Prediction and Causality:

How Can Machine Learning be Used for COVID-19?

Mark Crowley
Assistant Professor
Electrical and Computer Engineering
University of Waterloo

Submitted: February 18, 2021

Prediction and Causality

How Can Machine Learning be Used for COVID-19?

Abstract

Machine Learning is a very popular and very powerful set of analytical and algorithmic tools for helping human beings understand complex datasets and to take action to achieve their goals. This article will attempt to provide some perspective on the relevance of Machine Learning research for the global struggle with the Covid-19 disease, including understanding how machine learning practitioners and researchers view the usefulness of their toolbox and how they approach a problem such as Covid-19. An interesting fact about the current widely available ML tools is that they are especially good at one particular type of learning, supervised learning, where full knowledge of the "correct" inputs and outputs for the problem are fully known. This is a well-known limitation, which many ML researchers are working hard to rectify, but it means that it is often difficult to ML methods to answer questions in a field such as medicine where the datasets are often small and the amount of full knowledge, labelled input-output data needed is even smaller. There is a vocal, and growing, community within ML research that argues that in order to be truly beneficial to scientific pursuits and society at large, what is needed is for ML to have a greater focus on *causality*. We provide a basic introduction to the idea behind the causal hierarchy and how the existing ML methods relate to it.

What is Machine Learning?

They say that when all you have is a hammer, then all the world is a nail. For machine learning (ML) researchers and data scientists, aiming to help the world adapt and respond to the Covid-19 pandemic, the currently popular hammer is called Deep Learning. But in fact, there is a whole toolbox to available to be used, which requires a lot of expertise about those tools but also about the domain in question. The potential impact ML methods can have here is tremendous, but so are the challenges, since a rapidly changing pandemic involves a number of very difficult questions and constraints that most other domains popularly using ML do not have.

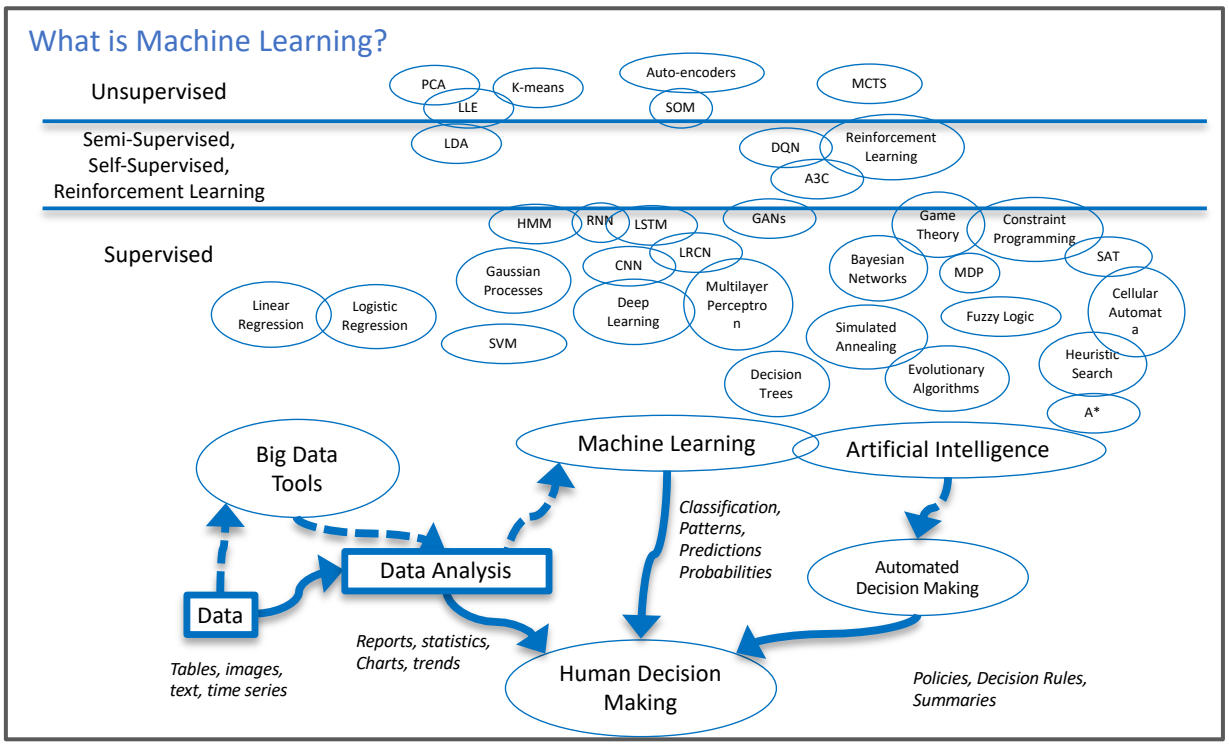

Figure 1: An overview of the pipeline from data to human decision making which includes AI/ML and many specific algorithms which are applicable depending on the datatype and situation.

If we start at the beginning, we have to ask what range of techniques and concepts are we talking about when we say, “Machine Learning”, and how do its practitioners see a problem such as the Covid-19 pandemic? Figure 1 provides my own overview of the virtual pipeline from data to human decision making which is the entire point of the study of ML and more broadly, Artificial Intelligence. Data can come in many forms such as numerical tables, continous signals, gigabytes of images, natural language text documents, etc. For each type data there are customized ways of preparing, cleaning up, and extracting the essential or most representative information to be used in further tasks. There are many specific algorithms which are applicable depending on the datatype and situation. While the field looks vast and complex, there are really only a small numbers of major patterns of algorithms based on underlying concepts or metaphors including : projections across high dimensional spaces, linear weighted functions with free variables to tune, trees that divide the data dimensions interatively and neural networks which take linear weighted functions to a combinatorial extreme and add non-linear activation thresholds. Regardless of the metaphor used, all of these approaches can be considered as a box that takes inputs data instances (images, patient records, sensor readings,…), represented somehow as a collection of numbers and producing an output, which is also a collection of numbers which can be interpretted as new data or probabilities. There are two important points to notice here. Firstly, there are many more methods

in the "supervised" category of algorithm. Secondly, all of these methods are focussed towards helping humans make decisions in the end, so any use should be in aid of that.

What is COVID-19 to a Machine Learning researcher?

From the point of view of an ML researcher or data scientist, the entire situation the world is in now with Covid-19 boils down to a very challenging problem of gathering widely distributed, and diverse data, full of uneccessary variations in standards, full of uncertainty and errors; and then being asked to make very important predictions about outcomes to incredibly complex systems which are not even fully understand by the scientists studying them. Great! AI/ML people love nothing more than an impossible problem. So there are two aspects here that we'll highlight. The first, is that we don't *need* to have AI/ML solve this full problem in order to help people, today, on the ground. This is exemplified by the kind of practical work being doing by ourselves and many others already. Second, is that in order to get beyond these very short-term, practical solutions to parts of the problem, we may need to rethink what ML is good for and what we actually need.

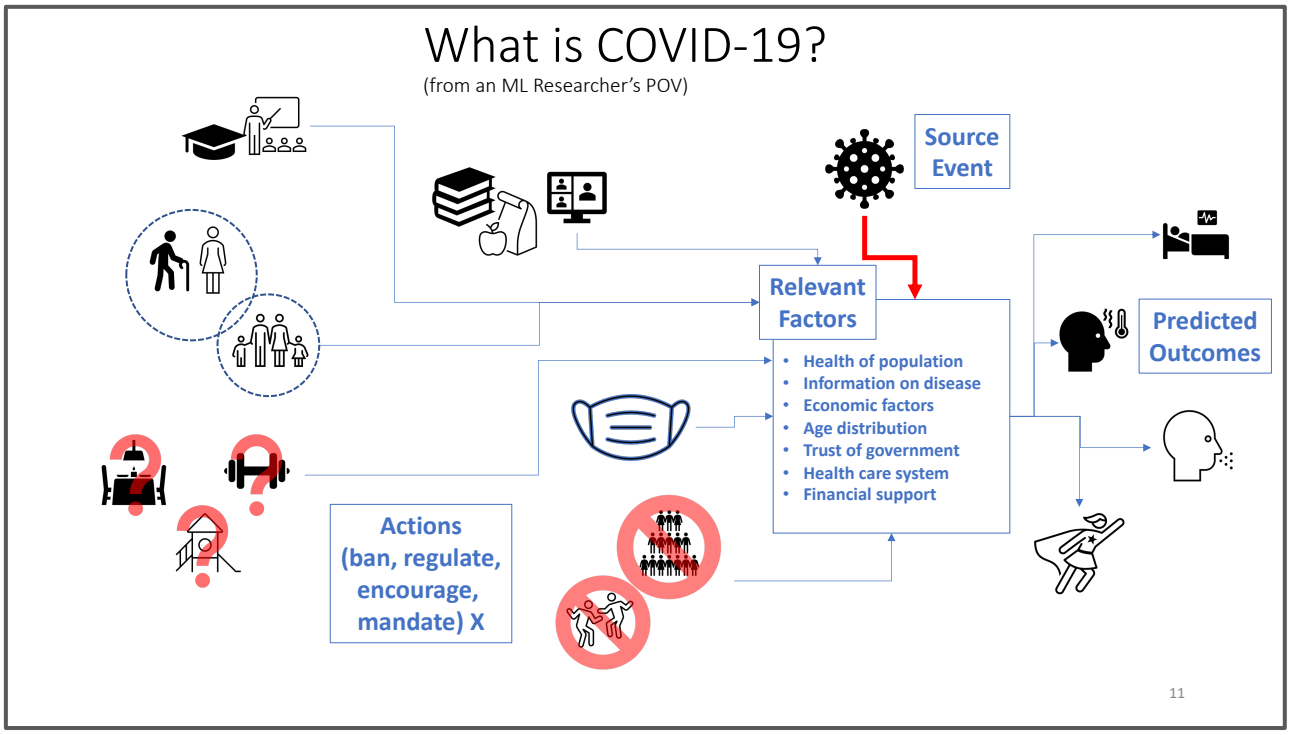

Figure 2 : A cartoon schematic of how COVID-19 can be seen as a process to analyze and learn about as a machine learning researcher.

How AI and ML are being used right now?

AI and ML should be, and are being, used for very practical, on-the-ground tasks to help with Covid-19, using approaches that don't necessarily require very complex AI/ML algorithms .

- *Spread Prediction:* Improving prediction spread models in one region given model of disease using in-region data. In our lab in summer 2020, we contributed to a model for improving prediction of ICU patient levels in Trillium/Peel region hospitals.
- *Mutation:* Analyse genetic variation patterns in the disease as it mutates.
- *Logistics:* Predict and manage shortfalls and of hospital supplies such as PPE, cleaning supplies and basic medicines.

More complex challenges and potential solutions exist than these. A particularly promising approach in the prediction/design category is *Anti-viral drug discovery* which attempt to aid the search for promising chemicals to test in various stages of drug development to speed up the process. Researchers in Mila[1], the machine learning institute in Montreal, are carrying out research on this with projects on learning relationships between drugs, chemicals, proteins on the outcome for good drugs. The main idea is to use machine learning models to propose new candidates, test them in large experimental setups, perhaps even using robotic drug labs. Data from these tests are then fed back into the ML models. Eventually, once confidence in a candidate is high enough, and after extensive human evaluation of the proposal, the decision could be made to try out clinical trials for the best candidates. In the future, this approach could become even more interactive, using machine learning methods such as Reinforcement Learning to design new molecules which could be useful drugs.

Since this topic has a focus on privacy, data and COVID, I'll focus on the possibilities for using ML to predict and understand the spread of the disease using data about people, government interventions and policies and specifically on the possibilities for **using ML to predict and understand** the spread of the disease **using data about people, government interventions and policies.**

ML's Dirty Not-So-Secret

Machine Learning is powerful, it's defeating world champions at their own games, it is being used to fly spaceships, drive cars. It is also optimizing addictive click-ability of social media apps and advertisements *really, really well*. ML is even recently making progress on solving the mind-

[1] https://mila.quebec/en/covid-19/

bogglingly complex challenge of protein folding! But…

What people don't say out loud as often, is that all of these wonders come almost all entirely from *associations*! Recall the scientist's primary warning:

> *"Correlation does not imply causation"*

Well, that applies to 90% of Machine Learning, including the recent darling success of ML, Deep Learning. It turns out that with enough data, and enough free parameters, correlation, or association if you prefer, is usually very learnable. The question is, what if we wanted more?

Multiple Routes to a Causal Model

Causality is at the very core of the scientific pursuit, so it's not surprising there are many different ways of approaching it in a reliable, and repeatable way.

Significance Tests / Hypothesis Testing

This is the bread and butter of much of science, but it's also highly restrictive. These types of approaches are often referred to as "Controlled Experiments". This approach is very widespread and standard. However, its results can be very easily misunderstood, or worse, purposely misused.

It is a good approach for direct and ccritical causal questions such as, in this case, "Is this vaccine effective at reducing severe outcomes?" or "Do masks make a significant difference in spread rate?"

The inflexibility of these methods makes them less useful for answering the following kinds of questions where we cannot run a controlled trial for practical or ethical reasons:

- Should we close restaurants and gyms?
- Or should we have closed schools instead?
- Or maybe we should have kept them all closed longer in the first place?

Tip-toeing Towards Causality

A more direct, interventional approach, is used in vaccine trials under the title of "Human Challenge Studies"[2] where healthy volunteers are given an experimental vaccine under review and then, once the immune response is expected to be high, such as a month later, to *purposely be infected* with active virus to see the effect. This is done in as controlled a manner as possible, but it is indeed a case of making your patient sick in order to test out your cure. From the perspective of causal, scientific practice, this is really the gold standard for determining the causal effect of some action. In physics and chemistry it is the *only way* to determine causal relationships. In medical problems, we almost always need to find ways to *infer causality* as well as possible using less direct methods.

A Note on the SIR Model

The classic approaches for epidemiological study and practice are models such as the *susceptible-infected-removed (SIR)* model for spread of a virus in a community. These are ODE models that capture the essential, *statistical* behaviour of a virus and how it spreads across the population.

> The Kermack-McKendrick model is an SIR model for the number of people infected with a contagious illness in a closed population over time. It was proposed to explain the rapid rise and fall in the number of infected patients observed in epidemics such as the plague (London 1665-1666, Bombay 1906) and cholera (London 1865). It assumes that the population size is fixed (i.e., no births, deaths due to disease, or deaths by natural causes), incubation period of the infectious agent is instantaneous, and duration of infectivity is same as length of the disease. It also assumes a completely homogeneous population with no age, spatial, or social structure. - Wolfram Mathworld[3]

It makes many simplifying assumptions about the world. And that is necessary for any model. But at what cost? My emphasis here on *statistical* is in contrast to *conditional* or *individual* which would be some approach that looks at the actual known effects of individuals, when available, and makes inferences.

---

[2] This article has a description of a detailed human challenge trial carried out in the UK for COVID-19, https://www.theguardian.com/world/2020/sep/24/uk-covid-19-vaccine-trial-set-to-infect-healthy-volunteers-with-virus.
[3] *https://mathworld.wolfram.com/Kermack-McKendrickModel.html*

Now, these ODEs are inherently already causal. They define how we understand disease to spread from in a population and how distance, risk profiles viral load and recovery/death rates. However, these are also *very simple* models, they cannot **learn** what these relationships should be from data. They do not **include** specific information which we have become so used to discussing now such as the impact of gyms vs. restaurants vs. schools, or the impact of various mask wearing policies

There are many other, more advanced forms of SIR, for example SEIR adds an "Exposed" step to the model[4]. SEIR model uses information such as birth and death rate, specific fatality rate caused by the virus directly or indirectly. The probability of transition for every contact between two people, rates of disease incubation progress and recovery for individuals.

The Causal Hierarchy

A different approach to understanding causal relationships comes from a giant of Artificial Intelligence and Machine Learning, Judea Pearl. Judea Pearl describes **the causal hierarchy** as a general relationship between *data, labels, knowledge and causality* (Pearl, 2009).
The hierarchy has three levels*:*

I. Association – passive observation
II. Intervention – active experiments, interventions
III. Counterfactuals – retrospective analysis, what-if scenarios

These three types of relationships are described in the Figures 3 and 4 with the mathematical notation Pearl uses which connect causality directly to conditional probability theory. Pearl was a central figure in the invention of *probabilistic graphical models* including the widely used Bayesian Network (Pearl, 1988). Even in that seminal paper, he put forward the importance of these models not only prediction and interpretation but also for inferring causal relationships in a system that might not otherwise be discernible. The "ball and stick" graphs in Figures 3 and 4 show one way to represent the three levels of the causal hierarchy as small graphical models. There are also some small examples to demonstrate the differences. Suffice it to say that Level II

[4] https://www.frontiersin.org/articles/10.3389/fpubh.2020.00230/full

reasoning is more difficult to do thank Level I, while Level III is often difficult for people to intuitively grasp, let alone build automated models to do this reasoning on our behalf.

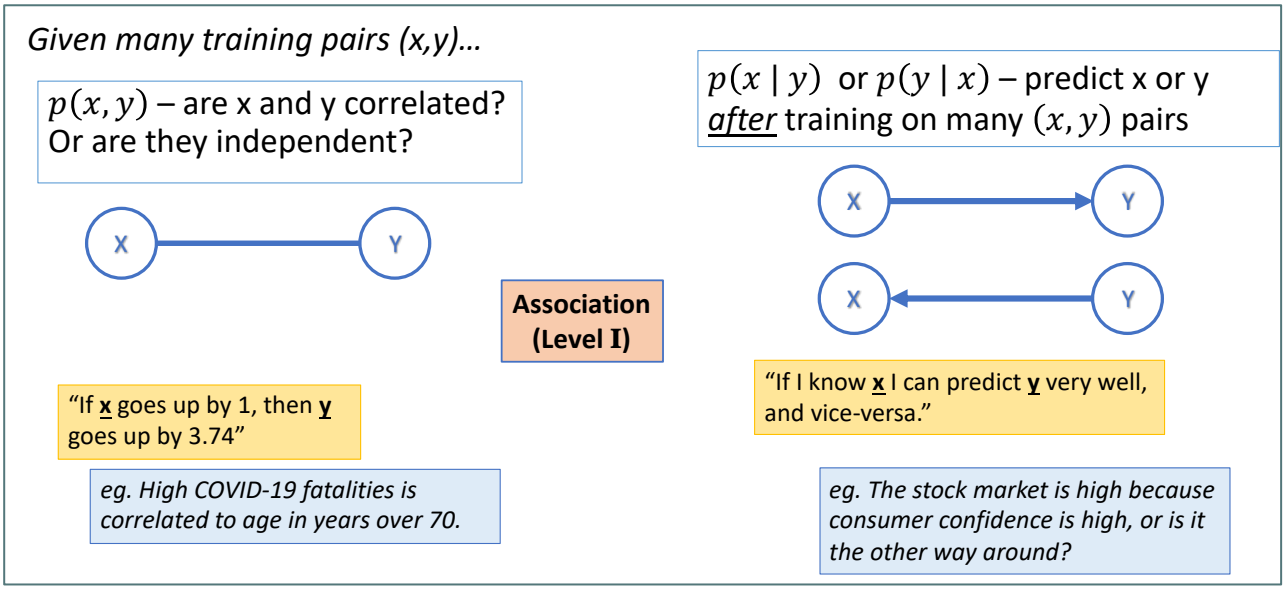

Figure 3: Visual depiction of Level I of the Causal Hierarchy as graphical models. Most Machine Learning results are restricted to this type.

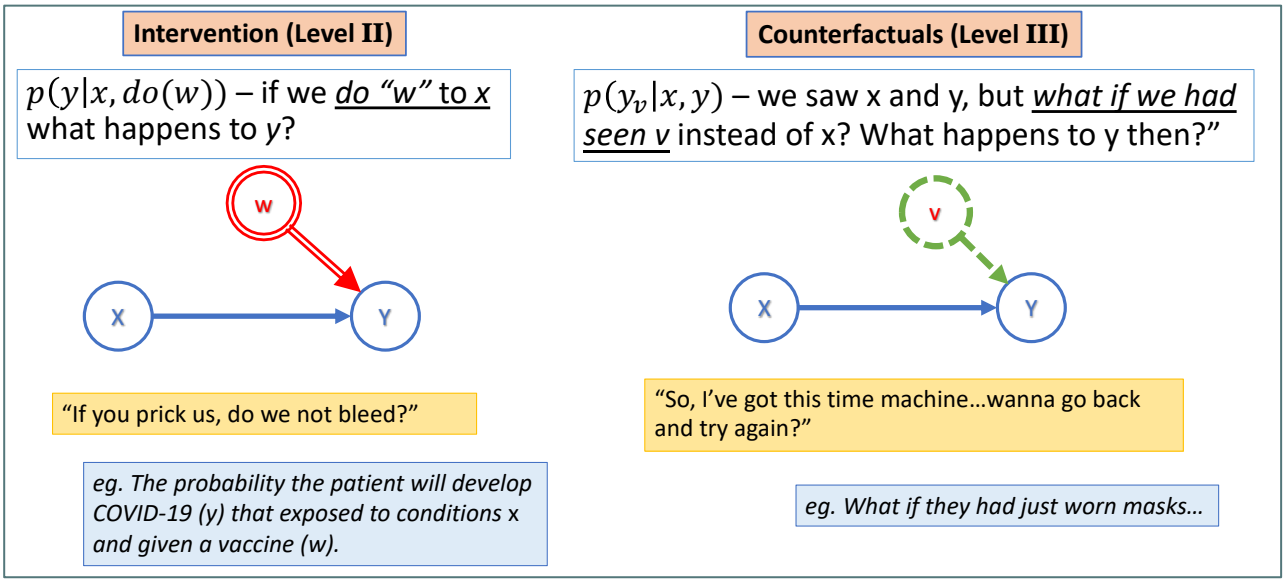

Figure 4: Visual depiction of Levels II and III of the Causal Hierarchy as graphical models. Some Machine Learning result in Decision Making and Reinforcement Learning fall in to Level II, very few methods deal with Level III.

Promising Challenges

Causal Inference methods are *fairly new* but they are being used in fields where causal answers are critical, such as medicine, infectious disease control, safety critical systems. However, in Machine

Learning they are not in the “hot” category right now. Deep Learning does not easily answer these questions, or at least, researchers in Deep Learning have not focussed much on enable such abilities. COVID-19 is a perfect moment to explore this further and see what can be achieved on a problem with a lot of data and where causal inferences will actually be critical to success.

Using the Experiment We Already Have…

The global response to COVID in each country, region and city is like **a thousand different experiments** being carried out at once on the same problem. These experiments are, of course, controlled by each region, but they are not *controlled experiments* in the statistical sense. In statistics, generally in science, a controlled experiment means that you attempt to determine a *causal link* by *controlling* the input population to be "identical", or more commonly, independently sample, from a known distribution.

Note that "identical" here only means "identical for all relevant variables". So, for example, if you know that hair colour does not have any impact on the question being studied then you would not need everyone in your data to belong to the same hair-colour group.

An experiment being "controlled" also, critically, means that *interventions* are taken, which you might also call *actions*. The whole goal of the experiment is to determine if the intervention has a notable effect which cannot be explained by anything else.

We cannot just take the results from Sweden (for example, where there was no mask mandate for many months in 2020) and make inferences about a Province in Canada without doing the following:

1. Matching the factors in common in both populations.
2. Analysing the differences and adding that to our model.
3. Using a causal graph structure to work out which questions can be answered.

Let a Thousand Experiments Bloom

So, a grand challenge for Machine Learning in this situation could be put forward based on three questions:

1. Can we *collect and sensibly represent* intervention data, all actions taken against COVID-19, by all levels of government, in different countries and regions?

2. Can we *fuse* all the demographic, economic, even socio-political information we have into a single, seamless database?
3. Finally, can we then learn a *useful* model that explains the differences in outcomes automatically?

This is partly a massive exercise in data organization and extraction. Much of this is being done already, the Oxford (TODO) database being the most notable example. Doing this well will also requires being able to answer all three levels of questions in the causal hierarchy.

The Factors Factor

One of the first questions we should always ask in a data analysis problem is "What are the relevant factors to our target question?" In the case of COVID-19 the usual biological and medical factors are being well-used already. But we've all seen how other factors are hugely influential on the way the disease spreads on the ground in any given region or country.

Some of the critical, and difficult factors not always being used include:

- demographic factors
- detailed age distribution
- health care system setup
- systemic racism, history of trust/distrust of government by segments of the population
- size of minority groups
- local temperature, season, weather
- wealth disparity in general
- group-specific education levels
- societal wealth gap
- dominance of anti-establishment feelings, resistance to strict central rules
- anti-vaccine culture
- religious control of state?

These are often difficult factors to define and measure. But they are real, societal factors that could explain why the same control policies don't have the same effect in different regions. There's no reason these factors can't be modelled, but it requires sociological expertise and causal inference to use this information fully.

Next Steps

It is clear there are many factors which could influence spread reduction. The most complex epidemiological models are still much simpler than the reality, as they should be! This is the only way to hope to understand the individual factors and to see what kind of data is missing. However, the world is currently undertaking a program of massively parallelized, unintentionally varied, experiments in every country and region of the world as we try to tackle the effects of this deadly disease. This offers us an opportunity for learning more, if we can use it right. Of course, the opportunity, and challenge, doesn't stop there either. Beyond modelling of virus spread, and

reduction strategies we can use the same ideas to consider the possible influence factors on vaccine effectiveness itself. This will include Medical and Epidemiological factors such as the particular mutation, the viral load in the population, the susceptibility to infection and serious outcomes of the population. Secondly, logistical factors will come into play which apply equally at national, regional, and local levels. This includes political will to distribute vaccines in the most logical way, and the ability to obtain, or produce, enough vaccine and distribute it. Thirdly, societal factors will apply, as they have with the strategies to mitigate spread of the disease in the first place. Some members of the public may not be willing to take the vaccine, or may be difficult to reach or inform about the proper ways to remain healthy until the vaccine deployment reaches where widespread immunity can actually impact the risks of daily life.

Conclusion

The main point of this article was to highlight how Machine Learning is being used and potentially could be used to do some small part to get humanity through this pandemic. ML is not a silver bullet to get better predictions or policies. For ML to be useful the *right questions* need to be asked and the *data being used needs to align* with those questions in some way. In other words, it needs to be reasonable that *the answer* being sought, it present, if very well hidden, deep within that data. No algorithm that produce information from nothing. Finally, it is important to remember, in this era of extreme hype in every area of human endeavour, that any analytical method or algorithm, no matter how advanced, is still usually only good at very specific types of questions. And for machine learning most of those questions are associational in nature, not causal.

References

TODO create proper references.

Old Text to Dump

Various Thoughts on Performing Causal Inference in Practice

The “Causal Hierarchy Theorem” states that in most cases, in order to answer questions about some level in the causal hierarchy, you need to already have a model for the lower level.

TODO: simplify these, drop if you don't need it.

- Learn Causal direction from data
    - functional decomposition – compare to Gaussian
    - Natural shocks as interventions
- Propensity Score Matching
    - Finding subpopulations with similar features but some intervention changes
    - Using the expected outcome vs the actual outcome to make a case for a causal relationship among factor
- Any hope for counterfactuals (III)?
    - One step at a time, a full (II) model is needed first.
- Causal graphs and Structural Equation Models allow us to
    - work out the right way to calculate probabilities
    - whether it is possible at all with the given information
- But you really need to
    - get *all the data in one place* and
    - know all these things about it
- This makes privacy a challenge…