

# Numeric Components and Data Mining: Part 1

## By NAG for DStar Numeric Components and Data Mining Part 1: Computer Arithmetic Driving Market for Numeric Components

This is first of a three-part series that discusses how numeric components underlie the success of emerging data mining and automated knowledge discovery tools. Parts II and III of this series will be published in subsequent issues and will feature two case studies where numerics matter.

With the increasing sophistication of data mining and automated knowledge discovery tools, the independent software vendors who create them need to take extra steps to ensure that they create accurate solutions in a reliable manner. One of the first criteria they must establish is that their software produces consistently correct results. To meet this goal, many are turning to the world's experts in numeric software, such as the Numerical Algorithms Group.

Whether they realize it or not, virtually everyone who uses computers for statistical, scientific or engineering applications is touched by the Numerical Algorithms Group (NAG). Founded in 1970 as a project within the UK's University of Nottingham, but soon moved to Oxford and was established as a not-for-profit organization. It quickly expanded beyond British shores to include peers from computational science academies and research institutions around the globe. Today the Numerical Algorithms Group includes more than 300 individuals recognized as the leaders in their field whose work is coordinated through offices in the United States, the United Kingdom, Japan and Germany.

The Numerical Algorithms Group's worldwide experts are drawn together by common desires to tame sometimes unwieldy mathematical constructs into terms understood by binary-bound and finitely defined computing machines. One needs only to consider the difficulties of handling round-off errors from manipulating numbers such as pi (3.14159265358) to glimpse into the magnitude of problems which command their attention.

NAG components have long been used in military and other mission critical applications. In this interview with Tony Nilles, vice president of Sales and Marketing for the US offices of NAG in Downers Grove, Illinois, the emerging trend of data mining and similar software specialists using numeric components is discussed.

*How is computer arithmetic driving the market for numeric components among data mining specialists?*

Commercial software providers and others developing data mining tools are faced with the same sorts of problems in computer arithmetic that pose potential obstacles to anyone attempting to automate numeric computations. Though it might at first seem counterintuitive (to say so), computers are inherently flawed in their arithmetic capabilities. While these limitations are always a potential problem when results matter, they are especially important in calculations with enormous datasets such as those in typical data mining applications.

*What are the inherent problems in computer arithmetic?*

In a computer, any computer, you can retain only a finite number of significant digits to represent a number. This means that whenever a result cannot be exactly represented, a round-off error is introduced. There are numeric methods that can keep track of how a computer handles these errors and can make certain that they don't cause flawed results. For example, even something as simple as the addition of numbers can result in different answers depending upon the sequence of this addition and whether the numbers are represented in ascending order of magnitude, descending order or in arbitrary order. In some cases the computer will lose track of the small amounts "rounded off" and in other cases the computer will be able to add them up. You have to be aware of these consequences and structure a computer's computations accordingly.

Numerical computing experts are also aware of the fact that computational formulae that are equivalent "on paper" don't always provide the same results on a computer. And there are inevitably formulas that work in theory but are simply impractical and inefficient for calculation in a binary computer. These are not trivial problems but they are often surmountable problems. NAG has devoted 30 years to solving these problems. These efforts have resulted in the creation of libraries of sophisticated computational strategies, error analyses and testing methods to overcome the obstacles of automating numerical analysis.

*Why is numeric accuracy especially germane to data mining and automated knowledge discovery?*

These combinations of data volume and methods "torture test" software, so to speak. The size of the datasets alone makes it more likely that when something goes wrong it will have far-reaching consequences and potentially be unwieldy to fix. You just have to get it right in the first place.

In addition, the speed of computation on such large datasets can be dramatically altered by using the right method for numeric computations. Of course, getting the wrong answer fast is obviously of no use. But once you have accurate numeric methods, there are often faster (and slower) methods to make the computations. NAG has traditionally worked in fields with datasets of equal or larger size to those being handled in data mining and knowledge discovery problems. For example, a trajectory analysis in a military application or scientific calculations in areas where there are many unknowns and many variables require handling of the largest datasets. This is where NAG components were first proven in a field-tested environment. Today's businesses that use data mining and automated knowledge discovery tools are able to access the computational feats that underlie the exploration of space. A computational tool that put the first man on the moon now helps project future customer behaviour for bottom-line oriented problems.

*How has this migration of computational tools from military applications, aerospace industries, and science to business occurred?*

Many of the first data mining software engineers were actually classically trained physicists and other scientific researchers who first worked on NASA projects, experimental studies in particle physics, and the like. They had honed their research skills using NAG components and, when they switched careers, they just brought these tools to bear on their work in commercial data mining. But even newcomers to the field without such prior experiences in military and scientific research are usually eager to find numeric components that will save them development time without jeopardizing the accuracy of their products.

*How do numeric components speed up development time of data mining and automated knowledge discovery tools?*

Whenever someone is engineering a data mining application, they will want to develop a proof of concept version of their applications. These are prototypes of their final solutions and they typically work with smaller amounts of data on smaller machines so that they can work the bugs out of the system without contending with enormous dataset sizes. Once these details are ironed out, the details are applied to the full dataset to scale up the solution. This is another reason why NAG's numeric components have become so much more in demand in the data mining and automated knowledge discovery fields. NAG has a wide variety of what are called scaleable algorithms. You can run an application on one CPU computer and find it does a great job but you also need to know that it will be repeatable with a multiple CPU hardware approach.

In fact, this ramping up of dataset size is a staple of data mining and automated knowledge discovery software development. The scale of problems being worked on today is far beyond what was being worked on even a few years ago. The nature of these projects is that they continuously evolve and you need to find ways to make this evolution more fluid. Data mining specialists are likely to be seeking ways to extend and/or enhance their existing software systems with minimal interruptions encountered by the end-user. A component approach is very helpful in this type situation because as user interfaces and methods for modelling data change, you can retain the stability of numeric components that work equally well in other configurations of the software.

Most data mining and automated knowledge discovery software engineers have the mathematics and computer science training to automate numerics. However, they realize that they can often bypass many man-years of development time by using components that have been extensively field tested by others. Numeric components can be easily inserted into existing systems so that software engineers don't need to start from scratch. This means they can leverage existing hardware and technology for the existing day-to-day business and extend the capabilities of these applications more easily.

*How do numeric components speed processing time for end-users?*

Numeric components have the added advantage of attaching to programs at a very low level. This means that there aren't a lot of wrappers or hierarchy that needs to be traversed before the data is processed. The computer has less baggage to contend with, and it therefore provides the user with faster and more efficient solutions.

*How do numeric components impact business enterprises that are involved in data mining?*

It all boils down to getting the correct answer as quickly as possible and having confidence in your results. Superior numerical methods can't compensate for a poor model design or "dirty data". But what extensively field-tested numeric components CAN do is eliminate any concerns about the accuracy of your computer arithmetic.

*Does NAG have computer routines specifically designed for data mining or automated knowledge discovery?*

In recent years, NAG has worked with many of the world's top data mining and automated knowledge discovery firms such as Informix and PeopleSoft, as will be described in the next articles in this DSstar series. In responding to various requests to help develop data mining tools, NAG has realized that there is a market for a toolbox of algorithms specific to the data mining industry. NAG's worldwide R&D efforts are currently devoting significant time to creating such a tool and we hope to unveil it sometime in 2001.

(Part II of this series on Numeric Components and Data Mining, will delve into how Informix integrates numeric components into their most recent releases.)

ALM Communications Inc,  
1454 West Glenlake Chicago,  
IL 60660-1802,  
773-973-2077,  
[alm@almcommunications.com](mailto:alm@almcommunications.com)