Chen *et al. J. Chin. Sociol.* (2021) 8:15
https://doi.org/10.1186/s40711-021-00152-z

The Journal of Chinese Sociology

RESEARCH Open Access

Social prediction: a new research paradigm based on machine learning

Yunsong Chen[1*], Xiaogang Wu[2], Anning Hu[3], Guangye He[1] and Guodong Ju[4]

*Correspondence:
yunsong.chen@nju.edu.cn
[1] Department of Sociology, Nanjing University, Nanjing, China
Full list of author information is available at the end of the article

Abstract

Sociology is a science concerned with both the interpretive understanding of social action and the corresponding causal explanation, process, and result. A causal explanation should be the foundation of prediction. For many years, due to data and computing power constraints, quantitative research in social science has primarily focused on statistical tests to analyze correlations and causality, leaving predictions largely ignored. By sorting out the historical context of "social prediction," this article redefines this concept by introducing why and how machine learning can help prediction in a scientific way. Furthermore, this article summarizes the academic value and governance value of social prediction and suggests that it is a potential breakthrough in the contemporary social research paradigm. We believe that through machine learning, we can witness the advent of an era of a paradigm shift from correlation and causality to social prediction. This shift will provide a rare opportunity for sociology in China to become the international frontier of computational social sciences and accelerate the construction of philosophy and social science with Chinese characteristics.

Keywords: Social prediction, Machine learning, Research paradigm, Quantitative research methods, Computational social sciences

Introduction

To Weber, sociology is a science that concerns itself with the interpretive understanding of social action as well as the causal explanation of its process and result (Weber 1968[1921]:4). Thus, sociologists have considered it their duty to describe and explain social processes and phenomena, seek the meaning and interpretation of society and social actions, or verify social hypotheses and theories. Consequently, description, interpretation, and statistical testing are traditional and mainstream approaches in sociological research.

In contrast, the prediction seems to have had no role within the discipline for a long time. This phenomenon also widely exists in social sciences such as economics, political science, and social policy, which has not drawn concern or reflection. However, scholars have already emphasized that a causal interpretation must serve as the "basis for predicting social phenomena" (Hempel and Oppenheim 1948:138). Hence, predictability is a necessary but not sufficient condition for the causal mechanism, as although predictability does not equate with causality, a prediction could be formed if there was causality.

Based on this rationale, the prediction should be the proper pursuit of sociological research since sociology looks for causality. In 2014, Duncan Watts published an article in the *American Journal of Sociology* to criticize the sociological tradition of pursuing the meaning of "common sense" and ignoring the value of prediction (Watts 2014). Watts argued that if sociology is a science, sociologists' interpretations must be evaluated by scientific standards; that is, the science must be able to make predictions.

Over the past century, sociologists have explored the meaning and explanation of society and social processes and devoted attention to providing theoretical guidance and evaluative examinations for social transformations. However, sociology and even the broader category of social science pay much less attention to ex ante prediction than ex post facto evaluation and interpretation (Hofman et al. 2017). A series of papers on the topic of "prediction and its limits" have been published by *Science* (Jasny and Stone 2017). Although these articles came from different disciplines of social sciences such as economics, sociology, and political science, scholars' consensus was that in contrast to natural science, theories and data of social science were rarely used to make predictions. Martin et al. (2016) highlighted that the significant complexity of the social system and the limited information (the lack of data and models) have made social science research unable to make predictions over a long period.[1]

In the history of social science research, scientific practices for predicting social phenomena or processes have remained absent. However, the concept of "social prediction"[2] has always lingered in the minds of scholars. As early as the 1940s, Kaplan (1940) proposed the concept of "social prediction," emphasizing that social science should predict social phenomena. In the early stage of reconstructing the sociology discipline in China after the reform and opening-up in the 1980s, Chinese scholars also proposed this concept (Yan 1986). However, in both international and domestic academia, substantive prediction research has not been fully developed in social science or aroused academic resonance due to the limitations of data availability and econometric methods. The long-term absence of social prediction in empirical research impedes quantitative researchers as they seek to exert scientific power in influencing policy and promoting the discourse power of media. Compared with scholars who emphasize theoretical criticism and interpretation, quantitative researchers are even more cautious, conservative, and uninteresting due to their limited capacity for predicting and anticipating. Indeed, the public and governments would not be satisfied merely with concept refinement, process interpretation, and statistical estimation.

With the availability of multisource data and the advance of computer processing performance, social prediction has dawned on quantitative social science research, which is at the forefront of interdisciplinary research. Currently, quantitative research in social science has gradually met the three requirements for high-precision prediction: data, computing power, and algorithms. In particular, with the development of computer algorithms for specific data (Athey 2018), social scientists have been able to process

[1] For example, the transformations in the Soviet Union and Eastern Europe are called the "Black Friday of social science." These historical events are such extremely rare cases that it is impossible to establish a dataset with sufficient historical samples. Therefore, scientific prediction based on data cannot be carried out.

[2] The society involved in "social prediction" in this article pertains to social science in general; the term is not limited to sociology.

large-scale social data, highlighting the academic value of social prediction (Hofman et al. 2017). The special issue of *Science* in 2017, which was devoted to predicting the social process, preliminarily sorted out this important developmental trend.

The rest of the article will review the historical development of the concept of social prediction and discuss the specific principles and methods of social prediction via algorithms. On this basis, the article will redefine social prediction in the contemporary era. In this sense, this article summarizes the disciplinary significance of social prediction from the perspective of policy significance and academic significance and especially aims to present the main contribution fields and direction of social prediction for contemporary social research. Furthermore, this article discusses the paradigm significance of social prediction. We suggest that social prediction represents a new subparadigm of social science research from the methodological perspective rather than strict ontology. With the emergence of big data and the improvement of computing power, the application of machine learning and the redefining of social prediction will boost the paradigm breakthrough of quantitative research in sociology and even social science research. The present moment may also be an important opportunity for Chinese sociology, especially quantitative social research, to catch up with international academia, which is of great significance for accelerating the construction of philosophy and social sciences with Chinese characteristics.

The history of social prediction

The rise and fall of prediction: early concepts and dilemma

The prediction has involved logical and sequential consideration of natural and social phenomena since ancient times (Goodman 1955). However, it is a modern development that prediction has become a scientific category. The prediction of social phenomena and human behavior has long been compared with the prediction of natural phenomena or animal behavior but has been regarded as a more difficult task. As early as in the 1940s, scholars suggested strengthening the work of prediction in social science and proposed the concept of social prediction (Kaplan 1940). Although Kaplan (1940) realized the difficulty of prediction, he believed that social behavior was even more predictable than natural phenomena at the microlevel.

However, since the mid-twentieth century, the development of social science has failed to make a substantial breakthrough in prediction. When Kaplan put forward the concept of prediction in social science, he predicted that it would be very difficult. In particular, there were four obstacles to be overcome: first, many influencing factors would be neglected; second, the sequence in the prediction may not be taken into account; third, it is difficult to capture accurate information on known variables; fourth, the complex association among social variables would be neglected (Kaplan 1940). Generally, these problems correspond to missing variables, logical chains, measurement error, and the complexity of social phenomena in terms of quantitative research in contemporary social sciences. Another development is the path model, which was popular in the 1970s and 1980s. The path model attempts to expose the implications of data and the complex associations among variables by including dozens of factors in the model. The method of variable packing is similar to the data-mining model of machine learning,

but the purpose and implementation of the path model are quite different from those of machine learning.

In the Chinese academia of the humanities and social sciences, some scholars have proposed the concept of social prediction and carried out exploration based on early warning indicators of social stability; some of this work was done as early as the 1980s (Yan 1986, 2005). In the *Sociology Dictionary* edited by Deng (2009), "social prediction" is defined as "speculation or analysis of possible social phenomena in the future… its purpose is to reveal the most important factors that determine the future development and the most important relationship of social phenomena, for decision-making reference." The dictionary even distinguishes "intuitive prediction" based on subjective experience and "quantitative prediction" based on data. Since the 1990s, scholars of the Chinese Academy of Social Sciences, namely, Xin Ru, Xueyi Lu, Peilin Li, Guangjin Chen, and Yi Zhang, have used the term "prediction" in a series of *Blue Books on Social Analysis and Prediction*. However, the related research methods are mainly data displays on time trends, intuitive trend prediction, and traditional regression. In this sense, empirical prediction with scientific standards based on data is almost blank, although the concept of social prediction has been used for a long time. Indeed, elaborations of its significance and difficulties are available in both international and Chinese academic publications.

The reasons for the poor development of social prediction are not complicated. Even in the twenty-first century, social scientists still cannot solve the problems put forward by Kaplan (1940). Due to the great complexity of social processes, one requires sufficient data, complex models, and strong computer processing capacity to make an ideally accurate and precise social prediction, which are the three dimensions of information limitations mentioned in the introduction. Due to these three bottlenecks, social scientists can hardly practice prediction. Instead, these scientists have shelved the ambition of prediction since the second half of the twentieth century. Since the data, models, and computing power required for accurate prediction were not ready, the whole social science community, especially quantitative research, has focused on the relationships among variables; that is, based on limited sample data, unbiased estimations of the pairwise relationships between variables are obtained by statistical models. The delicate implication of this compromise is that social scientists no longer have unrealistic confidence in social prediction as the early pioneers did. Instead, these scientists deliberately confined their work to analyze the pairwise relationships between variables (Hofman et al. 2017; Athey 2018). In short, because of the bottlenecks of data, models, and computing power, sociologists have abandoned the scientific practice of prediction.

Revisiting prediction: correlation, causality, and prediction

The abandonment of prediction has shaped the mainstream methods of quantitative research in the whole social science community. The primary efforts are spent determining the covariance between the independent variable X and the dependent variable Y through regression models. In other words, social scientists analyze whether the relationship between X and Y is statistically significant and has actual significance in economic and social scales based on certain statistical standards. They clarify whether and how the change of a unit of X is related to the change of Y. This correlation analysis, which focuses on covariance, cannot meet the ultimate academic mission of causal

explanation. Thus, social scientists have started to use the analytical approach of parameter estimation to go beyond correlation analysis and reach causal inference. With the development of advanced econometric methods and the cautious application of experimental methods, some disciplines, such as economics, have shifted their mainstream quantitative analysis almost completely to causal inference based on the "counterfactual" framework in recent years, that is, they explore whether and how a change in X causes a change in Y through the observed social data (Pearl 2000; Rubin 1974). Accordingly, a similar disciplinary evolution took place in sociology at the end of the twentieth century. The development of other disciplines, such as statistics and econometrics, provides ready-made methods and analytical models for causal inference in quantitative analysis. International sociologists have contacted and introduced appropriate methods at an early stage (Morgan and Winship 2007; Brand and Xie 2007). However, the mission of sociology also requires sociologists to realize Weber's statement of sociology by providing causal explanations of social processes and consequences (Weber 1968[1921]:4). In Chinese sociology, relevant model identification strategies have also been introduced and popularized (Chen 2012; Hu 2012).

Can sociologists provide a satisfactory response to Weber's emphasis on explanatory mechanisms and causality when the relationship between social index X and social index Y can be identified? Although sociologists are still upset and worried about the invasion of advanced econometric methods into the discipline, scholars at the forefront of the discipline have given a clear and determined answer that the covariant research of correlation analysis and causal analysis is not enough to constitute a sociological explanation in terms of science. Duncan Watts pointed out the paradigm crisis of sociological research, namely, overreliance on common sense. Many sociological explanations confuse understandability with causality, which does not meet the standard of scientific interpretation. If sociologists expect their interpretations to be scientifically legitimate, they must evaluate them according to scientific standards; that is, predictions must be made (Watts 2014:313). To make sociology more scientific, he stressed that it is necessary to sacrifice some seemingly straightforward and reasonable views. Even sociologists have to make the right and necessary choices between satisfying but unscientific stories and unsatisfying scientific explanations.

We ask how Watts' criticism can be comprehended with respect to traditional research methods of quantitative sociology. Watts directly targeted Weber, the pioneer of sociology. Weber argued that sociology is a science that provides an interpretative understanding of social action ("understanding" translates "verstehen" in German), and thus, it provides a causal explanation of its process and outcome (Weber 1968[1921]:4). However, Watts believes that interpretative understanding and causal interpretation are two different things. If the explanation provided by sociologists was causal, it could certainly be used to make prior predictions. However, the interpretive understanding only needs to be reasonable, and it can be just ex post. In other words, Watts emphasizes that predictability is a necessary but not sufficient condition for the establishment of a causal mechanism and is the most powerful means to verify the mechanism principle. However, sociologists tend to ignore the prediction or improvement of the accuracy of prediction. Instead, sociologists emphasize that prediction is not equal to causality, the generality of complex models is not strong, and it is difficult to improve the insight of

unexplained models. However, these are just irrelevant remarks that distract readers' attention. The key point is that although predictability does not necessarily equate with causality, causality can be adopted to make predictions.

Watts' criticism almost targets the general approach of all empirical sociological research, but his exposition of the relationship between causality and prediction is clear and accurate and conforms to the classic concept of causality (Hempel and Oppenheim 1948; Manski 2007). In a sense, Watts' view is a symbol of breaking between sociology as a discipline of social science and sociology as a traditional humanities discipline. Although the dichotomy between scientific and humanistic approaches to sociology has long existed with different characteristics and within different fields (Chen 2017), Watts' argument puts forward more stringent scientific academic standards for empirical research in sociology.

In summary, prediction is the main component of realizing the scientific research goal of quantitative sociology. Among correlation, causality, and prediction, correlation is the prerequisite for causality and prediction, while causality is a sufficient but unnecessary condition for prediction. If there was causality, an event must have been predictable, but not vice versa. Causality and prediction are both probabilistic at the social level, and they are both powerful approaches to comprehending social affairs. These concepts share the same foundation and same direction, and both are valued by empirical social science. The problem is that the discipline has been in a state of under-preparation for prediction for many years, although quantitative social science research has developed rather mature methods and models for correlation and causal analysis. Specifically, when the research goal is to provide a mechanical explanation of X with respect to Y, the traditional method to identify causality is sufficient. However, if the research goal was to predict the probabilities and intensities of certain phenomena in the future based on the existing data and methods (which is often the practical requirement of the social governance process for sociologists), traditional tools cannot achieve this task. Thus, the social prediction must be put on the agenda.

The realization of social prediction

With the development of machine learning, the emergence of large-scale social survey data and big data, and improved computer processing performance, the era of solving problems of data, algorithms, and computing power and realizing social prediction has dawned. From the methodological perspective, we argue that the main approach to achieving social prediction is machine learning.

The definition of machine learning

What is machine learning? Susan Athey of Stanford University, who was the first female economist to win the Clark Prize, defined machine learning in the context of social science such that its goal is to achieve tasks such as clustering, classification, and prediction by developing computer algorithms that are suitable for specific data (Athey 2018). More precisely, machine learning tries to constantly optimize the performance standards of statistical calculation programs to realize the task of learning, discover data

characteristics, and make statistical predictions based on a large number of data eigenvalues.[3] Generally, according to whether the target feature label has been given in the data set, machine learning can be divided into supervised learning and unsupervised learning, which refer to prediction and classification clustering tasks, respectively. Most of the predictions we are concerned about are derived from supervised learning (Li 2012).

Specifically, supervised learning could provide the predicted label results by testing the fitting effect of the given training set model and applying the model to the test set in the dataset with given features and labels. In the quantitative terminology of social science, when the covariate matrix (i.e., character X) and the explained variable (i.e., marker Y) have been given, an appropriate algorithm would be adapted to fit the conditional expectation of the corresponding covariate, and the fitting effect would be evaluated against the real value of the explained variable. This process requires that the original label (Y) provided by the dataset be objective and accurate, and the training set must be consistent with the covariate standard of the test set; that is, the data set has to be independent and identically distributed (IID).[4]

We take image recognition, which is the classic task in machine learning, as an example. To enable the computer to identify a car in a given picture, it is necessary for people to manually judge and mark some images first. If there were a car in the picture, it would be marked as 1. Otherwise, it would be marked as 0 to generate a training set with a "standard answer." "Whether there is a car in the picture" is the explained variable Y, and the covariant group X can be obtained by quantifying the image pixel information. By training the algorithm and adjusting the parameters, the properties of the explained variables can be predicted by using the covariate matrix information, and the prediction can be ideally accurate. Then, the algorithm obtained by training can be applied to the image library without manual marking to realize automatic recognition. For specific methods, supervised learning methods include regularized regression, regression trees, random forest, support vector machines, neural networks, naive Bayes classification, and ensemble learning.

Unsupervised learning is suitable for datasets without labels. When the covariate matrix is available and the explained variable is not, the algorithm will calculate the distance between different samples according to the given covariate information and cluster the samples. This kind of method is essentially a process of dimensionality reduction, which can be applied to unstructured data such as text, pictures, audio, and video and can expand the scale of the empirical data that are available in social science. For example, in image recognition, the algorithm directly processes the unlabeled image data set, calculates the similarity or difference degree of different pictures through the image pixel matrix data, and then makes the classification according to the principle of "minimum

[3] Machine learning and data mining are often used as equivalent concepts in many disciplines and academic practice. Comparatively, whereas machine learning emphasizes that computer programs apply existing data information to new research objects, which is the so-called "learning," data mining emphasizes the extraction and simplification of data features.

[4] The IID condition requires the training set and the test set to have the same probability distribution and to be independent of each other, so that the model obtained from the training set data can be applied to the whole dataset and the training effect can be guaranteed. To ensure the generalization capacity of the algorithm obtained from the training set, it is typical to set up multiple groups of random numbers to divide the training set or to use cross validation in practice.

intragroup distance and maximum intergroup distance." The interpretation of the category meaning is determined and defined manually. The common methods of unsupervised learning include K-means clustering, topic modeling, and community detection. Among them, latent Dirichlet allocation and other thematic modeling tools (Blei et al. 2003) have been widely used in cultural sociology. *Poetics*, which is a top journal of cultural sociology, published a series of studies based on thematic models in a special issue in 2013 (Mohr and Bogdanov 2013). In Chinese sociological academia, Ronggui Huang (2017) has also used this method to explore the topics with which labor is concerned.

Predictive principle of supervised learning

There are many specific methods for supervised learning, but the general model-fitting goal is quite different from the traditional model regression. The former category of methods aims at accuracy, that is, the difference between the prediction label and the real label is the smallest, while the latter type of procedure is meant to evaluate the impact of an independent variable on the dependent variable under the premise of controlling other variables (Athey 2018). Among numerous supervised learning algorithms, regularized regression based on linear models is widely used. Compared with the least squares model (OLS), the regularized regression model adds a penalty term to the regression coefficient. Specifically, the unbiased estimation of the OLS regression coefficient β is

$$\hat{\beta}_{\mathrm{OLS}} = \operatorname{argmin}(y - X\beta)'(y - X\beta) = (X'X)^{-1}X'y$$

The regression coefficient of the regularized regression is

$$\hat{\beta}_{\mathrm{OLS}} = \operatorname{argmin}\left\{(y - X\beta)'(y - X\beta) + \lambda\left[(1-\alpha)\beta'\beta + \alpha|\beta|\right]\right\}$$

when λ is 0, the penalty term is 0, which refers to an unbiased OLS regression; when λ is not 0 and α is 0, an L2 regularizer (∑P 1β2 J) is added to the parameter, where the term refers to ridge regression. When λ is not 0 and α is equal to 1, an L1 regularizer (∑P $1|\beta_j|$) is added to the parameter, where the term refers to Lasso regression. The other cases are all elastic net regression. Therefore, ridge regression and Lasso regression can also be regarded as special cases of elastic net regression.

Here, we describe how regularized regression can obtain more accurate predictions than OLS. Specifically, the error of a linear fitting of the model can be divided into three parts: bias, variance, and irreducible error. These errors represent the deviation between the fitting expectation and the real value, the dispersion of the fitting value, and the inevitable system noise, respectively. The deviation of OLS is constantly equal to 0 due to its inherent least sum of squares of residual error, while the regularized regression model reduces the variance and overall error by introducing bias to improve the prediction accuracy of the model (Athey and Imbens 2016). The modeling of machine learning does not consider theory much. Generally, the inclusion of more variables helps to increase the prediction accuracy. Therefore, machine learning models can include seemingly unrelated variables and sacrifice the theoretical nature of the model. In general, there is a bias-variance tradeoff between introducing bias to improve the fitting accuracy of the

model and continuing unbiased estimation to rely on previous theories, which directly reflects the difference between machine learning and traditional quantitative methods.

In addition to regularized regression based on linear models, other supervised learning methods have their own merits.[5] The input x is divided into many tree regions by the regression tree method, and then each output y is generated. Each node, i.e., each "leaf," corresponds to a prediction. The regression tree method divides the input X into many tree-like regions and generates the output Y in which each node (or leaf) is a prediction. With ample branches in the regression tree, an accurate prediction of the whole sample can be generated. A neural network is a "black box" algorithm designed to simulate biological neural systems. The input–output layer and hidden layer of the algorithm are composed of several simple cells at the same level, and multiple groups of interactive hidden layers construct the whole neural network. By increasing the number of hidden layers to conduct layer-by-layer training and adding convolution, the stability and accuracy of the learning effect of the algorithm can be improved. A support vector machine (SVM) based on the VC dimension (Vapnik–Chervonenkis dimension) obtains the maximum-margin hyperplane to achieve binary classification. The nonlinear classification task can also be realized by the kernel method. Bayesian classification adopts the classic Bayesian School of Statistics to classify samples by maximizing the prior probability. Ensemble learning integrates multiple learning results to obtain a more comprehensive, stable, and strong supervised model. Furthermore, the bagging method reduces the variance of the classifications by multiple return sampling, and the boosting method uses the previous classification error to modify the weight of the subsequent classification to optimize the classification. Readers can find more information in relevant references (such as Mitchell 1997, Li 2012).

Redefinition of social prediction: the perspective of social computing and machine learning

Based on machine learning, we try to define social prediction in the context of contemporary social science. Thus, social prediction uses spatiotemporal data to demonstrate social phenomena or processes; machine learning based on appropriate algorithms can accurately quantify and measure unknown information to provide information and a basis for social decision-making and research. Machine learning includes the vertical prediction of the future based on historical data and the horizontal prediction of other data based on sample data. We argue that social prediction is an important part of computational sociology. Computational sociology is a new field of quantitative sociology that describes, explains, and predicts complex social phenomena and processes based on complex models and social computing tools.[6] This field's research methods of social computing include social network analysis, simulation modeling, machine learning, and

[5] In fact, numerous algorithms have been developed by machine learning. This article focuses on the regularized regression method based on a linear model, since this method is the most similar to the traditional linear regression in social science. The method of exchanging deviation for precision can demonstrate the different functional orientations of machine learning and traditional quantitative models of sociology, economics and political science, and the predictive advantages of machine learning.

[6] In November 2019, Qiu Zeqi, Liang Yucheng, Chen Yunsong, Sun Xiulin, Hu Anning and Chen Huashan gave the basic definition of "computational sociology" for the first time at the preparatory meeting of the academic committee of computational sociology and the 2019 academic seminar.

big data analysis. Among these methods, the combination of big data, machine learning, and social prediction has special advantages. While the massive background information of observation objects provided by big data offers considerable convenience for model training, big data can provide outliers on a large scale. With the help of machine learning analysis technology, these outliers may contribute to theoretical innovation and policy implementation.

The disciplinary significance of social prediction

Since social prediction can be realized through machine learning, what is its main value for the development of social science, especially sociology as a new research field and research method? Based on our understanding of sociology and some of the latest literature, we classify the disciplinary value of social prediction into three dimensions: academic significance, governance significance, and discourse significance. Then, we analyze the contributions and limitations of machine learning to sociology.

The academic significance of social prediction

Because machine learning can process complex and diverse information content for social science and generate variable forms for analysis, it expands the research horizon of social science. In brief, social prediction based on machine learning has disciplinary significance for social science research in the following five respects.

First, the latent indicators of interest can be obtained through prediction. There are two types of data that are difficult to obtain through traditional surveys or big data in social science research. One type is the "subjective latent indicator," which often occurs because people are unwilling to disclose true individual information. The reasons include the sensitivity of the problem itself or social acceptance, such as personal unemployment, sexual orientation, venereal diseases, and religious beliefs. In the specific economic, social, and cultural context, this information is often deliberately hidden by respondents. At the social level, concealment makes researchers or social governance officials unable to obtain comprehensive, true, and accurate data about this kind of information. The other problem is "objective latent indicators," that is, there are objectively complex data measurements or heterogeneous group classification indicators that are difficult to detect directly. These latent indicators can be found by machine learning, which will provide new dependent variables or independent variables for academic research.

For the subjective latent indicators, as long as some people in our data provide these indicators truly and accurately, a social prediction based on machine learning can use the input data as a training set to accurately predict the people who are not willing to provide information or provide distorted information (in a sense, the result of this process can be regarded as a supplement to the missing values). The prediction accuracy depends on the size and independence of the samples and the optimization of the algorithm models. He et al. (2018) used a Baidu search to predict the regional distribution data of AIDS in China and employed the dynamic panel pooled mean group model based on the heterogeneity hypothesis because the prediction accuracy would be reduced with the dynamic generalized moment model. This model could make a more convincing and credible prediction when uncertainty was resolved by machine learning based on larger

samples. For the objective latent indicators, unsupervised learning (UML) is often used in variable generation. For example, in economics, unsupervised learning is used to analyze satellite images and generate measurements of data indicators of forest harvesting, environmental pollution, and night lighting (Donaldson and Storeygard 2016). Studies in sociology classify and analyze government documents (Mohr et al. 2013) and academic texts (McFarland et al. 2013). In addition, social network research with unsupervised learning has also attracted the attention of scholars.

Second, theoretical hypotheses can be generated via prediction. In traditional quantitative methods, the essence of the model is to include new main explanatory variables to test new theoretical hypotheses. In addition to theoretical intuition, whether to add variables to the model from a statistical perspective depends on stepwise regression, partial least squares, and AIC and BIC standard comparison. Scholars have summarized 21 varieties of traditional variable selection methods (Castle et al. 2009). However, with the machine learning method, some brand-new means could be adopted to examine and expand the influencing factors of the model. We can enhance the sociological imagination by discovering new explanatory variables and new explanatory dimensions and obtaining a new theoretical hypothesis. This process is exactly consistent with the idea of "bring theories back" advocated by big data analysts (Luo et al. 2018).

For the explanatory variables, if the effect of an independent variable X on the dependent variable Y is evaluated, Varian (2014) proposes using the same machine-learning algorithm to fit and predict Y when the variable is included and excluded and comparing the difference between the two fitting effects. If the fitting effect of the model containing X is better, the covariant correlation or even causal relationship between X and Y can be hypothesized theoretically, and the hypothesis can be verified by traditional quantitative methods. At the level of the explanatory dimension (i.e., a set of explanatory variables that are highly correlated conceptually and logically), machine learning can provide an important driving force for sociological imagination and realize the "precision difference analysis of grouped variables" method. Specifically, the data can be labeled in advance, and all the variables in the data are combined and packaged into their respective explanatory dimensions without ready-made theoretical guidance, which is ultimately unified into the fitting process of machine learning. The prediction results of the same algorithm with and without an explanatory dimension are compared one by one. Thus, we obtain the predictive ability of an explanatory dimension for the dependent variable. Once a new explanatory dimension has a good ability to predict Y, we can examine the specific variables of this dimension and excavate the most likely explanatory variable from it based on imagination and theory. In addition, the overall explanatory power or relevance of a new explanatory dimension to the dependent variables may trigger new sociological thinking and even inspire new theories and hypotheses, that is, attempts to discover the overall influence of the new dimension.

Third, prediction helps to conduct a causal inference. In social science, the counterfactual framework, which defines the causal mechanism, is essentially speculation and simulation of the nonreal world. Thus, when a certain influence is not applied or a certain processing factor is not changed, one asks what form the event will take. This problem is the precise problem at which machine learning is effective: constructing the state of an event that does not exist as accurately as possible with limited data (Athey 2015).

Therefore, many studies have tried to apply machine learning methods to causal inference problems, especially in the counterfactual construction process and extensions of the selection model (Green and Kern 2012; Hazlett 2014; Imai and Ratkovic 2013).

For example, after the first-stage regression of the linear instrumental variable model, the endogenous explanatory variable X should be predicted, and the predicted value should be included in the main model (Chen 2012). The prediction process can be replaced by machine learning methods. Application cases include Lasso regression (Belloni et al. 2012), ridge regression (Carrasco 2012), and neural network methods (Hartford et al. 2016). Other examples are the prediction of propensity values in propensity score matching (PSM) and the logistic model used in the standard methods (Hu 2012). With machine learning methods, the model hypotheses and restrictions are reduced, and the causal effect estimation is stabilized. Existing application cases include the boosting method (McCaffrey et al. 2004), the neural network method (Westreich et al. 2010), and the regression tree method (Diamond and Sekhon 2013). For another example, regarding the heterogeneity in causal effect (Xie et al. 2012), the machine learning method can also greatly improve the accuracy of the estimation, which demonstrates that we can make more accurate predictions of the counterfactual state of treated or untreated individuals without excessive hypotheses and constraints in the estimation process of the parameter model.

Athey (2018) predicted that, in general, machine learning technology would draw increasing attention to causal inference problems. We argue that in the causal inference of social science, the vast majority of counterfactual constructions can be completed by machine learning methods, and the difference between a counterfactual construction and an actual occurrence can be tested by quantitative methods. We advocate that in the process of constructing counterfactuals, the results of prediction with machine learning ought to be reported as well, which was mentioned by the review of machine learning published in *the Annual Review of Sociology* (Molina and Garip 2019).

Fourth, data proliferation can be realized by prediction. In empirical social surveys, randomly incomplete or missing sample data is a common but aggravating problem. The traditional processing methods are either deleting samples or supplementing data. Deleting data would not only reduce the sample size but also destroy the original sampling design. The method of supplementing data relies on subjective factors, mean values, or the prediction of traditional regression models with comprehensive information, though this method is well developed (Allison 2012). However, quantitative models are not good at accurate predictions, while machine learning can undertake this task. For instance, some scholars have tested the supplemental performance of different machine learning methods based on 15 datasets and found that the performances of support vector machines and naive Bayesian methods were relatively optimal (Farhangfar et al. 2008). Other scholars have tried to use a Gaussian mixture model to estimate the potential contribution of data and supplement data through an extreme learning machine method (a single-layer neural network method) (Sovilj et al. 2016). Sovilj's study evaluated six different datasets and found that the accuracy of supplementing values through machine learning is higher than the accuracy of traditional methods. Based on the existing research, we argue that the estimation of missing data values should adopt an appropriate machine learning method to achieve the best fitting effect or at least report the

fitting effect of machine learning estimation and other methods, among which the optimal result would be chosen.

Fifth, the prediction could promote theoretical innovation. Machine learning can provide powerful methods and new perspectives for scholars and help them expand their theoretical horizons and generate new academic knowledge. In the current research of machine learning in the field of social science, whereas the results given by the algorithm are not the goal of the research, they are a source of discovery that inspire scholars to develop and innovate at the theoretical level, to promote existing theories and to put forward new hypotheses. For example, Kleinberg et al. used the machine learning method to study the case texts of New York state courts (Kleinberg et al. 2017). They first trained the regression tree model to predict the decision of "bail or release" in New York state court cases and explained the contradiction between the model prediction results and the actual judgments through quasi-random experiments. The research shows that the results of recent cases significantly influence the judge, which will make the judge give more severe punishment to minor offenses. This discovery reveals the potential factors affecting the judge's decision-making behavior from a new perspective and thus promotes the theoretical development of the social-psychological process of judicial judgment.

The governance significance of social prediction

It is one of the most traditional and basic topics of social science to predict economic and social processes and guide practice through research. However, neither an early intuitionistic prediction nor covariance research relying on traditional quantitative models can meet the real demands of economic and social prediction. At the forefront of contemporary social science, other disciplines have begun to explore the issue of social prediction. Among these disciplines, algorithm optimization and even prediction competition (to establish an open-source platform for different teams to participate and contribute algorithms to find the optimal machine learning model) have been applied in social governance, which is worthy of attention from policymakers and social scientists. Here are three cases closely related to sociology.

First, the above practice could help socially disadvantaged groups. Professor Salgani, a sociology professor at Princeton University, and his colleagues used the big data of the "Fragile Families Challenge" project of Princeton University (the project tracks 5,000 American children and obtains 54 million data points of their physical and mental health, cognitive ability, social-emotional ability, education and living conditions, family composition, family stability, and family economic status) to conduct an open-platform algorithm competition of machine learning with six aspects of the social results of disabled children's performances, personalities, and life difficulties. More than 150 teams from 68 universities and scientific research institutes in seven countries have submitted prediction algorithms. In addition to applying the results of machine learning models to community services, these teams conduct in-depth learning on special case data of outstanding children growing up in some disadvantaged families to provide a

decision-making basis for improving the living standards of children in disadvantaged groups.[7]

Second, we have social inequality research. Scholars from the Imperial College of Technology published a paper in the *Scientific Reports,* in which they used in-depth learning of street image data to present and analyze the social, economic, environmental, and health inequalities in urban areas (Suel et al. 2019). The team focused on four British cities, including London, Birmingham, Manchester, and Liz, and used 525,860 pictures of London (corresponding to 156,581 postal codes) as the training set. With the government's statistics of housing conditions, average income, mortality, and morbidity, these researchers made relatively accurate predictions of the social stratification in the other three cities. This work's significance lies in training computer programs with visual signal tags in images of some characteristics of urban life (such as the housing quality and living environment) to predict inequality in cities without relevant data.

Third, we consider public health governance. In 2016, *American Economic Review* published a paper on improving urban governance by algorithmic competition (Glaeser et al. 2016). The authors cooperated with the Boston municipal government, Yelp (the largest review website in the United States, and it is similar to China's Popular Review "大众点评网"), and Driviendata (a well-known machine learning and data science competition platform in the United States that is similar to Kaggle and TopCoder) to train algorithms with comment text data on Yelp to predict the possibility of violations of hygiene and health regulations of restaurants in the Boston area. The authors conducted out-of-sample tests of 23 submitted and complete algorithms and compared the predicted results with the actual results of 364 restaurants. Their results show that the efficiency of health inspection can be greatly improved by using the winning algorithm and machine learning to determine the restaurants that are most likely to violate the regulations and need to be inspected.

In many other fields, such as international politics, criminology, and public health, social prediction based on machine learning has also been employed. For example, Perry (2013) used the random forest method to predict the occurrence of violent conflicts in Africa. Berk (2012) used machine learning to predict the risk of crime in several studies. Kleinberg et al. (2015) used the Lasso regression model to predict which patients with medical insurance can obtain the most benefit from joint replacement surgery. To a large extent, these studies have opened up new areas of social exploration and provided important references and data templates for improving social governance.

The discourse significance of social prediction

The restart and revival of social prediction have great significance in terms of discourse to contemporary Chinese sociology. Sociological research was originally imported from the West, which made the development of Chinese sociology closely follow the Western trends in terms of its discipline, approach, and method. The construction of philosophy and social science with Chinese characteristics requires Chinese discourse and a Chinese paradigm in the modern era. At the same time, in the innovation and application

[7] See more at http://www.fragilefamilieschallenge.org

of quantitative methods, there is still a significant gap between Chinese sociology (and other disciplines) and international sociology, although this gap has been greatly reduced in recent years. In fact, in some areas, Chinese sociology has been in line with the forefront of sociological research (such as sociological big data analysis). In this sense, it could help Chinese sociology actively obtain discourse power, serve China's social governance, and realize sinicization in the combination of internationalization and localization by seizing opportunities for machine learning and social prediction.

First, in the international field of computer science, the development of artificial intelligence and machine learning research is at an advanced level. In sociology and social science research, although some Western scholars emphasize the value of machine learning, they are generally confined in the introduction of methods and have not yet comprehended and promoted social prediction as a new method. Therefore, we should seize the historical opportunity, give full play to the scale and cooperation of Chinese academia, generate a batch of important academic works of social prediction, create research hotspots, and form theoretical schools.

Second, China is in a period of social transformation, and various social risks and contradictions are proliferating. Social prediction based on machine learning can greatly reduce the cost of social governance by forecasting and improving the ability of sociologists to comprehend social development and social change. It is helpful to provide better policy services in social governance and enhance the discourse power and contribution of sociology to the modernization of national governance.

Third, given the large population, vast territory, and characteristics of the governance system in China, it is possible to obtain large-scale, high-quality, and in-depth social information based on big data on the premise of respecting and protecting individual privacy. Machine learning based on big data enables the dual advantages of data and methods and forms an in-depth and detailed examination and overall perspective of China's social phenomena and social changes. From this perspective, it is possible for Chinese social scientists to reach the international frontier in the field of computational social sciences and to help comprehensively accelerate the construction of the discipline, academic system, and discourse system of philosophy and social sciences with Chinese characteristics based on machine learning and social prediction research with big data.

The paradigm significance of social prediction

In the development of natural science, scholars in the scientific community use different paradigms in different stages (Kuhn 1962). In social science disciplines, an orthodoxy is similar to a paradigm in different historical stages. Therefore, sociology has also experienced the development and evolution of the paradigm, which appears as a spiral rising from the conventional stage, the crisis stage, and the revolutionary stage to the new conventional stage. When the development of sociology steps into a specific stage and encounters problems that cannot be solved by the original methods or new methods and information materials appear, a new paradigm will appear and gradually be accepted and recognized by the academic community. For sure, the paradigm change in sociological research often does not equate to the "paradigm revolution" that Kuhn named in the field of natural science, such as the subversion of

Newtonian statics by relativity and the overthrow of Euclidean geometry by Riemannian geometry. Instead, social science presents complex coexistence of the new and the old and multiple dialogues.

Kuhn defined the paradigm more clearly at the end of his research career and summarized its content as three nested logical levels. Kuhn suggested that a paradigm refers to a prescriptive consensus of ontology and epistemology, a set of general rules of a theory or model, and a specific problem domain of a symbolic nature (Kuhn 1977). The three basic paradigms of contemporary sociological research are the empirical paradigm, interpretive paradigm, and critical paradigm founded by Durkheim, Weber, and Marx, respectively. In the past hundred years, sociology has transformed from speculation to the coexistence of speculative and empirical work, which demonstrates the core status of the two basic paradigms of interpretation and positivism and the important supplement of the logic mechanism and objective perspective of social realism to the individual reality and subjective difference of social nominalism. Within the empirical paradigm, qualitative interviews based on the field and quantitative analysis based on data complement each other, but there are significant differences in the epistemology, research rules, and problem domains. For example, qualitative analysis focuses on the depth of individual experience and theoretical refinement, while quantitative research focuses on sample size, causal mechanisms, and theoretical verification and falsification and emphasizes scientific attributes (Popper 1986). Therefore, qualitative work and quantitative work, in essence, constitute the subparadigm within the empirical paradigm.

With the rapid development of social prediction based on machine learning, we suggest that the basic paradigm of empirical work will be split into three pieces from the original dual peaks of qualitative and quantitative work to a trilogy of qualitative work, quantitative work, and quantitative prediction. Compared with the traditional quantitative mechanism (correlation and causality), the paradigm differences of social prediction are shown in the following aspects.

First, in terms of epistemology, the black box mechanism is introduced into prediction. Compared with traditional quantitative research that pursues explicit, clear, and theoretical explanations, the epistemological absolutism of prediction is weakened. Second, in terms of the problem domain, the prediction does not focus on the correlation and causal mechanism of the cause and result but takes the accurate estimation of the target variables as the goal. Third, in terms of the research method, the prediction reduces the dependence on theory and the focus on the counterfactual framework and instead relies on algorithms and data to train and test models. Fourth, in terms of the general rule, the prediction does not rely on the traditional hypothesis test rules and model identification techniques, such as the significance level of the regression coefficient. Instead, a series of new standards focusing on prediction accuracy are adopted, such as the F-score of accuracy and precision, the ROC (Receiver Operating Characteristic) curve with the true positive rate (TPR) and false positive rate (FPR) as axes and the area of AUC (Area Under roc Curve) enclosed by the lower part of the curve.

Of course, if we adopt a more cautious attitude towards the understanding of a paradigm, it can be emphasized that the paradigm significance of social prediction lies in

the fact that it has contributed to the subparadigm evolution of empirical sociology. In other words, the social prediction has experienced from qualitative work to qualitative plus quantitative work and then to a trilogy of qualitative work, quantitative correlation, and quantitative prediction.

Discussion and conclusion

Prediction has always been an indispensable element of scientific methods. A prediction can verify and evaluate the applicability and effectiveness of existing theories. Although the process in which prediction drives interpretation has been widely used in physics and other natural sciences, it has not been widely used in social science. Indeed, the complexity of human society is much greater than that of nature, and the available data and computing tools of the traditional social sciences are relatively lacking. In the past two decades, the rapid expansion of information and data in the network era has brought unprecedented opportunities to social science. As early as 2009, Lazer et al. predicted in *Science* that the era of computing social science would dawn (Lazer et al. 2009; Lazer and Jason 2017). In the past ten years, the network has developed, and research has accumulated rapidly. In accordance with Lazer's original expectation, a new wave is surging. Machine learning, which allows computers to work with data in new ways, has been widely used. In summary, there is an urgent need for current academia to embrace machine learning and other trends, and this article intends to point out its future use.

Machine learning provides new assistance for the prediction research of social science and creates conditions for the formation of new paradigms in social science. This article first reviews the historical development of social prediction, discusses the current path of social prediction with the principles and methods of machine learning, and provides theoretical thinking and empirical cases to show the value of prediction in social science. We emphasize that machine learning is helpful to expanding the research horizon of social science as it can obtain latent indicators, inspire theoretical hypotheses, generate causal inferences, realize data proliferation, and promote theoretical innovation. We believe that machine learning is a new paradigm of quantitative research in contemporary sociology and an important opportunity for Chinese sociology, especially computing sociology, to reach the international frontier, which realizes the development from correlation and causality to prediction. Seizing this historical opportunity is conducive to accelerating the construction of philosophy and social science with Chinese characteristics, developing social science theory, and elevating the service level of practice for socialist construction in the new era.

We fully realize that the data mining method that social prediction relies on cannot be perfect. The black box mechanism (such as the well-known Google Flu prediction; see Butler 2013) and prediction errors of machine learning are often criticized (Lazer et al. 2014). However, we believe that every method has its premises, assumptions, and limitations. The mission of sociologists is to ensure that these hypotheses are as close as possible to the specific research situation in a transparent, reasonable, and effective way and to make continuous progress on diminishing limitations. Therefore, Grimmer (2015) pointed out that data scientists should be not only computer scientists but also social scientists. We emphasize that social prediction based on the black box mechanism of machine learning does not imply the abandonment of the existing theoretical thinking

and empirical accumulations. In contrast, theoretical thinking and empirical accumulation play an important role in dismantling the black box. The black box mechanism of machine learning partly stems from its complex algorithms, subtle parameters, and multilayer encapsulations, all of which make it impossible to see the generating process of learning results from data information alone. Another contributing factor is the vastness and complexity of social phenomena and processes. In this sense, the black box mechanism will always accompany social prediction. However, in the process of dismantling the black box, the conclusions drawn from the black box mechanism can and should be interpreted from the theoretical perspective and further verified by empirical methods.

New methods and new paradigms will result in new problems and puzzles. In particular, the introduction of social prediction and machine learning may affect the theoretical saturation of sociology as a discipline. Similarly, sociology could potentially lose humanistic feelings and theoretical concern or even become a pure data mining game. We believe that such concerns are understandable, but they do not constitute grounds for rejection and exclusion. Excessive worry often derives from the lack of a comprehensive understanding of new methods and new paradigms and the lack of confidence in the strong theoretical tradition of sociology and the diversified expression of the humanistic spirit. Since its establishment a century ago, sociology has maintained its strong vitality and attraction with its open vision, tolerant mind, and interdisciplinary spirit.

However, we have to emphasize that these worries and doubts are not baseless. It is exactly this kind of vigilance rooted in the hearts of sociologists that has enabled the leaders of disciplines, advocates of paradigms, reformers of thinking, and innovators of methods to always maintain the spirit of academic reflection, respect, and perseverance of the theoretical tradition at every critical historical juncture. In this sense, excellent sociological research must be an example of the rational use of advanced methods with humanistic feelings and theoretical concern. Humanistic feelings, theoretical concern, and methods of the times are indispensable to the academic background, historical heritage, and contemporary pulse of sociology. Research without human feelings and theoretical concern cannot obtain historical respect and academic depth even if it uses powerful methods and data. Furthermore, research without scientific research methods cannot reach a real historical height regardless of how strong its theoretical consciousness and humanistic spirit are.

Therefore, like the introduction and emergence of every new thing and new field in the history of a discipline's development, the introduction and emergence of machine learning and social prediction into the toolbox and thinking mode of sociologists will neither change the research merit of sociology nor dim the traditional social research paradigms and methods. Causal inference in correlation research, big data, and computing sociology represented by machine learning constitute three frontier quantitative research areas in contemporary sociology. Sociologists' constant inquiry into causal mechanisms, continuous pursuit of the breadth and depth of social information, and unremitting exploration of social processes and phenomena certainty constitute the historical drive and endogenous power of these three frontier areas. Some of these three frontier areas have elaborated their research themes, while others have just emerged. As a new field of sociology, these areas will be nourished by the historical tradition of the discipline and thrive.

Abbreviations
IID: Independent and identically distributed; OLS: Least squares model; SVM: Support vector machine; UML: Unsupervised learning; PSM: Propensity score matching; ROC: Receiver operating characteristic; TPR: True positive rate; FPR: False positive rate; AUC: Area under ROC curve; AIC: Akaike's information criteria; BIC: Bayesian information criteria; VC dimension: Vapnik–Chervonenkis dimension.

Acknowledgements
N/A.

Authors' contribution
Dr YC designed and instructed the study, Dr XW, Dr AH, Dr GH, Guodong Ju contributed in writing and revising. All authors read and approved the final manuscript.

Funding
This research is supported by the major project of National Social Science Foundation—"Research on the development law and guidance strategy of big-data-driven cybersocialmentality" (19zda149).

Availability of data and materials
This paper focuses on theoretical framework and interpretation. There is no data oraffiliated material.

Declarations

Competing interests
The authors declare that they take full responsibilities for contents in this paper and nocompeting interests exist.

Author details
[1]Department of Sociology, Nanjing University, Nanjing, China. [2]Center for Applied Social and Economic Research, Hon Kong University of Science and Technology, Clear Water Bay, Hong Kong. [3]Department of Sociology, Fudan University, Shanghai, China. [4]Department of Social Policy, London School of Economics and Political Science, London, UK.

Received: 23 January 2021 Accepted: 24 August 2021
Published online: 01 September 2021

References

Allison, P.D. (translated by Lin Yuling) 2012. *Missing data*, Shanghai: Truth and Wisdom Press: 32–50.
Athey, S. 2015. Machine Learning and Causal Inference for Policy Evaluation. Paper Published at Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 10–13 August, Sydney, NSW, Australia. https://doi.org/10.1145/2783258.2785466.
Athey, S. 2018. The Impact of Machine Learning on Economics. In *The Economics of Artificial Intelligence: An Agenda from National Bureau of Economic Research* (pp. 507–547).
Athey, S., and Guido Imbens. 2016. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences of the United States of America* 113(27): 7353–7360.
Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica* 80(6): 2369–2429.
Berk, Richard. 2012. *Criminal Justice Forecasts of Risk: A Machine Learning Approach*, 27–41. New York: Springer.
Blei, D.M., A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3(4–5): 993–1022.
Brand, Jennie E., and Yu. Xie. 2007. Identification and Estimation of Causal Effects with Time-Varying Treatments and Time-Varying Outcomes. *Sociological Methodology* 37(1): 393–434.
Butler, D. 2013. When Google Fot Flu Wrong. *Nature* 494(7936): 155–156.
Carrasco, M. 2012. A Regularization Approach to the Many Instruments Problem. *Journal of Econometrics* 170(2): 1–16.
Castle, Jennifer L., Xiaochuan Qin and W. Robert Reed. 2009. How to Pick the Best Regression Equation: A Review and Comparison of Model Selection Algorithms. Working Papers in Economics 09/13, University of Canterbury, Department of Economics and Finance.
Chen, Yunsong. 2012. Logic Imagination and Interpretation: The Application of Instrumental Variables for Causal Inference in the Social Sciences. *Sociological Studies* 6: 192–216.
Chen Yunsong. 2017. Out of Fei Xiaotong's paradox: the dispute of methodology in Sociology. *Tsinghua Sociological Review*, 7: 1–12.
Deng, Zhiwei. 2009. *The dictionary of sociology*, 513. Shanghai: Cishu Press.
Diamond, A., and Jasjeet S. Sekhon. 2013. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics* 95(3): 932–945.
Donaldson, D., and Adam Storeygard. 2016. The View from Above: Applications of Satellite Data in Economics. *Journal of Economic Perspectives* 30(4): 171–198.
Farhangfar, A., K. Lukasz, and D. Jennifer. 2008. Impact of Imputation of Missing Values On Classification Error for Discrete Data. *Pattern Recognition* 41(12): 3692–3705.
Glaeser, E. L., H. Andrew, K. Scott Duke, and L. Michael. 2016. Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy. *American Economic Review* 106(5): 114–118.
Goodman, N. 1955. *Fact Fiction and Forecast*, 114–118. Cambridge: Harvard University Press.

Green, D.P., and H.L. Kern. 2012. Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly* 76(3): 491–511.
Grimmer, J. 2015. We Are All Social Scientists Now: How▪Big Data, Machine Learning, and Causal Inference Work Together. *Political Science and Politics* 48(1): 80–83.
Hartford, J., G. Lewis, B. K. Leyton and M. Taddy. 2016. Counterfactual Prediction with Deep Instrumental Variables Networks. arXiv preprint https://arxiv.org/abs/1612.09596.
Hazlett, C. 2014. Kernel Balancing (KBAL): A Balancing Method to Equalize Multivariate Distance Densities and Reduce Bias without a Specification Search." MIT Unpublished Manuscript.
He, Guangye, Yunsong Chen, Buwei Chen, Hao Wang, Li Shen, Liu Liu, Deji Suolang, Boyang Zhang, Guodong Ju, Liangliang Zhang, Sijia Du, Xiangxue Jiang, Yu Pan and Zuntao Min. 2018. Using the Baidu Search Index to Predict the Incidents of HIV/AIDS in China. *Scientific Reports* 8(1): 1–10.
Hempel, C.G., and P. Oppenheim. 1948. Studies in the Logic of Explanation. *Philosophy of Science* 15(2): 135–175.
Hofman, J.M., A. Sharma, and D.J. Watts. 2017. Prediction and Explanation in Social Systems. *Science* 355(6324): 486–488.
Anning, Hu. 2012. Propensity Score Matching and Causal Inference: A methodological review. *Sociological Studies* 1: 221–242.
Huang, Ronggui. 2017. Network Fields Cultural Identities and Labor Rights Communities: Big Data Analytics with Topic Model and Community Detection. *Society* 2: 26–50.
Imai, K., and M. Ratkovic. 2013. Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation. *Annals of Applied Statistics* 7(1): 443–470.
Jasny, B.R., and R. Stone. 2017. Prediction and Its Limits. *Science* 355(6324): 468–469.
Kaplan, Oscar. 1940. Prediction in the Social Sciences. *Philosophy of Science* 7(4): 492–498.
Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer. 2015. Prediction Policy Problems. *American Economic Review* 105(5): 491–495.
Kleinberg, J., A. Liang and S. Mullainathan. 2017. The Theory is Predictive, but is it Complete? An Application to Human Perception of Randomness. arXiv preprint arXiv: 1706.06974. https://arxiv.org/abs/1706.
Kuhn, T.S. 1962. *The Structure of Scientific Revolutions*, 383–394. Chicago: The University of Chicago Press.
Kuhn, T.S. 1977. *The Essential Tension*. Chicago: The University of Chicago Press.
Lazer, D., A. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. 2009. Social Science: Computational Social Science. *Science* 323(5915): 721–723.
Lazer, D., R. Kennedy, G. King, and A. Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343(6176): 1203–1205.
Lazer, D., and R. Jason. 2017. Data Ex Machina: Introduction to Big Data. *Annual Review of Sociology* 43(1): 19–39.
Li, Hang. 2012. *Statistical learning methods*, 10–67. Beijing: Tsinghua University Press.
Luo, Jiade, Liu Jifan, Yang Kunhao, and Fu. Xiaoming. 2018. Reintroducing Theory into the Triangle Dialogue of Big Data, Theory and Prediction Model. *Sociological Studies* 5: 1–19.
Manski, C.F. 2007. Partial Identification of Counterfactual Choice Probabilities. *International Economic Review* 48(4): 1393–1410.
Martin, T., J. M. Hofman, A. Sharma, A. Anderson and D. J. Watts. 2016. Exploring Limits to Prediction in Complex Social Systems. arXiv preprint https://arxiv.org/abs/1602.01013.
McCaffrey, D.F., G. Ridgeway, and A.R. Morral. 2004. Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods* 9(4): 403–425.
McFarland, D.A., D. Ramage, J. Chuang, J. Heer, C.D. Manning, and D. Jurafsky. 2013. Differentiating Language Usage Through Topic Models. *Poetics* 41(6SI): 1–19.
Mitchell, T. 1997. *Machine learning*. Seattle: McGraw▪Hill Education Press.
Molina, M., and F. Garip. 2019. Machine Learning for Sociology. *Annual Review of Sociology* 45: 27–45.
Morgan, S.L., and C. Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 15–31. New York: Cambridge University Press.
Mohr, J.W., and P. Bogdanov. 2013. Introduction-Topic Models: What They Are and Why They Matter. *Poetics* 41(6): 545–569.
Mohr, J. W, Wagner-Pacici, R., Breiger, R. L., Bogdanov, P. 2013. Graphing the grammar of motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics. *Poetics* 41: 670–700.
Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*, 41–61. Cambridge: Cambridge University Press.
Perry, C. 2013. Machine Learning and Conflict Prediction: A Use Case. *Stability: International Journal of Security and Development* 2(3): 1–18.
Popper, K. (translated by Fu Jizhong). 1986. *Conjectures and refutations: The growth of scientific knowledge*, Shanghai: Shanghai Translation Publishing House: 25–39.
Rubin, D.B. 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66(5): 688–701.
Sovilj, D., Eirola Emil, Miche Yoan, Bjork Kaj-Mikael, Nian Rui, Akusok Anton and Lendasse Amaury. 2016. Extreme Learning Machine for Missing Data Using Multiple Imputations. *Neurocomputing* 174: 220–231.
Suel, E., J. W. Polak, J. E. Bennett and M. Ezzati. 2019. Measuring Social, Environmental and Health Inequalities Using Deep Learning and Street Imagery. *Scientific Reports* 9(1): 1–10.
Varian, H.R. 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives* 28(2): 3–28.
Watts, D.J. 2014. Common Sense and Sociological Explanations. *American Journal of Sociology* 120(2): 313–351.
Weber, M. 1968/1921. *Economy and Society*. Translated and Edited by G. Roth and C. Wittich . New York: Bedminster Press: 183–187.
Westreich, D., J. Lessler, and M.J. Funk. 2010. Propensity Score Estimation: Neural Networks, Support Vector Machines, Decision Trees (CART), and Meta▪classifiers as Alternatives to Logistic Regression. *Journal of Cinical Epidemiology* 63(8): 826–833.

Xie, Y., J. Brand, and B. Jann. 2012. Estimating Heterogeneous Treatment Effects with Observational Data. *Sociological Methodology* 42(1): 314–347.
Yan Yaojun. 1986. On social science and social prediction. In *Social science and contemporary society*, Liu Zhongheng (ed), Shenyang: Liaoning People's publishing house: 21–57.
Yan, Yaojun. 2005. *Basic principles of social prediction*, 15–33. Beijing: Social Science Academic Press.

Publisher's Note