# PREDICTIVE ANALYTICS – DATA MINING

## Data Mining and „Big Data“

With the technical possibilities of storing and processing very large quantities of data within an attractive time-frame, new applications for data analysis and data mining are available, thus presenting interesting prospects in the fields of quality management and efficiency optimization.

For some time, we have been hearing and reading more and more about „Big Data“, „Hadoop“, „Map Reduce“ and others. There are a growing number of conferences on these subjects and companies are increasingly asking themselves if “Big Data” is important for them too.
But what is it all about? As the name suggests, “Big Data” is characterised by a large quantity of data, which has traditionally existed in companies. But that’s not all. Big Data often comes from outside of the usual corporate applications, for example through data that accrues during industrialised processes, for example through sensors. In industrial production especially, due to the increased focus on reducing downtime and on minimising maintenance and warranty claims, the analysis and utilization of this machine-generated data is practically essential.
Big Data also often consists of data which was rarely analysed until now, such as unstructured or half-structured information, for example from text, audio or video files.
Big Data came about with the growth and increasing importance of the Internet. Companies like Google, Facebook and eBay are faced with vast amounts of data which they need to store and process quickly. This requires new technical methods, and these have taken shape in the form of Hadoop (framework for scalable, distributed software) and the MapReduce algorithm from Google. HBase is a scalable database for very large quantities of data within a Hadoop cluster, and Hive extends Hadoop to data warehouse functionalities.
The data stored within this framework needs to be available for analysis too, in order to be able to make good use of it. For this, there are also different methods, such as accessing HDFS and Hive data, ODBC and massive parallel processing.

> “‚Traditional’ data mining methods have already largely anticipated what is now sold under the “Big Data” label.”

### Big Data is not used to refer to millions of data sets

However, one should avoid thinking of Hadoop et al. if queries from the warehouse take too long – also, Big Data is generally not mentioned even for millions of data sets. Performance problems can be solved with relatively little effort compared to implementing a complex Hadoop, for example with in-memory technologies or simply a better warehouse design.
In the future, Big Data will also play an important role in secure data analysis. Text mining, which has long been established, is also part of the Big Data theme. Distributed processing is a technology we will undoubtedly be hearing a lot about.
“Traditional” data mining methods have already largely anticipated what is now sold under the “Big Data” label. For some time now, it has been a matter of course for the different data sources in a data mining project to be collated (for example sensor data from production appliances), for free text to be processed and included in analyses or for picture and audio data to be integrated. In this respect, “Big Data” represents an almost logical continuation of developments that have been taking place for years in the field of analysis and considerably expands this technology. ●