COVID-19 Polymath Project Proposal

Christopher Strohmeier

March 24, 2020

1 Introduction

There is an abundance COVID-19 related data available. This data will be indispensable for the urgent tasks of understanding and preventing the spread of the virus as well as developing treatments. Moreover, there are a myriad of other e.g. economic, political, and mental health problems which have resulted from the pandemic. An organized collection of high-quality, clean, and accessible data sets will be indispensable in addressing these and many other issues.

1.1 Desirable Properties of a Database

Some of the properties such a database ought to possess include:

1. High Quality, Clean: data sets should be accurate, large, and relevant. While for the sake of flexibility it is important to have the original, "unclean" data set, ideally our collection should also contain "clean," versions of said data sets which are ready to be analyzed immediately.

2. Accessible: data sets should be easy to find. It should be extremely clear what a particular data set contains, what it does not, what kinds of data our database as a whole contains, what it does not. There should be detailed descriptions of the data, but there should also be an accompanying document to quickly explain the contents of the data.

3. Variety: the database should include a myriad of data relevant to combating COVID-19 and all of the problems it has caused.

4. Collaborative: this is a time for collaboration, not competition. In addition to having data with the desirable attributes mentioned above, the environment in which this data is distributed should allow for users to share ideas, easily share data sets and analyses of their own, find inspiration from existing projects, and work with one another.

1.2 Some Improvements on Kaggle COVID-19 Initiative

Unfortunately, as of this writing there does not appear to be a database satisfyng the above properties. The closest existing platform of which I am aware was created by the data science company Kaggle `https://www.kaggle.com/tags/covid19`. It's a start, but there is a lot to be improved upon. Regarding the desirable attributes above:

1. While there are some high quality data sets available, such as the scientific journal text data provided through kaggle `https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge`, almost none are clean in the sense of being immediately usable. This is a huge problem, as data cleaning can be an extremely time consuming process. Worse still, data scientists everywhere are likely cleaning the same data sets, and so there is a lot of wasted effort.

2. The data could be way more accessible than it currently is. For starters, the data sets available on Kaggle are not grouped together very well, if at all. Also it's nontrivial to determine whether or not a desired data set meetings one's needs exists. Perhaps data scientists proper would not have some of these accessibility issues, but making the data more accessible to a wider audience of mathematicians would be a tremendous help.

3. The list of questions related to COVID-19 put forth on Kaggle `https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks` are important, but very limited. One could argue that we should first deal with the most pressing issues of stopping the spread of COVID-19 and developing treatment first and worry about less urgent matters later, but basic economic concepts like law of diminishing marginal utility and comparative advantage suggest this thinking is misguided. In particular, while it is true it can be useful to approach the same problem such as "predict the spread of COVID-19" in different ways, if everyone is focusing on this one issue then other tasks are being neglected, and again there is the issue of work being repeated. Also many data scientists may have secondary interests in fields like economics or psychology. While they are still capable of contributing to the "most important problems" above, comparative advantage suggests there time may be better spent on problems which have sprung out of the pandemic which are more suited to their skill set. Some ideas like this are put forth below.

4. Kaggle is first and foremost a website for hosting data science competitions. While for the most part the attitude of kaggle users working on COVID-19 has been collaborative, the structure of the website is not optimally conducive to collaboration.

2 Proposed Database Format

*****It's worth reiterating that the ultimate function of this project is to organize a wide variety of high quality, clean, data, which address problems both directly and indirectly related to COVID-19, so that data scientists, mathematicians can get to work analyzing said data as quickly and efficiently as possible. Furthermore, the environment should enable collaboration so as to avoid wasted/overlapping effort, and a joint effort on some of the big problems, while also providing data in a form suitably flexible that anyone with a particular interest could go in and start analyzing what they need as soon as possible.

Taking for granted that we've agreed on what a "good" database should look like, there still remains the practical question of how to obtain such a database, or at least a close approximation.

2.1 Some Considerations for our Database

It's important to have a concrete plan for what we want to contribute. Below are some considerations to keep in mind.

1. Urgency: Some preliminary but satisfactory form of this database needs to be up and running as soon as possible.

2. Solid Foundation: we should be able to build upon our original platform easily. Adding new features should not involve modifying substantial portions of earlier versions of our platform.

3. Accessibility/Organization: it should be extremely easy to find what you're looking for.

4. Concrete Questions: e.g. "given this particular (high quality, clean) data set consisting of trends in the number of new reported COVID-19 cases each day and some other (high quality, clean) data set involving trending twitter topics pertaining to COVID-19, find some interesting correlation" as opposed to the far more general (but of course, still useful) questions that have been advertised on Kaggle.

5. Variety: the platform should include data sets useful for addressing not just the immediate problems pertaining to COVID-19 but also to other, related problems.

6. Inspiration: we could suggest interesting questions to work on. Even better, if this get to the crowd sourcing stage, users could discuss ideas in a forum-type format.

7. Data Cleaning Request: we should be able to request that a particular data set be clean, specify instructions for the cleaned dataset.

8. Our Results: Particularly in the initial release, we should make our results public. Aside from the intrinsic utility of the results, this would also get some attention and validate the platform.

2.2 ***Details of Format

This is one possible layout which attempts to address all of the considerations in the previous subsection.

1. Data Set Directory: all things data set related
   - (a) Data Set Wish List: forum for requesting datasets to be cleaned, possibly with upvote/downvote system. Might be important enough to make as a main directory.
   - (b) Data Collection 1 (e.g. Reported COVID-19 Cases)
   - (c) Data Collection 2 (e.g. Stock Market Data)
   - (d) Data Collection 3 (e.g. Twitter Data)
     - i. Twitter Directory 1
     - ii. Twitter Directory 2
       - one particular raw, uncleaned twitter data set
       - a clean version of this data set
       - .txt file with a quick summary of raw data, the cleaned data, some suggested (general) questions for analyzing data, more detailed description of data.
     - iii. Twitter Directory 3
     - iv. Twitter Directory 4
     - v. Twitter Directory 5
     - vi. Twitter Directory 6
     - vii. etc.
   - (e) Data Collection 4 (e.g. Mental Health Data)
   - (f) Data Collection 5 (e.g. Lung X-Ray Data)
   - (g) Data Collection 6 (etc.)
2. Problems Directory
   - (a) Directory containing most important questions
   - (b) "Braindump" forum for discussing different ideas, maybe with an upvote/downvote system, should make it easy for people to connect and collaborate
3. Results Directory
   - (a) Selected Results Directory
   - (b) Another forum for sharing results, again with a possible upvote/downvote structure.

3 COVID-19 Related Topics and Questions

Below are some possible COVID-19 related topics and questions. Needless to say understanding relationships between some of these ought to be interesting, e.g. lung x-ray and respiratory audio data together might make it easier to detect COVID-19 earlier, differentiate from other respiratory problems.

- Daily New Cases of COVID-19: understand trends in number of new reported cases, compare outbreak in different cities/countries.
- Twitter Data: understand trending topics pertaining to COVID-19. Do they indicate panic? Do they suggest that basic information about the disease is circulating effectively? Who is seeing what information? Understand qualitative difference between tweets which spread health and safety information effectively and those that do not.
- Media Data: Similar questions for twitter data.
- Scientific Journal Text Data: topic modeling: compare COVID-19 to other forms of coronavirus, other pandemics.
- Lung X-ray Data: enhance COVID-19 detection process. Save time and resources by releasing patients which do not have COVID-19 earlier.
- Respiratory Audio Data: breathing and cough abnormalities useful for detecting COVID-19?
- Stock Market Data: how does news about COVID-19 affect stock market?
- Supply/inventory data: understand how to allocate supplies? How to predict when supply will be depleted?
- Productivity: lot of self-help blogs about personal growth and productivity. Perhaps there's a useful way to synthesize insights from this large body of work, help people to be productive during quarantine, make the best of a bad situation.
- Mental Health Data: how best to connect people to mental health professionals? Study effects of isolation during quarantine? How best to keep people connected and mentally healthy?