Data Science, COVID-19 Pandemic, Privacy and Civil Liberties

Bhavani Thuraisingham
Computer Science Dept.
The University of Texas at Dallas
Richardson, TX, USA
bxt043000@utdallas.edu

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377966

***Abstract*— The world has seen pandemics, terrorism, hurricanes and other natural and man-made disasters. Each time such an event occurs we discuss technologies that can solve the problem and their impact on our privacy and civil liberties. Such discussions occurred after the 9/11 terrorist attacks and is happening now during the COVID-19 pandemic, the worst human crisis we have faced in a century. This paper discusses the applications of data science to detect and possibly prevent such pandemics and its impact on our privacy and civil liberties.**

Keywords— Data Science, COVID-19, Privacy, Security, Safety, Civil Liberties

I. Introduction

The COVID-19 pandemic is the worst we have seen in a century since the Spanish Flu (which is supposed to have originated in Kansas). Since then the world has seen many wars including World War II among others, other pandemics such as SARS and the Swine Flu, the terrorist attacks such as the 9/11 in the US and 7/7 in the UK and hurricanes like Katrina and Maria. But never before has the entire world been gripped with fear like we are now due to COVID-19. It is a nightmare that dwarfs everything else that has happened to us in a century. We dread going out in case we catch the virus and could die as a result of it and yet staying home night and day makes us restless depressed, and/or anxious.

The one bright spot we have in this case is technology. Due to technology we have been able to work from home. Video conferencing technologies such as ZOOM, WebEx, Blue Jean and Teams are helping us immensely not only to attend project meetings but also participate in conferences and panels as well as socializing with family and friends. It is technology that is keeping us sane. At the same time technology, especially Data Science and Artificial Intelligence (DS/AI), is helping to detect and possibly prevent the spread of COVID-19. Furthermore, DS/AI is also helping with the development of the vaccines that is the ultimate solution to the pandemic. While DS/AI is helping humans to survive during the pandemic, it also has its problems. This is because data collection, data storage, data analytics and data sharing are at the heart of the solution to the pandemic. But such data intensive activities can also cause serious privacy violations. That is, if the data about the individuals goes into the wrong hands, then it could cause the violation of the privacy of individuals and subsequently cause great harm such as being blacklisted by insurance companies or blackmailed by adversaries. While privacy is of the utmost importance to an individual, healthcare professionals have to gather as much information as possible to treat the individual for COVID-19. Therefore, we need a balance between safety and privacy.

This paper discusses the use of DS/AI for detecting and preventing COVID-19 and at the same time explores how the privacy of the individuals may be preserved. It draws parallels between 9/11 attacks and the COVID-11 pandemic. Back in 2002, we wrote a paper titled “Data Mining, National Security, Privacy and Civil Liberties” where we argued the need for data mining for counter-terrorism and yet discussed the violations due to data privacy [1]. We are faced with a similar situation today with the COVID-19 pandemic.

The organization of this paper is as follows. We set the stage first by discussing our previous work on data mining and counter-terrorism and its implications for privacy in Section 2. DS/AI for COVID-19 is discussed in Section 3. Privacy violations are discussed in Section 4. A note about Civil Liberties is provided in Section 5. Achieving a balance between DS/AI for COVID-19 and data privacy is discussed in Section 5. The paper is concluded in Section 6.

II. Data Mining, National Security, Privacy and Civil Liberties

We have conducted research on applying data mining for counter-terrorism problems [2]. At the same time, we have also investigated the privacy implications of data mining [3]. To understand the similarities between national security and the pandemic, we will review some of the discussions in [1]. While there are many similarities from a technology point of view, there are also some differences from a societal point of view.

Data mining is the process of posing queries against large quantities of data and extracting nuggets often previously unknown [4]. Data mining has evolved into Data Science, with the emergence of big data technologies, and now with the prominence of machine learning, it is considered part of AI. Nevertheless, data mining gained a lot of prominence after 9/11 when there was great interest to gather data about various individuals and apply data mining techniques to determine whether they were engaged in terrorist acts. For example, the system would examine all the associations of an individual together with his/her behavior patterns to determine if the person was suspicious or not.

This resulted in a lot of concern among privacy advocates and the ACLU (American Civil Liberties Union). The privacy advocates were extremely concerned about false positives that could result from data mining and as a result innocent individuals could be branded as terrorists. There was also a concern about a false sense of security due to data mining and whether it was worth sacrificing one's civil liberties in the name of national security.

Subsequently there was active research on combining data mining and privacy and the first paper on an area called privacy-preserving data mining was published [5]. The ideas are to introduce perturbation and randomization into the data and then carry out data mining while ensuring that the end results would be the same. This way, the individual data values can be hidden from the adversaries who want to obtain personal data. The initial work was on privacy-preserving association rule mining and then evolved into addressing various types of data mining techniques including decision trees [6].

There were many debates and panels on data mining, national security and privacy and I was involved in many of them at that time and was a strong supporter of data mining [7], [8], [9], [10]. This was partly due to the fact that I lost two people I knew in the 9/11 attacks; one was a sponsor and the other a colleague. However, over the years my position has changed due to a better understanding of technology and policy and I firmly believe that there has to be a tradeoff between data mining, national security and privacy and policies should guide the data mining process. More details on this will be given in a later section. Also, one of my more recent notes on this topic is given in [11].

III. DATA SCIENCE FOR COVID-19 DETECTION AND PREDICTION

Data mining combined with statistics and big data has evolved into Data Science over the past ten years. At the same time, machine learning has exploded as a field resulting in Artificial Intelligence gaining a lot of prominence. We will use the terms Data Science and Machine Learning interchangeably while Artificial Intelligence goes beyond machine learning to include planning and reasoning, among others.

Data is at the heart of Epidemiology and the study of infectious diseases. Data is used not only to detect and prevent the spread of infectious diseases but also to find treatments and vaccines to treat and possibly prevent them. Data is being collected on those infected with COVID-19, including information about their personal, work and travel details, their contacts and associations with other individuals as well as their social activities. The data being collected can be analyzed using various data science techniques to extract nuggets to detect the spread of the disease as well as to prevent it. For example, the graphs built for contact tracing purposes are analyzed using say link analysis techniques to determine the persons likely to get COVID-19 from the current infections. Clustering techniques can be used to determine the clusters of potential COVID-19 cases from the data gathered. Decision trees can be used for the classification of the individuals such as those from a certain ethnic origin in a certain county are more susceptible to COVID-19 infections. What we are hoping is that such data science techniques can be used to determine those who are asymptomatic and could be potential spreaders of the virus. For example, if those who are asymptomatic are tested positive for the virus then one can study the behavior patterns and the genetic markup of these people to understand the reasons as to why they may be asymptomatic. From this information, other asymptomatic people could be tested for the virus and their contacts warned.

While detecting the virus in patients is of the utmost importance, prevention is even more critical. Due to the impact on one's life even if a person survives the disease, it is better to prevent it than to get the virus. Therefore, one could gather data from countries that have very few cases of COVID-19 and learn the trends. Then the information gleaned can be used to prevent the disease in other parts of the world.

Data Science can also be used in developing treatments and vaccines. Data driven science and medicine has been widely accepted by scientists and medical doctors. Information about the DNA of the virus, the genetic makeup of the individuals susceptible to the virus, and all information pertaining to the disease as well as information in various databases can be integrated and analyzed to determine the treatments that could be useful to COVID-19 patients. Also, data from diseases such as the Spanish Flu and other infectious diseases such as SARS and the treatments and vaccines generated for these diseases can be used to develop new treatments and vaccines for COVID-19. In other words, every piece of data including data about the patients, the prior related infections diseases, treatment and vaccines that exists for the various infectious diseases as well as historical trends in the populations of humans and non-humans has to be gathered, stored and analyzed to provide solutions to the pandemic we are faced with today.

Data collection, data storage, data sharing (between different agencies and countries and data analysis (e.g., applying data science techniques) are activities that are critical f**or** developing solutions to handle the pandemic. These solutions could be medical solutions or behavioral and social solutions such as wearing a mask covering the nose and mouth, eye goggles, face shield, gloves, hair coverings and other protective equipment. For example, experiments could be conducted with one group of people wearing masks only and another group of people wearing masks and face shields to determine of the former group has a higher probability of getting COVID-19. Similarly, the number of people in social gatherings could be increased gradually and the outcomes studied. It is critical that such investigations need accurate and sound data if we are to analyze the data to produce useful results. In addition, it is also important that we carry out a thorough risk analysis under various scenarios and keep the public informed of the potential dangers.

IV. PRIVACY IMPLICATIONS AND POTENTIAL SOLUTIONS

Whenever data is collected about individuals and this data is stored, managed, shared and analyzed, there is a high probability that the individual's privacy is violated. The simple fact that a person may be tested positive for COVID-19 could result in the personal having a huge stigma and being shunned by everyone even if the person has recovered. On the other hand, we need to inform various individuals that they may have come into contact with a person who has COVID-19. One could do that without identifying the name of the person.

Another problem with collecting and analyzing data is contact tracing. One could examine the data stored in the smart phone of the person and from that data find out about all the contacts of the person. It is crucial that we inform the person's contacts. However, this means having to go through the data in his/her phone or finding out all the activities the person carried out over the past two weeks or so. This means the person's privacy would be violated when data is collected about him or her without his or her knowledge. The data collected may not just be about the person's symptoms or who he or she has contacted. It could also pertain to his/her genetic profile.

The privacy dilemma poses a huge challenge to those working in healthcare analytics. Denying access to the data means potentially millions of deaths. Having access to the data means violating the privacy of individuals that could result in serious consequences to the person such as being denied health coverage for a preexisting condition or using the data to determine the behavioral patterns of the person and subsequently blackmailing the person. Furthermore, in order to treat the person, the healthcare providers may need access to his or her genetic profile. From this information they may learn about potential diseases the person could get. The insurance companies may deny coverage for diseases the person may not have yet. Therefore, what should we do to ensure that we provide the best care to the COVID-19 patients but at the same time ensure their privacy? The solution may lie in privacy aware and policy-based data collection, storage, management, sharing and analysis.

We have conducted research on privacy aware and policy-based data collection, storage, management, sharing and analysis [12]. That is, policies guide the activities for the entire data life cycle. For example, what are the policies for collecting the data, storing the data, managing the data, sharing the data and analyzing the data? The various privacy-preserving data science techniques look at privacy for analysis purposes. Policies should guide this process. Similarly, policies should guide the other processes also such as with whom do I share the data I have collected about patients? What are the data sharing policies? Even with data deletion, we need policies as we need to ensure that the data is deleted properly.

The next step is to examine various aspects of data science activities including the privacy aware policy-based data life cycle process and explore how blockchain technologies and be securely applied for various distributed transactions involved in these activities. In addition, smart contracts in supply chains including data supply chain as well as executing financial transactions need to be explored. Finally, blockchain applications in cyber security need to be explored further including areas such as ransom-ware and adversarial machine leaning. We believe that blockchain is the glue that integrates data science with cyber security.

V. WHAT ABOUT CIVIL LIBERTIES?

The ACLU (American Civil Liberties Union) has published excellent articles on balancing contact tracing vs. privacy [13], [14]. These articles discuss the proposals put forward by smartphone companies such as Apple and Google on contact tracing using Bluetooth technology. They also strongly support policy-based data collection and analysis. For example, the article states the following:

"*Technology principles that embed privacy by design are one important type of protection. There still need to be strict policies (https://www.justsecurity.org/69444/how-to-think-about-the-right-to-privacy-and-using-location-data-to-fight-covid-19/) to mitigate against overreach and abuse. These policies, at a minimum, should include:*

- ***Voluntariness** — Whenever possible, a person testing positive must consent to any data sharing by the app. The decision to use a tracking app should be voluntary and uncoerced. Installation, use, or reporting must not be a precondition for returning to work or school, for example.*
- ***Use Limitations** — The data should not be used for purposes other than public health — not for advertising and especially not for any punitive or law enforcement purposes.*
- ***Minimization** — Policies must be in place to ensure that only necessary information is collected and to prohibit any data sharing with anyone outside of the public health effort.*
- ***Data Destruction** — Both the technology and related policies and procedures should ensure deletion of data when there is no longer a need to hold it.*

- ***Transparency** — If the government obtains any data, it must be fully transparent about what data it is acquiring, from where, and how it is using that data.*
- ***No Mission Creep** – Policies must be in place to ensure tracking does not outlive the effort against COVID-19.*

These policies, at a minimum, must be in place to ensure that any tracking app will be effective and will accord with civil liberties and human rights."

While privacy advocates are still concerned about the privacy violations, they also understand the need for data collection and analysis that would save lives. However, when we compare the ACLU's position with their position back in 2002, we see a change. This is partly due to the fact that we know more now about policy aware data collection and analysis. The concern back then was that we do not want to have a false sense of security at the expense of privacy. Furthermore, at that time certain individuals were detained by the government even when they had no connection to terrorism. However, the pandemic is quite different. Even if a person seems perfectly healthy it may be necessary to test the person as he/she may be asymptomatic. Therefore, there seems to be more tolerance towards testing and contact tracing with COVID-19 then there was in investigating people with respect to counter-terrorism.

Apart from privacy advocates, there are also people who refuse to wear masks and stop going to bars as they claim that their civil liberties are being violated. Although back in 2002, the general public did not raise as much concern when they were pulled from lines at airports for more extensive examination. This may be due to the reasons that we were able to see the gruesome terrorist attacks on television. With respect to COVID-19 we see what happens in hospitals and for some reason people do not seem as concerned about it because they may think that this would not happen to them. I believe that constant education is key to this problem. People have to be warned daily that by not taking precautions not only could they die, they could also infect others including their elderly relatives who have a much higher probability of dying. Wearing a mask or face shield causes some inconvenience, but it is worth making small sacrifices (or even big sacrifices) to save the human race. Politics should not interfere with human lives – not in 2019/2020 nor should it have been in 2001/2002.

VI. Balancing Safety and Security vs. Privacy and Civil Liberties

The discussions in the previous sections show that regardless of whether we are discussing national security or the pandemic, it's all about balancing safety and security (e.g., saving people's lives and protecting the individuals from terrorist attacks with privacy and civil liberties). Would it not be wonderful if we can have all: safety, security, privacy and civil liberties? But we know it is not possible. In order to be safe, we need to sacrifice some aspects of our civil liberties. The question is how much? That is our challenge.

Some would argue that if we don't have all the data to analyze and give advice to epidemiologists and physicians so that the COVID-19 patients get the proper medical care, we are in grave danger of having mass casualties. Those who believe that civil liberties come first would argue we are not a civilized society if the data is used for bad purposes to blackmail individuals and deny them health coverage. Furthermore, they argue that data getting into the wrong hands could also cause hostile acts like murder and rape. That is, knowing the whereabouts of individuals would give ideal opportunities for murders and rapists. In the end we need tradeoffs. During certain times we must focus on safety (e.g., pandemic worsening) and during some other times privacy should be given greater consideration (e.g., when the number of cases is few and far between).

There is now a global initiative called "AI for Good" [15] and the United Nations is also promoting this initiative. This initiative focuses on all the benefits of AI to help humans. But recently we wrote an article "Can AI be for Good in the midst of Cyber Security Attacks and Privacy Violations" [16]. For example, we argued that what happens if the AI techniques are attacked? We also discussed the privacy implications of AI. My focus in that article was on children's rights and preventing child abuse. We need to examine this initiative with respect to COVID-19. We need organizations such as ACLU and the United Nations to work together to handle the challenge we are faced with and that is how do we balance safety and security with privacy and civil liberties?

VII. Summary and Directions

This paper has discussed the applications of data science for science to the COVID-19 pandemic and then described the serious side effects such as violations of data privacy and civil liberties. It also provides an analogy to the dilemma we were faced with soon after 9/11 on the conflicts between national security and privacy. We argued that we need a privacy aware and policy-based framework for data collection, storage, management, sharing, analytics and even detection. We also mentioned that we should make small sacrifices such as wearing masks to save the human race from this global pandemic.

We have only discussed the problem and solutions at a very high level. We need to develop a conceptual framework and subsequently the detailed design to develop an architecture and a system that would carry out policy-aware activities for COVID-19. We need to take into consideration the guidance provided by organizations such as the ACLU and the activities of UN to develop such a framework. Our framework has to be flexible in the sense that during certain times, such as a massive increase in the number of COVID-19 cases, we must focus on data collection and analysis and during other times such as less cases we can give more attention to privacy and civil liberties. But we have to be careful as we cannot afford to be complacent. This is because if we relax sometimes even a little, then the infections could continue to explode. Therefore, we need an ideal balance between the two and that will be our challenge as we go forward to achieve a new normal that we are all comfortable with.

References

[1] B. Thuraisingham,:Data Mining, National Security, Privacy and Civil Liberties. SIGKDD Explorations 4(2)*: 1-5 (2002)*

[2] B. Thuraisingham, Web Data Mining with Applications to Counter-terrorism and Business Intelligence, CRC Press, 2003.

[3] B. Thuraisingham, Privacy-Preserving Data Mining: Development and Directions. J. Database Manag. 16(1)*: 75-87 (2005)*

[4] B. Thuraisingham, Data Mining: Technologies, Techniques, Tools and Trends, CRC Press, 1998.

[5] R. Agrawal, and R. Srikant, "Privacy-preserving Data Mining," Proceedings of the ACM SIGMOD Conference, Dallas, TX, May 2000.

[6] Li Liu, Murat Kantarcioglu, Bhavani M. Thuraisingham*:* Privacy Preserving Decision Tree Mining from Perturbed Data. HICSS 2009*: 1-10*

[7] B. Thuraisingham, Data Mining for Counter-Terrorism, Presented at the White House Office of Science and Technology Policy, February 2002.

[8] B. Thuraisingham, Data Mining for Counter-Terrorism, Presented at the Database Program panel at Stanford University, March 2002.

[9] B. Thuraisingham, Data Mining for Counter-Terrorism, Panel at the IFIP 113 Conference on Data and Applications Security, University of Cambridge, England, July 2002.

[10] B. Thuraisingham, Data Mining for Counter-Terrorism, Presented at the United Nations, September 2002 (also at the White House Office of Technology and Policy, 2002)

[11] B. Thuraisingham, Keeping Better Tabs on Suspicious Persons (NY Times Opinion Column) January 13, 2015 https://www.nytimes.com/roomfordebate/2015/01/12/when-known-jihadists-come-home/keeping-better-tabs-on-suspicious-persons

[12] B. Thuraisingham, et al, Towards a Privacy-Aware Quantified Self Data Management Framework. SACMAT 2018: 173-184

[13] A COVID-19 BALANCING ACT: PUBLIC HEALTH AND PRIVACY (EP. 97) https://www.aclu.org/podcast/covid-19-balancing-act-public-health-and-privacy-ep-97

[14] Apple and Google Announced a Coronavirus Tracking System. How Worried Should We Be? https://www.aclu.org/news/privacy-technology/apple-and-google-announced-a-coronavirus-tracking-system-how-worried-should-we-be/

[15] United Nations, AI for Good, https://en.wikipedia.org/wiki/AI_for_Good

[16] B. Thuraisingham, Can AI be for Good in the Midst of Cyber Attacks and Privacy Violations? A Position Paper. ACM CODASPY 2020: 1-4