# What is Data Mining, and How is it Useful for Power Plant Optimization? (and How is it Different from DOE, CFD, Statistical Modeling)

StatSoft White Paper, July 2007


*Abstract.* Data mining methodologies have been widely adopted in various business domains, such as database marketing, credit scoring, fraud detection, to name only a few of the areas where data mining has become an indispensable tool for business success. Increasingly data mining methods are also being applied to industrial process optimization and control. While the general approach is similar regardless of application (finding "nuggets" of new information in data), some specific methodologies and techniques for optimizing continuous processes, such as boiler performance in a coal-burning power plant, have proven particularly useful for those applications, and superior to existing traditional analytic approaches such as DOE (design of experiments), CFD (computational fluid dynamics), or statistical modeling. This paper will provide an introduction to data mining, and specifically contrast the methods used in data mining with traditional optimization techniques.

## Table of Contents

## *Introduction to Data Mining, and Comparison to other Methods*

Data mining, "knowledge discovery", or "machine learning" methods have many origins, drawing on insights from research on learning as it naturally occurs in humans (cognitive science), advances in computer science and algorithm design on how to best detect automatically patterns in "unstructured" data, engineering and advances in machine learning (e.g., neural networks), to name a few. While traditional statistical methods for analyzing data, based on statistical theories and models, are now widely accepted throughout various industries, data mining methods have only been widely embraced in business for a decade or two. However, their effectiveness for modeling and optimizing and improving "difficult" processes are making these techniques increasingly popular – and even necessary – in many real-world process application.

### What is Data Mining

Suppose you wanted to optimize a cyclone furnace (an older-type design for burning coal, still in use in many power plants) for stable high flame temperatures. Stable temperatures are necessary to ensure cleaner combustion, and less build-up of undesirable slag that may interfere with heat transfer. Typically, most power plants are equipped with very effective data gathering and storage technologies, so there are easy ways to extract the data that describe various parameter settings, as well as flame temperatures, on a minute-by-minute interval.

Traditional methods to approach this task – to optimize combustion to achieve stable flame temperatures in the presence of different loads, fuel quality, and so on – come down to the application of a-priori (CFD) models, or more or less trial-and-error parametric testing.

### CFD (Computational Fluid Dynamics) modeling

One approach is to use explicit theoretical (first principles) models, to understand (based on these usually complex and highly nonlinear models) how best to set certain parameters, distribute airflows, etc. to optimize performance. With an explicit theoretical knowledge (model) of how exactly various parameters affect flame temperatures, one can use standard computer optimization algorithms to identify optima, which "in the laboratory" can be expected to optimize for stable flame temperatures.

Typically, these methods are used to identify the parameter "boundaries" where to keep certain input parameters (controlled by operators, or closed loop control systems) to ensure stable operations. However, in practice, there are numerous obstacles that put limitations on the applicability, effectiveness, and usefulness of CFD methods to optimize furnace performance "in vivo", i.e., inside a "real" power plant.

First, theoretical, a-priori, physical models of furnaces will only model parameters that are known (consistent with models) to have an influence. If in a particular installation, there are other specific "noise factors" that effect performance, CFD will not "know about this", nor can CFD models accommodate various esoteric installation details in a real power plant.

Second, CFD models can be very complex, and indeed become practically impossible to optimize because of their complexity.

So what is often needed is a "simplification" of sorts, or a "proxy-model" ("stand-in") that can summarize how the parameter inputs such as over fired air (OFA) distribution, primary and secondary air flows, coal-flow, and so on will affect flame temperatures, and the variability in flame temperatures. Data mining methods can provide such "proxy models", as will be further explained later.

## DOE, parametric testing

Once the a-priori, theoretical CFD models exceeds its bounds/usefulness, and cannot improve further the performance of an actual (installed in a power plant) furnace, typically, the next approach that is applied is design-of-experiment (DOE) and parametric testing. In short, this is trial-and-error testing of the real furnace, moving one (or two, three) parameters at a time (e.g., changing OFA, stoichiometric ratio, etc.), and following an experimental "design" to ensure that the maximum amount of information about how these parameters affect the response (flame temperatures) can be extracted.

DOE methods have been used and refined for decades. In short, specific "configurations", "runs", or test plans of specific parameter settings are tried, and the results for various important response variables (e.g., flame temperatures) are recorded. Then linear models are fit to the data, and using those linear models, optimization is rather straight forward.

For example, consider a quadratic model of the form:

$$Flame\text{-}Temperature = 1950 + .3*P1 - .2*P2 + .1*P1^2 - .3*P2^2$$

where *P1* and *P2* are the parameter settings for two parameters (e.g., *P1*=stoichiometric ratio, *P2*=coal flow). This is fundamentally a simple linear model, fitted using multiple regression (least squares estimation).

There are a number of shortcomings using DOE methods. First, the models, similar to the one shown above, are very simple, and usually too simple. In practice, the relationships between the parameters and measurements in a dynamic system like coal-burning furnaces cannot be adequately summarized with such models.

Second, and perhaps more importantly, the specific DOE design that is chosen to conduct the testing will limit, and in some ways "pre-ordain" the results. For example, if an experimental design is used that specifically allows one to extract linear and quadratic effects (as shown above), when the "real" relationships between parameters and flame temperatures are cubic (involving $P^3$), then the results one gets are simply wrong and useless! Moreover, the data collected during the parametric tests usually will not permit the analyst to verify that the results are indeed wrong.

## Data Mining: Combining the "best of two worlds"

To summarize, in CFD modeling, a-priori theoretical models (e.g., about the parameters that control flame temperature) are used and optimized "in the lab" to determine the best settings for the real furnace, installed in a real power plant. The problem is that the "real world" often differs from what was envisioned and modeled in the "lab". DOE and parametric testing is trial-and-error testing of the real equipment, following specific experimental designs and plans. The problem with this approach is that the experimental designs and plans will limit, and "pre-ordain" the results (one only finds the patterns that are expected, and that *can* be found given the test plan). Data mining overcomes the limitations of both approaches.

**"Listening to the furnace."** In many ways, the furnace (or other equipment) is "talking" to the operators through the language of numbers, recorded into the production database. The task is to translate the data streams into actionable information that can be used to optimize the performance of the equipment.

Data mining starts with the real data, collected from the real equipment (furnace). In fact, the more data the better, so if hundreds of parameters are recorded and available for analysis, that is preferable to just looking at 5 or 10 parameters at a time. Data mining can consider and use *all* the data you are collecting already.

For example, we may extract 3 or 4 month of data describing actual operation of the furnace, as well as the flame temperatures and their "natural" up-and-down fluctuations. At first, this amount of data – literally thousands of points for hundreds of parameters – can be overwhelming. However, data mining methods are well equipped to handle large amounts of data, and to detect the useful patterns in those data that allow us to improve furnace performance.

**Finding patterns.** As mentioned earlier, data mining methodologies and algorithms have their origins in many different disciplines. For example, researchers on artificial intelligence (AI) have proposed various methods and techniques that can efficiently "mimic" how real people ("experts") can detect difficult hidden patterns in large amounts of complex data. As a practical matter, the data are submitted to these (data mining) algorithms, to extract two important pieces of information: Which parameters are important, for determining flame temperatures? How exactly do the important parameters effect flame temperatures?

The first question, which parameters are important, is answered by applying so-called feature selection and feature extraction methods. These methods test whether or not parameters, or combinations of parameters show any kind of systematic relationship to flame temperatures.

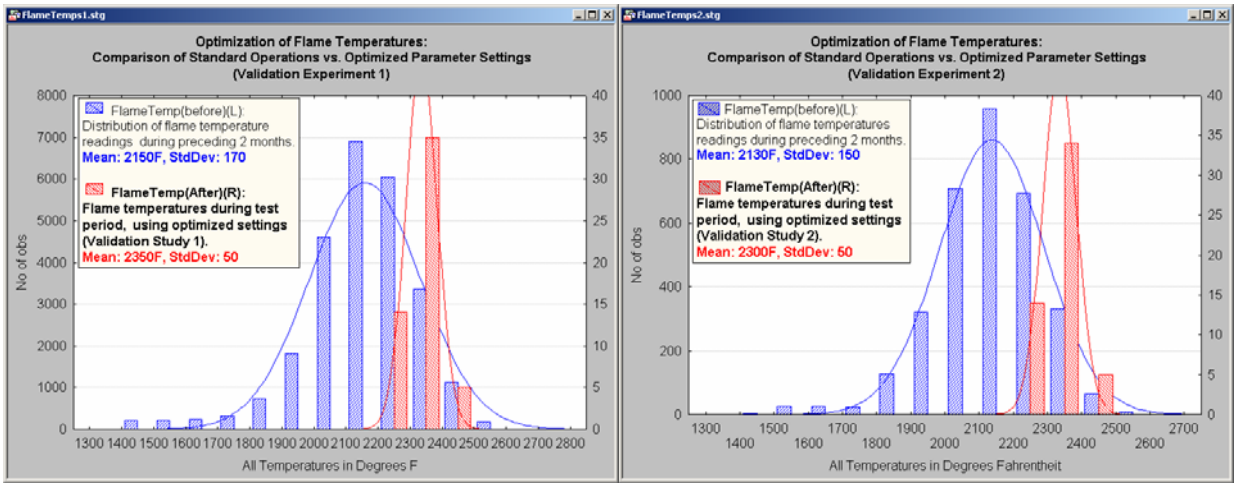
Once important parameters or combinations of parameters have been identified, then various algorithms can be applied to determine *how* exactly the important parameter affect flame temperatures.

**Optimization and deployment.** After information about the important parameters and the "patterns" how they affect flame temperatures has been extracted, the "deployment" or application of these models as well as optimization is straightforward. Essentially, the same techniques used to optimize CFD models are applied to optimize these "proxy-models", to

achieve, for example, stable and high flame temperatures. As a result, one would get specific values or ranges for a subset of the possibly hundreds of parameters that are recorded, which if when set as recommended, will yield consistent, stable, and desirable flame temperatures.

## How Effective is Data Mining?

Shown below is an example of a real data mining application to a furnace, with the goal to achieve stable flame temperatures above 2,100F. After going through the steps described above – extracting important parameters, and building and then optimizing data mining models, careful validation experiments were performed to verify that the recommendations from the data mining models indeed improve flame temperatures:




In short, there is little doubt that data mining methods were useful and successful in this application, to stabilize the flame temperatures. By making relatively small adjustments to a subset of specific parameters that operators routinely manipulated and adjusted to control the furnace. Specifically, minor adjustments were made to parameters such as coal flow stoichiometric ratios, and primary/secondary/tertiary airflows, to achieve much more robust operations and consistently higher flame temperatures.

To continue the metaphor of the furnace, that is "talking" to the operators through the data, data mining techniques were successful in "interpreting" this language, and using the information to identify the "sweet spot", where the furnace can provide consistent and robust performance.

## *Data Mining Methods and Algorithms*

So far the discussion has been around data mining "techniques" as a "black-box" approach. Next, we will provide a brief overview of typical algorithms. Detailed overviews of techniques and best practices can also be found here: http://www.statsoft.com/textbook/stathome.html.

## Feature Selection and Extraction Methods

These methods are used to identify among large numbers of parameters a subset of those that are appear to be useful for building prediction or proxy models, for relating parameters to important outcomes (e.g., flame temperatures). A particularly effective approach uses so-called recursive-partitioning methods, and multiple random starts, to identify which combinations of variables can be used to divide the (training) data so that the values of the outcome variable are most different between divisions.

Imagine an algorithm that starts by looking at all parameters one by one, and tries to split the data based on the respective parameter into one set where the flame temperature tends to be high, and another where it tends to be low. Next another variable is chosen to further split the subsets until we are left with several small samples or "nodes" that are as different from each other (with respect to flame temperature) as possible. This algorithm is applied over and over, using various random starts; during this process, the program keeps track of the parameters that appear to provide "effective splits" most of the time. Those are the important parameters.

**Finding interactions, combinations.** The most important aspect of this algorithm is that it is capable of identifying *interactions* between parameters, i.e., it can identify parameters that are only important if some other parameter is set within a specific range, while each parameter when considered by itself (one-by-one) has no obvious effect. In practice, all the "simple" patterns and effects of how parameters such as OFA settings affect flame temperatures are already known by experienced engineers and operators; however, even the most experienced engineers will not be able to sift through the thousands, or even millions of possible combinations of parameter settings to identify specific combinations where particular parameters "suddenly" (and usually unexpectedly) become important, and effective. Feature extraction alone often provides tremendously useful new insights in the operation of complex equipment.

## Predictive Modeling

There is a large number of predictive data mining modeling algorithms, too many to describe in detail here (see http://www.statsoft.com/textbook/stathome.html). In general, most of them can be considered "general approximators", i.e., algorithms that are capable of identifying and modeling any kind of relationship between parameters, no matter how complex, interactive, or nonlinear those relationships may be. In fact, many neural networks models can be considered general approximators, but data mining techniques encompass many different classes of algorithms that, compared to neural nets, are faster, more efficient, more accurate, easier to interpret and optimize, or have other desirable properties (e.g., tend not to extrapolate from the data, and provide much more realistic, conservative, and robust models).

## Deployment and Optimization

Perhaps one of the main advantages of data mining methods, compared to CFD and the simple application of neural networks methods, is that the algorithms were specifically developed to provide *practical solutions*. For example, in credit scoring, the business interest and goal is to

identify the combination of demographic variables and prior credit history that allows credit institutions to make rapid accurate decisions regarding credit risk; likewise, applying data mining to power plant optimization will yield practical solutions that can be quickly implemented in a cost effective manner. For example, a project may yield recommendations to make minor adjustments to stoichiometric ratios PA, SA, TA air flows and OFA settings to dramatically improve the robustness of e.g. flame temperature and a reduction in NOx and CO.

## *Summary*

Through this paper, we have attempted to describe what data mining is, how it is different from other data analysis technology, and how it can be applied to improve power plant operations in the real world. There is little doubt that these techniques will find increased applications in different industries, including the power industry; this is evident in many complex manufacturing applications (e.g., the manufacture of silicon wafers, or computer chips), where these methods have been enthusiastically embraced, because as a practical matter, they provide the *only* reasonable approach to manage and control highly complex processes, in order to meet customer demand for quality products. Today, the power industry is facing difficult demands for greater efficiency and cleaner operations, and data mining is sure to become an integral part of the necessary technologies to meet those demands.