

The A to Z of Analytics with Modeler



is for Automation

Why bother trying out loads of modelling techniques to see which one works best when [Modeler](#) can do that for you?

Modeler can test many permutations of the same algorithm and multiple instances of different methods before selecting the best performers according to a pre-specified criteria.

Oh and it will also automatically prepare your data so you can get the best results from your analysis.



Auto Classifier



Auto Numeric



Auto Cluster



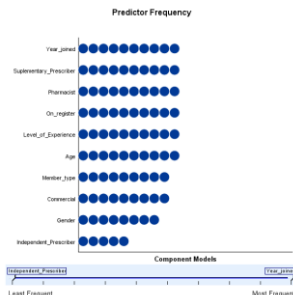
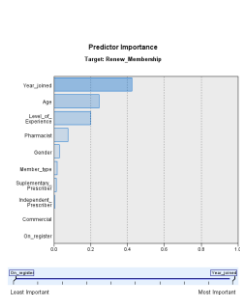
Auto Data Prep



B is for Boosting and Bagging

Boosting is a key technique in [Modeler](#) that can generate more *accurate* models. It works by building the same model multiple times but each time paying more attention to the cases where it failed to predict the outcome accurately. The final set of models are then combined together to create a single 'ensemble' model that produces one consolidated score for each record.

Modeler also offers Bagging (or Bootstrap Aggregation). This uses a similar method to Boosting but attempts to create a more *stable* model by combining the votes of several models together. Its particularly useful when you need more consistent results from your models.



Component Model Details

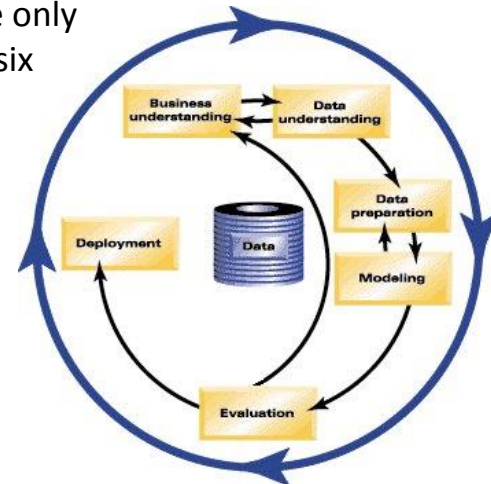
Model#	Accuracy	Method	Predictors	Model Size (Rules)	Records
1	90.0%	Boosting	9	63	3,600
2	89.7%	Boosting	9	59	3,600
3	89.0%	Boosting	9	57	3,600
4	87.3%	Boosting	9	70	3,600
5	84.3%	Boosting	10	75	3,600
6	89.0%	Boosting	9	69	3,600
7	89.0%	Boosting	9	66	3,600
8	89.0%	Boosting	9	59	3,600
9	84.3%	Boosting	9	63	3,600
10	87.7%	Boosting	9	71	3,600

C is for CRISP-DM

CRISP-DM is the Cross Industry Standard Process for Data Mining. It is the vendor-independent methodology adopted by thousands of organisations worldwide when they wish to develop and implement their own Predictive Analytics initiatives.

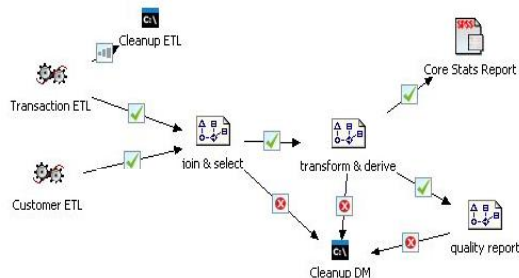
Modeler has a deep affinity with CRISP-DM as it is the only analytical workbench in its class that *fully* enables all six critical phases of the analytical life-cycle:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



D is for deployment

In analytics, the term 'Deployment' covers a *lot* of territory because there are simply so many ways in which the results can be fed back into the organization for better decision making. Whether it's saving models in [PMML](#) so they can be directly embedded in 3rd party systems, sending propensity scores to dashboards and BI tools, scoring warehouses with bulk loading, creating batch processes for scheduled execution or uploading streams to IBM's [Collaboration and Deployment Services](#) platform for coordination with other operational processes, [champion-challenger](#) modelling or even [real-time scoring](#), [Modeler](#) runs the *full gamut* of deployment options.



E is for Entity Analytics



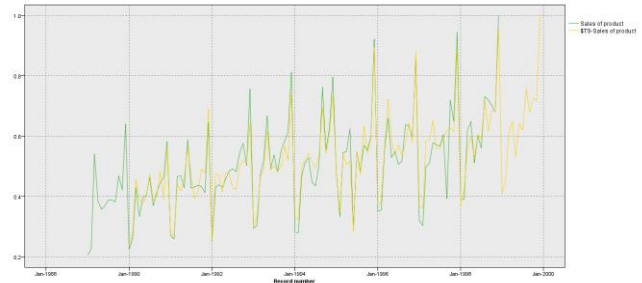
Entity Analytics

Being able to accurately resolve the identity of a person can be a serious problem for analysts and the organizations they work for. Apart from the issues around data quality and operational effectiveness, a vast amount of fraud is committed by individuals whose recorded identities are questionable. Entity analytics goes beyond simple duplicate detection by focussing on a number of contextual aspects of different records to identify those with a strong likelihood of belonging to the same entity. An entity can be a person, a company, a vehicle, a vessel or anything for which ambiguity might exist.

Entity Analytics is a key enhancement of [Modeler](#) Premium which combines powerful data cleaning capabilities with sophisticated approaches to fraud detection.

F is for Forecasting

Forecasting refers to a particular form of analytics where we wish to predict the values of something over a period of time. A more technical term for this approach is 'Time Series Analysis' which, among other things, is used to forecast the demand for products or services in the future. Once again, there are many different methods that may be employed to achieve this, however [Modeler](#) automatically identifies the time series method that best fits the historical data. Moreover, it can be used against *thousands* of different series - automatically selecting the best forecast model each time.



G is for Geospatial



Geospatial



Space-Time-Boxes



Spatio-Temporal



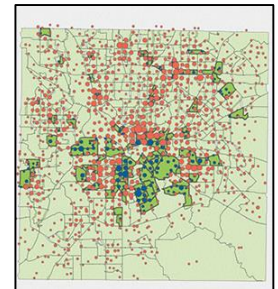
Reproject



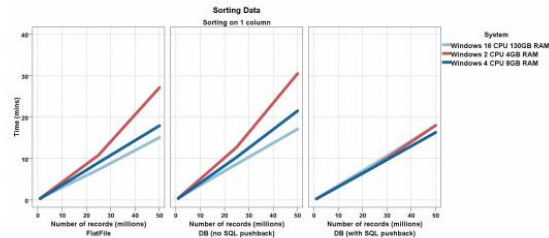
Map

Very often location matters. Unfortunately a lot of analytical techniques struggle to correctly exploit this information.

Using Modeler's [geospatial](#) analytical capabilities, retail chains can plan their expansion by predicting areas of higher demand. Police departments can identify associations between location and factors such as weather, footfall and time of day to predict when future crimes are most likely to happen, thus enabling them to increase patrols in those areas. Insurers can incorporate geospatial information such as population density, traffic volume and proximity to high crime areas into their risk calculations to optimize pricing for premiums.

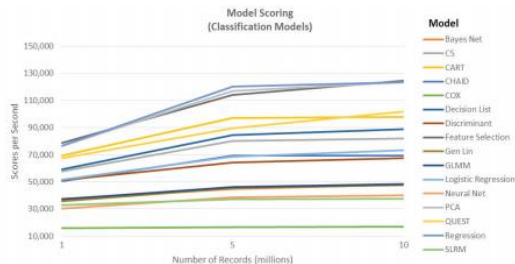


H is for Highly Scalable

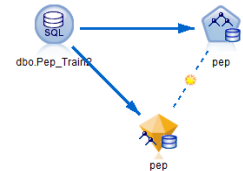


From a single desktop installation to an enterprise-wide deployment, Modeler scales and scales. In fact even when working within a simple [client-server](#) environment, Modeler's combination of SQL-pushback and parallel processing technologies mean that it is very nimble with large data volumes.

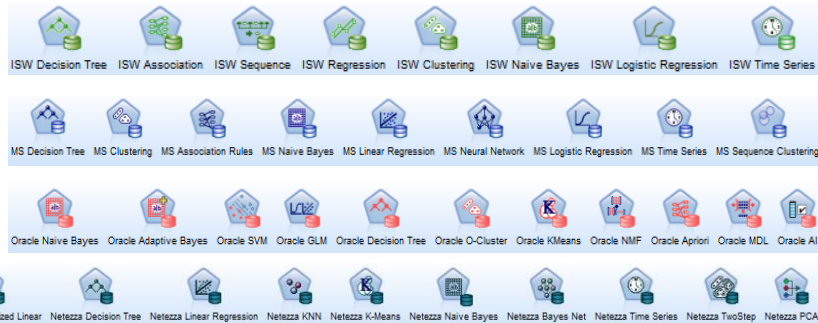
Furthermore, Modeler can harness IBM's own [Analytic Server](#) to address Big Data applications. Analytic Server allows organisations to distribute analytics processing into Hadoop environments with support for IBM InfoSphere® BigInsights™, Cloudera, Hortonworks and Apache Hadoop.



I is for *In-Database* Mining



When it comes to leveraging the analytical power of the warehouse, Modeler can access, configure and generate models using the native algorithms within IBM DB2 InfoSphere, Microsoft SQL Server (Analysis Services), Oracle and Netezza. As such, users can choose to build models without even extracting the data from the repository.

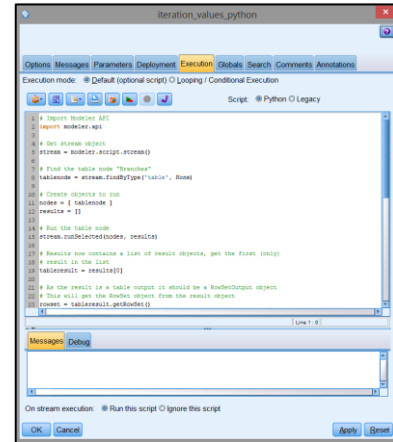


J is for Jython

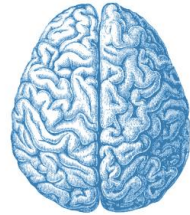


Modeler has supported scripting for many years as a way to automate and provide advanced control of stream execution. More recently it has included [Python](#) as the default scripting language. The implementation that Python uses in Modeler is called [Jython](#) which has some powerful capabilities of its [own](#) particularly with regard to providing an interface to Java.

Python is a well [documented](#) and popular language so it's relatively easy to find staff with the relevant skills to exploit this facility. Within Modeler, Python scripts have their own editor with colour-coded syntax, as well as a debugging tab and even auto-suggestion for function names (you can try this out by typing a letter or two and pressing CTRL+SPACE to see what functions start with those characters).



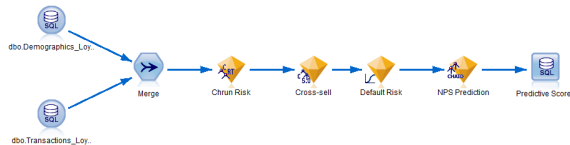
K is for Knowledge



Advanced analytical skills are currently in great demand so it's not surprising that experience with a major analytical platform like Modeler is highly prized. There are at least 8 official [courses](#) for Modeler from "Introduction to IBM SPSS Modeler" to "Advanced Predictive Modelling". Users can also attain professional accreditation by taking the "[IBM Certified Specialist - SPSS Modeler Professional](#)" certification exam. Alternatively there's an increasing amount of freely available [content](#) on the web devoted to providing help and education as well as user groups on sites like [LinkedIn](#) and [blog posts](#). There's even the occasional in-depth [book](#) available.

Last, but not least, there are even a few specialist [companies](#) ready to offer help and general no-strings-attached advice. 😊

L is for Licensing

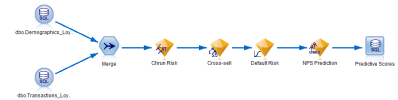


These days, there are a [lot](#) of options with regard to how you can license Modeler. The two primary forms of license are for **Authorized** users and **Concurrent** users. **Authorised** user licenses are tied to specific, named individuals. Organisations often buy a pack of these licenses. Alternatively, **Concurrent** user licences allow anyone in the organisation to use the product but stipulate the maximum number of people who can use it at any one time. There are also options regarding the length of the license. With **Perpetual** licences you pay a single one-off fee. This covers the cost of the software and your first year's maintenance. At the end of the year you can opt to renew your maintenance contract if you wish, or you can let it lapse. Alternatively you can have a **12 month fixed term** license or a **Rental** license if you only need it for a month or two. There are even Software as a Service (**SaaS**) options with no client installs required. Lastly you can sign up to a free 30 day [trial](#) just to try it out.



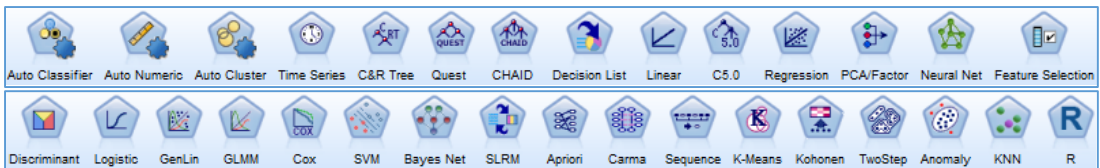


is for Modelling Techniques



Even when you ignore the option to include algorithms from Data warehouse platforms or SPSS Statistics or R – Modeler has a lot of [modelling techniques](#). These techniques address three broad areas.

Firstly, they cover **Predictive modelling** whereby category outcomes or actual numeric values are estimated. Secondly, they include **Association and Sequence** techniques that attempt to find events and categories that co-occur (such as product purchases). Lastly they cover **Segmentation** methods which enable analysts to identify similar (or dissimilar) groups within data. Such techniques are widely used to identify customers with similar requirements or behaviours as well as anomalous events or transactions.



N is for Nugget



“I could be bounded in a nutshell and count myself a king of infinite space”

Ever actually seen a predictive model? Normally they are expressed as reams of coefficients and esoteric fit statistics punched out as raw text in some suitably utilitarian font. For years the job of the analyst has been to pore over these results and make sense of the model before finally converting it into code so that it can be applied to new data for scoring.

In [Modeler](#) the model is a physical object. In fact it is a nugget. The entire model including the scoring code is encapsulated within a single portable object. You can browse the contents to assess it, interact with the graphical output, check the variable importance chart, generate new nodes to filter unused fields, convert it to PMML, annotate and archive it or simply attach it to a new data source node and watch it generate predictions.



is for Optimization



A prediction is not a recommendation. It's just a prediction. For example, if a customer is predicted to have a 23% likelihood of not renewing their contract, but an 18% chance of upgrading their current package and an 8% chance of defaulting on a payment – what is the right thing to do? This is one of the reasons organisations use IBM SPSS [Analytical Decision Management](#). ADM utilizes [optimization](#) technology in order to drive a decision that has the best chance of achieving an optimal outcome (such as maximising customer lifetime value). By combining business rules (such as eligibility controls) with predictions from [Modeler](#) and sending them to the optimization engine, they can ensure that each decision (even in real time) is smartest one they can make.

P is for PMML



Predictive Model Markup Language (**PMML**) is an XML-based file format developed by the [Data Mining Group](#) that enables applications to export and import models produced by different statistical or advanced analytical technologies. PMML models act as a sort of *lingua franca* for different software packages that utilise analytical models. Modeler can import and export PMML and indeed it provides PMML support for over 18 of its own model types.

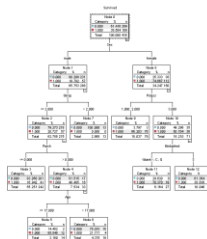
Some of the other organisations that also support various PMML standards include SAS, Angoss, Experian, MicroStrategy and Zementis.

Q is for Quinlan

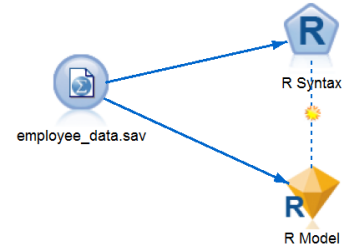


Ross Quinlan is the father of C5 and C5 is a truly excellent Decision Tree/Rule Induction algorithm based on the concept of [information entropy](#). It tends to be less well known than its (slower and less efficient) ancestor C4.5.

However, many experienced users of Modeler regard it as one of the best predictive algorithms in the group known as 'classifiers'. There are many reasons why it is notable, not least of all because it's fast, works well with target fields of unequal proportions, uses memory efficiently, frequently generates smaller trees that are easily converted to SQL and compared to other methods, it's often among most accurate on test samples. In short - it's a great all rounder.

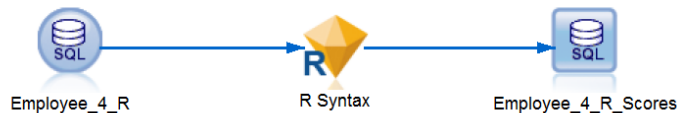


R is for R (of course)



[R](#) has shown itself to be the runaway success of open-source analytics in the last two decades. At last count, the main [site](#) hosting the archive of individual analytical routines and programs that analysts utilise when working with R code, contained over 7,000 individual ‘R Packages’.

For those who wish to combine the vast choice of analytical techniques that R offers with the power of Modeler’s intuitive visual programming interface, data manipulation capabilities and ease of use, Modeler provides a [direct interface](#) to R. It even has it’s own model-building node and produces output in the form of nuggets that can be combined with other procedures, browsed and used for scoring.



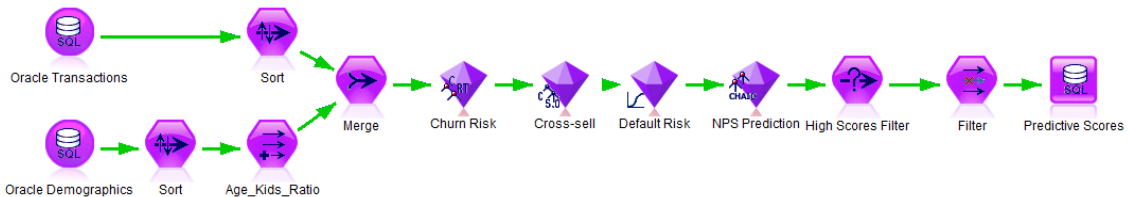


is for SQL Pushback



Let's say you are using [Modeler](#) to read and join a number of tables from a data warehouse. Then you attach some nodes to filter out the non-relevant records and fields before creating a series of newly calculated fields and aggregating the cleaned data. Finally, for good measure you attach a model nugget to generate predictive scores and punch out the results to a temporary table in the warehouse for campaign selection downstream.

You notice that when you run the stream it turns purple. That's because Modeler just converted the entire process to SQL and the data never even left the warehouse. That's called SQL Pushback.

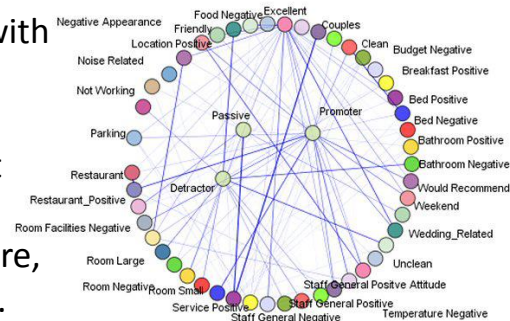


T is for Text Analytics



[Text Analytics](#) is a powerful component of [Modeler Premium](#) that enables users to transform unstructured data into structured. Using advanced linguistic technologies and [Natural Language Processing](#) (NLP), Modeler Text Analytics can extract key topics, themes and sentiments from a wide range of text sources such as reports, web pages, e-mails, and call centre notes. These extracted concepts can be used to automatically categorise free text.

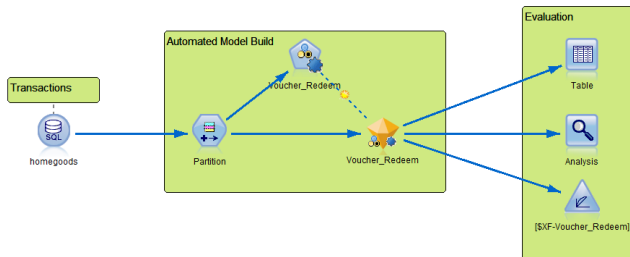
The categories themselves can then be combined with existing structured data, such as demographics or transactions and used to enhance the insight and accuracy of other modelling techniques. Using text analytics, organisations can build more effective models in areas such as customer churn, asset failure, fraud detection and estimating net promoter score.



U is for User Interface



A truly ground-breaking technology, [Modeler](#) (formerly known as [Clementine](#)) was the first analytics tool of its kind to use an icon-driven [Graphical User Interface](#) rather than requiring users to write code. Because advanced analysis is really a *process*, Modeler employs a *visual* programming interface that illustrates the entire analytical journey. In fact, Modeler's interface is one of the most compelling aspects of the software because it's so responsive and intuitive that it enables a kind of 'train of thought' approach to analytics. This approach helps users to be a lot more productive than using other more traditional interfaces that are based on writing code or selecting dialog boxes.





is for Victory Index

The [Hurwitz Victory Index](#) is an annually published, detailed assessment by Hurwitz and Associates of Advanced Analytics technologies and their respective vendors. The report examines the top trends for end users to consider and analyses 10 vendors across four key dimensions: vision, viability, validity and value.

With [Modeler](#) at the heart of its predictive analytics portfolio, in 2014 IBM was evaluated as a double victor receiving a victor rating in both 'Go to Market Strength' and 'Customer Experience Strength'.

"IBM has expanded its advanced analytics platform to provide customers with an integrated and holistic approach to managing big data and analytics" — Hurwitz & Associates 2014



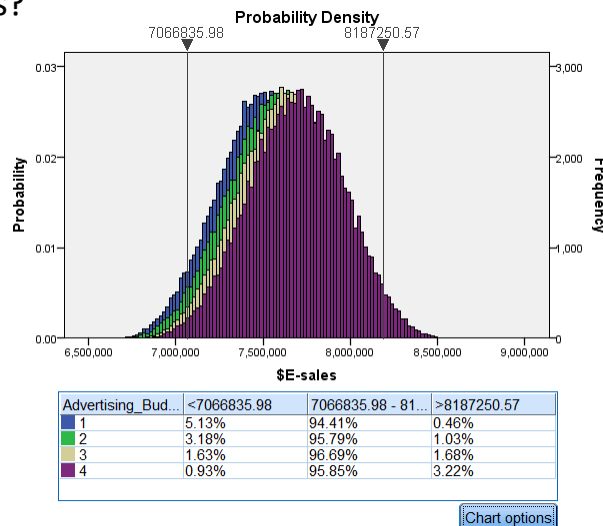


is for 'What if' Analysis

What if we increased our advertising budget by 10%? Or even 20%?

What would that do to our estimated revenues?

Modeler makes it easy to evaluate the impact of such changes. It generates simulated data that closely matches the original information. It uses the simulated data so that you can see how estimates or predictions are impacted by altering the values of key input factors such as budget or product demand or weather events. It actually goes beyond predictive modelling and allows business-minded analysts to test out different scenarios so they can make the best decisions for the future.



X is for eXtensions

Modeler's direct integration with R has encouraged contributors to IBM's [Predictive Analytics Community](#) to develop and share a [catalogue](#) of extensions that enable people to exploit R routines *without writing a single line of code*. The extensions themselves are installed as customised, configurable nodes that can be used alongside any other procedure in Modeler.

Currently the [catalogue](#) includes a number of extensions that enhance Modeler's Geospatial analytics such as links to Google Maps, reverse geocoding and geographical heat maps. The extensions are easy to download and install – there's even a [YouTube](#) channel that's shows you how to use them.



Y

is for $y = mx + c$



Otherwise known as the equation for a straight line. As such it's the function that underpins the granddaddy of predictive modelling: linear regression.

Modeler includes a classical multiple Linear Regression algorithm in its portfolio of predictive procedures as a matter of course. But it also contains another procedure simply called 'Linear'. This is an enhanced version of the Linear Regression routine that supports boosting and bagging, more criteria for variable inclusion/removal as well as a bunch of automatic data preparation routines such as missing value and outlier handling, date and time handling and merging of similar categories.



Linear



Regression



Z is for System z



Well it is an IBM product after all. Although Modeler *clients* run only on Windows, the Modeler [Server](#) options include Windows (2008 & 2012), AIX, Solaris and Linux (Red Hat, SUSE & Ubuntu).


Meanwhile, in the mainframe world, Modeler is now a core component of [IBM's Predictive Analytics on Linux for System z](#). Because so much of the data, from transactional to demographic details, human resource records to manufacturing reports, originates on the System z platform, it makes sense to bring the predictive analytics to the data. Using Modeler in this kind of environment has a number of benefits: because the data is effectively *in situ*, not only can the speed of data transformation and model building tasks be vastly increased but when coupled with [IBM's DB2 Analytics Accelerator](#), it becomes much more straightforward to execute high-volume scoring in real time applications.



Contact us:

+44 (0)207 786 3568

info@sv-europe.com

Twitter: @sveurope 

[Follow us on Linked In](#) 

[Sign up for our Newsletter](#)

Click [here](#) to find out about IBM SPSS Modeler