Corona Trend (EDA + Prediction) MVP

Using deep-learning methods (LSTM)

Summary

Details

Executive Summary

To better understand disease progression, we've developed a robust (and extensible) MVP forecasting engine using Deep Learning Method (LSTM) …

Objective

Build forecasts for COVID19 disease progression using Deep Learning Method (LSTM)

Complication

1. 3253 trends in 58 states (one trend per town)
2. 74 days of data (2020/01/22 – 2020/04/04)
3. Uni-variate (temporal) data for forecasting

Approach

Current Deliverable — Next Phase →

Understand data	Build MVP using 1 algorithm	Enhance Forecasting Process
1. Create Modeling Dataset 2. Distributions 3. Temporal trends 4. Build Dash framework to visualize data	1. LSTM on 1 trend	1. **Data enrichment** (exogenous data) 2. **Algorithm enrichment**

… with the following insights.

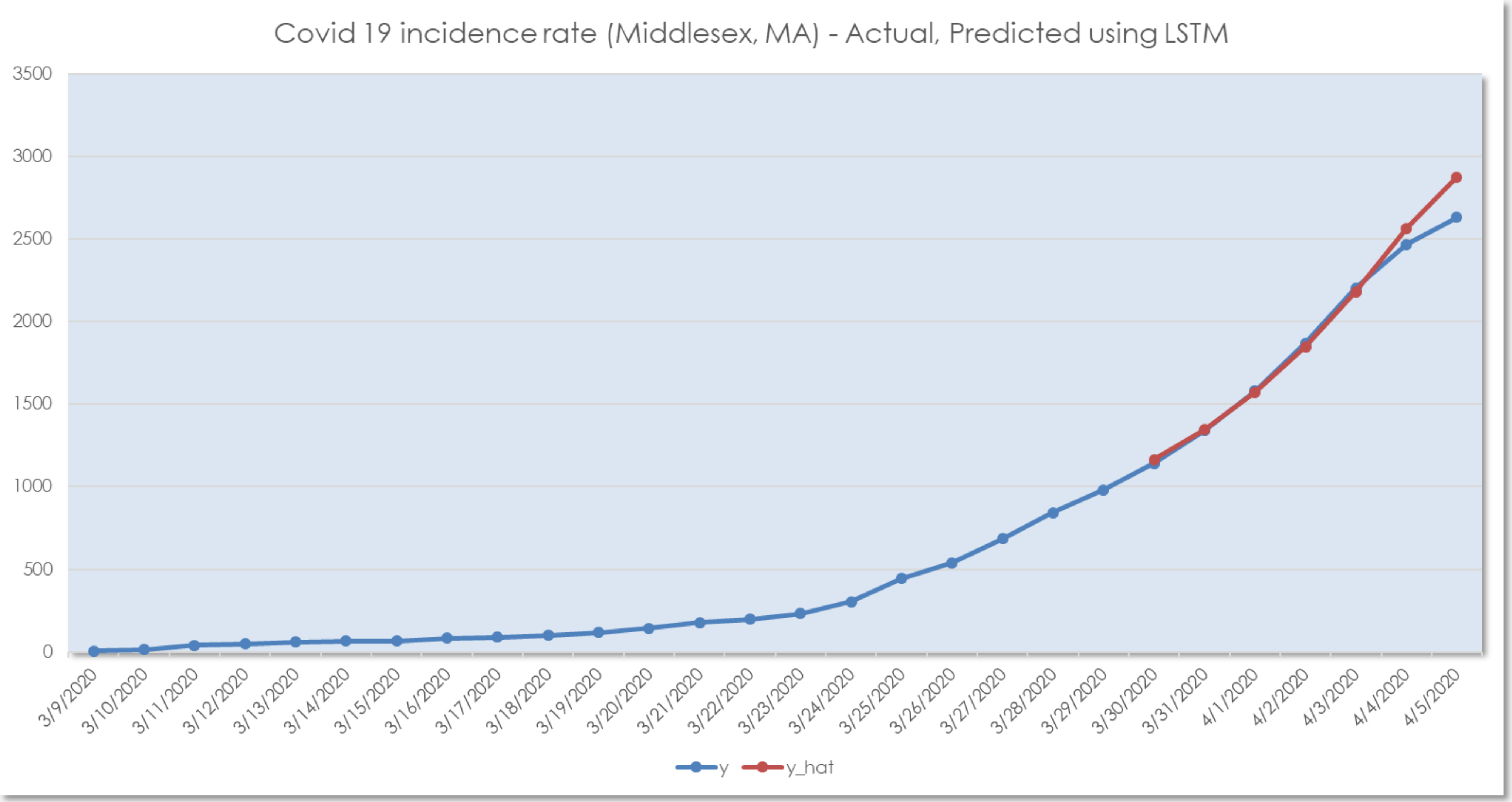
1) *Accuracy using LSTM on trend for Middlesex county using 7 hold-out days = **86%***

2) *The lower accuracies are in the recent days, where the trend seems to be getting better*

(1) ***Accuracy =*** *1 – abs(RMSE / Actual Volume)*

RMSE *= Root Mean Square Error, a measure of how good the forecasts are in out-of-time hold-out days (1 week)*

Actual trend vs. predicted using LSTM

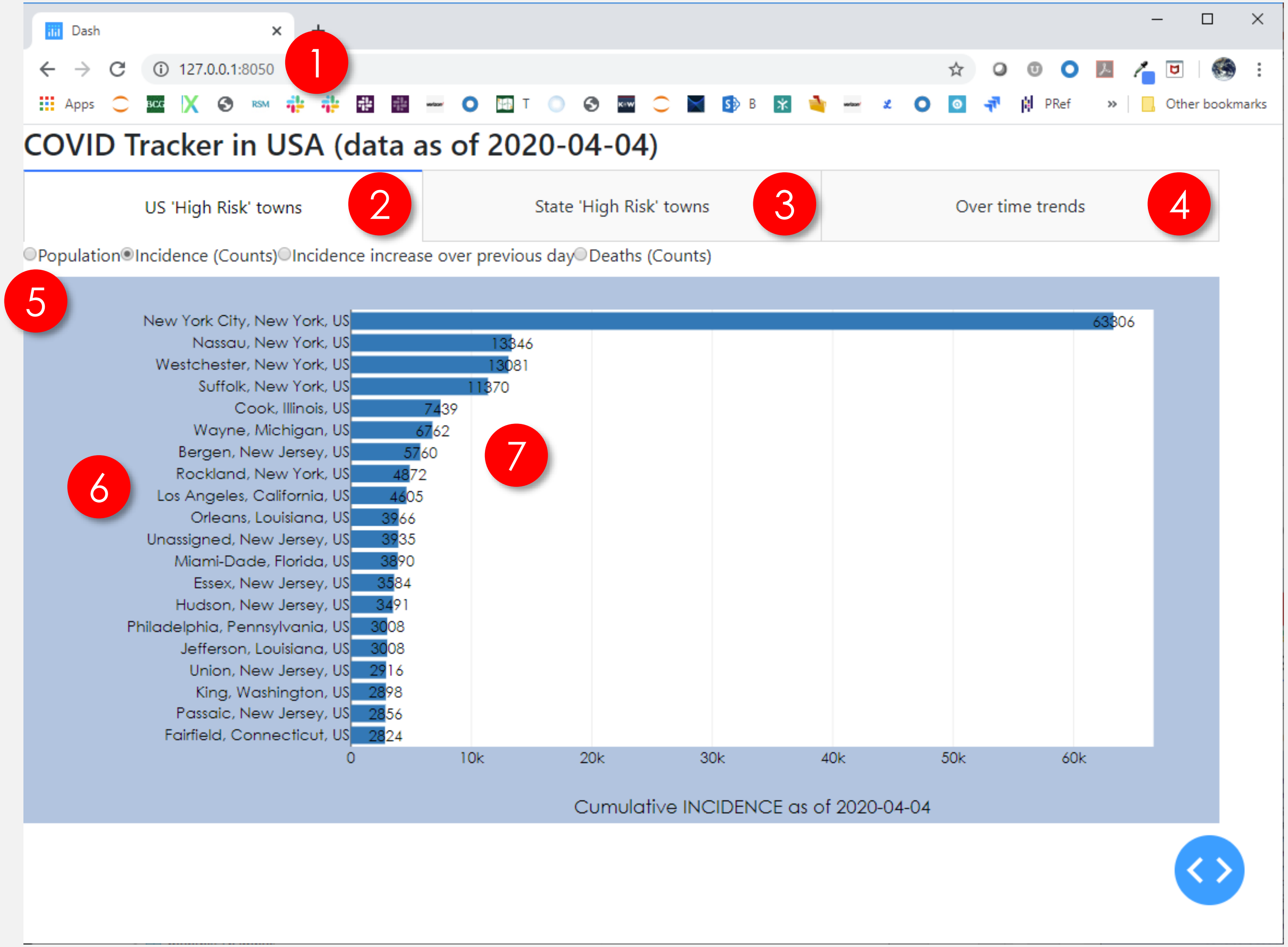


Summary

Details

Understand Data

Dashboard layout – top 20 high risk towns



1) When this dashboard is launched from the command line, one can view the results in a web-browser

2) US-level view (only looking at the high risk towns) looking at the grain of a town

3) State-level view (only looking at the high risk towns) looking at the grain of a town

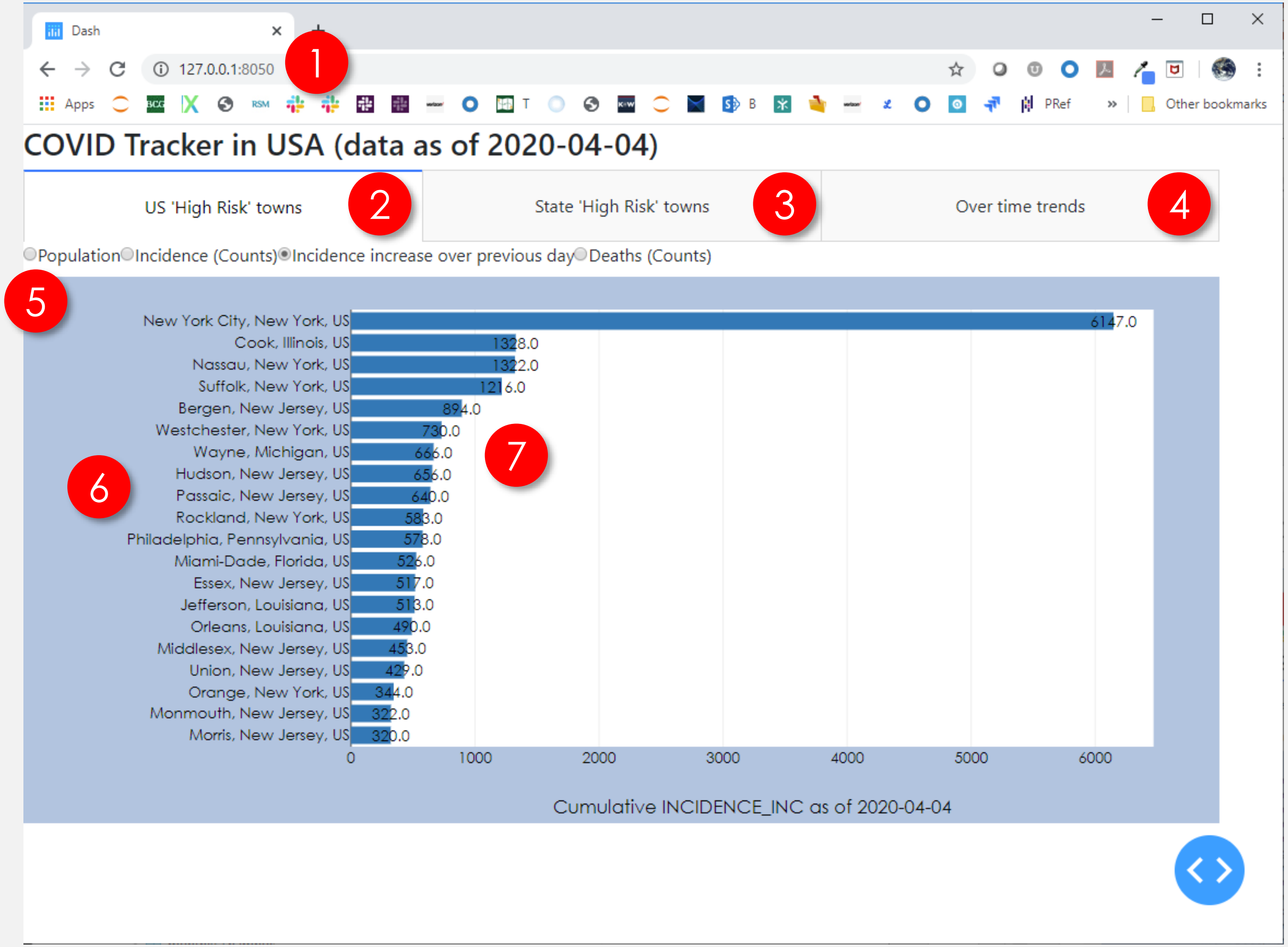
4) Town-level view to understand detailed trends

5) Various measures (population size, incidence, increase in incidence over the last day, deaths)

6) The towns ranked ordered by descending risk

7) A distribution of measure

US View – increase in incidence from previous day



1) *When this dashboard is launched from the command line, one can view the results in a web-browser*

2) *US-level view (only looking at the high risk towns) looking at the grain of a town*

3) *State-level view (only looking at the high risk towns) looking at the grain of a town*

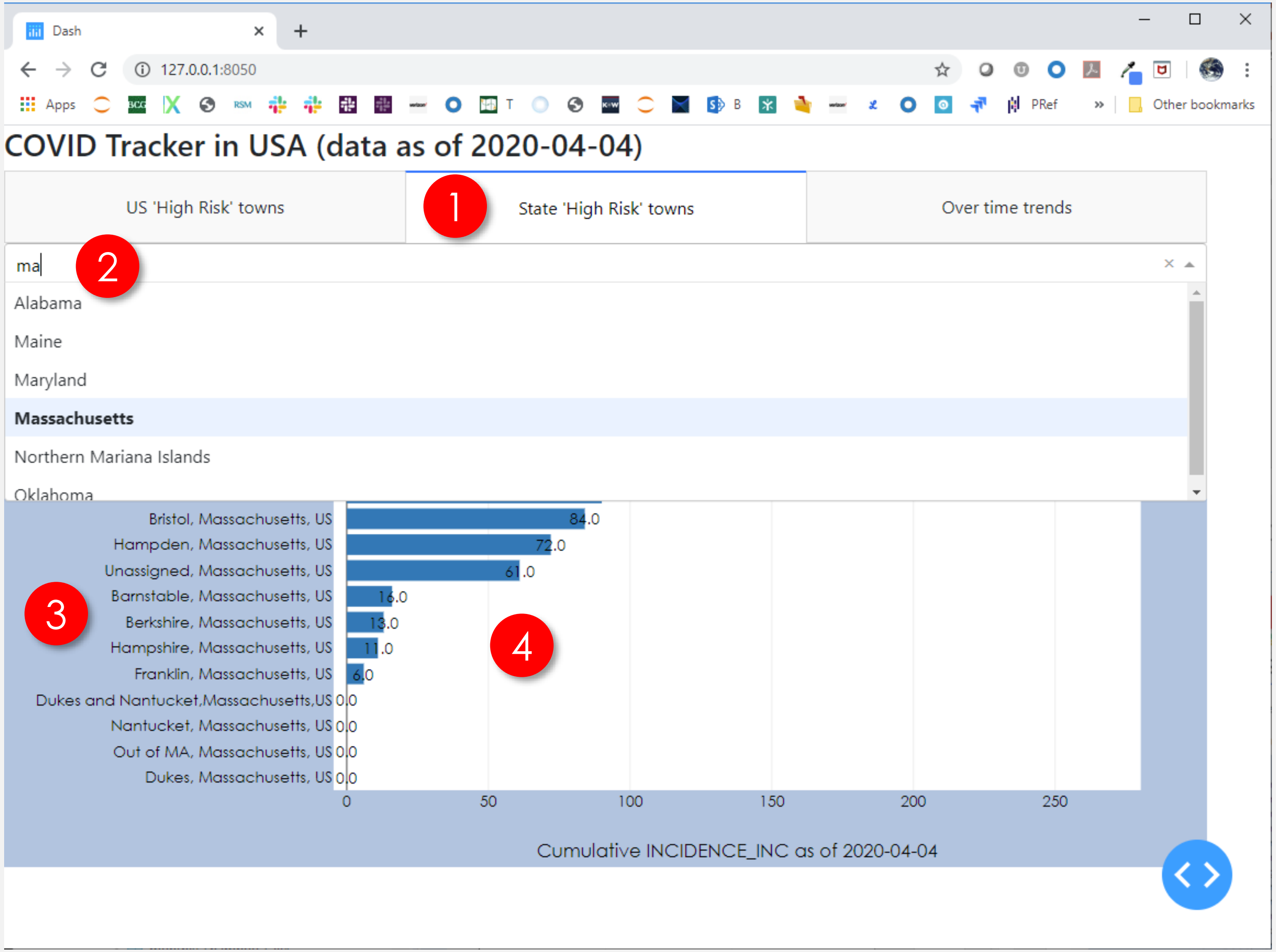
4) *Town-level view to understand detailed trends*

5) *Various measures (population size, incidence, increase in incidence over the last day, deaths)*

6) *The towns ranked ordered by descending risk*

7) *A distribution of measure*

State View (grain = town) – increase in incidence from previous day



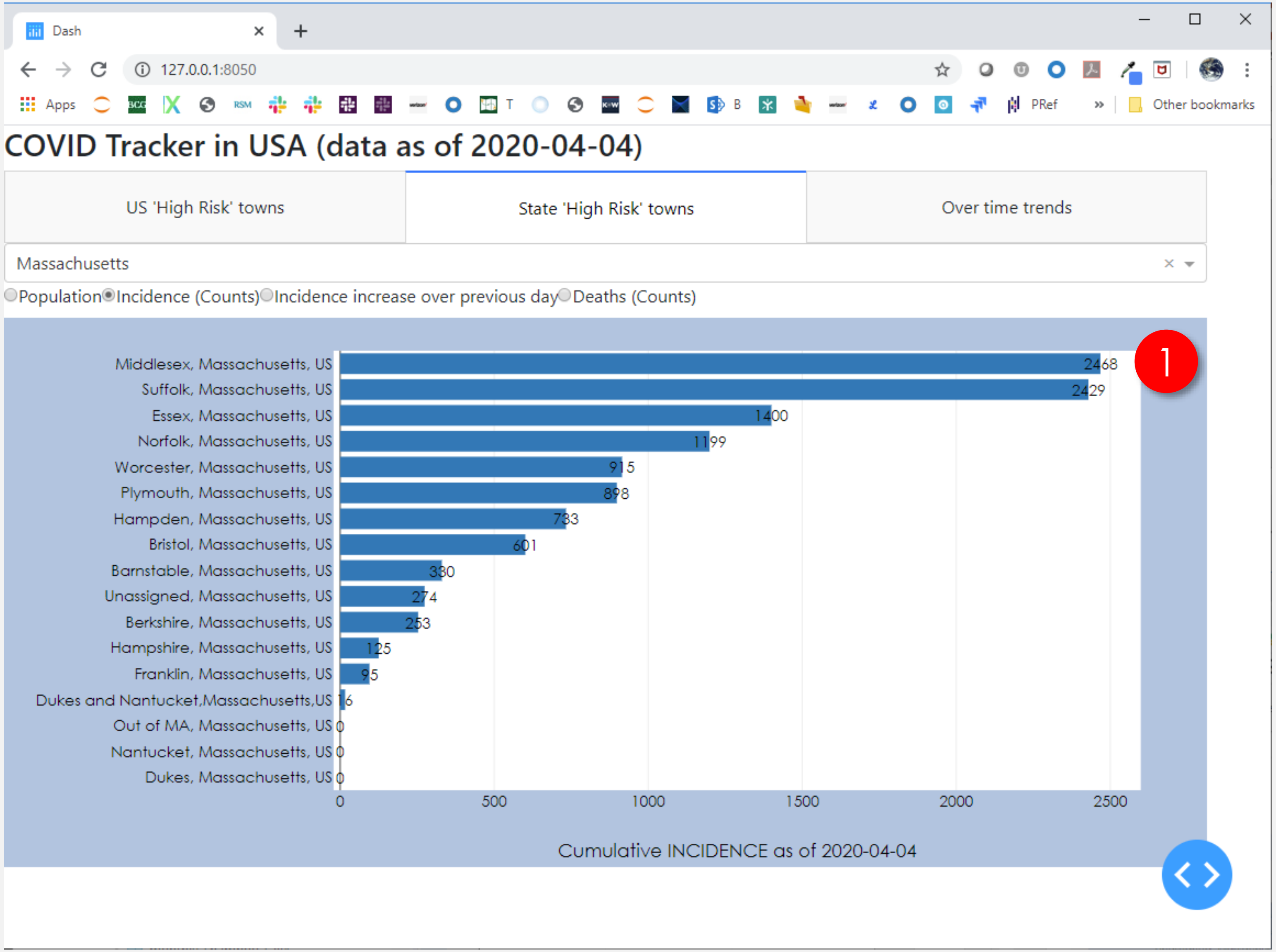
1) *State high risk town tab*

2) *Contextual filtering*

3) *Towns rank ordered by metric*

4) *The metric value*

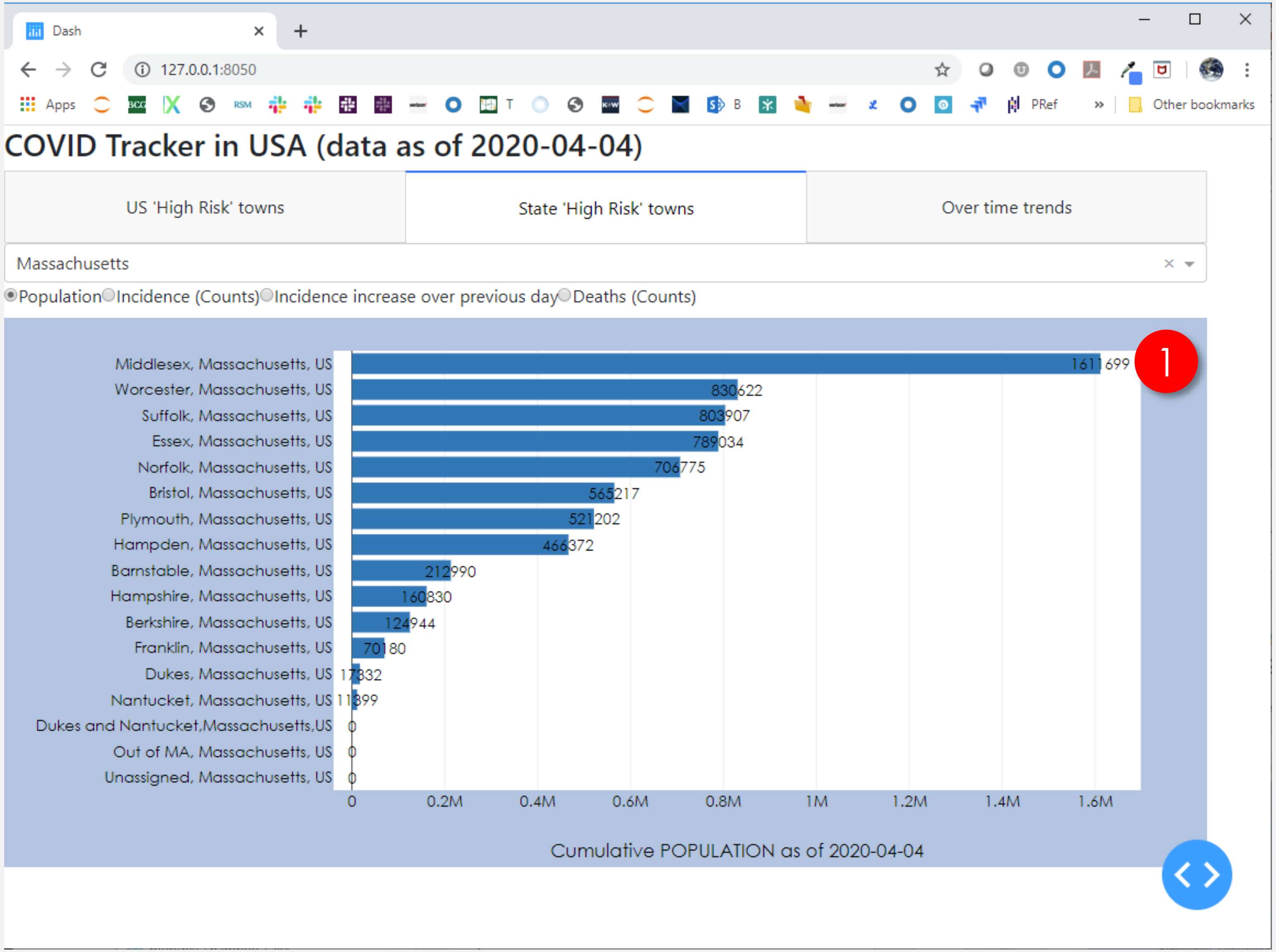
State View – cumulative incidence till 04/04/2020



1) *Middlesex the largest county by population has the highest number of cases*

2) *Suffolk county is very densely populated and has a lot of cases despite its lower population counts*

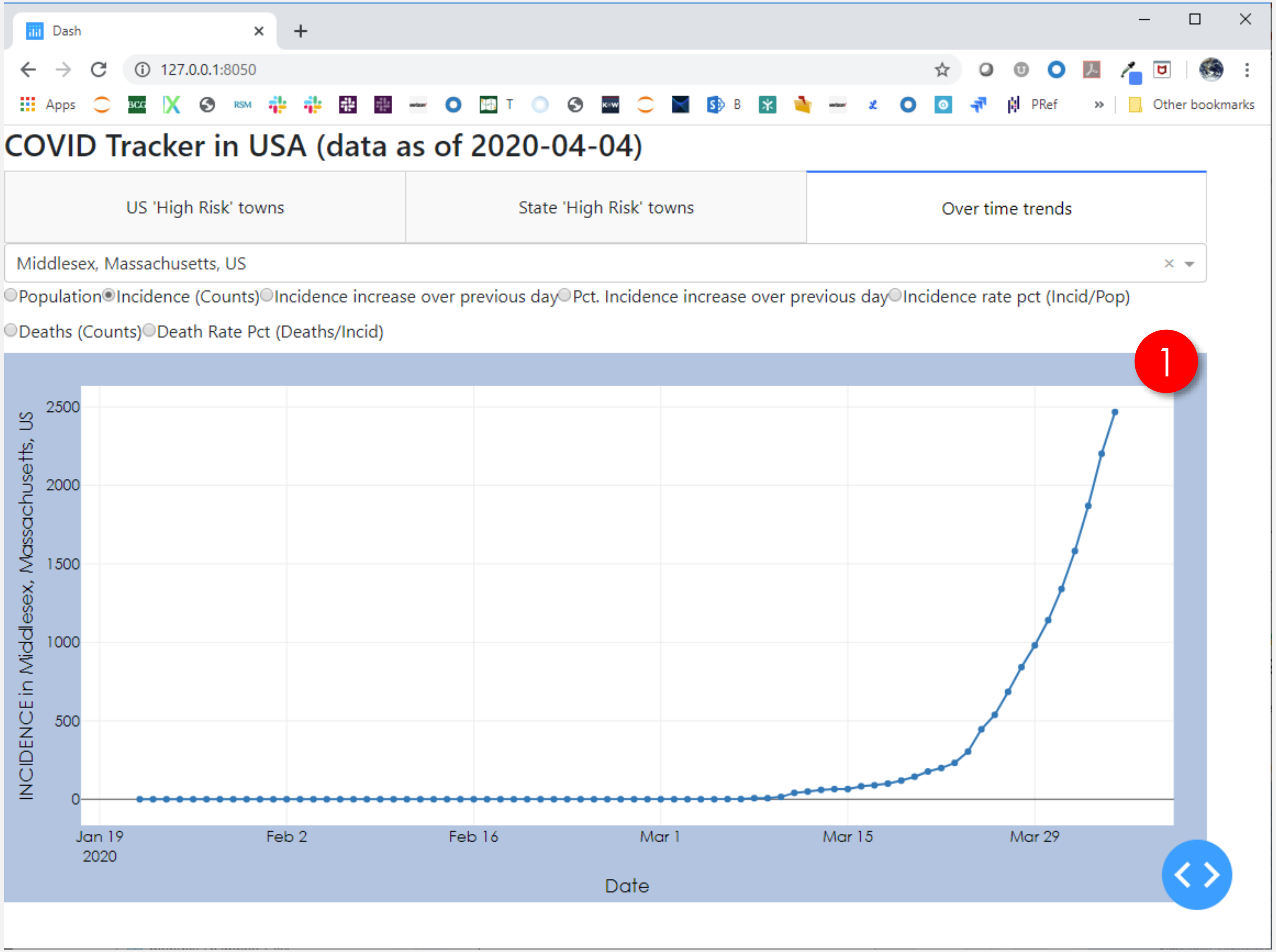
State View – population size in 04/04/2020



1) *Middlesex the largest county by population*

2) *Worcester county has nearly ½ the size of population as in Worcester*

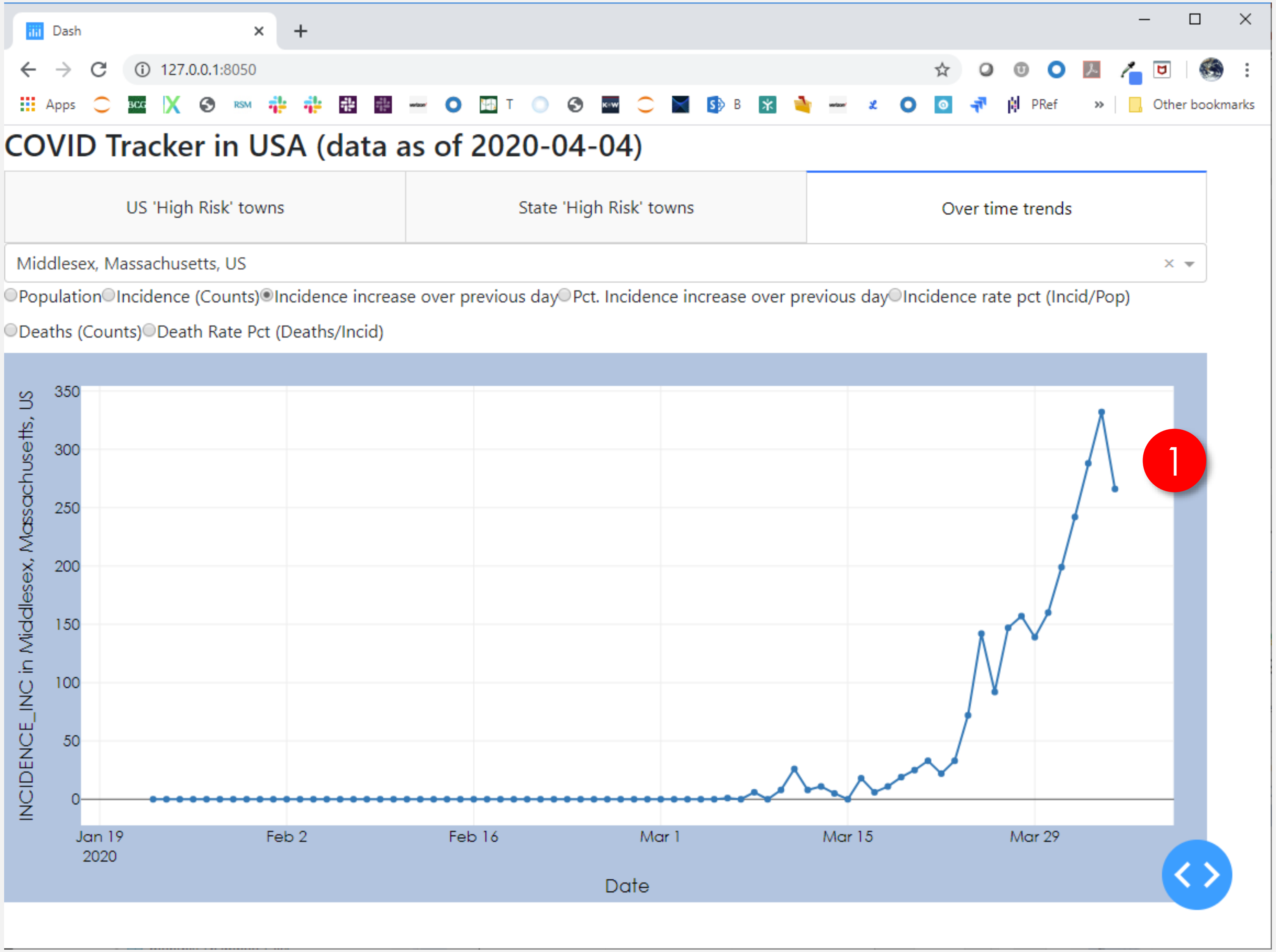
Town trend – incidence rate



1) *Middlesex the largest county by population has the highest number of cases*

2) *As of 04/04/2020, Middlesex county still is in its growth phase. Growth does not seem to be abating*

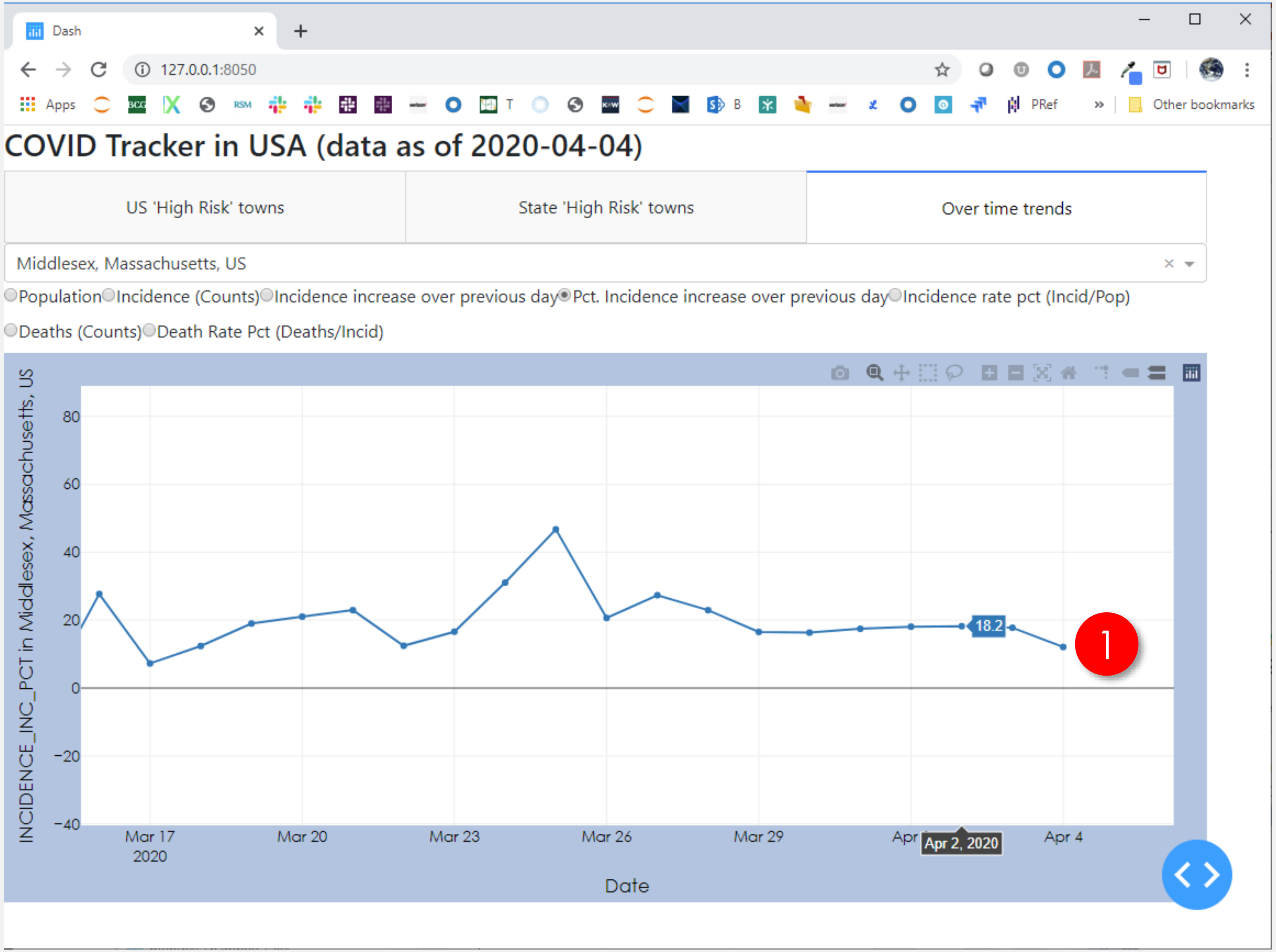
Town trend – incidence growth



1) *Day over day growth is still high confirming the previous hypothesis that Middlesex county is still in its growth phase.*

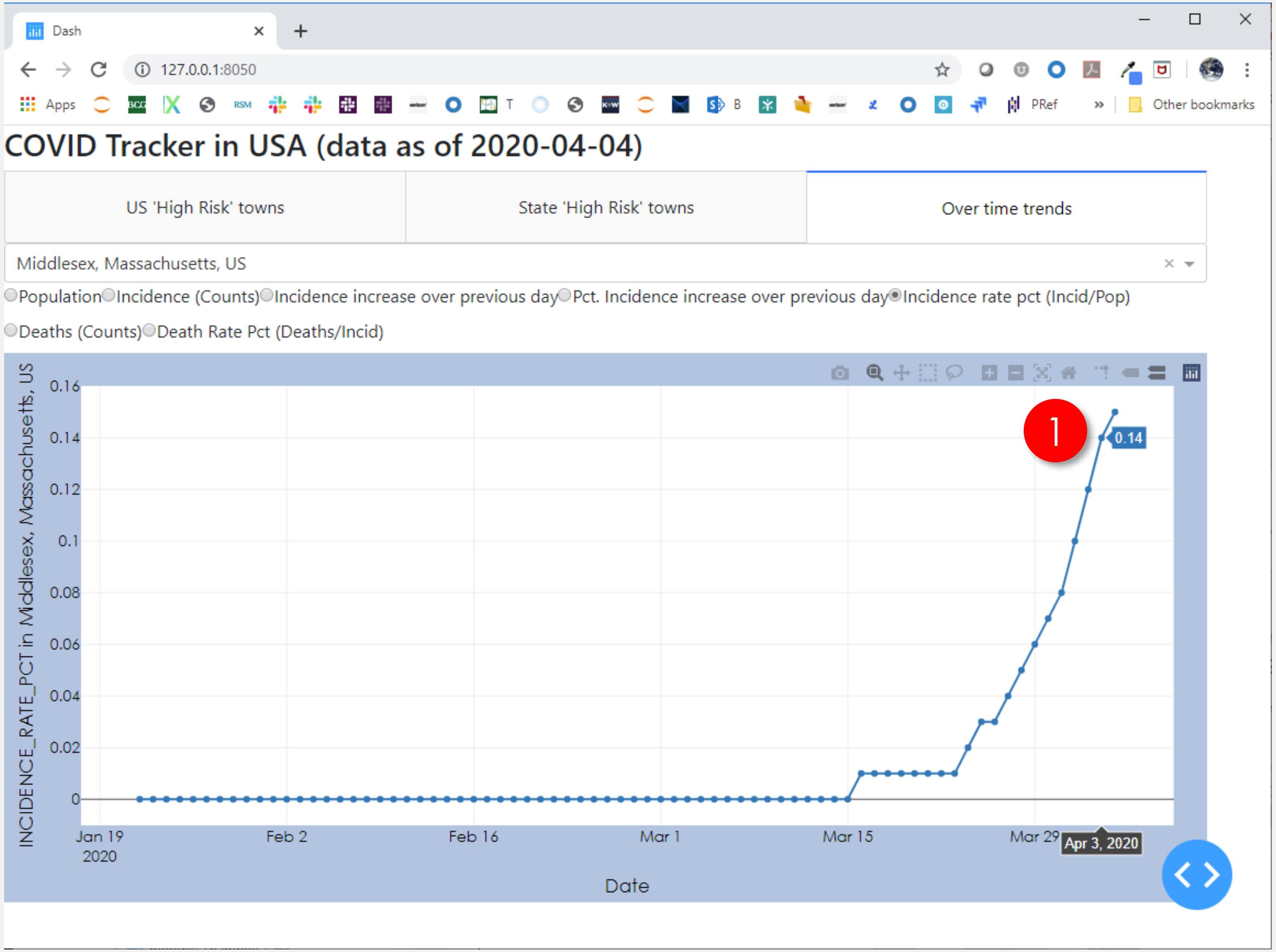
2) *Social distancing measures need to continue*

Town trend – incidence growth



1) *New incidences are growing around 18% day over day*

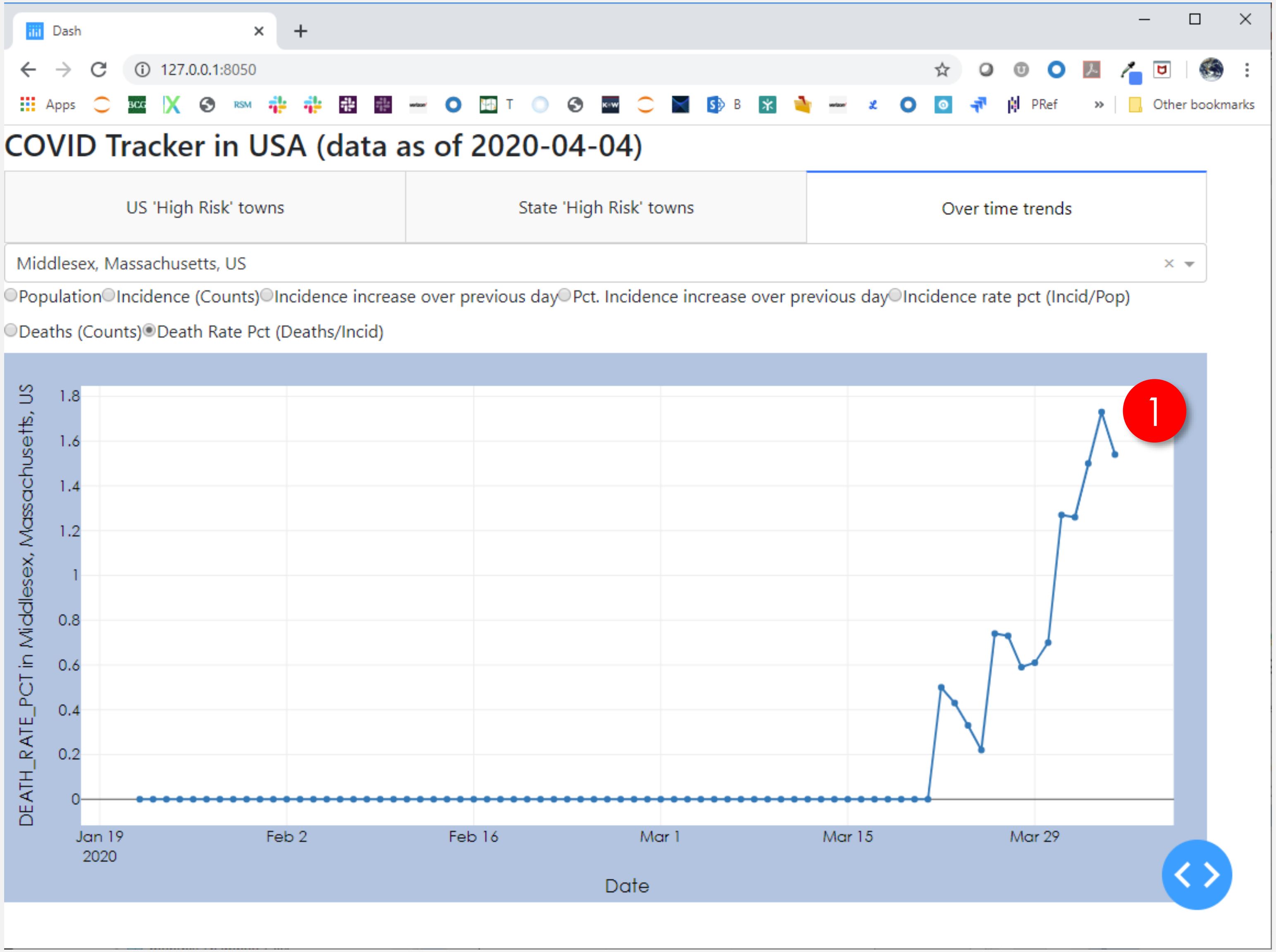
Town trend – incidence growth



1) *Only about 0.14% of population has tested positive. The numbers are relatively small*

Should lock-down measures be relaxed sooner rather than later, the new case incidence can significantly spike up

Town trend – incidence growth

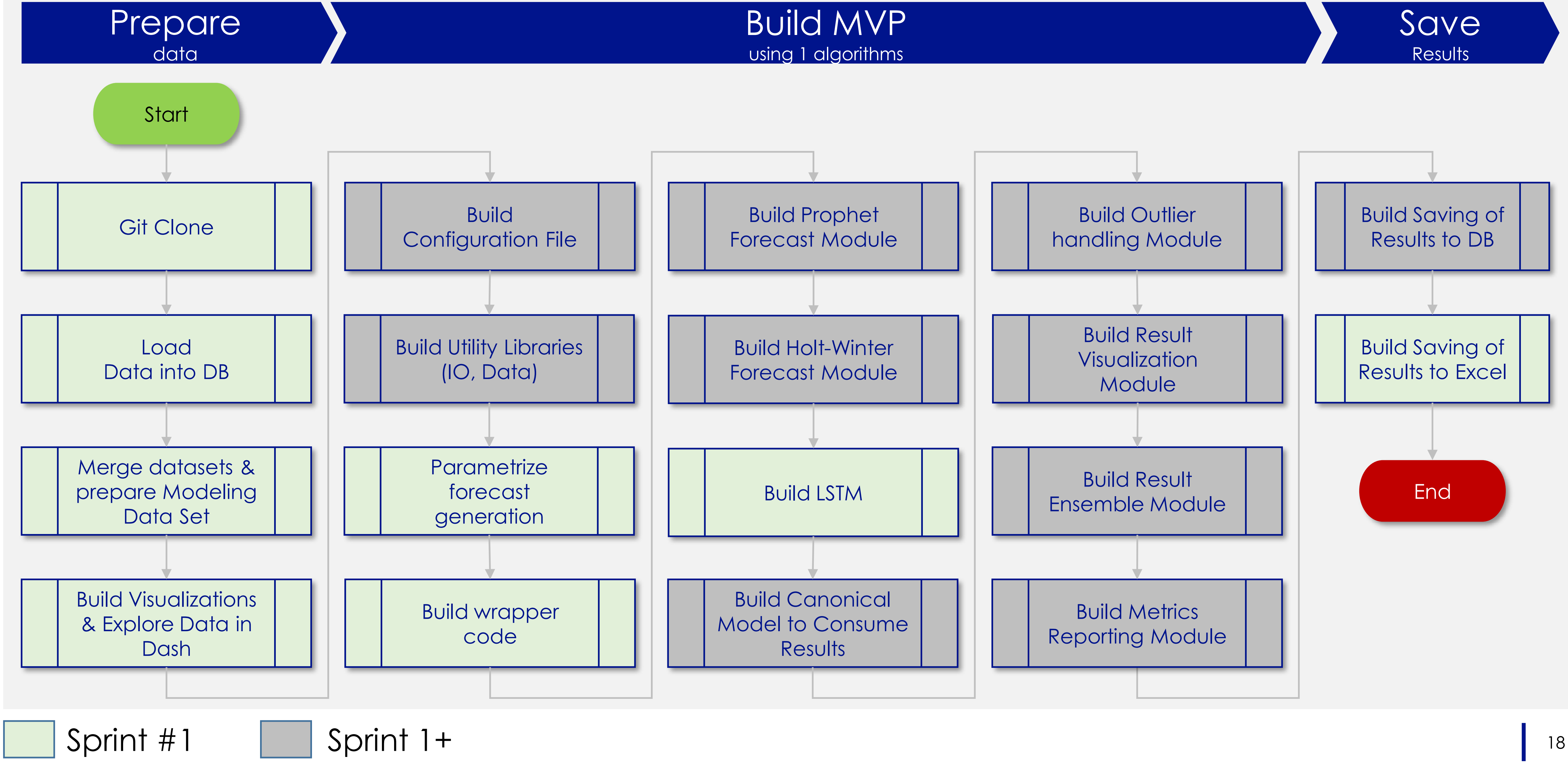


1) *Deaths rate is also climbing*

Build MVP

Algorithm development steps

Generating forecasts @ scale involves getting through 18 steps including …



Enhance Forecasting Process

Enriching components – data & algorithms

Data enrichment can occur through the inclusion of data sources such as …

#	Data Type	Example
1	**Population**	1. Size 2. Age distributions 3. Find similar towns around the globe
2	**Others**	1. Lock-down period start 2. Degree of lock-down 3. Exposed individuals and Susceptible Individual movements

… and algorithm enrichment can occur by evaluating output from algorithms including …

#	Algo. Type	Example
1	**Explainable**	1. Naïve (Moving average) 2. Multi-variate regression 3. Decision tree 4. Exponential Smoothing 5. ARIMA 6. ARIMAX
2	**Blackbox (with SHAP)**	Multi-variate scenarios (with exogenous variables) 1. Random Forest 2. Adaboost

Appendix – Files in package

Files in the package

#	Component	File(s)	Purpose
1	**Git Clone & Daily Pull**	1. Time_series_covid19_confirmed_us.csv 2. Time_series_covid19_deaths_us.csv	Raw data (https://github.com/CSSEGISandData/COVID-19)
2	**Data Prep**	1. 01_parse_covid_data.ipynb 2. 02_create_modeling_data_set.ipynb 3. 03_build_lstm_model.ipynb	1. Load data into SQL (for posterity) 2. Prepare data 3. Leverage output from data & model in Dash / Other viz. tools
3	**Data (& Result) Exploration**	1. 04_app.py (Launches a dashboard on the Web)	1. Understand the various dimensions of data 2. Use dimensions to infer patterns and improve forecasting algorithms
4	**Output**	1. Covid19_2020_04_05.csv 2. Covid_ms_data.xlsx	The modeling dataset for data prepared till 04/05/2020 The output of LSTM
5	**Discussion**	1. 05_corona_trend_mvp_2020_04_05_v2	This deck to summarize insights for broader discussion