European Review of Agricultural Economics Vol **47** (**3**) (2020) pp. 849–892
doi:10.1093/erae/jbz033
Advance Access Publication 21 August 2019

Machine learning in agricultural and applied economics

Hugo Storm[†,*], **Kathy Baylis**[‡] **and Thomas Heckelei**[†]

[†]*Institute for Food and Resource Economics, University of Bonn, Germany;* [‡]*Agricultural and Consumer Economics, University of Illinois, USA*

Received December 2018; editorial decision May 2019; final version accepted June 2019

Review coordinated by Salvatore Di Falco

Abstract

This review presents machine learning (ML) approaches from an applied economist's perspective. We first introduce the key ML methods drawing connections to econometric practice. We then identify current limitations of the econometric and simulation model toolbox in applied economics and explore potential solutions afforded by ML. We dive into cases such as inflexible functional forms, unstructured data sources and large numbers of explanatory variables in both prediction and causal analysis, and highlight the challenges of complex simulation models. Finally, we argue that economists have a vital role in addressing the shortcomings of ML when used for quantitative economic analysis.

Keywords: machine learning, econometrics, simulation models, quantitative economic analysis, agri-environmental policy analysis

JEL classification: C14, C45, C63

1. Introduction

Machine learning (ML) offers great potential for expanding the applied economist's toolbox. Recent overview papers have pointed to the potential for big data and ML to improve farm management (Raj *et al*., 2015; Shekhar *et al*., 2017; Coble *et al*., 2018; Kamilaris and Prenafeta-Boldú, 2018) and economic analysis more broadly (Einav and Levin, 2014; Varian, 2014; Bajari *et al*., 2015; Grimmer, 2015; Monroe *et al*., 2015; Athey and Imbens, 2016). ML tools are beginning to be employed in economic analysis (März *et al*., 2016; Crane-Droesch, 2017; Athey, 2019), while some researchers raise concerns about their transparency, interpretability and use for identifying causal relationships (Lazer *et al*., 2014). In this review paper, we introduce ML to

*Corresponding author: E-mail: hugo.storm@ilr.uni-bonn.de

applied economists by placing it in the context of standard econometric and simulation methods. We identify shortcomings of current methods used in agricultural and applied economics, and discuss both the opportunities and challenges afforded by ML to supplement our existing approaches.

What is ML? The terms ML, artificial intelligence (AI) and deep learning (DL) are often used interchangeably. ML is part of artificial intelligence, which in turn is a discipline in computer science. ML aims to learn from data using statistical methods. DL is a specific subset of ML that uses a hierarchical approach, where each step converts information from the previous step into more complex representations of the data (Goodfellow *et al*., 2016). Many of the newest advances in machine learning are in the area of DL (LeCun, Bengio and Hinton, 2015).

Why introduce ML to agricultural and applied economics now? First, data availability has dramatically increased in many different areas, including agriculture, environment and development (Shekhar *et al*., 2017; Coble *et al*., 2018). Along with helping process data from these novel sources, ML methods are well equipped to exploit large volumes of data more efficiently than traditional statistical methods. Second, since the early 2000s, the use of multiprocessor graphic cards (graphic processing unit, or GPU) has greatly sped up computer learning (Schmidhuber, 2015), and many ML methods can be parallelised and exploit the potential of GPUs. Third, the ML/DL research community from both academia and industry is rapidly developing the tools users need to apply these methods. Researchers have developed and improved algorithms that push the boundaries of ML/DL (Schmidhuber, 2015). The community has a strong open source tradition, including powerful DL libraries (e.g. tensorflow.org, pytorch.org) and pretrained models (e.g. VVGNet, ResNet), increasing the potential for adoption. Last but not least, economists have begun to realise that the predictive power of ML methods may not only be used as such, but can also improve causal identification (Athey, 2019).

How can ML be helpful for agricultural and applied economics? Our models often contain little prior information about functional form, have large potential heterogeneity across units of observation and frequently have multiple outputs. For example, imagine one wants to estimate the effect of a fertiliser subsidy on the yield of crops. Yield is determined by a complex combination of soil quality, weather, inputs, input timing and other management choices, replete with non-linearities and interactions. Or suppose one wants to ask how subsidies affect farm structure, where both policy and structure may be complex and multidimensional. In demand system estimation, one might have access to daily, product-level scanner data or data on housing sales to estimate preferences for local amenities, or one may want to estimate the effect of pollution on multiple measures of health. While our traditional methods have allowed us to approach these questions, ML increases the flexibility with respect to both data and functional form, as well as processing efficiency, opening up other avenues for analysis.

Often ML approaches are perceived as something special or even mysterious, potentially due to associated terms such as AI, neural networks (NN) or

DL, that create associations with human intelligence. As we lay out in Section 2, ML tools are 'just' statistical tools and in many ways are a natural extensions of the econometric toolbox. In the next section we introduce central ML approaches, not aiming for textbook coverage, but rather to present them from an applied econometric perspective highlighting similarities and differences with our traditional methods. One distinction is that ML focuses primarily on the predictive accuracy and forecast errors while econometricians focus on deriving statistical properties of estimators for hypothesis testing (see also Mullainathan and Spiess, 2017). In Section 2.1, we present the ML approach to predictive accuracy and to control for overfitting. We also present central supervised learning approaches for regression tasks (Section 2.2) and unsupervised learning approaches for dimensionality reduction (Section 2.3). Often there are concerns about ML models being a 'black box' and we reflect on the tradeoff between model complexity versus interpretability in Section 2.4, including tools to help interpret ML models.

Section 3 then takes a closer look at limitations of our current set of econometric tools and simulation methods, and explores to what extent ML approaches can overcome them. We frame this section in terms of current challenges faced in applied economic analysis; while there may be some overlap in the ML solutions, the problems being addressed are different. Functional forms employed in econometric analysis often lack theoretical grounding and are not sufficiently flexible to capture the multiple interactions, non-linearities and heterogeneity so common to biological or social processes in agricultural and environmental systems. ML tools allow for highly flexible estimation, address model uncertainty and efficiently deal with large sample sizes (Section 3.1). Our current methods limit the full use of novel unstructured data sources, such as remote sensing images, cellular phone records or text from news and social media. ML approaches may reduce the reliance on limited 'hand-crafted' features to make better use of the available data (Section 3.2). Similarly, ML offers opportunities in situations in which we have a very large number of potential explanatory variables or observe explanatory variables at high temporal or spatial resolution for which our current approaches to aggregate data into a standard panel form implies loss of information (Section 3.3). One common objection from economists is that ML tools are of only of limited use as they focus on prediction while economists are primarily interested in answering causal questions. While it is true that ML tools are primarily developed for prediction, there are recent contributions, particularly from economists, that exploit the prediction capabilities of ML tools for causal inference. We provide an overview of these approaches and how they can help to overcome limitations of the current tools for causal inference (Section 3.4). Beyond enhancing econometric methods, ML can help alleviate current constraints of simulation models. Partial or general equilibrium models or Agent Based Models (ABMs) are often computationally limited in their degree of complexity. Further, empirical calibration of equilibrium models or ABMs is challenging. ML methods are beginning to be employed to overcome these computational

limitations and to improve calibration (Section 3.5). In Section 4, we discuss potential limitations of ML approaches and what economists can add to overcome these limitations. Finally, we identify some relevant frontier developments in ML for economic analysis (Section 5).

While some of the issues reviewed in this paper have been raised in the general economics literature, and several authors have already highlighted the potential of 'big data' for agricultural economics, no overview on the existing and potential applications of ML methods for agricultural and applied economics analysis yet exists. We believe these methods hold particular promise for researchers in our field due to the frequent linkages with complex biological or physical processes, uses of non-traditional data sources such as those derived from remote sensing and the frequent use of simulation methods. While, like other reviews, we briefly introduce ML methods, we do so from the perspective of our standard econometric and simulation tools to aid understanding and appropriate application. Unlike earlier reviews, we highlight how ML tools can fill gaps in our existing methodological tool box, focusing on what long standing challenges they can solve. We place particular emphasis on NNs because despite holding significant potential for capturing complex spatial and temporal relationships, they are still not greatly used in economic analysis. Further, we review the application of ML tools in policy simulation, which, to our knowledge has not yet been extensively covered. We hope that relating ML methods to our current approaches and their shortcomings will allow this paper to serve as a guide for applied economists interested in expanding their methodological toolbox.

2. ML from an applied econometrics perspective

We begin this review by briefly introducing ML concepts, terminology and approaches. Our intention is not to give these topics a rigorous treatment, but instead to provide an intuitive introduction from a practitioner's perspective, laying out high-level connections to traditional econometric approaches, and to identify both the potential and limitations for empirical applications. Traditionally, ML and econometric approaches have different objectives. ML approaches are primarily intended for prediction tasks with the aim to obtain accurate predictions, while in econometrics we are usually interested in obtaining reliable estimates of marginal effects. This difference has important implications. For example, when the ML community refers to bias, variance or mean squared error (MSE), they are defined in terms of the prediction, i.e. the aim is to have an precise and unbiased prediction. While in econometrics we are usually interested in obtaining unbiased/consistent estimates of the coefficients. Importantly, a model that is unbiased in terms of the prediction might not necessarily be unbiased in terms of the coefficients. Another important difference is that in econometrics we are able to derive uncertainty estimates of the estimated coefficients and hence can use the estimates for hypothesis testing. Uncertainty estimates are usually not obtained for ML

methods, which is a substantial limitation of the approach and is an area of active research (see Section 5).

2.1. Regularisation/train-validation-test split approach to avoid overfitting

For prediction tasks, we aim to estimate models that generalise well, meaning that the estimated model generates accurate predictions for observations outside the employed sample. Models need to learn general relationships from the data but avoid 'overfitting', i.e. avoid learning aspects of the given sample that do not generalise to the population. Limiting overfitting is particularly important given the many parameters or non-parametric nature, and thus the high flexibility of many ML methods, which allows the model to fit very specific (nonlinear) relationships in the data.

In traditional econometrics, we are concerned about having 'sufficient' degrees of freedom, where more degrees of freedom reduce the standard errors around any single estimated coefficient. This approach inherently restricts the number of covariates (given a finite '*N*'), and thus limits the flexibility of a model. In the ML community, degrees of freedom are not explicitly considered and often ML methods contain a very large number of parameters and potentially negative degrees of freedom. In ML, limiting overfitting is typically done via regularisation. Regularisation in ML terms, controls the complexity (or capacity) of a model. Intuitively, the complexity of a model describes its ability to approximate a wide variety of functions. With increasing complexity, i.e. less regularisation, the risk of overfitting increases, while less complex, more regularised models might lead to underfitting (Hastie, Tibshirani and Friedman, 2009: 219–223; Goodfellow *et al.*, 2016: 107–117). In econometrics, the concern about overfitting is frequently overshadowed by the goal of obtaining accurate coefficient estimates. Regularisation often comes in the form of the selection of a parsimonious number of variables and the use of specific functional forms, without explicitly controlling for overfitting.

When regularising a model, one needs to make a trade-off between bias and variance, where, in prediction tasks, bias and variance refer to the prediction. Highly regularised (i.e. less complex, less flexible) models tend to have high (prediction) bias but low variance. As an extreme case of regularisation, think about predicting the outcome to be a constant, irrespective of explanatory variables. Less regularised, highly complex models tend to have low bias but high variance. Because MSE is the sum of squared bias and variance, the trade-off between bias and variance is embedded in minimising the MSE as the criterion for model selection.

One standard ML approach to find the appropriate level of model complexity is to split the available data set into a training, validation, and test set (see Hastie, Tibshirani and Friedman (2009), section 7 or Goodfellow *et al.* (2016) section 5.3 for textbook coverage, providing the basis for this

section). The training set is used for estimation (called 'training') the model and the validation set (also called the development or hold-out set), are used to monitor the out-of-sample prediction error. The out-of-sample prediction error, or the prediction error in the validation set, is monitored for different model specifications and different levels of model complexity. The model with the lowest out-of-sample prediction error in the validation set is then selected. The test set is then finally used to assess the out-of-sample prediction error of the selected model. Thus, it is important that the test set is neither used for training nor for model selection.

The train/validation/test approach can easily be applied in a data rich environment where setting aside a portion of the data is not a problem. When datasets are smaller, a common variation of the train/validation/test split approach is *k*-fold cross validation. This approach calculates the expected out-of-sample prediction error along with an estimate of the standard error of the out-of-sample prediction error in an iterative way. The general approach is to split the sample in *k* parts, each with equal number of observations.[1] Using these splits, we then estimate our chosen model *k* times; each time we use all the data except one of the *k* parts that we leave out. This left-out part is then used to derive the out-of sample prediction error. By averaging the out-of-sample prediction error over the *k* estimators, we obtain an estimate of the expected value of the out-of-sample prediction error.

In ML, *k*-fold cross validation is frequently used for model selection, or to select tuning parameters of a specific estimator (such as the learning rate of the numerical optimiser or the number and layer/neurons in a NN, discussed below). Cross validation is performed for each of the possible models or a range of tuning parameters as described above. The model/tuning parameters with the lowest expected out-of sample prediction error is then chosen as the final model. The final model is then estimated using the entire data set. It is interesting to contrast the cross validation or train/validation/test approach in ML, with the typical econometric approach to model selection, where variables may be given by theory or criteria like AIC or BIC are used.[2] Typically, in econometric forecasting, we tend to drop observations to measure the prediction error of our chosen model; not as part of a systematic model selection process. Conversely, this is part and parcel of ML methods.

2.2. Supervised approaches

Supervised approaches characterise the methods to estimate the conditional expectation of a dependent variable, or target in ML terms, given explanatory variables called features. Hence, supervised approaches include classical linear or limited dependent variable regression models. While a large variety of

1 In a simple cross section, splitting the data by random draw is straightforward; when dealing with time or spatially dynamic models one needs to take into account the data structure and the objective of the prediction task in order to decide on the most appropriate way to split the data.

2 One possible exception is the use of cross-validation for the selection of bandwidth in non-parametric kernel estimation, which is similar to the uses of the validation set in ML.

supervised ML approaches exist, we restrict ourselves to shrinkage methods, tree-based methods and neural networks, which hold particular relevance for applied economics.

2.2.1. Shrinkage methods

Shrinkage methods such as ridge regression or lasso are linear regression models that add a penalty term to the size of the coefficients, pushing coefficients towards zero. They can be used for prediction of continuous outcomes or classification and can efficiently be used on data sets with large numbers of explanatory variables. For coefficients to deviate from zero, variables have to substantially contribute to predictive power. The extent of shrinkage or regularisation can be tuned, where the optimal level is typically determined using cross-validation (see Varian (2014) for a brief discussion or Hastie, Tibshirani and Friedman (2009) for a more detailed exposition).

For the econometrician who is largely interested in finding the 'true' model and interpreting regression coefficients, newer variations of the lasso may be of specific interest. These new variations close in on the so-called 'oracle' property that offers good properties of model selection and coefficient estimation. One promising approach is the OLS post-lasso estimation (i.e. the penalty term pushing some coefficients to zero; Belloni and Chernozhukov, 2013). For a broader and more rigorous treatment of inference with lasso, including the oracle property and relevant sparsity conditions we refer to Tibshirani, Wainwright and Hastie (2015).

2.2.2. Tree-based methods

Decision trees can be used for both classification and regression. They use linear splits to partition the feature space (i.e. the space spanned by the explanatory variables), to maximise the homogeneity within the partitions created by each split, with the end of the sequential splits called 'leaves'. Once the tree is 'grown', one can use it to predict an outcome based on which side of each sequential split that observation's covariates fall, i.e. which 'leaf' it populates. The depth of a tree describes the number of splits, or nodes. Each split is sequentially chosen based on its contribution to the loss function. Trees are a useful tool for applied economists because they can easily be interpreted and are well suited to capture highly non-linear relationships. A disadvantage of trees is that they can be unstable and prone to overfitting, such that small changes in the data lead to substantial changes in splits. Even though they are well equipped to capture non-linearities, they are limited in capturing trully linear or smooth functions since, by construction, the resulting model is a step function. However, with sufficient data they can approximate any linear or smooth function arbitrarily well and, importantly, without the need to assume an underlying structure ex-ante.

Ensemble approaches such as random forests or gradient boosted trees combine the results of multiple trees in order to improve prediction accuracy and to reduce variance, at the cost of easy interpretability. Random forests

average the results of many deep trees grown on random subsamples of observations, and subsets of variables. Random forests can be thought of as being related to kNN methods with adaptive weighting (Lin and Jeon, 2006), where the predicted outcome of an out of sample observation is given by its neighbours defined by a weighting of its characteristics. Gradient boosted trees are additive models consisting of the sum of trees trained by repeatedly fitting shallow trees on the residuals (Efron and Hastie, 2016: 324). Given their additive structure, boosted trees are closely related to generalised additive models (GAMs) in traditional econometrics. However, estimation of GAMs is less efficient than gradient boosting when working with a large number of explanatory variables. These methods are currently among the most effective prediction techniques applied in many different areas (Hastie, Tibshirani and Friedman, 2009; Efron and Hastie, 2016: 347). Efron and Hastie (2016) argue that these methods are well suited as an 'off-the-shelf' ML prediction approach given their advantages to detect highly non-linear relationships, process quantitative and categorical data, are robust to highly non-normal data or outliers, provide an algorithmic treatment of missing data, irrelevant variables and consequently require relatively little preprocessing of the input data and comparatively little tuning during training. Additionally, they provide a ranking of the importance of each explanatory variable.

2.2.3. Neural networks

Next to tree-based methods, NNs are the most widely used, effective supervised ML approaches currently available. Sarle (1994) provides an early comparison between NNs and statistical models, including a overview of ML jargon. Goodfellow *et al*. (2016) provide a recent textbook on NNs, particularly deep neural networks (DNN), which is the basis for this section. As with any other supervised approach, including a classical regression, NNs are simply a mapping $\boldsymbol{y} = f(\boldsymbol{x}; \theta)$ from an input vector x to an output vector $\boldsymbol{y}$, governed by unknown parameters θ. Characteristically, the mapping consists of layers building a chain like structure of functions. A network with three layers would look like: $y = f(\boldsymbol{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\boldsymbol{x})))$. Deep neural networks refer to a NN with many layers. In a fully connected (or dense) NN, each layer has a structure given by $f^{(k)}(x) = \boldsymbol{h}^{(k)} = g^{(k)}(\boldsymbol{W}^{(k)\top}\boldsymbol{h}^{(k-1)} + \boldsymbol{b}^{(k)})$, with $\boldsymbol{h}^{(0)} = \boldsymbol{x}$, where $W^{(k)}$ is a matrix of unknown parameters and $b^{(k)}$ is a column vector of basis factors (similar to a constant in a regression). The column dimension of $W^{(k)}$ (or row dimension of $b^{(k)}$) specifies the size or number of neurons in each layer. The dependent variable, y, can include multiple jointly predicted characteristics per observation. Typical choices for the activation function $g^{(k)}$ are a rectified linear unit (relu) or a tanh transformation function. The weights of the NN are trained by minimising a loss function, such as mean squared error for regression or cross-entropy for classification. Note both linear and logit regression are special cases of a NN, when the NN has only one layer, y is of dimension one and we use a linear or logistic

activation function, respectively. From this perspective, NNs are already widely used in our profession![3]

While there are many NN architectures, the two most relevant for economists are convolutional neural networks (CNN) and recurrent neural networks (RNN). CNNs are well placed to process grid-like data such as 1D time-series data or 2D image data. CNNs get their name from the use of a convolutional operator in at least one of their layers, which is then called a convolutional layer. In a fully connected (dense) neural network, every unit in a hidden or output layer is connected to every unit (neuron) in the previous layer by the matrix multiplication $\boldsymbol{W}^{(k)\top}\boldsymbol{h}^{(k-1)}$. In a convolutional layer, by contrast, each unit looks only at a small fraction of units from the previous layer (thus, a sparse interconnection) and uses the same parameters at different locations (parameter sharing), thereby significantly reducing the number of parameters it needs to estimate. Intuitively, a convolutional layer in a time series model can be thought of as a collection of filters that are shifted across the time sequence; for example, one filter that detects cyclical behaviour and another that calculates a moving average. A crucial distinction between CNNs and classical time series models is that CNNs learn the parameters of the filter, i.e. the features that are useful to extract. In an image processing application, for example, a filter may learn to detect vertical edges in small locations of the image, while another filter detects horizontal edges, corners and curved lines. Each filter is then moved across the image to create a feature map (one from each filter) specifying where the features are present in the image. The next convolutional layer then combines the features (edges, corners etc.) into more complex structures (e.g. an eye, mouth or nose), providing maps of those features.

RNNs are an alternative to CNNs for processing sequential data, handling dynamic relationships and long-term dependencies. RNNs, particularly RNNs employing Long Short Term Memory (LSTM) cells, gained popularity and led to important advances in natural language processing (machine translation, speech recognition or speech synthesis (Schmidhuber, 2015)). The crucial feature of such RNN-LSTM models is that past information is carried across time using a cell state vector. During each time step, new incoming explanatory variables are encoded and combined with the past information in the cell state vector. Importantly, the model itself learns (i) in which way information is encoded and (ii) which encoded information can be forgotten (i.e. is not important for the prediction of later steps). As with CNNs, this approach differs from a classical AR process as it does not require the analyst to specify the lag structure and can capture more complex relationships. RNNs and CNNs both leverage the idea of parameter sharing, which allows them to detect a certain pattern irrespective of the pattern's location in a sequence or image. Both can be applied in a context of either very long time series or in a panel context with many short time series.

3 See also Hastie, Tibshirani and Friedman (2009) for laying out the similarities between projection pursuit regression and neural networks.

2.3. Unsupervised approaches

Along with prediction, another common use of ML is data grouping or clustering based on the characteristics of observations. Unsupervised approaches aim to discover the joint probability of (x) instead of E(y|x). Hence, they can be applied in situations where we lack labels, i.e. where we only have explanatory variables (features) and no dependent variable (outcome or label). These approaches are often used to reduce the dimensionality of data. Principal component analysis (PCA) is an unsupervised learning approach familiar to econometricians. These methods can also be applied to pre-define logical groupings of data for subsequent analysis, similar to cluster analysis or to generate an outcome of interest, such as defining the 'topic' of a news article.

As the number of potential descriptors increases, reducing dimensionality becomes more important. Unsupervised learning can also be applied to pre-train neural networks (see below). In these settings, the primary goal is to learn relevant relationships in the unlabelled data that can then be used in a second step for a supervised learning task.

Traditional dimensionality reduction approaches such as PCA rely on linear partitions of the variable space. ML approaches such as Autoencoders facilitate non-linear unsupervised learning. In general, an autoencoder is a NN consisting of an encoder and a decoder function. The encoder aims to map the inputs, x to an internal representation, $h = f(x)$, while the decoder, g, maps the internal representation h, to a reconstructed input $r = g(h)$, where r should be as close as possible to the input x. Usually, restrictions are placed on the function h, such that the autoencoder cannot simply copy x to r. Autoencoders can be interpreted as a non-linear generalisation of PCA (Hinton and Salakhutdinov, 2006).

Typically, autoencoders are simply fully connected neural networks, with the twist that the outputs are their own inputs making them an unsupervised approach. While copying input data to itself is not helpful on its own, restricting the internal layers of the neural net can provide an useful encoding of the data. For example, undercomplete autoencoders set the dimension of h to be lower than the dimension of x. Thus, the autoencoder is forced to learn an internal representation of x in a lower dimensional space. To be successful, the NN needs to be able to compress data with minimal information loss, by capturing only the most important features of x. Regularised autoencoders or denoising autoencoders are alternative specifications, see Goodfellow *et al.* (2016).

2.4. Model complexity versus interpretability

One common objection against the use of ML tools is that they are 'black-boxes' where the relationships learned by the model are not easily interpretable. Even though many ML methods are more complex than their linear regression counterparts, this is not an inherent problem of ML tools but it

rather reflects an unavoidable tradeoff between flexibility and interpretability faced by any method. As soon as we aim to reflect non-linearities, interactions or heterogeneity, model interpretation becomes more difficult. Consider tobit models that add flexibility to a linear regression to model censored observations at the cost that coefficients cannot be interpreted directly, and marginal effects depend on all explanatory variables. Quantile regression or locally weighted regression allow for more flexibility at the cost of complicating interpretability in the sense that the models generate a large number of marginal effects. This trade-off between flexibility versus interpretability also holds for simulation methods. For example, simple analytical economic models might be superior in terms of interpretability compared to, say, a complex computable general equilibrium model. The relevant question regarding interpretability is, therefore, not concerning ML tools versus 'traditional' methods but whether answering a certain research question requires a highly flexible complex model, able to reflect non-linearities, interactions, heterogeneity or dynamics.

While interpretability is fundamental for causal analysis, it can also be helpful for pure prediction tasks. Interpretability is helpful for debugging models or assessing whether the estimated relationships are plausible. Interpretability is also crucial to assess whether ML algorithms are discriminatory, for example when used by banks to determine lending or to give guidance for sentencing to courts (Molnar, 2018).

Numerous methods exist for interpretation, both in ML and econometrics. A good overview is presented by Molnar (2018), from which much of the following discussion is drawn. A more detailed discussion is provided in Appendix A1 (Appendix in supplementary data at *ERAE* online). One of the primary approaches for interpretation is to plot the implicit marginal effects of one or more specific characteristics such as often used for interpreting output from tobit or logit models. Both partial dependence plots (Hastie, Tibshirani and Friedman, 2009: 369) and accumulated local effects plots (Apley, 2016) compare outcomes of one or two variables against their predicted outcomes, whereas individual conditional expectation plots (Goldstein *et al.*, 2015; Molnar, 2018) generate them for an individual observation. Other methods use the model results to simulate marginal effects, as is frequently done with large simulation models for policy analysis. Shapley value explanations do this systematically, estimating the marginal effects by computing predictions drawn from the distributions of the other characteristics with and without the characteristic of interest.

If instead of generating marginal effects, one aims to understand how the model's predictions are correlated with specific inputs, one can use an interpretable model to estimate the relationship between inputs and the predicted outputs from the more complex model. When used for the whole distribution of the data, this is referred to as a global surrogate model. Local interpretable model-agnostic explanations (LIME) focus on understanding the predictions for a single data point (Ribeiro, Singh and Guestrin, 2016).

Another general approach is to determine how much influence each explanatory variable has on the resulting prediction. For tree-based models

the relative importance of predictor variables can be assessed by ranking the importance of different predictors (Hastie, Tibshirani and Friedman, 2009: 367). Fisher, Rudin and Dominici (2018) extended this approach to any other model. Flipping the idea of sensitivity tests on its head, a common approach in ML is to determine the smallest change of an explanatory variable that causes a change to a certain model prediction (Molnar, 2018). Other approaches develop so-called 'adversarial examples', identifying what characteristics of an observation need to change to generate a false prediction. Finally, there are approaches to explore the heterogeneous effects captured by a model by identifying a few representative data points, called 'prototypes' versus rare occurrences, called 'criticisms' using clustering algorithms (see Kim, Khanna and Koyejo, 2016). Then the model's predictions at these prototypes and criticisms are compared to their actual outcomes.

3. What ML can add to the agricultural economics toolbox

We explore the potential of ML by first highlighting specific limitations of current econometric and simulation methods, and identify areas where ML approaches may help fill those gaps. While some ML methods can be used to address multiple limitations, the limitations or challenges themselves differ. We hope that by highlighting multiple situations where ML methods may be useful will facilitate their broader use. As noted above, much of ML is focused on prediction and predictive tasks are highlighted in Sections 3.1–3.3. However, in Section 3.4 we lay out how the prediction capabilities of ML can be useful for causal analysis. Throughout the entire section we highlight current and potential applications of these methods in agricultural and applied economics.

3.1. Restrictive functional forms with little theoretical ground

The choice of model complexity should depend on the phenomenon under study and the specific research question. As noted above, many phenomena in agricultural and environmental economics are inherently non-linear, resulting from underlying biophysical, social or economic processes. For example, the effect of weather variables on yield (Schlenker and Roberts, 2009), groundwater extraction on pumping costs (Burness and Brill, 2001) or health effects of pollution (Graff Zivin and Neidell, 2013) are all likely to contain non-linearities. Other times we are interested in estimating relationships between observations, over time, space or social networks, and our current approaches usually impose some restrictive structure, such as pre-determined neighbours and structure of interaction in spatial econometrics, without strong grounds to justify these assumptions. Often, we are interested in estimating specific aspects of heterogeneity. For example, we may be specifically interested in the distributional effects of an intervention, such as the case of who reduces consumption in response to food warnings (Shimshack, Ward

and Beatty, 2007), or which children benefit from maternal health interventions (Kandpal, 2011). In most current approaches, applied economists estimate average effects or allow the effect to differ across dimensions or between a pre-defined, limited number of groups, or select groups ex-post, with the temptation to cherry pick those groups that conform to the researcher's priors or those that generate significant results. Pre-determining flexible ML processes to identify key dimensions or groups avoids this potential bias, and instead allows the data to determine heterogeneous responses across the population.

Economic theory rarely gives clear guidance about the specific functional form of the object one is trying to estimate. In many settings, it only provides information about shape restrictions such as curvature or monotonicity. Choosing a model that cannot capture non-linearities, interactions or heterogeneity and distributional effects might result in misspecification bias. This misspecification bias increases with the degree of non-linearity of the underlying process (Signorino and Yilmaz, 2003). While we worry a lot about potential endogeneity and think intensively about finding appropriate instruments or natural experiments to minimise potential bias in our estimates of treatment, we are often readily willing to make strong assumptions on functional form that themselves can introduce bias into our estimates.

3.1.1. Current econometric approaches

The current econometric toolbox already provides flexible models but in many cases computational demands limit their applicability for large datasets (large 'N') or high dimensional data (large 'K'). Recent examples of such approaches in our field are random coefficient models (Michler *et al*., 2019), quantile regression models (D'souza and Jolliffe, 2013; Lehn and Bahrs, 2018) or mixture models (Saint-Cyr *et al*., 2019). These approaches allow for more flexibility but still impose restrictive linear assumptions on the estimated relationships. Further, the 'flexible functional forms' advocated and used in demand or supply system estimation are only locally flexible but not across the domain of explanatory variables (Wales, 1977). This limitation considerably restricts their ability to represent heterogeneous responses to changes in the economic environment. Spline models, kernel and locally weighted regression models and GAMs add even more flexibility but their application is usually restricted to a limited number of explanatory variables (see Hastie, Tibshirani and Friedman (2009) for a detailed treatment of these methods, and Cooper, Nam Tran and Wallander (2017), Lence (2009) and Chang and Lin (2016) for applications from our field, Halleck Vega and Elhorst (2015) for flexible parametric specification and McMillen (2012) for semi-parametric approaches in spatial econometrics).

A similar constraint exists concerning numerical Bayesian inference approaches. Specifically, MCMC sampling methods such as Gibbs or Metropolis Hasting are limited in terms of their ability to deal with large

datasets and large numbers of variables (Blei, Kucukelbir and McAuliffe, 2017).

3.1.2. What ML can add

ML models are highly flexible and may be helpful in settings where other flexible models have computational problems due to the size of the dataset or the number of variables we want to consider. We identify three different approaches that are of particular relevance to applied economists: (i) ensembles of trees, particularly gradient boosting approaches, (ii) NNs and (iii) variational inference methods. While the first two approaches are ML methods that are both very flexible and efficient and can be generally applied to a large variety of tasks, variational inference is specifically relevant in the Bayesian context.

Gradient boosted trees (see Section 2.2.2) are emerging as some of the most effective prediction tools in many settings; for example, credit scoring (Lessmann *et al*., 2015; Xia *et al*., 2017) and corporate bankruptcy prediction (Jones, Johnstone and Wilson, 2017). While boosting is primarily used for tree-based approaches, it is not restricted to them. Fenske, Kneib and Hothorn (2011), for example, develop a Bayesian geoadditive quantile regression approach that is estimated with gradient boosting. In agricultural economics, März *et al*. (2016) apply this approach to farmland rental rates. Apart from being very flexible, the approach uses automatic data-driven parameter selection, allowing for different parameters across different quantiles. Their results reveal the existence of important non-linear, heterogeneous relationships between covariates and rental rates. Similarly, Ifft, Kuhns and Patrick (2018), find that these approaches outperform other ML and traditional econometric methods in predicting farmer credit demand.

NNs are also capable of capturing highly non-linear relationships. One important difference between NNs and tree-based methods is that using a NN is complex and usually requires the user to specify more attributes, such as the number of layers and neurons, and more tuning during training. With cross-sectional data, tree-based methods outperform NNs in several benchmark studies (Lessmann *et al*., 2015; Jones, Johnstone and Wilson, 2017). However, compared to tree-based methods, NNs offer more natural ways to deal with non-linear relationships beyond cross sectional data such as time series, panels or spatial data. Through their encoding of the sequential data in the hidden state vector in the case of RNN, or hidden layer in the case of CNN, NNs can uncover more complex non-linear dynamic relationships than the usual AR models. Cao, Ewing and Thompson (2012) find the univariate RNN outperforms the univariate autoregressive integrated moving average (ARIMA) model in the context of wind speed predictions. Recent studies apply a more complex LSTM RNN (Liu, Mi and Li, 2018b) or a RNN using convolution (Liu, Mi and Li, 2018a) for the same task using more complex data. Karlaftis and Vlahogianni (2011) review studies comparing the performance of NN and ARIMA models in the context of transportation

research, and report mixed evidence in terms of the superior performance of NN. However, most of the NN studies in their comparison were written before the recent innovations in DL methods. In our view, ML also holds potential for spatial econometric models. CNNs with their capability of handling 2D grid data seem particularly relevant. The CNN could be trained to detect the extent of the neighbourhood as well as the most relevant features in neighbouring characteristics without the need to pre-define the neighbourhood or the neighbourhood effects.

Other disciplines have actively debated the advantages and disadvantages of more flexible models such as neural networks. In political science, a controversy emerged in early 2000s that compared neural networks to logistic regressions (Beck, King and Zeng, 2000, 2004; de Marchi, Gelpi and Grynaviski, 2004). On one hand, de Marchi et al. (2004) question the superiority of NNs compared to logistic regression models, arguing that models should be as parsimonious as possible, and worrying about overfitting and the interpretability of NNs. Beck, King and Zeng (2004), on the other hand, note that NNs encompass logit models as a special case and argue that controlling for overfitting using a test set is superior to simply assuming that the logit model does not overfit. Most importantly, a logit model might require making unrealistic assumptions. In their context, for example, their restrictive model forces the probability of a conflict to be the same for all countries even though we would expect that effects are heterogeneous, depending non-linearly on variable interactions.

Variational inference (Blei, Kucukelbir and McAuliffe, 2017) is another ML approach that can increase model flexibility by allowing for a larger number of parameters. It can also efficiently deal with larger datasets. The basic idea of variational inference is to approximate complex distributions using more easy-to-compute distributions. It provides an alternative to MCMC sampling approaches, trading accuracy for computational efficiency (Blei, Kucukelbir and McAuliffe, 2017). Athey *et al*. (2018) use variational inference to estimate restaurant demand with a large number of latent variables that reflect unobserved characteristics, which would have challenged traditional methods. Using a similar approach, Ruiz, Athey and Blei (2017) estimate a sequential consumer choice model with latent attribute interaction using highly disaggregated shopping cart data that take into account interactions between individual grocery items.

3.2. Limited ability to extract information from unstructured data

Traditionally, economists work with data that are highly structured (e.g. cross sectional, time-series or panel). Increasingly, unstructured data such as images, text or speech have become available. We loosely define unstructured vs. structured data by distinguishing between data that can be processed in a spreadsheet (structured data) and those that cannot (unstructured data). Our econometric tool kit is only of limited use for the latter. Many ML

advances have been made specifically to derive information or variables (features) from unstructured data (LeCun, Bengio and Hinton, 2015). As such, ML can play an additional role in our discipline as a preprocessing step to derive variables for subsequent analysis using either ML or traditional tools. Most of the ML methods useful for processing unstructured data are also relevant in situations with (too) many explanatory variables or data at a very high temporal or geographical resolution. These cases are considered Section 3.3.

3.2.1. Current approaches

Many unstructured data sources, such as images from remote sensing (Donaldson and Storeygard, 2016), sensor data (Larkin and Hystad, 2017), text data from news (Baker, Bloom and Davis, 2015) or cell phone data (Dong *et al*., 2017) are already intensively used without the use of ML tools. Approaches rely on aggregating the data along hand-crafted features based on domain knowledge. For example, remote sensing data are used to derive a vegetation index (NDVI) (Bradley *et al*., 2007), or single measures such as night light intensity (Blumenstock, 2016; Bruederle and Hodler, 2018). Cell phone records are converted into specific indices (Dong *et al*., 2017; Steele *et al*., 2017). Equally, when working with text data, indices are typically derived based on the number of occurrences of certain terms or phrases (Antweiler and Frank, 2004; Gentzkow and Shapiro, 2010; Saiz and Simonsohn, 2013; Scott and Varian, 2013a,b; Heinz and Swinnen, 2015; Baker, Bloom and Davis, 2015; Baylis, 2015).

3.2.2. What ML can add

ML approaches can play an important role in making information from unstructured data sources available for economic analysis with an algorithmic approach. They can automatically extract the most relevant features for a task, and are potentially capable of deriving more complex features from the raw data missed by hand-crafting. This capability also opens up the opportunity to use novel data sources for economic analysis. For example, recent work uses Google Street View images to predict local demographics (Gebru *et al*., 2017).

We distinguish five different ML approaches to extract features from unstructured data. (i) If most data are labelled, i.e. observations include a dependent variable (outcome or label), end-to-end learning can be applied. If labelled data are scarce, i.e. for most observations we only observe the explanatory variables, (ii) unsupervised pre-training or (iii) transfer learning can be used. (iv) When dealing with more complex data such as networks or trajectories, approaches that automatically create a large number of features based on 'hand-crafted' rules can be applied. (v) We briefly describe uses of ML in text analysis.

3.2.2.1. End-to-end learning. If we have lots of labelled data we can use 'end-to-end learning', i.e. we can train a model using the raw data directly as

an input to predict the final variable of interest. The crucial point here is that we do not rely on hand-crafted features or variables, but let the ML algorithm, usually a DNN, learn to extract useful features from the raw data directly. This approach avoids the loss of information often implied by selection or aggregation in traditional approaches. For example, if we extract lumens per pixel in nightlights data in a hand-crafted approach, we a priori exclude the colour of the pixel or the pattern of lights, both of which might be informative for predicting economic activity. However, for end-to-end learning to work, the algorithm requires vast amounts of labelled data for training.

Authors have used CNN and RNN to derive crop-cover classifications from remotely sensed data (Ienco *et al*., 2017; Kussul *et al*., 2017; Minh *et al*., 2017; Rußwurm and Körner, 2017). The NNs were able to take into account the temporal dimension of the remote sensing data observed over time to allow them to discern between crop types. Rußwurm and Körner (2017) use remote sensing data (Sentinel 2 A images) as an input and a dataset of over 137,000 labelled fields in Bavaria, Germany to identify 19 field classes. But even in settings with less abundant labelled data, approaches close to end-to-end learning approaches might be feasible. You *et al*. (2017), for example, use multispectral remote sensing data to predict US county-level soybean yields. By making weak assumptions on the data generating process, they are able to reduce the dimensions of the input data. Specifically, they assume that the location of pixels in the input images does not matter when predicting average yield in a region. Hence, they convert the images to a histogram, counting the number of pixels with different intensities in the three RGB channels. After this preprocessing, they predict yields without deriving further hand-crafted features (as in Bolton and Friedl, 2013; Johnson, 2014) and thereby preserve the spatio-temporal structure of the data when applying their CNN and RNN models.

3.2.2.2. Unsupervised pre-training. One approach to make use of abundant unlabelled data and limited labelled data, is unsupervised pre-training of DNNs. Hinton, Osindero and Teh (2006) use unsupervised pre-training (or greedy layer wise training) to successfully train the first DNN. The idea is to train each layer of an NN in sequence in an unsupervised fashion. Each layer acts like an autoencoder that aims to map its input to itself while employing some form of regularisation. The model is therefore also called a stacked autoencoder. Once the first layer is trained (i.e. the first autoencoder), the learned encoding is given to the second layer (the second autoencoder), which is then trained and its encoding is given to the next layer. This process continues up to the second last layer whose output can be thought of as a representation of the input data. The last layer is then trained using the labelled data to match this learned representation to the target variable, usually involving only a small number of parameters. Training can stop here or it is possible to refine model parameters of all layers in a final supervised

training step using the labelled data. With this approach it is possible to train most of the parameters of the NN using only unlabelled data.

To help build intuition, compare unsupervised pre-training to using PCA in a binary classification. PCA performs unsupervised dimensionality reduction by mapping a high-dimensional input vector into a lower dimensional representation (or encoding). This encoding might then be used as explanatory variables in a simple (supervised) binary regression. Similarly, with an unsupervised pre-training approach, we train all layers up to the second last layer in an unsupervised way (the stacked autoencoder). The outcome of the second last layer is an encoding of the input variables, like the output of the PCA. The last layer takes this encoding as explanatory variables to be mapped to the dependent variables (i.e the binary target) just as in a simple binary regression. What is different between the two approaches is that PCA is less flexible compared to the stacked autoencoder as a feature extractor (see Section 2.3). Additionally, in an unsupervised pre-training approach we usually do not stop with the autoencoder trained in unsupervised way but rather use the trained weights as starting values for a supervised training in which all weights, including weights from the earlier layers, can be adjusted in a next step; hence the name 'pretraining'. Stacked autoencoder approaches are used in remote sensing (Zhang, Du and Zhang, 2015; Zhou *et al*., 2015; Othman *et al*., 2016; Liang, Shi and Zhang, 2017); Cheng, Han and Lu (2017) and Petersson, Gustafsson and Bergstrom (2016) provide an overview.

3.2.2.3. Transfer learning. An alternative approach to deal with limited labelled data is transfer learning. The general idea is that models and parameters trained in one context can be used in another. Typical applications are image classification or object recognition. Large models like VGG (Simonyan and Zisserman, 2014) or ResNet (He *et al*., 2016) are trained on vast datasets of labelled images (such as ImageNet). These models, along with their trained parameters, can be transferred to other image recognition tasks where only the last layer(s) are trained, or the pretrained parameters are used as starting values. At an intuitive level, even though a model is ultimately trained to distinguish between dogs and cats, the early layers of that network learn how to identify general structures in images such as edges, lines or circles that are also useful for other applications.

The idea of transfer learning is leveraged in several ways by Jean *et al*. (2016) to improve poverty predictions from remote sensing data. Prior work uses nightlight intensity (lumens per pixel) directly to predict poverty and economic activity (Blumenstock, 2016; Bruederle and Hodler, 2018). Jean *et al*. (2016) demonstrate that this does not capture variation in the low end of the poverty scale. They argue that daylight images may provide more information and propose to use transfer learning to combine it with the night light images. They start from a pretrained VGG16 model, using it to predict night time light intensity classes from daytime satellite images. This model is then used as a feature extractor, by removing the last layer of the CNN (the

layer that classified the light intensity). The extracted features (i.e. the outcome of the second last layer) are then used in a ridge regression to predict cluster level expenditures or assets using cluster-level wealth. Like stacked autoencoders (or greedy layer wise pretraining) this approach treats the layers of a NN as higher order representations of underlying raw data. In contrast to the stacked autoencoder approach, however, the model is not trained in an unsupervised fashion but trained to perform a different but related prediction task where labelled data (here night light images) are abundant. The assumption behind the approach is that characteristics in daytime images can be inferred from higher income regions (e.g. roofing material) that have a certain relationship to income or expenditure and that these relationships also extend to poorer regions. This approach can be more efficient than a stacked autoencoder. The latter aims to find representations of the input data that can recover the information in the input data as accurately as possible (i.e. to maintain as much information as possible) while the transfer learning approach aims to extract features that are most suitable to perform a related task, i.e. to maintain only the information relevant for that specific task – in this case, predicting nightlights. Head *et al*. (2017) explores to what extent the approach can be applied to other countries and measures of economic development.

3.2.2.4. 'Brute force' feature engineering. Blumenstock, Cadamuro and On (2015) propose a fourth approach to using complex raw data without the need for hand-crafted features. Their 'brute force' approach to feature engineering uses a deterministic finite automaton that automatically creates a large number of features, with the aim to capture as much variation in the raw data as possible. The created features are then used in a shrinkage regression to select the most promising features. They apply this approach to predict individual level poverty measures using phone record data and show that it outperforms an approach using hand-crafted features.

The basic idea of a deterministic finite automaton is that relatively simple rules (or 'grammar') are defined that specify how individual phone records can be summarised. The algorithm then extracts all features allowed by the grammar. While defining the rules requires more 'hand crafting' in comparison to end-to-end learning, transfer learning or unsupervised pre-training, it seems to hold potential in situations with particularly complex input data such as network data, trajectories, phone records or household level transactional scanner data.

3.2.2.5. ML for text analysis. ML can also improve the analysis of text data. Until recently, text analysis largely used hand-crafted features, but the unstructured nature of the data and the predictive nature of many of the research questions lend themselves to ML. For a recent review of text analysis in economics, see Gentzkow *et al*. (forthcoming) from which we draw much of the following discussion. Other recent reviews of uses of text

analytics exist: Evans and Aceves (2016) in sociology, and Grimmer and Stewart (2013) in political science.

Much text analysis can be grouped into three principal approaches. The first counts words or phrases and then predicts an outcome variable based on the those counts. One fundamental challenge of this approach is that due to the large number existing words/phrases we typically obtain a high dimensional input vector and ML approaches, particularly shrinkage methods, are applied to deal with this problem. For example, Gentzkow, Shapiro and Taddy (2016) measure partisanship in congress by analysing how easy it is to identify the party of a congressman from speech. Second, topic models are an unsupervised learning approach, where topics are unobserved and modelled as a weighted cluster of words or phrases that commonly appear together (see Blei (2012) for an overview). A given text is characterised in terms of a composition of such topics. These models are Bayesian models estimated using variational inference to deal with high dimensionality (see Section 3.1). They have been used to classify industries capturing changes over time, based on companies' product descriptions (Hoberg and Phillips, 2016) or to classify patents (Kelly *et al.*, 2018). Finally, text analysis can also gain from approaches central to ML natural language models used for machine translation speech recognition. One central feature of these approaches is word embeddings that map words and the relationships between them in a high-dimensional vector space. While currently there are only a few examples (Iyyer *et al.*, 2014) that use these approaches in a social science context, Gentzkow, Kelly and Taddy (forthcoming) have concluded that they 'can play a role in the next generation of text-as-data applications'.

3.3. Limited ability to deal with large number of explanatory variables

In many areas, economists have access to large datasets both in terms of the number of observations, N, and frequently with respect to the number of explanatory variables, K. In demand estimation, for example, one may observe a very large number of alternative products, sold at a number of stores, their prices and characteristics, that may affect demand for any one good (Bajari *et al.*, 2015). Similarly, geophysical data such as soil or weather data can include many observed characteristics (wind, temperature, precipitation, evaporation etc.) at highly granular spatial and temporal resolution often with variables misaligned over time and/or space in the sense that they are observed at different spatial or temporal resolution. Typically, economic theory and domain knowledge only provide weak guidance to selecting the specific variables that should be included in the model.

3.3.1. Current econometric approaches

The frequentist econometrics approach to dealing with issues of variable selection is to impose structure to select K, apply a general-to-specific testing

approach that is only feasible with $K < N$ or use model selection criteria such as AIC comparing all possible model combinations, which is only possible for small K. When K is large, and particularly when working with high-resolution data misaligned in space or time, data are typically aggregated by extracting hand-crafted features that are considered relevant, similar to the approach applied to unstructured data (see Section 3.2). The design of such aggregation measures requires specific domain knowledge and a loss of information is inevitable. Often, a mix of hand-crafted collapsing and test driven variable selection is employed. Alternatively, data driven dimensionality reduction techniques such as principal component analysis (PCA) are used. Bayesian variable selection or model averaging approaches are more flexible and theoretically consistent but are not routinely used in the profession. Many ML approaches, such as lasso, can be interpreted as Bayesian variable selection approaches.

3.3.2. What ML can add

ML can be useful in addressing large K problems. These methods are crucial when K exceeds N, but are frequently useful even when $N > K$. Several ML approaches that penalise model complexity such as lasso can be viewed as variable selection techniques (see Section 2.2.1). Other approaches such as trees perform internal variable selection and are well placed to deal with irrelevant explanatory variables. Further, the same approaches described in the previous section to deal with unstructured data can be applied in situations with large K. First, one can apply an unsupervised dimensionality reduction approach such as (stacked) autoencoders for greedy layerwise pre-training or as a feature extractor. For example, Li *et al.* (2016) use autoencoders to provide better air pollution predictions based on sensor data taking into account spatial and temporal dependencies, and avoiding the use of hand-crafted features. Zapana *et al.* (2017) use autoencoders for extracting features to characterise large climatological time series data. Liu *et al.* (2015), Saha, Mitra and Nanjundiah (2016) and Li *et al.* (2018) use autoencoders to derive forecasts of weather, monsoon and water quality, respectively. Autoencoders are also combined with RNNs to capture temporal dynamics and deal with missing observations (Bianchi *et al.*, 2018; Li *et al.*, 2018). Note that the extracted features do not have a direct interpretation (as with the components of a PCA), but can be used to trace back estimated marginal effects in terms of the original input variables.

Unsupervised feature extraction (i.e. dimensionality reduction) has the advantage that it can make use of unlabelled data. The disadvantage of these approaches is that they aim to preserve as much variation of the underlying data but do not take into account that some variation is more relevant than others for a given task. For example, for yield prediction, a certain variation in weather might be irrelevant (e.g. temperature outside of the growing season). End-to-end learning approaches can take into account which variation is most relevant but require that 'sufficient' labelled data are available, where

'sufficient' depends on the dimensions of the input data and the complexity of the problem. RNNs and CNNs are well-placed to handle large K, and are particularly applicable in cases where observations are misaligned in space or time. Intuitively, they also perform a form of dimensionality reduction: in RNNs, by encoding information in the cell state vector and in CNN by encoding information in the hidden layer of the network. In contrast to the unsupervised approach, the NNs do not aim to preserve as much variation as possible but to extract features that are relevant for the supervised prediction task. In our example, it might be sufficient to store the information that temperature is in the suitable range for crop growth but not the exact variation of temperature within that range. One disadvantage of RNNs in this situation is that while their architecture is good at memorising the temporal order of events, they are not well placed to detect at which place a certain event happens (which can be an advantage in other settings such as language modelling). Additionally, even though an RNN can theoretically memorise sequences of arbitrary length, in practice their performance deteriorates quickly once the input sequence becomes too long. Recently, novel CNN structures have been developed that have a much longer effective memory and can handle larger sequence lengths (Kalchbrenner *et al.*, 2016; Gehring *et al.*, 2017; Bai, Zico Kolter and Koltun, 2018). An additional advantage of a CNN structure is that the timing of an event can be more naturally preserved. The model could thus learn that a weather event in the winter has a different effect from one in the spring.

3.4. Causal inference and identification: linearity, lack of appropriate instruments and counterfactuals

The fundamental problem of causal inference is that we do not observe what would have happened to the treated observations without treatment (or the control observations with treatment). In some sense, this can essentially be thought of as a prediction problem where we need to predict the counterfactual.

Most econometric approaches to causal inference presume some structure. For example, difference in differences assumes parallel trends and common shocks have the same effect for treatment and control units. When evaluating a policy change in one region, assume that economic shocks have the same effect across it and other 'control' regions might be unrealistic, potentially biasing the estimate on treatment (Gobillon and Magnac, 2016 for a discussion; see Heckman, Ichimura and Todd, 1997). Current approaches have difficulty dealing with high dimensionality or flexibility, either in instruments or in the counterfactual. For example, difference in differences can fail when existent heterogeneity in the treatment effect is not modelled since the inclusion of location fixed effects will not remove the bias from the average treatment estimate (Gobillon and Magnac, 2016). While machine learning does not relieve the analyst from needing a good identification strategy, it can, for

some approaches, add flexibility to modelling selection or the effect of treatment and better model treatment heterogeneity. But the need for unbiased coefficients on treatment complicates the direct application of ML approaches as they can generate regularisation bias.

3.4.1. Current econometric approaches

There are many approaches for causal inference, and many excellent discussions of them exist (Angrist and Pischke, 2008). Here, we focus on a few approaches where ML has added flexibility. Estimating causal effects is easiest when treatment is exogenous. When treatment is randomised and covariates are balanced across treatment and control groups, one can use a simple average to get an estimate of the average treatment effect. More common is to control for baseline outcomes and use a difference in differences strategy to tighten up the standard errors around the treatment estimate.

When treatment is determined by observables, one can either explicitly model the selection process or match treatment and control on observables that determine treatment. Different versions of matching (nearest neighbour versus propensity score, for example) are simply different ways of collapsing a multi-dimensional object, made up of several matching variables, into a one-dimensional measure of proximity. A variant on the matching approach is the doubly robust regression, where treatment and control observations are first matched, and then the outcome is regressed against the controls and the treatment conducted by using the observations weighted by their propensity score from the matching. This approach is robust against misspecification in either the matching or the regression stage. Another variant of the matching approach is synthetic controls (Abadie, Diamond and Hainmueller, 2010), which match over pre-treatment outcomes, and is useful when one has few treatment units, but longer time series. Thus, the resulting counterfactual a weighted combination of multiple control observations. One constraint is that with many possible control observations, estimating a weight for each control may be problematic.

If treatment selection is based on time-invariant unobservables and one observes the treated observations' pre-treatment, one can simply apply a difference-in-differences approach, with unit fixed effects. One of the tricks with using ML methods in the context of fixed effects is that 'within'-transformations are not consistent in a non-linear setting, and errors are likely to be correlated within observations over time, which can require some modifications to standard ML methods, discussed below.

Last, in the case of endogenous regressors, one frequently uses instruments in two-stage least squares (2SLS). One constraint associated with 2SLS is that it assumes a linear relationship in both the first and second stage, as well as homogeneity of treatment (Hartford *et al.*, 2017). Non-parametric models relax these assumptions (Newey and Powell, 2003; Hall and Horowitz, 2005; Blundell, Chen and Kristensen, 2007; Chen and Pouzo, 2012), however,

these approaches are computationally limited in terms of the size of the dataset or the number of instruments or controls they can consider.

3.4.2. What ML can add

The predictive ability of machine learning in complex and high-dimensional settings can also be used to improve causal estimates. Along with several authors developing the potential of ML for causal inference in economics, discussion of causal analysis is currently emerging from the ML literature itself. Readers interested in a rigorous discussion of structural causal models from an ML community perspective are encouraged to consult Peters, Janzing and Schölkopf (2017). We briefly discuss five general types of ML approach for causal inference introduced in recent years, applicable to different settings defined by the assignment of treatment and the nature of treatment effect, as indicated in brackets:

i. Counterfactual simulation [exogenous treatment]
ii. Double Machine Learning [selection on observables, average effects]
iii. ML for Matching and Panel Methods [selection on observables, unobserved time-invariant characteristics]
iv. Causal forests [selection on observables, heterogeneous effects]
v. ML for IV and Deep IV [endogenous treatment]

3.4.2.1. Counterfactual simulation. Counterfactual simulation basically uses data on pre-treated and control observations to predict what would have happened to exogenously treated observations without treatment. Comparing this prediction to the actual outcome for the treated observations identifies the treatment effect. This general approach is certainly not new in itself, but the excellent predictive capabilities of ML offer considerable advantages with respect to the empirical specification in settings with big data and complex, high-dimensional control-outcome relationships (Varian, 2014: 21). These approaches can be used with randomised treatment or in quasi-experimental settings where treatment assignment is controlled for. For example, Burlig *et al.* (2017) combine panel data methods with lasso to predict a flexible counterfactual of high-frequency school energy consumption from pre-treatment data to estimate the effect of a programme to reduce school energy use.

No new methodology and no volume of data will change the fact that this approach only consistently identifies the treatment effect if treatment has been exogenously assigned to the units of observation. If this is not the case, confounding variables probably affect both the selection into treatment and the outcome variable causing a bias of the estimated treatment effect (see Section 3.4.2.2). Endogenous treatment requires special attention (see Section 3.4.2.5).

3.4.2.2. Double ML (DML). For the case where treatment is assigned based on a complex or non-linear combination of observables, ML may help

flexibly model selection. DML combines the predictive power of ML with an approach to address regularisation bias (Belloni *et al.*, 2016; Chernozhukov *et al.*, 2017, 2018a). Consider the following model where the outcome of interest is the additive effect of treatment plus some non-linear function of covariates (1), and those same covariates non-linearly determine treatment (2):

$$Y = D\theta_0 + g_0(X) + U;\ \ E[U|D, X] = 0 \tag{1}$$

$$D = m_0(X) + V;\ \ E[V|X] = 0 \tag{2}$$

where Y is the outcome variable, D the treatment, θ_0 the marginal treatment effect, and g_0 and m_0 are functions depending on controls, X. The error terms U and V are mean zero conditional on the respective right-hand side variables. A large dimension of X combined with a complex $g_0(X)$ complicates the estimation of $g_0(X)$ with standard econometric estimators, suggesting the employment of flexible ML tools such as lasso, random forests or NN. However, the regularisation bias inherent in such tools would render a naive application and subsequent estimation of θ_0 biased. The idea behind DML is to offset the regularisation bias by stripping out the effect of X from D. In a first step, split the sample, train $\hat{g}_0(X)$ based on equation (1) and $\hat{m}_0(X)$ based on equation (2) using one part of the sample. This step is responsible for the name of the approach. In the second step, regress $Y - \hat{g}_0(X)$ on the orthogonalised D: $\hat{V}D = (D - \widehat{m_0}(X))D$ to obtain $\hat{\theta}_0$ from equation (1) using the main sample. Note that by removing the influence of X on D and subtracting $\hat{g}_0(X)$ from the outcome in the second step removes the regularisation bias. Further, splitting the sample avoids the bias otherwise caused by overfitting (Chernozhukov *et al.*, 2018a: C4–C7). This approach is analogous to a linear instrumental variable (IV) estimator in classical econometrics, where one strips out the non-linear part of Y and includes the estimated residuals from a reduced form regression in the structural equation, and follows the spirit of 'debiased lasso' (Belloni and Chernozhukov, 2013; Belloni, Chernozhukov and Hansen, 2014). The DML approach is very flexible with respect to the ML technique applied in the first step. Chernozhukov *et al.* (2017) use k-fold cross fitting in their illustrative note but any supervised learning approach such as boosted trees, random forests or NNs could be used. DML can also be used to estimate the coefficient of an endogenous variable in a partially linear instrumental variables model, or the local average treatment effect (Chernozhukov *et al.*, 2017). While it holds much potential for settings with non-linear treatment assignment or outcomes that are non-linear functions of observables, note that this method assumes that the treatment effect itself is additive.

3.4.2.3. ML methods for matching and panel methods. When treatment is determined on observables, several authors also use ML approaches for matching to non-linearly control for selection into treatment (Nichols and McBride, 2019). Gradient boosted trees have been used for propensity score

matching in medical research (McCaffrey, Ridgeway and Morral, 2004; Lee, Lessler and Stuart, 2010). Simulated data demonstrate that boosted trees perform particularly well under non-linear and non-additive associations between covariates (Lee, Lessler and Stuart, 2010). Another approach to matching is an ML version of synthetic controls that can face challenges with a large number of potential control observations, which requires the estimation of many weights. Earlier approaches solved this problem by imposing restrictions on the weights, whereas Doudchenko and Imbens (2016) use an elastic net to estimate these weights, since fundamentally this is a prediction problem where control observations are being used to predict pre-trend treatment observations.

Dimension-reduction ML techniques for selection are frequently combined with doubly robust regressions to control for potential error in model specification (Belloni, Chernozhukov and Hansen, 2014; Farrell, 2015). Mullally and Chakravarty (2018) apply this approach to estimate the effect of a groundwater irrigation programme in Nicaragua.

As noted above, when treatment is determined by observables, a standard approach is to use panel methods for identification, setting up a difference in differences framework. Then one can control for time-invariant unobservables that might be correlated with the placement of treatment. Several authors have adapted ML methods for panel settings to allow for dimensionality reduction and more flexible functional forms. Belloni *et al*. (2016) suggest that there are two problems in naively applying regularisation to panel settings. First, they note that the assumption that many coefficients are effectively zero may conflict with the idea that individual heterogeneity is non-zero for most. We may want to allow for non-zero fixed effects for all units in the panel. Second, we generally assume that errors are correlated over time for the same individual, which may affect the number of explanatory variables selected using regularisation. Belloni *et al*. allow for the presence of unrestricted additive individual specific heterogeneity that is partialed out of the model before variable selection occurs using a version of the lasso estimator that allows for a clustered covariance structure. The authors further develop methods for inference that allow for model selection mistakes.

3.4.2.4. Causal forests. Double ML, matching and panel methods are capable of estimating average treatment effects, but we often care about individual responses to targeted interventions. 'Traditional' non-parametric approaches such as nearest-neighbour matching and kernel regression are quickly at their limit with more than a few covariates, use the same distance metric over all covariates, and can be highly sensitive to the addition of covariates, including those that do little to predict outcome. Wager and Athey (2018) introduce causal forests that are able to estimate considerably more complex models given sufficient data. Like random forests, causal forests choose covariates for the weighting depending on their predictive ability, and thus are robust to the addition of uninformative covariates. This approach

builds on Athey and Imbens (2016) who propose a data-driven approach to partitioning data into subgroups for treatment estimation, but takes it beyond group heterogeneity. Causal forests are able to consistently estimate heterogeneous treatment effects under unconfoundedness. Their algorithm grows 'honest' trees, estimating the splits based on one subsample and the treatment effects based on another. Even though Wager and Athey (2018) focus on causal inference, their paper is also the first to provide theoretically proven statistical inference procedures for random forests, which are also useful for generating confidence intervals in pure prediction tasks. In contrast to DML, causal forests are restricted to this specific ML method to control for covariates' influence on outcomes. In the case of randomised treatment, one can use a variety of ML algorithms to identify the most relevant groups over which to choose to estimate heterogeneous treatment effects (Chernozhukov *et al*., 2018b). Athey, Tibshirani and Wager (2019) extend their method of generalised random forests to estimate heterogeneous treatment effects with instrumental variables.

Chernozhukov *et al*. (2018b) apply several ML methods to estimate the heterogeneous effects of a randomised treatment on a microcredit intervention on borrowing, self-employment and consumption. They identify the most and least affected groups and the characteristics associated with them. Carter, Tjernström and Toledo (2019) use generalised random forests to estimate heterogeneous effects of a randomised small business programme in Nicaragua on farmer outcomes and find the largest effects for disadvantaged households. While they find small results overall, those households who were disadvantaged at the baseline benefited much more substantially from the programme, highlighting potential benefits to targeting. Rana and Miller (2019) use causal forests combined with matching to estimate heterogeneous effects of two types of forest management programme in India.

3.4.2.5. ML for IV and Deep IV. Prediction of counterfactual outcomes only identifies policy or treatment effects if predictors are not correlated with the error term, i.e. they are exogenous. Several papers adapt ML techniques for selecting a subset of a large number of instruments to predict the first stage of a linear IV regression. Belloni *et al*. (2012) develop a lasso-based method for estimating the first stage prediction in a linear IV estimation when one has a large number of potential instruments. Bevis and Villa (2017) use this approach to estimate long-run effects of maternal health on child outcomes, where they have a large number of potential instruments from weather outcomes during the mother's early life. Ordonez, Baylis and Ramirez (2018) use this approach to predict adoption of community forest management in Michoacan, Mexico to evaluate its effects on forest outcomes. They have multiple potential instruments from the location and activity or foresters that affect the supply of the community forest management plans.

While these methods address potential problem of selecting from a large number of instruments they still impose linearity on the first stage. Deep IV

(Hartford *et al.*, 2016) is a 2SLS-type approach that uses ML techniques to relax the restrictive linearity and homogeneity assumption of 2SLS and overcomes the computational limitations of non-parametric IV approaches. As with other ML approaches, it also offers an algorithmic approach for variable selection, which may be useful when facing a host of possible instruments.

To understand the concept of Deep IV, consider the structural equation (following Hartford *et al.*, 2016):

$$Y = g(D, X) + U \tag{3}$$

where Y is the outcome variable equal to the sum of a potentially non-linear function $g(D, X)$ and an error U with an unconditional mean zero. The vector of covariates X is exogenous, whereas the policy or treatment variable D is correlated with the error U such that $E[U|D, X] \neq 0$ and is therefore endogenous. If $g(D, X)$ was linear and an instrument was available, one could apply the typical 2SLS estimation. Here, for a counterfactual prediction, we would like to obtain a function predicting $E[Y|D, X]$ such that

$$h(D, X) = g(D, X) + E[U|X], \tag{4}$$

i.e. holding the distribution of U constant as D changes. If we had $h(D, X)$, we could identify the effect of changing policy from D_0 to D_1 through the counterfactual simulation: $h(D_1, X) - h(D_0, X) = g(D_1, X) - g(D_0, X)$. Non-structural ML applied to equation (3), however, would result in a prediction function where the expectation of the error is conditional on the policy variable, and instead obtain $E[Y|D, X] = g(D, X) + E[U|D, X]$, which is not equal to $h(D, X)$. Consequently, a counterfactual simulation with this conditional expectation leads to a biased estimation of the policy effect. The availability of a vector of instruments Z, defined as variables excluded from X, relevant for predicting D and uncorrelated with U, allows us to obtain the prediction function in equation (4). Due to the potential (and likely) non-linearity of $g(D, X)$, the approach is somewhat more involved than 2SLS but follows the same logic. Taking expectation of equation (3) conditional on X and Z gives us

$$E[Y|X, Z] = g(D, X|X, Z) + E[U|X, Z] = \int h(D, X) dF(D|X, Z) \tag{5}$$

where $F(D|X, Z)$ is the conditional distribution of treatment given the covariates and instruments. Hartford *et al.* (2016) suggest obtaining an estimate of $h(D, X)$ by first learning the distribution of treatment $\hat{F}(D|X, Z)$(first stage) and then the expectation of outcome given treatment $\hat{h}(D, X)$ from minimising a quadratic loss function of the difference between Y and the integral, given $\hat{F}(D|X, Z)$ (second stage). Note that if $g(D, X)$ and $F(D|X, Z)$ are linear, we return to the traditional 2SLS approach where the integral disappears and two sequential OLS estimates do the trick. Hartford *et al.* (2017) lay out an ML approach that uses a supervised ML in both the first and second stage.

The first stage estimation approach is a straightforward supervised prediction task where flexible ML tools, such as NN, can be used to predict complex non-linear effects of the instruments and controls on treatment. The second stage is also a supervised ML setting. However, training a NN for this task is more complex as it requires evaluating an integral to derive the gradients of the loss function during training. Hartford *et al.* (2017) propose a stochastic gradient descent approach using MC approximation that can efficiently be applied in a large dataset. The authors stress that the Deep IV approach can be applied using readily available ML techniques without customisation, thereby opening opportunities for applied economics research.

3.5. Limitations of simulation models for policy analysis

3.5.1. Problem statement / Current approaches

Apart from econometric applications, our profession also intensively uses computational simulation models, particularly for policy analysis. Policy-relevant models or modelling systems continue to increase in complexity due to demands like capturing agent heterogeneity or linking economic and bio-physical models. This complexity generates significant computational demands in application and calibration.

As a specific example, consider ABMs that are increasingly used as tools to analyse agricultural and environmental economic issues (Happe, Kellermann and Balmann, 2006; Manson and Evans, 2007; Rasch *et al.*, 2017). Even though they are well suited to analyse dynamic relationships and emergent phenomena arising from complex interactions between individual agents, their regional coverage, the number of agents or the modelled complexity of agents' behaviour is usually limited by–among other reasons–computational constraints. Despite substantial advances in recent years, calibration of ABMs remains a further challenge (Windrum, Fagiolo and Moneta, 2007; Fagiolo *et al.*, 2017).

3.5.2. What ML can add

ML has potential to address both computational demands of complex simulation models and their calibration. In both cases, surrogate modelling, also called meta-modelling or response surface modelling, offers opportunities. A surrogate model approximates the mapping between inputs and outputs of an underlying complex model. The basic aim of the surrogate model is to approximate the behaviour of the underlying model while being computationally cheaper to run. What makes this approach potentially more powerful compared to previous meta-modelling approaches is that the accuracy and dimensionality of the prediction is only restricted by the amount of data generated by the model to be approximated. Other disciplines have started to exploit this approach. Surrogate models are intensively used in engineering (Forrester, Sobester and Keane, 2008; Koziel and Leifsson, 2013), natural science such as water resources modelling (Razavi, Tolson and Burn, 2012)

and weather forecasting (Kim *et al.*, 2015). Established approaches include approximations using polynomial models, radial base function models, kriging, multivariate adaptive regression splines and support vector machines (Forrester, Sobester and Keane, 2008; Kleijnen, 2009). Recently, random forests and NN are also being used (Gong *et al.*, 2015). NN are of particular interest because they can handle multi-output models (Razavi, Tolson and Burn, 2012). Recent advances in ML make the surrogate modelling approach more compelling, for example by using RNN or CNN (e.g. Guo, Li and Iorio, 2016) to handle sequential or spatial data and to reflect model dynamics (see Section 3.3). Appendix A2 provides a more detailed exposition about how surrogate modelling can be beneficial for ABM modelling of farm structural change (Appendix in supplementary data at *ERAE* online).

Surrogate models can also be used for model calibration and are intensively applied for this purpose in water resource modelling (Razavi, Tolson and Burn, 2012; Asher *et al.*, 2015), land surface models (Gong *et al.*, 2015), building-energy demand (Nagpal *et al.*, 2018) and material science (Mareš, Janouchová and Kučerová, 2016). Similarly, they are also used for sensitivity analysis for complex models of physical systems (Tripathy and Bilionis, 2018). The basic idea of using surrogate models for calibration is that in a first step, a surrogate model is trained on a sample of simulated model outcomes and then in a second step, a calibration is performed based on that surrogate model to find the parameter values that most closely match the empirically observed data. This approach still requires a relatively large number of runs of the underlying model to generate the sample to train the surrogate model. To alleviate that problem, approaches such as adaptive sampling (Wang *et al.*, 2014; Xiao, Zuo and Zhou, 2018) or iterative calibration are available (Lamperti, Roventini and Sani, 2017). A challenge related to both issues is the choice of an appropriate loss function used to compare model outcomes with surrogate model outcomes or observed characteristics, particularly for dynamic models (Barde, 2017; Guerini and Moneta, 2017; Lamperti, 2018).

There may be potential for improving calibration by leveraging ideas from Generative Adversarial Nets (GANs) (Goodfellow *et al.*, 2014). GANs train a generator, such as for images, together with a discriminator model. The generator aims to learn to generate images that are similar to actual images while the discriminator aims to learn how to efficiently distinguish between generated images and actual images. Feeding discriminator outcomes back to the generator improve its performance in an iterative approach. In the context of model calibration, the model generator could explore in which way to tune the parameters of the model such that the generated output data is as close as possible to the observed data, while the discriminator is trained to distinguish generated from observed data. The advantage of such an approach would be that no criteria for comparison would need to be specified a priori, with the discriminator learning itself which features are most useful for detecting generated data; while the generator would aim to mimic the observed data as closely as possible. Such an approach could, in principle, be applied for

calibration where we aim to generate observations that closely replicate observed inputs and outputs, as well as for surrogate modelling where we aim to generate observations from the surrogate model that are as close as possible to the output generated by the true model.

Further connections between theory-based simulations models and ML approaches are discussed in physics (de Bezenac, Pajot and Gallinari, 2017; Karpatne *et al*., 2017). They propose hybrid models where outcomes from theoretically based simulation models are used to initialise or pre-train ML models, to complete missing observational data or for statistical downscaling of simulation model results.

4. What economists can add to ML

Novel ML approaches have led to important breakthroughs and many disciplines are exploring the potential of ML, including economics. One central challenge facing ML is to unite data-driven ML methods with the amassed theoretical disciplinary knowledge (Karpatne *et al*., 2017). Why is this relevant? Why are purely data driven models not sufficient (as argued by Anderson *et al*., 2008)? Despite the increase in data availability, in many applications, we still face a shortage of data and their labels. An example is Blumenstock, Cadamuro and On (2015), mentioned above, who, despite of having billions of phone records, can only link them to less than 900 survey respondents for whom labelled data are available. Even with lots of data, the information contained in the data might be insufficient for prediction or identification, for example when dealing with rare events, when the variation in the outcome variable is small, or if outcomes are very noisy. Even 'big data' might be insufficient when dealing with highly complex processes and non-stationary patterns that change dynamically, such as in climate science (Karpatne *et al*., 2017). In all of these settings, the risk of picking up spurious correlations and discovering relationships that do not generalise is high. The Google flu prediction is an example in this respect (Lazer *et al*., 2014).

Theoretical knowledge can help with these data challenges in two ways. First, theoretical domain knowledge is necessary to understand why a model works and if it has learned plausible relationships. For this, models need to be interpretable (see Section 2.5). Understanding why a model works is also crucial to assess when it will stop working. Second, incorporating theoretical knowledge can increase the efficiency of ML approaches (see Section 3.1), particularly in the described settings where the information in the data is limited and processes are complex.

Karpatne *et al*. (2017: 2) call for a 'novel paradigm that uses the unique capability of data science models to automatically learn patterns and models from large data, without ignoring the treasure of accumulated scientific knowledge'. In this respect, econometrics has a natural role to play, as an approach that uses statistical methods and combines them with theoretical knowledge to answer economic questions. The development of approaches to include theoretical or prior knowledge in novel ML approaches is a relatively

young research field, with several contributions from climate and material science (Faghmous and Kumar, 2014; Faghmous *et al.*, 2014; Ganguly *et al.*, 2014; Wagner and Rondinelli, 2016; Karpatne *et al.*, 2017; Sheikh and Jahirabadkar, 2018). Economists should closely follow and contribute to these developments.

Another set of problems are issues surrounding the data themselves. While novel data sources hold exciting potential, they often come with issues of selection bias. For example, cell phone data are only available for those with access to cell phones; the quality of labels may vary by country or region. Economists are trained to think about these selection problems and theoretical knowledge is useful to assess their importance and to handle them. Finally, the scarcity of labelled (ground truth) data versus the abundance of unlabelled data often constrains the usefulness of 'big' data for economics. As noted above (Section 3.2), unsupervised learning approaches can help to some extent but transfer learning approaches, as in Jean *et al.* (2016), which take into account theoretical understanding of the underlying processes, might be more efficient.

5. Frontier

We conclude by highlighting a few current developments in ML that are particularly relevant for agricultural and applied economics. As noted in the causal section (3.4), applying ML methods to causal analysis is a new and growing field. Thinking carefully about how to control for unobservables in highly non-linear settings and comparing ML to traditional identification methods are all areas that are ripe for investigation.

The imposition of structural information when training ML models may improve their predictive performance. As discussed in Section 4, methods are available that allow the use of disciplinary knowledge when training ML models. Economic theory often provides information on the curvature of behavioural functions (production frontiers, profit functions) or the sign of marginal effects. Such additional structural information may especially help in situations with limited data availability and complex interactive relationships between features.

The combination of supervised and unsupervised approaches to improve predictions (and thereby also causal analysis) seems to be currently underexplored in applied economic analysis (see Section 3.2). The potential to combine high resolution biophysical data with limited amounts of labelled economic data may offer many additional opportunities to enrich our models. Questions such as estimating land use choices driven by climate change, or estimating nutrient emissions over space could significantly benefit from allowing for more complexity in the biophysical components of our models. A better understanding about the performance of the ML methods in this context could also inspire targeted data collection of labelled data for this purpose.

The recent engagement of economists with ML tools is generating increased attention to the derivation of statistical properties of ML estimators, which is crucial for appropriate statistical inference in the field. But new approaches in probabilistic programming developed within the ML community (Tran *et al*., 2017; Bingham *et al*., 2019) with clear Bayesian interpretations of model outcomes may offer a natural way of combining ML with procedures for statistical inference. An interesting promise of probabilistic programming is to move from the case-specific development of variational inference procedures to a generic approach only requiring the specification of a probabilistic economic model from which one can generate a random sample (Ghahramani, 2015).

Recent ML advances also hold potential for simulation models. First, apart from supervised and unsupervised learning, reinforcement learning approaches comprise a third class of ML algorithms. In reinforcement learning the aim is to learn optimal behaviour facing a reactive environment. These approaches recently gained popularity due to, among others, their successful application in playing the board game 'Go' (Silver *et al*., 2016). Reinforcement approaches are relevant in situations where a function can be specified that provides a reward for a chosen action in a given situation. The algorithm learns by choosing different actions and observing the associated rewards. As such, reinforcement learning is an optimisation approach. However, they are particularly well suited for sequential setups where agents take multiple actions in sequence and previous actions influence the outcome of following actions and feedback is not instantaneous but delayed. They can also handle an uncertain environment with outcomes that are not deterministic. Reinforcement learning is increasingly used in game-theoretical settings but with limited policy relevance so far (Fudenberg and Levine, 2016; Chen, 2017). Further development may have potential for models with learning agents in more descriptive, policy relevant models where, for example, agents make optimal strategic choices learning from their own experience and information provided by their environment (networks). Second, GANs (Section 3.5) pose an interesting opportunity to calibrate simulations models to available data without having to select, a priori, specific, limited features of the data to calibrate to. The interplay between generator and discriminator algorithms would allow the approach to learn what features matter in distinguishing model outcomes from observations and to exploit complex data structures for this purpose. Knowing the purpose of the simulation model may help restrict the calibration approach for targeted performance, but the GANs might make restricted choices of features less ad hoc and the resulting simulation model more generally valid.

Last, a new and active area of ML research facilitates the distributed training of models on multiple datasets, where these datasets do not need to be shared. Given machine learning's powerful abilities to derive information from data, merely removing personal identifiers has been shown to be insufficient to preserve participants' identities. Further, data breaches are becoming more common, raising concerns for academics collecting or analysing

confidential data. Privacy-preserving machine learning may be important to economists in the future, both to allow for the use of confidential data and to facilitate collaboration.

In summary, machine learning methods already have demonstrated great potential in improving prediction and computational power in economic analysis. The next few years will undoubtedly see more of these tools tailored and applied to economics. While it may be difficult to keep up with all of the advances as they appear, we hope that this article gives readers an entry-point with which to start to engage these exciting methods (Appendix A3, in supplementary data at *ERAE* online, provides additional hints on how to get started).

Acknowledgements

Thanks go to an anonymous reviewer, Patrick Baylis, Robert Brunner, Svetlana Fedoseeva and the participants at the workshop in environmental economics and data science for their comments and suggestions. Hugo Storm and Thomas Heckelei acknowledge support from the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob

Funding

The research for this publication by Hugo Storm is funded by the Deutsche Forschungsgemeinschaft under grant no. STO 1087/1-1. The research contributed by Kathy Baylis was in parts funded by the USDA Hatch project number ILLU-470-333.

Supplementary data

Supplementary data are available at *European Review of Agricultural Economics* online.

References

Abadie, A., Diamond, A. and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* 105: 493–505.

Anderson, C., Allain, R., Niiler, E., Barber, G., Gonzalez, R., Dreyfuss, E. and Klarreich, E. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired*. Magazine from June 23, 2008, www.wired.com/2008/06/pb-theory/.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance* 59: 1259–1294.

Apley, D. W. (2016). Visualizing the effects of predictor variables in black box supervised learning models. arXiv:1612.08468 [stat.ME].

Asher, M. J., Croke, B. F. W., Jakeman, A. J. and Peeters, L. J. M. (2015). A review of surrogate models and their application to groundwater modeling. *Water Resources Research* 51: 5957–5973.

Athey, S. (2019). The impact of machine learning on economics. In: Agrawal, A., Gans, J., and Goldfarb, A. (eds.), *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press (p. 507–547).

Athey, S., Blei, D., Donnelly, R., Ruiz, F. and Schmidt, T. (2018). Estimating heterogeneous consumer preferences for restaurants and travel time using mobile location data. *AEA Papers and Proceedings* 108, 64–67.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America* 113: 7353–7360.

Athey, S., Tibshirani, J. and Wager, S. (2019). Generalized random forests. The Annals of Statistics 47(2):1148–1178.

Bai, S., Zico Kolter, J. and Koltun, V. (2018). Convolutional sequence modeling revisited. Invited paper at the 6th International Conference on Learning Representations, April 30 - May 3, 2018, Vancouver, BC, Canada.

Bajari, P., Nekipelov, D., Ryan, S. P. and Yang, M. (2015). Machine learning methods for demand estimation. *The American Economic Review* 105: 481–485.

Baker, S. R., Bloom, N. and Davis, S. J. (2015). Measuring economic policy uncertainty. Working Paper Series. https://doi.org/10.3386/w21633

Barde, S. (2017). A practical, accurate, information criterion for nth order Markov processes. *Computational Economics* 50: 281–324.

Baylis, P. (2015). Temperature and temperament: evidence from a billion tweets. Energy Institute Working Paper.

Beck, N., King, G. and Zeng, L. (2000). Improving quantitative studies of international conflict: a conjecture. *The American Political Science Review* 94: 21–35.

Beck, N., King, G. and Zeng, L. (2004). Theory and evidence in international conflict: a response to de Marchi, Gelpi, and Grynaviski. *The American Political Science Review* 98: 379–389.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica: Journal of the Econometric Society* 80: 2369–2429.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* 19: 521–547.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives: A Journal of the American Economic Association* 28: 29–50.

Belloni, A., Chernozhukov, V., Hansen, C. and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics* 34: 590–605.

Bevis, L. E. M. and Villa, K. (2017). Intergenerational transmission of mother-to-child health: evidence from Cebu, the Philippines, Working paper. https://sites.tufts.edu/neudc2017/files/2017/11/Intergen_Bevis_Villa.pdf, last access August 1, 2019..

Bianchi, F. M., Livi, L., Mikalsen, K. Ø., Kampffmeyer, M. and Jenssen, R. (2018). Learning representations for multivariate time series with missing data using Temporal Kernelized Autoencoders. arXiv:1805.03473 [cs.NE].

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P. and Goodman, N. D. (2019). Pyro: deep universal probabilistic programming. *Journal of Machine Learning Research* 20:1–6.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* 55: 77–84.
Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association* 112: 859–877.
Blumenstock, J. E. (2016). Fighting poverty with data. *Science* 353: 753–754.
Blumenstock, J., Cadamuro, G. and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science* 350: 1073–1076.
Blundell, R., Chen, X. and Kristensen, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica: Journal of the Econometric Society* 75: 1613–1669.
Bolton, D. K. and Friedl, M. A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology* 173: 74–84.
Bradley, B. A., Jacob, R. W., Hermance, J. F. and Mustard, J. F. (2007). A curve fitting procedure to derive inter-annual phenologies from time series of noisy satellite NDVI data. *Remote Sensing of Environment* 106: 137–145.
Bruederle, A. and Hodler, R. (2018). Nighttime lights as a proxy for human development at the local level. *PLoS ONE* 13: e0202231.
Burlig, F., Knittel, C., Rapson, D., Reguant, M. and Wolfram, C. (2017). Machine learning from schools about energy efficiency. Working Paper Series. https://doi.org/10.3386/w23908
Burness, H. S. and Brill, T. C. (2001). The role for policy in common pool groundwater use. *Resource and Energy Economics* 23: 19–40.
Cao, Q., Ewing, B. T. and Thompson, M. A. (2012). Forecasting wind speed with recurrent neural networks. *European Journal of Operational Research* 221: 148–154.
Carter, M. R., Tjernström, E. and Toledo, P. (2019). Heterogeneous impact dynamics of a rural business development program in Nicaragua. *Journal of Development Economics* 138: 77–98.
Chang, H.-H. and Lin, T.-C. (2016). Does the minimum lot size program affect farmland values? Empirical evidence using administrative data and regression discontinuity design in Taiwan. *American Journal of Agricultural Economics* 98: 785–801.
Chen, S.-H. (2017). *Agent-Based Computational Economics: How the Idea Originated and Where It Is Going*. Routledge, New York.
Chen, X. and Pouzo, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica: Journal of the Econometric Society* 80: 277–321.
Cheng, G., Han, J. and Lu, X. (2017). Remote sensing image scene classification: benchmark and state of the art. *Proceedings of the IEEE* 105: 1865–1883.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *The American Economic Review* 107: 261–265.
Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters. *Economic Journal* 21: C1–C68.
Chernozhukov, V., Demirer, M., Duflo, E. and Fernández-Val, I. (2018b). Generic machine learning inference on heterogenous treatment effects in randomized experiments. Working Paper Series. https://doi.org/10.3386/w24678
Coble, K. H., Mishra, A. K., Ferrell, S. and Griffin, T. (2018). Big Data in agriculture: a challenge for the future. *Applied Economic Perspectives and Policy* 40: 79–96.

Cooper, J., Nam Tran, A. and Wallander, S. (2017). Testing for specification bias with a flexible Fourier transform model for crop yields. *American Journal of Agricultural Economics* 99: 800–817.

Crane-Droesch, A. (2017). Technology diffusion, outcome variability, and social learning: evidence from a field experiment in Kenya. *American Journal of Agricultural Economics* 100: 955–974.

de Bezenac, E., Pajot, A. and Gallinari, P. (2017). Deep learning for physical processes: incorporating prior scientific knowledge. arXiv:1711.07970 [cs.AI].

de Marchi, S., Gelpi, C. and Grynaviski, J. D. (2004). Untangling neural nets. *The American Political Science Review* 98: 371–378.

Donaldson, D. and Storeygard, A. (2016). The view from above: applications of satellite data in economics. *The Journal of Economic Perspectives: A Journal of the American Economic Association* 30: 171–198.

Dong, L., Chen, S., Cheng, Y., Wu, Z., Li, C. and Wu, H. (2017). Measuring economic activity in China with mobile big data. EPJ Data. *Science* 6: 29.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: a synthesis. Working Paper Series. https://doi.org/10.3386/w22791.

D'souza, A. and Jolliffe, D. (2013). Food insecurity in vulnerable populations: coping with food price shocks in Afghanistan. *American Journal of Agricultural Economics* 96: 790–812.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Einav, L. and Levin, J. (2014). Economics in the age of big data. *Science* 346: 1243089.

Evans, J. A. and Aceves, P. (2016). Machine translation: mining text for social theory. *Annual Review of Sociology* 42: 21–50.

Faghmous, J. H., Banerjee, A., Shekhar, S., Steinbach, M., Kumar, V., Ganguly, A. R. and Samatova, N. (2014). Theory-guided data science for climate change. *Computer* 47: 74–78.

Faghmous, J. H. and Kumar, V. (2014). A Big Data guide to understanding climate change: the case for theory-guided data science. *Big Data* 2: 155–163.

Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A. and Roventini, A. (2017). Validation of agent-based models in economics and finance. LEM Working Paper Series.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 189: 1–23.

Fenske, N., Kneib, T. and Hothorn, T. (2011). Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association* 106: 494–510.

Fisher, A., Rudin, C. and Dominici, F. (2018). All models are wrong but many are useful: variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. arXiv:1801.01489 [stat.ME].

Forrester, A., Sobester, A. and Keane, A. (2008). *Engineering Design via Surrogate Modelling: A Practical Guide*. John Wiley & Sons.

Fudenberg, D. and Levine, D. K. (2016). Whither game theory? Towards a theory of learning in games. *The Journal of Economic Perspectives: A Journal of the American Economic Association* 30: 151–170.

Ganguly, A. R., Kodra, E. A., Agrawal, A., Banerjee, A., Boriah, S., Chatterjee, S., Chatterjee, S., Choudhary, A., Das, D., Faghmous, J., Ganguli, P., Ghosh, S., Hayhoe, K., Hays, C., Hendrix, W., Fu, Q., Kawale, J., Kumar, D., Kumar, V., Liao, W., Liess, S., Mawalagedara, R., Mithal, V., Oglesby, R., Salvi, K., Snyder, P. K., Steinhaeuser,

K., Wang, D. and Wuebbles, D. (2014). Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques. *Nonlinear Processes in Geophysics* 21: 777–795.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L. and Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. *Proceedings of the National Academy of Sciences of the United States of America* 114: 13108–13113.

Gehring, J., Auli, M., Grangier, D., Yarats, D. and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. arXiv:1705.03122 [cs.CL].

Gentzkow, M., Kelly, B. T. and Taddy, M., forthcoming. Text as data. *Journal of Economic Literature*.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? Evidence from U.S. daily newspapers. *Econometrica: Journal of the Econometric Society* 78: 35–71.

Gentzkow, M., Shapiro, J. M. and Taddy, M. (2016). Measuring polarization in high-dimensional data: method and application to congressional speech. www.nber.org/papers/w22423, Working Paper Series. https://doi.org/10.3386/w22423

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature* 521: 452–459.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: interactive fixed effects and synthetic controls . *The Review of Economics and Statistics* 98: 535–551.

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2015). Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* 24: 44–65.

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Dai, Y., Ye, A. and Miao, C. (2015). Multi-objective parameter optimization of common land model using adaptive surrogate modeling. *Hydrology and Earth System Sciences* 19: 2409–2425.

Goodfellow, I., Bengio, Y., Courville, A. and Bengio, Y. (2016). *Deep Learning*. Cambridge: MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (eds), *Advances in Neural Information Processing Systems, Vol. 27*. Curran Associates, Inc. 2672–2680.

Graff Zivin, J. and Neidell, M. (2013). Environment, health, and human capital. *Journal of Economic Literature* 51: 689–730.

Grimmer, J. (2015). We are all social scientists now: how big data, machine learning, and causal inference work together. *PS Political Science Political* 48: 80–83.

Grimmer, J. and Stewart, B. M. (2013). Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21: 267–297.

Guerini, M. and Moneta, A. (2017). A method for agent-based models validation. *Journal of Economic Dynamics & Control* 82: 125–141.

Guo, X., Li, W. and Iorio, F. (2016). Convolutional neural networks for steady flow approximation. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. ACM, New York, NY, USA, pp. 481–490.

Hall, P. and Horowitz, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics* 33: 2904–2929.

Halleck Vega, S. and Elhorst, J. P. (2015). The SLX model. *Journal of Regional Science* 55: 339–363.

Happe, K., Kellermann, K. and Balmann, A. (2006). Agent-based analysis of agricultural policies: an illustration of the agricultural policy simulator AgriPoliS, its adaptation and behavior. *Ecology and Society* 11(1): 49.

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. arXiv:1612.09596 [stat.AP].

Hartford, J., Lewis, G., Leyton-Brown, K. and Taddy, M. (2017). Deep IV: a flexible approach for counterfactual prediction. In: D. Precup, Y. W. Teh (eds), Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, International Convention Centre, Sydney, Australia, pp. 1414–1423.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer Science & Business Media.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 770–778.

Head, A., Manguin, M., Tran, N. and Blumenstock, J. E. (2017). Can human development be measured with satellite imagery? In: Proceedings of the Ninth International Conference on Information and Communication Technologies and Development. ACM, p. 8.

Heckman, J. J., Ichimura, H. and Todd, P. E. (1997). Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *The Review of Economic Studies* 64: 605–654.

Heinz, M. and Swinnen, J. (2015). Media slant in economic news: a factor 20. *Economics Letters* 132: 18–20.

Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation* 18: 1527–1554.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313: 504–507.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *The Journal of Political Economy* 124: 1423–1465.

Ienco, D., Gaetano, R., Dupaquier, C. and Maurel, P. (2017). Land cover classification via multitemporal spatial data by deep recurrent neural networks. *IEEE Geoscience and Remote Sensing Letters* 14: 1685–1689.

Ifft, J., Kuhns, R. and Patrick, K. (2018). Can machine learning improve prediction – an application with farm survey data. *International Food and Agribusiness Management Review* 1–16.

Iyyer, M., Enns, P., Boyd-Graber, J. and Resnik, P. (2014). Political ideology detection using recursive neural networks. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Presented at the Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1113–1122.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science* 353: 790–794.

Johnson, D. M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Enivornment* 141: 116–128.

Jones, S., Johnstone, D. and Wilson, R. (2017). Predicting corporate bankruptcy: an evaluation of alternative statistical frameworks. *Journal of Business Finance & Accounting* 44: 3–34.

Kalchbrenner, N., Espeholt, L., Simonyan, K., van den Oord, A., Graves, A. and Kavukcuoglu, K. (2016). Neural machine translation in linear time. arXiv:1610.10099 [cs.CL].

Kamilaris, A. and Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: a survey. *Computers and Electronics in Agriculture* 147: 70–90.

Kandpal, E. (2011). Beyond average treatment effects: Distribution of child nutrition outcomes and program placement in India's ICDS. *World Development* 39: 1410–1421.

Karlaftis, M. G. and Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies* 19: 387–399.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. and Kumar, V. (2017). Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering* 29: 2318–2331.

Kelly, B., Papanikolaou, D., Seru, A. and Taddy, M. (2018). *Measuring Technological Innovation over the Long Run*. Cambridge, MA: National Bureau of Economic Research, https://doi.org/10.3386/w25266.

Kim, B., Khanna, R. and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. In: D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett (eds), *Advances in Neural Information Processing Systems, Vol. 29. Curran Associates, Inc.* 2280–2288.

Kim, S.-W., Melby, J. A., Nadal-Caraballo, N. C. and Ratcliff, J. (2015). A time-dependent surrogate model for storm surge prediction based on an artificial neural network using high-fidelity synthetic hurricane modeling. *National Hazards* 76: 565–585.

Kleijnen, J. P. C. (2009). Kriging metamodeling in simulation: a review. *European Journal of Operational Research* 192: 707–716.

Koziel, S. and Leifsson, L. (2013). Surrogate-based modeling and optimization. Applications in Engineering, Springer, New York Heidelberg Dordrecht London.

Kussul, N., Lavreniuk, M., Skakun, S. and Shelestov, A. (2017). Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters* 14: 778–782.

Lamperti, F. (2018). An information theoretic criterion for empirical validation of simulation models. *Econometrics and Statistics* 5: 83–106.

Lamperti, F., Roventini, A. and Sani, A. (2017). Agent-based model calibration using machine learning surrogates. LEM Working Paper. https://doi.org/10.2139/ssrn.2943297

Larkin, A. and Hystad, P. (2017). Towards personal exposures: how technology is changing air pollution and health research. *Current Environmental Health Reports/ Statistics Canada, Canadian Centre for Health Information: Rapports sur la Sante / Statistique Canada, Centre Canadien d'information sur la Sante* 4, 463–471.

Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science* 343: 1203–1205.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature* 521: 436–444.

Lee, B. K., Lessler, J. and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* 29: 337–346.

Lehn, F. and Bahrs, E. (2018). Quantile regression of German standard farmland values: do the impacts of determinants vary across the conditional distribution? *Journal of Agricultural and Applied Economics* 50: 453–477.

Lence, S. H. (2009). Do futures benefit farmers? *American Journal of Agricultural Economics* 91: 154–167.

Lessmann, S., Baesens, B., Seow, H.-V. and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *European Journal of Operational Research* 247: 124–136.

Li, X., Peng, L., Hu, Y., Shao, J. and Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research International* 23: 22408–22417.

Li, Z., Peng, F., Niu, B., Li, G., Wu, J. and Miao, Z. (2018). Water quality prediction model combining sparse auto-encoder and LSTM network. *IFAC-PapersOnLine* 51: 831–836.

Liang, P., Shi, W. and Zhang, X. (2017). Remote sensing image classification based on stacked denoising autoencoder. *Remote Sensing* 10: 16.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association* 101: 578–590.

Liu, J. N. K., Hu, Y., He, Y., Chan, P. W. and Lai, L. (2015). Deep neural network modeling for big data weather forecasting. In: W. Pedrycz and S.-M. Chen (eds), *Information Granularity, Big Data, and Computational Intelligence*. Chamonix: Springer International Publishing, 389–408.

Liu, H., Mi, X. and Li, Y. (2018a). Smart deep learning based wind speed prediction model using wavelet packet decomposition, convolutional neural network and convolutional long short term memory network. *Energy Conversion Management.* 166: 120–131.

Liu, H., Mi, X.-W. and Li, Y.-F. (2018b). Wind speed forecasting method based on deep learning strategy using empirical wavelet transform, long short term memory neural network and Elman neural network. *Energy Conversion Management* 156: 498–514.

Manson, S. M. and Evans, T. (2007). Agent-based modeling of deforestation in southern Yucatan, Mexico, and reforestation in the Midwest United States. *Proceedings of the National Academy of Sciences of the United States of America* 104: 20678–20683.

Mareš, T., Janouchová, E. and Kučerová, A. (2016). Artificial neural networks in the calibration of nonlinear mechanical models. *Advances in Engineering Software* 95: 68–81.

McCaffrey, D. F., Ridgeway, G. and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* 9: 403–425.

McMillen, D. P. (2012). Perspectives on spatial econometrics: linear smoothing with structured models. *Journal of Regional Science* 52: 192–209.

Michler, J. D., Tjernström, E., Verkaart, S. and Mausch, K. (2019). Money matters: the role of yields and profits in agricultural technology adoption. *American Journal of Agricultural Economics* 101(3): 710–731.

Minh, D. H. T., Ienco, D., Gaetano, R., Lalande, N., Ndikumana, E., Osman, F. and Maurel, P. (2017). Deep recurrent neural networks for mapping winter vegetation quality coverage via multi-temporal SAR Sentinel-1. arXiv [cs.CV].

Molnar, C. (2018). *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*, ebook, Leanpub, n.p.

Monroe, B. L., Pan, J., Roberts, M. E., Sen, M. and Sinclair, B. (2015). No! formal theory, causal inference, and big data are not contradictory trends in political science. *PS Political Science Politic* 48: 71–74.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *The Journal of Economic Perspectives: A Journal of the American Economic Association* 31: 87–106.

Mullally, C. and Chakravarty, S. (2018). Are matching funds for smallholder irrigation money well spent? *Food Policy* 76: 70–80.

März, A., Klein, N., Kneib, T. and Musshoff, O. (2016). Analysing farmland rental rates using Bayesian geoadditive quantile regression. *European Review of Agricultural Economics* 43: 663–698.

Nagpal, S., Mueller, C., Aijazi, A. and Reinhart, C. F. (2018). A methodology for auto-calibrating urban building energy models using surrogate modeling techniques. *Journal of Building Performance Simulation* 1–16.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica: Journal of the Econometric Society* 71: 1565–1578.

Nichols, A. and L. McBride. (2019). Propensity scores and causal inference using machine learning methods. Presentation in the Track session "Machine Learning in Applied Economics" at the annual meeting of the Agicultural and Applied Economics Association (AAEA), Atlanta, July 21–23.

Ordonez, P., Baylis, K. and Ramirez, I. (2018). Factors that affect the management of common pool resources: the case of community forest management in Michoac, Mexico International Association of Agricultural Economists (IAAE) 2018 Conference, July 28-August 2, 2018, Vancouver, British Columbia.

Othman, E., Bazi, Y., Alajlan, N., Alhichri, H. and Melgani, F. (2016). Using convolutional features and a sparse autoencoder for land-use scene classification. *International Journal of Remote Sensing* 37: 2149–2167.

Peters, J., Janzing, D. and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.

Petersson, H., Gustafsson, D. and Bergstrom, D. (2016). Hyperspectral image analysis using deep learning – a review. In: 2016, 6th International Conference on Image Processing Theory Tools and Applications (IPTA), IEEE, pp. 1–6.

Raj, M. P., Swaminarayan, P. R., Saini, J. R. and Parmar, D. K. (2015). Applications of pattern recognition algorithms in agriculture: a review. *International Journal of Advanced Networking and Applications* 6: 2495.

Rana, P. and Miller, D. C. (2019). Machine learning to analyze the social-ecological impacts of natural resource policy: insights from community forest management in the Indian Himalaya. *Environmental Research Letters* 14: 1–12.

Rasch, S., Heckelei, T., Storm, H., Oomen, R. and Naumann, C. (2017). Multi-scale resilience of a communal rangeland system in South Africa. *Ecological Economics : The Journal of the International Society for Ecological Economics* 131: 129–138.

Razavi, S., Tolson, B. A. and Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research* 48: 559.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv:1606.05386 [stat.ML].

Ruiz, F. J. R., Athey, S. and Blei, D. M. (2017). SHOPPER: a probabilistic model of consumer choice with substitutes and complements. arXiv:1711.03560 [stat.ML].

Rußwurm, M. and Körner, M. (2017). Multi-temporal land cover classification with long short-term memory neural networks. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-1/W1: 551–558.

Saha, M., Mitra, P. and Nanjundiah, R. S. (2016). Autoencoder-based identification of predictors of Indian monsoon. *Meteorology and Atmospheric Physics* 128: 613–628.

Saint-Cyr, L. D. F., Storm, H., Heckelei, T. and Piet, L. (2019). Heterogeneous impacts of neighbouring farm size on the decision to exit: evidence from Brittany. *European Review of Agricultural Economics*, 46:237–266.

Saiz, A. and Simonsohn, U. (2013). Proxying For unobservable variables with Internet document-frequency. *Journal of the European Economics Association* 11: 137–165.

Sarle, W. S. (1994). *Neural Networks and Statistical Models*, Proceedings of the Nineteenth Annual SAS Users Group International Conference, April, 1994.

Schlenker, W. and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences of the United States of America* 106: 15594–15598.

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks: The Official Journal of the International Neural Network Society* 61: 85–117.

Scott, S. L. and Varian, H. R. (2013a). Predicting the present with Bayesian structural time series. Available at SSRN 2304426. https://doi.org/10.2139/ssrn.2304426

Scott, S. L. and Varian, H. R. (2013b). Bayesian variable selection for nowcasting economic time series. NBER Working Papers 19567.

Sheikh, R. and Jahirabadkar, S. (2018). An insight into theory-guided climate data science – a literature review. *Advances in Data and Information Sciences*. Singapore: Springer, 115–125.

Shekhar, S., Schnable, P., Le Bauer, D., Baylis, K. and Waal, K. V. (2017). Agriculture Big Data (AgBD) challenges and opportunities from farm to table: a Midwest Big Data Hub Community Whitepaper. White Paper for the US National Institute of Food and Agriculture.

Shimshack, J. P., Ward, M. B. and Beatty, T. K. M. (2007). Mercury advisories: Information, education, and fish consumption. *Journal of Environmental Economics and Management* 53: 158–179.

Signorino, C. S. and Yilmaz, K. (2003). Strategic misspecification in regression models. *American Journal of Political Science* 47: 551–566.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* 529: 484–489.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs.CV].

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J. and Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of the Royal Society, Interface/The Royal Society* 14, https://doi.org/10.1098/rsif.2016.0690.

Tibshirani, R., Wainwright, M. and Hastie, T. (2015). *Statistical Learning With Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC.

Tran, D., Hoffman, M. D., Saurous, R. A., Brevdo, E., Murphy, K. and Blei, D. M. (2017). Deep probabilistic programming. arXiv:1701.03757 [stat.ML].

Tripathy, R. and Bilionis, I., (2018). Deep UQ: learning deep neural network surrogate models for high dimensional uncertainty quantification, *Journal of Computational Physics* 375:565–588.

Varian, H. R. (2014). Big Data: new tricks for econometrics. *The Journal of Economic Perspectives: A Journal of the American Economic Association* 28: 3–28.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113: 1228–1242.

Wagner, N. and Rondinelli, J. M. (2016). Theory-guided machine learning in materials science.. *Frontiers in Materials* 3: 2271.

Wales, T. J. (1977). On the flexibility of flexible functional forms: an empirical approach. *Journal of Econometrics* 5: 183–193.

Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z. and Miao, C. (2014). An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environmental Modelling & Software* 60: 167–179.

Windrum, P., Fagiolo, G. and Moneta, A. (2007). Empirical validation of agent-based models: alternatives and prospects. *Journal of Artificial Societies and Social Simulation* 10: 8.

Xia, Y., Liu, C., Li, Y. and Liu, N. (2017). A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications* 78: 225–241.

Xiao, N.-C., Zuo, M. J. and Zhou, C. (2018). A new adaptive sequential sampling method to construct surrogate models for efficient reliability analysis. *Reliability Engineering & System Safety* 169: 330–338.

You, J., Li, X., Low, M., Lobell, D. and Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17).

Zapana, R. A., del Alamo, C. L., Quenaya, J. F. L. and Valdivia, A. M. C., 2017. Characterization of climatological time series using autoencoders. In: 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI). ieeexplore.ieee.org: 1–6.

Zhang, F., Du, B. and Zhang, L. (2015). Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing: A Publication of the IEEE Geoscience and Remote Sensing Society* 53: 2175–2184.

Zhou, W., Shao, Z., Diao, C. and Cheng, Q. (2015). High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder. *Remote Sensing Letters* 6: 775–783.