Predicting Flattening of COVID-19 first wave

Abstract:

In December of 2019, the first outbreak of COVID-19 was detected in mainland China, eventually spreading to every continent in the world except Antarctica. Named Coronavirus Disease 19 (COVID-19) by the World Health Organization (WHO), this highly contagious disease was caused by the virus SARS-CoV-2. With a transmission rate greater than SARS or common flu, continued tremendous efforts will be needed to successfully combat this disease. In this article, we review the role that data science is playing in this war. Data science, combined with statistical analysis, computer science and computational biology, is helping in myriad ways with applications including epidemiology, drug discovery, and molecular design for diagnostic and therapeutic purposes. A number of data driven models, mathematical models, correlations and predictive models have been developed for COVID-19. Challenges faced by the data scientists today have been highlighted. Finally, open-source datasets sources are mentioned that can be potentially used in diagnostics and evaluation of health policies.

Introduction:

Determining where the next surge in coronavirus cases could happen is critical knowledge for everyone, and is of particular importance to government officials and public health practitioners who are entrusted with making decisions that affect the lives and safety of many.

Data science specialists have also concluded that graph databases are instrumental in showing them how COVID-19 spreads. A graph database shows links between people, places or things. Scientists refer to each of those entities as a node, and the connections between them are the "edges." The results give a visual representation of the relationship between things, if any.

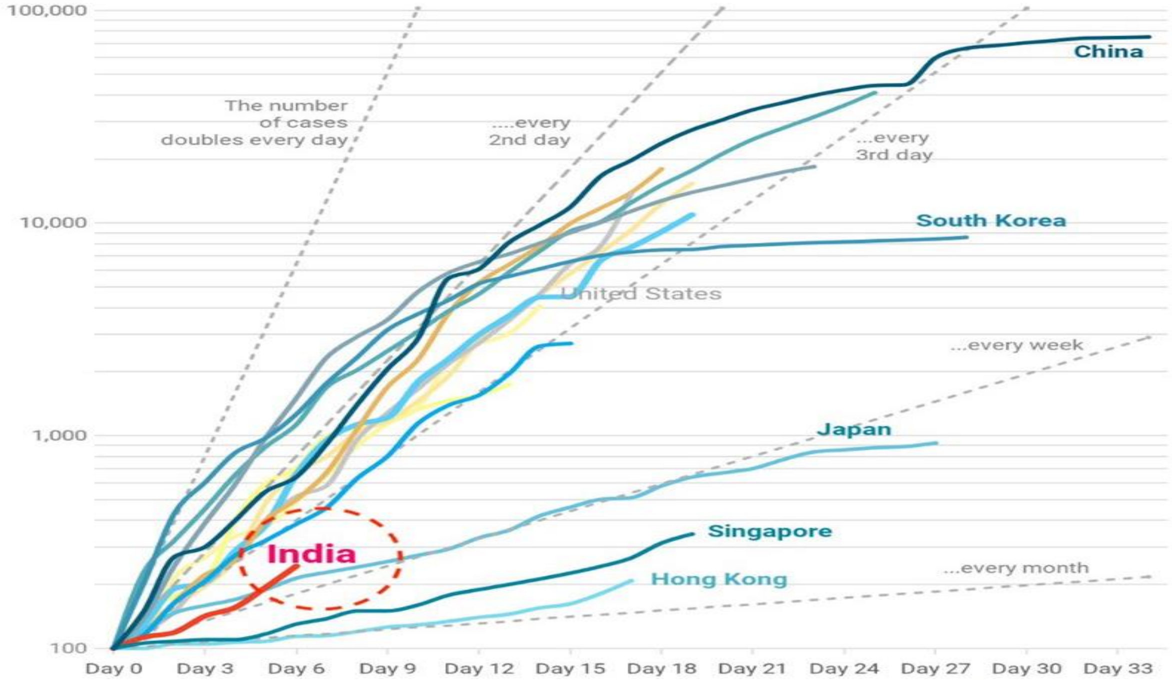

Figure 1: India compared to others

Methodology:

So, we chose country, population, population density, literacy rate, health GDP (spending on health infrastructure), testings in first 10 days and testings in first 40 days of exploding virus in that particular country, hospital beds (per million), recovery rate for 20 days and 60 days been taken as input to get the output as the weeks to flatten the curve. This model used the regressor prediction algorithm of first order.

Any data driven model or data analysis relies on the amount of data available in addition to their accuracy and reliability. As per the requirement of deep learning, the available datasets for COVID-19 are small. For audio analysis, there is no publicly available data whereas for imaging analysis and textual analysis, limited data is present. This is due to scattered sources at international, national, and hospital levels. Data security, privacy and ethics play a major role in data sharing between the sources. It is a challenge to collate and bring all data to a uniform platform. Also, the criticality of time has to be highlighted. It is seen that by the time the data is collected, collated and reaches a platform, it becomes outdated. This is perhaps the biggest hurdle in creating an analytical approach towards the problem and raises doubts on the accuracy of the models. There is a need to devise techniques taking data uncertainty in account, for example inclusion of Bayesian methods. To understand the long-term impact of the coronavirus, more multidisciplinary research is needed.