Wolfgang Pietsch
(Technische Universität München)

The prospects of data-intensive science for dealing with causal complexity

I analyze novel epistemic strategies that data-intensive science provides for dealing with causal complexity. To clarify the relevant concept of complexity I closely follow Sandra Mitchell's account in ‘Komplexitäten’ (2008) and link her position to a difference-making account of causality. On this basis, I argue that data-intensive methods can identify complex causal relationships illustrating this claim with several widely-used algorithms. I briefly discuss if the results of these algorithms can be interpreted as law-like.

Mitchell (2008) identifies various aspects of complexity in systems sciences like biology and sociology. Most of these characteristics can be linked with causality: (i) the causal relationships are complicated, in particular there are usually many contributing factors instead of one dominating cause; (ii) the causal dependencies are generally non-linear, sometimes also non-functional; (iii) causal interaction takes place between different levels of ontology; (iv) many causal relationships are strongly context-dependent; (v) the composition of causes does not follow simple addition laws. Furthermore, Mitchell points out several related aspects concerning evolutionary complexity, in particular historical contingency, path-dependence, and variability over time as opposed to stability. Simple modularity is in general not realized for such complex systems mainly due to context-dependency and non-additivity.

As Meinard Kuhlmann (2011) has rightly stressed, notions of complexity discussed in philosophy of science such as the causal complexity sketched by Mitchell mostly differ from the dynamical complexity that has been treated extensively in physics in recent decades, chiefly in chaos theory. Dynamical complexity arises in often quite simple systems of equations exhibiting non-linearity and feedback processes. A typical example is the logistic map that models a wide range of phenomena from population dynamics to the growth of tumors. Dynamically complex systems generally lack some of the features of causally complex systems, e.g. dependence on a large number of variables, context-dependence or non-modularity. On the other hand, causally complex systems cannot be modeled as relatively simple systems of equations and do not necessarily exhibit the complicated dynamical patterns familiar from chaos theory including exponentially diverging trajectories, dynamical attractors, bifurcation patterns dependent on a control parameter etc.

In her discussion of causal complexity, Mitchell refers to a difference-making account of causality broadly in the tradition of Mill's methods. Such an account can be linked with the basic features of causal complexity mentioned above. As stressed for example in Baumgartner & Graßhoff (2004), difference-making accounts presuppose a homogeneity assumption that causally relevant background conditions do not change during experiments or observations. This establishes the ceteris-paribus character of causal laws. Also, a difference-making account can in principle deal with arbitrary non-linear dependence as well as with non-additivity provided a sufficient amount of data is available. Finally, difference-making is flexible enough to allow for interlevel-causation as it does not require a preferred formulation in terms of fundamental properties.

Having clarified the relevant basic concepts, we can now address the advantages of data-intensive methods for dealing with causal complexity. In this, a definition of data-intensive science is employed that is relative to the complexity of the examined phenomenon and to the research question: 'In data-intensive science, sufficient data is available to represent all states of a phenomenon that are relevant to a specific research question'. The definition directly implies the hyper-inductive approach presupposing few theoretical assumptions that is typical for data-intensive science. The main reason why this mode of research is good at dealing with causal complexity is quite simple. The approach is non-reductive and its methods are able to work with a large number of variables as well as a large number of instances, i.e. observations of the considered phenomenon in different states.

I briefly discuss various algorithms to substantiate this claim. Classificatory trees and naïve Bayes classification allow handling complex causal relationships that involve a large number of relevant parameters. I will discuss under which circumstances these methods can identify causes in the sense of difference-makers and will also argue that they are more likely to identify causes, when the number of considered parameters and the number of instances increases. As another example, nearest-neighbor approaches used for example in non-parametric regression or non-parametric density estimation can deal with non-linearity and even with non-functionality.

The relationships produced by such algorithms have several properties that reflect causal complexity: (i) they potentially involve a large number of parameters and (ii) do not adhere to simple functional dependencies. Relatedly, these relationships (iii) often hold only in a few instances, i.e. their applicability is very limited. Furthermore, there is no reason why (iv) a hierarchy of increasing universality should exist into which these laws can be systematically integrated. To conclude, I will briefly discuss whether such relationships can be considered scientific laws.

References:

Baumgartner, M.; Graßhoff, G. (2004). Kausalität und kausales Schließen. Bern: Bern Studies in the History and Philosophy of Science.

Kuhlmann, M. (2011). Mechanisms in dynamically complex systems. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), Causality in the Sciences (pp. 880-906). Oxford: OUP.

Mitchell, S. (2008). Komplexitäten − Warum wir erst anfangen, die Welt zu verstehen. Frankfurt a.M.: Suhrkamp.