0.1 Literature review

Within data social science there exist widely differing definitions of big data and machine learning. This section seeks to clarify the use of these terms within the scope of this paper.

0.1.1 On big data & machine learning

Historically, big data has been a term reserved for data that was unable to be processed by extant software. However, recent increases in computational power has enabled data processing of hitherto unheard of quantities of data (Lazer & Radford 2017). Today, big data is no longer too cumbersome to analyse. Matthew Salganik (2018) instead mentions ten typical characteristics of big data. Salganik's tentative definition suggests that big data is not as single entity, but includes many different types of systems. Among the most important features of Salganik's definition of big data is that the data has a high frequency of observations and is continuously being generated. Since the data is always-on it is also drifting, meaning that the structure and population it represents is ever-changing. It is therefore important to understand that big data is not a naturally occurring system, but driven by the engineered purpose of the system. This algorithmic confounding forces the scientist to be careful regarding any observed human behaviour that is extracted from a single digital system.

In the analysis of big data it is often useful to employ machine learning as it can identify and predict various nonlinear relationships in big data sets, that otherwise would remain hidden. Data scientists' have for the most part optimised the predictive capabilities of the algorithm applied, and as a consequence they have often ignored or trivialised machine learning's potential in causal modelling (Varian 2014).

The technical and theoretical challenges faced by big data and machine learning research are important to consider, when employing the tools they provide. One of the major discussions revolve around machine learning's predictive capabilities. Chris Anderson (2008) argues that since the computing power and the scale of data has increased exponentially, our reliance of scientific models could become obsolete. Instead of focusing on the theoretical implications of observations, scientists should, according to Anderson, focus on the statistical outputs: In the age of big data, correlation perhaps should supersede causality and consequently social data scientists should not try to develop coherent models or unified theories to explain a social phenomenon.

However, many social data scientists argue against this point of view. Justin Grimmer (2015) claims, that correlations extracted from big data cannot stand alone. Large quantities of data is not sufficient to make scientifically valid causal inferences. It requires a rigorous research design and clear theoretical assumptions, in order to yield scientifically accurate estimates. Indeed social sciences greatest contribution to big data research, comes from the organized framework provided by rigorously tested theory (Einav & Levin 2014).

The contribution from the social sciences to machine learning help create new methods that will be able to utilize the strengths of machine learning to help solve causal inference problems within the framework of a well defined theory (Athey 2018). These new approaches could help define what variables to ma-

nipulate and how to properly use machine learning within the framework of theoretical assumptions. Ultimately, big data and machine learning could increase the scope of the social scientist's field, not only by delivering new data and methods, but by helping the social scientist to focus on new questions (Mullainathan & Spiess 2017)

0.1.2 Ethical concerns in the petabyte age