PROJECT TITTLE : COVID 19 VACCINE ANALYSIS

**Introduction:**

The use of data science methodologies in medicine and public health has been enabled by the wide availability of big data of human mobility, contact tracing, medical imaging, virology, drug screening, bioinformatics, electronic health records and scientific literature along with the ever-growing computing power [1–4]. With these advances, the huge passion of researchers and practitioners, and the urgent need for data-driven insights, during the ongoing coronavirus disease 2019 (COVID-19) pandemic [5], data science has played a key role in understanding and combating the pandemic more than ever.

COVID-19, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [6], has swept the globe and claimed over 3.4 million lives as of 19 May 2021. Because of its enormous impact on global health and economies, the COVID-19 pandemic highlights a critical need for timely and accurate data sources that are both individualized and population-wide to inform data-driven insights into disease surveillance and control. Compared with responses to previous epidemics such as SARS, Ebola, HIV and MERS, the COVID-19 pandemic has attracted overwhelming attention from not only medicine and public health professionals but also experts in other data and computational sciences fields that in previous epidemics were more peripheral [7,8].

The COVID-19 pandemic presents a platform as well as a rich data source for mathematicians, physicists and engineers to contribute to disease understanding from data-driven and computational perspectives. Some of these data were unavailable in previous epidemics, while other data were available, but their potential had not been fully unleashed. The public health systems established by many countries’ Centres for Disease Control (CDCs), including those proven to be effective in the past, were easily outflanked by the SARS-CoV-2 virus due to its very high transmissibility and the ever-increasing global human mobility. Within only a few weeks of the virus being reported it was apparent that conventional public health practices had failed in containing it. Looking back, there were notable deficiencies in the public health systems [7,8], including (a) the slow response to highly contagious viruses, particularly if the symptoms resembled those of seasonal influenza and other mild infectious diseases; (b) the lack of reliable data at critical points (such as early outbreak and mutant strains); (c) slow and disorganized data collection; (d) policy decision-making based on political expediency but not scientific evidence; (e) slow and incomplete manual contact tracing; (f) the conflict between the effectiveness of contact tracing and the invasion of privacy; and (g) difficulty in identifying effective drugs to treat COVID-19 patients.

Many of these deficiencies can be addressed by creatively mining big data related to people’s behaviours and opinions, the biological structure of drugs, human interactomes and the constantly mutating virus. The threat of the pandemic has resulted in the whole scientific community being mobilized to combat COVID-19, resulting in many successful and innovative applications. These applications required the capabilities of not only experts in one field but collaborations between people with diverse professional backgrounds. A difficult year has passed, yet it was also a remarkable year of the rise of interdisciplinary data-driven research on emerging infectious diseases. It is therefore important to summarize the progress that has been made so far, and to lay out a blueprint of an emerging field of using data science and advanced computational models to confront future infectious diseases.

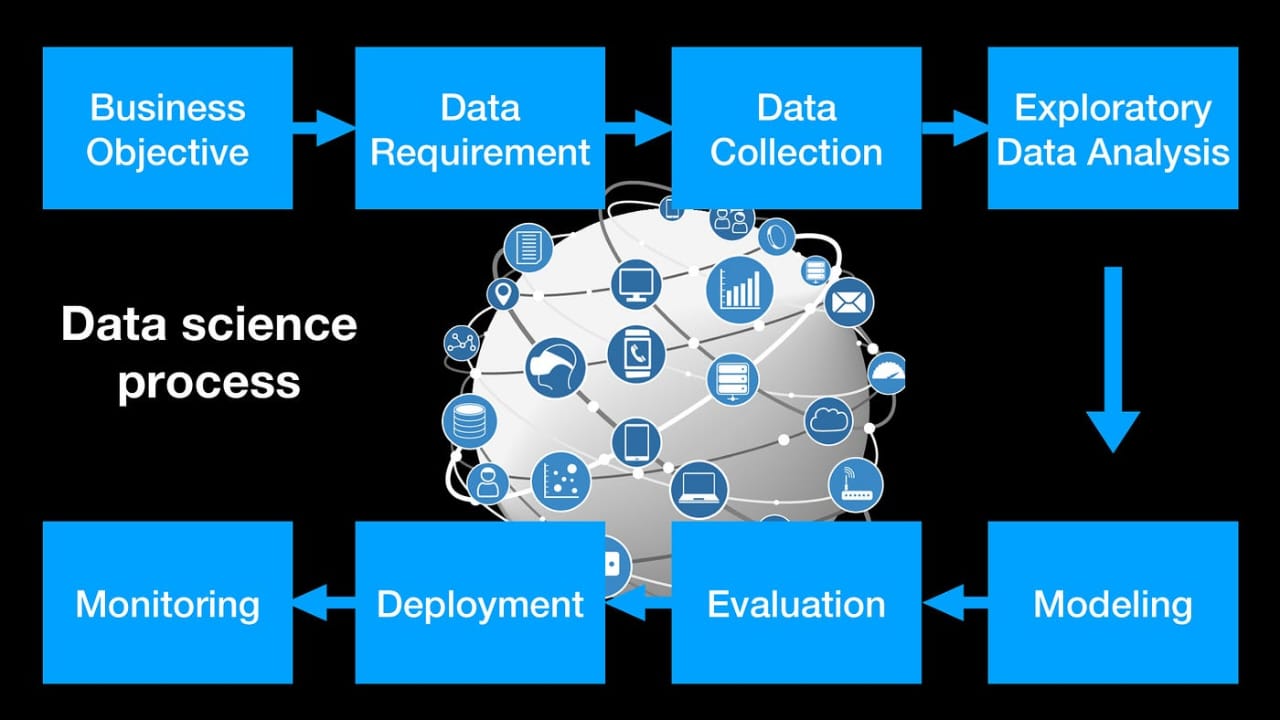
In this article, we briefly summarize the important progress made during the COVID-19 pandemic. There have been over 400 000 coronavirus-related publications in 2020 alone [9]. The list of papers we reviewed here (see table 1) is by no means complete, nor is it meant to be. Instead, we selected a set of typical and representative publications and discuss how these approaches shed light on how data science will be an indispensable tool in the ongoing war against the COVID-19 and future epidemics. The selection process is as follows. First, we used the keyword combination (‘COVID-19’ \*OR ‘2019-nCov’) \*AND (‘data science’ \*OR ‘artificial intelligence’) to retrieve all related papers during 1 January 2020 to 31 May 2021 from Web of Science by Clarivate Analytics. Second, we used the same keyword combination to further retrieve additional conference papers from DBLP (a computer sciences bibliographic database). Third, we ranked the retrieved papers in terms of the number of citations and the impact factor of the journals. Fourth, we manually added a small number of papers that we agreed to be representative but not in the highly cited list. Fifth, the authors and five PhD students manually selected the papers to review. We prioritized the representative papers published in top-tier journals.

**Modelling human mobility**

The success of using human mobility data to estimate the epidemiological parameters of the disease translates to other tasks. Travel restriction has been a popular control measure around the world in response to restricting the spread of SARS-CoV-2. Similarly, Gatto et al. used nationwide census mobility fluxes to quantify the effect of local non-pharmaceutical interventions (NPIs) and support the spatio-temporal planning of emergency measures in Italy [14]. However, a number of studies concluded that travel restriction might not be the most effective approach to containing the virus. Lai et al. and Kraemer et al. used open-source anonymized human movement data (Baidu migration data, https://qianxi.baidu.com/, derived from Baidu users) to evaluate the effect of NPIs in containing the COVID-19 epidemic in China. It found that early detection and timely isolation of infected patients was more effective than travel restrictions and contact reductions [15,16].

A number of companies provide individual or aggregated mobile phone-derived mobility data. In a representative study using aggregated mobile phone users data (provided by SafeGraph, https://www.safegraph.com/), Chang et al. developed dynamic mobility networks to simulate the COVID-19 outbreak in 10 major metropolitan areas in the USA [17]. Not only did the model predict the superspreader points of interest would account for a majority of the infections but this work also revealed risk inequities that disadvantaged groups suffered, for instance they had a higher risk of infection because they could not reduce their mobility as sharply. Liu et al. reported similar findings from a retrospective analysis of the anonymized daily mobile phone location data in China [19]. Two studies using commercial data (SafeGraph, Pei et al. [18], Teralytics https://www.safegraph.com/, Badr et al. [20]) reported that social distancing played a central role in mitigating COVID-19 transmission in the USA.

In examining the effect of NPIs in a city or smaller country, agent-based models are useful because of their flexibility and high granularity in modelling travel patterns. To better model the travel tendencies in a city, census and demographic data are required, especially when individualized mobility data are absent. For example, Koo et al. used national census data to build an agent-based model of the COVID-19 transmission in Singapore [21]. Similarly, Aleta et al. used mobile phone, census and demographic data to build an agent-based model of the COVID-19 transmission in Boston [22]. A recent study took a more aggressive approach, where Zhou et al. constructed an agent-based model with 7.55 million agents representing each citizen in Hong Kong [23]. The authors collected open government data including demographics, public facilities and functional buildings, transportation systems and travel patterns (based on census), and also incorporated the real-time human mobility patterns provided by Google’s Community Mobility Report (https://www.google.com/covid19/mobility/). The entire city of Hong Kong was split into 4905 500 m×500 m grids (refer to figure 1 for an illustration). This very detailed model was used to identify the high-value grids for targeted interventions with low disruption of the whole city.



Human mobility data are useful in informing responsive and adjustable NPIs, which can maintain economic productivity. Leung et al. used digital transactions for transport to enable real-time and accurate nowcast and forecast of COVID-19 epidemics in Hong Kong [24]. Successful application of such real-time predictions has the potential to maximize economic productivity. Yang et al. proposed a simple optimization scheme that considers both the reduction in infections and the social disruption in New York City, and concluded that tight social distancing measures in public places was the key to protect the elderly who are most vulnerable to experiencing severe disease, or death [25]. In a study in Italy, Bonaccorsi et al. modelled mobility restrictions as a shock to the economy by harnessing a near-real-time Italian mobility dataset provided by Facebook. These researchers found that mobility contraction was stronger in municipalities with greater inequality and lower income per capita, and they subsequently called for fiscal measures that targeted poverty and inequal mitigation [26].

On a global scale, Chinazzi et al. proposed a metapopulation disease transmission model that considered both air transportation and ground mobility across 3200 sub-populations in 200 countries and regions. They suggested that early detection, hand washing, self-isolation and household quarantine were more effective than travel restrictions at containing the virus [27]. Gilbert et al. used global air travel data to estimate the risk of COVID-19 importation per African country, as well as the preparedness of each country [28].

Facing a global pandemic, coordination between countries/regions is apparently a key in reducing cross-border transmissions. Ruktanonchai et al. examined the coordinated relaxation of NPIs across Europe by estimating human movements among European countries by using mobile phone data. They found that coordination of on–off NPIs is indeed important to containing the outbreak across Europe [29].

**Manual and digital contact tracing**

Contact tracing is an indispensable method to identify and isolate at-risk people, in an attempt to reduce infections in the community. During the COVID-19 pandemic, most public health practice has still relied on conventional manual contact tracing. Although such data are rarely made publicly available for research due to privacy concerns, there have been good empirical and modelling studies using it. Bi et al. analysed a complete dataset of 391 cases and 1286 of their close contacts in Shenzhen City (provided by Shenzhen CDC), China, during 14 January 2020–12 February 2020, and demonstrated that contact tracing significantly reduced the reproduction number and thus prevented a localized outbreak [30]. Zhang et al. analysed survey data for Wuhan City and Shanghai City, as well as detailed contact tracing data in Hunan Province (provided by Hunan CDC), and constructed a transmission model to evaluate the impact of NPIs on transmission [31]. They concluded that the NPIs implemented in these places had successfully controlled the COVID-19 outbreak.

Conventional manual contact tracing has major challenges, such as recall bias and time delay. The wide adoption of smartphones makes the novel digital contact tracing techniques a promising supplement to, if not replacement of, manual contact tracing [32,33]. This is particularly relevant to SARS-Cov-2, which is highly infectious. Ferretti et al. used a mathematical model to explore the feasibility of controlling the epidemic using conventional manual contact tracing by questionnaires versus digital contact tracing, and concluded that manual contact tracing is not feasible. Thus, the use of digital contact tracing is potentially more effective in stopping the epidemic given the high proportion of people using smartphones [34].

In developed countries/regions, there appear to be no technical obstacles for effective digital contact tracing because current smartphones are mostly equipped with GPS and Bluetooth [84]. Both Google and Apple have implemented frameworks in smartphones to assist in contact tracing and exposure notifications (figure 2). Since COVID-19 is likely to become endemic, digital contact tracing may eventually become a common public health practice. However, the wide implementation of digital contact tracing has not been particularly successful except for a few countries in East Asia [85]. There are many controversial issues including privacy concerns, accuracy, connection to health authorities, and other cultural and political factors [85,86]. In many lower- and middle-income countries/regions, where citizens are less technologically savvy, manual contact tracing is still playing the dominant role in containing the epidemic.

**Feature Engineering :**

Identify relevant data sources related to Covid-19 and vaccines. Preprocess the data, handle missing values, and perform data cleaning. Extract relevant features from the dataset. These features could include factors like demographics, vaccine types, efficacy rates, geographical data.

**Model Training :**

Choose appropriate machine learning algorithms for your analysis. Common choices include regression, decision trees, random forests, or neural networks. Split your data into training and testing sets to evaluate the model's performance accurately. Train your chosen model(s) using the training data. Adjust hyperparameters and algorithms as needed for better accuracy.

**Evaluation :**

Evaluate the trained models using the testing dataset. Common evaluation metrics include accuracy, precision, recall, F1-score, or area under the ROC curve (AUC-ROC). Compare the performance of different models and select the one that best fits your project's objectives. Interpret the results and draw conclusions about the Covid-19 vaccine data. Visualizations and statistical analyses can aid in understanding the patterns in the data.

Remember to document your process thoroughly, including any challenges faced and decisions made regarding feature selection and model choices. If you encounter specific issues during any of these steps, feel free to ask for more detailed assistance.