Chen and Chen
The Journal of Chinese Sociology (2024) 11:17
https://doi.org/10.1186/s40711-024-00218-8

The Journal of Chinese Sociology

RESEARCH Open Access

Computing grounded theory: a quantitative method to develop theories

Zhuo Chen[1] and Yunsong Chen[1*]

*Correspondence:
yunsong.chen@nju.edu.cn

[1] School of Social and Behavioral Sciences, Nanjing University, 163 Xianlin Road, Qixia District, Nanjing, Jiangsu, China

Abstract

The inductive logic of grounded theory and the principle of avoiding theoretical preconceptions are significantly different from the deductive logic and hypothesis testing of traditional quantitative research. Based on the limitations of theory production in quantitative research, this paper proposes a Computing Grounded Theory (CGT) approach that directly quantitatively assists theories. With the help of machine learning and attribution algorithms, CGT identifies variables that have not been the focus of previous studies based on the predictive power of the independent variables to propose new theoretical hypotheses, following the principle that causality is a sufficient and unnecessary condition for predictability. This paper systematically discusses CGT's basic idea, logical premise, and methodological foundation while providing an empirical example. This method bridges the gap in the theoretical production of quantitative research and is of great value in theory, discipline, knowledge systems and social governance.

Keywords: Computing grounded theory (CGT), Grounded theory, Machine learning, Attribution algorithms, Quantitative research

Introduction

Sociology quantitative research based on objective data and models has formed a vital supplement to the traditional sociological research methods that have long been rooted in logical speculation and historical contexts. With the implementation of large-scale social surveys and the popularization of data models, it has become an important paradigm for sociological research. It has gradually become a methodological paradigm commonly followed by quantitative sociologists to extract statistical inference and causal identification from multisource data with regression models, test the falsification of theoretical hypotheses, and pursue scientific, normative, and causal explanations in quantitative research by the sociological community.

Quantitative research was originally a general term for analyzing and researching quantified data (Scott and Marshall 2009: 538). With the solidification of the paradigm, especially in the context of the dichotomy of qualitative and quantitative research, the academic community has gradually confined quantitative research to a single aspect that deploys deductive methods as logic, theoretical verification as the purpose, and statistical inference as the means. This method of hypothesis testing undoubtedly exceeds pure

philosophical speculation that does not involve social phenomena; however, it seems to have gradually lost the initiative in theoretical production and development in long-term competition with qualitative research. Qualitative researchers constantly observe, discover, and refine new concepts and theories, forming the pioneer of theoretical development. In contrast, quantitative researchers conduct postpositional statistical tests on theories or hypotheses based on literature and the sociological imagination. Quantitative researchers value Karl Popper's definition of science, and, consequently, often become lost in their self-appreciation of the importance of falsification testing, unconsciously neglecting the value of data and models for directly inspiring theory and the potential application of inductive logic in quantitative research.

Is there a new logical approach and model that allows quantitative scholars not only to conduct postpositional scientific tests but also to directly generate hypothetical theory with data? In fact, there is a long history of developing theoretical explanations with quantitative data, such as in Émile Durkheim's classic study on suicide. There are various feature selection methods used in statistics to meet that end. However, quantitative data is infrequently used to develop theory after decades of institutionalization of quantitative research methods. This is probably because random or even ergodic brainstorming of $N \times (N-1)$ pairwise correlations over the variable list may generate a large number of unfounded or even absurd hypotheses, and loop testing of various X–Y combinations with traditional regression models cannot solve many problems, such as limiting the number of control variables, selecting combinations, and multicollinearity. Therefore, the direct generation of theory with data and models has been neglected by quantitative scholars for a considerable period.

However, it is possible at present with the increase in large-scale social survey data and the application of machine learning in sociology. In this article, we propose a quantitative theoretical production method based on a large amount of data and machine learning models: for a given Y and a large number of explanatory variables X, the predictive ability of X for Y is quantitatively analyzed with a supervised learning method. By utilizing the logical relationship between causality and predictability, we can explore and screen for many X with strong predictive power, directly leading to the generation of theory, finding potential new X with theoretical value for Y, and helping sociologists generate, develop, and revise theories. Although this method is a typical computational social science method, its logical starting point shares similarities with the core principle of grounded theory. It goes beyond the preconceptions of theory and digs into the data itself without making any theoretical hypotheses, thereby transcending the logic of deduction verification and conducting theoretical research with empirical cases. Therefore, we call this "Computing Grounded Theory".[1]

This article will first briefly present the hypothesis testing methods of traditional quantitative research and then provide a detailed introduction to the specific logic and approach of Computing Grounded Theory. The article then further examines the possibility of Computing Grounded Theory from both theoretical and methodological

[1] We noticed a method in academia called "Computational Grounded Theory" (https://doi.org/10.1177/0049124117729703), which involves using computer processing to analyze textual data. To distinguish, we use the term "Computing Grounded Theory".

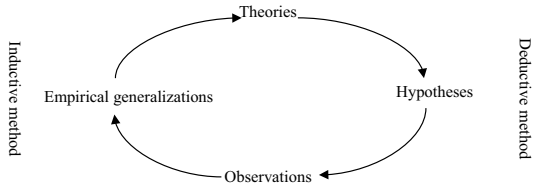

Fig. 1 Two logics of Wallace's the Wheel of Science. Legend: This figure is adapted from Wallace, Walter L. 1971. *The logic of science in sociology*, p. 18, with simplifications. The Wheel of Science illustrates that sociological research is an endless, spiral process that includes theory construction and testing

perspectives and uses the sense of well-being as an example to demonstrate the reliability and validity of Computing Grounded Theory. Finally, this article concludes with a critical review of the methodological significance and potential problems that may arise from Computing Grounded Theory.

The limit of falsification: the bottleneck of the theoretical generation of traditional quantitative research

Hypothesis testing and the wheel of science

In the past four decades, the mainstream pattern of quantitative sociology research has been based on survey questionnaire data and the use of multiple regression models to statistically infer whether there is a correlation or causality between the explanatory and dependent variables. Yusheng Peng once vividly compared quantitative research to "Westernized eight-legged essays" (Peng 2010:180). He noted that mainstream social science journals adopt a relatively standardized "template style" format, with each part of eight sections performing its own duties interconnectedly—question, literature, hypothesis, measurement, data, method, analysis, and conclusion (Peng 2010). Scholars from other countries have made similar summaries of hypothesis testing models by analyzing the content of the *American Sociological Review* (Wells and Picou 1981). More interestingly, the hypothesis-testing paradigm of quantitative research is not unique to sociology, which has permeated various disciplines, such as economics, political science, and psychology (Lin 1995). Although the relevant parts can be merged or refined, the basic principle is to falsify the proposed null hypothesis.

However, testing theory is not the entire work of scientific research. Walter Wallace proposed the wheel of science (Fig. 1) in *The logic of science in sociology*, pointing out that sociological research is a cyclic, spiral, and endless process that includes theory construction and testing (Wallace 1971). It is obvious that the quantitative paradigm of hypothesis testing is confined to the right half of the scientific wheel. It was originally a complete path of scientific research from theory construction to theory testing, but with the distinction between quantitative and qualitative research, theory construction seems to have become the exclusive mission of qualitative research, and quantitative research increasingly adheres to theoretical verification as the norm. In fact, as Merton suggested, empirical research goes far beyond the passive function of testing theories. It not only confirms or refutes hypotheses but also performs at least four functions in the development of theories: creating, revising, transforming, and clarifying theories (Merton 2006).

The historical origin of hypothesis testing

The hypothesis testing paradigm originates from the positivist tradition and was strengthened after the standardized movement of quantitative research by the Columbian school. Paul Lazarsfeld and colleagues (1967) and Samuel Stouffer (1962) advocated for the purpose orientation and scientific movement of using empirical materials to validate theories. Stouffer wrote the book *Social Research to Test Ideas*, which further popularized the method of using data to verify theory throughout the field of quantitative research. This methodological tradition does not regard traditional theoretical discourse, which contains a large amount of metaphysical speculation and untested assertions, as precise scientific knowledge, since it is just empty statements that do not improve reliable judgments about social facts. Through academic cultivation and methodological training, quantitative sociological researchers have gradually developed an empirical personality that requires them to constantly revise their ideas about society and commit to improving the effectiveness of social science by answering substantive questions (Pawson 2000).

There is a consensus in sociology that the quantitative paradigm of hypothesis testing bridges the gap between theory and experience, ensuring the scientificity of conclusions. However, if we took the analytical approach of theory first—data validation next as granted without reflection, this methodology, which used to be a force for knowledge liberation, would easily be transformed into constraints on the productive creativity of theory. In fact, using quantitative data for theory exploration is not new. As early as forty years ago, analytical methods and models for automatically selecting variables from data were available. Statistically, methods such as forward selection, backward selection, and stepwise regression are used to select the most suitable variables for the model. Subsequently, partial least squares regression based on feature dimension reduction, AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) standard comparisons based on information criteria, and regularization methods such as ridge regression and LASSO (Least absolute shrinkage and selection operator) regression have emerged. Although these methods have been somewhat criticized either for unscientific variable selection criteria or for the equivalence between correlation and causality (Rubin 1974), quantitative scholars have been inspired by the data itself in the actual research process. However, many researchers do not present their research according to the actual research process after obtaining new discoveries from the data; instead, they then tends to write their induction discovered hypothesis as if it had been thought up before the research began, and then prove it according to the logic of hypothesis testing (Glaser 2008; Wu and Li 2020).

Consequences of hypothesis testing only

The hypothesis testing approach of quantitative research constrains the knowledge production of quantitative research in two ways.

1. A lack of quantitative exploratory research

The academic community has gradually formed a common impression among practitioners and bystanders for a long time that empirical research is used to verify

theories, and theories are obtained through researchers' whims (Merton and Barber 2011). Although it has contributed to the common prosperity of different research paradigms, it has also led to the postposition or even absence of quantitative research in the field of scientific discovery: quantitative research has refined existing theories but has rarely produced new theoretical constructions (Charmaz 2009).

2. Excessive reliance on common sense and the absence of insight in quantitative research

The hypotheses validated by quantitative research are primarily derived from existing theories or sociologists' common sense and inspiration. The contradiction of common sense lies in its ability to help us understand the world while also weakening our ability to understand (Watts 2011). Interestingly, on the one hand, sociologists need to doubt and verify the scientific nature of common sense with a disciplinary mission of challenging common sense; on the other hand, they have to select possible explanatory variables from existing common-sense stereotypes when establishing hypotheses, which often leads to doubts — using complex methods to verify common sense (Liu and Gong 2020:155).

Computing grounded theory: theory generation with machine learning

The fundamental idea of computational is to bridge a reverse path from data to theory, utilizing the predictive power of machine learning and interpretable attribution algorithms and directly generating the mechanism theory of established dependent variables based on the law that causality is a sufficient and unnecessary condition for predictability. This section will provide a detailed discussion of the basic ideas, logical premises, and methodological foundations of Computing Grounded Theory.

The basic ideas of Computing Grounded Theory

As shown in Fig. 2, Computing Grounded Theory includes the following basic six steps.

The first step is to set the dependent variable. Based on the data from the social survey questionnaire, research Subject Y was selected based on research interests and needs. Theoretically, we cannot determine Y beforehand; thus, each non-preassigned variable becomes the predicted object Y and is analyzed with the ergodic exploratory method.

The second step is to prepare high-dimensional data. Social survey data are often high-dimensional, with hundreds or even more variables. Each of these numerous indicators may be a potential cause of Y, which implies the possibility of grounded theory. Data at different dimensions can be matched, and even irrelevant features can be included.

The third step is to carry out social prediction. Based on high-dimensional data, supervised learning methods, such as support vector machines, random forests, gradient boosting trees, and neural networks, are used to train the prediction model of Y. Algorithms can be diverse. As long as relatively good prediction results can be achieved, complex or interpretable algorithms could also be considered.

The fourth step is to compare the predictive performance. With the interpretability algorithms of machine learning models, attribution analysis is performed on the black box models generated by predictions, and possible causes are searched for based on the

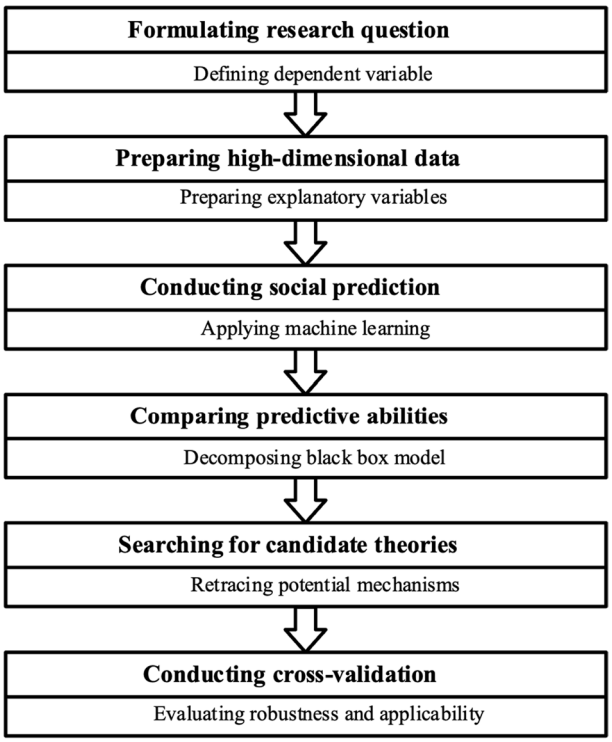

Fig. 2 Overall path of Computing Grounded Theory based on machine learning. Legend: Fig. 2 illustrates the six basic steps of Computing Grounded Theory. The first two steps are preparatory steps, and the third and fourth steps represent core technical steps involving the combination of machine learning and interpretable machine learning. The final two steps focus on theory generation

ranking of X's predictive power on Y. The basic idea is to determine whether disrupting certain features affects the accuracy of model prediction and how changing features will affect the prediction results.

The fifth step is to search for potential theories. Social associations not addressed in previous studies are searched for based on a set of Xs sorted by predictive power. They can be compared with existing research to validate or clarify theories. Similar explanatory terms can also be classified as abstract concepts or inductive theoretical propositions.

The sixth step is cross-validation to verify the robustness of the results of Computing Grounded Theory and the applicability of the theoretical hypotheses. Different data, other machine learning models and attribution algorithms are adopted to calculate the same dependent variable. Alternatively, other hypotheses derived from the new theory are reexamined to mutually verify and complete the wheel of science.

In brief, this method is vastly different from quantitative research on hypothesis testing. It does not presuppose theoretical hypotheses but relies solely on algorithms and data to train predictive models. By accurately estimating and ranking the predictive power of X on Y, it provides a set of potential theoretical hypotheses for possible causal relationships, ensuring an indiscriminate open attitude to data and precise goal orientation toward empirical problems. According to the four-quadrant framework of social science research methods proposed by Hofman and co-authors (2021), grounded

computational theory is a comprehensive modeling approach that integrates prediction and interpretation. According to the five major types of social science methods classified by Zhou and colleagues (2022), the basic ideas for Computing Grounded Theory are "exploratory research based on (big) data" (Zhou et al. 2022:133). If other data are used to further validate the exploratory theoretical hypothesis after it is generated, it belongs to the "integrated research of exploration followed by verification"(Zhou et al. 2022:143).

In fact, scientific research assisted by machine learning is currently a cutting-edge research trend. *Nature* once predicted that artificial intelligence would guide scientific intuition under the cover title of "AI-guided intuition". Specific paths for using similar methods to guide intuition and propose guessing have also emerged in the fields of mathematics (Davies et al. 2021), economics (Ludwig & Mullainathan 2024), and management (Shrestha et al. 2021).[2] Moreover, relevant empirical research has combined machine learning and interpretable AI methods, such as identifying potential factors that lead to anxiety (Nemesure et al. 2021) and exploring potential variables that improve the survival rate of patients with breast cancer (Moncada-Torres et al. 2021). Chinese scholars, such as Jiade Luo, have used relevant methods to revise and clarify the theoretical model of Chinese network circles (Luo et al. 2021), while Zhou and coauthors (2022) have proposed a similar analytical strategy for team innovation ability. On the basis of these relevant empirical explorations, we could standardize and systematically refine this approach and propose a complete method and practice approach that hinges on the possibility of its application in sociological research, focusing on its universal methodological value and significant differences from traditional research models, and substantively connecting the interpretability of algorithmic models and theoretical generation algorithmic models at the methodological level.

The logical premises of computing grounded theory

As a method of theory generation, Computing Grounded Theory has clear logical premises, comprising two aspects.

One is the inductive logic of grounded theory. Grounded theory arises from the reflection of empirical research on the paradigm of quantitative hypothesis testing. Grounded theory's founders noted that sociology places too much emphasis on theoretical verification, and "attempts to close the gap between theory and research have concentrated principally on the improvement of methods for testing theory" (Glaser and Strauss 1967: vii). Researchers should discover theories from data to bridge the discrepancy between empirical research and theoretical research. Grounded theory advocates extracting theories directly from empirical materials with a stepwise induction method and comparing them with existing theories and research. Avoiding preconceived ideas or speculations before the analysis is an important principle to ensure the effectiveness of grounding.

Notably, Barney Glaser, as the founder, emphasized that grounded theory is a universal methodology that applies to both qualitative and quantitative data. "There is no

[2] The article also proposes the approach to construct theory with predictive models. The difference is that the article advocates to combine inexplicable complex algorithms and interpretable algorithms that constrain model complexity to balance the accuracy and interpretability of model predictions. We argue that the balance between prediction and interpretation does not necessarily need to be achieved through constraints on algorithm complexity. The interpretable AI algorithm itself can explain any black box model and avoid any unnecessary or even misleading variable screening.

fundamental clash between the purposes and capacities of qualitative and quantitative methods or data,"... "each form of data is useful for both verification and generation of theory" (Glaser and Strauss, 1967: 17–18). However, with the development of grounded theory, people have found that it still seems more suitable for qualitative research. Anselm Strauss, another proposer of grounded theory, even regarded it as an exclusive tool for qualitative research (Strauss and Corbin 1994). The reason is not difficult to comprehend: the depth and interpretability of qualitative data are often more conducive to directly proposing theoretical hypotheses with sociological imagination, while quantitative data, as a numerical indicator, are characterized by highly simplified abstraction, and their inherent mathematical and statistical correlations are difficult to discover through intuitive means. In fact, to overcome the stereotype in the academic community that grounded theory is only applicable to qualitative data, Glaser has written a manual titled "Doing Quantitative Grounded Theory" to elaborate upon the steps of quantitative grounded theory. The basic idea is: "Saturating core index with all possible two-variable runs, discovering relationships among the theoretically relevant consistency indices, summation indices and single questionnaire items, and then generating conceptual hypotheses. The next non-neglectable step is elaboration analysis. That is to make three or more variable analyses in order to saturate categories further by developing their properties and thereby achieving a denser theory". (Glaser 2008: 54). However, the problem is that a large number of variables are difficult to directly correlate with the human brain. When using statistical methods, there is actually a lack of clear selection criteria for which variables should be included in the model. In particular, when the number of independent variables is large, there may be many problems, such as insufficient degrees of freedom or collinearity. In short, the logic of quantitative grounded theory is feasible, but there is currently no suitable method to carry out a convincing application.

The other aspect is the predictable logic of causality. The predictability and causal mechanisms between social phenomena are two different but highly correlated categories. According to Max Weber, sociology is a science that provides causal explanations of behavioral processes and outcomes (Weber 1968), in which sociological theories can be understood as the causal relationship between indicators. According to this logic, the dependent variable in sociological indicators must have predictability for the independent variable. This is because predictability is a necessary but insufficient condition for establishing causal relationships. It is also the most basic means to verify the principle of a mechanism (Watts 2014).

However, due to limitations in mathematical and statistical tools, sociologists often do not pay much attention to predictions. When discussing the concepts of cause and effect, correlation, and prediction in sociology, sociologists often use evasions: sometimes they emphasize that prediction is not equal to cause and effect, while abandoning the logic that cause and effect can inevitably be predicted. Alternatively, they argue that complex regression models that incorporate too many independent variables are not concise enough, or they criticize algorithmic models that can make data predictions unexplainable due to black box processes. Duncan Watts has summarized and strongly criticized these types of arguments (Watts 2014).

One of the logical foundations of Computing Grounded Theory is to fully utilize the important relationship between prediction and causality; that is, causality is a sufficient but unnecessary condition for prediction. This means that if an X can predict Y well, X may indeed cause Y. Although this relationship is only possible, not inevitable, its probability of forming causality is much greater than that of nonpredictive associations. In the context of sociologists gradually confining the focus of their discipline to two-variable analysis and abandoning social prediction (Hofman et al. 2017), it is important for quantitative research to generate theories with the predictive power of machine learning.

The implementation of computing grounded theory

Computing Grounded Theory allows for the interaction of dozens, hundreds, and even thousands of variables, which is much more stable and reliable in theory discovery by comparing the predictive power of the relevant eigenvalues of algorithmic models than through human thinking. The specific implementation process includes two aspects: social prediction and predictive power comparison.

Social prediction: fitting algorithms with supervised learning

Traditional quantitative regression models are good at correlation and causal inference rather than prediction. Therefore, what kind of model is most suitable for predicting complex social processes? Leo Breiman, a renowned statistician, divided statistical modeling methods into two groups: data models and algorithmic models. The data model assumes that the data follow a certain functional distribution f (x) in advance (such as a linear regression model) and then fits and estimates the parameters of the assumed f (x). The algorithmic model does not assume any distribution characteristics of the data and aims to find a function g (x) through which y can be predicted (Breiman 2001a). In fact, this classification precisely highlights the fundamental difference between traditional econometric models in sociology and machine learning. Breiman further noted that the mindset of data models that are widely used in social and behavioral sciences emphasizes unbiased estimation of model parameters rather than predictive accuracy. In other words, the widely recognized practice model in social science is not to inquire whether specific data and models can predict certain interesting results but to inquire whether specific coefficients in idealized models are statistically significant and their impact direction.

However, there are two obvious problems with data models. First, the data must meet certain assumptions to fit specific parameter models. Taking linear regression as an example, the relationship between independent and dependent variables needs to be linear, the respective variables are not multicollinear, the residuals should follow a normal distribution, the perturbation term should meet the same variance, and there should be no autocorrelation. However, in complex and diverse societies, certain data are too strict. Therefore, the academic community has adopted an ostrich policy, gradually shifting its focus to statistical significance and maintaining an open or suspenseful attitude toward whether the data meet model assumptions (Freedman 1991). Second, the conclusion is about the mechanism of the model rather than the mechanism of the facts. Imposing simple parameter models on the data generated by complex systems can lead to a loss of accurate and critical information. Model errors or the introduction of a large

amount of discretion by researchers in data analysis can result in potential biases (Simmons et al. 2011). If the model cannot simulate natural situations, the conclusion would be incorrect (Breiman 2001a).

The algorithmic model, represented by machine learning, provides a rather powerful alternative solution for the above problems. The implicit epistemological assumption of the algorithmic model is that the intrinsic mechanism of factual data is unknown and complex, and the key is to find an algorithm that can predict y well with x, that is, to use the algorithm to fit the data. Algorithmic models often adopt nonlinear, nonparametric estimation methods to adjust the complexity of the model through one or more hyperparameters. Including data complexity in machine learning provides the analyzed data with arbitrary distributions without any assumptions. We suggest that this liberation will advance theory generation in at least two ways.

First, it reflects the nonlinear data relationships that exist in real social processes. The linear assumption of data models often hardly reflects the social reality. Although the model's simplicity can be advantageous, simplification is only a means rather than an end. Most machine learning fitting processes do not have to fulfill existing function settings but rather aim to pursue prediction accuracy as the highest objective (Breiman 2001a).

Second, it reflects high-dimensional complex data relationships in real social processes. Traditional econometric models can incorporate only limited explanatory variables. Supervised learning algorithms can simultaneously consider thousands of different factors and various complex interaction patterns in a single learning model (Linthicum et al. 2019). The influencing factors of a social phenomenon are numerous and complex, and incorporating more potential 'causes' increases the likelihood of discovering new explanatory dimensions.

Comparison of predictive power: attribution algorithms for the interpretability of black box models

Although machine learning has challenged the various limitations of existing statistical models and improved the simulation of the true state of things, it is widely criticized for its black box process, resulting in inexplicability. However, an increasing amount of evidence in recent machine learning literature suggests that the contradiction between predictive accuracy and interpretability is not as severe as imagined. With the urgent demand for the interpretability of complex models, an increasing number of methods for disassembling black boxes have been invented and implemented (Ribeiro et al. 2016). A widely cited paper by Harvard professors noted that the explanatory analysis of the black box model in machine learning is an effective method for revealing interpretable factors based on data (Doshi-Velez and Kim 2017).

The SHapley Additive exPlanations (SHAP) method could be an example of a detailed explanation of the specific mechanism of the black box model. This method calculates the Shapley value of each X based on alliance game theory and considers it an indicator of its significance. Given that the number and sequence of different participants affect the final overall return, this method calculates the difference in overall return among various states, such as including and excluding the participants, for each combination and records it as the marginal contribution of the individual participant by exhausting

the arrangement and combination of various participants. Then, it calculates the mean of the marginal contribution of the participant for various arrangements and combinations and records it as the Shapley value of the participant (Shapley 1953). The sum of the Shapley values of all participants represents the overall return.

The specific calculation formula for the Shapley value of each participant i is:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S))$$

N is the set composed of all participants. |N| is the number of participants included in this set. S is a combination of participants and a subset of N. $S \subseteq N \setminus \{i\}$ represents all subsets of set N after removing participant i. $v(S \cup \{i\}) - v(S)$ is the average marginal contribution of the overall return with participant i to the total return without participant i, which appears $|S|!(|N| - |S| - 1)!$ in the total ranking. Data scientists have developed the SHAP algorithm to improve computational efficiency, which approximates the Shapley value through conditional expectation functions. The specific techniques will not be detailed here. The Shapley value fully considers the interaction between variables and has a solid foundation in game theory. It is a fair allocation method with simultaneous effectiveness, symmetry, virtuality and additivity (Lundberg and Lee 2017).

In addition to the Shapley value, there are many other methods for interpretability analysis of black box models, such as permutation feature importance, which measures the significance of a particular feature by comparing the changes in model prediction errors before and after permutation (Breiman 2001b). Another example is the partial dependence plot, which is a graphical representation that shows how a single feature influences the predicted output of a machine learning model while controlling for the effects of other features (Zhao and Trevor 2021). Alternatively, an interpretable local surrogate model can be used to simulate the original black box model (Ribeiro et al. 2016). Establishing and developing these methods have provided the possibility of rebalancing the accuracy and comprehensibility of predictions, laying a solid methodological foundation for Computing Grounded Theory.

Practice and standards for computing grounded theory: example of theoretical construction

Research question and data

We use subjective well-being to demonstrate how Computing Grounded Theory could help to inspire and clarify the theory of well-being. The data used in this case are extracted from the Chinese General Social Survey (CGSS) of 2017, which includes a total of 12,582 samples and 783 variables, providing a relatively comprehensive and reliable dataset for the calculation and analysis of well-being. The dependent variable of this study is "Overall, do you think life is happy – very unhappy, relatively unhappy, hard to say, relatively happy, or very happy?", and the independent variables are all variables in the questionnaire except for the dependent variable.[3]

[3] The well-being score and the frequency of depression are excluded since they are almost alternative questions for the dependent variable.

Research methods and steps

The first step is data preprocessing. First, a binary Y helps improve the accuracy of the algorithm's prediction. We mark "very unhappy, relatively unhappy, and hard to say" as 0, representing the unhappy samples, and record the responses of "relatively happy" and "very happy" as 1, representing the happy samples. Second, we convert the categorical variables into dummy variables. Next, we delete variables with more than 30% missing values. Finally, since the number of class 1 samples is significantly greater than the number of class 0 samples, data imbalance may lead to algorithm bias. We use bootstrap sampling to oversample the minority class and achieve rebalancing between the two classes.

The second step is model training. We used the gradient boosting algorithm XGBoost to train the prediction model with 1000 subdecision trees and other default parameters. After 70% of the training set converges iteratively, the remaining 30% of the test set shows a model accuracy of 0.92, with a recall rate of 0.86 and an F1 score of 0.92. The overall performance of the model is satisfactory.

The third step is model attribution. The analysis uses SHAP, a model-agnostic method for interpreting any machine learning model, to identify the most influential factors and their impact on predictions. Specifically, we calculate a SHAP value for each independent variable X in each case. The connotation of this indicator is as follows: for this case, how much of an average marginal contribution will it make to the predicted results when this X is added compared to when it is not added? A positive value means that the addition of X leads to an increase in well-being, while a negative value means that the addition of X leads to a decrease in well-being.

Primary findings

Figure 3a shows the top 20 variables extracted by the attribution algorithm that contribute the most to predicting well-being, measured by the average absolute value of the SHAP value on each X for all cases, that is, the average marginal contribution of variable X. Figure 3b shows the details of the influence of different predictive variables in the form of scatter plots. Each point in the graph represents a real sample. For each row, the color represents the magnitude of feature X of the variable in that row. The darker the color of the point, the larger the X.[4] The horizontal axis represents the size of the SHAP value. The more points with the same SHAP value, the larger the cross-sectional area of the honeycomb, and the thicker it appears. Overall, the graph can reflect the way and magnitude of interactions among variables, as well as the distribution of individual cases. Taking the sense of fairness as an example, the scatter plot shows that cases with a greater sense of fairness (black dots) often concentrate on the right side of the horizontal axis; that is, a positive SHAP reflects an increase in well-being, while cases with a lower sense of fairness (gray dots) often concentrate on the left side of the horizontal axis, where the SHAP is negative and well-being is thus reduced. This indicates that a sense of fairness has a typical positive impact on well-being.

[4] For the convenience of readers, the magnitude of variable eigenvalues is arranged from small to large by default. For instance, the fairness eigenvalues are 1–5, where 1 represents very unfair and 5 represents very fair.

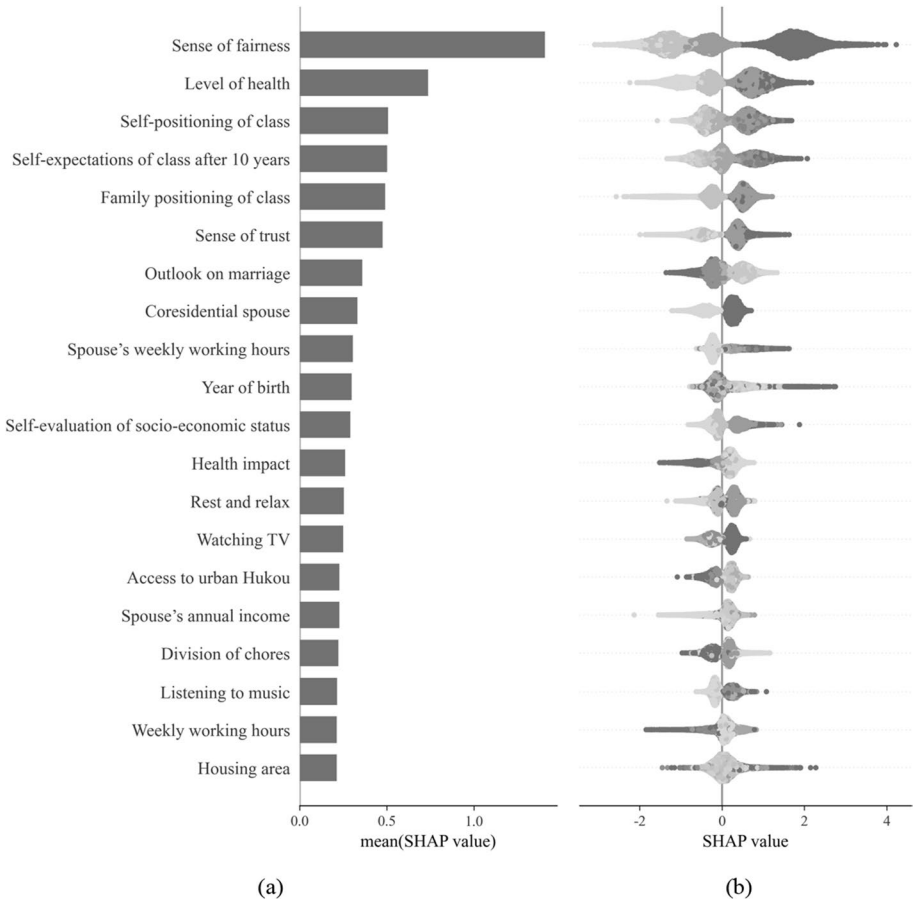

Fig. 3 Top 20 factors with the largest average marginal contribution (SHAP value) to wellbeing prediction. Legend: In 3**a**, the average absolute value of the SHAP value for each variable (X) across all cases is depicted, representing the average marginal contribution of each variable. Figure 3**b** shows the details of the influence of different predictive variables in the form of scatter plots. Each point in Fig. 3**b** represents a real sample. For each row, the color represents the magnitude of feature X of the variable in that row. The larger the X is, the darker the color of the point. The horizontal axis represents the size of the SHAP value. The more points with the same SHAP value, the larger the cross-sectional area of the honeycomb and the thicker it appears. Taking the sense of fairness as an example, the scatter plot shows that cases with a greater sense of fairness (black dots) often concentrate on the right side of the horizontal axis; that is, a positive SHAP reflects an increase in happiness, while cases with a lower sense of fairness (gray dots) often concentrate on the left side of the horizontal axis, where the SHAP is negative and happiness is thus reduced. This indicates that a sense of fairness has a typical positive impact on happiness

Figure 3 shows the top 20 variables that have the greatest impact on predicting happiness. The most significant predictor of wellbeing is the sense of fairness, with a SHAP value of approximately 1.4, followed by the level of health. For the convenience of induction, we categorize five primary influencing dimensions of wellbeing based on the similarity of variable meanings: (1) subjective cognition: sense of fairness and trust; (2) subjective and objective status: self-positioning of class, self-expectations of class after 10 years, family positioning of class, self-evaluation of socioeconomic status, and housing area; (3) demographic and health factors: level of health, year of birth, and health impact; (4) marriage and family: coresidential spouse, spouse's weekly working hours, spouse's annual income, and division of chores between husband and wife; and (5) lifestyle: rest and relaxation, watching TV, listening to music, and weekly working hours.

Theoretically, we can generalize all categories by layer and extract higher-level concepts and an overall theoretical model of wellbeing. It is also possible to further explore and compare a certain variable or specific dimension that has not been previously studied, explore common factors and covariate laws, and summarize theoretical hypotheses at the micro level. Given that the above variables and dimensions involve multiple disciplinary fields and have been discussed in previous theoretical and empirical studies (Liu et al. 2012; Qiu and Li 2012; Diener et al. 2018), we prioritize predictive power and select the variable "spouse's weekly working hours", which ranks among the top ten in terms of predictive power but has not been previously studied, for demonstration.

New discovery of wellbeing: searching for new variables with strong predictive power

The process of generating theoretical hypotheses from Computing Grounded Theory is composed of the following steps: (1) generating empirical propositions of potential hypotheses; (2) eliminating false correlations and establishing causal relationships; (3) inducing and conceptual refinement of relevant categories; (4) discussing existing theories and logical derivations; and (5) summarizing theoretical propositions and using other data methods for revalidation. The first step is to directly rank the predictive power of variables and discover factual propositions about the correlation between variables. However, the proposition has not yet established a rational understanding based on causality between phenomenon and essence. We can further deploy steps 2–5 to fill the gap between proposition and theory to increase the theory's scientificity.[5]

We first propose the empirical proposition. According to Fig. 3a, the variable "spouse's weekly working hours" ranks 9th in prediction, but previous studies have not paid sufficient attention to this variable. We present the relationship between the two variables as a new empirical proposition: a spouse's working hours can affect the other spouse's subjective well-being.

The second step is to adopt double machine learning (Chernozhukov et al. 2018) to exclude other possible confounding variables as much as possible, purifying the relationship between the two variables. Using all other questionnaire variables as confounding variables, four algorithms—Lasso, random forest, decision tree, and XGBoost—all showed a significant causal relationship between the two variables. Due to the word limit of this article, detailed results are not presented here.

The third step is to seek other variables that are highly similar to the connotation of X, observe whether they have explanatory stability and logicality, and then create a concept or a set of concepts to induce a unified understanding of the relationship patterns among the data to eliminate the predictive power caused by data randomness. In this example, "a spouse's weekly working hours" refers to a spouse's time allocation between work and family. We screen other similar variables with large SHAP values that indicate the time allocation of husband and wife to work and family: "weekly working hours (ranked 19)' and 'face-to-face communication time in the family (ranked 21)".

We further compare the above three variables to generate theoretical intuition. Figure 4 shows the variation curve of the SHAP values of the three variables through

[5] These steps are not indispensable, but rather to complement each other and increase the scientificity of theoretical hypotheses.

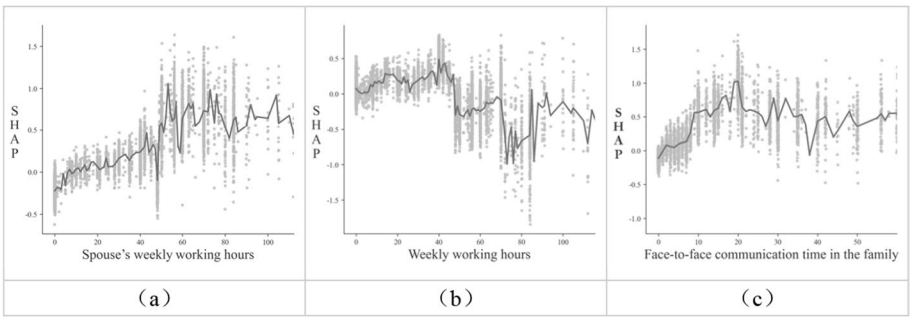

Fig. 4 Partial dependence plots of the new theory of family distance. Legend: The gray dots represent samples; the horizontal axis represents the true value of the relevant feature X of the sample; the vertical axis represents the average marginal contribution of X corresponding to the sample, which is the SHAP value. The black line represents the line of the mean SHAP value of X at each value, and the change in the line reflects the change in the relationship between the two variables. Fig. 4a, Fig. 4b, and Fig. 4c respectively reflect the partial dependence plots for the variables "spouse's weekly working hours", "weekly working hours" and "face-to-face communication time in the family"

the partial dependence plots of the entire sample. It shows that the weekly self-working hours of both spouses increase their wellbeing within 0 to 40 h as working hours increase. However, after working for more than 40 h, there is a completely opposite result of the working hours of both spouses: longer working hours by spouses can significantly improve wellbeing, but an increase in self-working hours can significantly reduce wellbeing. This implies that if spouses take on more social responsibilities for the family and reduce the amount of time they spend at home, this would improve their partner's wellbeing. We conceptualize this phenomenon as "family distance". Moreover, excessive family distance, which means that spouses work more than 60 h per week, will weaken the positive influence of family distance on people's wellbeing. Similarly, the "face-to-face communication time in the family" also shows a peak distribution, where the best length of time is approximately 20 h per week, and the positive influence of the communication is relatively weakened when family members communicate with each other for less than or more than 20 h. By comparing the above variables, we find that improving well-being requires a certain family distance between spouses, but this distance should not be too close or too far. Thus, we construct the core theoretical hypothesis that family distance affects the wellbeing through variable comparison and conceptualization.

For step four, we need to engage in a dialogue with existing theories and logically derive detailed mechanisms of causal relationships between variables, including influence pathways (mediating effects) and heterogeneity of different group influences (moderating effects), to form an abundant series of logically progressive hypothesis propositions.[6] Psychological research suggests that the allocation of time, individual autonomy, and connections with others are important factors that affect individuals' wellbeing (Becker 1965; Reis et al. 2000). The family distance hypothesis proposes that spouses need to maintain an independent and balanced state of time allocation and communication with

[6] We also adopt the Computing Grounded Theory among the male and female groups separately, to provide more evidence to enrich the theory. Due to word limit of this article, it will not be specifically presented in the article. Researchers can conduct more refined analysis based on different groups such as gender, urban and rural areas, and occupation to further inspire and enrich theoretical hypotheses.

others. A short family distance indicates that family members spend more time together, directly squeezing the other party's autonomy and increasing the risk of conflict between spouses. Long family distance leads to estrangement from the family. At the same time, we consider other pressures that accompany family distance.

Existing studies have shown that, due to the emphasis on work and personal responsibility, unemployed individuals are seriously stigmatized in society. Unemployed people are often considered lazy, useless, or unreliable (Brand 2015). In addition to the explicit consequences of earning income, being employed also has significant 'nonmonetary benefits', including providing a time schedule for a day and defining an individual's status and identity (Jahoda 1981). This means that people with shorter working hours also suffer from stigmatization within their families and the loss of nonmonetary costs. On the one hand, spouses with short working hours may be considered lazy and unsuccessful, and their partner may experience a strong sense of deprivation, while a spouse who works longer hours will be considered hardworking, reliable, or successful, and their partner will develop relative satisfaction.

On the other hand, excessive family distance is also not conducive to improving one's wellbeing. Excessive work by spouses means a decrease in contact with their significant others, and as communication between spouses is an important mediating variable for balancing conflict and marital satisfaction (Carroll et al. 2013), excessive family distance often leads to emotional alienation and the accumulation of conflicts. Meanwhile, if family distance is too great, the spouse may have to assume too much family responsibility, leading to the transfer and imbalance of the family obligation distribution (Bianchi et al. 2000).

Due to word restrictions, this section is just a case demonstration of Computing Grounded Theory and does not use other data to verify the generated theoretical hypotheses.[7] Based on the above calculation and analysis results, the family distance theory is summarized; that is, too long or too short of a family distance is not conducive to improving people's wellbeing. We further express this as a series of hypotheses with logical progression.

(1) The family distance of a spouse can affect an individual's wellbeing, but this relationship is nonlinear.
(2) A close family distance between spouses can compress one's autonomy time and increase coexistence conflicts. An appropriate family distance will increase one's autonomy and reduce coexistence conflicts. However, long family distances can reduce communication opportunities for family members, leading to emotional alienation and the accumulation of conflicts.
(3) The family distance of spouses affects wellbeing by influencing couple identification. Too close family distance between spouses can lead to a decrease in their

[7] We advocate that researchers should use multiple methods to verify theoretical hypotheses after obtaining theoretical inspiration, including cross validation of multiple algorithmic models, cross validation of algorithmic models and data models, cross validation of other data, and cross validation of other theoretical propositions based on logical reasoning that can all be used as methods to validate theories. Researchers should choose appropriate validation methods according to real needs. Meanwhile, we do not suggest that the results of the algorithmic model must fully correspond to the results of the data model. As discussed earlier, when comparing data models and algorithmic models, due to their essential differences in variable quantity, parameter settings, and data relationship pattern assumptions, the relationship between the two models should be complementary rather than competitive.

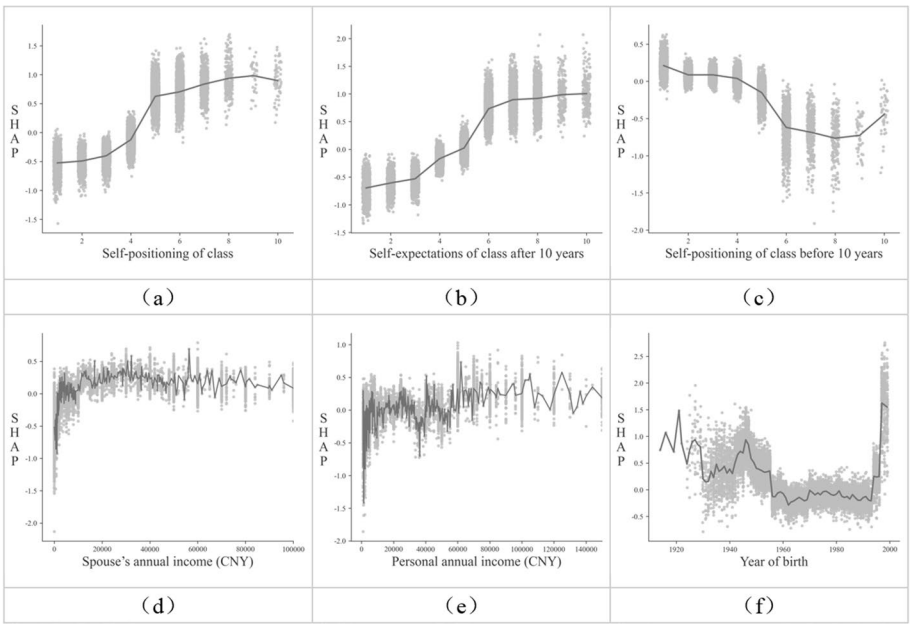

Fig. 5 Partial dependence plots of the SHAP values of related variables. Legend: The gray dots represent samples; the horizontal axis represents the true value of the relevant feature X of the sample; the vertical axis represents the average marginal contribution of X corresponding to the sample, which is the SHAP value. The black line represents the line of the mean SHAP value of X at each value, and the change in the line reflects the change in the relationship between the two variables. Fig. 5a, Fig. 5b, and Fig. 5c display the partial dependence plots for the variables "self-positioning of class," "self-expectation of class after 10 years," and "self-positioning of class before 10 years," respectively. Fig. 5d, Fig. 5e, and Fig. 5f similarly show the partial dependence plots for "spouse's annual income (CNY)," "personal annual income (CNY)," and "year of birth," respectively

sense of identification, and coresidence can lead to a sense of deprivation. A moderate or slightly distant family distance can lead to an increase in identification with a spouse and relative satisfaction with coresidence.

(4) Family distance affects the distribution of family rights and obligations, and a far family distance can force the partner to take on more family responsibilities. Due to the imbalanced distribution of family obligations, conflicts between the two parties accumulate, thereby reducing wellbeing.

Refined development of theory: exploring multiple patterns of complex relationships

According to the previous section, the curve of the SHAP value of "a spouse's weekly working hours" exhibits a nonlinear pattern. Therefore, is there any common pattern other than this complex relationship between X and Y? We also selected some variables and plotted partial dependence plots of the SHAP values (Fig. 5). It is apparent that we can find many details that regression analysis models cannot discover, and these details are vital for further expanding, supplementing, validating, and clarifying the theory. We obtain five basic patterns of complex relationships with the variation in SHAP values of X.

First, there is a "ladder" distribution. The impact of X on Y changes rapidly after a certain turning point and then tends to flatten out, similar to going up a staircase.

Typical variables include "self-positioning of class", "self-expectation of class after 10 years", and "self-positioning class before 10 years" (Fig. 5a-c). Among them, the key turning point of "self-positioning of class" is 4 (Fig. 5a); that is, if self-positioning is above 4, its impact on wellbeing is positive, and there is no significant difference between classes (the SHAP value is 0.6–0.8). Once the positioning is below 4, it quickly becomes negative (approximately −0.2), and the impact of lower social classes (1–3) has not changed significantly (maintained at approximately −0.5). More interestingly, this turning point is different from the expected class turning point for the future (Fig. 5b): the latter has a turning point of 5. This subtle difference means that people who believe they are currently in the middle class of society (= 5) will feel good, but they have higher expectations for the future, while the average marginal contribution of being in the fifth class to wellbeing in the future is only 0.

The second is the "⌒" distribution. The impact of X on Y increases sharply in the early stage and tends to flatten out in the later stage, with "personal annual income" and "spouse's annual income" as examples (Fig. 5d-e). This is consistent with the well-being saturation theory: the positive impact of income on well-being tends to have a decreasing average marginal contribution. This has important implications for social governance policies: poverty alleviation programs should allocate limited funds to the most disadvantaged groups.

The third is the "valley" distribution. The impact of X on Y is relatively high at both ends, while in the middle, it is relatively low, forming a valley shape, with a typical manifestation of "year of birth" (Fig. 5f). The wellbeing of elderly people born before 1955 and young people born after 1995 is significantly greater than that of middle-aged people. In addition, the relationship between age and wellbeing among middle-aged people was not significant, and the SHAP values were almost horizontally distributed. These results echo the groundbreaking literature on age and wellbeing in recent years (Blanchflower & Oswald 2008).

Fourth, there is the "peak" distribution. The impact of X on Y is relatively high in the middle group but gradually decreases on both ends to form a peak, such as the influence of "spouse's weekly working hours" (Fig. 4a) and "face-to-face communication time in the family" (Fig. 4c), which will not be repeated here.

The fifth is the "homogeneity–heterogeneity" effect. The homogeneity effect is manifested as having a consistent impact on the wellbeing of the same group of people, with a small intragroup SHAP variance. The heterogeneity effects manifest as significant differences in the impact on wellbeing among the same group of people, with a large variance in SHAP within the group. Taking "weekly working hours" as an example (Fig. 4b), the SHAP values for working hours ranging from 0 to 40 h are all within −0.5–0.5, whose distribution is relatively uniform, and the impact of work time on wellbeing is relatively homogeneous. The SHAP values for 70–80 h are distributed between −1.5–0.1. with a significant effect of heterogeneity on wellbeing. This suggests that people with shorter working hours are generally happier, while those with longer working hours may be happier or less happy, and there may be other important interaction variables that affect wellbeing.

Robustness testing: solving the rashomon effect

Data and algorithms are highly significant in the calculation process. A considerable number of scholars have noted the Rashomon effect of algorithms, namely, the internal heterogeneity caused by different parameter settings and the external heterogeneity caused by different algorithms (Breiman 2001a; Hu et al. 2021). Is there a Rashomon effect in Computing Grounded Theory? To what extent does it exist? This section tests these questions from the following three perspectives.

First, there was heterogeneity in the data. Robust results will not vary significantly with changes in data volume and composition. We use bootstrap self-sampling to randomly select 50%, 60%, 70%, 80%, 90%, and 100% of the original balanced data for calculation and conduct grounded computational theory.

Second, there is heterogeneity in the prediction algorithms. Robust results should be similar under different prediction algorithms. We compare the calculation results of five algorithms: XGBoost, CatBoost, LightGBM, gradient boosting, and random forest.

Third, there is heterogeneity in the algorithm parameters. Different internal parameters of the same algorithm may also lead to different analysis results. We replace the internal parameters of the XGBoost algorithm, including the maximum tree depth (max_depth), regularization coefficient (alpha), learning rate, subsample, and so on.

Under each condition, we obtain a table that includes all features and their means of SHAP absolute values. We calculate the Pearson correlation coefficient for the SHAP results calculated by different conditional models, as shown in Fig. 6. Overall, the training results of these models are highly similar; the correlation coefficients calculated by the pairwise models are basically above 0.95, and the significance of the correlation coefficients is 0.000. The heterogeneity of the data and the internal heterogeneity of the algorithm parameters are basically nonexistent. There is a certain degree of heterogeneity in the prediction algorithms, but the minimum also reaches above 0.88. We also calculate the Spearman correlation coefficient based on rank, and the analysis results are highly similar to the Pearson coefficient, so we do not report them here. In brief, in the case of wellbeing, Computing Grounded Theory has a considerable degree of robustness.

Recommended technical standards for grounded computational theory

No unified standard for machine learning training methods has been applied in the field of social sciences. Therefore, we selected 60 highly cited papers in the social sciences from the core collection of the Web of Science with the search keyword "machine learning" and summarized information such as the number of variables, sample size, model selection, and model evaluation indicators commonly used in training models, providing empirical reference standards for algorithmic model training.[8]

(1) Number of samples. According to the statistical results of the literature, the median sample size was 1,888, and the median sample size after 2015 was 11,196. After ensuring sample availability and representativeness, we suggest that the sample size

[8] Due to world limit, specific screening and statistical results will not be presented in detail. Readers can contact the author for more detailed results.

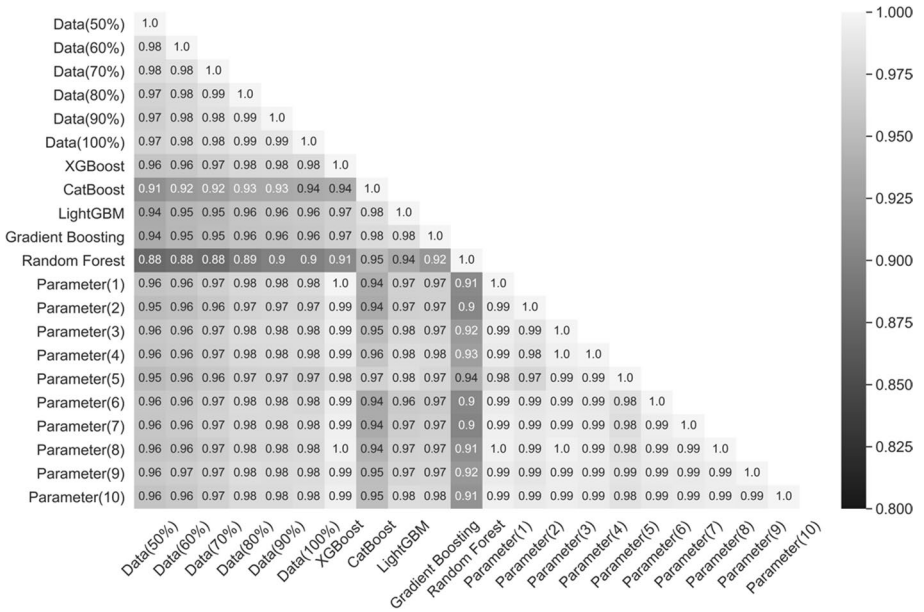

Fig. 6 Results of the robustness test of the computing grounded theory. Legend: The X and Y axes represent 21 different scenarios, including different data, models, and parameters, for which SHAP values were calculated. The value in each small square represents the Pearson correlation coefficient calculated for each pairwise combination of these scenarios, with color indicating the magnitude of the coefficient. Lighter colors indicate greater similarity between the SHAP results calculated by the models for these two scenarios

for exploratory analysis should be greater than 2000, and the sample size for exploratory and validation analysis should be even larger.

(2) Sample balance. The sample size should be adjusted according to the number of categories and difficulty level of the variables to be predicted. It is particularly worth noting that the sample size for rare categories should not be too small. For imbalanced samples, resampling should be used to balance the various sample sizes (Chen et al. 2022).

(3) The number of features, that is, the number of X variables used for training. According to the literature, the largest number reached 1,821, and the median number of all papers was 22.5. Many X variables will bring better training results and be conducive to discovering new potential theories. However, it is also important to consider that certain models may be sensitive to data noise.

(4) Algorithmic model selection. Among the 60 papers, the most commonly used algorithm was random forest (29%), followed by support vector machine (26%). Neural network and gradient enhancement algorithms account for approximately 17% and 15%, respectively. Most papers have adopted more than one algorithm and compared the performance results of the models. Therefore, we suggest comparing the predictive performance and results of multiple algorithms, selecting the optimal model as much as possible, and conducting robustness tests.

(5) Model prediction performance. The median accuracy of the model in the existing literature is 0.79.[9] The effectiveness of Computing Grounded Theory is based on the accuracy of model predictions. Given that most of the predicted variables in existing papers are binary variables, we suggest that the accuracy of binary variables should be greater than 0.8. This criterion for the accuracy of continuous variables can be appropriately reduced.

Multiple values of computing grounded theory

The development of Computing Grounded Theory to traditional quantitative research is multifaceted. At the data level, the indicators included in the model are no longer a limited group of variables; rather, as many indicators are included as possible. At the target level, the statistical significance of model coefficients is no longer emphasized, but the accuracy of social prediction and the interpretability of mechanisms are rebalanced. From the perspective of observation, this approach is no longer limited to the number and direction of regression coefficients but rather involves careful exploration of the nonlinear relationships and population heterogeneity effects between variables. These innovations can bring various values to grounded computational theory.

The value of theoretical generation: discovering potential patterns

Compared to traditional data models, algorithmic machine learning methods can overcome the limitations of model form and variable selection and consider various interaction relationships among variables. The breaking of the bottleneck in the number of independent variables and the limitation of relationships allows us to gain a more robust ability to search, think, and test explanatory variables. This means that as long as the data itself are abundant, Computing Grounded Theory can guide researchers to form new theoretical hypotheses by discovering new explanatory variables (Chen et al. 2020, 2021). For a given dataset, by performing the method of Computing Grounded Theory only once, we can filter and compare hundreds of indicators across the entire survey dataset.

Value of theoretical development: capturing complex relationships

Traditional econometric methods use data fitting models, which can easily lead to the loss of, or even errors in, key information (Varian 2014). Grounded computational theory uses hyperparameters to fit data. As long as the model simulates real social situations as much as possible, it will fully capture the complex relationships between variables, liberate the linear shackles of traditional econometric models, and validate or develop theories. The previous case clearly demonstrates its ability to reveal and explain complex relationships and reminds us that the pairwise relationship of data in the real world is far from as neat and uniform as we expected: the SHAP curve has almost not obeyed any distribution of a straight line.

[9] The model evaluation indicators used in different papers vary. To ensure consistency, the highest values of model accuracy, precision, recall rate, AUC, and R^2 are used as the approximate reference for model accuracy.

The value of the discipline paradigm: the second sociological imagination

Hofman and Watts et al. published an article in the 2021 issue of *Nature* calling for the integration of interpretation and prediction in the computational social sciences (Hofman et al. 2021). They noted that research activities that integrate interpretive and predictive thinking are valuable, but existing research is scarce, and this field should receive more attention than it has so far. This method is a new attempt to integrate the explanatory and predictive aspects of social science. For quantitative research paradigms, mastering Computing Grounded Theory is equivalent to obtaining a supplement beyond the sociological imagination proposed by Mills. Mills' sociological imagination is the enhancement of an analytical perspective based on personal experience (Mills 2017), while Computing Grounded Theory provides a data-based thinking ability to directly assist in the production of theory through algorithmic models. The second type of sociological imagination contains the immense power that drives new theoretical discoveries and unleashes potential.

The value of knowledge systems: autonomous knowledge production

Grounded computational theory naturally has stronger power in the production of systematic knowledge: there are many new theoretical hypotheses that can be extensively inspired by data and more subtle mechanisms and relationship features that can be simultaneously discovered through detailed contribution analysis of predictive power for theoretical development and clarification. A real autonomous knowledge system requires a tool that is capable of discovering complex relationships and extracting theories from large-scale, spatiotemporal, and high-throughput data. Computing Grounded Theory is undoubtedly one of the most important components of such tools.

The value of social governance: finding intervention factors

Sociology is a practical science, and the public and governance entities are often not satisfied with concept extraction, process interpretation, or statistical judgment. This means that the disciplinary mission of quantitative sociology cannot be confined to verifying theoretical hypotheses; but it must also master the ability to identify key intervention factors for social phenomena to provide advice and suggestions for those who serve the country. Computing Grounded Theory is a problem-oriented approach to social prediction, and by simulating existing social phenomena, it has important practical value in identifying key intervention variables for social governance.

Conclusion

As an important research paradigm in the field of sociology, quantitative research is deeply rooted in the tradition of positivist methodology and tends toward single-path dependency on hypothesis testing. The overemphasis on theoretical validation in quantitative research may overlook the enormous theoretical energy inherent in the data itself. Based on this, this article proposes a method for theoretical production based on quantitative data: Computing Grounded Theory. With the predictive power of machine learning and the interpretability of attribution algorithms, Computing Grounded Theory

can open up a gate for developing quantitative research in terms of exploring potential relationship patterns and capturing nonlinear relationships and for opening up the path from empirical observation to theoretical production.

With respect to the development history of quantitative methods, more than forty years ago, there was an academic trend in the social sciences of exploring the relationships among variables based on data and models; however, no mature research paradigm had yet been developed. The reason for this is that the inclusion of many variables makes it difficult to satisfy assumptions and leads to problems such as multicollinearity. Second, the method of selecting variables by deleting or adding a single indicator is only a partial rather than a globally optimal solution, and changes in control variables can cause significant disturbance to the results. Third, the presupposed function mode has difficulty exhausting the complex relationships and action modes among variables. Forty years later, we once again call for quantitative research to fill the gaps in theory production, learn from historical lessons, and address the enormous theoretical energy contained in the data. The advantages of grounded computational theory include the following: first, the algorithmic model can break the assumptions and relationship patterns made by the model, fully incorporate a large number of variables and consider the complex relationship patterns among variables; second, interpretable machine learning can utilize the algorithmic power to obtain the global optimal solution while fully considering various permutations and combinations of variables; third, the importance ranking of variables based on predictive power is closer to the category of causal relationships in analytical logic than simple variable correlations and; fourth, mining and visualizing various nonlinear relationship patterns among variables provide more solid and detailed information for guiding theoretical intuition.

While celebrating the capability of algorithms and data in theory generation, it is worth noting that this article does not negate "traditional" quantitative methods and their value. Each method has its own premises, assumptions, and limitations. They are all important components of quantitative sociological methods. We emphasize that Computing Grounded Theory is not a rejection of theory but rather a departure from the limitations of existing theories and common sense, creating opportunities for proposing new hypotheses. Computing Grounded Theory does not exclude the validation of theories but emphasizes the generation of theories from data as a scientific step before quantitative scholars test theories.

We are fully aware that new analytical methods often propose new research questions. The challenge of Computing Grounded Theory is as significant as the knowledge production it can bring. These challenges include the following: first, the limitations of the data. Variables can never be exhausted. Although Computing Grounded Theory is trying to broaden the data dimension for analysis, it is not far from "dancing with the shackles of data availability". Second, there are limitations to social prediction. There has always been skepticism about the predictability of complex social phenomena (Taleb 2010). Due to insufficient data, model limitations, and the inherent unpredictability of complex social systems, grounded computational theory is not applicable to all research scenarios. Third, there is heterogeneity in grounded computational theory. The position of knowledge production for researchers shifts from the front-end to the back-end, with data and models being pushed to a significant position and potentially leading to potential biases. Fourth, correlation is not

causation. Predictability is not equivalent to causality, and further exploration is needed to determine the causal relationship and implicit impact mechanism chain.

Any method has a long development process, which is constantly being tested and corrected by practice and the scientific community. There are many areas that need to be explored and improved in the future for Computing Grounded Theory. For example, the recommended standards and specifications for Computing Grounded Theory need to be further tested and improved, its applicable scenarios and reliability and validity need to be explored, and the dialogue between Computing Grounded Theory and statistical inference and causal inference methods needs to be promoted. Meanwhile, the Computing Grounded Theory proposed in this article is mainly based on the analysis of structured data. With the continuous progress in diverse forms of big data and artificial intelligence, it is also worth considering whether and how to apply grounded computational theory to massive unstructured data and more complex deep learning algorithms. As a new paradigm of quantitative research that blends qualitative paradigm thinking and logic, Computing Grounded Theory appeals for more inclusion, practice, and review in the academic community. We call for more empirical testing and exploration and more active exploration and application of Computing Grounded Theory in current sociological research. Only when Computing Grounded Theory can effectively generate more concepts and theories for contemporary sociology and generate more autonomous knowledge for sociology can we have a deeper understanding of the power and limitations of this method.

Abbreviations

CGT	Computing grounded theory
AIC	Akaike information criterion
BIC	Bayesian information criterion
LASSO	Least absolute shrinkage and selection operator
SHAP	SHapley additive exPlanations
XGBoost	EXtreme gradient boosting
CatBoost	Categorical boosting
LightGBM	Light gradient boosting machine

Acknowledgements
We appreciate the anonymous reviewers for their important comments and suggestions.

Author contributions
YC and ZC conceived and designed the study; YC instructed the study, ZC performed the data analysis; ZC and YC wrote the paper and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding
This article is sponsored by the major project of the National Social Science Foundation, "Research on the Development Law and Guiding Strategies of Big Data Driven Network Social Psychology" (19ZDA149).

Availability of data and materials
We based our study on data, publicly available of the Chinese General Social Survey (CGSS,2017) http://cgss.ruc.edu.cn/

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 21 November 2023 Accepted: 19 May 2024
Published online: 24 June 2024

References

Becker, Gary S. 1965. A theory of the allocation of time. *The Economic Journal* 75 (299): 493–517.

Bianchi, Suzanne M., Melissa A. Milkie, Liana C. Sayer, and John P. Robinson. 2000. Is anyone doing the housework? Trends in the gender division of household labor. *Social Forces* 79 (1): 191–228.
Blanchflower, David G., and Andrew J. Oswald. 2008. Is well-being U-shaped over the life cycle? *Social Science & Medicine* 66 (8): 1733–1749.
Brand, Jennie E. 2015. The far-reaching impact of job loss and unemployment. *Annual Review of Sociology* 41: 359–375.
Breiman, Leo. 2001a. Statistical modeling: The two cultures. *Statistical Science* 16 (3): 199–231.
Breiman, Leo. 2001b. Random Forests. *Machine Learning* 45: 5–32.
Carroll, Sarah June, E. Jeffrey Hill, Jeremy B. Yorgason, Jeffry H. Larson, and Jonathan G. Sandberg. 2013. Couple communication as a mediator between work-family conflict and marital satisfaction. *Contemporary Family Therapy* 35: 530–545.
Charmaz, Kathy. 2009. *Constructing grounded theory: A practical guide through qualitative analysis*. London: Sage Publications.
Chen, Yunsong, Guangye He, and Fei Yan. 2021. *Understanding China through big data: Applications of theory-oriented quantitative approaches*. London: Routledge.
Chen Yunsong, Guangye He, and Guodong Ju. 2022. The hidden sexual minorities: Machine learning approaches to estimate the sexual minority orientation among Beijing college students. *Journal of Social Computing* 3 (2): 128–138.
Chen Yunsong, Xiaogang Wu, Anning Hu, Guangye He, and Guodong Ju. 2020. Social prediction: A new research paradigm based on machine learning. *Sociological Studies* 35 (3): 94–117.
Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21 (1): C1–C68.
Davies, Alex, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli. 2021. Advancing mathematics by guiding human intuition with AI. *Nature* 600 (7887): 70–74.
Diener, Ed., Oishi Shigehiro, and Tay Louis. 2018. Advances in subjective well-being research. *Nature Human Behaviour* 2 (4): 253–260.
Doshi-Velez, Finale, and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 https://doi.org/10.48550/arXiv.1702.08608.
Freedman, David A. 1991. Statistical models and shoe leather. *Sociological Methodology* 21: 291–313.
Glaser, Barney G. 2008. *Doing quantitative grounded theory*. Mill Valley: Sociology Press.
Glaser, Barney G., and Anselm L. Strauss. 1967. *Discovery of grounded theory: Strategies for qualitative research*. New Brunswick: Aldine Transaction.
Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. Prediction and explanation in social systems. *Science* 355 (6324): 486–488.
Hofman, Jake M., Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J. Salganik, Simine Vazire, Alessandro Vespignani, and Tal Yarkoni. 2021. Integrating explanation and prediction in computational social science. *Nature* 595 (7866): 181–188.
Hu Anning, Xiaogang Wu, and Yunsong Chen. 2021. Analysis of heterogeneous treatment effect: New opportunities and challenges with machine learning techniques. *Sociological Studies* 36 (01): 91–114.
Jahoda, Marie. 1981. Work, employment, and unemployment: Values, theories, and approaches in social research. *American Psychologist* 36 (2): 184–191.
Luo Jiade, Xin Gao, and Tao Zhou. 2021. A new paradigm of combining big data and survey data based on the theoretical perspective. *Sociological Studies* 26 (02): 69–91.
Liu Junqiang, Moulin Xiong, and Yang Su. 2012. National sense of happiness in the economic growth period: A study based on CGSS data. *Social Sciences in China* 12: 82–102.
Lazarsfeld, Paul F., William H. Sewell, and Harold L. Wilensky. 1967. *The uses of sociology*. New York: Basic Books.
Lin Yifu. 1995. Localization, standardization, and internationalization - Celebrating the 40th anniversary of the founding of economic research. *Economic Research Journal* 10: 13–17.
Linthicum, Kathryn P., Katherine Musacchio Schafer, and Jessica D. Ribeiro. 2019. Machine learning in suicide science: Applications and ethics. *Behavioral Sciences & the Law* 37 (3): 214–222.
Liu Runze, and Yixuan Gong. 2020. Review and Reflection: The misuse of quantitative research in public administration. *Journal of Public Management* 17 (01): 152–158.
Ludwig, Jens, and Sendhil Mullainathan. 2024. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics* 139 (2): 751–827.
Lundberg, Scott M., and Su-In, Lee. 2017. A unified approach to interpreting model predictions. *In Proceedings of the 31st International Conference on Neural Information Processing Systems*: 4768–4777.
Merton, Robert K. 2006. *Social theory and social structure (Translated by Tang Shaojie and Qi Xin)*. Jiangsu: Yilin Publishing House.
Merton, Robert K., and Elinor Barber. 2011. *The travels and adventures of serendipity*. Princeton: Princeton University Press.
Mills, C. Wright. 2017. *Sociological imagination (Translated by Li Kang)*. Beijing: Beijing Normal University Press.
Moncada-Torres, Arturo, Marissa C. van Maaren, Mathijs P. Hendriks, Sabine Siesling, and Gijs Geleijnse. 2021. Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. *Scientific Reports* 11: 6968.
Nassim Nicholas Taleb. 2010. *The black swan: The impact of the highly improbable*. New York: Random House.
Nemesure, Matthew D., Michael V. Heinz, Raphael Huang, and Nicholas C. Jacobson. 2021. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific Reports* 11: 1980.
Pawson, Ray. 2000. Middle-range realism. *European Journal of Sociology* 41 (2): 283–325.
Peng Yusheng. 2010. The structure of empirical social research. *Sociological Studies* 25 (02): 180–210.
Qiu Haixiong, and Gan Li. 2012. A review of studies on well-being from multiple perspectives. *Sociological Studies* 27 (02): 224–241.

Reis, Harry T., Kennon M. Sheldon, Shelly L. Gable, Joseph Roscoe, and Richard M. Ryan. 2000. Daily well-being: The role of autonomy, competence, and relatedness. *Personality and Social Psychology Bulletin* 26 (4): 419–435.
Ribeiro, Marco Tulio., Sameer, Singh., and Carlos, Guestrin. 2016. Why should i trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*: 1135–1144.
Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5): 668–701.
Scott, John, and Gordon Marshall. 2009. *A dictionary of sociology*. Oxford: Oxford University Press.
Shapley, Lloyd S. 1953. A value for n-person games. *Contributions to the Theory of Games* 2 (28): 307–317.
Shrestha, Yash Raj, Vivianna Fang He, Phanish Puranam, and Georg von Krogh. 2021. Algorithm supported induction for building theory: How can we use prediction models to theorize? *Organization Science* 32 (3): 856–880.
Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22 (11): 1359–1366.
Stouffer, Samuel A. 1962. *Social research to test ideas*. New York: The Free Press of Glencoe.
Strauss, Anselm., and Juliet, Corbin. 1994. Grounded theory methodology: an overview. Handbook of qualitative research. Thousand Oaks: Sage.
Varian, Hal R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28 (2): 3–28.
Wallace, Walter L. 1971. *The logic of science in sociology*. Chicago: Transaction Publishers.
Watts, Duncan J. 2011. *Everything is obvious: once you know the answer*. Australia: Currency.
Watts, Duncan J. 2014. Common sense and sociological explanations. *American Journal of Sociology* 120 (2): 313–351.
Weber, Max. 1968. *Economy and society*. New York: Bedminster Press.
Wells, Richard H., and J. Steven Picou. 1981. *American sociology: Theoretical and methodological structure*. Washington, DC: University Press of America.
Wu Suran, and Minghui Li. 2020. Grounded theory: History and logic. *Sociological Studies* 35 (02): 75–98.
Zhao Qingyuan, and Trevor Hastie. 2021. Causal interpretations of black-box models. *Journal of Business & Economic Statistics* 39 (1): 272–281.
Zhou Tao, Xin Gao, and Jiade Luo. 2022. Social science research methods driven by social computing. *Sociological Studies* 37 (05): 130–155.

Publisher's Note