Journal of Current Research and Review
Volume 13, Number 7; July-2022;
ISSN: 2836-5615 | Impact Factor: 6.19
https://zapjournals.com/Journals/index.php/jcrr/index
Published By: Zendo Academic Publishing

TRANSFORMING RESEARCH: EXPLORING THE INTERPLAY OF DATA MINING, MACHINE LEARNING, AND KNOWLEDGE DISCOVERY

[1]Bond, Robert M., and [2]Fariss, Christopher J

Article Info

Keywords: Knowledge discovery, data mining, machine learning, causal discovery, prediction, big data.

Abstract

The emergence of big data has necessitated the development of new analytical methods in the interdisciplinary field of knowledge discovery and data mining. This approach goes beyond traditional statistical approaches and employs deductive and inductive processes to extract new knowledge from vast amounts of data. By considering a larger number of joint, interactive, and independent predictors, data mining addresses causal heterogeneity and enhances prediction capabilities. Rather than challenging conventional model-building approaches, data mining complements them by improving model goodness of fit, uncovering hidden patterns, identifying nonlinear and non-additive effects, and providing valuable insights into data developments, methods, and theory. Additionally, data mining enriches scientific discovery by revealing valid and significant findings. Machine learning, on the other hand, leverages models and algorithms to learn from data, particularly when the explicit model structure is unclear or achieving good performance is challenging. Recent developments incorporate this predictive modeling paradigm with the classical approach of parameter estimation regressions, resulting in improved models that combine explanation and prediction. In this era of big data, knowledge discovery and data mining have revolutionized research processes across various fields, including the social sciences. These projects require domain knowledge from experts in diverse disciplines, as well as expertise in data processing, database technology, and statistical and computational algorithms. Data mining technologies enable the discovery of previously hidden patterns, fostering innovation and the development of new theories. This paper explores the epistemological contributions of data mining to theory innovation and discusses the implications of big data. By situating knowledge discovery and data mining within the philosophical and methodological

[1] University of California Davis, USA

[2] University of California Davis, USA

traditions of scientific research, we highlight their strengths and challenges. We provide a systematic explanation of key procedures in supervised and unsupervised machine learning, model selection and assessment, and machine learning development. Through the integration of basic research principles, classical statistics, and machine learning, we demonstrate potential pathways for discovering new knowledge through data mining. Furthermore, we review empirical research to illustrate the research processes and the contribution of knowledge discovery in theory advancement and innovation.

Introduction

We have entered an era of big data. The exponential growth in the amount and complexity of data from the internet, mobile phones, wearable devices, computers, and recording equipment have provided new opportunities to study human behaviors, discourse, and interactions. The rise of big data and the wealth of knowledge concealed in the data mine invisible to human eyes require a paradigm shift in the research process. Knowledge buried in this data mine necessitates data mining technologies to discover interesting, meaningful, and robust patterns. This alternative method of research relative to the traditional approach has had profound effects on all research fields, including the social sciences.

Knowledge discovery and data mining as a research paradigm have emerged as a structural transformation in the nature of research in the social sciences. Knowledge discovery projects require deep domain knowledge from experts such as sociologists, psychologists, economists, political scientists, and linguists, as well as in-depth knowledge about data processing, database technology, and statistical and computational algorithms to discover valid and meaningful knowledge. Data mining provides us with a series of new technologies to assist us in revealing previously hidden patterns, which have the potential to help us innovate and develop new theories, thus promoting revolutionary influences on new theory development in many disciplines.

In this paper, we first assess the contribution of data mining to theory innovation from an epistemological view and discuss the rise of big data and its implications. We situate the new frontier of knowledge discovery and data mining in the philosophical and methodological traditions of scientific research and clarify both the strengths and challenges of data mining. We offer systematic explanations of key procedures in supervised and unsupervised machine learning, selecting and assessing analytical models, and developing machine learning. With each procedure, we apply basic research principles, classical statistics, and machine learning to illustrate potential pathways for discovering new knowledge through data mining. We also review empirical research to illustrate research processes and the contribution of knowledge discovery in theory advancement and innovation.

1. Knowledge discovery: integrating theory and data

1.1. Dialectic relationship between theory and data

Knowledge discovery is a dialectic research process that is both deductive and inductive. A deductive approach is “top-down,” starting from theories and concerned with testing hypotheses, while inductive research derives patterns from data and is more open- ended. One common misconception of data mining is the belief that “datadriven” or “data fishing” research does not require theoretical guidance. Such a misunderstanding has hindered the social science research community from accepting and adopting this approach, and hesitation, doubts, and resistance still exist. The reciprocal relationship between theory and research can be characterized by deductive and inductive research approaches. While the relationship between theory and research differs between the two approaches, they complement each other. Most social research involves inductive and deductive processes at some point in the project. The two processes are often combined to form a cycle from theory to data and back to theory. This cycle is also the case for knowledge discovery.

Such a research process combining inductive and deductive methods is not new in social science research. Grounded Theory is a general method that uses systematic research to generate systematic theory that involves both inductive and deductive phases of research (Holton and Walsh 2017). It is a well-established research tool that enables researchers to employ rigorous research procedures for data collection and analysis to develop conceptual categories. This investigation process proceeds in two stages, encompassing both inductive and deductive research dimensions. Researchers first immerse themselves in data and apply an inductive approach to generate substantive patterns from the data. They do so by discerning and conceptualizing hidden social patterns and structures from observational data in a specific research field through the process of constant comparison. From this newly developed theory, they shift to the second stage of the grounded theory process by designing new research, collecting data, analyzing data, and testing the theory.

Although the grounded theory process has been practiced mostly in qualitative research, the scientific field of data mining has also developed and matured by carefully trying to guard against the pitfalls of data-driven approaches. As soon as the data mining field was taking shape in the mid-1990s, there were efforts to initiate and establish the Knowledge Discovery Process (KDP) model by defining rigorous procedures to guide users of data mining tools in knowledge discovery (Fayyad et al., 1996; Anand and Büchner 1998). These models involve multiple iterative steps, and loops connect any two steps. Most importantly, all the models start with understanding domain knowledge, indicating a theory-guided process. The next steps are selecting a dataset, data preprocessing, data reduction, choosing a data mining method, selecting a data mining algorithm, conducting data mining, interpreting patterns, and consolidating discovered knowledge. Similar to the dialectic relationship between theory and data, the process can reverse to an earlier step to revise and reconsider in light of new insights. An excellent example of such practice is the “Computational Grounded Theory” research procedure (Nelson 2020). The initial step is to use an inductive approach to efficiently discover interesting and valid patterns within a large corpus of text using unsupervised learning methods (e.g., lexical selection or topic modeling). After researchers conduct a guided deep reading of the text data to confirm computer-detected patterns as valid, the research transitions into deductive natural language processing tools.

1.2. *Knowledge discovery and big data*

Recorded data is growing at an unprecedented scale in industry, government, and civil society. Analysis and distilling knowledge from big data now drive many aspects of our society, including retail, financial services, insurance, wireless mobile services, business management, urban planning, science and technology, governance, law, social sciences, and humanities.

Big data is a multifaceted and complex concept consisting of information, technology, methods, and impacts (Dumbill 2013; Mauro et al., 2016). The foundation of big data is digitalized, compiled, and stored information from the internet, retail records, tax forms, recording devices, and texts (Seife 2015). The enormous size and high complexity of the data require computer storage, data processing, and data mining technologies to process, store, manage, and facilitate data analyses. Distributed data storage, cloud computing, data mining, and artificial intelligence are required technology components of big data (Manyika et al., 2011). Big data require new analytical methods beyond the traditional statistical approaches to discover new knowledge from the data mine. They include a series of new and old processing and analytical methods such as language processing, neural networks, network analysis, pattern recognition, predictive modeling, spatial analysis, statistics, supervised and unsupervised learning, and simulation (Manyika et al., 2011). Big data has impacted many dimensions of our society and will continue to bring changes to laws, guidelines, and policies on the utility and management of personal information. Insights gained from big data have revolutionized how we conduct business, governance, research, design, production, human interactions, and daily life.

Approaches to data mining contribute to research issues in big data in several ways. Data mining provides automatic and semi- automatic data cleaning, compressing, and visualization methods to preprocess big data. Machine learning methods enable analyses of heterogeneous, multimedia, unreliable, and contradicting data.

Unsupervised machine learning aims to make sense of a large amount of data by revealing patterns, correlations, and classifications to reduce measurement dimensions, while supervised learning derives interactions and linear and nonlinear relationships among high-dimension data effectively and efficiently.

1.3. *Knowledge discovery and data mining in computational social science*

Knowledge discovery in the database (KDD) emerged from the necessity of analyzing big data. KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad et al., 1996, page 84). Data mining techniques and terminology draw from three academic fields. The oldest source is statistics, which provides well-defined techniques to identify relationships between variables systematically. Data visualization, descriptive statistics, correlation, frequency tables, multivariate exploratory techniques, advanced linear and generalized linear models, etc., are classical methods in statistics. A second and modern source of knowledge discovery technique is Artificial Intelligence (AI) of machine learning and Artificial Neural Network (ANN). Machine learning “trains” computers to recognize patterns in data. ANNs model after human brains as structures of numerous interconnected processing elements working in unison to process information (Shu 2020). The last foundation is database systems that allow it to store, access, and retrieve huge amounts of data, thus providing support for the platform for information processing and mining.

In the social sciences, computational social science (CSS) emerged as a new interdisciplinary area of research from the synergy of domain knowledge in the social sciences, information technology, big data, data management, social computing, and transdisciplinarity (Lazer et al., 2009; Mason et al., 2014). It involves collaboration and coordination among scholars from different disciplinary training: social scientists provide insights on research knowledge, deciding on data sources and collection methods, while statisticians and computer scientists develop appropriate mathematical models and data mining methods and provide computational knowledge and skills. CSS is based on collecting and analyzing big data and using digitalization tools and methods such as social computing, social modeling, social simulation, network analysis, online experiments, and artificial intelligence to research human behaviors, collective interactions, and complex organizations (Watts 2013). CSS provides an unprecedented ability to analyze the breadth and depths of vast amounts of data, thus affording us a new approach to understanding individual behaviors, group interactions, social structures, and societal transformations. Computational social science methods consist primarily of social computing, online experiments, and computer simulations (Conte 2016). Social computing uses information processing technology and computational methods to conduct data mining and analysis to reveal hidden collective and individual behavioral patterns. Online experiments use the internet as a laboratory to break free of the confines of conventional experimental approaches and use the online world as a natural setting for experiments that transcend time and space (Bond et al., 2012; Kramer et al., 2014). Computer simulations use mathematical modeling and simulation software to set and adjust program parameters to simulate social phenomena and detect patterns of social behaviors (Bankes 2002; Gilbert et al., 2006; Epstein 2006).

Data mining constitutes an important method in computational social science (CSS) by providing methods of social computing and online experiments and computer simulations (Conte 2016; Watts 2013). Knowledge discovery approaches contribute to social computing through its information processing technology and computational methods to conduct data mining to reveal hidden patterns of collective and individual behaviors and shed new light on causal discovery and theory innovation.

2. Knowledge discovery and data mining

2.1. New insights from predictive analyses

The data mining approach automatically or semi-automatically implement elements of causal inquiry (Morgan and Winship 2015; Brand et al. 2023). The popular classical statistical models in social sciences tend to analyze the average impact of explanatory variables on all cases. The classical statistical models often do not routinely recognize or evaluate causal heterogeneity when the same predictors generate differential outcomes for different groups with different combinations of other conditions and methods. They are also less likely to consider independent alternative causes when these causes are of little interest to the specific research field of the participating researchers. Since the data mining approaches focus on their models' prediction and predictive power, they automatically and semi-automatically consider a larger number of joint, interactive, and independent predictors, often effectively improving the predictive power of models over traditional single-predictor models. Both classical statistical approaches and data mining play the same role in scientific research: providing information on correlations among variables or associations among entities. In addition, data mining can efficiently filter complex and multiple correlations among variables to help us identify complexity, interactions, and heterogeneity in discovering potential causal relationships. Domain experts should be able to elaborate on the causal mechanism based on the associations between variables discovered from the data. Data mining does not challenge the conventional model-building approach. Rather, it plays an important complementary role in improving model goodness of fit, revealing valid and significant hidden patterns in data, and enriching our discovery.

Data mining approaches can be used to identify heterogeneous treatment effects automatically or semiautomatically by differentiating subpopulations, matching the treatment group and the control group on their likelihood of being in the treatment group, and addressing potential omitted variable bias (Brand et al., 2023; Molina and Garip 2019). Decision trees minimize the errors in treatment effects to estimate treatment effects for subgroups (Athey and Guido 2016), and random forests calculate the average across trees to allow individualized treatment effects (Wager and Athey 2018). To match treatment and control groups, researchers estimate and match the propensity scores indicative of the likelihood of being in the treatment group conditional on other observed inputs using neural networks (Westreich et al., 2010) and regression trees (Diamond and Sekhon 2013; Wyss et al., 2014). Machine learning approaches can measure sensitivity to misspecification to address the impact of unobserved variables correlated with the treatment and outcome (Athey and Guido 2015) and generate instrument variables using neural networks (Molina and Garip 2019).

Data mining contributes to discovering potential causality in multiple ways (Shu 2020). Data mining uncovers new, sometimes unexpected, ways of conceptualization conducive to theory innovation and knowledge discoveries. Confirmatory analysis of model building tends to emphasize some specific causal mechanisms as a way to test one or a few theories. These models focus on a small number of explanations consisting of simple, functional forms because such models are considered straightforward, parsimonious, elegant, and theoretically appealing. These attractive features aside, such models also suffer from some flaws. Such simple single-cause models assume an average effect linking the input variable to the output, often ignoring to shed light on the mechanism of causal heterogeneity. On the one hand, using one or two theories can often provide partial explanations for the outcomes and does not exhaust all the explanations; thus, such models usually have low predictive power. On the other hand, data mining is concerned with offering a thorough or complete (to the extent of available data) account of the event under investigation. This approach does not shy away from a rich analysis of multiple, complicated, and nuanced explanations as they all contribute to the strong predictive power of the resulting model. Data mining models provide a full account of the event or outcome to the fullest permitted by information buried in data to provide maximum accuracy in prediction for the future. Both approaches have advantages and disadvantages. The data mining approach provides an alternative and complementary in aiding causal discovery.

2.2. *Features of data mining*

Although data mining will never replace the confirmatory analysis of model-building commonly used in the social sciences, nor it is contrary to the current models, they are different in multiple ways. Data mining emphasizes complex causal heterogeneity, considers a variety of nonlinear and joint effects, and tends to estimate complex and elaborate models. Data mining aims *to* maximize a model's predictive power, provides methods capable of analyzing non-numerical data such as text, images, and voice, and uses convenience samples, but sometimes is unable to provide information on the causal mechanism. Table 1 compares data mining characteristics with the classical approach. Our purpose is to compare and contrast the classical statistical approach vs. data mining to provide insights into the contributions of data mining as complementary to the classical approach. Because classical statistics is the foundation of data mining and machine learning, it is not possible to completely separate these two fields as they have been mutually influencing, inspiring, learning, and adopting each other during their development and evolution. We strive to accurately reflect the characteristics and roles they each play based on the most "typical" or "common" social science practices in the two traditions.

2.2.1. Complex causal heterogeneity

The machine learning approach assists with the evaluation of heterogeneous treatment and treatment effects. Some researchers use propensity-based heterogeneity analysis to overcome the limitation of the traditional regression models, but this approach has several shortcomings: it depends on researchers' existing insights of heterogeneity that may overlook unexpected sources of meaningful group heterogeneity; it cannot exactly pinpoint which variables are the source of heterogeneous effect because the propensity is measured as a summary of all the confounding variables, and it introduces uncertainty to the model and estimation. Machine learning approaches overcome some of these limitations by considering all the interactions between the treatment and confounding variables instead of preselecting confounding variables and presetting models (Brand et al., 2021; Hu et al., 2021; Lundberg and Brand 2022). For example, decision trees enable researchers to identify subgroups that respond differently to treatments by partitioning the sample to minimize heterogeneity in within-leaf treatment effects. This can be achieved through their ability to search over high dimensional functions of covariates and their interactions (Brand et al., 2021). Such practices follow the logic of treatment and notreatment, "matching" and "simulation" to approximate individual treatment effects, and these procedures are usually automatically implemented in ML.

Data mining automates searching and evaluating heterogeneous joint predictors. New data mining techniques can generate and estimate thousands of interactions and combinations among predictors to achieve high prediction of outcome variables. For example, association rule mining performs well in identifying associations among combinations of attributes in the form of multivariate associations that provide novel or unexpected combinations of prior conditions that co-occur before the outcomes. Discovering such combinations in the data mine directs researchers to investigate interesting joint effects among explanatory variables. Decision tree programs automatically consider large numbers of combined effects among predictors to improve the prediction models. Neural networks also automatically generate functions equivalent to interactions or combined effects among predictors to substantially increase the predictive power of their models. Data mining also provides the tools to detect causal heterogeneity. Unsupervised machine learning uncovers hidden patterns and classifies the data such that different predictor variables are applied to different sub- populations. For example, an unsupervised learning approach revealed four primary waves of Mexican immigrants, some of which had never before been detected (Garip 2017). After clustering immigrants into different groups and waves, it becomes evident that their migration actions result from different pulling and pushing factors. These findings shed light on the causal heterogeneity in the Mexican immigration process and give rise to different theoretical accounts of these diverse experiences. Past research has tested migration theories on all Mexican immigrants without recognizing their causal heterogeneity, and it is no surprise that they have produced confusing evidence. By discovering the hidden pattern of heterogeneity among migration waves and groups, these findings reconcile some of the ongoing theoretical debates that have produced conflicting results. Instead of applying these theories universally to account for all migration patterns, we are informed by the results from machine learning that theories should be applied selectively to different population segments at different historical times as the explanatory power of divergent theories is bound by the type of population, the larger context, and historical times.

Table 1

Characteristics of data mining and classical approaches.

Data Mining	Classical/Conventional Approach
Complex causal heterogeneity	Average main effect
Nonlinear and joint effects	Linear main effects
Complex & “complete” models	Parsimonious/“simple” models
Maximizing model prediction power	Estimation of predictor coefficients
Convenience Samples	Probability Samples
Non-numerical data (text, images, voice, etc.)	Non-numerical data belong to qualitative methods
Function form among variables could be opaque in some methods	Function forms among variables are easy to interpret

2.2.2. Nonlinear and joint effects

Data mining provides automatic or semi-automatic tools to search for nonlinear relations and increase model prediction accuracy. Data mining procedures can automatically generate breakpoints for continuous independent variables to capture a nonlinear effect between the dependent and independent variables. Automatic binning and discretization of variables that reflect nonlinear effects. Optimal binning by predictive value involves partitioning the variable based on the predictive values of a target variable by choosing the boundaries for each bin to differentiate the cases in each bin from the other bins in terms of their functions on the outcome variable (Witten et al., 2011). This helps increase the prediction accuracy of each of the explanatory variables and thus the power of the entire model. Visualization tools enable researchers to view the relationship among variables from any angle with any rotation to detect nonlinear relations and to visualize how values of an outcome variable are affected jointly by changes in the value of other variables in a model.

2.2.3. Complex and "complete" models

To achieve the purpose of successful prediction with new data, data mining models can be complex, large, and computationally costly. The measures of model performance pay scant attention to the size and complexity of models and thus tend not to penalize those large and convoluted models (Shu 2020). The most typical method to evaluate models for prediction purposes is cross-validation. Researchers divide data into several training sets to build the model and a testing set to verify if the model can predict the outcomes. Mean squared differences between the observed and predicted values are used to measure the prediction accuracy. Researchers repeat this procedure multiple times using the training sets and calculate the mean squared difference overall iterations to calculate a general value of the standard deviation. Data mining uses measures calculated from confusion tables to evaluate model performance and select classification schemes for discrete outcomes. To measure the overall model performance, researchers use the product of sensitivity and specificity, while accuracy, error rate, sensitivity, and specificity are measures of specific dimensions of models. Cost-benefit matrices are used in conjunction with confusion tables to calculate profit to evaluate model performance. A confusion table in combination with the cost-benefit matrix can also produce a series of curves, including the cumulative response curve, life chart, ROC curve, and recall-precision curve, to aid in measuring model performance. Machine learning is complementary to classical statistics, not superior nor a replacement. ML offers a different epistemology, new perspective, and alternative approach to informing causal discovery. Estimating "complex and complete" models may provide new insights and ideas to the current research model, but not necessarily without the costs of large complex models, difficulties in interpretation, and computational complexity, to name a few.

2.2.4. Model prediction power

Data mining emphasizes accuracy in model prediction. An important measure of success in data mining models is their ability to accurately predict outcomes in real-world applications because data mining originated from Artificial Intelligence's preoccupation with applied predictive models, such as prediction of insurance fraud, illness diagnoses, pattern recognition, etc. (Shu 2003; 2020). To achieve this purpose, data mining uses a variety of approaches to make predictions more accurate to achieve this purpose. By combining different methods to maximize the overall predictive power, data mining uses various tools from machine learning, AI, database, and statistics on a vast amount of data with complex models to provide much more predictive power than conventional statistical models (Shu 2003).

2.2.5. Convenience samples

Data mining employs several methods to reduce sampling bias associated with data from "natural" samples. "Convenience" samples are usually records of nonrandom human activities that are neither population censuses nor well-designed probability samples. They are often obtained from online databases documenting activities such as online purchase records, Twitter posts, fitness records from track devices, GPS records of mobility, credit card purchases, Uber or Lyft travels, Facebook interactions, and insurance claims. These "natural" data also evolve with social and historical dynamics, thus reflecting different "population" sampling schemes. Their relationship with the population is unknown. We don't know if samples are representative of the population or the probability of each member of the population being included in the sample. As a result, traditional significant tests are not suitable for these "convenience" samples. By no means convenience sampling bias can be resolved without using a representative random sample However, researchers can employ several methods such as multisource empirical replication and cross-validation methods to mitigate these issues.

Empirical replication is the primary way to reduce bias in convenience sampling. Researchers use different types of data to pragmatically test, verify, and modify the model to prevent overfitting and guarantee its reliability, validity, and generalizability (Peterson and Merunka 2014). Although replication using data from a different data source does not correct the errors in the original dataset, a replication analysis provides empirical evidence for the original investigation. A study has shown that efficiency can be significantly improved by incorporating the primary convenience sample with a relatively small, presumably less expensive, random sample (Hedt and Pagano 2011). Finally, when random samples are unavailable, researchers can pool multiple convenience samples from the same population of interest into one more generalizable meta-sample (Winton and Sabol 2021). For example, a multi-source meta-sample can consist of convenience samples such as students, crowdsourcing, professional panels, and social network sites. Researchers may compare sample results from each source to findings from the meta-analytic (Winton and Sabol 2021).

Cross-validation divides datasets into training samples to train the model, tuning samples to adjust and modify the model, and testing samples to verify the model. Multiple rounds of cross-validation may be performed using different sample partitioning, and the results are averaged over the rounds. If researchers are interested in generalizing their findings to a population, cross-validation may not be able to reduce the convenience sample problem. When primary data collection is impossible, cross-validation can be the only option to reduce sample bias by mimicking the replication process using simulations.

The importance of the convenience sampling problem largely depends on the research goal, target population, and computational capacity (Glymour et al., 1997; Kim et al., 2018). When researchers try to create an easily computable representation of how the data are distributed in a particular database, using a convenience sample is more tolerable because the goal is to be fast and "convenient." On the contrary, data mining needs to pay more attention to the sampling issues when the goal is to provide a basis for policy or develop a predictive model. Some institutional researchers may not need to be concerned with generalizability because they usually have access to the entire institution database and are only interested in learning about such a population in the institution. Such a sample is the population, and external validity is not a concern. Access to the total population data is not uncommon in some "big data" studies where sample and population data are identical.

2.2.6. Variety of data (numerical, text, images, video, and audio)

In addition to numerical data, data mining provides tools to preprocess unstructured data and various methods and models to mine non-numerical data. Text mining and linguistic analysis of digitalized books, journal bibliographies, online searches, news portals, print media, TV broadcasts, online media, and online forums have emerged as some of the most innovative and stimulating research areas in the social science and humanities in the last couple decades (Moody 2004; Michel et al., 2011; Van de Rijt et al., 2013; Nelson 2021; Provost and Faucett 2013). These data types are complicated with varying lengths, certain order, and messy structure and are plagued with ungrammatical sentences, misspellings, unexpected abbreviations, random punctuations, domainspecific terminologies, and context. Procedures of pre-processing data play a vital role in converting messy and unstructured inputs into formats conducive to data mining algorithms. Data mining algorithms can treat individual words as terms (bag of words), include sequences of adjacent works as terms (N-gram sequences), recognize common named entities (name entry extraction), and model the set of topics as clusters of words (topic models) (Shu 2020). Using a database of over 136K protests and 300K Weibo images of protests in China, Zhang and Peng (2022) preprocessed these images to size and pixel represenation conducive for extracting intermediate representations in preparation for cluster analysis.

2.2.7. Causal mechanisms

Some data mining models are opaque about the complex relationships among variables. Although data mining can provide a great deal of prediction power, some models, such as random forest and deep-learning neural networks, are so complex that it is unclear how the input variables are connected with the output. It is sometimes unattainable to derive causal relationships from the connection between theoretical framework and empirical evidence from the data from certain data mining models. A complex multilayer neural network can be very successful in predicting outcomes, but it often sheds little light on the causal process and the relative influences of each variable on the outcome. As powerful as their predictive power in dealing with a variety of data and information, especially in pattern recognition, some AI models process the relationship between predictors and outcomes in a "black box." The actual causal mechanisms, including which predictors act, in what ways, and by how much, are usually invisible. To make these machine learning models useful for social science research, we must make the causal processes in these models perceptible and meaningful. The primary interest of social science research is the detection of important predictors, their functions in this process, and insights into the causal mechanism. One such effort is sensitivity analysis procedures to measure input variable importance in predicting the outcome variable to shed light on the relative roles played by the input variables (Shu 2020).

3. Knowledge discovery and machine learning

3.1. Machine learning: supervised and unsupervised

Machine learning refers to the process in which computer systems use computer science and statistical techniques to progressively improve their performance on data analytical tasks by learning from data (Samuel 1959). Machine learning builds models and algorithms by learning and improving from data. Machine learning is best for computing tasks when the explicit model structure is unclear and algorithms with good performance are difficult to attain. When researchers lack complete information to design explicit models to specify the relationships among variables or cases, they learn from the data to discover the hidden patterns and structures.

They derive models from these patterns and structures to inform theory building.

Machine learning encompasses two main types of tasks: supervised learning and unsupervised learning. The main difference between these two types of machine learning is that supervised learning starts with knowledge of the output values. In contrast, unsupervised learning does not have explicit outputs to predict. Supervised learning aims to find and learn a computational function that best approximates the relationship between attributes and the outcome variable in the data. Unsupervised learning does not have explicit outputs to predict, and its objective is to infer and reveal the hidden structure in data (Shu 2020).

3.1.1. Unsupervised machine learning

The goal of unsupervised learning is to model the underlying structure from data when researchers have only input data and do not have corresponding outputs or outcome variables. The process is called unsupervised learning because, in the absence of explicit outcomes to predict, no supervision or teaching takes place. Without feedback from data, algorithms learn on their own to discover interesting structures in the data.

Unsupervised learning includes two types of problems—classification (cluster analysis, latent class analysis, and topic modeling) and association (association analysis and sequence analysis). They are similar in that they all analyze connections within the data, but differ in the type of connections on which they focus. On the one hand, cluster analysis, latent class analysis, and topic modeling emphasize the links among data instances or words and aims to discover inherent groupings of data cases or words. The clusters, latent classes, and topics are groups of individuals or subjects, or words sharing similar traits. On the other hand, the association is interested in the connections among attributes and attempts to reveal rules that describe relationships between variables in the data. Associations describe strong links or connections between traits or characteristics of the subjects in the data.

3.1.1.1. Unsupervised classification: cluster analysis, latent class analysis, and topic modeling. Cluster analysis, latent class analysis, and topic modeling are some of the more frequently applied methods of unsupervised classification. These methods separate people, organizations, products, behaviors, attitudes, objects, words, cities, or countries into relatively homogeneous subgroups of clusters or latent classes or topics. A cluster, a latent class, or a topic is a collection of cases or words similar to each other and dissimilar to cases in other clusters or classes. We assume that these clusters, latent classes, and topics reflect some underlying mechanism that causes some cases to bear a stronger resemblance to each other than to the remaining cases, which belong to other clusters, classes, or topics. With an enormous amount of data, it is also beneficial to use unsupervised learning methods to reduce high-dimension data, thus decreasing the search space for the downstream algorithm.

Cluster analysis and latent class analysis are usually performed as preliminary or preparatory steps in a data mining process. The resulting clusters or classes are then used as inputs in subsequent analytical procedures either as predictors or outcomes. Because of the complex composition of traits that jointly shape groups of similar cases, it is useful to differentiate subjects into groups sharing similar traits to analyze a package of characteristics simultaneously. By grouping data instances into clusters, classes, or topics of a series of characteristics, these types of unsupervised learning perform dimension reduction by summarizing high-dimensional data, which is beneficial for additional analyses and interpretations.

Another advantage of first allocating subjects into different clusters or classes and next analyzing these groups allows researchers to test theoretical ideas and observe heterogeneous outcomes in these groups. Thus, they can derive under what circumstance and to which group some theories hold while others ideas do not apply. Instead of mixing all cases and applying theories indiscriminately to all cases, cluster analysis and latent class analysis can effectively aid in the process of separating causal heterogeneity and informing theory development. Cluster analysis and latent class analysis assign membership to cases by maximizing the similarity among the group and minimizing the similarity between groups. They both see clusters or classes as reflecting some mechanism that causes some instances to resemble each other. They are useful methods to differentiate subjects into groups with similar traits and enable researchers to analyze simultaneously a package of characteristics to account for causal heterogeneity. These two approaches are different because latent class analysis allows for a more flexible framework than cluster analysis. The latent class method allows a different error structure that does not assume equal variances across classes or covariances between variables are zero, as cluster analysis does (Hagenaars and McCutcheon 2002)). The latent class method also provides a broad methodology framework that allows the incorporation of background information as covariates, directional relationships between variables as expressed in causal models, and transition in class membership over time (Muthen 2004´).

Cluster analysis requires that all variables in the analysis be normalized so that distance can be measured appropriately. Range normalization and Z-score standardization are the most commonly used methods. Cluster analysis uses measures of similarity and distance to divide the instance space into regions based on the closeness of cases in the instance space using at least five ways to measure such distances: Euclidean, Manhattan, Jaccard, Cosine, and Edit distances (Deza and Deza 2006). Euclidean distance uses the Pythagorean Theorem to calculate the distance between two points in space. Manhattan distance sums up the differences along all the dimensions between the two cases. Jaccard distance measures the proportion of all the characteristics not shared by two cases. Cosine distance measures the distance between two documents by representing the difference in the proportion of occurrence of each word in a text regardless of the text length or word count. Edit distance measures the number of edit operations required to convert one string or sequence to make it identical.

Cluster analysis divides the instance space into clusters. The two most commonly used algorithms are the methods of K-means and hierarchical cluster analysis. The K-means approach focuses on the clusters by representing each cluster by the geometric center (centroid) of a group of cases assumed to be in the same cluster (MacKay 2003).

Hierarchical clustering sees clusters as formed in a hierarchical order by grouping cases based on their similarity. Both methods are based on a choice of linkage functions to determine the distance between clusters: single linkage, complete linkage, or average linkage. The single linkage is the nearest neighbor approach determined by the two most similar cases from the two clusters. This approach tends to result in long and slender clusters, with some heterogeneous cases clustered together. Complete linkage uses the farthest-neighbor approach, the maximum distance between two cases in the two clusters. This approach thus tends to result in compact and sphere-shaped clusters. Average linkage uses the average of all cases in one cluster from all the cases in another cluster and forms clusters that tend to have approximately equal within-cluster variability.

Frequently used methods to measure cluster goodness and model validation are the pseudo-F test, Silhouette value, and cluster validation. The pseudo-F test measures the ratio of the distance between the clusters versus the spread of cases within the cluster. It is often used to find the optimal number of clusters. The Silhouette value gauges the goodness of the cluster assignment for a particular case in the data after considering cluster cohesion and separation. Cluster validation safeguards the analysis from random noise to ensure that the model has good generalization and avoids model over-overfitting by splitting the data into a training set and testing set to compare model performance (Reitermanova 2010).

The unsupervised clustering method has been used in migration studies to identify unique migrant types and migration behaviors (Bail 2008; Garip 2012; 2017). Garip (2012) employs a K-means clustering algorithm, which reveals four distinct Mexican-to-United States migrant clusters that would otherwise be overlooked by a supervised regression method. In Garip's study, each cluster of migrants possesses a specific configuration of individual, household, and communal characteristics corresponding to a specific migration theory. The clustering analysis helps visualize multiple migration mechanisms that work simultaneously for different segments of the migrants.

Scarborough and Sin (2020) use hierarchical cluster analysis to study how gender norms vary across geographical locations. The model classified commuting zones into clusters with similar gender norm dynamics. This inductive approach shows that gender norms can be conceptualized into four types rather than the conventional differentiation of traditional versus egalitarian attitudes. These geographical clusters also demonstrate that geographical locations have contextual effects on gender norm dynamics.

Clustering analysis extends to other types of complex data, such as images. Zhang and Peng (2022) employ an unsupervised image clustering method to classify images about protests in China and climate change. They use a pre-train transfer learning model that groups images into several "meaningful" clusters. This model distinguishes a cluster of images of people in rural environments from images of people in urban environments or images of landscapes without people. Although image clustering does not fully replace human interpretation of images, it helps social scientists quickly capture theoretically significant categories from large-scale image databases (Zhang and Peng 2022).

Latent class is another approach to classifying cases (Bacher 2000; Hagenaars and McCutcheon 2002; Gorunescu 2011). Like cluster analysis, the latent class method assumes an underlying statistical model that can be used to identify groups of individuals that are similar to a categorical latent variable. However, it differs from cluster analysis in two ways. The variables used in the latent class analyses consist of two components: the true measures at the latent level and measurement error partitioned into within-class residual variance. Each case is assigned to one class or group with a probability, recognizing the uncertainty in the classification. These probabilities are computed from both the model and the patterns of observed scores (Hagenaars and McCutcheon 2002). The latent class analysis usually uses two iterative estimation functions of the maximum likelihood method and the maximum- posterior method to assign cases to the classification scheme of an underlying latent variable. They both derive log-likelihood functions from the probability density function underlying the latent class model (Hagenaars and McCutcheon 2002). The estimation procedure produces model parameters, including the probability of membership in latent classes, means, variances, and covariances for each latent class. The optimal solution is the one with the minimum number of classes possible while achieving an acceptable model fit (Clogg 1995). Several criteria are often used to evaluate model fit. First, the Akaike information criteria (AIC) and the Bayesian information criteria (BIC) are heuristic indices that use "parsimony criteria" to offer comparative evidence to evaluate different solutions. Second, entropy, computed as the maximal probability density distribution underlying the latent class cluster model, indicates how well the model predicts class memberships (Akaike 1977). Third, a better-fitting solution will have a higher average class membership probability across all classes. Last, substantive meanings of each latent class should be valid and meaningful as informed by centroid and descriptive information for latent classes.

Social scientists extensively use latent class analysis to identify hidden heterogeneity among subpopulations. Lee and others (2017) used LCA to identify distinct profiles of childhood abuse victims. The latent class variable in their study is the combination of the severity and multiple types of abuse, e.g., emotional, sexual, or physical abuse. They found that women and lower socioeconomic status individuals were exposed to a greater risk of severe and multiple types of abuse. In a geopolitical study, Soehl and Karim (2021) use LCA to identify four latent classes of nationalism (ardent, disengaged, liberal, restrictive) from four dimensions (identification, exclusionism, pride, and hubris) and 26 measures of nationalism. The LCA reveals a qualitative difference between the two groups: one takes pride in their nation while having an inclusive attitude, and another is not patriotic while having an exclusive attitude toward outsiders. These two unexpected latent groups may never be revealed if the categories had been predetermined. This study concludes that a nation-state with turbulent geopolitical history is less likely to have restrictive nationalism (less pride in the nation and anti-immigration). A different study on American nationalism also used the latent class method to uncover two additional nuanced forms of nationalism (Bonikowski and DiMaggio 2016): one group of young and educated Americans with unpatriotic sentiment and another group of older and less educated Americans embracing all forms of nationalism. Gender scholars have employed LCA to identify new dimensions of gender norms and gender attitudes that have not been previously conceptualized. Scarborough, Sin, and Risman (2019) used LCA to study whether gender egalitarianism has grown with each generation or if Millennials are stalling the gender revolution. LCA models identified six latent classes based on four gender attitude indicators: women holding public office, women's family social roles, and two indicators about maternal employment on children's wellbeing. They labeled these latent classes as strongly egalitarian, egalitarian, strongly traditional, traditional, pro-public anti-private ambivalent, and anti-public pro-private ambivalent. The last two ambiguous latent variables captured respondents with mixed response patterns - e.g., those who support women in the public sphere but oppose gender equality in the private sphere or vice versa. They found that although younger generations are becoming more liberal, Millennials are slightly more likely to be anti-public ambivalent than previous generations. The latent classification revealed that gender revolution has a diverse pathway that older generations of traditionalists were first replaced by ambivalent before they transitioned to be more egalitarian. In another study, Scarborough and others (2021) applied LCA to study the intersectionality of gender and racial attitudes based on a set of eight established survey items. Aided by goodness-of-fit tests, LCA models identified four salient racial and gender attitudes clusters among those eight indicators: racial structuralism/gender egalitarianism, new racialism/gender egalitarianism, racial structuralism/gender ambivalent, and new racialism/gender traditionalism. The classification of racial and gender attitudes reveals that individuals with liberal gender attitudes do not necessarily uphold egalitarian attitudes on race.

Topic modeling is a method of unsupervised classification of a corpus of documents that we want to classify into natural groups. This approach is similar to unsupervised cluster analysis, which identifies “natural groupings” of cases using numerical data. The main idea of the topic model is to model the set of topics in a corpus separately. Each document constitutes a sequence of words, which are used as inputs to a classifier to be mapped to one or more topics. These topics are learned from the word data via unsupervised data mining. Documents are then characterized in terms of these topics and the component words.

There are two general methods for creating topic models: matrix factorization methods (e.g., latent semantic indexing) and probabilistic topic models (e.g., latent Dirichlet allocation). Latent Dirichlet allocation (LDA) is a popular method for fitting a topic model. It is a mathematical method for estimating both relationships simultaneously: the words associated with each topic, and the topics that describe each document. LDA sees each document as consisting of various topics, and each topic consists of various words. Using words and documents as input, LDA learns to identify the topics through unsupervised data mining. LDA allows documents to consist of a mixture of topics, overlapping with each other in content as measured in both words and topics. LDA follows two principles: every topic is perceived as a mixture of words, and each document is seen as a mixture of topics. Each document contains many words which are from several topics in different proportions.

LDA topic modeling has been used to map out the large-scale social media content of mental health discourses on Twitter (Pavlova and Berkers 2020). The topic modeling method detects each mental health topic’s contextual theme and topic. It also analyzes the mechanisms through which each discourse becomes popular and unpopular. Traditional media such as newspapers were more likely to frame mental health discourse with a stigma-related theme, labeling a mental health hospital escapee as a “violent psychopath” rather than a “patient”. User-generated content that focuses more on the *feeling* is correlated with positive sentiments, higher engagement, a diverse audience, and online solidarity.

The topic modeling method can be innovatively combined with event history analysis to study U.S. sociology Ph.D. students’ career success (Heiberger et al., 2021). Based on 80,000 sociology dissertations and publications, it analyzed how students’ specialization strategy (e.g., topic choice, focus, novelty, and consistency) affected their career outcomes. Topic modeling detects themes from the qualitative data and thematic combinations (novelty) and trends (consistency) over time. Students whose theses are closely related to identity, statistical method, and race are associated with a higher chance of becoming mentors, indicative of achieving an influential career, and so does a consistent publication theme.

3.1.1.2. Unsupervised Co-occurrences: association rule learning and sequence analysis. Co-occurrence grouping is a method for finding associations between entities. The study of associations between characteristics is called association rule learning, affinity analysis, or market basket analysis. It seeks to discover affinities among traits and find rules for quantifying the relationship between these attributes. It originated from marketing research on consumer behavior based on the theory that if consumers buy a certain group of items, they are more (or less) likely to buy another group of items.

Association rule learning is a powerful tool for detecting relationships among the outcome and numerous candidate explanatory variables in large amounts of data (Luma-Osmani et al., 2020). When the dataset is broad and complex, the search space for combinations of attributes is large. Association rule learning is very good at identifying associations among combinations of attributes in the form of multivariate associations. It can lead researchers to discover which attributes and combinations of attributes occur most frequently with some outcomes. Such discoveries of combined effects of several variables that collectively lead to some outcomes, expected and unexpected, direct researchers to further investigate interesting joint effects among variables. These Association rules provide useful leads on the direction of the causal relationship between the antecedents and consequents. Combined with other methods, they are effective tools in suggesting fruitful directions in the search for potential causal relationships and aid in theory innovation.

Association rules consist of three parts: a pair of antecedent and consequent events, support, and confidence. The *antecedent* and *consequent* are two events or attributes that occur in sequence. The support is the proportion of times when both antecedent and consequent occur. Confidence is the proportion of times that both antecedent and consequent appear when the antecedent occurs. Two indices measure the strength of association. *The lift* measures the frequency by which an association occurs compared to random chance, which is the ratio of the probability that both antecedent and consequent occur together versus the probability that they occur randomly together. *Leverage* is the difference between these two probabilities rather than the ratio.

In order to handle a large number of potential rules, the association rule algorithm reduces the search space using an a priori algorithm (Hand et al., 2001). The association rules rank the rules according to their prevalence and confidence, gauged by the measure of usefulness which is the product of support and confidence. Two comparisons are used to account for the prior probability.

First is a comparison of the advantage of using the association rule over not using it by converting lift into a *confidence ratio*. The proportion of the improvement from using the association rule over not using it. The second approach is calculating the *confidence difference*, which is the difference between the proportion of improvement from using the association and that of not using it.

Association rule learning is advantageous over correlation and conventional multivariate analysis in several ways. First, association rule learning can generate the rules automatically and semi-automatically in the order of association strength. A conventional correlation matrix presents associations in the order of how the variables are entered into the analysis; researchers need to manually evaluate all the associations, which could become unattainable when the number of variables is large, and the correlation matrix is vast. Second, association rule learning effectively identifies associations among combinations of attributes and generates association rules in multivariate associations, thus providing insights into novel combinations of prior conditions that co-occur before the outcomes. In contrast, a correlation matrix can only provide bivariate correlations. Third, association rules learning is nonparametric statistics that is more flexible and can be applied to a variety of data with various distributions. At the same time, conventional multivariate models such as correlation and generalized multiple regression analyses require the assumptions of multivariate normality homoscedasticity and independent residuals. Lastly, unlike regression analysis, in which dichotomous variables need to be mutually exclusive or exhaustive and produce effect parameters that must be interpreted relative to an omitted category, association rule learning can take dichotomous variables with any combination of attributes that may be interesting and relevant.

Association rule learning has been used to investigate the multiplexity of cultural meanings in social networks (Gondal 2022). This study investigates how different relationship roles (e.g., parents, siblings, coworkers, neighbors) and foci (workplaces, neighborhoods, and group memberships) change the meaning of interpersonal relationships. For example, an alter classifies their advisor as a friend. If that advisor has a spouse who works in the same organization, the alter will likely perceive the advisor’s spouse as a friend as well. In this case, the antecedent and consequence relationship is (advisor * spouse * same workplace → friend). The arrangement of different relationship roles implies different group dynamics and inequalities.

Sequence analysis (Blanchard et al., 2014) is a variant of association rule learning that can identify sets of events that generally occur in sequence. Associations may also include a time lag between elements such as statuses, events, objects, or actions as it is not required for all the items to occur simultaneously. The positions of the elements are fixed and ordered by either passage of time or by some other natural order. The method can be adapted to analyze a sequence of events over time. While antecedents and consequents are inputs and outputs, in association rule learning, the set of traits or actions (called itemset) leading to the final item set forms the antecedent sequence. In contrast, the last item forms the consequent sequence, and they are combined and treated as an entity of their own. Instead of searching for antecedents and consequents as in association rule learning, sequence analysis mining ordered items based on the time sequences or any natural order in sequences.

An alternative technique from association analysis for comparing sequences is optimal matching (Blanchard et al., 2014). OM defines the distance between two sequences as the ‘‘Levenshtein distance’’ (edit distance as introduced in cluster analysis), the number of operations that transform one sequence into the other (Levenshtein 1966). The resulting distance matrix measuring similarity or dissimilarity between each pair of sequences is then used as input for a cluster analysis or multidimensional scaling. Using the OM approach to convert the dissimilarity between pairs of sequences into numerical values often has a little theoretical basis for assigning different values to transition between the sequences.

Sequence analysis is different from time-series analysis and event history analysis. Time-series analysis assumes that characteristics and actions at earlier times have bearings on characteristics and actions at later times and that these trajectories of influences are explicitly specified in the models by researchers. Sequence analysis allows data to present sequences, a series of ordered statuses/ traits/actions, to emerge from the data without a preset notion of the pattern. Unlike treating the last status/trait/action as the ultimate outcome variable, sequence analysis regards the series of ordered statuses/traits/actions as individual entities of their own, and the research focuses on the ordered character of all elements together. Similar to time-series analysis, event history analysis assumes that events are stochastically generated from earlier to later times, while sequence analysis views the data holistically, treating sequences as whole entities. Both time series analysis and event history analysis derive probabilistic inferences of the causal mechanisms between preset covariates and outcomes of longitudinal nature. In contrast, sequence analysis uncovers patterns of sequences and measures the dissimilarity of sequences. These approaches complement each other in that sequence analysis can provide insightful information on patterns of ordered elements, while time-series and event history analyses can be used to provide insights into the causal mechanism. They can be combined to provide a multidimensional investigation into the pattern of event sequences and their causal mechanisms.

Sequence analysis provides an effective tool for studying women’s early labor market trajectories (AnyadikeDanes and McVicar 2010). Unlike event history, which examines one particular type of transition, such as returning to work after childbirth, sequence analysis allows holistic analysis of women’s diverse career paths. Using inter-sequence distances, the model has reduced more than 1600 permutations of unique career paths into a few meaningful classifications. Combining the results from multinomial logit models, the sequence analysis model illustrates predicted proportions of British women fall into one of the career pathways. The classifications are indicated by “achievers vs. underachievers,” “disabled,” and “middle class vs. working class."

Sequence analysis can also study the romantic relationship congruence among young adults in sub-Saharan Africa (Frye and Trinitapoli 2015). It was too complicated for conventional regression models to analyze different sequences of relationship activities embedded with different relationship ideals because the key dimensions of related activities and ideals that predict congruence or relationship well-being were unclear. Couples who had premarital sex are qualitatively different from those who were married before sexual intercourse. These two marriage-sex sequences reflect divergent relationship ideals and affect relationship wellbeing differently.

3.1.2. Supervised machine learning

Supervised learning differs from unsupervised learning in that it learns from a training dataset similar to a process of a teacher supervising a student's learning. The target variables, the correct answers, are known in supervised learning. Machine learning progressively improves predictions on the training data, compares them to the actual outcome variables, and makes corrections with supervised feedback. Supervised learning uses a series of input variables to predict an output variable to identify the function that can closely reproduce the outcome variable using new input data and testing the dataset.

3.1.2.1. Parameter learning: regression models. **Regression models** belong to methods of supervised parameter learning. Researchers specify the inputs, the outputs, and the functional form while machine learning searches for the best-fitting parameters. The input variables are chosen based on researchers' domain-specific knowledge, which is informed by theories and accumulated knowledge about the subject matter. The functional forms are also pre-specified according to prior knowledge about the mechanisms by which the inputs influence the output. These mathematical functions linking inputs and outputs could take many forms, but most researchers set up a linear model structure that treats the outputs as a weighted sum of all input values. Machine learning searches for the best-fitting weights or parameters, for the input attributes. This group of parameter learning models includes linear regression for numerical outputs, logistic regression for categorical outputs or continuous dichotomous variables, and linear discriminants such as support vector machines for classification. We will not elaborate on this approach of parameter learning as a form of supervised machine learning since it is already well-known to readers.

3.1.2.2. Decision trees and random forests. **Decision tree analysis** classifies data in a treelike manner (Breiman et al., 1984). It starts with the root node with all data points, assigns each possible outcome to a branch, and decides whether each branch leads to another decision node or a terminating leaf node. A decision tree uses a recursive divide-and-conquer procedure by starting with the whole data set and repeatedly applying variable selection to create the purest possible subgroups using the available attributes (Ross 1986). It uses logical statements as classification rules, and the entire tree can be expressed as a set of rules.

Decision tree analysis usually uses chi-squared, Gini index, information gain, and variance reduction when deciding how to split a node into two or more sub-nodes. The chi-squared method uses the sum of squares of standardized differences between the observed and expected frequencies of the target variable. The Gini index approach assigns cases to the class to achieve the highest class purity. The information gain method uses entropy to define the degree of disorganization in a decision tree system. The variance reduction method is for continuous target variables and uses the analysis of variance to decide on the best variable to use for the split.

Necessary procedures and data quality are required to safeguard data overfitting and erroneous decision rules. The approaches of pre-pruning and post-pruning are procedures to prevent overfitting. Pre-pruning, also called constraining tree size, stops growing the tree before it gets too large by using various parameters to define a criterion for a correct final tree size criterion. Post-pruning first grows a tree, then prunes it either by reducederror pruning or rule-based pruning. The quality of training data is also important for training. Decision tree learning, as a form of supervised learning, require a high-quality training set. The training set should contain large variations, providing rich and varied data for the decision tree algorithm to learn and derive decision rules. When the training set is systematically lacking, decision trees trained by such a limited dataset may be erroneous when predicting future outcomes using data with greater variety.

Decision trees have been a popular data mining method because they are easy to use, implement, and compute. The tree induction procedures are elegant, and most data mining packages include some type of tree induction technique. Decision trees also tend to produce better-fitting models than regression models. They identify relationship heterogeneity by building interactions and nonlinear relations to maximize prediction accuracy. A decision tree is a nonparametric method, flexible, and can be applied to various data, not constrained by missing data or data types. Decision trees need to balance their high accuracy in prediction and data overfitting. Decision trees are usually straightforward to understand when the tree structure is simple. However, such models are difficult to interpret when they produce models of many variables, a complex combination of many variables, and a huge data set with many branches.

A life-course study applies the decision tree technique to simultaneously analyze the timing, sequencing, and quantum (number of events) of major life-course events based on data from Austria and Italy (Billari et al., 2006). The study aimed to detect the divergences in life course patterns between Austria and Italy by analyzing a series of life course events, including leaving home, union/marriage, first childbirth, first employment, and education completion. The decision tree model provides a flexible approach that allows events to happen simultaneously, in different order/sequences, and with varying duration/timing.

Random Forest uses an ensemble approach that grows multiple decision trees to classify new objects based on attributes, with each tree giving a different classification (Breiman 2001a). A powerful model is formed by combining a group of weak models by forming a forest from multiple decision trees. The forest chooses the classification with the most votes in the forest based on the number of tree "votes" for each data point's membership in a class.

Random Forest has advantages and disadvantages. It is a versatile machine learning method considered a comprehensive tool for data science problems. It can perform classification tasks, conduct dimensional reductions, treat missing values and outlier values, and perform many other essential data exploration tasks. The Random Forest algorithm can handle large data sets with higher dimensionality. It can identify the most significant variables from thousands of input variables. It is considered one of the more effective dimensionality reduction methods. It is also an effective method for estimating missing data and maintains model accuracy when missing data are highly prevalent in the data set. Random Forest contains methods for balancing errors in data sets where classes are imbalanced. On the other hand, Random Forest pays the price for its versatile abilities by building large, complex, and hard-to- interpret models. Researchers have little control over what the model does other than specifying different parameters.

A recent empirical study used a random survival forest to conduct a dimensional reduction of circumstantial factors that substantially improved infant mortality rates in six Asian countries (Aizawa 2021). The random survival forest approach predicted the counterfactual probability of infant mortality in the 2010s would face if they shared the same household characteristics, family socioeconomic status, and prenatal healthcare use of the 1990s. The study reveals that small families, adequate prenatal care, long gestation period, and improved living standard are the main contributors to improving infant survival rates.

Another study employed a random survival forest to study union dissolution in Germany (Arpino et al., 2022). The random survival forest model provides markedly higher predictive accuracy than a conventional regression model of discrete-time event history logit analysis. Results from random survival forest suggest that couples' life satisfaction and women's percentage of housework are the most important predictors of union dissolution, whereas three factors–couples' income, women's percentage of income, and if the woman is wealthier than the man–have low predictive power.

3.1.2.3. Artificial neural networks and deep learning. **Artificial neural networks** are machine learning approaches that imitate the complex learning systems in animal brains through closely interconnected sets of neurons (Neal 1992; Garson 1998). Dense networks of interconnected simple neurons can perform complex learning tasks such as pattern recognition and classification. An artificial neural model collects information from a data set, combines it in a mathematical function, inputs the result into an activation function to produce an output response, and sends this output to neurons downstream. This process is then repeated in downstream neurons. A neural network consists of input layers, hidden layers, and output layers. The input layer takes the predictor variables. The hidden layers, which can be multiple, process and transform information from the input layer. The output layer produces the outcome.

Neural networks learn in several steps through a back-propagation algorithm. It initially randomly assigns weights for each of the layers. After comparing the initial predicted value produced by the network to the actual outcome value in the training set by calculating squared errors, the algorithm uses a gradient-descent optimization method to locate the weights that will minimize the errors. The algorithm then adjusts these weights iteratively using comparisons of the generated outcome with the actual outcome variables. After repeating this learning process until the best-fitting values are found for all weights, the values are combined to produce an outcome. The backpropagation then works backward from the prediction error and divides it across the various connections in the neural network.

A simple neural network functions similarly to logistic regression but generally outperforms it because neural networks incorporate functional form and variable complexities through multiple hidden layers. The hidden layers take the input variables, assign weights, combine these weighted values, transform them using a specified mathematical function, and produce an output. These mathematical functions in the hidden layers usually take the form of either the hyperbolic tangent function or the sigmoid function. They are both versatile in describing various functions, including nearly linear, curvilinear, and nearly constant behaviors. The sigmoid function is also called squashing because it always produces outputs bounded between 0 and 1. The hyperbolic tangent function always produces outputs bounded between − 1 and 1 and zero-centered, allowing easy optimization.

Neural networks have advantages and disadvantages compared to other approaches. They are robust to noisy data containing uninformative or erroneous instances. Networks contain many artificial neurons, with weights assigned to each connection, so the network can learn to work around uninformative or erroneous examples in the data set. Neural networks outperform other approaches in datasets with noise levels as high as 30% (Goldberger and Ben-Reuven 2017). Some deep neural networks perform very well with extremely noisy data labels at 50% (Han et al., 2018). They also have superior predictive capacity over regression models and even decision trees. However, the models generated by neural networks are relatively difficult to interpret due to a large number of nonlinear behaviors in the hidden nodes. Sensitivity analysis provides one approach to mitigating this shortcoming by determining the relative importance of each input variable on the output target variable. Neural networks tend to overfit with data, are also unstable, and may generate different models with identical training data, variables, settings, and validation data. Lastly, neural networks can be computationally greedy and thus expensive in computing time with big data.

Neural networks have various creative applications in solving classification problems in social science research. Sianes and others (2014) employ neural networks to rate the performance of wealthy countries in helping developing countries in poverty reduction. The traditional approach of ranking rich countries' commitment is arbitrary and can lead to biased estimation. The artificial neural network method uses an adaptive learning process to approximate a classification based on a country's real contribution with both quantitative and qualitative variables. Such a neural network model is flexible with data and has produced the highest classification accuracy than other ordinal ranking methods.

Neural networks can effectively detect sarcasm on Twitter (Zhang et al., 2016). A bi-directional gated recurrent neural network first captures syntactic and semantic content in tweets locally, and then a pooling neural network extracts contextual information from historical tweets automatically. This neural network modeling provides more than 90% accuracy and substantially outperforms a discrete model that searches for simple sarcasm indicators with less than 80% accuracy.

Deep-learning neural networks can achieve powerful deep learning by using many nodes and layers in a deep neural network. Modern deep-learning networks are distinguished by their depth, that is, the number of layers through which data must pass in a multistep learning process. Neural networks deeper than three layers (including input and output) qualify as "deep" learning. In such networks, each layer of nodes trains on features based on the previous layer's output. The deeper the data move forward in the neural net, the more complex the features the nodes can recognize since they aggregate and recombine features from the previous layers. These deep neural nets can discover latent structures within unlabeled, unstructured, and complex data, including pictures, texts, and video and audio recordings. Deep learning is the best approach for processing and clustering raw, messy, and multidimensional data although other supervised learning approaches such as logistic regression, decision tree, or random forest may perform well when the data are less noisy and simple. For example, deep learning can take images, emails, and news articles and cluster them according to their similarities. In the 2017 ImageNet Large Scale Visual Recognition Challenge, in which research teams develop algorithms on tens of thousands of images and compete to achieve higher accuracy on visual recognition tasks, most teams using deep learning approaches achieved greater than 95 percent accuracy on more than 14 million images classified into 17,000 categories (ImageNet 2017).

Deep-learning neural networks can detect protest events on heavily censored Chinese social media (Zhang and Pan 2019). The researchers use two-stage classification deep-learning and cross-validate their results with other protest data sets. Despite the aggressive social media censorship, the deep-learning model detected more than 136,000 offline protest events from 500,000 microblogging posts. These collective action events on social media are mostly related to rural and land disputes, while few are related to ethnic and religious conflicts.

Deep-learning neural network method produces a domain knowledge-aware risk assessment model to detect and calculate suicide risk (Xu et al., 2021). Multiple experienced counselors and social workers independently annotate the conversation and classify each conversation into crisis and non-crisis using data from Hong Kong online counseling service. This domain knowledge is then incorporated to train a deep-learning model to analyze linguistic patterns for suicide prevention. This deep-learning model significantly outperforms a baseline model that only detects keywords in identifying suicidal risks.

4. Conclusions

Knowledge discovery is a dialectic research process that is both deductive and inductive. Data mining is sometimes misconceived as "data-driven" or "fishing expedition" without theory guidance, contributing to a misunderstanding and hindrance in the social science research community i recognizing the contribution of this approach. While the relationship between theory and research differs between deductive and inductive research approaches, they complement each other. The two processes are often combined to form a cycle from theory to data and back to theory. Such a combination is not new in social science research, as Grounded Theory is a general method that uses systematic research to generate a theory that involves both inductive and deductive phases of research.

The rising tsunami of data derived from digitalized, compiled, and stored information from the internet, records, forms, recording devices, and texts not only requires technologies of computer storage and data preprocessing but also necessitates data mining technologies to process, manage, and analyze a large amount of data from "convenience" samples. Big data require new analytical methods beyond the traditional statistical approaches to discover new knowledge from the data mine. Knowledge discovery and data mining emerged from such a necessity. Three academic fields form the foundation. Statistics provides well-defined techniques to systematically identify relationships between variables, including data visualization, descriptive statistics, correlations, frequency tables, multivariate exploratory techniques, and advanced linear and generalized linear models. Artificial Intelligence of machine learning forms a modern foundation by training computers to recognize patterns in data. Database systems make it possible to store, access, and retrieve huge amounts of data, providing support for information processing and mining.

Both classical statistical approaches and data mining play the same role in scientific research: providing information on correlations among variables or associations among entities to shed light on causal discovery. In addition, data mining can efficiently filter complex and multiple correlations among variables to help us identify complexity, interactions, and heterogeneity in the causal relationship. Data mining does not challenge the conventional model-building approach but plays an important complementary role in improving model goodness of fit, revealing valid and significant hidden patterns in data, identifying nonlinear and nonadditional effects, providing insights into developments of data, methods, and theory, and enriching our discovery.

Data mining contributes to the discovery of potential causality in multiple ways (Shu 2020). Data mining uncovers new, sometimes unexpected, ways of conceptualization conducive to theory innovation and knowledge discoveries. Data mining is concerned with offering a thorough or complete account of the event under investigation. This approach does not shy away from a rich analysis of multiple, complicated, and nuanced explanations as they all contribute to the strong predictive power of the resulting model. Data mining models provide a full account of the event or outcome to the fullest permitted by information buried in data to provide maximum accuracy in prediction for the future. Data mining emphasizes complex causal heterogeneity, considers a variety of nonlinear and joint effects, and tends to estimate complex and elaborate models. Data mining aims to maximize a model's predictive power, provides methods capable of analyzing non-numerical data such as text, images, and voice, and uses convenience samples, but sometimes cannot provide information on the causal mechanism.

Machine learning builds models and algorithms by learning and improving from data. Machine learning is best for computing tasks when the explicit model structure is unclear and algorithms with good performance are difficult to attain. Researchers learn from the data to discover hidden patterns and structures and derive models from informing theory building. Machine learning encompasses supervised learning and unsupervised learning. Unsupervised learning models the underlying structure from data with no outcome variables. Unsupervised learning includes two types of problems—classification and association. Cluster analysis, latent class analysis, and topic modeling emphasize the links among data instances or words and aims to discover inherent groupings of data cases or words. Association is interested in the association among attributes and attempts to reveal rules that describe relationships between variables in the data. Supervised learning progressively improves predictions on the training data, compares them to the actual outcome variables, and makes corrections with supervised feedback.

The difference between the classical statistical approach and supervised machine learning results from two paradigms (Breiman 2001b; Donoho 2017). Models of conventional parameter regression are derived from a data modeling culture or generative modeling, while other supervised learning (decision tree, random forest, neural networks, and deep learning) from the algorithmic modeling culture or predictive modeling (Breiman 2001b; Donoho 2017). On the one hand, the data analysis culture (or generative modeling) is based on models with explicit specifications representing the mechanisms "by which nature works" (Breiman 2001b). Researchers in this tradition first propose a stochastic model that could have generated the data and then estimate the parameters of this model from the data. The focus is almost exclusively on explanations of how inputs are linked to the outcomes. On the other hand, the algorithm modeling culture (predictive modeling) does not require models that specify the connection mechanisms between inputs and outputs, and the objective of the estimation is to best approximate the true responses, thus deriving insights into how the outcome is related to its predictors. In this paradigm, the model that could have generated the data is considered unknown. The central objective is to learn from the current data to search for a model with the highest predictive accuracy on new data. Predictive modeling produces complex models with strong performance with new data, but complex models generated by random forest and deep learning neural networks are opaque and provide little insight into the mechanism linking the inputs to the output.

Predictive modeling in machine learning may offer several advantages. They permit the inclusion of a large number of variables and the exploration of nonlinear and interactive links between the inputs and outputs. These features usually result in a high out-of- sample predictive accuracy. In addition to the explanatory power model, predictive power is another significant dimension in inspiring theory building and validation and promoting improvements in methods and data collection (Hofman et al., 2017; Salganik et al., 2020), knowledge discovery and theory innovation (Shu 2020). For example, conventional regress models, such as discrete-time event history models, perform poorly with a low out-of-sample predictive accuracy, while random survival forest models provide considerably superior predictive accuracy in predicting marital union dissolution (Arpino et al., 2022). Another example is that machine learning models of random forest, elastic net, and gradient-boosted trees have improved out-of-sample predictions for three life outcomes in GPA, grit, and layoff compared to baseline models (Rigobon et al., 2019). However, when operating without theoretical guidance, researchers should not mistake data-driven machine learning models as causal models because neither their parameters nor their predictions necessarily have a causal interpretation. It is questionable that such knowledge informs trustable conclusions about causal relationships. Researchers need to employ causal approaches to derive causal structures (Brand et al., 2023).

When the two paradigms of generative and predictive modeling join forces to produce improved models, it increases the possibility of more robust models that combine both explanation and prediction (Hofman et al., 2017). The classical approach of parameter estimation regression models provides interpretable explanations. Predictive modeling in supervised machine learning generates insights into the possible complex ways a large set of predictors may lead to the outcome. This information improves model specification and improves their predictive accuracy. Such new insights into the nonlinearity, interactivity and causal heterogeneity between the inputs and outputs can combine with regression models to produce models with improved predictive accuracy while providing an interpretation of findings from predictive approaches.

References

Aizawa, Toshiaki, 2021. Decomposition of improvements in infant mortality in asian developing countries over three decades. Demography 58 (1), 137–163. https:// doi.org/10.1215/00703370-8931544.

Akaike, Hirotugu, 1977. In: Krishnaiah, R. (Ed.), On the Entropy Maximization Principle" in P, Applications of Statistics. North-Holland, Amsterdam.

Anand, Sarabjot S., Büchner, Alex G., 1998. Decision Support through Data Mining. FT Pitman Publishers.

Anyadike-Danes, Michael, McVicar, Duncan, 2010. My brilliant career: characterizing the early labor market trajectories of British women from generation X. Socio. Methods Res. 38 (3), 482–512.

Arpino, Bruno, Le Moglie, Marco, Mencarini, Letizia, 2022. What tears couples apart: a machine learning analysis of union dissolution in Germany. Demography 59 (1), 161–186. https://doi.org/10.1215/00703370-9648346.

Athey, Suan, Guido, imberns, 2015. A measure of robustness to misspecification. Am. Econ. Rev. 105, 476–480.

Athey, Susan, Guido, Imbens, 2016. Recursive partitioning for heterogeneous causal effects. Proc. Natl. Acad. Sci. 113, 7353–7360.

Bacher, Johann, 2000. A probabilistic clustering model for variables of mixed type. Qual. Quantity 34, 223–235.

Bail, Christoper A., 2008. The configuration of symbolic boundaries against immigrants in Europe. Am. Socio. Rev. 73, 37–59.

Bankes, Steven C., 2002. Agent-based modeling: a revolution. Proc. Natl. Acad. Sci. USA 99 (Suppl. 3), 7199–7200. https://doi.org/10.1073/pnas.072081299.

Billari, Francesco C., Fürnkranz, Johannes, Prskawetz, Alexia, 2006. Timing, sequencing, and quantum of life course events: a machine learning approach. Eur. J.

Popul. 22, 37–65.

Blanchard, Philippe, Bühlmann, Felix, Gauthier, Jacques-Antoine (Eds.), 2014. Advances in Sequence Analysis: Theory, Methods, Applications. Springer, New York.

Bond, Robert M., Fariss, Christopher J., Jason, Jones J., Kramer, Adam D.I., Marlow, Cameron, Settle, Jaime E., Fowler, James H., 2012. A 61-million-person experiment in social influence and political mobilization. Nature 489 (7415), 295–298. https://doi.org/10.1038/nature11421.

Bonikowski, Bart, DiMaggio, Paul, 2016. Varieties of American popular nationalism. Am. Socio. Rev. 81, 949–980.

Brand, Jennie E., Xu, Jiahui, Koch, Bernard, Geraldo, Pablo, 2021. Uncovering sociological effect heterogeneity using tree-based machine learning. Socio. Methodol. 51 (2), 189–223.

Brand, Jennie, Zhou, Xiang, Yu, Xie, 2023. Recent developments in causal inference and machine learning. Annu. Rev. Sociol. Breiman, Leo, 2001a. Statistical modeling: two cultures (with discussion). Stat. Sci. 16, 199–231. Breiman, Leo, 2001b. Random forests. Mach. Learn. 45, 5–32.

Breiman, Leo, Friedman, Jerome H., Olshen, R.A., Stone, Charles J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

Clogg, Clifford, 1995. Latent class models" in. In: Arminger, G., Clogg, C.C., Sobel, M.E. (Eds.), Handbook of Statistical Modeling for the Social and Behavioral Sciences. Springer, New York.

Conte, Rosaria, 2016. Computational social and behavioral science. In: Cecconi, F. (Ed.), New Frontiers in the Study of Social Phenomena. Springer, Cham. https:// doi.org/10.1007/978-3-319-23938-5_1.

Deza, Elena, Deza, Michel-Marie, 2006. Dictionary of Distances. Elsevier Science, Amsterdam www.elsevier.com/books/dictionary-of-distances/deza/978–0- 444–52087–6.

Diamond, Alexis, Sekhon, Jasjeet S., 2013. Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev. Econ. Stat. 95, 932–945.

Donoho, David, 2017. 50 Years of data science. J. Comput. Graph Stat. 26, 745–766. https://doi.org/10.1080/10618600.2017.1384734.

Dumbill, Edd, 2013. A revolution that will transform how we live, work, and think: an interview with the author of big data. Big Data 1 (2), 73–77. https://doi.org/ 10.1089/big.2013.0016.

Epstein, Joshua M., 2006. Remarks on the foundations of agent-based generative social science. Handb. Comput. Econ. 2, 1585–1604. https://www.sciencedirect. com/handbook/handbook-of-computationaleconomics/vol/2/suppl/C.

Fayyad, Usama, Piatesky-Shapiro, Gregory, Smyth, Padhraic, 1996. Knowledge discovery and data mining: towards a unifying framework. KDD-96 Proceedings 82–88.

Frye, Margaret, Trinitapoli, Jenny, 2015. Ideals as anchors for relationship experiences. Am. Socio. Rev. 80, 496–525.

Garip, Filiz, 2012. Discovering Diverse Mechanisms of Migration: the Mexico–US Stream. 1970–2000. Popul. Dev. Rev. 38:393–433. Garip F. 201. On the Move: Changing Mechanisms of Mexico–U.S. Migration. Princeton Univ. Press, Princeton, NJ.

Garip, Filiz, 2017. On the Move: Changing Mechanisms of Mexico-U.S. Migration. Harvard University Press, Cambridge, MA.

Garson, David G., 1998. Neural Networks: an Introductory Guide for Social Scientists. Sage, Thousand Oaks, CA.

Gilbert, Nigel, Matthijs den Besten, Bontovics, Akos, Bart, G., Craenen, W., Divina, Federico, Eiben, A.E., Griffioen, Robert, Hevízi, Gyorgy, L´ orincz, Andras, ˜ Paechter, Ben, Schuster, Stephan, Schut, Martijn C., Tzolov, Christian, Vogt, Paul, Lu, Yang, 2006. Emerging artificial societies through learning. J. Artif. Soc. Soc. Simulat. 9 (2), 9. http://jasss.soc.surrey.ac.uk/9/2/9.html.

Glymour, Clark, Madigan, David, Pregibon, Daryl, Smyth, Padhraic, 1997. Statistical themes and lessons for data mining. Data Min. Knowl. Discov. 1 (1), 11–28. https://doi.org/10.1023/A:1009773905005.

Goldberger, Jacob, Ben-Reuven, Ehud, 2017. Training Deep Neural-Networks Using A Noise Adaptation Layer. 5th International Conference on Learning Representations, Toulon, France. ICLR 2017).

Gondal, Neha, 2022. Multiplexity as a lens to investigate the cultural meanings of interpersonal ties. Soc. Network. 68, 209. https://doi.org/10.1016/j. socnet.2021.07.002.

Gorunescu, Florin, 2011. Data Mining: Concepts, Models and Techniques. Springer.

Hagenaars, Jacques A., McCutcheon, Allan L., 2002. Applied Latent Class Analysis. Cambridge University Press, Cambridge, UK.

Han, Bo Han, Yao, Quanming, Yu, Xingrui, Niu, Gang, Xu, Miao, Hu, Weihua, Tsang, Ivor W., Sugiyama, Masashi, 2018. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels." 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). Montr´eal, Canada.

Hand, Davis J., Mannila, Heikki, Smyth, Padhraic, 2001. Principles of Data Mining. MIT Press, Cambridge, MA.

Hedt, Bethany L., Pagano, Marcello, 2011. Health indicators: eliminating bias from convenience sampling estimators. Stat. Med. 30 (5), 560–568. https://doi.org/ 10.1002/sim.3920.

Heiberger, Raphael H., Munoz-Najar, Galvez S., McFarland, Daniel A., 2021. Facets of Specialization and its

Relation to Career Success: An Analysis of U.S. Sociology, 1980 to 2015." American Sociological Review 86 (6), 1164–1192.

Hofman, Jake M., Sharma, Amit, Watts, Dunchan J., 2017. Prediction and explanation in social systems. Science 355, 486–488.

Holton, Judith A., Walsh, Isabelle, 2017. Classic Grounded Theory: Applications with Qualitative and Quantitative Data. Sage.

Hu, Anning, Wu, Xiaogang, Chen, Yunsong, 2021. Analysis of heterogeneity effects: opportunities and challenges of machine learning. Sociol. Stud.

ImageNet, 2017. ImageNet large scale visual recognition challenge. Retrieved. https://imagenet.org/challenges/LSVRC/2017/. (Accessed 16 July 2022).

Kim, Hwalbin, Jang, S. Mo, Kim, Sei-Hill, Wan, Anan, 2018. Evaluating sampling methods for content analysis of twitter data. Social Media + Soc. 4 (2), 2056305118772836 https://doi.org/10.1177/2056305118772836.

Kramer, Adam D.I., Guillory, Jamie E., Hancock, Jeffrey T., 2014. Experimental evidence of massive-scale emotional contagion through social networks. Proc. Natl. Acad. Sci. USA 111 (24), 8788–8790. https://doi.org/10.1073/pnas.1320040111.

Lazer, David, Pentland, Alex, Adamic, Lada, Aral, Sinan, Barabasi, Albert-L´ aszl´ o, Brewer, Devon, Christakis, Nicholas, Contractor, Noshir, Fowler, James, ´

Gutmann, Myron, Jebara, Tony, King, Gary, Macy, Michael, Roy, Deb, Marshall Van Alstyne, 2009. Computational social science. Science 323 (5915), 721–723. https://doi.org/10.1126/science.1167742.

Lee, Chioun, Coe, Christopher, Ryff, Carol D., 2017. Social disadvantage, severe child abuse, and biological profiles in adulthood. J. Health Soc. Behav. 58 (3), 371–386.

Levenshtein, Vladimir, 1966. Binary codes capable of correcting deletions, insertions, and reversals. Dokl. Phys. 10, 707–710.

Lundberg, Ian, Brand, Jennie E., 2022. The Effect of Income on Educational Outcomes: the Nonlinear and Heterogeneous Effects of a Continuous Treatment. Paper presented at the annual conference of the American Sociological Association, Los Angeles, CA.

Luma-Osmani, Shkurte, Ismaili, Florije, Zenuni, Xhemal, Raufi, Bujar, 2020. A Systematic Literature Review in

Causal Association Rules Mining," 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 48–54. https://doi.org/10.1109/ IEMCON51383.2020.9284908.

MacKay, David J.C., 2003. Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge.

Manyika, J., Chui, M., Brown, B., et al., 2011. Big Data: the Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.

Mason, Winter, Vaughan, Jennifer Wortman, Wallach, Hanna, 2014. Computational social science and social computing. Mach. Learn. 95 (3), 257–260.

Mauro, De, Andrea, Greco, Marco, Grimaldi, Michele, 2016. A formal definition of big data based on its essential features. Libr. Rev. 65 (3), 122–135. https://doi.org/ 10.1108/LR-06-2015-0061.

Michel, Jean-Baptiste, , Yuan Kui Shen, Aviva Presser Aiden, Veres, Adrian, Gray, Matthew K., 2011. The google books team, joseph P. Pickett, dale hoiberg, dan clancy, peter norvig, jon orwant, steven pinker, martin A nowak, erez lieberman aiden. Quantit. Anal. Cult. Using Millions Digitized Books." Sci. 331, 176.

Molina, Mario, Garip, Filiz, 2019. Machine learning for sociology. Annu. Rev. Sociol. 45 (1), 27–45.

Moody, James, 2004. The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999. Am. Socio. Rev. 69, 213–238.

Morgan, Stephen L., Winship, Christopher, 2015. Counterfactuals and Causal Inference: Methods and Principles for Social Research, second ed. Cambridge University Press, Cambridge.

Muthen, Bengt, 2004. Latent variable analysis: growth mixture modeling and related techniques for longitudinal data. In: Kaplan, D. (Ed.), Handbook of Quantitative´ Methodology for the Social Sciences. Sage, Newbury Park, CA.

Neal, Radford M., 1992. Connectionist learning of belief networks. Artif. Intell. 56 (1), 71–113. https://doi.org/10.1016/0004-3702(92)90065-6.

Nelson, Laura K., 2021. Cycles of conflict, a century of continuity: the impact of persistent place-based political logics on women's movement form. Am. J. Sociol. 127 (1).

Nelson, Laura, 2020. Computational grounded theory: a methodological framework. Socio. Methods Res. 49 (1), 3–42.

Pavlova, Alina, Berkers, Pauwke, 2020. Mental health discourse and social media: which mechanisms of cultural power drive discourse on twitter. Soc. Sci. Med. 263, 133250.

Peterson, Robert A., Merunka, Dwight R., 2014. Convenience samples of college students and research reproducibility. J. Bus. Res. 67 (5), 1035–1041. https://doi. org/10.1016/j.jbusres.2013.08.010.

Provost, Foster, Faucett, Tom, 2013. Data Science for Business: what You Need to Know about Data Mining and Data-Analytic Thinking. O'Rilly Media Inc.

Reitermanova, Z., 2010. Data Splitting. WDS'10 Proceedings of Contributed Papers 1, 31–36 www.mff.cuni.cz/veda/konference/wds/proc/pdf10/WDS10_105_i1_ Reitermanova.pdf.

Rigobon, Daniel, Jahani, Eaman, Suhara, Yoshihiko, Al-Ghoneim, Khaled, Alghunaim, Abdulaziz, Pentland, Alex, Almaatouq, Abdullah, 2019. Winning models for GPA, grit, and layoff in the fragile families challenge. Socius 5, 1–10.

Ross, Quinlan J., 1986. Induction of decision trees. Mach. Learn. 1 (1), 81–106.

Salganik, Matthew J., Lundberga, Ian, Kindel, Alexander T., Ahearn, Caitlin E., Ghoneim, Khaled Al, Almaatouq, Abdullah, M Altschul, Drew, Brandb, Jennie E., Carnegie, Nicole Bohme, Compton, Ryan James, Datta,

Debanjan, Davidson, Thomas, Anna, Filippova, Gilroy, Connor, Goode, Brian J., Jahani, Eaman,

Kashyap, Ridhi, Kirchner, Antje, McKay, Stephen, C Morgan, Allison, Pentlande, Alex, Polimis Louis Raes, Kivan, Rigobon, Daniel E., Roberts, Claudia V., Stanescu, Diana M., Adaner Usmani, Yoshihiko Suhara, Wangz, Erik H., Adem, Muna, Alhajri, Abdulla, AlShebli,

Bedoor, Amin, Redwane, Amosy, Ryan B.,

Argyle, Lisa P., Baer-Bositis, Livia, Buchi, Moritz, Chung, Bo-Ryehn, Eggert, William, Faletto, Gregory, Jeremy Freese, Zhilin Fan, Gadgil, Tejomay, Gagn e, Josh,

Gao, Yue, Halpern-Manners, Andrew, Hashimy, Sonia P., Hausen, Sonia, He, Guanhua, Higuera, Kimberly,

Hogan, Bernie, Ilana, M., Horwitz, Lisa M., Naman Jainx, Hummel, Jin, Kun, Jurgens, David, Kaminski, Patrick, Karapetyan, Areg, H Kim, E., Leizman, Ben, Liu, Naijia, M oser, Malte, Mack, Andrew E.,

Mahajan, Mayank, Mandell, Noah, Marahrens, Helge, Mercado-Garcia, Diana, Mocz, Viola, Mueller-Gastell, Katariina, Musse, Ahmed, Niu, Qiankun, Nowak Hamidreza Omidvar, William, Or, Andrew, Ouyang, Karen, Pinto, Katy M., Porter, Ethan, Porter, Kristin E., Qian, Crystal, Rauf, Tamkinat, Sargsyan, Anahit,

Schaffnery, Thomas, Schnabel, Landon, Schonfeldz, Bryan, Sender, Ben, D Tang, Jonathan, Tsurkov, Emma, Loon, Austin van, Varol, Onur, Wang, Xiafei,

Wang, Zhi, Wang, Julia, Wang, Flora, Weissmany, Samantha, Whitaker, Kirstie, Wolters, Maria K., Woon, Wei Lee, Wu, James, Wu, Catherine, Yang, Kengran,

Yin, Jingwen, Zhao, Bingyu, Zhu, Chenyun, Brooks-Gunn, Jeanne, Engelhardty, Barbara E., Hardt, Moritz, Knox, Dean, Levy, Karen, Narayanany, Arvind, M Stewarta, Brandon, J. Watts, Duncan, McLanahan, Sara, 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. Proc. Natl. Acad. Sci. USA 117 (15), 8398–8403.

Samuel, Arthur, 1959. Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 3 (3), 210– 229.

Scarborough, William J., Sin, Ray, 2020. Gendered places: the dimensions of local gender norms across the United States. Gend. Soc. 34 (5), 705–735.

Seife, Charles, 2015. Big data: the revolution is digitized. Nature 518, 480–481.

Scarborough, William J., Joanna, R Pepin, Lambouths, Danny L., Kwon, Ronald, Monasterio, Ronaldo, 2021. The intersection of racial and gender attitudes, 1977 through 2018. Am. Socio. Rev. 86 (5), 823–855.

Scarborough, William J., Sin, Ray, Risman, Barbara, 2019. Attitudes and the stalled gender revolution: egalitarianism, traditionalism, and ambivalence from 1977 through 2016. Gend. Soc. 33 (2), 173–200.

Shu, Xiaoling, 2003. In: Lewis-Beck, M., Bryman, A., Liao, T.F. (Eds.), Artificial Intelligence" in the Sage Encyclopedia of Social Science Research Methods. Sage Publications.

Shu, Xiaoling, 2020. Knowledge Discovery in the Social Sciences: A Data Mining Approach. University of California Press, Oakland, CA.

Sianes, Antonio, Dorado-moreno, Manuel, Hervas-martínez, ´ C´esar, 2014. Rating the rich: an ordinal classification to determine which rich countries are helping poorer ones the most. Soc. Indicat. Res. 116 (1), 47–65.

Soehl, Thomas, Karim, Sakeef M., 2021. How legacies of geopolitical trauma shape popular nationalism today. Am. Socio. Rev. 86 (3), 406–429.

Van de Rijt, Arnout, Shor, Eran, Ward, Charles, Skiena, Steven, 2013. Only 15 minutes? The social stratification of fame in printed media. Am. Socio. Rev. 78 (2), 266–289.

Watts, Duncan J., 2013. Computational social science: exciting progress and future directions. T*he Bridge on Frontiers of Engineering* 43 (4), 5–10.

Wager, Stefan, Athey, Suan, 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. 113, 1228–1242.

Westreich, Daniel, Lessler, Justin, Funk, Michele Jonsson, 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J. Clin. Epidemiol. 63, 826–833.

Winton, Bradley G., Sabol, Misty A., 2021. A multi-group Analysis of convenience samples: free, cheap, friendly, and fancy sources. Int. J. Soc. Res. Methodol. 1–16. https://doi.org/10.1080/13645579.2021.1961187.

Witten, Ian H., Frank, Eibe, Hall, Mark A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, third ed. Elsevier.

Wyss, Richard, Alan, R., Ellis, M Alan Brookhart, J Girman, Cynthia, Funk, Michele Jonsson, LoCasale, Robert, Til Stürmer, 2014. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. Am. J. Epidemiol. 180, 645–655.

Xu, Zhongzhi, Xu, Yucan, Cheung, Florence, Cheng, Mabel, Lung, Daniel, Law, Yik W., Chiang, Byron, Zhang, Qingpeng, Paul, S., Yip, F., 2021. Detecting suicide risk using knowledge-aware natural language processing and counseling service data. Soc. Sci. Med. 283, 114176.

Zhang, Han, Pan, Jennifer, 2019. CASM: a deep learning approach for identifying collective action events with text and image data from social media. Socio. Methodol. 49 (1), 1–57.

Zhang, Han, Peng, Yilang, 2022. Image clustering: an unsupervised approach to categorize visual data in social science research. Socio. Methods Res. https://doi.org/ 10.1177/00491241221082603.

Zhang, Meishan, Zhang, Yue, Fu, Guohong, 2016. Tweet sarcasm detection using deep neural network.” Paper presented at the COLING 2016 - 26th International Conference on Computational Linguistics.

Proceedings of COLING 2016: Technical Papers 2449–2460.