Social Science Data Analysis: The Ethical Imperative

Forthcoming in ***Ethical Data Mining Applications for Socio-Economic Development***, **Hakikur Rahman and Isabel Ramos (eds.)**

Anthony Scime

The College at Brockport, State University of New York, USA

Gregg R. Murray

Texas Tech University, USA

ABSTRACT

Social scientists address some of the most pressing issues of society such as health and wellness, government processes and citizen reactions, individual and collective knowledge, working conditions and socio-economic processes, and societal peace and violence. In an effort to understand these and many other consequential issues, social scientists invest substantial resources to collect large quantities of data, much of which are not fully explored. This chapter proffers the argument that the privacy protection and responsible use are not the only ethical considerations related to data mining social data. Given (1) the substantial resources allocated and (2) the leverage these "big data" give on such weighty issues, this chapter suggests social scientists are ethically obligated to conduct comprehensive analysis of their data. Data mining techniques provide pertinent tools that are valuable for identifying attributes in large data sets that may be useful for addressing important issues in the social sciences. By using these comprehensive analytical processes, a researcher may discover a set of attributes that is useful for making behavioral predictions, validating social science theories, and creating rules for understanding behavior in social domains. Taken together, these attributes and values often present previously unknown knowledge that may have important applied and theoretical consequences for a domain, social scientific or otherwise. This chapter concludes with examples of important social problems studied using various data mining methodologies including ethical concerns.

Keywords- Big Data, Data Mining, Ethical Imperative, Social Problem Analysis, Social Sciences.

INTRODUCTION

Social and economic development are driven by human behavior, and the social sciences are the academic and professional disciplines that study society and human behaviors. These disciplines include but are not limited to: anthropology, communication, criminology, economics, education, history, human geography, law, linguistics, political science, psychology, social work, and sociology. Social scientists address such questions as (Giles, 2011): "How can we induce people to look after their health?" "How do societies create effective and resilient institutions, such as governments?" "How can humanity increase its collective wisdom?" "Why do so many female workers still earn less than male workers?" And "Why do social processes, in particular civil violence, either persist over time or suddenly change?" In an effort to understand the issues in which they are interested, social scientists invest vast resources to ask subjects thousands of questions in surveys and observe millions of behaviors that usually are systematically archived but often are not analyzed in a comprehensive manner.

The protection of privacy and responsible implementation are not the only ethical considerations related to data mining social data. Given (1) the substantial resources allocated and (2) the leverage these "big data" give on such weighty issues, this chapter suggests social scientists are, in addition, ethically obligated to conduct comprehensive analysis of their data, an obligation that data mining techniques can facilitate (National Science Foundation, 2009; Rosenthal, 1994). To support this argument, the chapter begins with a background discussion about resources invested in social data and then a review of some major data sources used in the social sciences. This is followed by examples of how data mining has been used in analysis of political science, social work, sociology, health science, education, and criminal justice issues. These examples only scratch the surface of what can be done with the collected data to meet this ethical obligation. The section on future research directions notes the problem of under-analyzed data will get worse, but new technologies will be developed to solve or mitigate the problem. We conclude on the hopeful note that as data mining becomes main stream for social science analysis the backlog of social science data will decline, and, more importantly, more social problems will be resolved.

BACKGROUND

Social scientists allocate substantial resources to data collection. 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 non-government organizations spend billions of dollars more collecting data on society. This investment represents not only a financial expenditure but also the expenditure of countless days, weeks, and months of researcher, participant, and administrator time and effort. The results of this immense investment are often embodied in extensive data sets. Minimal efficiency demands reasonable output from this substantial input in data collection. This is not a new concept. Rosenthal (1994, 130) contends there is a larger obligation:

> [D]ata 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.

Further, *these data sets may contain answers to some of society's pressing issues*. Very broadly, the challenge to social investigators as stated in the National Science Foundation's mission statement is lofty: "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 social issues.

More specifically, though, these extensive data sets offer substantial leverage on the issues they address. For example, Oatley and Ewart (2003) worked with the West Midlands Police, UK, to develop a decision support system designed to reduce burglary and other crimes. The system used neural networks and a Bayesian belief network based on details of victims, offenders, locations, and the specific victimizations.

In education, a number of schools are using data mining to improve recruitment, attendance and degree completion, curriculum, and fund raising. North Carolina State University is using data mining to find and retain donors and to match them to specific projects. Wichita State University has used data mining in admissions to focus recruitment funds on improving matriculation versus simply identifying acceptable students, with the overall result that successful recruitment has increased by 26% after two years. Based on their analyses, Wichita State admission staff no longer travel out of state to recruit. Further, they more narrowly target their high school visits and college fairs, reducing the number of the latter by 14 per year. Finally, they reduced direct mailings to high school students by over one-third. Telford College in the United Kingdom mined student demographic, attendance, and financial data to create individual student profiles used to track and report attendance, with attendance reports sent to students via mobile applications. As a result, Telford's attendance rates have improved 4% and degree completion rates have improved 3% from 2008-2011. Based on this success, Telford is analyzing courses with high dropout rates to further improve attendance and completion (Media Education Group, 2012).

On a much larger and prominent scale, the 2012 Obama U.S. presidential reelection campaign used data mining and its vast voter data set—which included merged information from pollsters, fundraisers, field workers, consumer databases, social media, and mobile contacts—to "[help] Obama raise $1 billion, [remake] the process of targeting TV ads, and [create] detailed models of swing-state voters that could be used to increase the effectiveness of everything from phone calls and door knocks to direct mailings and social media" (Scherer, 2012; for a fuller account see Issenberg, 2012). According to reports, the analyses helped the Obama campaign exceed fundraising expectations and increased ad buying efficiency by 14% (Scherer, 2012).

Given the combined allocation of resources and the potential to address effectively important social issues, we assert there is an ethical imperative to exploit these and other social data to their fullest extent.

SOCIAL SCIENCE DATA MINING

In general, data mining is one technique to explore an otherwise overwhelmingly large mass of social, behavioral, and economic data and expose the knowledge it contains. It is a data-intense analytical technique that is designed to exploit large data sets, like many of the data sets collected for social science. Data mining involves the analysis of data to find interesting patterns,

confirm and probe previously known relationships, and detect previously unknown relationships in the data. Data mining models not only predict the results of a future event, but they also can provide knowledge about the structure and interrelationships among the data. It is these interrelationships that can lead to a better understanding of the data.

Social Science Data

There are a vast number of extensive data sets in the social sciences that record the social and economic characteristics and changes that occur in society. For instance, the General Social Survey contains data on American social trends collected every two years since 1972 and includes more than 5,300 attributes with time trends on almost 2,000 attributes (NORC, 2010). The Uniform Crime Reporting Program, which has been collecting data since 1930, includes measures on such diverse issues as crime levels as well as law enforcement administration, operation, and management (Federal Bureau of Investigation, 2004). The Baccalaureate and Beyond Longitudinal Study follows 11,000 students who completed their BA/BS degree to assess their education and work experience (Wine et al., 2005). The American National Election Studies are a series of surveys on political participation collected approximately biennially since 1948 that include pre- and post-election studies of both presidential and midterm elections (American National Election Studies, 2010). The Database of Political Institutions (2010) contains cross-country data on political institutions, such as measures of tenure, stability, checks and balances, identification of parties with the government coalition or the opposition, and fragmentation of opposition and government parties in legislatures.

Other examples of extensive social science data include the National Survey on Child and Adolescent Well-Being, which presents data on children and their caregivers collected over a 36-month time period (National Survey of Child and Adolescent Well-Being, 2006); the Global Terrorism Database (Global Terrorism Database, 2009), which contains more than 80,000 cases on domestic and international terrorist events worldwide from 1970 to 2007; the Correlates of War (2009) project, which collects and disseminates quantitative data on international relations; and the World Development Indicators (2010), which includes more than 800 indicators on the world's population, environment, economy, states, and markets. Table 1 summarizes these few examples of the large number of social data sets that are collected.

Table 1. Sample Social Data Sets

Data set	Years	Cases/records	Variables	Focus	Source
American National Election Studies	1948 - present (biennially)	48,000+	900+	Voting, public opinion, and political participation.	American National Election Studies, 2010 http://www.electionstudies.org/studypages/download/datacenter_all.ht

					m
Baccalaureate and Beyond Longitudinal Study	Beginning 1992	17,000+	500+	Students who completed their baccalaureate degree in 1992-93 to assess their education and work experience	Wine et al., 2005 http://nces.ed.gov/pubsearch/index.asp?searchcat2=pubslast6month&HasSearched=1
Database of Political Institutions	1975 – present	178 countries	127	Cross-country data on political institutions	Database of Political Institutions, 2010 http://www.nsd.uib.no/macrodataguide/set.html?id=11&sub=1
General Social Survey, 1972-2008 Cumulative Data Set (March 2010)	1972 – present (biennially)	52,000+	5,300+	Data on contemporary American society and attitudes	NORC, 2010 http://www.norc.uchicago.edu/projects/gensoc1.asp
Global Terrorism Database	1970-2008	87,000+	120	Terrorist events.	Global Terrorism Database, 2009 http://www.start.umd.edu/start/data/gtd/
National Survey on Child and Adolescent Well-Being	1999-2000	5,500	20,000+	Children and families in 97 child welfare systems	U.S. Department of Health and Human Services, 2001 http://www.acf.hhs.gov/programs/opre/research/project/

					national-survey-of-child-and-adolescent-well-being-nscaw-1997-2010
Panel Study of Income Dynamics	1968 – present	65,000	22 broad areas	Economic well-being of American families.	Institute for Social Research, 2010 http://psidonline.isr.umich.edu/
World Development Indicators	1960 - present	210 economies	900+	World population, economy, states, and markets	World Development Indicators, 2010 http://data.worldbank.org/data-catalog/world-development-indicators

As this brief introduction to social science data shows, the social science domains and their data are complex, because they reflect human behavior, which is varied and complex. The importance of these data sets lies in their breadth and depth and the possibility they contain a significant key to understanding a social concern. One of the problems with data sets collected over long periods of time is changing interests and methods of the data collectors. These changes can result in data sets over the years becoming broader, with more attributes, and sparser, with increasing missing values. Data mining techniques are well suited to managing analytically the breadth, depth, and sparseness of this type of data, particularly in regard to reducing data dimensionality and identifying useful relationships.

For instance, Scime and Murray (2007) data mined more than 900 attributes from the American National Election Studies data set to identify 13 measures that successfully predict for which party an individual will vote in a presidential election or whether that individual will vote at all. This is consistent with other approaches to data reduction and rule identification. For example, Fu and Wang (2005) reduced data dimensionality using a separability-correlation measure that ranks the importance of variables to improve classification and the usefulness of rules, while, Murray, Riley, and Scime (2007) reduced data dimensionality by iteratively creating classification models.

Beyond reducing data dimensionality, data mining techniques produce models of the data and domain in the form of individual rules, which may uncover consequential relationships among the values of the attributes. There may be a large number of rules, from which the most useful ones must be identified. Rules can be selected and reduced mechanically and with various levels of guidance from a domain expert. There has been extensive work on the mechanical selection and reduction of classification and association rules, in which the use of domain expertise is limited (Jaroszewicz & Simovici, 2004; Deshpande & Karypis, 2002; Freitas, 2000; Padmanabhan & Tuzhilin, 2000). Rajasethupathy, Scime, Rajasethupathy, and Murray (2009) improved the usefulness of rules by identifying "persistent rules," which are those rules identified by both classification and association mining. But the domain expert often plays a primary role. For instance, in Perception-based Classification (PBC; Ankerst, Ester & Kriegel, 2000) and Iterative Domain Knowledge and Attribute Elimination (Scime & Murray, 2007) the expert and the computer interact together to identify and reduce rules.

The data sources discussed above are available to responsible researchers. These researchers have an ethical obligation to use the data mined models responsibly. Implemented models may affect policy and, hopefully, change and improve society. But as society changes the models likely will be less applicable. There is an ethical responsibility to monitor and survey society continually, data mining again to find the current models and changing policy accordingly. This is a never ending process.

Social Science Mining Applications

Data mining techniques have been used to shed light on a number of important social issues, such as in political science and social work. These studies include analyses of the American National Election Studies (2010) and National Survey on Child and Adolescent Well-Being (2006), and a data set constructed from the Global Terrorism Database (2009), the Correlates of War (2009), the Database of Political Institutions (2010), and the World Development Indicators (2010). These studies were designed to answer questions such as, What factors affect an individual's decision to turn out to vote? What are the long-term predictors of terrorist events? And, How do we improve outcomes for children who enter the child welfare system, especially maltreated children?

Smaller but no less significant data sets have been created by researchers to study issues in sociology, health science, education, and criminal justice, among others. These studies were conducted to answer questions about emotions, residential patterns, time shifting, smoking and alcohol addictions, epidemic containment, student success, learning technology, and crime adjudication and victimization.

In each case the researchers identified a social problem and used data mining techniques to understand the problem and provide possible solutions. They used data mining because of its ability to find interesting patterns, confirm and probe previously known relationships, and detect previously unknown relationships in the data. The resultant data mining models predict the results of a future event or provide knowledge about the structure and interrelationships among the data, which can lead to a better understanding of the data, the problem, and the domain.

Political Science: How Can We Predict Voting and Terrorism?

Political science is the study of the governance of individuals and nations and of the power relations between and among them. It seeks knowledge of political and governmental behavior and institutions. Ultimately it addresses how individuals and society use resources. Besides being used by political campaigns to improve messaging and fundraising, as discussed above, data mining has contributed to our understanding of voting behavior and terrorism.

The American National Election Studies (ANES) is an ongoing, long-term series of public opinion surveys intended to produce research-quality data for researchers who study the theoretical and empirical bases of American national election outcomes using voting behavior, public attitudes, and measures of political participation. The ANES collects data on items such as voter registration and choice, social and political values, social background and structure, partisanship, candidate and group evaluations, opinions about public policy, ideological support for the political system, mass media consumption, and egalitarianism. The ANES has conducted pre- and post-election interviews of a nationally representative sample of adults every presidential and midterm election year since 1948, except for the midterm election of 1950.

The ANES data set is used primarily in the field of political science and contains a large number of records (more than 47,000) and attributes (more than 900). Following an attribute elimination process using classification data mining, researchers reduced the number of attributes necessary to predict the presidential vote choices of individuals to 13 attributes (Scime & Murray, 2007). That is, there are 13 specific survey questions that effectively predict whether individuals will vote and, if they do vote, the party candidate that will get their vote. The results of such a survey will be correct 66% of the time, substantially outperforming previous studies using statistical techniques showing only 51% accuracy.

Another important issue in political science, and in conducting valid public opinion surveys in particular, is the likelihood of a citizen voting in an election. Again using the ANES, but selecting a different set of attributes and instances, the researchers identified two survey questions that together can be used to categorize citizens as likely voters or non-voters. These results met or surpassed the accuracy rates of previous non-data mining models. The two items correctly classify 78% of respondents over a three-decade period. Additionally, the findings indicate that demographic attributes are less salient than previously thought by political science scholars (Murray, Riley, & Scime, 2009).

Terrorism analysts have associated a number of social, political, and economic conditions at the national level with the likelihood that a nation will fall victim to a terrorist event. These conditions generally fall into at least one of four broad categories: level of democracy, economic development, modernization, and social fractionalization. A project was designed to capture long-term predictors of terrorist events that have persisted, and are likely to persist, over time. The researchers constructed a unique data set comprised of terrorism events and measures of social, political, and economic contexts in 185 countries worldwide between the years of 1970 and 2004. The attributes were selected and instances constructed from the Global Terrorism Database (2009), Correlates of War (2009), Database of Political Institutions (2010), and the World Development Indicators (2010).

The unique data set contained 126 attributes and 5431 instances. Analysis by data mining reduced the number of attributes to 22 statistically significant attributes. Further analysis of this same data set using association mining found that the level of democracy is an integral part of

the explanation for terrorism. The results suggest that more democratic states are more likely to suffer a terrorist attack. The contra theory, which asserts that higher levels of democracy lead to fewer terrorist attacks, is not supported. Economic development and modernization are also not supported as significant factors leading to terrorist attacks (Scime, Murray, & Hunter, 2010).

Social Work: How Can We Improve the Lives of At-Risk Children?

Social work is the discipline that studies how to improve individual and group quality of life and wellbeing using social interventions. Research is conducted in areas such as human development, social policy, public administration, psychotherapy, program evaluation, and international and community development. Few if any areas of social work are more important than child welfare. Data mining has led to an understanding of how the lives of at-risk children can be improved.

The National Survey on Child and Adolescent Well-Being (NSCAW) is a rich collection of data designed to represent children and families who enter the child welfare system. The NSCAW data set contains descriptive information on children, families, parents and caregivers, community environmental factors, and caseworker/intervention factors that may shed light on ways to improve outcomes for children who enter the child welfare system. NSCAW tracks service interventions over time to assess the effects of support on parenting in families of origin or in foster families. These data are appropriate for analysis of child welfare outcomes such as the safety, permanence of care, and well-being of children.

The NSCAW data set is used in the fields of social work and child welfare and contains many fewer records (5,501) but many more attributes (more than 20,000) than the ANES. Each record is a composite of data collected by survey from the children's biological parents, caregivers, teachers, caseworkers, and administrative records. Multiple temporal attribute values were collected over the period of the study from face-to-face and telephone interviews and assessments conducted as follow-up with children, parents, non-parent caregivers, and teachers, as well as data on school engagement and performance. Agency and system-level data were added to the records from data collected from caseworkers, agency administrators, and the States. Using data mining, researchers reduced the number of attributes necessary to understand child placement conditions in homes to eight with 84% accuracy. As important, the resulting rules identify meaningful relationships regarding the living arrangements of maltreated children, suggesting important inflection points pertaining to the living arrangements for these children (Scime, Murray, Huang, & Brownstein-Evans, 2008).

Sociology: How Can We Understand Emotions? Explain Residential Patterns? Save People Time?

Sociology is the study of society's institutions, interactions, and relationships. It includes social stratification, social class, culture, social mobility, religion, secularization, law, and deviance. Data mining has been used to help us understand social relationships, residential living patterns, and people's use and saving of time.

Social network environments are places where emotions are expressed. Sentiment analysis is the determination of the feelings or attitudes of a speaker or a writer. Thelwall, Wilkinson, and Uppal (2010) data mined public comments on a major social networking site to

determine the strength of positive and negative emotions and the relationship to gender. They found that positive emotion is expressed by women in about two thirds of the comments. This result corresponds with previous research where it was found that females tend to use positive emotions more than males. However, negative emotions are found to be not associated with gender and are clearly rarer than positive emotions. Further study of social networks may lead to an understanding of the importance of emotions in communication and strategies for emotion use.

Spielman and Thill (2008) explored the relationship between social similarity and geographic proximity using data mining and a geographic information system. They identified a number of social patterns by looking at New York City's complex demographic structure as represented in census data. The results bring into question Tobler's First Law of Geography, which states that geographically close individuals or objects are more similar than distant individuals or objects. Tobler's first law of geography defines the difference between geography and other disciplines, such as sociology. The first law is the basis for spatial auto-correlation and geo-statistics, important in Geographic Information Systems (GIS). Spielman and Thill (2008) point out that additional case studies are needed before the first law is found no longer viable.

Significant changes are occurring in people's use of time. With the proliferation of digital video recorders people are time shifting; that is, they are taking advantage of technology to postpone television watching. This has an effect on advertisers as well as viewers. Data mining can analyze viewing patterns, to include when viewing takes place to understand the demographic and behavioral characteristics of viewers (Spangler, Gal-Or, & May, 2003).

Health Science: How Can We Reduce Smoking and Alcoholism? The Emergence of Epidemics?

Health Science is a complex discipline studying the determinants of personal and societal health. It includes health education, promotion and behavior change, substance abuse counseling, health care administration, and public health. Data mining has contributed to knowledge about smoking and alcoholism, the need for emergency cesarean procedures, cancer treatment, and epidemic containment.

Data mining examined factors that might be predictive of healthcare staff's advising elderly residents on tobacco cessation. The analyses showed that healthcare workers' license level, beliefs regarding effectiveness of giving advice, and administrative policy on authority to give advice were predictive of not giving advice to encourage elderly residents to stop smoking (Watt, Lassiter, & Scheidt, 2009).

Among the factors that affect the recovery of alcohol addicts are the characteristic of the client and the counselor. Data mining records containing the client's stage of recovery, addiction types, location, year, sex, marital status, ethnicity, employer, age, counselor, relative counselor/client sex, and relative counselor/client ethnicity 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, non-white female with multiple addictions meeting with a black female counselor. The results also identified the most effective counselors (Burn-Thornton & Burman, 2009). These results can influence policy on the assignment of clients to counselors and future study of the techniques used by the best counselors.

Hospital surveillance data are used to forecast emerging epidemics. Using data mining, surveillance data were analyzed to classify patients into three groups based on the criticality of their condition: high, mid, and low. This model also provides precautionary measures to be taken to reduce the level of damage caused by an epidemic (Hasan, & Rahman, 2009).

Education: How Can We Improve Student Behavior? Student Success? Classroom Technology?

Education as a discipline looks to improving the learning experiences of students. This requires the study of many facets often linked to other disciplines but always in a school or learning setting. Data mining has identified issues in monitoring student behavior, student success determinants, and technology use for learning.

Educators are asked to use data to identify systemic issues in schools, which can result in the development of systemic interventions aimed at mitigating identified issues. One such issue is student discipline. By identifying the student attributes associated with particular discipline issues, predictions can be made concerning the groups of students involved, when discipline problems occur, and what problems are most significant. A dataset was constructed from one school's disciplinary data. This data set contains records of the 35,272 disciplinary problems occurring in the 2008-2009 school year by students in grades 8-11. Included in the data set were student demographics, the discipline problem, and the day of occurrence. Data mining identified the who, what, and when for discipline problems at this school (Scime & Reiner, 2012).

Colleges want to admit students who will grow, succeed, and graduate. The admitted students need to be qualified academically and emotionally. Selection of the best students that will attend is the goal of the admissions process. Similar to the work at Wichita State discussed above, Chang (2006) mined college admissions data to create a model that predicted admissions yield more accurately than solutions based on logistic regression. The results were found to be actionable and practical at the individual level, making the model highly desirable to enrollment decision makers.

Likewise, scholarships often are targeted toward to students who will most likely benefit. Data mining applied to scholarship data has provided insight into the characteristics of successful scholarship recipients. The results can be used as guidelines for selecting new recipients from among applicants as well as identifying potential applicants (Media Education Group, 2012; Francia & Sanders, 2009). Likewise, by identifying the characteristics for failure to maintain a scholarship, proactive efforts can be taken to correct and address problem areas (Francia & Sanders, 2009).

Data mining has been used to analyze the usage patterns of one comprehensive college's learning management system (LMS). The data included counts of all the features the LMS offered its users for the Fall and Spring semesters of the academic years beginning in 2007 and 2008. The data set contained 9430 records, one for each class section, and 20 attributes, the features. Various data mining techniques were applied to evaluate which LMS features are used most commonly and most effectively by instructors and students. These results could be used by the institution, as well as similar institutions, for decision making concerning feature selection and overall usefulness of LMS design, selection, and implementation or to identify feature areas needing additional training in their use (Swanger, Whitlock, Scime, & Post, 2012).

Student retention in higher education is a continuing problem. Whether due to absenteeism, as at Telford above, or for personal or academic reasons, students leave school. With sufficient warning school leaders can mitigate the loss of students with the appropriate interventions. But, the students and the reasons need to be identified. Daimi and Miller (2009) demonstrated how classification mining can advise institutions on student retention. Data mining helps institutions understand retention risks, provides a list of students most likely to leave, and enlightens student retention policies.

Criminal Justice: How Can We Improve Equal Treatment Under the Law? Protect Victims of Crime?

Criminal justice studies the reasons for and costs of criminal behavior. This includes law enforcement and the judicial system. In addition to the crime reduction system set up in the UK discussed above, data mining has informed this discipline on inequities in judicial decisions and the effects of crime on victims.

A contextual analysis of state and federal judicial decisions from 1998 to 2010 resulted in a data set of adjudication differences between celebrities and non-celebrities. This data set contained 105 records and five attributes (Carroll & Scime, 2012). This study considered the effect of exogenous influences on judicial decision-making and the impact of celebrity status. The findings suggest that celebrities do not receive special treatment but in fact are convicted at a higher rate than non-celebrity defendants.

A Bayesian belief network classifier that predicts victimization was developed using data from the National Crime Victimization Survey. The National Crime Victimization Survey is the United States' annual collection of criminal victimization information. The results predicted the value of the victimization attribute with 99% accuracy (Riesen & Serpen, 2009). This research showed how Bayesian belief network and data mining in general may be use to predict victimization in the criminal justice domain.

FUTURE RESEARCH DIRECTIONS

Social science data sets are often large and complex, and advances in data collection and processing have made the analysis of such massive data sets viable. Data mining techniques more fully exploit data collections and, importantly, shed a brighter light on some important questions in the social sciences. Social scientists with access to these vast stores of data recognize the untapped knowledge they likely hold and are beginning to realize that data mining is a valuable tool for uncovering this important knowledge.

Information scientists will continue to develop and improve data mining algorithms for social science data analysis. New algorithms designed to find new social knowledge, find and prove or disprove social theories, and predict human behavior will continue to emerge. Together social and information scientists can assess data to improve the human condition.

For instance, a new avenue for social interaction is the Internet and Web 2.0. Social media and other Web 2.0 application use have created new, vast, and fast growing data collection (Wilson, Gosling & Graham, 2012). New technologies in data mining are being developed to analyze these social networks. Technologies being developed or improved include network

analysis, streaming analysis, temporal database analysis, text mining, and content analysis, among others. Currently these techniques are primarily focusing 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, Wilkinson, and Uppal (2010).

Further, government entities around the world collect and control vast stores of data. E-Government is the use of computing and the Internet by government agencies to collect, store, and calculate data in the execution of their duties. The concept of e-Government is to provide more efficient and effective service to citizens, businesses, and other parts of the government. Data mining will be one aspect of the analysis of these collected data, and the resulting models may usefully inform policy makers. It is incumbent on these policy makers to use the data ethically for the whole of the country's citizens.

CONCLUSION

Social science research is motivated by the desire to understand and solve human problems. Each of the above examples demonstrates the ability of data mining techniques to address key concepts that may be uncovered in large social science data sets. The studies uncovered attributes that effectively predict and identify significant factors that can address a problem in a domain that involves people or organizations. The results of these analyses suggest that data mining is useful for more fully analyzing important but under-evaluated data collections and for addressing consequential questions in the social sciences. Hence, it seems reasonable to conclude there is an obligation to "hold up these data to the light in many different ways" (Rosenthal, 1994: 130).

Modern social scientists are fortunate to have access to vast stores of data. The data sets are often so immense, though, that comprehensive analysis using conventional techniques is all but impossible. Data mining techniques can be employed to find useful patterns in these data. While data mining is not the only analytical method and it is not well recognized in the social sciences, it is a collection of techniques that may be able to provide an easily interpretable and implementable analysis on data where other methods are not suitable. This will be increasingly vital as advances in data collection and processing accelerate the creation of massive data sets and, therefore, advance the opportunities for addressing important societal issues.

Today, social science data are ubiquitous and the tools needed to analyze them (e.g., computing hardware and software) are easy to use, inexpensive, and fast. The costs to exploit more fully these data are plummeting. Given the pressing nature of some of these problems and the massive investment made in collecting these data, one can very easily argue there is an ethical imperative to analyze these data as comprehensively as possible.

Privacy concerns and responsible use are not the only ethical considerations related to data mining in the social sciences given the substantial resources allocated to data collection and the leverage these "big data" give on important social issues. In this chapter we have suggested there is an ethical imperative to analyze collected data. We have provided examples of data mining analysis in political science, social work, sociology, health science, education, and criminal justice. The future use of data mining in the social science domains has the potential to influence government, business, and non-governmental organizations in policy development that

can enhance, if used and pursued ethically, the socio-economic development of citizens around the globe.

REFERENCES

American National Election Studies. (2010). Center for Political Studies, University of Michigan, Ann Arbor, MI.

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

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

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

Chang, L. (2006). Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research*, 131, 53-68.

Correlates of War. (2009). Retrieved April 5, 2010, from http://www.correlatesofwar.org/.

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

Database of Political Institutions. (2010). Retrieved September 1, 2010 from http://www.nsd.uib.no/macrodataguide/set.html?id=11&sub=1.

Deshpande, M. & Karypis, G. (2002). Using conjunction of attribute values for classification. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (pp. 356-364). McLean, VA.

Federal Bureau of Investigation (2004). Uniform crime reporting handbook, revised edition. U.S. Department of Justice. Washington, DC: Federal Bureau of Investigation.

Francia, G. A. & Sanders, C. (2009). Applied data mining in a scholarship program. In *Proceedings of the 2009 International Conference on Frontiers in Education* (pp. 336-340). Las Vegas, NV.

Freitas, A. A. (2000). Understanding the crucial differences between classification and discovery of association rules – A position paper. *ACM SIGKDD Explorations Newsletter*, 2(1), 65-69.

Fu, X. & Wang, L. (2005). Data dimensionality reduction with application to improving classification performance and explaining concepts of data sets. *International Journal of Business Intelligence and Data Mining*, 1(1), 65-87.

Giles, J. (2011). Social science lines up its biggest challenges. *Nature*, 470, 18-19.

Global Terrorism Database (2009). Retrieved May 4, 2009 from http://www.start.umd.edu/start/data/gtd/.

Hasan, F. R. & Rahman, R.M (2009). Mining ICDDR,B hospital surveillance data using decision tree classification algorithm. In *Proceedings of the 2009 International Conference on Information & Knowledge Engineering* (pp. 290-296). Las Vegas, NV.

Institute for Social Research (2010). Panel study of income dynamics. Retrieved August 26, 2010 from http://psidonline.isr.umich.edu/.

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

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

International Conference on Knowledge Discovery and Data Mining (pp. 178-186). Seattle, WA.

Liu, S., Rusen, I. D., Joseph, K. S., Liston R., Kramer, M.S., Wen, S. W., & Kinch, R. (2004). Recent trends in caesarean delivery rates and indications for caesarean delivery in Canada. *Journal of Obstetrics and Gynecology Canada*, 26(8), 735-42.

Media Education Group (2012). *Building a Smarter Campus: How Analytics is Changing the Academic Landscape*. IBM. Retrieved November 15, 2012 from https://www-01.ibm.com/software/analytics/education/.

Murray, G. R., Riley, C. & Scime, A. (2007, May). *A new age solution for an age-old problem: mining data for likely voters*. Paper presented at the 62nd Annual Conference of the American Association of Public Opinion Research, Anaheim, CA.

Murray, G. R., Riley, C. & Scime, A. (2009). Pre-election polling: Identifying likely voters using iterative expert data mining. *Public Opinion Quarterly*, 73(1), 159-171.

National Science Board (2012). *Science and Engineering Indicators 2012.* Retrieved July 7, 2012 from http://www.nsf.gov/statistics/seind12/c5/c5s1.htm.

National Science Foundation (2009). *National Science Foundation FY 2005 Performance Highlights*. Retrieved August 25, 2010 from http://www.nsf.gov/pubs/2010/nsf10002/nsf10002.pdf.

National Survey of Child and Adolescent Well-Being (2006). U.S. Department of Health and Human Services; Administration for Children and Families; Office of Planning, Research, and Evaluation.

NORC (2010). GSS study description. National Organization for Research at the University of Chicago. Retrieved September 1, 2010 from http://www.norc.uchicago.edu/projects/gensoc1.asp.

Oatley, G.C. & Ewart, B. W. (2003). Crimes analysis software: 'pins in maps', clustering and Bayes net prediction. *Expert Systems with Applications*, 25, 569–588.

Padmanabhan, B. & Tuzhilin, A. (2000). Small is beautiful: discovering the minimal set of unexpected patterns. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 54-63). Boston, MA.

Rajasethupathy, K., Scime, A., Rajasethupathy, K. S. & Murray, G. R. (2009). Finding "persistent rules": combining association and classification results. *Expert Systems With Applications*, 36(3P2), 6019-6024.

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

Rosenthal, R. (1994). Science and ethics in conducting, analyzing, and reporting psychological research. *Psychological Science*, 5(3), 127-134.

Scherer, M. (2012). Inside the secret world of the data crunchers who helped Obama win. Time. Retrieved November 11, 2012 from http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/.

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

Scime, A., Murray, G. R., Huang, W. & Brownstein-Evans, C. (2008). Data mining in the social sciences and iterative attribute elimination. In D. Taniar, (Ed.), *Data Mining and Knowledge Discovery Technologies* (pp. 308-332). Hershey, PA: IGI Publishing.

Scime, A., Murray, G. R. & Hunter, L. Y. (2010). Testing terrorism theory with data mining. *International Journal of Data Analysis Techniques and Strategies, 2(2)*, 122-139.
Scime, A. & Reiner, S. (2012). Finding interesting classification rules: an application from education. In *Proceedings of the 2012 International Conference on Data Mining* (pp. 37-43). Las Vegas, July.
Spangler, W. E., Gal-Or, M. & May J. H. (2003). Using data mining to profile TV viewers. *Communications of the ACM*, 46(12), 66-72.
Spielman, S. E. & Thill, J-C. (2008). Social area analysis, data mining, and GIS. *Computers, Environment and Urban Systems*, 32(2), 110-122.
Swanger, T., Whitlock, K., Scime, A., & Post, B. (2012). "ANGEL mining" in R. Babo & A. Azevedo (Eds.), *Higher education institutions and learning management systems: Adoption and standardization* (pp. 94-115). Hershey, PA: Information Science Reference.
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(1), 190–199.
U.S. Department of Health and Human Services (2001). *Safety, Permanence, Well-being: Child Welfare Outcomes 2001 Annual Report*. Washington, National Clearinghouse on Child Abuse and Neglect Information.
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. In *Proceedings of the 2009 International Conference on Data Mining* (pp. 561-565). Las Vegas, NV.
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.
Wine, J.S., Cominole, M.B., Wheeless, S., Dudley, K., & Franklin, J. (2005). 1993/03 Baccalaureate and beyond longitudinal study (b&b:93/03) methodology report. U.S. Department of Education. Washington, DC: National Center for Education Statistics.
World Development Indicators (2010). Retrieved September 1, 2010 from http://data.worldbank.org/data-catalog/world-development-indicators.

ADDITIONAL READINGS

Ashrafi, M.Z., Taniar, D., & Smith, K. (2007). Redundant association rules reduction techniques. *International Journal of Business Intelligence and Data Mining*, 2(1), 29-63.
Barber, B. & Hamilton, H. J. (2003). Extracting share frequent itemsets with infrequent subsets. *Data Mining and Knowledge Discovery*, 7(2), 153–185.
Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., & Lakhal, L. (2000). Mining minimal nonredundant association rules using frequent closed itemsets. In *Proceedings of the 1st International Conference on Computational Logic* (pp. 972-986). London, UK.
Bay, S. D. & Pazzani, M. J. (1999). Detecting change in categorical data: Mining contrast sets. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining* (pp. 302-306. San Diego, CA.
Bayardo, R. J. & Agrawal R. (1999). Mining the most interesting rules. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining* (pp. 145-154). San Diego, CA.

Breiman, L., Freidman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth and Brooks.

Broder, A. Z., Charikar, M., Frieze, A.M., & Mitzenmacher, M. (1998). "Min-wise independent permutations," In *ACM Symposium on Theory of Computing* (pp. 327-336). Dallas, TX.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Weiner, J. (2000). Graph structure in the web," *Computer Networks*, 33(6), 309–320.

Calders, T., & Goethals, B. (2007). Non-derivable itemset mining. *Data Mining and Knowledge Discovery*, 14(1), 171-206.

Carter, C. L., Hamilton, H. J., & Cercone, N. (1997). Share-Based measures for itemsets. In In *Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery* (pp. 14-24). Trondheim, Norway.

Chan, R., Yang, Q., & Shen, Y. (2003). Mining high-utility itemsets. In *Proceedings of the 3rd IEEE International Conference on Data Mining* (pp. 19-26). Melbourne, FL.

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *ACM Symposium on Theory of Computing* (pp. 380-388). Montreal, Canada.

Dong, G. & Li, J. (1998). Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. In *Proceedings of the 2nd Pacific Asia Conference on Knowledge Discovery in Databases* (pp. 72-86). Melbourne, Australia.

Gaber, M. M., (2010) *Scientific data mining and knowledge discovery — Principles and foundations*, New York, NY: Springer.

Gay, L. R., Mills, G. E., & Airasian, P. (2006). *Educational research: Competencies for analysis and applications* (8th ed.). Upper Saddle River, NJ: Pearson.

Geng, L., & Hamilton, H.J. (2006). Interestingness Measures for Data Mining: A Survey. *ACM Computing Surveys*, 38(3), No.9.

Gibbons, P. B. (2001). Distinct sampling for highly-accurate answers to distinct values queries and event reports. In *International Conference on Very Large Databases* (pp. 541-550). Rome, Italy.

Hamilton, H. J., Geng, L., Findlater, L., & Randall, D. J. (2006). Efficient spatio-temporal data mining with GenSpace graphs. *Journal of Applied Logic*, 4(2), 192-214.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W., (Eds.). (1985). E*xploring data tables, trends, and shapes*. New York, NY: Wiley.

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

Kancherla, K., Chilakapati, R., Suryakumar, D., Cousins, J., Dorian, C., & Mukkamala, S. (2009). Lung cancer detection and classification (using microarray data). In *Proceedings of the 2012 International Conference on Bioinformatics and Computational Biology* (pp. 168-176). Las Vegas, NV.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A.I. (1994). Finding interesting rules from large sets of discovered association rules. In *Proceedings of the Third International Conference on Information and Knowledge Management* (pp. 401-408). Gaithersburg, Maryland.

Leedy, P. D., & Ormrod, J. E. (2010). *Practical research: Planning and design* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Li, G. & Hamilton, H. J. 2004. Basic association rules. In *Proceedings of the 4th SIAM International Conference on Data Mining* (pp. 1166-177). Orlando, FL.

Liu, B., Hsu, W. & Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (pp. 27-31). New York, NY.

Lu, S., Hu, H., & Li, F. (2001). Mining weighted association rules. *Intelligent Data Analysis*, 5(3), 211-225.

Murray, G. & Scime, A. (2010). "Microtargeting and electorate segmentation: data mining the American National Election Studies. *Journal of Political Marketing*, 9(3), 143-166.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining Knowledge Discovery*, 2(4), 345-389.

Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., & Lakhal, L. (2005). Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, 24(1), 29-60.

Scime, A., Rajasethupathy, K., Rajasethupathy, K.S., & Murray, G. R. (2011). "Finding persistent strong rules: Using classification to improve association mining" In A. V. S. Kumar, (Ed.), *Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains* (pp. 85-107). Hershey, PA: Information Science Reference.

Sears, L., Hashemi, R. R., Sears, P., & Smith, M. (2009). Video mining: Theatrical upstaging detection. In *Proceedings of the 2009 International Conference on Information & Knowledge Engineering* (pp. 560-566). Las Vegas, NV.

Sekaran, U. (2003). *Research methods for business* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Shen, Y. D., Zhang, Z., & Yang, Q. (2002). Objective-oriented utility-based association mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining* (pp. 426-433). Maebashi City, Japan.

Srikant, R., & Agrawal, R. (1995). Mining generalized association rules, In *Proceedings of the 1995 International Conference on Very Large Data Bases* (pp. 407-419). Zurich, Switzerland.

Stefanakos, S. (2009). Using social ties to predict missing customer information. In *Proceedings of the 2009 International Conference on Data Mining* (pp. 42-47). Las Vegas, NV.

Swan, G. & Scime, A. (2010). Winning baseball through data mining. In *Proceedings of the 2010 International Conference on Data Mining* (pp. 151-157). Las Vegas, NV.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*, Upper Saddle River, NJ: Addison-Wesley.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hätönen, K., & Mannila, H. (1995). Pruning and grouping of discovered association rules. In *Proceedings of ECML-95 Workshop on Statistics, Machine Learning, and Discovery in Databases* (pp. 47-52). Heraklion, Crete.

Vaillant, B., Lenca, P., & Lallich, S. (2004). A clustering of interestingness measures. In *Proceedings of the 7th International Conference on Discovery Science* (pp. 290-297). Padova, Italy.

Webb, G. I. & Brain, D. (2002). Generality is predictive of prediction accuracy. In *Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop* (pp. 117-130). Tokyo, Japan.

Zbidi, N., Faiz, S., & Limam, M. (2006). On mining summaries by objective measures of interestingness. *Machine Learning*, 62(3), 175–198.

Zhang, H., Padmanabhan, B., & Tuzhilin, A. (2004). On the discovery of significant statistical quantitative rules. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining* (pp. 374-383). Seattle, WA.

Zhong, W., Chow, R., He, J., Stolz, R., & Dowel, M. (2009). Multi-level support vector machines for classifyinglarge chronic disease datasets. In *Proceedings of the 2012 International Conference on Bioinformatics and Computational Biology* (pp. 370-375). Las Vegas, NV.

KEY TERMS AND DEFINITIONS

Association Mining – A data mining method that discovers frequent patterns, associations, correlations, or causal structures among sets of attributes in data sets. A frequent pattern is a pattern (set of attributes or a sequence) that occurs with some pre-established frequently in the data set.

Attribute – A characteristic of an instance in the data. Also known as a(n) data element, field, item, data field, data item, and column, among other things.

Classification Mining – A data mining method that constructs a model of the data's behavior used to determine the expected classification of future instances. The model constructed from the data is a decision tree. The decision tree consists of decision nodes and leaf nodes, beginning with a root decision node, connected by edges. Each decision node is an attribute of the data and the edges represent the attribute values. The leaf nodes represent the dependent variable; the expected classification results of each data instance.

Data Dimensionality Reduction – The act of selecting attributes and instances to simplify the data without reducing the classification capabilities of the resultant model.

Domain Expert – A person with a strong theoretical foundation in the specific field for which the data were collected. A domain expert understands the practical implications of the data and can interpret the effect on the domain from the rules resulting from the data mining.

Record – A set of attributes that together define a single, unique entry in the data. Also known as a instance, entity, row, case, transaction, etc.