COVID-19 and computational sciences: data variety can be an enabler for good science– if properly utilized

Mayank Kejriwal (Viterbi School of Engineering, University of Southern California)

Abstract

This brief argues for recognizing the variety (rather than just volume) of data as a valuable enabler for rigorous computational science, facilitating robust, data-driven policymaking during a global crisis such as COVID-19. Drawing on peer-reviewed findings and citations from both our own research group and others to make our case, we present three concise recommendations for further incentivizing research into this important issue.

Although an unprecedented crisis of our century, a silver lining of the COVID-19 pandemic was that it showcased the power of both medical science and technology to address global challenges at a pace never witnessed before in human history [1,2]. Technology played an important role both because of the abundance of data [3,4], as well as the easy access to inexpensive, open-source and customizable software that anyone with reasonable coding skills could set up [5]. In the early days of the pandemic, the focus was rightfully on building practical applications that could be useful to medical professionals. More recently, however, the data has been used to facilitate computational social science to better understand the pandemic from multiple viewpoints [6].

Mining useful insights and trustworthy information from raw data has been a longstanding problem in the Artificial Intelligence (AI) community, and in an age of misinformation, is not trivial for humans either [7]. While scale can sometimes be an issue for some methods and infrastructures, well-designed algorithms and cheap compute power partially help alleviate this issue [8]. A larger problem involves deriving trustworthy findings when the provenance of the data itself is uncertain, and when various (often unknown) biases may be potentially present[1].

Even partial solutions to these problems could significantly help advance the state-of-the-art, and lead to greater uptake of advanced technology among non-technical stakeholders and skeptical domain experts. In the case of COVID-19, these would not include hospitals, nurses and doctors, but also pharmaceutical companies and government agencies that are respectively looking to allocate scarce resources (for developments of vaccines and cures, and running of clinical trials) and make difficult policy decisions in a high-risk pandemic.

In contrast with the core AI application areas, the broader computational sciences community, including computational social sciences, have been actively looking into issues involving various kinds of biases, trust and the 'science' of dealing with heterogeneous, and often, observational, data [9]. One of the key findings, presented subsequently, is that even with biases, many datasets are still useful for evaluating *appropriately formulated* hypotheses. The challenge then is not always to eliminate the 'problems' with a dataset, or to discard it, but to instead design the correct hypotheses that it can be used to evaluate. Here we refer to 'dataset' broadly as not just a single set of records, but potentially, a collection of independent (and possibly, integrated) datasets that researchers are relying on for their findings.

The core thesis of this brief is that data variety[2] needs to be recognized as an enabler for rigorous and ambitious science, rather than as an unwanted technical challenge to be smoothened over. In the next section, we present some examples of this in action during COVID-19. As some of the findings will lead us to subsequently conclude, appropriately using the different kinds of data, each with its own set of biases and characteristics, for doing both predictive analytics and hypothesis-driven science, requires different groups of scientists and practitioners to proactively collaborate together.

Findings

In an article published in the Harvard Data Science Review [2], we cited a range of COVID-19 computational studies spanning both social science and technology-centric areas such as information retrieval, knowledge graphs and natural language processing [10,11]. While there were only a few examples where heterogeneous data sources were used in the same framework or system architecture, there was also promising evidence

[1] A non-comprehensive set of examples being sampling bias, selection bias and non-response bias.

[2] Recall that the 4Vs of Big Data are usually held to be volume, variety, veracity and velocity. Traditionally, there has been much focus on volume, and more recently, in the age of rapid social media dissemination, veracity and velocity. There has been far little focus on variety, on the other hand, except in very specific application areas.

of different datasets being utilized in a piecemeal fashion in separate communities to answer narrowly tailored questions [12,13]. Despite having varying degrees of quality, volume and other such characteristics, the datasets clearly have some value, depending on the community. At the same time, there is reason to suspect that, because of the siloed nature of these communities, with little overlap in the sets of authors, researchers or (in some cases) even institutions, much more is possible than what is currently being done.

In the next section, we draw on our practical experience running a research group to enumerate some recommendations that will make this kind of collaboration more likely. Below, we point to several examples of how different datasets can provide their own value, using work generated both in our own research group, as well as those of others.

As one specific example, in our group, we recently used Twitter data for a variety of useful purposes, while being aware of the data's limitations and the broader ethical question of privacy. In [14], for instance, we recognized the findings of previous authors' research that Twitter tended to be biased toward more metropolitan and heavier populated areas. Hence, rather than collect and analyze Twitter data broadly, under the implicit (but faulty) assumption that a random sample from Twitter reflects a random sample from the general population, we primarily focused on metropolitan areas. We were also careful to describe the data collection and provenance in replicable detail in the article, which was published after undergoing peer review, and we presented examples on how the data could be used to make bounded statements. In another study, recently accepted after peer-review, we showed how Twitter data could be appropriately utilized to hypothesize about causes of vaccine hesitancy [15].

In some instances, data variety can be leveraged to mine insights from small quantities of data, or by combining aggregate statistics from two or more independent survey studies. Such work may be deemed too preliminary to meet the bar for a high-quality peer-reviewed venue, but could still be published as preprints. An example is a piece on food insecurity published on the PsyArxiv preprint server [16]. In that piece, we used aggregate data from a specific Gallup COVID-19 survey, which was rigorously designed and implemented a few months after the pandemic was declared. We combined this data with data collected by the United States Census Bureau to showcase the disparate impacts of food insecurity in different metropolitan areas, and its correlation to subjective wellbeing. We found, worryingly, that *loneliness* was best associated with food insecurity, in line with more qualitative sociological research on the subject [17].

Examples of research from other groups that have also sought to leverage data variety include the work in [12,18]. We note that these strands of research are largely systems-driven, and although this is valuable from an engineering and applied perspective, we advocate that data variety can also be an asset in support of rigorous and robust science, especially when investigating complex global phenomena, such as COVID-19.

Policy Recommendations and Conclusion

Based on the arguments laid out at the outset of the brief, and the findings and examples from the previous section, we propose three recommendations that could proactively incentivize the development of effective approaches for maximizing *scientific* utility from a broad and varied set of data sources:

- First, more government-funded research programs are needed from the likes of the US National Science Foundation to enable scientists and engineers to proactively utilize data variety for evaluating a robust and more comprehensive range of scientific hypotheses than is currently possible. Research also needs to be funded into the appropriate statistics and data science methodologies for such hypotheses.
- Second, there needs to be a better recognition in the data science and AI communities that ingesting and handling varied datasets for scientific decision making is a valuable research agenda. From a policy standpoint, this could be enabled by sponsoring more conferences and symposia on how best to use data variety in support of good science. At the same time, there also needs to be focus among academics on communicating these findings to the general public.
- Third, there need to be concerted efforts at national and international levels to share data more proactively using the FAIR (Findability, Accessibility, Interoperability, and Reuse) principles [19]. For certain sensitive collections of datasets, or due to regulatory hurdles, novel technologies (such as privacy-preserving federated machine learning [20]) should be considered and incentivized.

We conclude with the sobering note that, in these polarized times, trust is being eroded at an alarming rate in established institutions and practices, such as science. In these times, it is all the more important for us to continue investing in research that enables us to

strengthen our methods and statistics, while taking on grander scientific questions. Properly implemented, scientific methodology and statistics both offer us avenues for trusting the insights gleaned from our datasets with more rigor. It may also help us communicate our results better to the general public, and restore the trust that is being eroded.

References

[1] Renu, N. (2021). Technological advancement in the era of COVID-19. *SAGE Open Medicine*, *9*, 20503121211000912.

[2] Kejriwal, M. (2020). Knowledge graphs and COVID-19: opportunities, challenges, and implementation.

[3] Hamzah, F. B., Lau, C., Nazri, H., Ligot, D. V., Lee, G., Tan, C. L., ... & Chung, M. H. (2020). CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. *Bull World Health Organ*, *1*(32), 1-32.

[4] Shuja, J., Alanazi, E., Alasmary, W., & Alashaikh, A. (2021). COVID-19 open source data sets: a comprehensive survey. *Applied Intelligence*, *51*(3), 1296-1325.

[5] Wu, T., Hu, E., Ge, X., & Yu, G. (2020). Open-source analytics tools for studying the COVID-19 coronavirus outbreak. *MedRxiv*.

[6] Zhang, J. J., Wang, F. Y., Yuan, Y., Xu, G., Liu, H., Gao, W., ... & Chen, K. C. (2021). Guest Editorial Computational Social Systems for COVID-19 Emergency Management and Beyond. *IEEE Transactions on Computational Social Systems*, *8*(4), 928-929.

[7] Jin, F., Wang, W., Zhao, L., Dougherty, E., Cao, Y., Lu, C. T., & Ramakrishnan, N. (2014). Misinformation propagation in the age of twitter. *Computer*, *47*(12), 90-94.

[8] Cano, A. (2018). A survey on graphic processing unit computing for large-scale data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1), e1232.

[9] Alvarez, R. M. (Ed.). (2016). *Computational social science*. Cambridge University Press.

[10] Kejriwal, M., Knoblock, C. A., & Szekely, P. (2021). *Knowledge Graphs: Fundamentals, Techniques, and Applications*. MIT Press.

[11] Kejriwal, M. (2019). *Domain-specific knowledge graph construction*. Cham: Springer International Publishing.

[12] COVID*GRAPH. (2020, March 30). *We build a knowledge graph on COVID-19*. ODBMS. http://www.odbms.org/2020/03/we-build-a-knowledge-graph-on-covid-19/

[13] Nagpal, A. (2020, April 27). Yahoo Knowledge Graph Announces COVID-19 Dataset, API, and Dashboard with Source Attribution [Blog post]. https://developer.yahoo.com/blogs/616566076523839488/

[14] Melotte, S., & Kejriwal, M. (2021). A geo-tagged COVID-19 Twitter dataset for 10 North American metropolitan areas over a 255-day period. *Data*, *6*(6), 64.

[15] Luo, Y., & Kejriwal, M. (2021). Understanding COVID-19 Vaccine Reaction through Comparative Analysis on Twitter. *arXiv preprint arXiv:2111.05823*.

[16] Kejriwal, M., & Shen, K. (2021). Affective Correlates of Metropolitan Food Insecurity and Misery during COVID-19.

[17] Gaines-Turner, T., Simmons, J. C., & Chilton, M. (2019). Recommendations from SNAP participants to improve wages and end stigma. *American journal of public health*, *109*(12), 1664-1667.

[18] Bragazzi, N. L., Dai, H., Damiani, G., Behzadifar, M., Martini, M., & Wu, J. (2020). How big data and artificial intelligence can help better manage the COVID-19 pandemic. *International journal of environmental research and public health*, *17*(9), 3176.

[19] Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., ... & Schultes, E. (2020). FAIR principles: interpretations and implementation considerations. *Data Intelligence*, *2*(1-2), 10-29.

[20] Agrawal, R., & Srikant, R. (2000, May). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 439-450).