

Data Mining FAQ*

Robert L Grossman[†]
Open Data Group

August 28, 2011

Question: What is data mining?

Data mining is the semi-automatic discovery of patterns, changes, associations, anomalies, and other statistically significant structures in large data sets in order to gain knowledge and to support processes that improve decision making.

Question: What is analytics?

Over the last few years, within business, the term analytics has been used for the process that turns data into knowledge and actions that improves decision making.

Question: Hmm ... I thought that was the goal of business intelligence?

That's correct. It is confusing. In the middle 1990's and continuing for several years, the term *business intelligence* was used in this way. Today, though, the term business intelligence usually refers to the more narrow tasks of producing reports from data and performing ad-hoc queries on data that is stored in data warehouses and datamarts in order to gain insights about data and to improve decision making.

Prior to the middle 1990's, building statistical models was the domain of statistics. As the amount of data grew and it become a challenge for conventional statistical applications to manage the data and to build statistical models from it, the disciplines of data mining, and then business intelligence, and then predictive modeling emerged.

Recently a new term is starting to be used – *big data analytics*. Although the names have changed, and the software tools and algorithms have slowly improved over time, by and large, best practices in the field have remained relatively stable.

Some of these best practices include: gathering all relevant data, understanding the data, building models using the data, and then deploying and integrating these models to make decisions. As this is done, it is always essential to check and to double check that the various regulations governing how we collect data and how we use data once it has been collected have been followed.

For example, when building models that are used in Internet applications, it is important to check that the data that is used has been collected in a way that is compliant and consistent with the privacy policies posted on the web sites from which the data has been collected.

Question: Why is data mining important?

There is more and more digital data being collected, processed, managed and archived every day, yet the number of people available to analyze the data has remained approximately the same. This gap — between the amount of data and the number of people trained to analyze it — has been growing each year. The goal of data mining and related technologies is to automate in part the analysis of this data. Data mining algorithms, software tools, and systems are critical to a wide variety of problems in business, science, health care, engineering and national defense.

*Copyright Robert L. Grossman, 1999-2011.

[†]Robert Grossman is also a faculty member at the University of Chicago.

Question: What are some commercial success stories in data mining?

Data mining has been applied successfully in a number of different fields, including:

- selecting which online ads to display by Google;
- product recommendations, such as when Amazon suggests books that might interest you and when Netflix suggests movies that might interest you;
- detecting credit card fraud by FICO;
- pricing non-standard insurance by Progressive;
- identifying suitable prospects, setting the amount of credit offered, and determining the associated interest rate for offers of pre-approved credit cards by Capital One, American Express and other credit card issuers.

Question: What are the historical roots of data mining?

From a business perspective, data mining's roots are in direct marketing and financial services, which have used statistical modeling for at least the past two decades. From a technical perspective, data mining emerged as a separate discipline approximately twenty years from various fields, including a) statistics, b) machine learning, c) databases, and d) high performance computing. One of the main conferences in data mining – The International Conference on Knowledge Discovery and Data Mining (KDD) – started in 1994. One of the main conferences on machine learning – the International Conference on Machine Learning (ICML) – started in 1980.

Question: Has data mining and big data had an impact on science?

The answer is definitely yes, but the impact of data mining and big data on science is outside the scope of this FAQ.

Moore's law increases the capability and drives down the cost of the integrated circuits that power not only computers and digital devices for customers but also the sensors that, collect scientific data. This has resulted in an exponentially increasing amount of scientific data being produced each year.

The impact of big data has been well understood within the high energy physics community for over twenty years, within the astronomy community for at least a decade, and more recently is beginning to be recognized within the biological sciences community. As an example within biology, the cost to sequence genomes is decreasing faster and the quantity of data being produced is growing faster than would be predicted by Moore's law.

The SIAM International Conference on Data Mining started in 2001 and is one of the conferences that focuses on the applications of data mining to science. (Disclosure: I was one of the founders of the conference.)

Question: What is a model? What are some of the different categories of models used in data mining?

In this FAQ, by a model, we mean a statistical or data mining model. Loosely speaking a model is the result of applying an algorithm to data.

There are several different categories of models in data mining, including:

- Predictive models. These types of models predict how likely an event is. Usually, the higher a score, the more likely the event is. For example, how likely a credit card transaction is to be fraudulent, how likely a visitor to a web site is to click on an ad, or how likely a company is to go bankrupt.
- Summary models. These models summarize data. For example, a cluster model can be used to divide credit card transactions or airline passengers into different groups depending upon their characteristics.

- Optimization models. Although optimization is a field in its own, it is worth mentioning optimization models. The goal with these types of models is to make choices so as to maximize (or minimize) a fixed function (usually called the objective function). For example, an optimization model might assign packages to trucks so as to minimize the time required to deliver all the packages.

Question: Are there other categories of models?

There are several other categories of models, and two of them are probably worth discussing briefly: graph models and association models. In both cases, these models can be used to summarize data and to make predictions, so they also fit into the taxonomy above.

- Graph models. These types of models uncover certain structures in data represented by nodes and links. As an example, in a network model describing Facebook friends, nodes might be individuals and directed edges with weights might represent the likelihood that one friend will contact another friend in the next 72 hours. As another example, a credit card fraud ring may surreptitiously collect credit card numbers at a pawn shop and then use them for online computer purchases. Here the nodes are credit card accounts and merchants and the links are credit card transactions.
- Association models. Sometimes certain events occur frequently together. For example, purchases of certain items, such as beer and pretzels, or a sequence of events associated with the failure of a component in a device. Association models are used to find and to characterize these co-occurrences.

Question: How does data mining work?

When data mining is used to make *predictions* about future events, there are two main steps. First, some historical data is analyzed and used to build a *model*. As just mentioned, by a model, we mean a statistical or data mining model. Some of the specific different types of models that are used to make predictions are described below. Second, this model is then applied to new data to make predictions. Often the model is embedded in an operational system. As an example, with online advertising, a response model is used to predict whether an individual will click on an ad that is displayed.

When data mining is used to *summarize* data, one or more statistical or data mining algorithms are applied to the data resulting in a model. A common type of summary is a cluster which groups together similar data records. The basic idea is that data records within a cluster are all closer to each other than they are to data records belonging to other clusters. For example, in direct marketing it is common to cluster consumer behavior in this way. A dataset containing data records describing millions of individuals might be clustered into a few dozen different clusters. For example, 18-25 urban males might be in one cluster, while 25-40 year old suburban families with children might be in another cluster.

You can think of data mining as the process of applying a statistical or data mining algorithm to a dataset to produce a model. That is the dataset is the *input* and the model is the *output*.

Question: I hear the phrase “empirically derived and statistically valid” applied to models. What does that mean?

Decisions based upon models derived from data are usually expected to be empirically derived and statistically sound. That is, first, they must be derived from the data itself, and not the biases of the person building the model. Second, they must be based upon generally acceptable statistical procedures. For example, the arbitrary exclusion of data can result in models that are biased in some fashion.

Also, when building models that are statistically valid, a process is used to evaluate the performance of a model. When this is done, one model (sometimes called the Champion Model) can be compared to the new model (sometimes called the Challenger Model) and the better performing model can be selected.

Question: What are the major steps in data mining?

1. **Define the objective.** The first step and one of the most important is to identify a specific problem that would benefit from a statistical or data mining model and to make sure that if a proper solution were deployed that something useful would result.
2. **Identify and obtain all the required data.** The next step is to identify the different datasets and data streams necessary in order to build the model and to deploy the model. Then this data must be obtained. Even though this step sounds simple, it is often the case that it can be challenging to obtain the required data for modeling since the modeling group and the group that is responsible for the data are often in different parts of the organization.
3. **Clean and explore data.** A very important step is to clean and explore the data in order to prepare the data for data mining and statistical modeling. The main goal with this step is *understand the data*. This is usually the most challenging step.
4. **Build a datamart.** The next step is to put in place an infrastructure to manage the data. In practice, except for small datasets, it is usually necessary to put the data into a database or datamart so that it can be more easily managed during the modeling process. For very large datasets that do not fit into a database or datamart, specialized distributed file systems, such as Hadoop, are used.
5. **Derive features.** It is rare for a model to be built using only the attributes present in the cleaned data; rather, additional attributes sometimes called *features* or *derived attributes* are usually defined. As a single example, a stock on the S&P 500 has a price and an earnings associated with it, but the ratio of the price divided by the earnings is more important for many applications than either single attribute considered by itself.
6. **Build the model.** Once the data is prepared and datamart is created, one or more statistical or data mining models are built. After the models are built, they are validated. To validate a model, a statistical process is used to measure the accuracy and effectiveness of the model.
7. **Post-process the output of the model.** It is common to normalize the outputs of data mining models and to apply business rules to the inputs and the outputs of the models. This is to ensure that the scores and other outputs of the models are consistent with the over all business processes the models are supporting.
8. **Deploy the model.** Once a statistical or data mining model has been produced by the steps above, the next step is to deploy the model in operational systems. Almost always, various reports are created to measure the effectiveness of the models.
9. **Maintain the model.** On a periodic basis, say monthly or yearly, a new model is built and compared to the existing model. If required, the old model is replaced by the new model.

Question: This is too complicated. Isn't there a simpler way to describe this?

Yes, think of data mining as consisting of three basic steps, that you can remember with the mnemonic DDM, which stands for Define, Deploy and Model.

1. **Define.** Identify a problem that would benefit from a statistical or data mining solution and that if effectively deployed would further an organization's strategic objectives. That is choose a modeling opportunity that is a good strategic fit for your company or organization. Next make sure that there is not another model that would bring greater value to the organization.

2. **Deploy.** Forget about the model initially and instead assume that you have built a great model and figure out how the scores produced by the model can lead to actions that can be integrated into the required business processes and either increase revenue, decrease risk, or optimize a business process. Finally, make sure that you can get all the data required to build the model.
3. **Model.** Finally, gather the data, explore the data, identify useful features, build a model and validate it. If you have done the previous steps correctly, you should be able to deploy the model and the model should bring value.

Question: What are the differences between predictive models and rules?

Predictive models use historical data to predict future events, for example the likelihood that a credit card transaction is fraudulent or that a visitor to a web site will click on an ad.

Rules are quite different. It is useful to distinguish between three types of rules:

1. business rules
2. subjective rules
3. data driven rules

Business rules ensure that business processes follow agreed upon procedures. For example, business rules may dictate that a predictive model can use only the first three digits of a zip code not all five digits.

Often organizations do not have on hand technical staff that are comfortable building statistical and data mining models. In this case, they may use *subjective rules*. Sometimes these are called heuristics or golden gut rules. As an example, a heuristic may be to offer an ad related to automobiles if the visitor to the web site has been referred to from a site that has content related to automobiles. Sometimes these types of heuristics can be quite effective.

There are two critical differences to keep in mind though when using subjective rules instead of models. First, models are *validated using data* in the sense that there is a process that measure the effectiveness of a model and steadily improves it. Subjective rules may or may not be validated, but in most cases are not validated. Second, models are derived from data using a statistical or data mining algorithm. In contrast, subjective rules are developed using *common sense* or business experience.

Typically, business rules are used to preprocess the data to make sure that the data being used in a model complies with all the requirements and processes. Business rules are also typically used to post-process the results of a model. For example, if a model is used to detect a data quality problem, then a business rule might be used to silence an alert from a model unless the estimated monetary value associated with the data quality problem is above a threshold.

To summarize, business rules ensure that business processes are being followed and predictive models ensure that historical data is being used most effectively. A good best practice is to use both business rules and statistical models, but to minimize the use of subjective rules.

Question: When I talk to the IT Organization they keep mentioning architecture. How does this related to data mining?

Loosely speaking, an *architecture* for an IT system describes the various components of the system, the inputs and outputs of each component, and how the system interfaces to the other systems that it interacts with. There are standards for many IT systems and an architecture specifies which standards are used.

Architectures for analytic systems are important since one business unit usually builds the models, while another business unit usually deploys the models (most often the IT group). Modelers generally use specialized software (such as R, SAS or SPSS), while the IT group uses databases and languages such as Java, Python and C++. Because of this, it can be challenging to deploy models within an organization.

Over the past decade, a standards-based analytic architecture has emerged as a best practice for building and deploying analytic modelers. (Disclosure: I have been one of the many people involved in the development of this architecture and of standards associated with it.)

This architecture is based upon using a standard format for statistical and data mining models called the Predictive Model Markup Language or PMML. PMML allows applications written in different languages and running on different systems to easily import and export models.

With an analytic architecture based upon PMML, one application called the *model producer* can build the model, while another application, called the *model consumer* can deploy the model in operational systems.

For example, R, SAS, SPSS, and many other statistical and data mining applications can export model in PMML. There are several so called scoring engines available that can consume PMML models and that can be easily integrated with operational systems.

An important advantage of this approach is that predictive models can be easily update since all the consumer needs to do is read a new PMML file. In practice, if models are hard coded in Java or C++ it can be labor intensive to update models. Since good accuracy require fresh models on fresh data, using PMML can improve the accuracy in practice when predictive models are deployed.

Question: What determines the accuracy of predictive models?

The accuracy of a predictive model is influenced most strongly by the quality of the data and the freshness of the model. Without good data, it is simply wishful thinking to expect a good model. Without updating the model frequently, the model's performance will generally decay over time.

Accuracy is measured in two basic ways. Models have false positive rates and false negative rates. For example, consider a model predicting credit card fraud. A false positive means that the model predicted fraud when no fraud was present. A false negative means that the model predicted that the transaction was ok when in fact it was fraudulent. In practice, false positive and false negative rates can be relatively high. The role of a good model is to improve a business process by a significant degree not to make flawless predictions. Only journalists and pundits make flawless predictions.

Question: What are the major types of predictive models?

Although there are quite a large number of different types of predictive models, the majority of applications use one of the following types of models.

1. Linear models. For many years, especially before the advent of personal computers, these were the most common types of models due to their simplicity. They divide data into two different cells using a line in two dimensions and a plane in higher dimensions. Quadratic models are similar but use a curve instead of a line to divide the data.
2. Logistic models. Logistic models are used when the predicted variable is zero or one, for example predicting that a credit card transaction is fraudulent or not. Logistic models assume that one of the internal components of the model is linear. Computing the weights that characterize a logistic model is difficult by hand, but simple with a computer.
3. Trees. Trees are a type of nonlinear model that uses a series of lines or planes to divide the data into different cells. Trees consist of a sequence of if ... then rules. Because of this, it is easier to interpret trees than other types of nonlinear models such as neural networks.
4. Neural Networks. Neural networks are a type of nonlinear model broadly motivated ("inspired by" is the phrase Hollywood uses) by neurons in brains.
5. Support Vector Machines. Support vector machines use what are called kernel functions to separate data into two classes. Using kernel functions, a nonlinear classifier can be found by computing a hyperplane in a higher dimensional linear space that separates the two classes. The higher dimensional linear space is a transformation of the original space.

6. Hybrid Models. It is common to combine one or more of the models above to produce a more powerful model.

Question: What is the difference between a linear and nonlinear model?

Models can be thought of as a function that takes inputs, performs a computation, and produces an output. The output is often a score, say from 1 to 1000, or a label, such such as high, medium, or low. A very simple type of model, called a linear model, uses the n input features to split the space of features into two parts. This is done using an $(n-1)$ -dimensional plane. For example, 2 features can be separated with a line, 3 features with a plane, etc. Most data is not so simple. Any model that is not linear is called a nonlinear model. Logistic models, tree based models and neural networks are common examples of nonlinear models.

Question: Which type of model is the accurate?

That is a common question, but there is no one best model. Different data requires different types of models. The accuracy of a model depends more on the quality of the data, how well it is prepared, and how fresh the model is than on the type of model used. On the other hand, there are some important differences between different types of models. Nonlinear models are generally more accurate than linear models. Linear models were more common in the past because they were easier to compute. Today this is no longer relevant given the proliferation of computers and good quality statistical and data mining software. Neural networks were very popular in the 80's and early 90's because they were quite successful for several different types of applications and because they had a cool name. Today, a variety of other methods are also commonly used, including tree-based methods and support vector machines. For example, tree-based methods are generally considered easier to build, easier to interpret, and more scalable than neural networks.