Mike Sutherland
Write up:

## Data Preprocessing

I will describe the steps taken to preprocess the data. Since this is complete it will basically entail describing what is in the existing code.

- Acquire data and data source; data is pulled from URL every time the script is run so up-to-date models are trained daily.
- Convert to datetime
- Create “days since start” var which is the days since the first day of the dataset
- Convert days since start to double because matlab gpr does not work on timeseries data
- Describe which data was pruned
    - Low quality
    - States without data
    - New Jersey because there are outliers

## Model

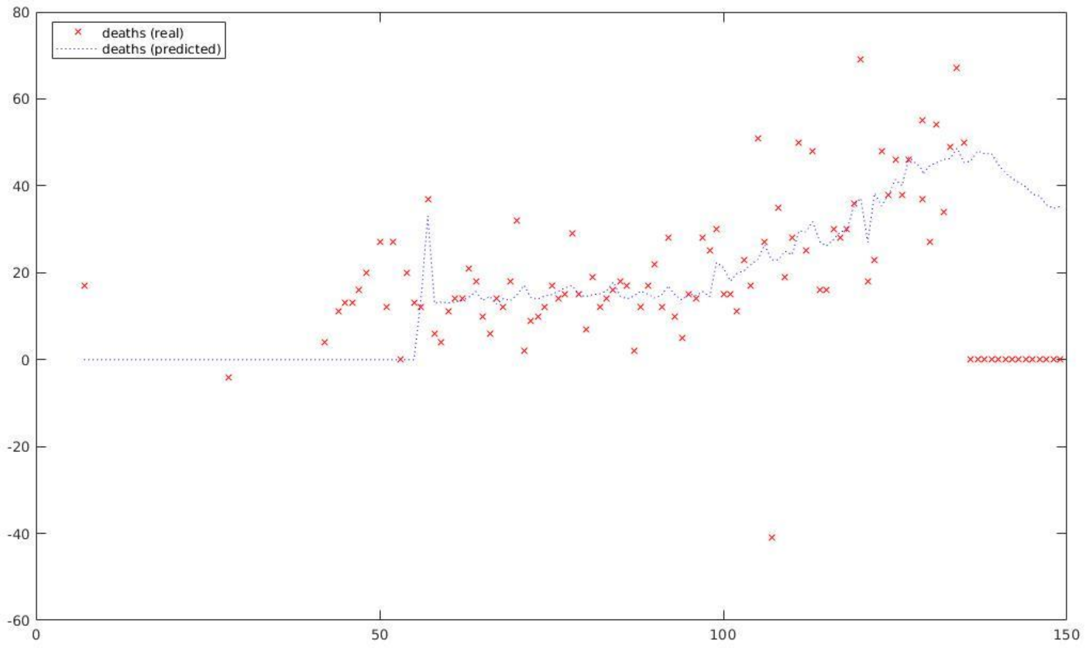
I will describe the model used, how it was trained, how kernel parameters were implemented. Maybe a brief overview of how GPR works, and:

- How kernel param was found using bayesian optimization
- How model was trained with pairs of inputs and outputs
    - Inputs: increase in hospitalizations, ICU population, ventilator population, death increase that day, new positive cases that day.
    - Inputs are taken for a sample day *k* from *k-14* days ago in the dataset
    - Outputs: deaths 14 days later.
    - The model is trained on input output pairs for day k in the dataset for a total of n days
    - Each model is for one state, because e.g. someone is not getting sick in New York and then dying in Colorado 14 days later
- Assumptions made:
    - We can predict 14 days into the future (this param is arbitrary)
    - The processes mapping future deaths to icu stays, patients on vent, etc. are gaussian distributions with independent timescales
    - More about assumptions (I need to think deeply about what assumptions were made)

# Results

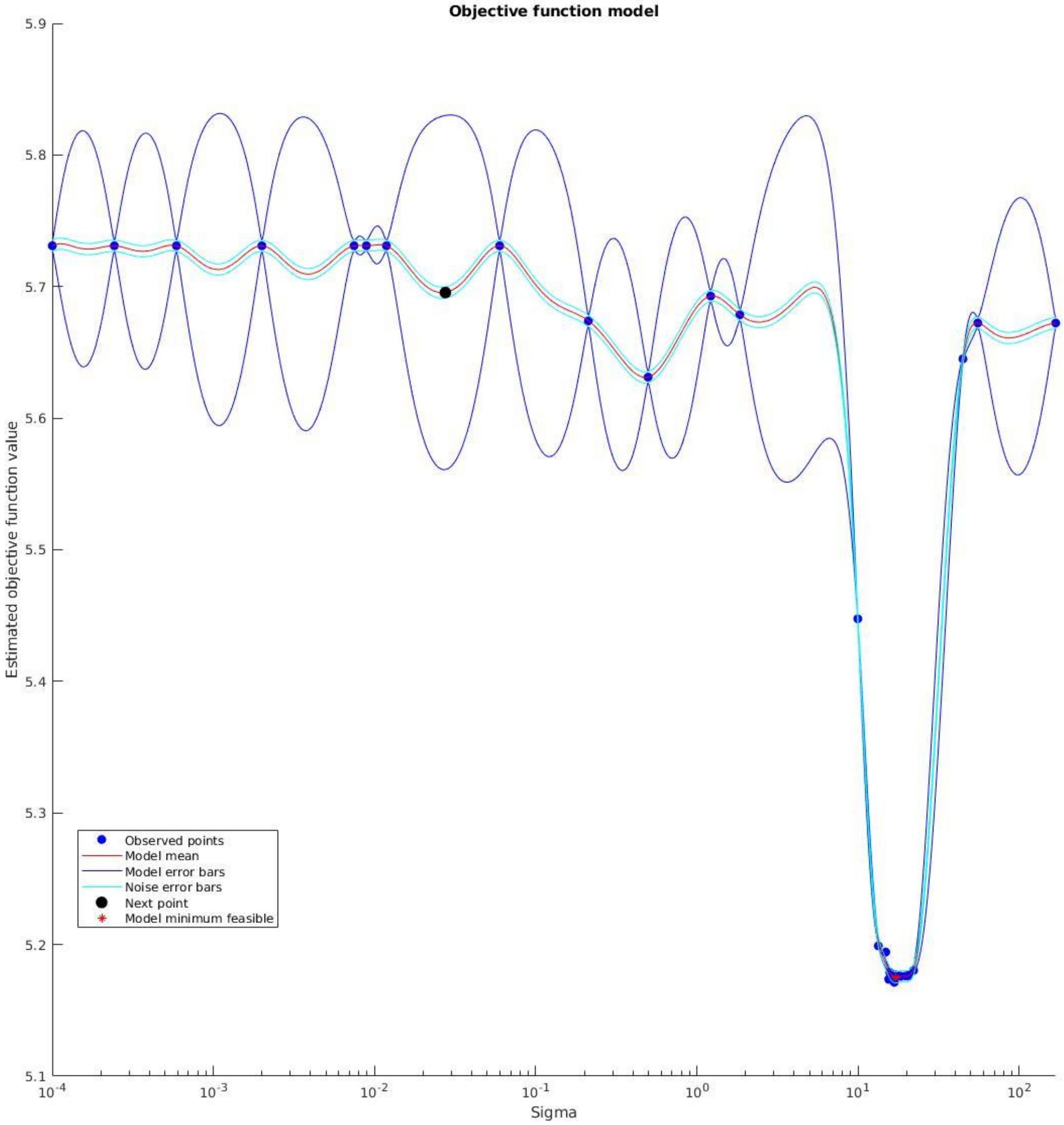
Charts so far:

Result: deaths pretty closely match predictions



x axis is days since start, y axis is no. of deaths.

Result: acquisition function value for finding the kernel vs. number of iterations of optimization



x axis is sigma (essentially, timescale of causation observed in the data), y axis is acquisition function value (essentially, error in the fit, so lower is better)

We can see that the timescale is ~10^1 days which seems to confirm our gut check that covid deaths follow new positive cases/icu stays/etc on the order of 10^1 days.

## TODO:

- Prettify charts

- Add mean/std/95% ci curves
- Proper labels to all charts
- Detailed interpretation of model results and potential flaws to this model (there are many)
- More figures; plot the prediction variables as well