Big data and prediction: four case studies

Robert Northcott, Birkbeck College

Abstract

Has the rise of data-intensive science, or 'big data', revolutionized our ability to predict? Does it imply a new priority for prediction over causal understanding, and a diminished role for theory and human experts? I examine four important cases where prediction is desirable: political elections, the weather, GDP, and the results of interventions suggested by economic experiments. These cases suggest caution. Although big data methods are indeed very useful sometimes, in this paper's cases they improve predictions either limitedly or not at all, and their prospects of doing so in the future are limited too.

Keywords

Big data; prediction; case studies; explanation; elections; weather

1. Introduction: prediction, big data, and case studies

Accurate prediction has long been possible in the laboratory and within engineered artefacts. But in unshielded field contexts it has usually been thought difficult, if not impossible, because it requires taking account of every relevant factor. Usually, the over-abundance of such factors makes accurate prediction infeasible. Moreover, many of these factors will likely be transient or sui generis and thus difficult to capture for theories or causal models, which by their nature tend to focus instead on factors common to many contexts. In reaction, most field sciences have therefore concentrated not on prediction but instead on the development of a

repertoire of theories, models and mechanisms. These, it is hoped, can provide explanation and understanding even in the absence of accurate prediction.

This methodology is dominant in economics, and increasingly so in ecology, political science, sociology, and many other fields. But big data advocates challenge it. In particular, both the amount of data available and our ability to analyze it have increased enormously in recent years. As a result, accurate prediction of many field phenomena has become possible for the first time. Notable examples include the discovery of the CRISPR technology for genome editing in living eukaryotic cells (Lander 2016), how to get the cheapest airline tickets, Amazon's personalized suggestions of new purchases, and prediction of which manhole covers will blow or which rent-controlled apartments will have fires in New York City (Mayer-Schoenberger and Cukier 2013) and Facebook and Google's experiments regarding page design and marketing. New analytical techniques include various forms of machine learning and algorithmic methods. Neural nets, for example, are behind rapid recent advances in natural-language translation (Lewis-Kraus 2016).

The stakes are high. If big data revolutionizes our ability to predict, it is claimed, then, that this should lead to a transformation: a new priority for prediction over explanation or causal understanding. Because the new predictive successes have usually come via algorithmic or black-box approaches that preclude theoretical interpretation, the traditional emphasis on theory impedes progress in prediction and, accordingly, should be abandoned. The most eye-catching versions of this argument have heralded the 'death of theory' altogether (Anderson 2008) and a new paradigm for scientific method (Hey et al. 2009). The sheer number of successful new predictions, it is claimed, makes the case for a huge methodological re-orientation (Mayer-Schoenberger and Cukier 2013). Does it?

Two clarifications: first, 'big data' is a vague phrase. Some interpret it narrowly to refer only to particular machine learning techniques. I will interpret it more broadly, in the spirit of the American Association of Public Opinion Research's definition: 'an imprecise description of a range of rich and complicated set of characteristics, practices, techniques, ethical issues, and outcomes all associated with data' (Japec et al. 2015, 840). A broad interpretation gives 'big data' every chance to prove its worth. My eventual conclusion, that its prospects (in one respect) are limited, is, then, stronger. Although a huge range of techniques fall under big data so understood, these techniques have sufficient core features in common that we may usefully assess their impact and prospects as a group.

Second, big data has chalked up many impressive applications in field science already, with the likelihood of many more in the future (Japec et al. 2015, Foster et al. 2017). My focus, however, will just be on better prediction, and thus any big data method relevant to that, such as predictive analytics.[1]

The heart of this paper will be four case studies of prediction of field phenomena, namely political elections, weather, GDP, and economic auctions. All of these cases are well understood by philosophers of science, having been closely studied by them for other purposes. Why case studies? There exist general analyses already, by philosophers of science and others, of what conditions are necessary for big data predictive methods to succeed (section 6). Case studies serve to stress-test such analyses against practical realities: when are the necessary conditions satisfied? When they are satisfied, are they sufficient? There is no

[1] I therefore will not discuss the many important political and ethical issues raised by big data methods.

substitute for local detail. This also enables us to assess better the role that is left for theory, and whether there really is no hope for causal understanding.

General analyses implicitly promise that they will shed light on big data's prospects in pressing actual cases. This paper proceeds the other way around, so to speak, by starting with pressing actual cases and then examining *in those cases* how effective big data methods actually are. Informed by this dive into particularity, we may generalize out again to get a better sense of big data's prospects more widely.

Big data advocates have naturally cited various success stories, but are those stories representative? In this paper's cases, there is no presumption as to the potential efficacy or otherwise of big data methods. They are therefore neutral tests in this regard. The first three of them – elections, weather and GDP – are of interest because of their independent importance. The fourth – economic auctions – is of interest because it is an example from social science of successful field prediction based on the extrapolation of results from the laboratory. This is an influential method that may become much more widespread.

Overall, the paper's thesis is 'generalist' in that, as the case studies show, the same factors are positives and negatives for big data's prospects across contexts. It is also generalist in that these factors tend to be positives and negatives for all big data methods alike. Which of these factors are actually pertinent in any given case, though, varies case by case. Accordingly, the prospects for big data also vary case by case. After going through each case study in turn (sections 2 to 5), the paper summarizes big data's prospects in each (section 6), before assessing more generally big data's promise for prediction and causal understanding (sections 6 to 8).

2. First example: Political elections

There are several approaches to predicting the results of political elections.[2] By far the most successful is opinion polling.[3] Polls use the voting intentions of an interviewed sample to serve as a proxy for those of a population. How might things go wrong? The most familiar way is sampling error: small samples can be unrepresentative flukes. But sampling error is not the only, nor even the most important, source of inaccurate predictions.[4] A far bigger issue for pollsters is to ensure that their samples are appropriately balanced. Results will be biased if a sample is unrepresentative of the voting population with respect to, for instance, age, gender, race, or income, since these factors correlate with voting preference. This is quite different from sampling error: if a sampling procedure over-selects for men, say, then that cannot be alleviated just by making the sample bigger. Pollsters must decide exactly which factors to allow for. Should they rebalance, for instance, for declared political affiliation or for degree of interest in politics? Mistaken treatment of these latter factors has been the source of errors in recent US and UK election polling. Further decisions are necessary too: how hard and in what way to push initially undecided respondents for their opinions; how hard and in what way to pursue respondents who decline to participate; whether to sample face-to-face or by phone or online, and (in the latter cases) whether to

[2] See (Northcott 2015) for more details and references regarding this case.

[3] This is true even though polls are not perfectly reliable. The main alternative is to predict on the basis of 'fundamental' variables that recur from election to election, most commonly macroeconomic ones such as growth in GDP, employment or real wages. It is uncontroversial that polls predict better. The alleged compensating advantage of models based on fundamentals is that at least they can explain, or provide understanding of, election results whereas polling cannot. I think, in fact, that neither approach explains or provides understanding (Northcott 2015, 2017), but I will be concerned only with prediction here.

[4] Almost 25% even of late polls of US presidential elections miss the final result by more than their official 95% confidence interval, yet the expected miss rate given sampling error alone should be only 5%.

interview or to let respondents fill out answers alone; how to assess how firmly held a respondent's preference is; and how to assess the likelihood that a respondent will actually vote.[5] Exactly how pollsters tackle such issues has been shown to significantly influence the accuracy of their predictions (Sturgis et al. 2016, Wells 2018).

Separately from such 'internal' issues, the systematic *aggregation* of polls improves predictive accuracy significantly. One obvious reason is that aggregation increases effective sample size and therefore reduces sampling error. But mere aggregation is no cure for incorrect sample balancing because the optimal balancing procedure may not be the industry average. To assess the chance that all polls are systematically skewed in the same way requires sophisticated aggregation rather than taking the results of individual polls at face value. Overall, aggregation requires a second layer of method, quite distinct from that required to conduct a single poll.

What role for big data in polling? Clearly, more data has helped: polling predicts better today in part simply because there is more polling data (Arguably, there were no reliable political polls at all until after World War Two). As with weather forecasting (section 3), improved analytical methods have also helped – polling aggregation is one example.

In all of our case studies, a crucial question is: how much *could* prediction be improved by the application of big data methods in the future? What is big data's upper limit? In the case of elections, the answer, alas, is that predictive paradise will remain elusive. Let us see why.

[5] So far, it is doubtful that new-technology methods such as automated 'robo-calling' or online surveys are any better predictively than more traditional live-interviewer methods (https://fivethirtyeight.com/features/which-pollsters-to-trust-in-2018/).

Political campaigns increasingly use sophisticated big data methods. These have mainly taken the form of 'microtargeting' voters. Extensive data can now be collected about individual voters' consumption patterns, media preferences, demographic characteristics and so on, and algorithms track how these factors correlate with political preference and likeliness to vote. Obama's 2008 campaign, for instance, was tracking over 800 different voter variables as early as the Iowa caucuses in January. Campaigning material and tactics are tailored accordingly, at the level of individual voters, in order both to change voter preferences and (especially) to increase supporter turnout. Such microtargeting, which first became prominent in the Bush 2004 campaign, is an example of a 'theory-free' big data approach displacing a more traditional model-based one. Might it enable campaigns, or anyone, to predict election results better than they can with opinion polls?

The key would be to identify correlations between the effect variable, i.e. actual votes, and the putative cause variables, i.e. consumption patterns, media preferences and so on. But there is a major epistemological roadblock: the limited sample of past elections means that public results are insufficient for training predictive algorithms, yet no other voting data are available because the secret ballot means that individuals' votes are unknown.[6] This data limitation threatens all big data techniques.

An obvious response might seem just to *ask* individual voters how they voted. Some may answer falsely but, the reasoning goes, so long as most do not then we may establish correlations sufficiently well to generate accurate predictions. However, the salient comparison is whether we can predict *better* with big data methods than we already do with opinion polls. It now seems that in order to predict at all with big data methods, we must rely

[6] It is acknowledged by practitioners that machine learning requires a lot more data than are available in most cases (Foster et al. 2017, 172-3). Elections seem to be an example of this.

on asking voters for whom they voted. But pollsters do that already, so where is the comparative advantage?[7] It could only come from new factors that add predictive value over and above existing sample balancing by pollsters.[8] But this seems a dubious hope (see below).

Moreover, the value of this kind of augmented polling of voters seems inevitably limited because it does not address the biggest difficulties that polling methods actually face. It is one thing to know what a voter's political preference is; it is quite another to know whether they will actually vote. For example, polls in the 2015 UK general election were unusually inaccurate. Subsequent investigation revealed the main cause: pollsters assessed likeliness to vote by, roughly speaking, just asking voters themselves. However, errors in voters' self-assessment correlated with political preference, which led to biased samples. It would have been better to rely on historical rates of turnout for particular demographic groups (Sturgis et al. 2016). For the 2017 UK general election, therefore, most polling firms switched to this latter method. However, their predictions were again unusually inaccurate. Investigation revealed, roughly speaking, that the solution would have been to switch back to the 2015 methodology (Wells 2018). In other words, there was a reversal regarding which method was optimal.

[7] True enough, pollsters more often ask how a voter *will* vote rather than retrospectively how they *have* voted. But it is not clear that the reliability is less in the former case; indeed, it might even be greater because in the latter case respondents' answers can be sensitive to post hoc perceptions of the result, perhaps via a desire to be seen to have voted for the winning side or simply to have voted at all (Issenberg 2016, 193).

[8] It is not enough that, say, media preferences in isolation correlate with voting preference. Rather, the issue is whether balancing samples for media preferences adds predictive value over and above existing balancing for, e.g., gender and race. That is, the new variables must impact on voting independently (at least in part) of how the pollsters' existing ones do.

This example illustrates a fundamental problem – namely *non-stationarity* in the underlying causal processes (section 6). In general, such non-stationarity is a threat to all big data techniques. In this case, the causal processes that relate various demographic variables to turnout were not stable between 2015 and 2017. The problem is that non-stationarity cannot be overcome by knowledge of past correlations. The same issue arises with other electoral variables. Do, say, the percentages of blacks, women, the rich, sports fans, and so on, that vote for a particular party stay constant across elections? Historically, they often have not.[9]

One response to such non-stationarity is to stick to short-term forecasting within a single campaign, on the assumption that correlations are more likely to remain stable within this shorter timeframe. However, even within a single campaign there are many relevant non-stationarities, especially during primaries when voter preferences – and thus the correlation between them and the various predictor variables – are especially volatile. Problems can arise during general elections too, as with temporary surges of opinion after notable events. Moreover, the effectiveness of a particular campaigning tactic can fade quickly with repeated use (Issenberg 2016).

Campaigns adopt two approaches, in part to identify, and thus to counter, such non-stationarity. First, they often run daily polls to help calibrate their inferences from data regarding consumer preferences and so on. This is a sensible tactic. But the relevant point for our purposes is that the accuracy of any election predictions inferred from such daily polls still has an upper limit given by those polls' accuracy. So, again, there is no reason to expect a dramatic outperformance of regular public polls.

[9] For this reason, the non-stationarity problem also applies to countries with mandatory voting, even though the specific problem of predicting differential turnout does not.

The second approach to counter non-stationarity is to utilize a campaign's extensive *non*-polling information, namely voters' responses to doorstep, phone and other interactions. Such responses play a huge role, in conjunction with polling, in calibrating a campaign's microtargeting algorithms and sometimes in altering them mid-campaign. These voter responses are predictive of voting behavior, of course, but again what matters here is whether they are *better* predictors than regular polling. As yet, there is little convincing evidence that they are better (see below). The fundamental problem remains the same, namely that campaigns cannot observe individuals' actual votes.[10]

Perhaps, it might be objected, non-stationarity itself can be addressed by big data methods, at least in principle. Presumably, any non-stationarity is a result of other causal processes, and these other processes might themselves generate trackable correlations. However, this observation is not terribly helpful because it does not provide any actionable advice beyond the truistic 'look for variables that are not subject to non-stationarity'. What matters is whether there is non-stationarity with respect to variables actually tracked. Moreover, there is no guarantee that it offers a solution, even in principle. For it is quite conceivable in any particular domain, especially in hugely complex social domains, such as elections, that the underlying causal processes are so fragile and fast-changing that they never do generate correlations that are trackable.

[10] Similar remarks apply to the increasingly frequent use of randomized experiments and trials, which is the other major innovation of recent campaigns (Issenberg 2016). Such experiments usually test particular campaigning tactics, with efficacy measured either by changes in turnout (which can be observed) or by changes in proxies for actual votes such as opinions expressed on the doorstep or in focus groups. Again, there is no reason to think that such experiments enable us to predict overall election outcomes better than polls do (Note also that, as with the auctions case in section 5, the relevant data in these experiments is *created* in a theory-informed way).

Returning to actual practice, political campaigns certainly have ways of identifying likely supporters and also of estimating how likely those are to vote. But then, so do opinion pollsters. Can campaigns predict overall election outcomes better than pollsters do? So far, there is only fragmented, anecdotal evidence for that (e.g. Issenberg 2016, 324-5, 348). Against that is evidence (admittedly also anecdotal) of precisely the opposite: for example, all sides in the 2018 US presidential election, 2017 and 2015 UK general elections and 2016 UK Brexit referendum were privately surprised by the results. Neither is there any indirect evidence of insider special knowledge, such as telltale activity on political betting markets.

Finally, there are two other alternatives, in some ways more in keeping with big data methods generally. First, might one just adopt the 'n equals all' approach of asking *every* voter how they will vote? But such an interviewing marathon is not a realistic prescription. Moreover, even if it were realistic, it would still address only one of polling's lesser problems, namely sampling error. So, 'n equals all' is no panacea.

Second, might one be able to predict elections, not by asking voters anything, but instead by tracking indirect indicators such as the number of Google searches of candidates? Alas, this method's record is not encouraging, either for predicting elections or for predicting other phenomena such as flu outbreaks. Social media users are often unrepresentative of the target groups. Moreover, there is a new source of non-stationarity, namely that Google's search algorithms themselves change frequently (Lazer et al. 2014).

In conclusion, the prediction of elections has improved although predictive accuracy is still limited. More data has helped. But there are important limitations on how much useful data can ever be available, given the problems of infrequent elections and widespread non-

stationarity. These limitations hamper all big data techniques alike, and can be expected to continue to in the future.

3. Second example: Weather

Earth's weather system is widely believed to be chaotic, so that weather outcomes are indefinitely sensitive to initial conditions (Lorenz 1969). Moreover, it has also been argued that weather predictions are indefinitely sensitive to *model* errors too – that is, even tiny inaccuracies in a model can lead to very large errors in the predictions made by that model (Frigg et al. 2014). These difficulties are ominous. Yet, despite them, weather forecasting has improved significantly.[11] Hurricane paths, for instance, are predicted more accurately and more in advance, and temperature and rainfall predictions are more accurate too. Overall, a few years ago the reliability of seven-day forecasts had become equal to that of three-day forecasts 20 years earlier (Bechtold et al. 2012), and progress has continued since.

Several factors together explain this achievement.[12] The first is a huge increase in the quality and quantity of available *data* since the launch of the first weather satellites in the 1960s. Temperature, humidity and other reports are of ever greater refinement both horizontally (currently increments of 20km squares) and vertically (currently 91 separate altitude layers). Over 10 million observations per day are inputted into the models of leading forecasters.

[11] I will use the terms 'prediction' and 'forecast' interchangeably. There is no uniform usage of these two terms across different sciences. 'Prediction', for instance, may denote any of: in-sample consequences of a model; extrapolation to new subjects; deterministic future earthquake claims; probabilistic future climate claims. Conversely, 'forecast' may denote respectively: forecasts strictly of future, out-of-sample data; forecasts, based on past data, only for known subjects; probabilistic future earthquake claims; deterministic future weather claims.

[12] See (Northcott 2017) for more details and references regarding this case.

The second factor is the forecasting *models*. At the heart of these models are differential equations of fluid dynamics that have been known for hundreds of years. They are assumed to govern the fiendishly complex movements of air in the atmosphere, and how those are impacted by temperature, pressure, the Earth's rotation, the cycle of night and day, and so on. However, in practice these equations are insufficient to generate accurate weather forecasts. Moreover, refining the equations from first principles is not an effective remedy for that. Instead, a whole series of ad hoc additions have been made in order to accommodate the impacts of various specific factors, such as mountains, clouds, or the coupling of air movements and ocean currents. The exact form that these additions take has been determined by a trial-and-error process (Jung et al. 2010, Bechtold et al. 2012). They are under-determined by fundamental theory and indeed they sometimes contradict it.[13]

Third, new *analytical methods* have been developed. The most notable innovation dates from the late 1990s when models began to feature stochastic terms. This enabled the running of multiple simulations to generate probabilistic forecasts. In turn, this 'ensemble method' overcame the problem of chaos: although any one simulation may go seriously askew, it has been found from experience that, as in many chaotic systems, errors tend to cancel out over many iterations. As a result, the probabilistic forecasts are unbiased.

Fourth, available *computing power* has hugely increased, while interacting with other advances to enable the new data to be exploited fully. Thus, the ensemble method of forecasting was infeasible until sufficient computing power became available, because not

[13] The finding that duly refined models can still predict accurately has been the brute empirical solution to the problem of sensitivity to model error mentioned earlier. Simply put, it turns out that, after testing, the models do predict well despite being literally false in many details.

enough simulations could be run in timely fashion. The increase in data and computing power have also together enabled the development and exploitation of more sophisticated models. And additional data is not just collected blindly; rather, experience of what kind of data most improves the accuracy of models' predictions has informed the choice of instruments on new satellites.

With this background in place, let us return to our main concern: what role has big data played? First, weather forecasting's improvement has not been the result of any change in the underlying theory of fluid dynamics. Instead, the forecasting model has been repeatedly tweaked – and in such a way that it has lost easy theoretical interpretation. Different features of the model interact in complex ways so that adjustments are tested holistically in brute instrumentalist fashion. The case thus instantiates the stereotypical big data priority for predictive success over causal transparency.

The improvements in forecasting accuracy are certainly due in part to exploitation of more and better data. They are also due to improvement in data analysis techniques, especially the use of ensemble forecasting.[14] On the other hand, they are not due to these factors alone. Moreover, the improvements are limited: even now, forecasts more than seven or eight days ahead cannot beat the baselines of long-run climate averages or simple extrapolation from current conditions.

As in the case of elections, a crucial question is: how much *could* weather forecasting be improved by the application of big data methods in the future? What is big data's upper limit?

[14] Knüsel et al. (2019) show how big data techniques can also be combined with theory as part of hybrid methods, which are then useful for various subsidiary tasks in the process of prediction, such as finding proxies for missing data or modeling clouds or vegetation.

First, the availability of even more data will indeed likely help. But any new data must be collected by new physical instruments, which requires choices about which instruments to deploy and where. While, as noted, these choices are in part informed by the forecasting model, they also require theory external to it. Thus, background theory is necessary for new data to improve prediction. Moreover, if the weather system is indeed chaotic, only probabilistic forecasts will ever be possible. How accurate such forecasts could eventually become, how far in advance, is unknown.

Second, big data might also improve weather forecasting via the development or application of new methods rather than simply via more data. One possibility is that weather could be 'blindly' predicted by machine learning techniques instead of by, as currently, a model adapted from physical theory.[15] At first sight, the case does seem to satisfy the conditions necessary for such techniques to succeed (sections 6 and 7). To my knowledge, this approach has never been tried. It is hard to assess its potential in advance. There is one thing in its favor, comparatively speaking: since there is little capacity for causal inference from current weather models anyway (section 8), the opportunity cost of a black-box alternative is reduced.

4. Third example: Gross Domestic Product

Predicting GDP has proved very difficult.[16] One benchmark is to assume that the growth rate of real GDP will stay the same as now. Currently, 12-month forecasts barely outperform this benchmark. 18-month forecasts don't outperform it at all. Forecasts also persistently fail to predict turning points, i.e. when GDP growth changes sign. In one study, in 60 cases of

[15] I thank Eric Martin for this suggestion.

[16] See (Betz 2006) for more details and references regarding this case.

negative growth the consensus forecast was for negative growth on only three of those occasions (Loungani 2001).

The record shows little or no sustained difference in the success of different forecasters, despite widely varying methods. These methods include: purely numerical extrapolations, both informal (chartists) and formal (usually univariate time series models improved by trial and error); non-theory-based economic correlations, both informal (indicators and surveys) and formal (multivariate time series); and theory-based econometric models, which sometimes feature hundreds or even thousands of equations. There is no improved return from sophistication, and in particular no superiority of econometric over other methods (Betz 2006, 30-38). Moreover, unlike in the weather case, the forecasting record has not improved in 50 years despite vast increases in available data and computing power in addition to theory development.

The induction is, therefore, that more data and computing power will not improve matters. Given the complexity of what determines a country's GDP, no existing forecasting method likely captures all of the generating processes. Moreover, it seems likely that, as in the elections case, the generating processes are non-stationary. If so, unless it changes over time in the right way, no single predictive method will work for long, including any generated by machine learning techniques. The difficulty applies to any big data approach.

Besides non-stationarity, GDP forecasting also faces other potential difficulties (Betz 2006, 101-108):

1) The economy is an *open* system. In other words, it is continuously impacted by non-economic variables, such as election results, that inevitably do not appear in economic

forecasting models. (As long ago as 1928, Oskar Morgenstern pointed out that economic prediction requires prediction also of non-economic variables.)

2) The economy is a *reflexive* system, in other words forecasts may themselves affect the economy in such a way as to impede the task of forecasting it.[17]

3) *Measurement errors* are large.[18] GDP can only be estimated by aggregating meso-level inputs and the details of that process require many statistical estimates and subjective judgments. Methods for seasonal adjustment introduce further imprecision. One symptom of these difficulties is significant discrepancies between different measuring methods. Another symptom is the large size of revisions, which are typically greater than 1% – comparable to the average forecast error.

4) The economy might be a *chaotic* system, in which case at best only probabilistic forecasts are possible.

In addition, one recent argument holds that confirmation of causal hypotheses in macroeconomics requires knowledge of *unobservable variables*, in particular of agent expectations, and is therefore necessarily infeasible (Henschen 2018). If so, and if accurate forecasting requires a verified causal model (which admittedly it might not), then macroeconomic forecasting too is necessarily infeasible.

It may well be that several or even all of these difficulties are significant. No big data method is a plausible solution for any of them. Accordingly, even if non-stationarity is somehow overcome, big data is not a plausible savior of GDP forecasting.

[17] This is why many rational expectations models deem it impossible to forecast systematically better than a random baseline. Similar pessimism is applied – perhaps more convincingly – to other economic variables besides GDP, such as exchange rates and stock prices. Forecasts of these latter two are, like those of GDP, both unimpressive and not improving.

[18] Data quality is a major difficulty for big data analyses generally (section 7).

5. Fourth example: Economic auctions

Laboratory experiments are increasingly common in social science (Kagel and Roth 2016). In turn, extrapolation from these experiments is an increasingly common guide to field interventions. Such interventions are implicitly predictions of the interventions' own effects, by those who make the interventions. Can big data methods help? I will consider here one well-studied case, namely the US government spectrum auctions from the mid-1990s.[19]

The radio spectrum is the portion of electromagnetic spectrum between 9 kilohertz and 300 gigahertz. In the USA, parts of the radio spectrum that are not needed for governmental purposes are distributed via licenses by the Federal Communications Commission (FCC). In the early 1990s, the FCC acquired the right to do this using competitive market mechanisms such as auctions. That left it the formidable task of designing such auctions. The importance of doing this well is best illustrated by the embarrassments that occur when it is done badly. Examples of that include: an Otago university student winning the license for a small-town TV station by bidding just $5 (New Zealand 1990); an unknown outbidding everyone but then turning out to have no money, thus delaying paid television for nearly a year as do-over auctions had to be run (Australia 1993); and collusion and a subsequent legal fight resulting in four big companies buying the four available licenses for prices only one-fifteenth of what the government had expected (Switzerland 2000). In contrast, the FCC's series of seven auctions from 1994 to 1996 were a remarkable success. They attracted many bidders, allocated nearly two thousand licenses, and raised $20 billion, an amount that surpassed all

[19] For more details and references regarding this case, see Guala (2005), Alexandrova (2008), and Alexandrova and Northcott (2009).

government and industry expectations. Even the first auctions passed off without a glitch, and there was reason to believe that licenses were allocated efficiently.

How was this success achieved? A wide range of goals was set by the government besides revenue maximization, such as efficient and intensive use of the spectrum, promotion of new technologies, and ensuring that some licenses go to favored bidders such as minority- and women-owned companies. Exactly what design would achieve these goals was a formidable puzzle for teams of economic theorists, experimentalists, lawyers, and software engineers. The country was eventually subdivided into 492 basic trading areas, each of which had four spectrum blocks up for license. The eventual auction mechanism put all of these licenses up for sale simultaneously as opposed to sequentially, in an open rather than sealed-bid arrangement. Bidders placed bids on individual licenses as opposed to packages of licenses. When a round was over, they saw what other bids had been placed and were free to change their own combinations of bids. Bidders were also constrained by a number of further rules, such as upfront payments, maintaining a certain level of activity, increasing the values of their bids from round to round by prescribed amounts, and caps on the amount of spectrum that could be owned in a single geographical area. The full statement of the auction rules was over 130 pages.

Game-theoretical models revolutionized the auction literature in the 1980s. However, the final spectrum auction design was not derived (or derivable) from game theory alone. Indeed, no single model covered anywhere near all of the theoretical issues mentioned above. And in addition to instructions covering entry, bidding, and payment, much work also had to be put into perfecting other features such as the software, the venue and timing of the auction, and whatever aspects of the legal and economic environment the designers could control. Many

experiments and consequent ad hoc adjustments and fine-tuning were essential. These took the form of extensive testing in laboratory settings with human subjects. The results often took designers by surprise. For example, in some circumstances – and against theoretical predictions – 'bubbles' emerged in the values of the bids. These in turn were unexpectedly sensitive to the availability of information about rival bidders' behavior. Chief experimental investigator Charles Plott commented:

> Even if the information is not officially available as part of the organized auction, the procedures may be such that it can be inferred. For example, if all bidders are in the same room, and if exit from the auction is accompanied by a click of a key or a blink of a screen, or any number of other subtle sources of information, such bubbles might exist even when efforts are made to prevent them. The discovery of such phenomena underscores the need to study the operational details of auctions. (Plott 1997, 620)

Experiments showed that the impact of any particular auction rule tended to be dependent both on which other rules were included and also on the details of its implementation. Theory alone was typically unable to predict the impact of any given rule individually. Because individual rules did not have stable effects across different environments the performance of any particular *set* of rules had to be tested holistically, and moreover, tested anew with every significant change in environment. This resembles the holistic testing of weather forecasting models. The eventual result of a complex testing process was the perfection of one auction design as a whole, i.e. of a set of formal rules and practical procedures together.[20]

In this way, extensive laboratory investigation was the basis for field predictions, namely of the outcomes of the eventual auctions.

[20] A very similar analysis applies to the even more successful 2000/1 spectrum auctions in the UK.

What was the role of big data methods? It was the work in the experimental testbeds, mired in messy practical details, that was crucial. The key was not new data about bidders or other data that could be collected from existing sources, nor was the key better analysis of such data. Rather, the relevant new data had to be *created* by running experiments and trials. A lot of the benefit from these experiments, as Plott makes clear, came in the form of practical know-how. This was what made the difference.

What about future prospects for big data methods? We have asked this question for each of our examples. The auctions case reveals an implicit assumption. Unlike in many big data success stories, there was no prospect here of simply applying big data techniques to a stock of pre-existing data. Instead, because the relevant data were actively and purposefully created, it was necessary to *decide* what data to create. Prospects for prediction depended on these decisions, and therefore they depended too on the background theory essential to making those decisions.

The type of predictive progress in the auctions case is different too. Much of the 'data' relevant to predictive success were practical know-how, which by its nature tends to be context- and task-specific. Accordingly, progress takes the form of predictive success in one task and then another task and then another, and so on. There is no trend of a greater *degree* of predictive success, rather only a greater *scope*.

The details of the auction case do not bode well for big data advocates. Success required intricate knowledge of the context of application and active creation of relevant data. Both of these require background theory and are not a matter of better machine learning or data

mining. Thus, if the auction case is indicative, big data methods will not improve the derivation of field predictions from laboratory experiments.

6. Conditions for big data predictive success

Summarizing the upshots of the four cases: prediction of weather and elections has improved somewhat. In both cases, more data are part of the reason, as are improved analytical techniques, although sophisticated machine learning methods are not. GDP prediction has not improved; neither more data nor more sophisticated techniques to analyze that data have helped, and they do not seem likely to in the future. With economic auctions, accurate prediction about the impact of interventions requires fresh data to be created with each application, and progress is with regard to scope rather than accuracy; big data techniques of data analysis are irrelevant. Overall, the picture is therefore mixed: more data does help sometimes (not surprisingly), but it is not a panacea anywhere because lack of data is one of, but not the only, constraint on predictive success. New data analysis techniques have been valuable in some cases, but machine learning methods have not played a role.

What determines if big data methods succeed? There has been much work on this question, by both philosophers of science and practitioners. The surveys by the philosopher Wolfgang Pietsch are a useful starting point (2015, 2016).[21] Pietsch discusses several predictive methods that are widespread in data science. One is classificatory trees, which use a number of parameters to determine whether a certain instance belongs to a particular group. Examples include: predicting on the basis of demographic variables which candidate a voter will prefer; predicting on the basis of surf history, cookies and past purchases which product a consumer

[21] For references and more detailed discussion, see these Pietsch papers.

will prefer; and using genetic and environmental factors to predict whether a patient will suffer a certain disease. A second method is nonparametric regression, which, roughly speaking, seeks to account for data using minimal modeling assumptions, thereby allowing great flexibility as to the eventual predictive model's functional form. This method has become feasible only recently because it is so computationally and data intensive. (The contrast between nonparametric and parametric regression is similar to that between data and algorithmic models.)

Pietsch identifies four conditions necessary for such investigations to predict successfully (2015, 910-11):

1) Vocabulary is well chosen, i.e. parameters are stable causal categories

2) All potentially relevant parameters are known

3) Background conditions are sufficiently stable

4) There are sufficient instances to cover all potentially relevant configurations

Label these the *Pietsch conditions*. These conditions apply to big data methods generally, i.e. to techniques of machine learning and data mining.

To see the need for Condition 1: suppose variable X perfectly correlates with Z but Y does not. So, we may predict Z by tracking X. But suppose instead we track only a composite variable X + Y. Then we will fail to predict Z accurately, missing the chance to exploit X. In our four case studies though, satisfying this condition was not the relevant constraint.

The importance of Condition 2 is obvious. Arguably, whenever full predictive success is absent, we cannot be sure this condition is satisfied. GDP and elections are especially clear examples.

Condition 3 refers to non-stationarity, which we have come across already. Any correlation that might be exploited for prediction is presumably generated by some underlying causal process. If that process is unstable then its exploitation may become impossible. Exactly this problem severely limits the efficacy of big data methods for predicting GDP and, to some extent, elections. By contrast, the relative stability of the causal processes underlying the Earth's weather enable big data methods to be much more effective there.

Condition 4 is that the available dataset must be sufficiently rich. Ideally, it should include all relevant configurations of cause and effect variables, else some predictive patterns may be missed. (In practice, even less than this ideal might still enable accurate prediction in a limited range.) Satisfying this condition was the biggest problem in the elections case: the relatively small number of elections is insufficient for selecting between all of the many possible causal hypotheses. Again, weather is a contrast case, because we have ample records of every relevant combination of weather causes and outcomes.

7. Augmenting the Pietsch conditions

In the well-known big data success stories, the Pietsch conditions are satisfied: the underlying generating process is stable enough, the training set of combinations is rich enough, and the variables are well-chosen enough, that we can infer reliably predictive patterns. For example, the process that causes some rather than other New York City manholes to blow seems to be relatively stable, and it was possible to collect a large enough dataset to identify the relevant correlations (Mayer-Schoenberger and Cukier 2013). Thus, the Pietsch conditions do illuminate actual cases. Nevertheless, the case studies enable us to address several further issues.

First, most simply, are the Pietsch conditions actually satisfied in important cases? It seems that they were in the weather example but not in our other ones.

Second, why might the Pietsch conditions sometimes not be satisfied? One reason is a system being open, thus threatening the stability condition because an effect may be unpredictably influenced by unmodeled factors. A system being open also threatens the condition that all relevant parameters are known (elections, GDP). Another reason concerns the sufficient data condition, which may be threatened either because too few iterations exist of the relevant event (elections) or because the relevant data are too contextual (auctions).[22]

Third, are the Pietsch conditions *sufficient* for accurate prediction? One lesson of the case studies is that they are not. The GDP case illustrates well further possible barriers, such as measurement error or a system being chaotic.[23] The weather and election cases highlight the necessity sometimes of other factors too, such as the availability of the new techniques of ensemble forecasts and polling aggregation.

Perhaps these various difficulties can all be recast simply as failures to satisfy Pietsch's conditions: an open system implies either non-stationarity or incorrect vocabulary; reflexivity implies non-stationarity; and measurement errors, unobservable variables, and too few events

[22] A further potential difficulty is a system being reflexive (section 4), as many social systems may be, leading again to a failure of stationarity. It has been suggested that this may apply to elections and GDP, although if it does it is not clear how significantly.

[23] The examples also illustrate remedies for some of these difficulties. In the weather case, for instance, the ensemble method makes possible the (probabilistic) prediction even of a chaotic system.

each implies the unavailability of sufficient (accurate) data.[24] But the point is that most of the real work now consists in assessing when and why the conditions will actually apply, so there is no substitute for supplementary local investigation.

This contextualist moral also leads to the recognition that prediction may often be an amalgam of various methods, some of which rely on large datasets and make use of big data methods such as machine learning, while others do not. Even if big data does not improve overall predictions, it might still be helpful for certain local modeling tasks where theory is scarce but data are not. Thus, often the impact of big data on prediction in any given domain is not all-or-nothing. Knüsel et al. (2019) illustrate this in the context of climate science. There are examples in our case studies too. For instance, as noted, political campaigns successfully use big data methods to predict and influence many aspects of voter behavior.[25]

Fourth, the importance of *background theoretical knowledge* is underlined. Such knowledge is often an essential guide to choosing the Pietschian correct vocabulary, as in both the election and weather cases. It is often crucial too for correcting non-stationarity, and for offering guidance regarding extrapolation (section 8; see also Knüsel et al. 2019, 199). That

[24] The distinctions between Pietsch's different conditions are themselves fuzzy, as sometimes the same issue can be assigned to more than one of the categories. For example, suppose that in the 1980s in the UK buying a Ford Fiesta car predicts support for Conservatives but that by 1997 it predicts support for Labour. This is a case of non-stationarity. But suppose we re-describe the situation in terms of a more fundamental causal relation, perhaps that 'middlebrow voters vote for middlebrow politicians'. This new relation is plausibly stable across the different elections. If so, rather than non-stationarity, the initial problem would become a case of incorrect vocabulary, or perhaps that not all relevant parameters are known. Generally, any open system is vulnerable to disruption by unmodelled variables, and thus a problem of non-stationarity is always vulnerable to being re-classified in this way.

[25] See also https://www.facebook.com/business/success/rick-scott-for-florida#u_0_0 and https://www.facebook.com/business/success/snp for controversial (claimed) examples. I thank an anonymous referee for raising the issues in this paragraph.

is, theoretical knowledge is beneficial even just for the narrow goal of better prediction.[26] Election predictions, for example, can be improved by incorporating systematic turnout differences between US midterm and presidential years, or by understanding why turnout patterns changed between the 2015 and 2017 UK elections. And theory informed the experiments that gathered the relevant data in the auctions case. It is also a commonplace that background theory and knowledge of the data-generating mechanism are often essential to handling likely data errors effectively (Foster et al. 2017, 180).

A note of caution though: although background theory is thus indispensable, at the same time theory alone does not predict successfully in the field. All of the case studies confirm that. The weather forecasting models require many ad hoc adjustments that go beyond, or even contradict, basic theory; models of elections based on fundamentals are out-predicted by opinion polling; theory-based forecasts of GDP fare no better than those based on other methods; and the spectrum auction design was not derived, or derivable, from game-theoretical models but rather used those models only as heuristic starting points to be repeatedly refined by sui generis experiments. As a result of this need for extra-theoretical input, predictive success when it is achieved is local and hard to extrapolate to new contexts. Evidence suggests that this pattern is typical of field cases generally (Tetlock and Gardner 2015). Recent work suggests that machine learning methods likewise predict best by avoiding theory-driven models (Mullainathan and Spiess 2017). In our case studies, more data do not counter this anti-theory trend; if anything, they exacerbate it.

There have also been analyses from big data practitioners themselves. There is some overlap between these and Pietsch's conditions. Correct choice of variables is recognized as key for

[26] Sabina Leonelli (2016) compellingly emphasizes the same point for the field of systems biology.

making machine learning methods work, for instance. However, it is notable that practitioners have markedly different emphases. Perhaps the biggest issue for them is ensuring data quality, which is a catch-all for a range of more specific issues, such as whether data are representative, whether there are measurement errors, and whether the data capture what we want them to – often data are repurposed, being the products of instruments and methods not designed with data scientists in mind (Foster et al. 2017, 276-285; Japec et al. 2015, 848-850). Such concerns are captured by the Pietsch conditions at best only implicitly.

In our case studies, however, while data quality issues such as the reliability of polling evidence and weather measurements are concerns, they are not the biggest constraints on predictive accuracy. Conversely, what is in practice the biggest such constraint, namely non-stationarity, is comparatively neglected by practitioner analyses.[27] In this way, the case studies augment them too.

8. Theory and causal understanding

The case studies also shed light on causal understanding. Advocacy for big data methods has often celebrated those methods' *lack* of connection to causal inference: the 'death of theory' heralds an emphasis instead exclusively on correlation and prediction (Mayer-Schoenberger and Cukier 2013, Hey et al. 2009). But this is not quite right, as theoretical analysis and case studies together reveal.

[27] This despite the fact that non-stationarity is a classic concern in statistics, with many associated diagnostic tests. Knüsel et al. (2019) is one exception to the pattern, as they do take what they call 'constancy' in the data to be in practice the most important condition to satisfy.

Pietsch (2016) shows how some big data methods can offer causal inference – up to a point. In particular, examining patterns of covariation enables these methods to identify INUS causes in the sense of (Mackie 1980), even though the available evidence is only observational.[28] If we assume a stable causal background then INUS causes in turn license interventions because actual variations can stand as proxies for the relevant counterfactual ones. Thus, one benefit of theory can be achieved even by 'theory-free' big data methods. On the other hand, these same methods are vulnerable to spurious correlations in the same way as Mackie himself noted for INUS causes, and as practitioners recognize too (Foster et al. 2017, 277-279). The best defense against such spurious correlations is to import background knowledge – and so theory reappears. Moreover, INUS causes merely mark patterns of covariation. They do not provide mechanistic or other underlying explanations of those patterns, nor therefore any understanding of why they hold, nor therefore any guidance as to when they will extrapolate to new contexts. To plug these gaps, again background theory is required.

The case studies demonstrate that, in practice, causal understanding is hard to deliver. Even given relative predictive success (weather, elections, auctions), our ability to explain has increased very little. Testing of the weather model and the auction design was holistic, militating against assigning causal responsibility to particular factors and thus against causal explanations.[29] Generally, the theoretical demonstration of the possibility of causal inference by big data methods turns out to be mostly inapplicable to our cases, because an algorithmic search for correlations from which INUS causes can be inferred is too simple a method to be

[28] 'INUS' stands for an *I*nsufficient but *N*ecessary part of an *U*nnecessary but *S*ufficient condition.

[29] Lenhard (2018) argues that holistic testing tells against causal inference in complex simulation models generally.

useful. Rather, there are breakdowns of stationarity (elections, GDP, auctions), or insufficient data (elections), or the data need to be created (auctions).

The picture is not quite wholly bleak though. Causal understanding is occasionally achievable in the weather case (Northcott 2017). At root, this is because of the exceptional quantity of data available. For example, recently, there have been extensive changes to the model's treatment of convection schemes in the tropics and to its treatment of the radiative properties of ice clouds. These changes have, of course, been thoroughly tested for their impact on predictive success and refined accordingly. But in addition, the data allowed modelers to test whether the two changes composed non-linearly or not. In this case, it was found that the non-linear – i.e. interactive – effects were relatively small. Accordingly, empirically verified changes in model outputs (i.e. successful predictions) could now be attributed to particular changes in model inputs; that is, some causal transparency was returned.

Some limited extrapolation was possible in the auction and election examples too, as work from earlier cases helped with later cases such as the UK spectrum auction of 2000/1 and later US presidential elections. Still, even then this extrapolatory help was far from infallible, as witnessed by the failure of the 2000 spectrum auction in Switzerland and (relative) failure of forecasting of UK parliamentary elections. Usually, new models are needed each time (Northcott 2017).

9. Conclusion

In the right circumstances, the Pietsch conditions are satisfied, and big data methods do significantly advance field prediction. They enable the best possible use to be made of a

given body of data, and they promise still more predictive success as new techniques and more data become available. But often the binding constraint is neither lack of data nor our inefficient use of them. Instead, prediction may be hindered by a system being non-stationary (elections, GDP, auctions), or chaotic, open, or reflexive (weather, GDP, perhaps others). These hinder all big data methods alike. Or a lack of relevant data cannot feasibly be remedied (elections), or new data can, in whole or in part, only be collected with non-big-data methods (auctions, elections, weather, GDP). Such problems may be frequent and unfortunately may affect those cases that we most want to predict.[30]

There is still a need for theory and thus for human experts – in part because it is this that enables some predictive progress even in the face of the difficulties above. Theory and experts ubiquitously inform both the correct choice of variables to analyze and the collection of data in the first place. They are essential to the 'internal' operation of prediction too. Nevertheless, despite this continued role for theory, the possibilities for causal inference are often very limited.

Our case studies illuminate all of these issues. Overall, they suggest caution about whether prediction, and thus scientific method generally, will really be revolutionized by big data.

[30] Are social sciences more prone than natural sciences to these problems? I do not yet see convincing evidence of that. Natural language translation and internet company experiments are success cases from social science. The only problem specific to social science is reflexivity, and it is not clear how often reflexivity is predictively significant.

References

Alexandrova, A. (2008). 'Making Models Count'. *Philosophy of Science* 75, 383-404.

Alexandrova, A., and R. Northcott (2009). 'Progress in Economics: Lessons from the Spectrum Auctions', in H. Kincaid & D. Ross (eds.), *The Oxford Handbook of Philosophy of Economics*, Oxford University Press, 306-337.

Anderson, C. (2008). 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired Magazine*, June 23, 2008.

Bechtold, P., P. Bauer, P. Berrisford, J. Bidlot, C. Cardinali, T. Haiden, M. Janousek, D. Klocke, L. Magnusson, T. McNally, F. Prates, M. Rodwell, N. Semane, and F. Vitart (2012). 'Progress in Predicting Tropical Systems: The Role of Convection', ECMWF Research Department Technical Memorandum no. 686.

Betz, G. (2006). *Prediction or Prophecy?* (Wiesbaden: Deutscher Universitaets Verlag)

Foster, I., R. Ghani, R. Jarmin, F. Kreuter, and J. Lane (2017). *Big Data and Social Science.* (Boca Raton, Florida: CRC Press)

Frigg, R., S. Bradley, H. Du, and L. Smith (2014). 'Laplace's Demon and the Adventures of His Apprentices', *Philosophy of Science* 81: 31-59.

Guala, F. (2005). *Methodology of Experimental Economics.* Cambridge: Cambridge University Press.

Henschen, T. (2018). 'The in-principle inconclusiveness of causal evidence in macroeconomics', *European Journal for Philosophy of Science* 8.3, 709–733.

Hey, T., S. Tansley, and K. Tolle (eds) (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery.* (Redmond, WA: Microsoft Research)

Issenberg, S. (2016). *The Victory Lab.* (Broadway, New York)

Japec, L., F. Kreuter, M. Berg, P. Biemer, P. Decker, C. Lampe, J. Lane, C. O'Neil, and A. Usher (2015). 'Big data in survey research: AAPOR Task Force Report', *Public Opinion Quarterly* 79.4, 839– 880.

Jung, T., G. Balsamo, P. Bechtold, A. Beljaars, M. Koehler, M. Miller, J-J. Morcrette, A. Orr, M. Rodwell, and A. Tompkins (2010). 'The ECMWF Model Climate: Recent Progress Through Improved Physical Parametrizations', *Quarterly Journal of the Royal Meteorological Society* 136: 1145–1160. (ECMWF Technical Memorandum No 623)

Kagel, J., and A. Roth (eds) (2016). *The Handbook of Experimental Economics, Volume 2.* (Princeton)

Knüsel, B., M. Zumwald, C. Baumberger, G. Hirsch Hadorn, E. Fischer, D. Bresch, and R. Knutti (2019). 'Applying big data beyond small problems in climate research', *Nature Climate Change* 9, 196-202.

Lander, E. (2016). 'The Heroes of CRISPR', *Cell* 164, 18-28.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). 'The Parable of Google Flu: Traps in Big Data Analysis', *Science* 343, 1203-1205.

Lenhard, J. (2018). 'Holism, or the Erosion of Modularity: A Methodological Challenge for Validation', *Philosophy of Science*, 832-844.

Leonelli, S. (2016). *Data-Centric Biology*. (University of Chicago Press)

Lewis-Kraus, G. (2016). 'The Great AI Awakening', *New York Times Magazine* 14/12/2016.

Lorenz, E. (1969). 'Three Approaches to Atmospheric Predictability', *Bulletin of the American Meteorological Society 50: 345–349.*

Loungani, P. (2001). 'How Accurate are Private Sector Forecasts? Cross-country Evidence from Consensus Forecasts of Output Growth', *International Journal of Forecasting* 17, 419-432.

Mackie, J. (1980). *The Cement of the Universe*. (Clarendon)

Mayer-Schoenberger, V. and K. Cukier (2013). *Big Data*. (John Murray)

Mullainathan, S., and J. Spiess (2017). 'Machine Learning: An Applied Econometric Approach', *Journal of Economic Perspectives* 31.2, 87–106.

Northcott, R. (2015). 'Opinion Polling and Election Predictions', *Philosophy of Science* 82, 1260-1271.

Northcott, R. (2017). 'When are Purely Predictive Models Best?', *Disputatio* 9.47, 631-656.

Pietsch, W. (2015). 'Aspects of Theory-Ladenness in Data-Intensive Science', *Philosophy of Science* 82, 905-916.

Pietsch, W. (2016). 'The Causal Nature of Modeling with Big Data', *Philosophy and Technology* 29, 137-171.

Plott, C. (1997). 'Laboratory Experimental Testbeds: Application to the PCS Auction', *Journal of Economics and Management Strategy* 6.3, 605-638.

Sturgis, P. Baker, N. Callegaro, M. Fisher, S. Green, J. Jennings, W. Kuha, J. Lauderdale, B. and Smith, P. (2016). 'Report of the Inquiry into the 2015 British general election opinion polls', London: Market Research Society and British Polling Council.

Tetlock, P., and D. Gardner (2015). *Superforecasting.* (Random House)

Wells, A. (2018). http://ukpollingreport.co.uk/blog/archives/10002

Acknowledgements
For helpful feedback, I would like to thank: two anonymous referees; audiences at the University of Kent, the Centre for the Future of Intelligence at Cambridge, the American Philosophical Association, and Birkbeck College; and Wolfgang Pietsch, Katie Creel, Adrian Currie, Jacob Stegenga, Anna Alexandrova, and Rune Nyrup.