Chapter 1

Introduction

1.1 Motivation

Many of the most important questions in science and in our personal lives are about the outcomes of doing something. Will asking people to pay upfront at the doctors reduce long term health expenditure? If we developed a drug to suppress particular genes, could we cure MS and would delaying teen-aged pregnancies improve the outcome for their kids.

These are hard questions because they require more than identifying a pattern in data. Correlation is not causation. Causal inference has proven so difficult that there is barely any consensus on even enduring questions like the returns to education or the long-term consequences of early life events – like teenage pregnancy, where the variables involved are susceptible to human intuition and understanding.

We now live in a world of data. Hours of our lives are spent online, where every click can be recorded, tiny computers and sensors are cheap enough to incorporate into everything and the US Institute of Health is considering if all infants should be genetically sequenced at birth. Such data gives us a window into many aspects of our lives at an unprecedented scale and detail but it is messy, complicated and often generated as a by product of some other purpose. It does not come from the controlled world of a randomised experiment.

The rise of big data sets and powerful computers has seen an explosion in the application of machine learning. From health care, to entertainment and self driving cars; machine learning algorithms will transform many industries. It has been suggested that the impressive ability of statistical machine learning to detect complex patterns in huge data sets heralds the end of theory [10] and that we may be only a short step from "The Singularity", where artificial intelligence exceeds our own and then grows exponentially.

However, despite the huge advances in machine learning (in particular deep learning), machine learning algorithms are effective only within narrow problem settings. Getting them to generalise to even slightly different problems or data sets remains very challenging. Deciding how we should act or what policies we should implement requires us to be able to predict how a system will behave if we change it. The correlations detected by standard machine learning algorithms do not enable us to do this, no matter how many petabytes of data they are based on. As machine learning algorithms are incorporated into more and more of the decision making processes that shape the world we live it, it is critical to ensure that we understand the distinction between causality and prediction and that we develop techniques for learning how to act that are as effective as those we have for pattern recognition.

1.2 What is causality?

The notion of causality has been widely debated in science and philosophy [84, 117, 129, 125, 108, 182, 74, 38] but is still widely viewed as poorly defined. This has led to a reluctance among applied researchers in many fields to refer to causality in their work, leading them instead to report that variables are *related*, *correlated* or *associated*. However, the magnitude, direction and even existence of an association depends on what other variables we adjust for (or include in a regression). Avoiding formalising causation, the real question of interest, leaves it up to the reader to determine via "common sense" the implications of the reported associations.

There are two ways in which an association detected in a dataset may be non-causal. The first is that the variables concerned may not be related at all, and the association has arisen by chance in that data sample. Given a finite data on enough variables, there is a high probability of finding some that appear correlated even though they are completely unrelated. For example, based on data from the years 1999 to 2009, the age of miss America is strongly correlated with the number of murders (in the US) by steam, hot vapours and hot objects [177]. We would not expect this relationship to hold if we were to look at a new sample of data. This form of spurious correlation also has serious repercussions. It lies at the heart of major problems with the process of scientific research: researchers are incentivised to detect effects and thus to explore many possible paths in the process of analysing data and studies that fail to find an effect are less likely to be published. As a consequnce, the likelyhood that reported effects have arisen by chance is underestimated, leading to the conclusion that "most published scientific results are false" [88]. This issue is also highlighted by recent crises in replication [120]. This issue can be ameliorated by getting more data and by seperating learning models from evaluating their performance, for example by evaluating models on a stict hold-out set or on the extent to which their results can be replicated.

However, a strong association, observed in multiple indpendent studies may still not be causal. The correlation can arise because both variables are consequences of some other, unmeasured factor. For example, the reading ability of children under twelve is strongy correlated with their height, because older children are taller and can read better. However height is not a cause of reading ability because if we were to intervene to increase a child's height, for example by giving them growth hormones, we would not expect dramatic improvements in their reading. Similarly, extra tutoring in reading will not make a child grow taller. This problem is fundamentally different to the issue of spurious correlations arising by chance in finite datasets. Obtaining more (even infinitely many more) samples without directly intervening in the system to manipulate the variables does not allow us to seperate causation from correlation.

The key distinction between a real, but non-causal, association and a causal relationship is in what happens if we intervene in the system and change one of the variables. In this thesis, I take an interventionalist viewpoint of causality: any model or approach designed to predict the outcome of intervening in a system is causal. This viewpoint captures the types of questions that motivate this thesis. How can we change the way we do things to obtain better outcomes.

Causality is often liked to explaination; understanding how and why things happen. I view explaination in terms of compression and generalisabilty: how much information about the world a model can capture. This creates a hierachy in the degree to which models are explainitory, rather than a simple binary distinction. A standard predictive model encodes all the information we need to predict some output given inputs provided the system generating the data does not change. A high-level causal model might be able to predict the outcome of a specific intervention holding all else fixed. More detailed causal models could predict the outcome for a wide range of combiations of interventions conditional on a range of contexts. By considering conditional interventions within our definition of causal questions we also capture mediation: the study of

through which pathways one variable causes another [174]. Finally, a model that can distil how elements interact to mathematical equatons like Newton's laws can be used to predict what will happen in an astounding range of settings, including many we have never previously observed [1].

Gelman [67], Gelman and Imbens [68] make a distinction between forward causal inference, the types of "What if" questions I focus on in this thesis, and reverse causal questions, asking why something occurs. The former aims to identify the effect of a known cause. The latter can be viewed as identiying causes of an effect. They regard forward causal inference as well defined within the counterfactual and graphical model frameworks for causal inference, which we describe in chapter 2. However, state that "A reverse causal question does not in general have a well-defined answer, even in a setting where all possible data are made available." I view this as overly pessimistic, depending on how "all possible data" is defined. The goal of identifying the causes of an effect can be formalised within the graphical causal model framework. Solving this problem is certainly much more challanging than identifying the effect of a specific intervention on given outcome, since it requires us to test or infer the effect of interventions on many different variables. These practical difficulties may well be overwealming, particularly in fields such as social science and economics where datasets are relatively small, systems are complex, the variables are difficult to directly manipulate and even relatively simple "What if" questions are hard to resolve conclusively. However, this does not mean that the problem of identifying causes of effects is ill-posed in principle. It can be viewed as a form of causal discovery: the attempt to learn the structure of the causal relationships between variables, on which there is a rich liturature which we review briefly in section 3.4.

There has traditionally been a large gap between researches in machine learning who focus on prediction, use largly non-interpretable models and researches in statistics, social science and economics who (at least implicitly) aim to answer causal questions and tend to use heavily theory driven models. However, there is relatively little awareness, particularly within the machine learning and data science communities, of what constitutes a causal problem and the implications of for the training and evaluation models. In the next section we emphasise the subtley that can exist in determining if a problem is causal by examiming some typical examples.

1.3 What makes a problem causal?

Machine learning is in the midst of a boom, driven by the availability of large datasets and the computation resources to process them. Machine learning techniques are being applied a huge range of problems, in both industry and achademia. The following examles are inteded to capture the bredth of problems that machine learning algorithms are actively beeing applied to. Which, if any, of these problems require causal inference?

- Speech recognition (for systems like Siri or Google)
- Image classification
- Forecasting the weather
- Identifying spam emails
- Automated essay marking
- Predicting the risk of death in patients with pneumonia.
- Predicting who will re-offend on release from prison

[1] Although Newton's laws are not fully general, as they break down for extremely massive objects

- Customer churn modelling
- Demand prediction for inventory control
- Playing Go

The question is disingenuous because I have not posed the problems in sufficient detail to determine if causality is an important considertion. In particular, I failed to specify how any model we might build would be used: what actions would be taken in response to its predictions.

Consider speech recognition. You say something, which causes sound waves, which are converted converted to a digital signal that Siri maps to words. Whatever action Siri takes is unlikely to change the distribution of words you use, and even less likely to change the function that maps sound waves to text (unless she sends you a DVD on elocution). A similar argument could be made for many applications of machine translation and image classification.

In image classification we do not particularly care about building a strong model for exactly how the thing that was photographed translates to an array of pixels, provided we can be fairly confident that the process will not change. If we develop a discriminative model that is highly accurate at classifying cats from dogs, we do not need to understand its internal workings (assuming we have strong grounds to believe that the situations in which we will be using our model will match those under which it was trained).

What about forecasting the weather? If you are using a short term forecast to decide whether to pack an umbrella causality can be ignored - your decision will not effect if it actually rains. However, longer term climate forecasts might (theoretically) lead us to take action on emissions which would then change the weather system. For this we need a (causal) model that allows us to predict the outcome under various different interventions.

Identifying spam and automated essay marking both involve processing text to determine an underlying (complex) attribute such as its topic or quality. In both cases, there is inherent competition between the algorithm and the people generating the text. As a result, decisions made by the algorithm are likely to change the relationship between the features it relies on and the true label. Spammers and students will modify their writing in order to optimise their results. A standard supervised learning approach can only work if the resulting change in the mapping from features to label is sufficiently gradual. There are two key ways of ensuring this. The first is to limit people's ability to observe (and thus react to) decisions made by the algorithm. The second is to use a model in which the features are related to the outcome in such a way that they cannot be manipulated independently.

This example also highlights a connection between causal models and transparancy in machine learning. If we are using a non-causal model to make decision effecting people, there will be a trade-off between the performance and transparancy of the model; not because the requirement for transparancy restricts us to simple models but because revealing how the model works allows people to change their behaviour to game it.

What about predicting the risk of death in patients with pneumonia? Suppose the goal is to build a model to decide who should be treated in hospital and who can be sent home with antibiotics. If we assume that in hospital treatment is more effective for serious cases, this seems like a straightforward prediction problem. It is not. Depending on how the decision to admit was previously made and what features are included (or omitted) in the model, the relationship between those features and the outcome may change if the model is used to make admition decisons. Caruana et al. [39] found exactly this effect in a real data set. The model learnt that people suffering asthma were *less* likely to die from pneumonia. They realised this was because doctors were treating such patients very aggressively, thus actually lowering their risk. The issue is not with model, it performed very well at the task for which it was trained, which is to predict

wo would be likely to die under the original admition and treatment protocols. However, using it to decide how to *change* these protocols could kill. The actual question of interest in this case is what happens to patients with charachteristics X if we assign them treatment according to decision rule Z.

Predicting which customers will leave or who will re-offend if granted parole also fit within the category of problems where you wish to identify a group for which a problem will occur and target some treatment to them (hospitalisation, loyalty reward, more monitoring & support on parole, etc). Predictive models can be applied to such problems provided it is known which treatment option will be most effective for the target group and deciding who to treat on the based of the model predictions won't change the relationship between features and outcome.

Demand prediction seems like a relatively straightforward prediction problem. These models use features such as location, pricing, marketing, time of year, weather, etc to forecast the demand for a product. It seems unlikely that using the model to ensure stock is available will itself change demand. However, depending on the way demand is measured, there is a potential data censoring issue. If demand is modelled by the number of sales, then if a product is out of stock demand will appear to be zero. Changing availability does then change demand.

Playing Go (and other games) is a case with some subtleties. At every turn, the AI agent has a number of actions available. The board state following each action is deterministic and given by the rules of the game. The agent can apply supervised machine learning, based on millions of previous games, to estimate the probability that each of these reachable board states will lead to a win [2]. Supervised learning can also be applied to learn a policy $\mathrm{P}(a|s)$, the probability of selecting action a given board state s. This allows the agent to estimate the likelyhood of winning from a given starting state by simulating (many times) the remainder of the game, drawing actions from $\mathrm{P}(a|s)$ for both players. Google's Alpha Go, which in May 2017 beat the then strongest human player [118], incorporates a compbination of these approaches [155]. The supervised learning was enhanced by having the agent play (varients) of itself many times, so that its estimates of value for each board state and of the likelyhood an opponent will play as given move are based on a combination of replicating the way humans play and on the moves that led to a win when playing itself.

The problem of playing go is causal from the interventionalist perspective. The agent wishes to learn the probability of a win given an action they take. However, there are some special characteristics to the go problem that make it amenable to a primarily supervised learning approach. The actions the agent has to explore are the sames ones as human players explored to generate the trainging data, and both have the same objective - to win the game. In addition, the state of the board encapsulates all the information relevent to selecting a move. These factors make it reasonable to conclude that selecting moves with an algorithm will not change the value of a board state or the probability of given move by the opponent by a sufficiently large margin to invalidate the training data.

Having considered these examples we can now identify some general aspects of problems that require causal (as opposed to purely predicive) inference. A predictive model may be sufficient if, given the variable(s) being predicted, it is clear which action is optimal and if selecting actions on the basis of the model does not change the mapping from features to outcomes. The second requirement is particularly difficult to satisfy when an algorithm is making important decisions effecting individual people. Think about problems like credit scoring and parole decisons. There are strong ethical grounds for demanding transparancy, but if the goals of society and the individuals are not perfectly aligned and there is any posibility that people can manip-

[2] This is a challenging pattern recognition problem. There are around 2×10^{170} legal board positions in Go, ([171]), so the algorithm cannot simply memorise the proportion of times each state leads to a win, it must identify higher level features of the board state that are associated with winning.

ulate features independtly of the outcome, there will be a conflict between model accuracy and transparancy. It is rare to build a model without any intent to make some kind of decion on the basis of its results. Thus, I argue we should assume a causal model is required until we can justify otherwise.

1.4 An overview of this thesis and its contributions

Causal questions are everywhere. Statistics, economics, social science, artificial intelligence and machine learning. Techniques and for addressing such problems have developed in parralel within many diciplines. These techniques can usefully be categorised into two broad approaches, reinforcement learning and observational causal inference.

In reinforcement learning, under which we include traditional randomised experiments, we learn the outcome of actions by taking them. We take the role of an agent, capable of intervening in the system, and aim to develop algorithms that allow the agent to select actions optimally with respect to some goal. A particular strand of research within reinforcement learning are multi-armed bandit problems. They describe settings in which there are a set of available actions, the agent repeately decides which to select and then observes the outcome of the chosen action. They capture problems such as a doctor deciding which treatment to prescribe to a patient or a search engine selecting which advertisement to display to a user, where the agent faces the same set of choices repeatedly and is able to assess the value of outcome of the actions.

The approach of learning the outcome of an action by taking it has plays a key role in advancing our knoweledge of the world. However, we frequently have access to large bodies of data that have been collected from a system in which we did not have any control over what actions were taken, or perfect knowledge of the basis on which those actions were chosen. Estimating the affect of an action from such observational data sets is the problem addressed by observational causal inference.

Observational causal inference can be viewed as a form of transfer learning. The goal is to leverage data obtained from one system, the system we have observed, to estimate key characteristics of another, the system after we select an action via some policy that may differ from the process driving which actions occur in the orignial system. This is impossible without some assumptions about how the the orignal system and the system after intervention are related to one-another. The key to observational inference is a framework that models how actions change the state of the world and thus allows us to map information collected in one setting to another.

Both multi-armed bandits and observational causal inference can be seen as extensions to the concept of randomised controlled trials. Bandit algorithms deal with the sequential nature of the decision making process, causal inference with the problem randomisation is not always feasible, affordable or ethical. The similarities between the problems that these techniques have been developed to address raises the question of if there are problems best addressed by a combination of these approaches and how they can be combined.

The goals of this thesis are: to clarify the distinction between prediction and causal inference and highlight when and how we can leverage advances in prediction to improve causal inference and to draw together the key approaches to solving causal problems, from both the interventionalist and observational viewpoints. In chapter two, I describe the existing frameworks for formalising causality developed withing economics, statistics and machine learning and discuss how they relate to one another. Chapter three reviews the key approaches to causal inference from observational data. Chapter four highlights the interventionalist viewpoint, covering traditional randomised experiments and the closely related multi-armed bandit problems and highlights

connections between the algorithms applied to bandit problems and those for estimating causal effects from observational data.

In chapter five, I develop a framework that unifies the causal graphical model approach for inference in observational settings with the sequential experimental approach encapsulated by multi-armed bandits. The causal models allow us to represent our knowledge of how variables are related to one-another in a very natural way and induces an interesting and novel form of structure between the different actions modelled in the bandit problem. I develop a new algorithm that can exploit this structure as a first step towards a unified approach to decision making under uncertainty.

Chapters two to four predominantly review the existing liturature with the goal of making connections between the different viewpoints and approaches. They do not contain novel technical results but present the material in a unified way. Chapter five is a novel contribution. It formally connect causal graphical models and multi-armed bandit problems, demonstrates this leads to a form of structure that had not previously been analysed within the bandit liturature and describes an algorithm that can leverage this structure to make better decisons.