Data Mining

GREGG R. MURRAY and ANTHONY SCIME

Abstract

This essay introduces data mining as an analytical technique for novice to professional social and behavioral scientists. It presents data mining, which is also known as, among other things, *data analytics* and *predictive analytics*, as an effective tool for researchers who are interested in the analysis of "big data" as well as small, unique data sets. It addresses foundational elements of data mining such as how to avoid "data dredging" and the importance of theory as embodied in researcher domain expertise. It also briefly defines and describes classification analysis, association rules, and clustering, which are the major methodologies among a large number of methodologies that constitute data mining. This essay identifies analytical problems and data for which the techniques are best suited. It goes on to highlight a number of cutting-edge studies that relied on data mining techniques in disciplines such as criminal justice, education, health sciences, linguistics, political science, and sociology. This essay concludes with a review of key considerations for future research to include discussions of the burgeoning of new analytical techniques and new data sets and sources, the importance and protection of data-source privacy, and the ethical obligation researchers have to exploit to their fullest extent the costly data on social and behavioral issues collected by scientists and society.

INTRODUCTION

Data mining is the exploration and analysis of typically large quantities of data using mathematical algorithms and computer learning techniques to detect hypothesized and previously unknown relationships and patterns. Often, but not always, the objective of data mining is to predict behaviors and outcomes, identify correlations and patterns, and group similar cases. *Data mining* (also known as, among other things, *data analytics*, *Knowledge Discovery from Data*, *predictive analytics*, and *data/pattern analysis*) is a term used to describe a number of analytical techniques that can be employed to identify meaningful relationships in data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). As a discipline, data mining has its origins in statistics, artificial intelligence, and machine learning.

Emerging Trends in the Social and Behavioral Sciences. Edited by Robert Scott and Stephen Kosslyn.
 ISBN 978-1-118-90077-2.

Data mining is an inductive process that allows researchers to evaluate data in-depth to discern complex patterns, generate rules, uncover previously unknown relationships, spark new ideas, refocus questions, make predictions, and develop or verify theories. Data mining takes advantage of large quantities of data and allows those data to suggest which relationships are and are not worthy of attention, including theoretically well-known relationships that may not persist when forced to compete with a variety of other variables (e.g., Murray, Riley, & Scime, 2009). It takes research out of the convenient yet narrow beam of the drunkard's streetlight that is often trained by theory developed before the pertinent data were collected or methods created and places it under the bright glow of a multivariate method that acknowledges that previously ignored variables and relationships often have predictive and, therefore, previously ignored theoretical value (Burton, 2010; Cohan, 2012).

While data mining most often creates decision trees, clusters, and association rules to be used as analytical tools, other widely used data mining techniques include regression analysis, time series analysis, and hazard functions and survival analysis. All of these techniques may be implemented in various ways using a number of different algorithmic methodologies. Very broadly, data mining is well suited for:

- variable interaction identification, profiling, and segmentation: identifying heterogeneous effects or the common characteristics of groups of individuals who behave in similar ways and then analyzing differences between the groups;
- association and link analysis: understanding which items, characteristics, behaviors, or risks typically occur together;
- outlier/extreme case analysis: analyzing rare behaviors or events or detecting individuals that are likely to behave contrary to expectations; and
- trend analysis: describing the actions of or differences between typical individuals over time.

Data mining techniques make possible the analysis of "big data," the popular current hallmark of which are social media and marketing data. More formally, *big data* refers to "various forms of large information sets that require special computational platforms in order to be analyzed" (Research Trends, 2012, p. 1). Although data mining is associated with "big data" analysis of hundreds of thousands or millions of cases and hundreds or thousands of variables, it is also often used on low-volume data sets as small as tens of cases and variables. While the tools for big data analysis are still in the developmental stage, the algorithms for data mining more typical

low-volume data sets are better established. However, these algorithms are not yet formalized in easy-to-use packages, unlike many statistical techniques.

Data mining has been applied to and produced noteworthy results in a number of disciplines including sociology, criminal justice, health science, linguistics, and political science. For example:

- Butterfield (2010) shows data mining techniques are used to determine how the English language is used and how it has evolved.
- Jakulin, Buntine, La Pira, and Brasher (2009) used data mining to identify clusters of cohesive voters in the US Senate, which indicates more complex voting blocs than the simple two-party divide.
- Green and Kern (2012) employed data mining to identify heterogeneous treatment effects systematically in survey experiments.
- Faisal, Olatunji, and Ghouti (2009) developed a data mining approach to classify and detect gasoline types used in arson cases that proved to be more consistent and accurate than previous models based on other methods.
- Hasan and Rahman (2009) used classification mining to forecast emerging epidemics and provide precautionary measures to be taken to reduce the level of damage caused by an epidemic.

FOUNDATIONAL RESEARCH

Induction, "Data Dredging," and Multiple Comparisons

Few if any graduate students in the social and behavioral sciences avoid lectures on the dangers of "data dredging." This sin of data analysis is marked by trolling around one's data to the point that relationships emerge because of statistical chance not the empirical manifestation of theory. The end results of such efforts are often misleading models that explain nothing more than the data to which they are overfit. Good-faith inductive analyses, too, often face related criticism driven by the concern that the white swans we always see do not preclude the possibility of the black swan that is rare but possible (Popper, 2002 [1959]).

Social and behavioral sciences guard against these concerns by employing the "scientific method," a classic form of which is hypothetico deduction. Hypothetico deduction requires scholars *prior to analysis* to describe their theory, present pertinent expectations in the form of hypotheses, identify appropriate tests for their hypotheses, and establish criteria for accepting or rejecting their hypotheses. As noted by Burton (2010), this type of procedure reduces the probability of misleading overfitting of data and preserves the ability to falsify hypotheses.

Principled scholars who use data mining techniques acknowledge these problems and employ a variety of techniques, in particular train-test methods (e.g., Burton, 2010; Green & Kern, 2012), to assuage such concerns. Using the basic train-test approach, a researcher randomly divides the data set into two independent sub-sets: training and testing. The training data set is used to build the model inductively, while the testing data set is used to verify that the model is not overfit (or underfit) to the data. This train-test approach is designed to confirm that the model is robust and that it can be reliably applied to new data in the domain of the study. The difference in predictive success of the model between the two data sets indicates the extent to which the training-built model fits the data. As noted by Green and Kern (2012), the use of split samples (Hastie, Tibshirani, & Friedman, 2009; Izenman, 2008) and exploratory and confirmatory analysis (e.g., Tukey, 1977) to avoid overfitting is not new, and this approach allows one to develop hypotheses inductively. The need to split the data set for train-test procedures means that larger data sets are better suited for data mining techniques (Green & Kern, 2012), although low-volume data sets are also viable.

The Vital Role of the Social Scientist (i.e., the "Domain Expert")

It is easy to infer incorrectly from popular but uninformed descriptions of data mining that it is an atheoretical or black-box endeavor. However, effective data mining is not atheoretical. Knowledgeable researchers understand technical methodologies are not sufficient, because data are often complex and the relationships among them are often theoretically complicated. Therefore, a domain expert is vital to effective analysis for a number of reasons (DuMouchel & Pregibon, 2001).

First, the social and behavioral science domains are complex, because they reflect human behavior. Their data are specialized and often require domain expertise in order to develop meaningful research questions, identify and select the relevant data, guide the data mining process, and interpret the results. Once the research question is established and the data source is identified and selected, the domain expert must drive the data cleaning and data mining processes. Minimally, the domain and its associated data may contain redundant or overlapping data, or they may exclude important variables. The knowledge and familiarity of a domain expert is required to unravel the data to achieve valid and reliable results (Anand, Bell, & Hughes, 1995; Hofmann & Tierney, 2003; Scime & Murray, 2007).

Second, expert knowledge increases model usefulness and fit and informs the selection of the appropriate model representation of the data and domain (Osei-Bryson, 2004). Recognizing this, Scime and Murray (2007), Hofmann

and Tierney (2003), and Ankerst, Ester, and Kriegel (2000) propose iterative processes that combine domain expertise with the data mining process.

Third, data mining techniques produce models of the data and domain often in the form of IF-THEN rules. There may be a large number of rules, from which the most useful or interesting ones must be identified. The selection of the interesting rules from rule sets is important to understanding and discovering new knowledge in the domain. Rules can be selected and reduced mechanically, but the domain expert often plays the primary role (Geng & Hamilton, 2006). For instance, researchers can direct the mining process step-by-step as in the Perception-based Classification System (Ankerst *et al*., 2000), and they can also decide to apply a combination of data mining methodologies (e.g., Jaroszewicz & Simovici, 2004).

An Overview of Data Mining Techniques

Data mining techniques may be either "unsupervised" or "supervised." Unsupervised analysis is designed to detect meaningful relationships between variables and their values in the data, whereas supervised analysis is designed to detect the relationship of one variable (i.e., the dependent variable) with the remaining variables in the data. For example, a researcher wishing to predict voting behavior may use supervised data mining to segment citizens into factions based on the presidential party for which they will vote (Murray & Scime, 2010), and use unsupervised data mining to find citizen's characteristics associated with the presidential party for which a citizen will vote (Rajasethupathy, Scime, Rajasethupathy, & Murray, 2009). It is not uncommon for multiple techniques to be used in concert to analyze a data set.

While there are a number of data mining techniques, the most common are: classification, association, and clustering. Classification analysis, a supervised technique, constructs a classification tree model, finding a path to a predetermined dependent or target variable for each case. In addition to classifying discrete or categorical variables, continuous variables may also be used. A classification decision tree contains branches that can be converted into IF-THEN rules unique to the dataset, but applicable to future similar datasets. Classification evolved from two sources. In statistics, CHAID (chi-squared automatic interaction detection) (Kass, 1980) is a well-known classification method that uses the chi-squared statistic to determine model structure. On the other hand, the best known machine learning classification algorithm is C4.5 (Quinlan, 1993), which uses information gain to define the model.

Association mining, an unsupervised technique, finds patterns from discrete variables in data. Association analysis directly produces IF-THEN rules that relate a set of variables and their values with another set of variables/values. Apriori (Agrawal, Imieliński, & Swami, 1993) is the

predominant association mining algorithm. Association produces many rules, and domain expertise and special techniques are needed to reduce the rule set to those that are interesting.

The final major technique is clustering. Clustering is used to find groupings of data that show where cases occur in the multidimensional problem space, where each variable is either discrete or continuous and represented as a dimension. It is used to segment a heterogeneous population into homogeneous subgroups (i.e., clusters) using measures of similarity. A popular clustering algorithm is *k*-means (MacQueen, 1967). Cluster analysis is often used in profiling as well as being employed before other data mining techniques to help identify salient variables for the later analyses.

It is appropriate to note interrelationships in data can be evaluated using other methods as well, such as theoretically built regression models. However, data mining holds advantages over nondata mining approaches other than its greater suitability for larger data sets. For example, Andoh-Baidoo and Osei-Bryson (2007) showed classification trees are more insightful than regression in identifying the interactions of predictor variables. More specifically, data mining holds at least four significant advantages over regression analysis. First, regression requires all important relationships to be identified before testing, which limits the discovery of previously unknown relationships (Burton, 2010; Cohan, 2012) and increases the effects of analytical bias through variable selection (Popper, 2002 [1959]). On the other hand, data mining identifies both previously known and unknown relationships and reduces variable-selection bias by including more variables in the analysis. Second, regression requires missing values to be estimated or data to be eliminated, whereas data mining algorithms maintain the integrity of the data by accounting for missing data. Third, data mining provides direct knowledge of how changes to the variables can change the result. Finally, data mining produces output that is easily converted into specific, actionable rules that, unlike regression, identify groups of like individuals.

CUTTING-EDGE RESEARCH

Data mining techniques have been used to address a number of interesting social and behavioral issues using both large and small data sets. In each case the researchers identified a pertinent issue and used data mining techniques to understand the problem and, in some cases, provide possible solutions. The resulting models provide knowledge about the structure and interrelationships among the data and/or predict the results of a future event, both of which can lead to a better understanding of the data, the problem, and the domain.

Political Science: Voting and Terrorism

The American National Election Studies (ANES) is a large data set based on an ongoing series of public opinion surveys intended to collect research-quality data on the theoretical and empirical bases of American national election outcomes using voting behavior, public attitudes, and measures of political participation. The ANES includes items on voter registration and choice, social and political values, partisanship and ideology, opinions about public policy, social background, mass media consumption, and egalitarianism. The ANES data set contains more than 47,000 cases and more than 900 variables. Addressing a core issue in political science, Murray and Scime (2010) created a data mined model that segmented the electorate using only 13 variables to predict their vote, including abstention. The model's 66% accuracy substantially outperforms previous statistical models, which show only 51% accuracy.

An important issue in political science and in conducting reliable public opinion surveys in particular is the likelihood of an individual voting in an election. Using the ANES and data mining, researchers identified two survey items that together can be used to categorize individuals as likely voters or nonvoters. The two items correctly classify 78% of respondents over a three-decade period, a result that met or surpassed the accuracy rates of previous-non-data mining "likely voter" models. In addition, the results indicate that demographic attributes are less salient than previously found by political scientists (Murray *et al*., 2009).

Terrorism analysts connect a number of social, political, and economic conditions at the national level to the likelihood that a nation will face a terrorist event. These conditions generally fall into at least one of four broad categories: economic development, level of democracy, modernization, and social fractionalization. With the intent of identifying long-term predictors of terrorism, researchers constructed a smaller, unique data set comprised of terrorism events and measures of social, political, and economic contexts in 185 countries occurring between the years of 1970 and 2004. The data set contained 126 variables and 5431 cases or country-years. Classification and association mining results showed that more democratic states are more likely to suffer a terrorist attack. In addition, economic development and modernization were not supported as significant factors leading to terrorist attacks, a finding that challenges some previously proffered theory (Scime, Murray, & Hunter, 2010).

Sociology: Emotions and Residential Patterns

Individuals readily express their emotions and sentiments in social media. Thelwall, Wilkinson, and Uppal (2010) mined public comments on a major

social media site to assess the strength of positive and negative emotions and their relationship to gender. They found that women express positive emotions in about two thirds of their comments. This result corresponds with previous research showing females tend to use positive emotions more than males. However, they also found that negative emotions are not associated with gender and are clearly less frequent than positive emotions.

Spielman and Thill (2008) evaluated the association between social similarity and geographic proximity using data mining and a geographic information system. They identified several social patterns by looking at New York City's complex demographic structure as presented in census data. The results challenge Tobler's First Law of Geography (Tobler, 1970), which states that geographically close individuals or objects are more similar than distant individuals or objects.

Health Science: Smoking and Addiction

Health Science is a complex discipline studying the determinants of personal and societal health. Researchers used data mining to examine factors that might lead to a better understanding of what advice healthcare workers give to elderly residents on tobacco cessation. The analyses indicated that healthcare workers' license level, beliefs regarding effectiveness of giving advice, and administrative authority to give advice were predictive of not giving advice to encourage elderly residents to stop smoking (Watt, Lassiter, & Scheidt, 2009).

The characteristics of the client-counselor relationship can affect the recovery of clients from addiction. Data mining client records found that the fastest recovery is by a white, married male with a single addiction meeting with a white female counselor. The slowest recovery is by a single, nonwhite female with multiple addictions meeting with a black female counselor. The results also highlighted a number of characteristics of highly effective counselors (Burn-Thornton & Burman, 2009).

Education: Student Behavior and Retention

It is vital for educators to identify and address systemic issues in schools, such as student discipline. Researchers constructed a dataset from one school's disciplinary data containing records of 35,272 disciplinary problems occurring in one school year. The data include student demographics, the discipline problems, and the day of occurrence. Using classification analysis, the researchers identified the characteristics of students who likely caused disciplinary problems, what problems they were likely to cause, and when the problems were likely to occur, which together allowed school officials to

anticipate and respond more effectively to these problems (Scime & Reiner, 2012).

Student retention in higher education is an ongoing problem. Students leave school for a variety of reasons including absenteeism, personal reasons, or academic challenges. With sufficient warning school administrators can reduce the loss of students by appropriate interventions. Daimi and Miller (2009) demonstrated how classification mining can inform colleges and universities on student retention. They found that data mining can help institutions understand retention risks, provide a list of students most likely to leave, and strengthen student retention policies.

CRIMINAL JUSTICE: EQUAL TREATMENT UNDER THE LAW AND VICTIMIZATION

A core tenet of democracy is equal treatment for all citizens under the law. Researchers used a contextual analysis of state and federal judicial decisions from 1998 to 2010 to build a small, 105-case, five-variable data set of adjudication differences between celebrities and non celebrities (Carroll & Scime, 2012). This study considered the effect of a number of influences on judicial decision-making including the impact of celebrity status of the defendant. The results suggest that celebrities do not receive special treatment but in fact are convicted at a higher rate than noncelebrity defendants.

Another vital issue for civil society is the protection of citizens from predators. Researchers developed a Bayesian belief network classifier that predicts victimization using data from the National Crime Victimization Survey, the United States' annual collection of criminal victimization information. The results predicted victimization with 99% accuracy (Riesen & Serpen, 2009). This research showed how a Bayesian belief network and data mining in general may be use to predict victimization in the criminal justice domain and possibly lead to more effective defenses and interventions.

KEY ISSUES FOR FUTURE RESEARCH

More Data and More Tools

As has been demonstrated here, data mining is a sound, principled, and viable technique for conducting social and behavioral analyses. Data mining techniques robustly exploit data collections and, importantly, shine a broader, brighter light on social and behavioral data. Social scientists are beginning to recognize the untapped knowledge that data mining makes accessible, and they are starting to realize data mining is a valuable tool for building knowledge.

Information scientists continue to develop and improve data mining algorithms for social science and behavioral data analysis. New algorithms designed to find new social knowledge, find and assess social theories, and predict human behavior will continue to emerge. Together social and information scientists can fruitfully assess data to improve their understanding of their domain of interest.

For instance, social media and other Web 2.0 application use have created new, vast, and fast growing data collections (Wilson, Gosling & Graham, 2012). New technologies in data mining are being developed to analyze these social networks and the vast data they provide. Technologies being developed or improved include network analysis, streaming analysis, temporal database analysis, text mining, and content analysis, among others. Currently these techniques primarily focus on analyzing the Web 2.0 platform's efficiency and effectiveness, but there is movement to analyze the content of the media, as in Thelwall *et al*. (2010) work on emotional expression and gender.

Privacy and Data Mining

The collection and use of personal information in databases brings up the issue of privacy. Solove (2007) identifies two sources of privacy violations that may be relevant to a discussion of data mining. First, data may be collected that is considered private. However, one must recall that data mining does not collect data, but it processes data that has already been collected. Second, personal information may be used in a manner for which it was not originally intended. This could be where data mining is a concern.

The concern is that individual people will be identified and harmed by an invasion of their privacy. This can be done in data mining just as it can be done in any analytical effort. However, there are procedural protections outside the research process. Institutional review boards (IRB) are tasked with safeguarding individual rights to ethical treatment in research, including rights to privacy and confidentiality. This includes individuals who may be represented as cases in data. It is incumbent on the expert or researcher to remove from the data identifying variables or values, and to have their procedures approved by the appropriate IRB. Of course, because the data mining process is designed to discover new, unknown, and interesting patterns, it is possible that an unexpected combination of variables/values could lead to the identification of individuals. Therefore, the expert/researcher must also review the data mining results for privacy violations and then take the appropriate, ethical actions.

An Ethical Imperative

At colleges and universities in the United States more than $4.4 billion was spent on social science research and development in fiscal year 2009 (National Science Board, 2012). Governments and nongovernment organizations spend billions of dollars more collecting related data. This investment represents a financial expenditure as well as the expenditure of countless hours of researcher, participant, and administrator time and effort. The results of this immense investment are often embodied in extensive data sets such as the American National Election Studies (48,000-plus cases and 900-plus variables), the Baccalaureate and Beyond Longitudinal Study (17,000-plus cases and 500-plus variables), and the General Social Survey (52,000-plus cases and 5300-plus variables). Minimally, efficiency demands reasonable output from this substantial input in data collection (Scime & Murray,). This is not a new concept. Rosenthal (1994, p. 130) contends

> Data are expensive in terms of time, effort, money, and other resources … If the research was worth doing, the data are worth a thorough analysis, being held up to the light in many different ways so that our research participants, our funding agencies, our science, and society will all get their time and their money's worth.

The challenge to investigators is greater than simple efficiency, though. As stated in the National Science Foundation's mission statement, the lofty goal is "to promote the progress of science; to advance the national health, prosperity, and welfare; [and] to secure the national defense" (National Science Foundation, 2009). In this endeavor, the large quantities of data collected may contain the keys to resolving important issues. In some sense, then, there is an ethical imperative to exploit the data to their fullest extent. Data mining is one way to explore the sometimes inaccessible mass of data and expose the knowledge it contains.

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings: 1993 ACM SIGMOD International Conference on Management of Data*. Washington, DC, pp. 207–216.

Anand, S. S., Bell, D. A., & Hughes, J. G. (1995). The role of domain knowledge in data mining. *Proceedings of the Fourth International Conference on Information and Knowledge Management*, Baltimore, MD, pp. 37–43.

Andoh-Baidoo, F. K., & Osei-Bryson, K. (2007). Exploring the characteristics of internet security breaches that impact the market value of breached firms. *Expert Systems with Applications*, *32*, 703–725.

Ankerst, M., Ester, M., & Kriegel, H. (2000). Towards an effective cooperation of the user and the computer for classification. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, MA, pp. 179–188.

Burn-Thornton, K. & Burman, T. (2009). Factors which influence the recovery of alcohol addicts: A second follow up study. *Proceedings of the 2009 International Conference on Data Mining*, Las Vegas, NV, pp. 165–170.

Burton, M. J. (2010). *A defense of machine learning procedures in quantitative political analysis: Modeling and validation*. Poster presented at the 2010 meetings of the Midwest Political Science Conference, April 22–25, Chicago, IL.

Butterfield, J. (2010). *Damp squid: The English language laid bare*. Oxford, England: Oxford University Press.

Carroll, B., & Scime, A. (2012). Mining for the truth: Analyses of celebrity adjudication decisions. *National Social Science Journal*, *39*, 1–7.

Cohan, F. M. (2012). Science needs more *Moneyball*. *American Scientist*, *100*, 182–185.

Daimi, K. & Miller, R. (2009). Analyzing student retention with data mining. *Proceedings of the 2009 International Conference on Data Mining,* Las Vegas, NV, pp. 55–60.

DuMouchel, W. & Pregibon, D. (2001). Empirical Bayes screening for multi-item associations. *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, pp. 67–76.

Faisal, K., Olatunji, S. O., & Ghouti, L. (2009). Classification of premium and regular gasoline using support vector machines as a novel approach for arson and fuel spill investigation. *Proceedings of the 2009 International Conference on Artificial Intelligence,* Las Vegas, NV, pp. 345–350.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *99*, 27–34.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys*, *38*, article 9.

Green, D. P., & Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly*, *76*, 491–511.

Hasan, F. R., & Rahman, R. M. (2009). Mining ICDDR, B Hospital Surveillance Data using decision tree classification algorithm. *Proceedings of the 2009 International Conference on Information & Knowledge Engineering*, Las Vegas, NV, pp. 290–296.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (*2nd edition*). New York, NY: Springer.

Hofmann, M. & Tierney, B. (2003). The involvement of human resources in large scale data mining projects. *Proceedings of the 1st International Symposium on Information and Communication Technologies*, Dublin, Ireland, pp. 103–109.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. New York, NY: Springer.

Jakulin, A., Buntine, W., La Pira, T. M., & Brasher, H. (2009). Analyzing the U.S. Senate in 2003: Similarities, clusters, and blocs. *Political Analysis*, *17*, 291–310.

Jaroszewicz, S. & Simovici, D. A. (2004). Interestingness of frequent itemsets using Bayesian networks as background knowledge. *Proceedings of the Tenth ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, pp. 178–186.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*, 119–127.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Murray, G. R., Riley, C., & Scime, A. (2009). Pre-election polling: Identifying likely voters using Iterative Expert Data Mining. *Public Opinion Quarterly*, *73*, 159–171.

Murray, G. R., & Scime, A. (2010). Microtargeting and electorate segmentation: Data mining the American National Election Studies. *Journal of Political Marketing*, *9*, 143–166.

National Science Board. (2012). *Science and Engineering Indicators 2012*. Retrieved from http://www.nsf.gov/statistics/seind12/c5/c5s1.htm.

National Science Foundation. (2009). *National Science Foundation FY 2005 Performance Highlights*. Retrieved from http://www.nsf.gov/pubs/2010/nsf10002/nsf10002.pdf.

Osei-Bryson, K. (2004). Evaluation of decision trees: A multicriteria approach. *Computers and Operations Research*, *31*, 1933–1945.

Popper, K. (2002[1959]). *The logic of scientific discovery*. London, England: Routledge.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco, CA: Morgan Kaufmann.

Rajasethupathy, K., Scime, A., Rajasethupathy, K. S., & Murray, G. R. (2009). Finding "persistent rules": Combining association and classification results. *Expert Systems With Applications*, *36*, 6019–6024.

Research Trends (2012). Special issue on big data. *Research Trends*, *30*(September).

Riesen, M. & Serpen, G. (2009). A Bayesian belief network classifier for predicting victimization in national crime victimization survey. *Proceedings of the 2009 International Conference on Artificial Intelligence,* Las Vegas, NV, pp. 648–652.

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science*, *5*, 127–134.

Scime, A., & Murray, G. R. (forthcoming). Social science data analysis: The ethical imperative. In H. Rahman & I. Ramos (Eds.), *Ethical data mining applications for socio-economic development*. Hershey, PA: IGI Global.

Scime, A., & Murray, G. R. (2007). Vote prediction by iterative domain knowledge and attribute elimination. *International Journal of Business Intelligence and Data Mining*, *2*, 160–176.

Scime, A., Murray, G. R., & Hunter, L. Y. (2010). Testing terrorism theory with data mining. *Data Analysis Techniques and Strategies*, *2*, 122–139.

Scime, A. & Reiner, S. (2012). Finding interesting classification rules: An application from education. *Proceedings of the 2012 International Conference on Data Mining,* Las Vegas, NV, pp. 37–43.

Solove, D. J. (2007). "I've got nothing to hide" and other misunderstandings of privacy. *San Diego Law Review*, *44*, 745–772.

Spielman, S. E., & Thill, J. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, *32*, 110–122.

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, *61*, 190–199.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*, 234–240.

Tukey, J. W. (1977). *Exploratory data analysis*. New York, NY: Pearson.

Watt, C. A., Lassiter, J. W., & Scheidt, D. M. (2009). The use of logistic regression analyses and data classification mining to examine variables predictive of long-term healthcare staff giving cessation advice. *Proceedings of the 2009 International Conference on Data Mining*, Las Vegas, NV, pp. 561–565.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, *7*, 203–220.

FURTHER READING

Han, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

Issenberg, S. (2012). *The victory lab: The secret science of winning campaigns*. New York, NY: Crown Publishers.

Kantardzic, M. (2011). *Data mining: Concepts, models, methods, and algorithms* (2nd ed.). Hoboken, NJ: Wiley.

Kumar, A. V. S. (2011). *Knowledge discovery practices and emerging applications of data mining: Trends and new domains*. Hershey, PA: Information Science Reference.

Roiger, R., & Geatz, M. (2003). *Data mining: A tutorial based primer*. New York, NY: Addison-Wesley.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston, MA: Pearson Education.

Whitten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

GREGG R. MURRAY SHORT BIOGRAPHY

Gregg R. Murray is an Associate Professor of Political Science at Texas Tech University. He received his PhD from the University of Houston in 2003. Murray's research focuses on political behavior as well as the development of techniques to analyze large-scale data sets. His research related to data mining has appeared in *Public Opinion Quarterly*, the *Journal of Political Marketing*, *Expert Systems with Applications*, the *International Journal of Data Analysis Techniques and Strategies*, the *International Journal of Business Intelligence and Data Mining*, and a number of book chapters. For more information on Murray and his research, see www.GreggRMurray.com.

ANTHONY SCIME SHORT BIOGRAPHY

Anthony Scime holds a doctorate from George Mason University in Information Systems and Education. Currently he is an Associate Professor of Computer Science at The College at Brockport, State University of New York. He has data mined with social scientists in political science, criminal justice, social work, and education. Their work in data mining and the social sciences has been published in *Expert Systems with Applications*, the *International Journal of Business Intelligence and Data Mining, Public Opinion Quarterly, Elections and Exit Polling in the 21st Century*, the *Journal of Political Marketing*, and the *National Social Science Journal*. Idea Group Publishing published his book *Web Mining: Applications and Techniques*. His current research interests also include data mining measures of interestingness and computing education.

RELATED ESSAYS

To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure *(Methods)*, Robert Boruch and Alan Ruby
Ambulatory Assessment: Methods for Studying Everyday Life *(Methods)*, Tamlin S. Conner and Matthias R. Mehl
Models of Nonlinear Growth *(Methods)*, Patrick Coulombe and James P. Selig
Quantile Regression Methods *(Methods)*, Bernd Fitzenberger and Ralf A. Wilke
Ethnography in the Digital Age *(Methods)*, Alan Howard and Alex Mawyer
Participant Observation *(Methods)*, Danny Jorgensen
Structural Equation Modeling and Latent Variable Approaches *(Methods)*, Alex Liu
Digital Methods for Web Research *(Methods)*, Richard Rogers