

*Data Mining*  
*Tools for Exploring Large Datasets*

*Robert Stine*

*Department of Statistics*

*Wharton School, University of Pennsylvania*

*[www-stat.wharton.upenn.edu/~stine](http://www-stat.wharton.upenn.edu/~stine)*

Modern data mining combines familiar and novel statistical methods to identify reproducible patterns in wide data sets. Wide data sets are distinguished by having many columns (or variables), often more columns than rows. The objective in this domain is prediction. If you can predict new data accurately or better than alternatives, then you've made a contribution. Rather than building a model that relates one or two experimental results to a response, data mining involves searching for patterns. Such searches may scan thousands of features, looking for the few that are predictive of the response. The search might be entirely automated or allow expert insight. Once a vile thing to be accused of, data mining has become respectable, useful, and necessary.

These lectures introduce data mining through a combination of lectures and examples. You'll see examples that look for patterns in voting behavior, patients at risk of a disease, prospective job candidates, and credit applications that reveal fraud. In each illustration, the goal is prediction. Rather than interpret a pattern found in one set of data, the objective is to predict new data. Interpretation is fun, but we'll exercise considerable restraint to avoid confusing association with causation. Even if you stick to simple models, concepts from data mining can help determine whether you've missed an important feature of your data.

Data mining does not require exotic hardware or software. Today's PC would have been a supercomputer in 1999. You can explore large datasets quite well with nothing more than regression and a laptop. That's basically what we'll do in the first week of class. Once you grasp the fundamentals, you'll appreciate the strengths and weaknesses of exotic methods. We'll start with regression, and then logistic regression, classification and regression trees, and a bit of neural networks and cluster analysis.

You need to do data mining to learn data mining. For this class, we'll use a combination of R and JMP from SAS. JMP handles very large data sets and includes an extensive collection of algorithms for building and assessing regression models with data mining tools such as trees and neural networks. The software is highly interactive and graphical. Two class sessions (Thursdays) are devoted to lab time so that you can try some of these tools yourself in a supervised lab.

Participants who attend these lectures are encouraged to work in a group on a project associated with an ICPSR data set. The work on that project will help you learn how the tools work. Possible topics for this project include data from studies related to elections, health outcomes, criminal sentencing, economics, and social activities.

## ***Guide to Planned Lectures***

### *Lecture 1. Introduction, Exploring Large Data Sets*

Good data analysis begins by looking at the data. That can be hard to do when the dataset has hundreds of columns, but it's not impossible. This lecture introduces data mining and explores several datasets using *interactive graphics*, a key strength of the JMP software. We'll borrow ideas from *multivariate analysis* (e.g., principal components, multidimensional scaling, and cluster analysis) to help find interesting views of the data. The objectives are to gain familiarity with the data, spot unusual patterns, recognize collinear variables, and form conjectures. Hypothesis *generation*, not just testing, is an important aspect of data mining. Multivariate analysis is also useful for creating new features to add as explanatory variables. Even though we may start with a wide data set, the columns in the data may not be the right ones to use in a data mining analysis. This lecture will also introduce some of the data sets used in the class, particularly the survey of voters in the 2008 US presidential election.

### *Lecture 2. Models for Prediction*

This lecture explores the strengths and weaknesses of regression analysis for building predictive models. Predictive models aim to predict or classify new observations of the response as accurately as possible. Regression analysis is the most commonly used methodology in statistics and it's the benchmark for data mining as well. Much of the success that you can have with a regression model comes from the fact that we have so many useful diagnostic graphics that reveal whether a regression performs well. We'll also consider a close relative of least squares regression, *logistic regression*, that's often used when the response variable is categorical. Logistic regression is often used to correct for *calibration* problems in a linear regression. In those cases, it is often better to calibrate the original regression model rather than switch methods.

A key challenge in data mining is finding the right set of explanatory variables to search. We may start from those that are part of the study results, but we usually need to expand this set to obtain a predictive, parsimonious model. The simplest additions are interactions, products of the original explanatory variables. Before we can exploit interactions in data mining, we need to be acquainted with what they are and what they do in a model. Interactions carry important interpretations, such as the identification of heterogeneous subsets and nonlinearities. Interactions also introduce potential anomalies such as sparse data and outliers that fool the uninformed.

### *Lecture 3. Data Mining with Regression*

Wide data sets challenge the thoughtful modeler. Even the best researcher with a clear theory is going to wonder what other variables might be predictive of the response. Since most collections of possible explanatory variables are highly correlated, it can be very hard (and risky) to impose a strong interpretation on the estimates. That's where methods that automatically *search for predictors* (explanatory variables that don't really

explain but do predict the response) become necessary. The best known, most disparaged, and yet perhaps most useful is the greedy search provided by *stepwise regression*. Even if you have your heart set on one of the new algorithms, you should always set a baseline for comparison with stepwise regression.

The introduction of automatic search techniques increases the risk of *over-fitting*, adding variables to a predictive model that are not in fact predictive. Over-fitting is very common; we've all seen applications in which a model fits the observed sample well but predicts new data poorly. This lecture considers how stepwise regression works and what can be done to effectively control its search. *Cross-validation* and the roughly equivalent criterion known as *AIC* are often recommended, but we'll come down on the side of methods related to the *Bonferroni* criterion. We'll also consider shrinkage methods like *ridge regression* (or the more modern *lasso*) that pull estimates toward zero to reduce the effects of outlying cases.

#### Lecture 4. Lab Session

Hands on time in the Michigan Lab.

#### Lecture 5. Using Regression More Effectively

Typical implementations of stepwise regression work well, but to be truly successful, we need to do a bit of customization. After you've seen how to fool stepwise regression with a few carefully chosen variables, you'll start to recognize what can happen in your data. Once you've seen the problems (biased variance estimates, leverage points, weird sampling distributions), you'll also recognize how easy it is to fix stepwise regression to avoid these problems. That's part of the appeal of building on a tool that we understand: you can understand when it's going to fail and fix the problem. Some patches to stepwise even make it run faster than ever. A very useful adjustment changes the way that we calculate the accuracy of slope estimators, dropping the fancy formula you've come to know for a simpler *sandwich* alternative.

#### Lecture 6. Streaming Feature Selection and Alpha Investing

At some point, you'll discover that you want features for data mining that the software you buy shrink-wrapped doesn't provide. You can either pay someone to write the code for you, write the software yourself (such as in R), or find someone who's already done it for you. We'll visit a web site in class that implements an enhanced version of stepwise regression. The software implements several of the enhancements described in prior lectures and adds a new wrinkle: *streaming feature selection with alpha investing*. This approach to searching for predictive features scales up to huge problems and allows the investigator to guide the search providing strategies for exploring the space of possible predictors while avoiding over-fitting.

### Lecture 7. Alternatives to Regression: Neural Networks, Classification and Regression Trees

Regression models produce an equation that “explains” how the model predicts new cases. Alternative methods such as *neural networks* and *regression trees* work differently. Neural networks, and the closely related projection pursuit regression, blend several regression models together using heuristics borrowed from engineering and biology. Regression trees drop the equations altogether and instead build predictions by binning data and averaging. Both alternatives to regression have their advantages, and we’ll use them as diagnostic tools for seeing if we’ve missed anything with our models.

### Lecture 8. Classification and Regression Trees

Regression trees (*a.k.a.*, CART with the closely related *classification trees*) are different enough from regression models that they deserve more time to appreciate. This lecture uses trees as alternatives to regression models, considering what they are very good at (*e.g.*, finding interactions and producing a set of rules that is often more appealing than the concept of an equation) and what they don’t do so well (*e.g.*, smooth patterns). Part of the appeal of using trees is the elegant implementation in JMP.

### Lecture 9. Lab Session

Hands on time in the Michigan Lab.

### Lecture 10. Further Topics, including a Glimpse of the Future and Pet Tricks

We’ll use this last class to cover topics that come up along the way and to handle the inevitable overflow from prior lectures. We’ll also discuss an ongoing project that applies regression to *text mining*: modeling textual data rather than numbers. Once you see how that’s done, you’ll see how to expand the scope of models to images and other types of novel data.

Trees are neat, but if you like them, you’ll want to learn how to grow *forests*. A random forest isn’t the latest method for drawing woods in an animated film, it’s instead a method for growing and combining many regression trees. Random forests are an example of *model averaging* in which you combine several different models to arrive at a prediction (other methods are called *boosting* and *bagging*). Ensemble approaches use, for example, *bootstrap resampling*, to create multiple data sets that can be used to fit alternative models. Averaging the predictions of these models usually produces a better prediction than any one model alone. The ideas apply widely beyond trees to virtually any predictive model.

As for the pet tricks, you’ll have to be there!

## References

If you would like some reading to accompany these lectures, or perhaps afterwards, here are a few papers, books, and web sites that you might find useful.

Berk, R A (2008) *Statistical Learning from a Regression Perspective*. Springer, New York.

Berk starts from the regression point of view and includes trees and ensemble methods like bagging and random forests. There's also a bit on smoothing via regression.

Berk, R A (2006) An introduction to ensemble methods for data analysis. *Sociological Methods & Research*, **34**, 263–295.

This paper describes ensemble methods like bagging using the example of random forests. There are examples with cross-validation as well.

Breiman, L (2001). Statistical modeling: the two cultures. *Statistical Science*, **16**, 199-215.

Statistics missed the boat, sticking to asymptotic estimates for small samples as the world of computing and large data bases exploded. Several good discussions accompany this article.

Breiman, L, J Friedman, R Olshen, and C J Stone (1984). *Classification and Regression Trees*. Wadsworth, Belmont CA.

This classic popularized the use of tree-based models and the use of cross-validation in picking good models. Still a good read.

Chatfield, C (1995). Model uncertainty, data mining, and statistical inference. *Journal of the Royal Statistical Society, Series A*, **158**, 419-466.

Over-fitting is a serious problem when the data suggest the model, often leading to wildly optimistic promises of prediction accuracy.

Foster, D P and R A Stine (2004). Variable selection in data mining: building a predictive model for bankruptcy. *Journal of the American Statistical Association*, **99**, 303-313.

Using 3,000,000 months of credit card activity and 67,000 possible features, we show how to use least squares regression to build a model that predicts as well (or better) than the computational learning algorithm known as C4.5. The algorithm is now implemented in the SAS enterprise miner software package.

D P Foster and R A Stine (2008).  $\alpha$ -investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society B*, **70**, 429–444.

This paper shows how to use a theory to guide the allocation of the chance for false positives over a sequence of tests, as in the case of stepwise regression.

Friedman, J (2001). The role of statistics in the data revolution. *International Statistics Review*, **69**, 5-10.

Jerry Friedman is one of the most creative modelers in statistics. Here's his take on how statistics can play a more central role in a world of large data sets.

Friedman, J and B Silverman (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, **31**, 3-a lot.

This paper shows that nonparametric regression can be viewed as a special case of stepwise regression. It's all in the choice of the  $x$  variables that the search considers.

Hand, D J, H Mannila, and P Smyth (2001). *Principles of Data Mining*. MIT Press, Cambridge MA.

There are a lot of books on data mining, but most talk more about assembling the data and are written for computer science. Assembling the data is a hard part of the problem, and if you get it wrong, it does not matter how you do the modeling. This book describes some of those issues, but goes on to summarize the statistical methods as well. A nice high-level view of models and patterns in general.

Hand, D J, G Blunt, MG Kelly, and N M Adams (2000). Data mining for fun and profit. *Statistical Science*, **15**, 111-131.

If my predictions are more accurate than yours, I profit and you lose. It's no wonder that data mining has become important in the business world. Be they pharmaceuticals like Merck and Pfizer or web sellers like Amazon, a large chunk of the value of many firms lies in their proprietary data.

Hastie, T, R Tibshirani, and J Friedman (2001). *The Elements of Statistical Learning*. Springer, New York.

Much of the initial development of methods for handling large data sets began outside of statistics, in an area known as computational learning or, more boldly, knowledge discovery. Computer science was quick to see the importance of getting value from large databases, and forged ahead. They certainly came up with better names (e.g., neural network versus projection pursuit regression).

Linoff, G S, and M J A Berry (2011). *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management* (3<sup>rd</sup> Edition). Wiley, Indianapolis.

Emphasizes role in business decision making, with 100+ pages introducing the role of data mining. Wide scope of methods with discussion of presentation, communication of results, and privacy issues.

MacKay, DJC (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge.

This text introduces many of the principles that guide modern research on modeling within computer science, nicely called "machine learning." The text introduces the foundations by developing the connection between statistics and information theory. These connections provide heuristics for the computing algorithms.

Stine, R A (2004). Model selection using information theory and the MDL principle. *Sociological Methods & Research*, **33**, 230-260.

You have to think about the scope of the search when you build a model by finding the best fit possible using anything from a wide database. Information theory (ideas that show how to fit more data on your computer disk or make cellular telephones work) turns out to offer a very useful paradigm for judging models.

Tan, P N, M Steinback, and V Kumar (2006). *Introduction to Data Mining*. Pearson, Boston.

This is a wide-reaching, medium-level math introduction. The presentation is not nearly so technical as Hastie. It includes many of the ideas that are now used in computer science for modeling, such as the so-called kernel trick and support vector machines. (It has little discussion of regression or logistic regression.)

Torgo, L (2011). *Data Mining with R*. CRC Press, Boca Raton.

Introduces data mining and R five case studies, but some of the data is really small (like  $n = 7$ ). Examples include nature, stock trading, fraud detection, and microarrays. The methods include the usual ones (like regression), but reach out to support vector machines (SVM) and multivariate adaptive regression splines (MARS).

Witten, I H, E Frank, and M A Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3<sup>rd</sup> Edition). Morgan Kaufman, Amsterdam.

The emphasis here is on trees and association rules. The coverage is terse, but comprehensive with much of the terminology mentioned. Technical material is included, but confined to boxes. The examples are closely tied to the WEKA software tools. Examples use small data sets.

[www.kdnuggets.com](http://www.kdnuggets.com)

This web site links to software, data and conferences, including the data sets used for its annual competitions to see what sort of data mining software can predict a hold-back sample the best.