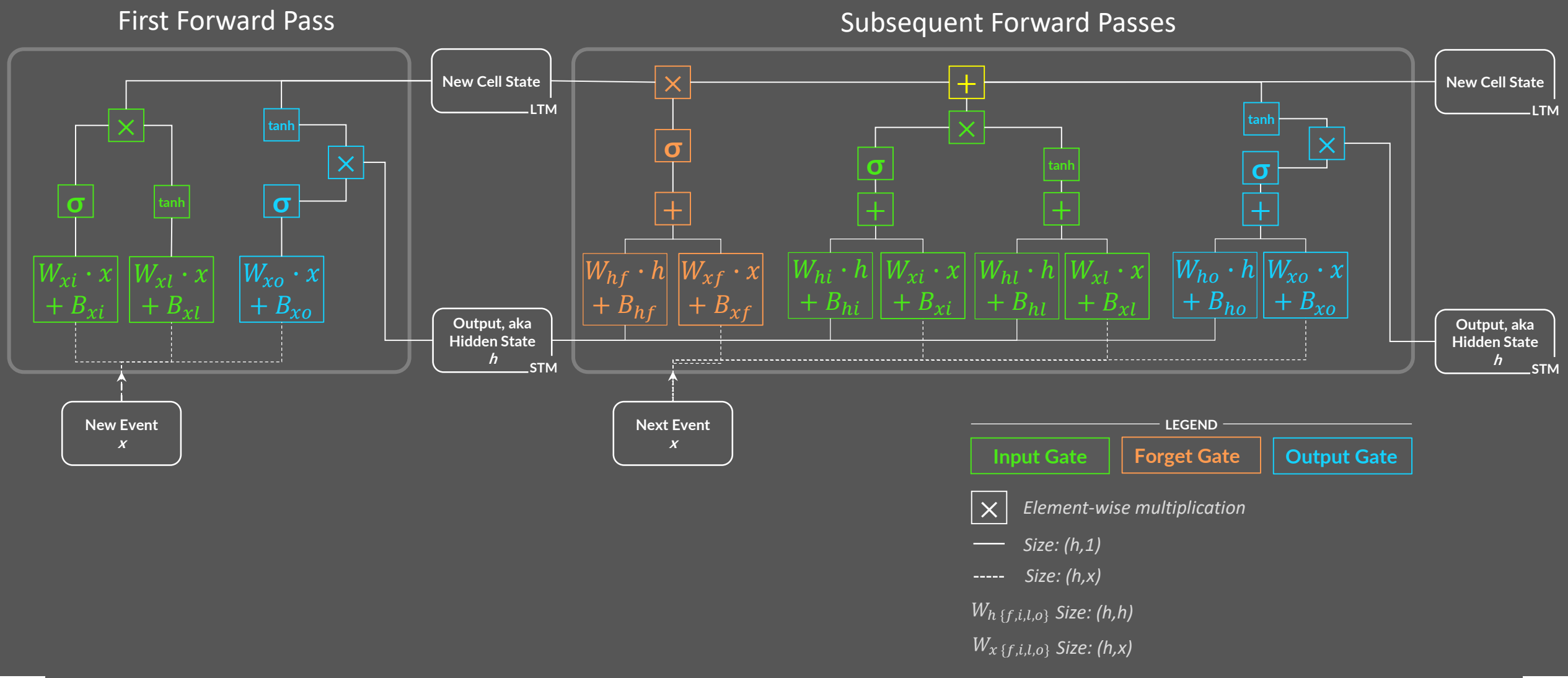


Following a “favorable” weather event, we’d expect to see reports of Dengue Fever in 2-3 weeks.

2 Time Series Techniques

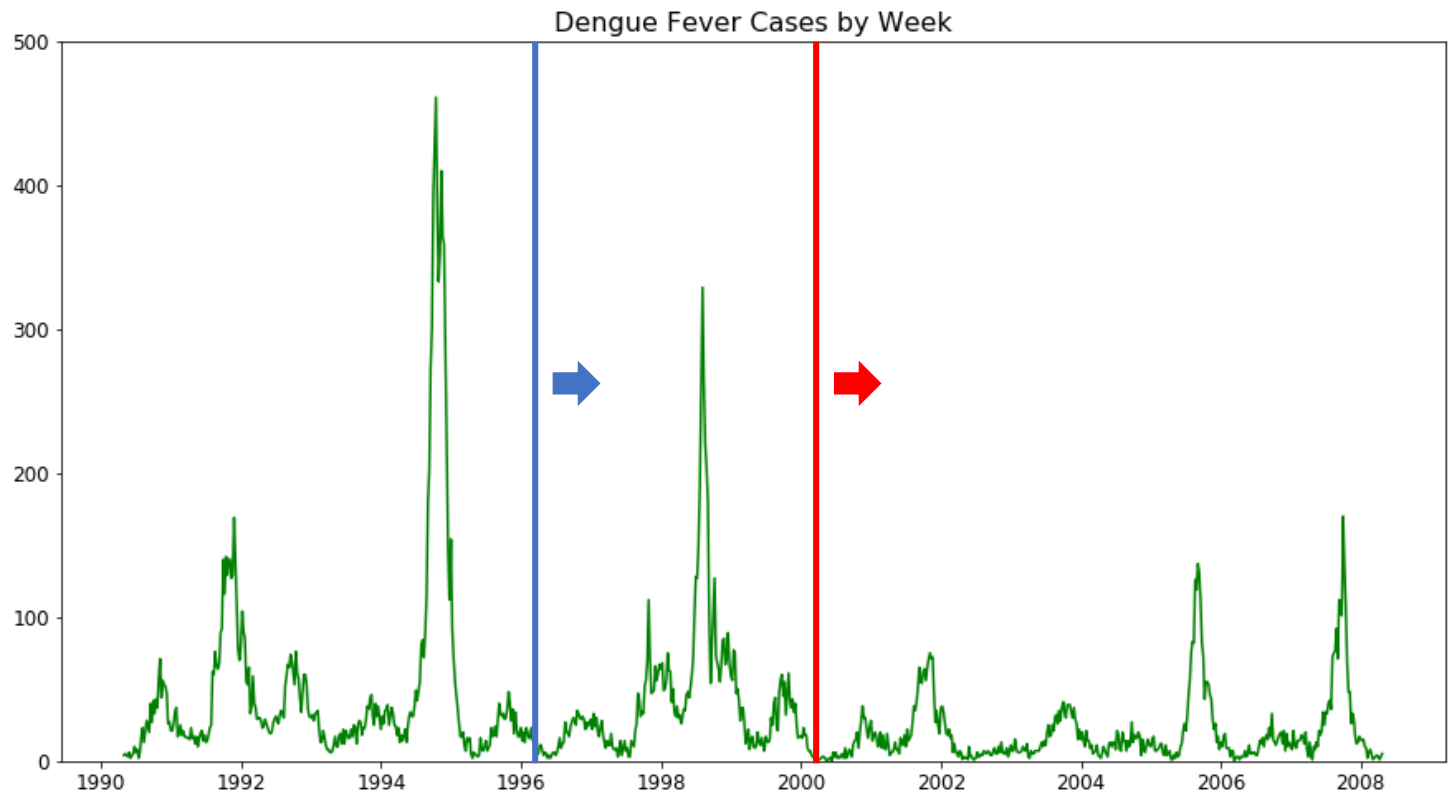
1. Long short term memory neural network **(LSTM)**
2. Supervised Learning methods with lagged features

LSTM Logic, from the ground up



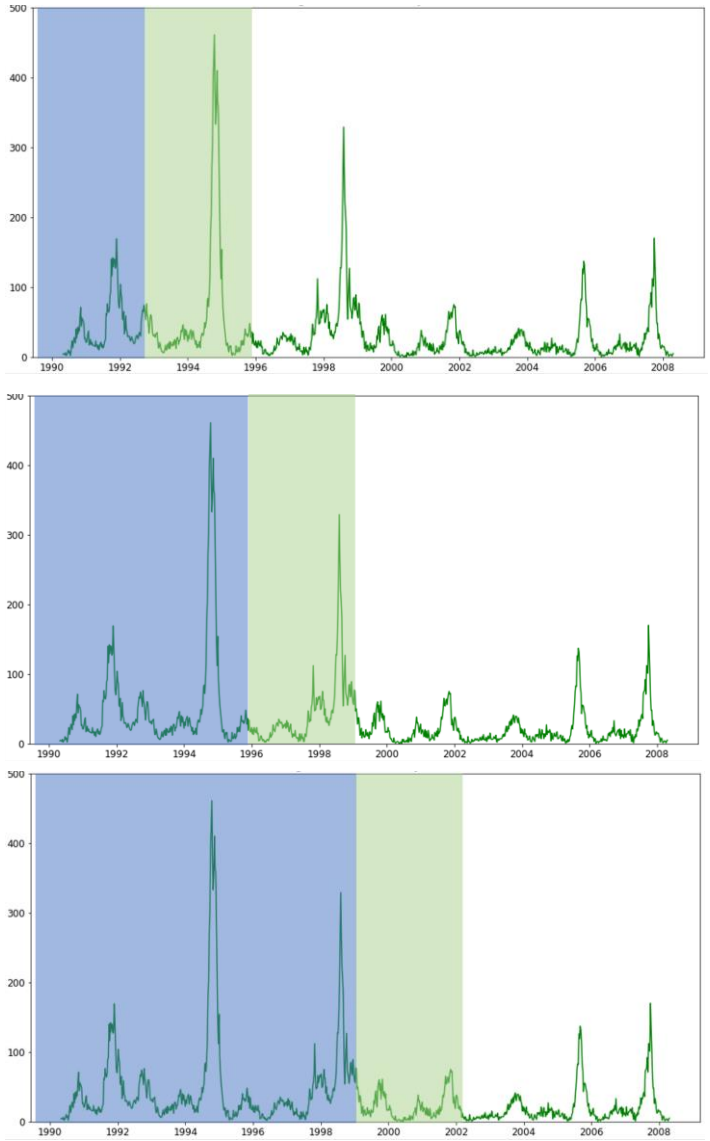
Assumes hidden and cell states are initialized with zeroes

Challenge: Choosing Validation Data (San Juan)



- No random sampling – order must be maintained for correct training
- Split to right of red line: Validation data has no big outbreak, so validation loss consistently lower than training loss
- Split to right of blue line: sparse training data, and risk of splitting up the leading indicators of the second outbreak

Walk Forward Validation



etc…

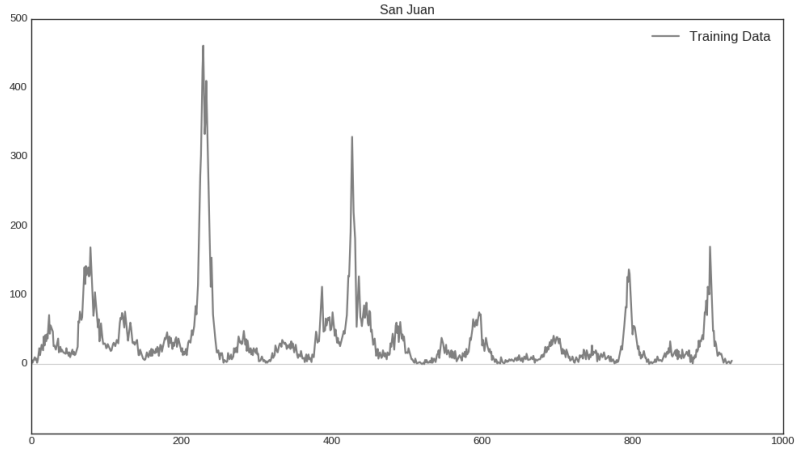
1. Pick the smallest viable training size n
2. Train with the first n samples, and predict sample(s) n+x
3. Increase n by x and repeat

By keeping the validation size constant (n+x regardless of training size), you can summarize loss across all points and know you are comparing apples to apples.

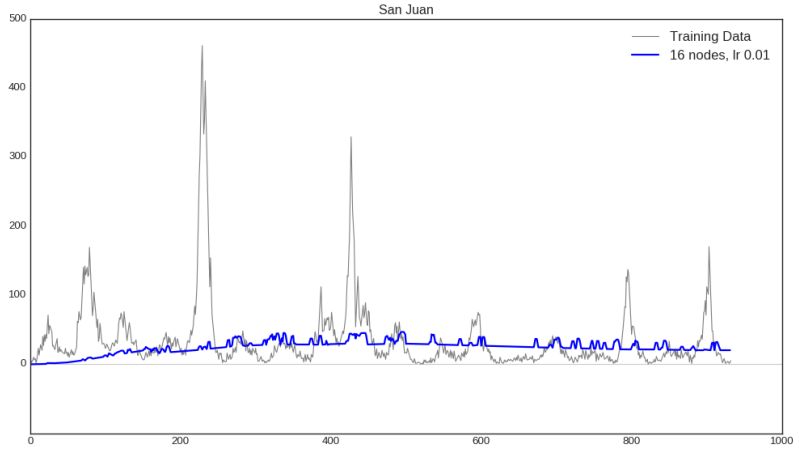
Training Data | Validation Data

LSTM Training

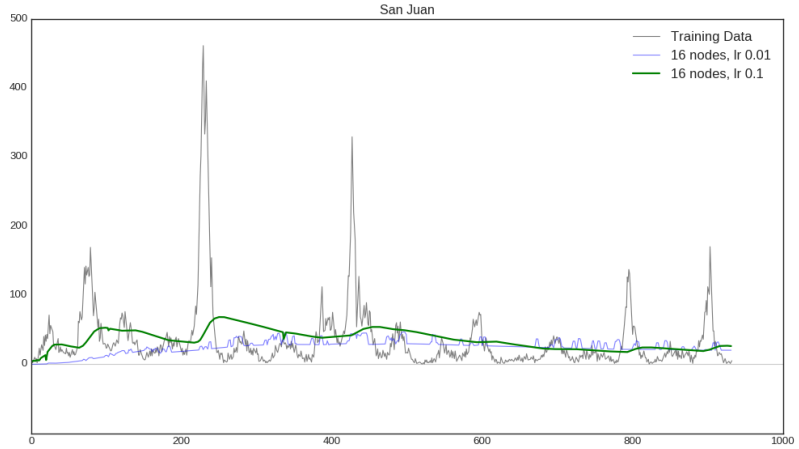
Training Data for comparison



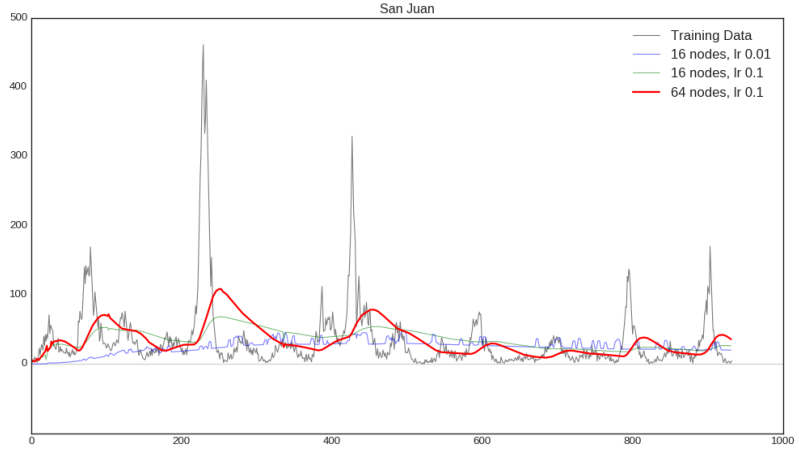
First attempt: learning slowly

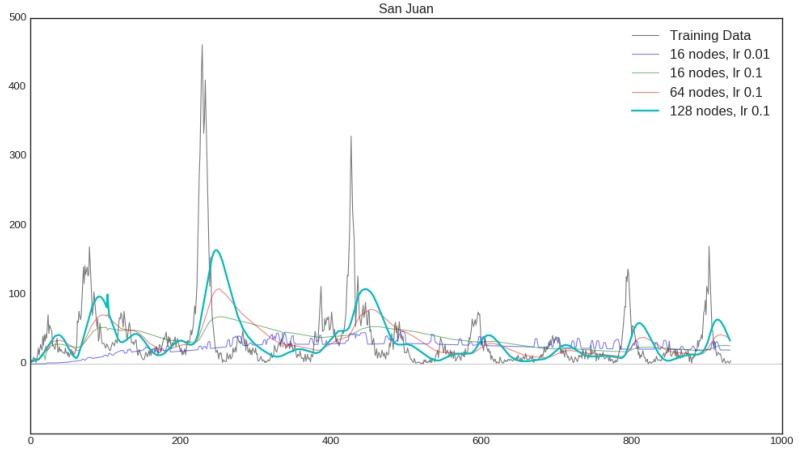


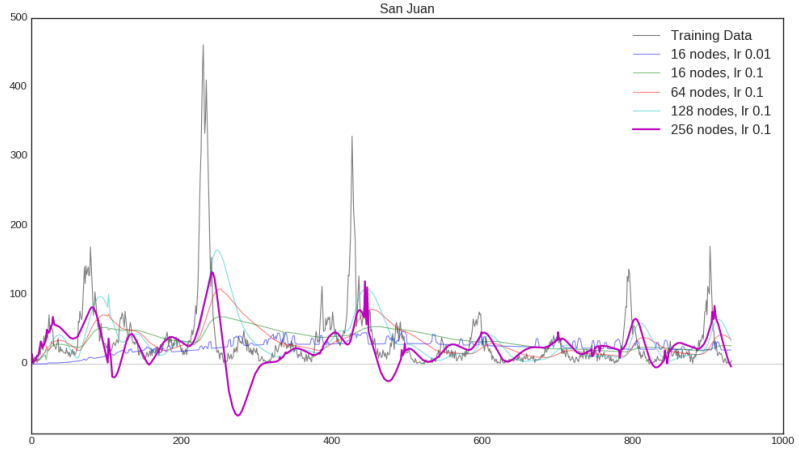
Higher learn rate helps identify increases



Larger Hidden Layers helped up to a point, but overfit past 128 nodes



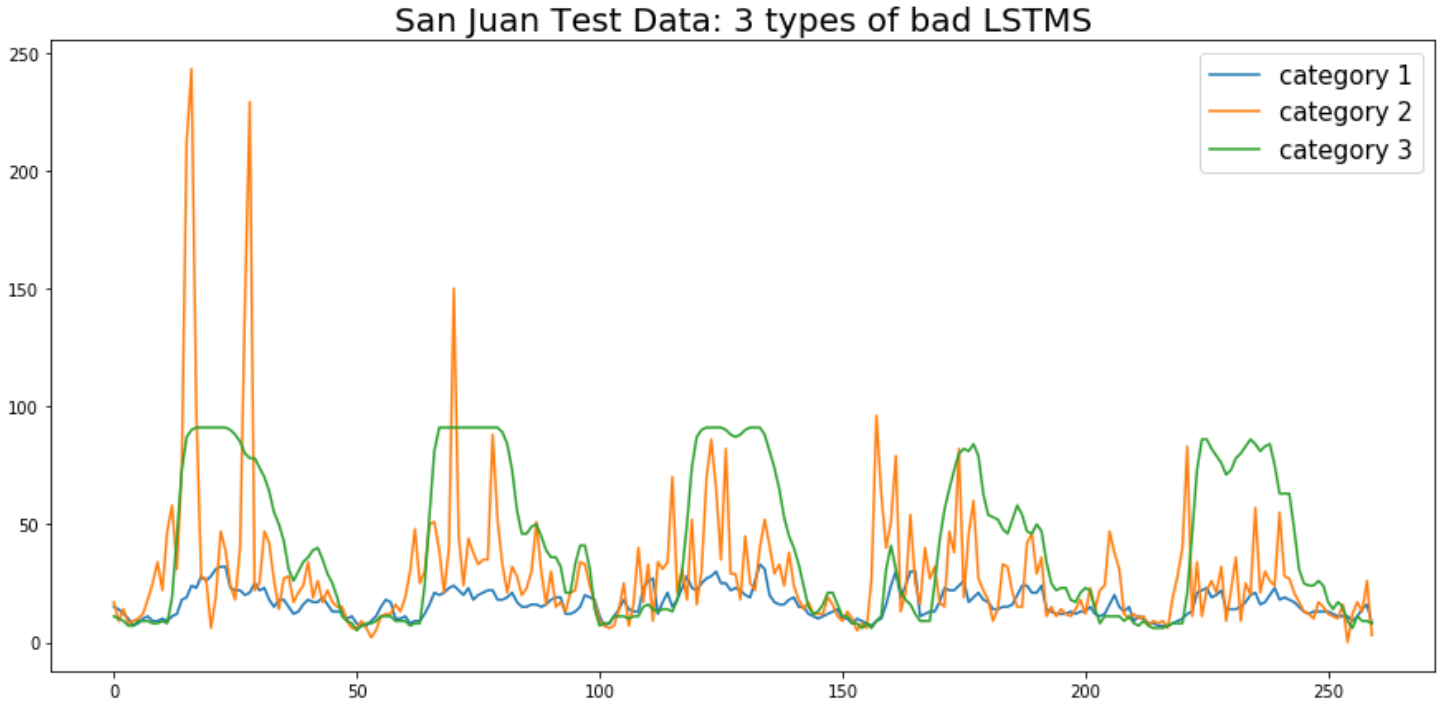




LSTM Performance

Ultimately, all the LSTM architectures I built fell into one of three categories:

1. Models that didn't predict any outbreaks (likely underfit)
2. Models that predicted nearly constant outbreaks (likely overfit)
3. Models that seemed to predict when an outbreak would happen, but continued predicting outbreak numbers for dozens of weeks after the fact. The error from these extended high predictions made these models unusable.



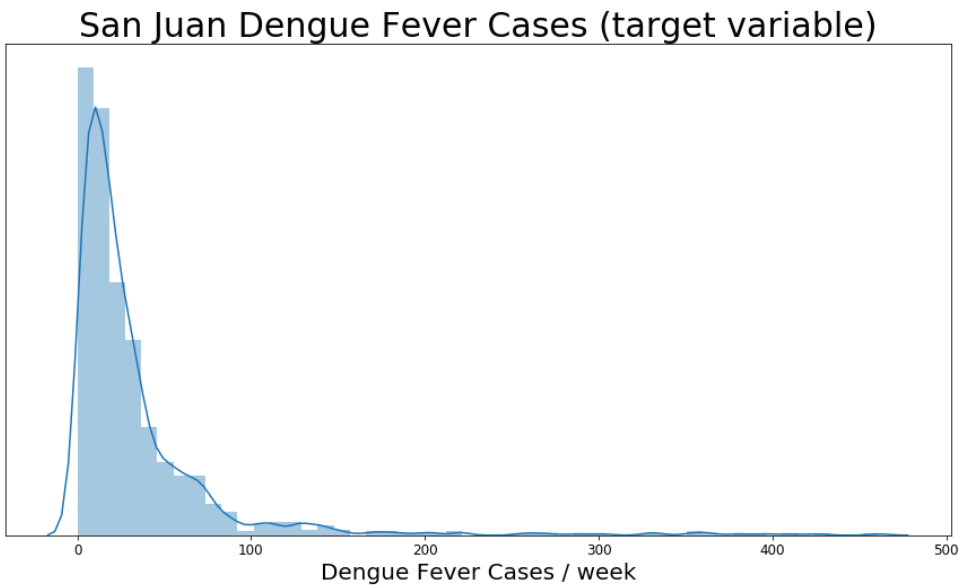
Time Series Modelling using Lagged Features

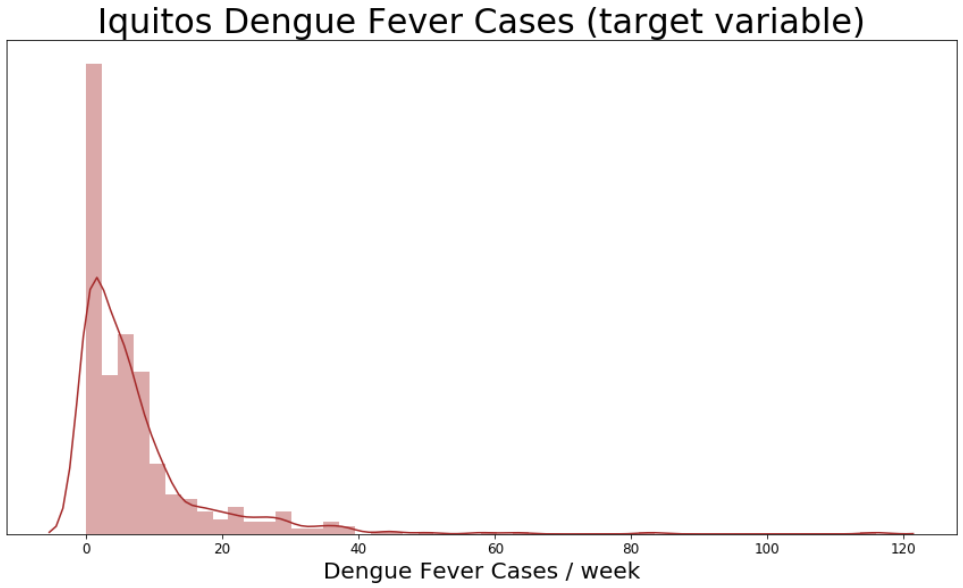
- Lag Features by n periods
- Use with models that can ignore or penalize features (Random Forest, Lasso Regression, etc.)

ID	FeatureA	FeatureB	FeatureC	FeatureA_lag1	FeatureB_lag1	FeatureC_lag1
1	A1	B1	C1	Need to impute	Need to impute	Need to impute
2	A2	B2	C2	A1	B1	C1
3	A3	B3	C3	A2	B2	C2
4	A4	B4	C4	A3	B3	C3
5	A5	B5	C5	A4	B4	C4
6	A6	B6	C6	A5	B5	C5

Random Forest Training

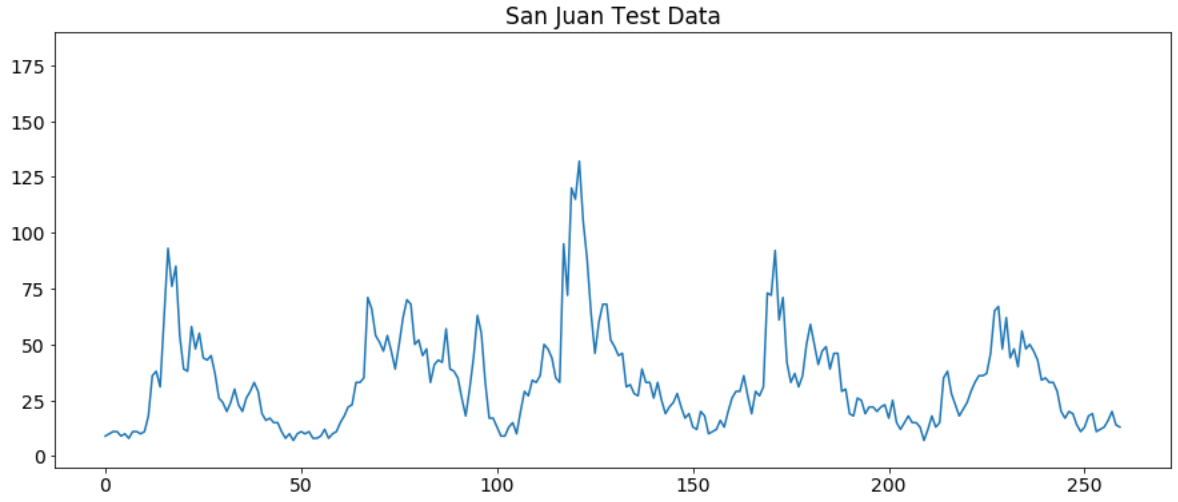
- Model Choice: Random Forest Regressor:
  - No assumptions of linearity or normality, which would be problematic given the distribution of the target variable



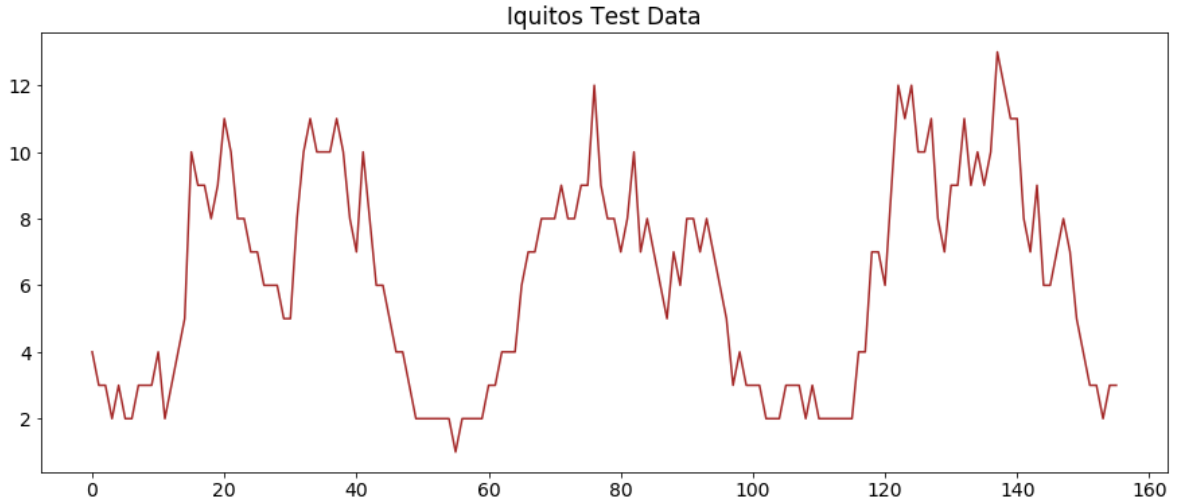


  - Can “ignore” features that do not contribute to meaningful nodes in the trees. This is helpful since feature-lagging creates so many extra features

Random Forest Training



<u>San Juan</u> Random Forest Regressor via cross validated grid search	Maximum tree depth: 35 Maximum features considered per split: 5 Minimum samples to create a leaf: 3 Minimum samples to split a node: 2 Trees: 100

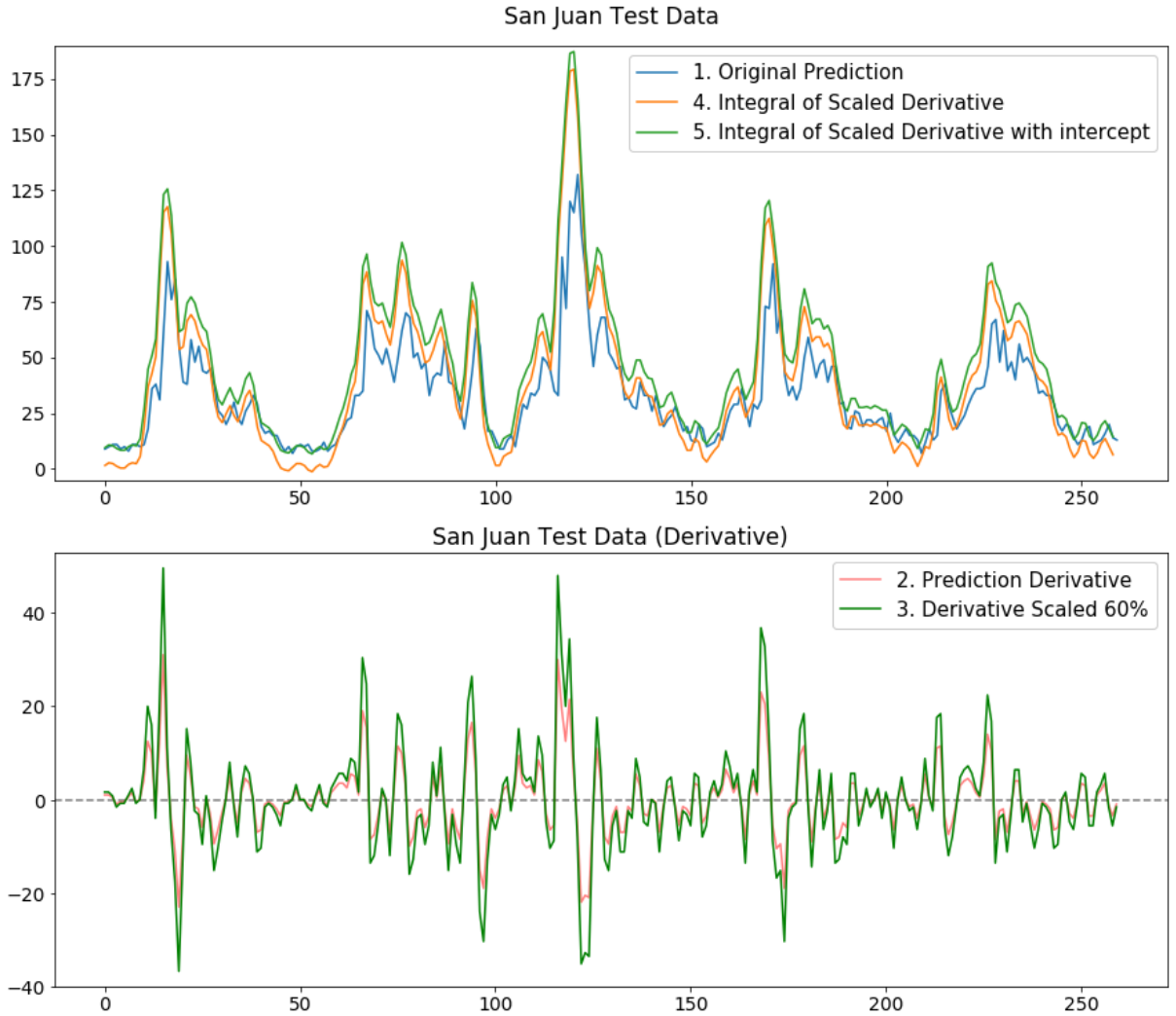


<u>Iquitos</u> Random Forest Regressor via cross validated grid search	Maximum tree depth: 10 Maximum features considered per split: 2 Minimum samples to create a leaf: 2 Minimum samples to split a node: 2 Trees: 300

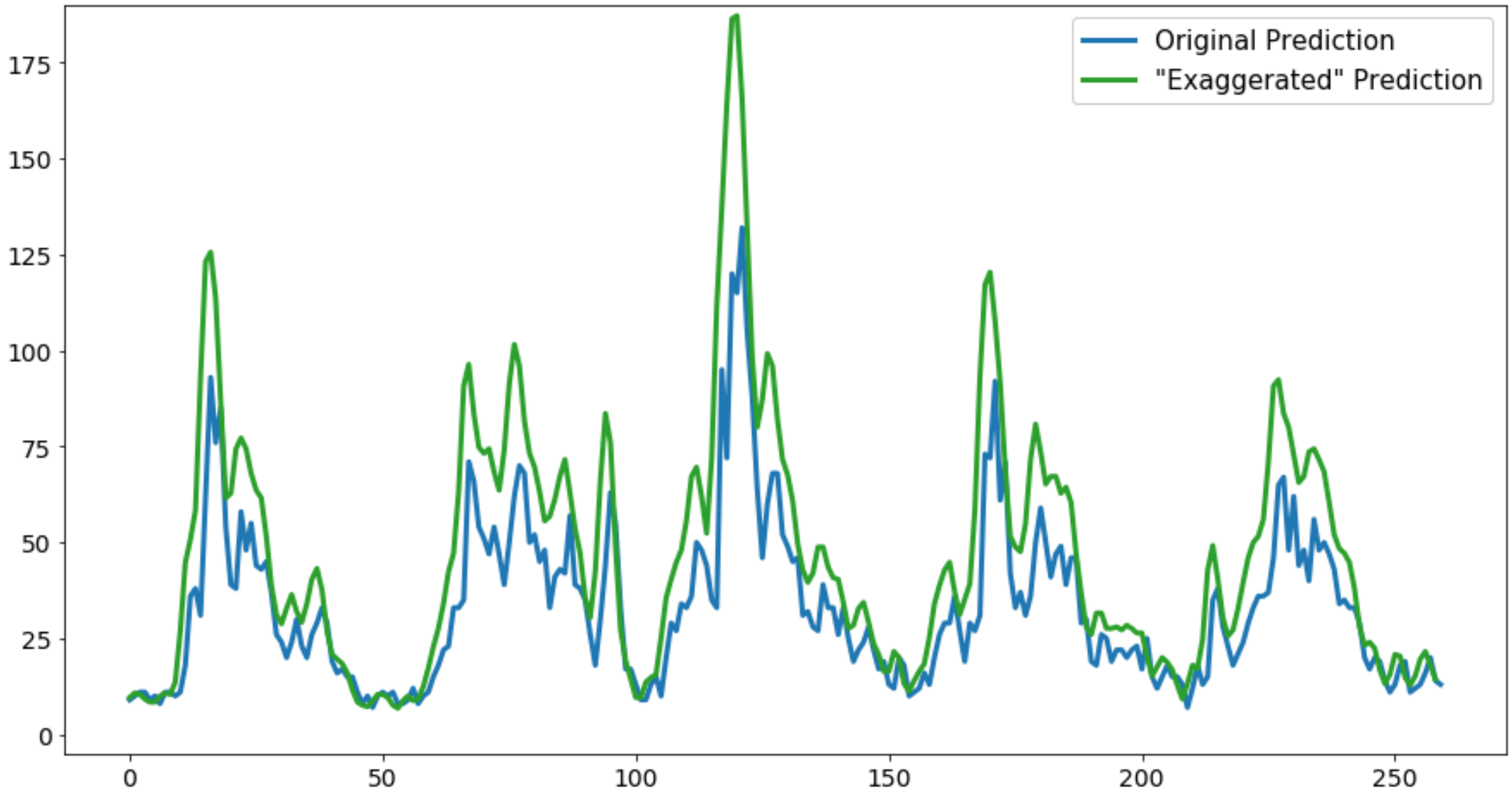
Peak Boosting

Reminder: the value of a disease outbreak predictor would be finding the outliers. Most models were unlikely to predict extreme values, so the following was done to exaggerate large predictions:

1. Predict test data with model
2. Take derivative of the predictions
3. Scale derivative values
4. Take Integral of scaled derivative
5. Add “intercept” (the first predicted value)



Peak Boosting



Performance

- San Juan Best Model:
  - Random Forest Regressor with 100 trees and maximum depth of 35
  - Prediction derivative scaled up by 60% to exaggerate predicted outbreaks
- Iquitos Best Model:
  - Random Forest Regressor with 300 trees and a maximum depth of 10
  - Prediction derivative scaled up by 10% to exaggerate predicted outbreaks

Combined Mean Absolute Error of 24.0, good for 440th / 5910 competitors (93rd percentile)

DRIVENDATA

Resources

LSTMs:

- Download The LSTM Reference Card and code an LSTM forward pass with just NumPy!
- Free Udacity course on deep learning (including RNN's) in pytorch
- Helpful article on LSTM's and GRU's
- Interesting podcast (TWiML&AI) suggesting you can trim an LSTM down to ONLY a forget gate
- Walk Forward Validation guide by Dr. Brownlee (Machine Learning Mastery)

Dengue Fever and *Aedes Aegypti* mosquitoes:

- Juliano, Steven A et al. "Desiccation and thermal tolerance of eggs and the coexistence of competing mosquitoes."
  Oecologia vol. 130,3 (2002): 458-469. doi:10.1007/s004420100811
- Yukiko Higa, Nguyen Thi Yen, Hitoshi Kawada, Tran Hai Son, Nguyen Thuy Hoa, Masahiro Takagi
  Journal of the American Mosquito Control Association (1 March 2010)

More about this project:

- Full writeup: gregcondit.com/articles/dengue-fever
- Github: /conditg/deng-ai