# 19

# Data Mining

*JACOB FURST, DANIELA STAN RAICU, AND LEONARD A. JASON*

Data mining, the subject of this chapter, has been most frequently used in the physical sciences (Kutz, 2013). However, as we shall show, it has also been successfully applied by social science investigators of community-level phenomena. Because they can be used to uncover patterns and relationships within large samples of people, organizations, or communities that would not otherwise be evident because of the size and complexity of the data, data mining methods are particularly appropriate for research on social problems.

Increasingly, as researchers, we are confronted with ever-larger data sets, and, as we bring diverse voices (e.g., consumers, community-based groups, government officials, and media and electronic sources) into our work, the complexity will inevitably increase (Dhar, 2013). With these vast new reservoirs of information, there is a need for us to develop methods to understand the dynamic transactions that occur between individuals and their social environments. Data mining is one method that helps us understand such voluminous data in new and more efficient ways. The IBM computer that was used on the television program *Jeopardy* in 2011 to defeat master human players had 16 terabytes of memory, an unimaginable amount of memory capacity at that time, but such an amount may be on desktop computers within the next 10 years (Harris, 2008). We are quickly having access to more and more powerful programs to process and search for solutions, ones in which computers actually learn and then provide us with ways of better understanding these large data sets. These processes are ones with which social scientists are now dealing and which could help solve some formerly intransigent social and community problems.

Social problems that could benefit from the use of data mining include detecting underlying communities, analyzing behaviors, and discovering evolutionary patterns in a community (Wang, Tong, Yu, & Aggarwal, 2012). For example, several studies (Davidson, Gilpin, & Walker, 2012; Ferdowsi, Settimi, & Raicu, 2010; Jiang, Ferreira, & Gonzalez, 2012) offered new perspectives for urban and transportation planning, as well as emergency response systems. Jiang et al. (2012) analyzed activity-based travel survey data from the Chicago metropolitan area to learn when, where, and how individuals interact with places in metropolitan areas. Ferdowsi et al. (2010) employed socioeconomic and housing data for the city of Chicago to help understand social changes of urban areas leading to the gentrification or abandonment of communities.

In this chapter, we will provide an overview of one method of data mining that uses decision trees to predict a classification (e.g., negative outcomes of high-risk neighborhoods in a community), based on successive binary choices of risk factors. At each branch point of the decision tree, a characteristic is examined (e.g., gang activity within a community), and the decision tree determines whether a characteristic is important in the outcome or classification. In data mining, multiple characteristics are reviewed, and an algorithm is ultimately developed that best predicts outcomes. We will then illustrate the application of this method to a chronic health condition, showing how computer-generated algorithms were developed to help guide community organizations and government bodies in arriving at more valid and less stigmatizing ways of characterizing patients.

## INTRODUCTION TO DATA MINING AND DECISION TREES

Data mining is the process of discovering hidden, implicit, nontrivial, and useful patterns from large amounts of data. Figure 19.1 indicates that this process is an iterative and interactive sequence of steps that includes domain understanding, data collection, data preprocessing, data reduction, pattern discovery, and pattern evaluation for knowledge extraction. In the first step, domain experts and data mining experts formulate the research question or problem to be addressed using data mining. In the second step, the data are either collected or extracted from data resources, such as data warehouses, data marts, and databases. Third, because data rarely come in a clean format, a preprocessing step is required to do a number of functions, including, for instance, removing duplicates, filling in missing values, and solving any inconsistencies in the attributes. The process of collecting and preprocessing the data is time consuming and usually takes between 60% and 80% of the entire data mining process. Once the data are cleaned, a reduction in the number of attributes or number of cases may be necessary if the number of attributes is too large compared to the number of cases or the number of cases is too large to allow efficient modeling of the data. The next step, pattern discovery, employs such techniques as machine learning, artificial intelligence, and statistics to uncover patterns in the data.

Traditionally, these techniques can be *supervised* or *unsupervised*, depending on the availability of labeled data. If all the data samples have known labels, then supervised techniques can be used; supervised techniques use the known labels of existing data to create models to predict the unknown labels of new data. For example, a body of historical medical data, including patient symptoms and diagnosis, could be used to create a supervised learning model to diagnosis new patients based on their symptoms. If the data samples do not have known labels, then unsupervised techniques need to be applied to learn from the data based on the similarities among the cases; unsupervised techniques separate the data into similar categories based solely on relationships between the features of the data samples. To extend the earlier example, if the historical patient data had no diagnosis, unsupervised techniques could be used to separate the patients into similar symptom groups. Techniques for supervised learning include neural networks, decision trees, Bayesian classifiers, and support vector machines (Kotsiantis, 2007). Clustering techniques, including partitioning and hierarchical techniques (Ghahramani, 2004), are the most popular ones for unsupervised learning.

In the rest of this section, we will focus on decision trees as a machine learning technique for classification. Machine learning is one of the major disciplines used to support data-driven (i.e., empirical) research, research in which the data are too many for a reasonable hypothesis to be formulated a priori, making hypothesis-driven research impractical. A decision tree is a method of machine learning that is primarily focused on the task of data classification: predicting the category (or label) of data samples based on the attributes (or features) of the samples. Therefore, a decision tree is a supervised machine learner; that is, there must be samples with a known label from which to construct a

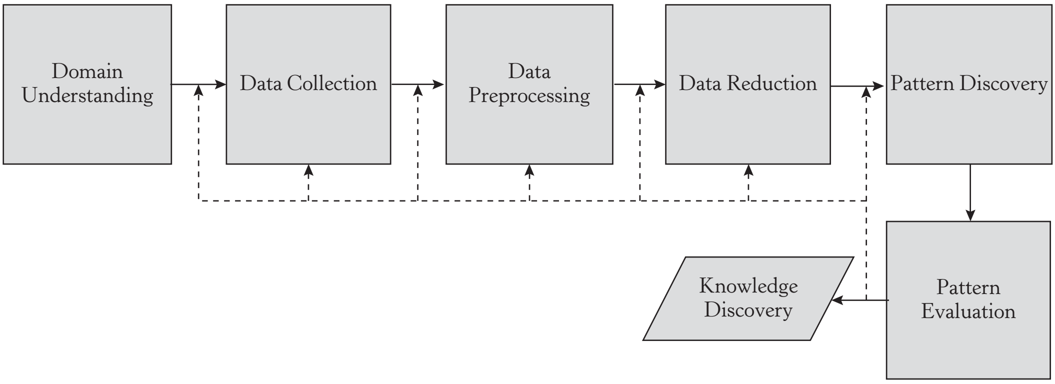

**FIGURE 19.1**: An overview of the data mining process.

model (the decision tree), on which future samples (with unknown labels) can be classified.

A decision tree is constructed by examining the features and labels of the data set and deriving a split of the data set based on a single feature and corresponding threshold of that feature that improves some measure of data consistency or classification accuracy. That is, the decision tree splits the data based on the value of some feature, such that the accuracy or consistency of the resultant subsets is better than the accuracy or consistency of the original data set. A decision tree will generally perform a comprehensive search of all features and all possible threshold values to determine the best split of the data. The measure of consistency or accuracy will depend on the kind of classification tree being used and the input of the user. The two most common methods of measuring "goodness of split" are Gini impurity and information gain (Breiman, Friedman, Olshen, & Stone, 1984). In most cases, the two metrics will behave similarly. After each split, the two subsets are recursively analyzed to determine if improvement can be made by splitting them. The tree will stop splitting when no further improvement can be gained. Consistent with the terminology used in general computer science data structures theory, subsets of data in the tree are called nodes, and the number of splits required to reach a node starting from the root is called the depth of the tree. The original data set has the special designation of root node and is at level zero.

Although accuracy, defined as the number of correctly classified cases over the total number of cases, is in general used to evaluate the performance of a classifier, there are other performance measures that can be employed as well. Specifically, in the biomedical and health care domains, when the interest is in the performance with respect to the positive class (it has the disease) versus the negative class (it does not have the disease), sensitivity and specificity are used. Sensitivity is the ratio between the number of correctly classified positive cases (true positives) over the total number of positive cases. Specificity is the ratio between the number of correctly classified negative cases (true negatives) over the total number of negative cases. A Receiver Operator Characteristic (ROC; Green & Swets, 1966) is used to visualize the relationships between specificity and sensitivity and to determine the best combination of parameters for the highest possible sensitivity and specificity.

Decision trees come in a variety of types, depending on the intended outcome and the method of building the tree. Classification trees are used to predict a discrete numerical or categorical label, while regression trees are used to predict a continuous numerical label. Frequently, the terms C&RT, CART, or Classification & Regression Tree (Breiman et al. 1984) are used to include both categories. CHAID (Kass, 1980) is a variation that allows for more than a single split at each node of the tree. It can be helpful if the data are missing values, as a split can involve a feature threshold value (or values), as well as a node for missing values (which cannot be determined to be above or below a threshold).

Among the most important advantages of decision trees is that they make no assumptions about the distribution of the underlying data. In particular, features do not have to be normally distributed for the tree to generate accurate and robust results. This can be especially important when the number of samples is very small. Decision trees are generally easy to understand and interpret. Using thresholding on feature values to split the data set into two more consistent data sets is an intuitive idea and easy to demonstrate. The features and their corresponding thresholds can also be stated as a Boolean logic decision rule, which can be easily and quickly applied to new cases.

Decision trees have built-in feature selection. A decision tree model can be easily analyzed to determine which features were important for the classification. This can refine and simplify further data collection and provide insights into properties of the data beyond the classification results.

Although decision trees require minimal data preparation, they do have a number of constraints that are important to remember when interpreting the results of classification: (a) They will generally overfit the data. (b) They use a "greedy" strategy. (c) They can be very sensitive to input parameters. (d) They can be sensitive to label sets of unequal size.

Overfitting is caused by the recursive nature of the construction of the decision tree. That is, because the tree stops splitting only when no further improvement on purity can be gained, a decision tree will always predict the known label set perfectly. If new (unclassified) elements do not match the original, labeled set perfectly, they will be misclassified. Thus, most decisions trees are limited in their growth, so as to find a balance between the predictive accuracy on the known set and the predictive accuracy on unknown elements.

There are generally three ways in which to limit the growth of a decision tree. The first method restricts the minimum size of a node before it can be split. The input parameter that restricts this is called the parent node size parameter. The second method restricts the minimum size of a child node resulting from a split. The input parameter that restricts this is called the child node size parameter. The third method limits the depth of the tree, not allowing nodes to split past a certain depth. There also exist pruning techniques, which do not limit the initial growth of the tree, but postprocess the finished tree to remove splits that are likely overfit. We used growth-limiting parameters, rather than pruning, in the case study to be presented.

A "greedy" strategy is an iterative solution that will always make decisions that are the best at the moment, without regard for previous or potential future decisions of the solution. With decision trees, this shows up in two significant ways. First, the decision tree will choose the single best feature on which to generate a split. If two features are highly correlated, and might produce very similar splits, the decision tree will choose the better of the two. The second feature may then not be optimal for subsequent splits and may not show up at all in the resulting decision rules, leading to an incorrect conclusion about the possible importance of the two correlated features. Second, the decision tree must choose a single feature for a split; the tree cannot choose, for instance, a pair of features and a double threshold that might be better than either of the features alone. There has been some initial unpublished work in the area of choosing feature pairs, but it has not yet established its value. In general, and for decision trees in particular, a "greedy" algorithm cannot guarantee a globally optimal solution.

As mentioned in the paragraph on overfitting, a typical tree will have three input parameters: parent size, child size, and depth. Although the depth parameter rarely needs to be used if set at a high level initially, the parent and child size parameters are important to prevent overfitting, and the classification results of a tree can be highly dependent on them. Also, there are no currently recognized solutions for finding the best pair of parent/child size parameters, and there are not even any common heuristics for choosing them. Most researchers choose parent and child size parameters initially as some fraction (e.g., 10% and 5%, or square root of the number of cases for the parent and half of that for the child node) of the total data set size and then try variations close to that fraction and compare results of models built on different parameter sizes. This can be a time-intensive and inconclusive approach to classification.

Finally, trees can be sensitive to disparities in the size of label sets, with greater disparities resulting in ever-worse decision tree models. In particular, a decision tree will almost always favor the classification of data items as belonging to the largest label set. There are two common techniques to overcome this bias: oversampling and undersampling (He & Garcia, 2009). In oversampling, the less-dominant label set provides multiple copies of each element to the creation of the model, such that the size of the two label sets is equal for the model. Some oversampling techniques create new elements from the smaller label set; this should be attempted only when one is confident about the underlying distributions of one's data features. Undersampling chooses a random, smaller set from the dominant label set, such that both label sets provide an equal number of samples to the decision tree model. To avoid undersampling bias, it is recommended that one run multiple trials, with a new, random undersample conducted in each trial.

Despite these shortcomings, a decision tree can produce very accurate and robust results on many data sets. There are a number of refinements to the basic strategy that can be used to gain more improvement from decision trees. The first of these involves the use of three subsets of the original data, termed the training set, the testing set, and the evaluation set (Fig. 19.2). The training set is used to create an initial model, which is then used to classify the elements of the testing set. Based on some comparison of performance (typically the difference between accuracies on the training and testing sets), a new set of input parameters will be used to create a new model on the training set, which will be used to again classify the elements of the testing set, leading to a new evaluation of parameters. This will cycle until the desired goal (typically near-equal accuracies on the training and testing sets) is attained, at which point the model will be tested on the evaluation set, that is, used to classify the elements of the evaluation set. This will provide

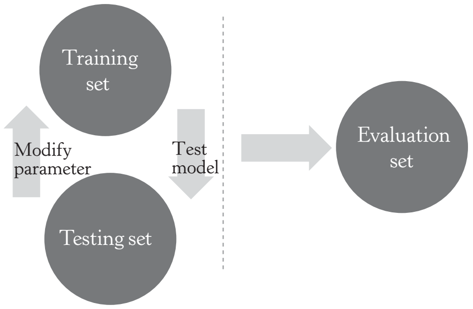

**FIGURE 19.2**: Block diagram of model creation.

the most robust predictor of the decision tree accuracy on unknown elements, as no elements of the evaluation test were used in the creation of the final model. The effectiveness of the predictor is suspect only if new elements come from a completely different data distribution than the original data. As mentioned earlier, in the creation of training, testing, and evaluation subsets, it is best to maintain a balance of labels in each set.

Although the method of testing, training, and evaluation can produce very reliable results, it can be difficult to implement if the size of the original data set is very small. In this case, a technique called *n*-fold cross-validation is typically used. In *n*-fold cross-validation, the original data set is broken into *n* distinct subsets of data. For any single fold, the remaining fraction of the data outside the fold is used as a training set, and the fold itself is used as a testing set. This is done for each fold, and results are typically reported as the average accuracy over all the folds. Where a final model is also presented, it is typically the model that performed the best on its training fold. Because there is no evaluation subset in cross-fold validation, it is recommended to use it carefully if one intends to tune parameters of the decision tree models.

The final variation on decision trees is becoming more common in machine learning in general, as improvements in technology allow ever-more complex models. The basic idea is to create an ensemble of classifiers (Dietterich, 2000), in which multiple different trees are created, with a final classification a result of the combination of the classification results from all members of the ensemble. Common ensemble techniques for decision trees are boosting and bagging, although both are beyond this chapter. Next, to illustrate the application of decision trees, we will present their usage in classifying patients with chronic fatigue syndrome (CFS) and myalgic encephalomyelitis (ME).

## CASE STUDY

Data mining could be used to help legitimize a group of individuals who have been stigmatized by labels and inappropriate case definition criteria. In our case study, we will focus on CFS and ME, whose scientific validity many health care professionals continue to doubt. The social construction of this disorder as a psychogenic illness of neurotic women, similar to earlier depictions of multiple sclerosis, has contributed to the negative attitudes that health care providers have toward those with this syndrome (Jason et al., 1997). This has had a serious negative impact on patients with this illness. For example, investigators have found that 95% of individuals seeking medical treatment for CFS reported feelings of being misunderstood because of the illness or the treatment (Green, Romei, & Natelson, 1999). Patients had been characterized as predominantly European American, middle-to-upper class women, and this perpetuated a myth that CFS was a “yuppie flu” disease, affecting middle-class and affluent people. Epidemiological research has shown that is a myth, as those with this illness are more likely to be minorities and of lower socioeconomic status (Jason et al., 1999).

For now, we will focus on how to identify who has and who does not have ME or CFS. Data mining can help with this important objective. Although this might appear to be a topic appropriate for a more traditional clinical domain, rather than one within the community field, this question has important public-policy implications because, if ambiguities occur in case definitions, investigators might select samples of patients who are different on fundamental aspects of this illness. Impediments to replicating findings across different laboratories would make it exceedingly difficult to estimate the prevalence of the illness, consistently identify biomarkers, or determine which treatments help patients.

The issue of diagnosis becomes important because many patients have been considered by their health care professionals to have a primarily affective disorder, which patients feel has stigmatized them, just as patients with cancer would feel undermined if health care professionals felt that they only had a psychogenic disease. Major depressive disorder is an example of a primary psychiatric disorder that has often been confused with CFS. Some patients with major depressive disorder also have chronic fatigue and CFS-like symptoms that can occur with depression (e.g., unrefreshing sleep, joint pain, muscle pain, impairment in concentration). Because fatigue and such symptoms are also defining criteria for CFS, some health care professionals and scientists have used an inadequate CFS case definition to conclude that ME and CFS are really psychiatric illnesses (Barsky & Borus, 1999). However, several ME and CFS symptoms, including prolonged fatigue after physical exertion, night sweats, sore throats, and swollen lymph nodes, are not commonly found in depression. In addition, although fatigue is the principal feature of CFS, fatigue does not assume equal prominence in depression (Friedberg & Jason, 1998). Moreover, illness onset with CFS is often sudden, occurring over a few hours or days, whereas primary depression generally shows a more gradual onset. In summary, CFS and major depressive disorder are two distinct illnesses, although they share a number of common symptoms. If one uses appropriate measures, it is possible to successfully differentiate these two disorders (Hawk, Jason, & Torres-Harding, 2006).

It is also important for case definitions to have high sensitivity and specificity, particularly for disorders with low prevalence rates such as CFS (about 4.2 in a thousand) (Jason et al., 1999). As an example, in a city of 1,000,000, with a true CFS rate of 4.2 per thousand, there would be 4,200 CFS cases. According to Bayes' theorem (Jaynes, 2003), if a case definition had a 95% rate of sensitivity, it would correctly identify 3,990 of these cases. However, if the case definition had 95% specificity, there would be more than 49,000 individuals who did not have CFS but were identified as having it. Clearly, being able to identify true negatives with precision is of high importance with low prevalence illnesses, such as CFS.

Criteria for the current CFS (Fukuda et al., 1994) case definition required a person to experience 6 or more months of chronic fatigue of a new or definite onset, but it used polythetic criteria, that is, a set of symptoms in which all do not need to be present to make a diagnosis. Because the Fukuda et al. (1994) criteria require only four symptoms out of a possible eight, critical CFS symptoms such as postexertional malaise and memory and concentration problems were not required for a patient to receive a diagnosis of CFS. This has increased the heterogeneity of the population, and, when similar biological findings have not emerged in different laboratories, it has been easy to jump to the conclusion that this illness is really psychologically determined.

In part as a reaction against the vague Fukuda et al. (1994) criteria, another consensus clinical case definition was developed, called the Canadian Clinical ME/CFS criteria (Carruthers et al., 2003). This ME/CFS case definition does specify core symptoms, including postexertional malaise; impairment of memory and concentration; unrefreshing sleep; arthralgia and/or myalgia; and several autonomic, neuroendocrine, and immune manifestations. However, the Canadian ME/CFS criteria require seven specific symptoms or domains, and requiring larger numbers of symptoms can inadvertently increase the rate of psychiatric comorbidity of the group that meets criteria. In addition, these criteria were based on consensus rather than empirical methods. Domains have the disadvantage of being less precise, as symptoms of both high and low prevalence could exist within a particular domain (Jason et al., 2014). At the present time, both the Institute of Medicine and the Office of Disease Prevention have committees focused on this issue of what case definition is, and there is considerable controversy among the scientific community regarding how to proceed. Patients have been clamoring for change and have rejected the commonly used Fukuda et al. (1994) CFS criteria, preferring the Canadian ME/CFS criteria (Carruthers et al., 2004). However, there continues to be scientific skepticism regarding this case definition, with respect to both the theoretical justification for their seven domains and the measurement of the domains.

Statistical selection techniques can be used to develop an empirical case definition, which would go beyond current consensus-based approaches. The problem for investigators is that there are many possible symptoms that might be included in such a case definition, but it is unclear which ones best distinguish between patients and healthy people, and, therefore, which symptoms are most characteristic of the illness. Methods to resolve this issue have important policy implications, as all science is built on the construction of case definitions, and, if they are not reliable and valid, then the diagnostic criteria might not successfully identify patients, which will hamper efforts to estimate prevalence, etiology, prevention, and treatment. Data mining techniques can help compare and contrast case definitions, as well as determine the types of symptoms that may be most useful in accurately diagnosing illnesses. In particular, data mining can uncover patterns in the data that would not be evident to human observers because of the size and complexity of the data.

In our case study, decision trees were used to analyze 54 common symptoms among patients with CFS, with all variables being placed into the analyses, rather than one item or domain or a limited group of items or domains. In this effort, decision trees helped determine which symptoms (and, implicitly, which questionnaire items) were most effective at accurately classifying participants as patients or controls.

For our case study, decision trees consist of a series of successive binary choices (branch points) that ideally result in an accurate classification of participants. At each branch point of the tree, all of the symptom variables are examined to determine which symptom has the greatest effect on the entropy of the classifications. Here, entropy indicates the certainty of the diagnosis. The symptom selected at each branch point is the one that best predicts classifications at that point in the tree; it is used to split all of the cases into two groups. This process is repeated, and more symptoms are chosen, until the resulting series of branch points produces groupings of correctly classified participants.

SPSS Statistics software was used to build our decision tree models. To construct the models, a Classification and Regression Tree (CART) algorithm was applied to a training set consisting of 66% of the cases, stratified to reflect the distribution of patient and control groups. The value of the model was measured by evaluating its classification performance when applied to cases reserved for testing (34% of the data), allowing this technique the ability to be generalized to new data.

Given the unbalanced distribution of the two classes (CFS versus non-CFS) and the fact that learning algorithms are biased toward the majority class, we conducted an experiment with similar numbers of participants in groups by taking a random undersample of 80 patients with CFS along with the 80 controls. We created 100 sets of randomly chosen patient data to analyze. For most analyses, only three to five variables (symptoms) were needed to classify participants. The analyses suggested the selection of four symptoms: fatigue or extreme tiredness, difficulty finding words to express thoughts, physically drained/sick after mild activity, and unrefreshed sleep (Jason et al., 2015).

The findings of this study suggest that core symptoms of this illness are fatigue, postexertional malaise, neurocognitive issues, and unrefreshing sleep. These results are theoretically compatible with other studies, such as Hawk et al.'s (2006) investigation, which found that these domains were able to successfully differentiate patients with CFS from major depressive disorder. Other symptoms, such as pain, autonomic, immune, and neuroendocrine symptoms, are less prevalent, but still important, and scores on these domains could also be specified as secondary areas of assessment. This data mining study suggests that empirical methods can be used to help determine which symptoms to include in the case definition.

## CONCLUSION

In this chapter, we reviewed data mining as a strategy to handle large amounts of data. In our case study, data mining methods were used to propose ways to develop a more empirical, rather than consensus-based, ME and CFS case definitions. The scientific enterprise depends on reliable, valid methods of classifying patients into diagnostic categories, and this critical research activity can enable investigators to better understand etiology, pathophysiology, and treatment approaches for ME and CFS, along with other disorders.

It is easy to become overwhelmed when confronting complex problems or power holders, such as in the case definitions of ME and CFS. However, by using advanced computational methods, and focusing on one small piece at a time, tangible change and success in the public-policy arena can be achieved. In part because of such research as that presented in the case study, the third author of this chapter was appointed the chairperson of the Research Subcommittee of the Chronic Fatigue Syndrome Advisory Committee, which makes recommendations regarding CFS to the US Secretary of Health and Human Resources. In this capacity, he was able to work on other policy-related issues, such as the inappropriate name given to this illness, an expanded case definition that the Centers for Disease Control (CDC) introduced, and leadership issues at the CDC regarding its program of CFS research. This policy work has taken more than 20 years, working with a number of coalitions involving patient organizations and scientific organizations.

Because of the third author's focus on sophisticated data-analytic methods with the case definition, he was invited to be a member of Health and Human Services' Department of Disease Prevention's Pathway to Prevention planning workshop that will focus on ME and CFS case definitions and has given an invited talk at the Institute of Medicine's commission to review the ME and CFS clinical case definitions. In each of these venues, the use of data mining strategies has been emphasized as one way to help investigators, patient organizations, and government bodies improve their decision making on complicated issues such as the ME and CFS case definitions.

In general, data mining provides a powerful tool to help both practitioners and researchers in uncovering patterns in the data that are not obvious to human observers and, consequently, cannot be analyzed using typical statistical analysis of hypothesis acceptance or rejection. In fact, data mining is opening up a new era of research, in which experimentation is data driven rather than hypothesis driven. Indeed, this new paradigm makes machine learning an ideal tool for community-based research for a number of reasons. First, unlike the exact sciences, community-based research rarely has easily discovered hypotheses, and the questions surrounding the interesting problems often cannott be represented simply using verifiable hypotheses. For instance, in our case, the question of "What symptoms are important for the definition and diagnosis of ME and CFS?" could be formulated as simply verifiable hypotheses, but we would have had to propose each possible subset of symptoms as the correct one and then use traditional statistical analysis to accept or reject each hypothesis. Given the existence of 54 symptoms in our survey instrument, this would have generated on the order of $10^{15}$ hypotheses to check. Instead, data mining provides a tool by which we can limit the number of possible hypotheses in a rigorous, empirical way. Second, where stigma or cultural avoidance issues enter into the research, data mining methods provide an objective method of investigation, in contrast to hypothesis-driven research, in which even the choice of hypothesis can have unfortunate social consequences. When the data determine your hypothesis, it is hard to argue that research bias exists. It is not the case that data-driven research is completely without bias, but it is harder to introduce bias when using automatic methods on source data. Third, despite the frequently intense algorithmic and analytical complexity of machine learning, faster and cheaper computers are becoming ever more prevalent, and one can confidently expect that data mining will be effectively available in mobile devices in the near future, either executed on one's phone or through quick and efficient cloud connections to powerful servers. For example, technological advances have allowed applying data mining to model public health on a population scale. Several studies have showed that, using large amounts of Twitter data, it is possible to track and predict influenza (Collier, Son, & Nguyen, 2011; Krieck, Dreesman, Otrusina, & Denecke, 2011) and also detect affective disorders such as depression (Golder & Macy, 2011).

Data-driven research is becoming increasingly more common. When the volume of data becomes so large that it is difficult for humans to discern patterns, then data-driven research can be effectively used to discover underlying issues in a relatively objective and empirical way. Note that it is not necessary to have an enormous sample in order to have a large volume of data; in community-based research in particular, it can be the case that the number of samples is relatively small, but the data on each sample are enormously rich. Although this can present a challenge to machine learning, the use of feature selection techniques, such as decision trees, can reduce the complexity of the sample data and allow for confident predictions on a small sample size. Furthermore, although much human data do distribute normally, much do not, and machine learning techniques, such as decision trees, that do not rely on assumptions about the distribution of the underlying data can effectively uncover patterns without normality.

Machine learning techniques, when used for classification, offer a number of other advantages that may be desirable in community-based research. Although classification typically predicts a categorical label, the underlying probability of prediction can be maintained, and probabilistic classification can be used. Thus, for example, rather that reading the output of a decision tree to say, "This patient has CFS," one can reference the underlying probabilities to suggest, "This patient has a 65% chance of having CFS." Such uncertainty can have positive impacts in human research, in which certainties may actually be detrimental to promoting cultural or policy change.

Furthermore, many machine learning techniques, and decision trees in particular, offer a variety of parameters that can be tuned for particular applications. Although such parameter tuning can contribute uncertainty to the final results, it does offer the possibility of leveraging the machine learning to focus on accuracy, specificity, or sensitivity. For example, in medical research, there is often a focus on specificity; the cost of missing a pathology in a diseased patient is much higher than the cost of misdiagnosing a healthy patient. Medical research will often sacrifice sensitivity for small increases in specificity. However, as we have seen in the case of CFS, and as is true in community-based research more generally, a focus on sensitivity might be more appropriate; allocating resources most efficiently or avoiding social stigma might argue in favor of not mislabeling pathology. The parameter tuning of machine learning allows us to generate models that focus on the best measure of effectiveness for a particular problem or situation.

In general, machine learning provides a powerful, flexible way of investigating data that allows researchers to uncover patterns that are not immediately obvious to human observers, in a way which preserves as much objectivity as possible and allows the data to directly determine results. Especially in the case of community-based research, in which standard methods from the physical sciences may not be directly applicable to the cultural environment of the richness of the data, machine learning can be a very effective method for discovery.

## REFERENCES

Barsky, A. J., & Borus, J. F. (1999) Functional somatic syndromes. *Annals of Internal Medicine, 130,* 910–921.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole.

Carruthers, B. M., Jain, A. K., De Meirleir, K. L., Peterson, D. L., Klimas, N. G., Lerner, A. M., . . . van de Sande, M. I. (2003). Myalgic Encephalomyelitis/chronic fatigue syndrome: Clinical working case definition, diagnostic and treatments protocols. *Journal of Chronic Fatigue Syndrome, 11,* 7–115.

Collier, N., Son, N. T., & Nguyen, N. M. (2011). OMG U got flu? Analysis of shared health messages for bio-surveillance. *Journal of Biomedical Semantics, 2*(Suppl 5), S9.

Davidson, I., Gilpin, S., & Walker, P. B. (2012). Behavioral event data and their analysis, *Data Mining Knowledge Discovery, 25,* 635–653.

Dhar, V. (2013). Data science and predictions. *Communications of the ACM, 58,* 64–73.

Dietterich, T. G. (2000). Ensemble methods in machine learning. J. Kittler & F. Roli (Ed.) *First international workshop on multiple classifier systems, lecture notes in computer science* (pp. 1–15). New York, NY: Springer Verlag.

Ferdowsi, Z., Settimi, R., & Raicu, D. S. (2010, July). *An application of clustering techniques to urban studies.* Paper presented at the 2010 International Conference on Data Mining, Las Vegas, NV.

Friedberg, F., & Jason, L. A. (1998). *Assessment and treatment of chronic fatigue syndrome.* Washington, DC: American Psychological Association.

Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., & Komaroff, A. (1994). The chronic fatigue syndrome: A comprehensive approach to its definition and study. *Annals of Internal Medicine, 121,* 953–959.

Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, G. Raetsch, & U. von Luxburg (Eds.), *Advanced lectures on machine learning* (pp. 77–112). New York, NY: Springer Verlag.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and day length across diverse cultures. *Science, 333,* 1878–1881.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Green, J., Romei, J., & Natelson, B. J. (1999). Stigma and chronic fatigue syndrome. *Journal of Chronic Fatigue Syndrome, 5,* 63–75.

Harris, R. (2008). The 16 TB RAM PC: When? *ZDNet.* Retreived June 2015, from http://www.zdnet.com/article/the-16-tb-ram-pc-when/

Hawk, C., Jason, L. A., & Torres-Harding, S. (2006). Differential diagnosis of chronic fatigue syndrome and major depressive disorder. *International Journal of Behavioral Medicine, 13,* 244–251.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21,* 1263–1284.

Jason, L.A., Kot, B., Sunnquist, M., Brown, A., Evans, M., Jantke, R., Williams, Y., Furst, J., & Vernon, S.D. (2015). Chronic fatigue Syndrome and myalgic encephalomyelitis: Toward an empirical case definition. *Health Psychology and Behavioral Medicine: An Open Access Journal*, *3*, 82–93..

Jason, L. A., Richman, J. A., Friedberg, F., Wagner, L., Taylor, R. R., & Jordan, K. M. (1997). Politics, science, and the emergence of a new disease: The case of chronic fatigue -syndrome. *American Psychologist*, *52*, 973–983.

Jason, L. A., Richman, J. A., Rademaker, A. W., Jordan, K. M., Plioplys, A. V., Taylor, R. R., & Plioplys, S. (1999). A community-based study of chronic fatigue syndrome. *Archives of Internal Medicine*, *159*, 2129–2137.

Jason, L. A., Sunnquist, M., Brown, A., Evans, M., Vernon, S. D., Furst, J., & Simonis,V. (2014). Examining case definition criteria for chronic fatigue syndrome and Myalgic Encephalomyelitis. *Fatigue: Biomedicine, Health, and Behavior*, *2*, 40–56.

Jaynes, E. T. T. (2003). *Probability theory: The logic of science*. New York, NY: Cambridge University Press.

Jiang S., Ferreira, J., Jr., & Gonzalez, M. C. (2012). Discovering urban spatial-temporal structure from human activity patterns. In *Proceedings of the Association for Computing Machinery SIGKDD International Workshop on Urban Computing* (pp. 95–102). New York, NY: Association for Computing Machinery

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

Krieck, M., Dreesman, J., Otrusina, L., & Denecke, K. (2011). A new age of public health: Identifying disease outbreaks by analyzing tweets. In *Proceedings of Health WebScience Workshop, ACM Web Science Conference*. Koblenz, Germany: Association of Computing Machinery

Kotsiantis, S. B. (2007). Supervised machine learning: A review of techniques. *Informatica*, *31*, 249–268

Kutz, J. N. (2013). *Data-driven modeling & scientific computation: Methods for complex systems and big data*. Oxford, England: Oxford University Press.

Wang, F., Tong, H., Yu, P., & Aggarwal, C. (2012). Guest editorial: Special issue on data mining technologies for computational social science. *Data Mining and Knowledge Discovery*, *25*, 415–419.