



White Paper

Making More Decisions Intentionally
&
Competing on Analytics in the Real World



Extreme Data Mining

Executive Summary

Leading companies in varied fields – finance, retail, telecommunications, and consumer products among them – are increasingly using data mining (or predictive analytics in new parlance) as a source of competitive differentiation.

As the volume of data in enterprise data warehouses increase, companies wanting to get value out of their data asset are turning to data mining (DM) to produce models for every business process and every decision. Deploying data mining on a large scale – which we call *Extreme Data Mining* – poses specific opportunities and challenges.

Historically, data mining has been in the hands of small teams of expert statisticians who only produce a few models per year for core business issues. However, as companies transition to competing on analytics and decide to embed predictive analytics into the workflow of their business processes, the economics of traditional data mining break down.

Enterprise data warehouses in production now range from a few terabytes to petabytes in size, and contain millions of records and thousands of variables; for example, 5,000 variables on 150 million customers and prospects. Businesses who want a return on that investment are looking well beyond reporting and basic statistics [1]. Often, a review of business activities reveals the opportunity to significantly enhance business results through 100s or 1000s of predictive models per year.

Traditional data mining tools do not scale to these requirements as they rely on expert staff for even the most basic analytics. Fortunately, a new kind of math, championed by KXEN, is helping leading companies achieve Extreme Data Mining.

Recent KXEN customer successes with Extreme Data Mining include:

- A major US broadband services provider advancing from 5 cross-sell models per year to 1,600, with response rates rising from 1.5 percent to 5.5 percent.
- A European wireless communications provider upgrading from 2 to 700 churn prevention models per year, reducing overall annual churn from 26 percent to 21 percent.
- A direct marketing firm building 300 propensity-to-buy models, increasing response rates by 2.5 times over former methods.
- A leading direct-to-consumer insurance company building hundreds of models on click-stream data, enabling its e-commerce agents to connect with three times as many customers and to close twice the business using real-time predictive scoring, raising their effectiveness six fold.

Extreme Data Mining requires looking at analytics from an industrial viewpoint, handling very large data volumes (records and variables), the ability to produce robust and reliable models with little manual intervention, and utilizing technology that is self-optimizing and self-configuring. KXEN delivers Extreme Data Mining through the extensive application of a new kind of mathematics, Vladimir Vapnik's SRM framework [2].

KXEN is proven new technology deployed at 500 global sites. It has transformed companies' decision-making, or multiplied the great results they have already achieved with analytics. We would be delighted to discuss how we may assist you in achieving great results through analytics – please contact us at www.kxen.com if you would like to schedule a call with your local KXEN Sales Consultant.

Françoise Fogelman Soulié
francoise@kxen.com

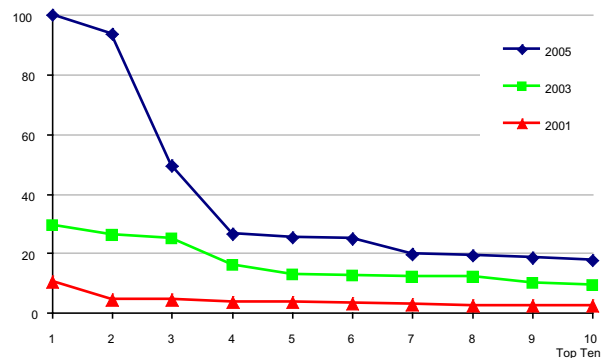
Data Mining In Companies Today

Data mining is widely used in companies today, mostly for CRM, fraud detection, credit scoring, web mining ... (see [3] for example). Yet, data mining is still mostly seen as an art to be practiced only by experts (statisticians, data miners, analysts).

The development of data mining-based applications is strongly linked to the ever-increasing availability of large volumes of data. In the last ten years, companies have invested heavily in the implementation of very large data warehouses, where the larger databases nowadays reach 20 to 100 terabytes and comprise millions of customers and thousands of variables. Sizes typically keep tripling every two years (Fig 1).

After such investments (in the \$ 100 million range), companies need to see measurable returns. Data mining can bring them just this, if the company is willing to put in place the right organization. Successful "analytics competitors" [1] need to invest in collecting lots of data of course, but also to commit to basing all decisions on data, which will require top management commitment and employees capable of handling data and producing models for every business process in the company (which will probably mean hundreds of models per year in the company).

Figure 1. Database size in TBytes
(from <http://www.wintercorp.com>)



Enterprise Performance Management approaches such as e.g. Six Sigma, Baldrige, Balanced Score Card, Kaizen ... [4] provide systematic methodologies for improving key business processes. Actually investigating all key business processes may lead to a very large number of required models, for example, a major European wireless communications company [5] has identified the needs for 716 models per year (Fig. 2).

Figure 2. Number of models needed at a major communications company

Domains	# Analysis / Year
Segmentations	
$2 \cdot 2 \cdot 10$	40
Churn in General	
$2 \cdot 3 \cdot 2 \cdot 3$	36
Churn per Product	
$2 \cdot 3 \cdot 2 \cdot 4 \cdot 10$	480
Cross sell : segments* offers	
$2 \cdot 4 \cdot 10$	80
Acquisition	
$2 \cdot 4 \cdot 10$	80

The production of a data mining model will usually involve a triangular relationship (Fig.3).

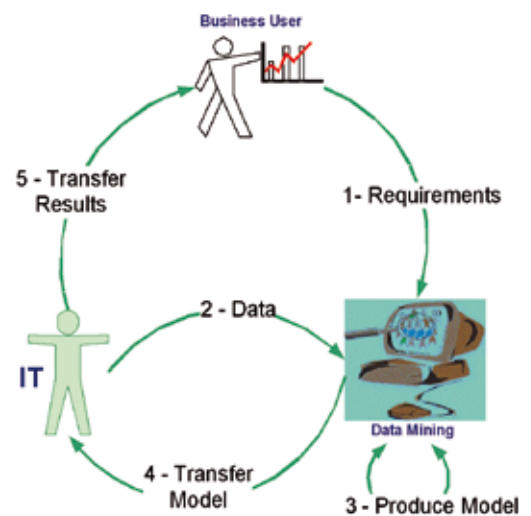
- The business users have an operational role (for example, they are in charge of designing / planning / executing marketing campaigns). They will list their requirements and will use the model results in their business processes;
- The Analytics Department serves the needs of 20-50 business users, producing 5-10 models / year with 5-10 data mining experts. Data mining experts usually are a very scarce resource in the company; they will produce the models in line with business users requirements;
- The IT department is in charge of maintaining the data bases and executing the models transferred to them by analysts to produce the results (e.g. scores) used by the business users.

This process will result in a typical 3-8 weeks delivery time and only a handful of predictive models, which is hardly compatible with the agility needed in a competitive market.

What is needed is industrial data mining at a greatly accelerated pace – we call it *Extreme Data Mining*. This requires that a company put in place a *model factory* capable of churning out hundreds of models per year on terabytes of data (millions of lines – customers – and thousands of variables). The right data mining technology is indeed a key ingredient and it will have to provide employees flexibility, ease-of-use and productivity.

Unfortunately, most of the time companies fall short of achieving such feats: data mining tools are so complex that only a few expert analysts can handle them, producing only a few models per year, with the result that most decisions are based not on data-driven analysis, but simple rules of thumb.

Figure 3. People involved in a Data Mining project



Requirements for Extreme Data Mining

Extreme Data Mining needs three ingredients: the right culture, the right people and the right technology [1]. Critical to the latter is the right data mining tool, which should have the following characteristics.

1. Data manipulation

- Data sources usually come in heterogeneous formats (continuous, nominal, ordinal, text, image, speech, video ...). The data mining tool should be able to handle all these formats, with as little intervention as possible from the user. It might be necessary to automatically recode some of the variables;
- Data quality is often poor, especially with large volumes, where multiple sources might have very different quality levels. The DM tool has to automatically handle missing values, and get robust results with respect to outliers;
- Data volume continues to grow exponentially, tripling in size every two years. One should be able to handle millions of rows (e.g. customers) and thousands of variables. This growth in data volume should not require the user to:
 - Hand-pick 'relevant' variables. This process takes time and requires expert knowledge, and is thus not suitable for handling thousands of variables;
 - Look individually at each variable and study its statistical properties, for the same reason as before;
 - Duplicate or move data to produce models. The DM tool should be able to work directly off the enterprise data warehouse – after spending millions of dollars on your Enterprise Data Warehouse; you don't want to rebuild yet another data store for data mining.

2. Model production

- Model calibration should be reasonably fast. Algorithms requiring days of computation should be ruled out, as well as algorithms not scaling well with respect to number of variables and examples.
- Model application. For some applications, real time or "right-time" [5] is necessary. For example, for web behavior scoring, a model has to be applied in real-time to data extracted from the click-stream and the result transferred on-the-fly as the Internaut goes on clicking. DM tool scoring must thus execute quickly.
- Model industrialization
 - Export. After a model is produced, it may have to be used on a regular basis. It will thus be necessary to automatically export this model to a production environment. This has to be done fast and accurately, and not as a hand-coding project. "*Velocity is always king! Good enough and deployed is always better than perfect and in the lab !*" [6];
 - Control. When a model is regularly applied on new data, one has to make sure that the structure of the data has not changed. The DM tool has to provide ways to easily check for deviations or changes in the nature of the data (model drift).
- Methodology. Data mining is about solving a problem with an unknown solution. A solid methodology is required to define the business objectives, identify measures of success, analyze the data and build predictive models, select the right improvement solution, and finally control and verify the results.

3. Users

- Users know their business. They know which issues and problems are key, what data are used / generated by their activity, and the business value of the results delivered by a model. The DM tool must allow them to easily express their business questions in non-statistical jargon;
- Business users do not know statistics (and do not want to learn!). So they cannot tell which is the best algorithm to be used, how to manipulate / recode data, how to select variables according to statistical significance, how to handle outliers, missing data, how to 'decode' results, or how to evaluate the statistical validity of a result. The DM tool must do all of this for them (automatic algorithm selection, data coding and selection, missing data and outliers handling, guaranteed validity of results) and the model must be self-explaining with business-explicit reports (avoiding statistical jargon).

4. Integration

- Technical
 - Analytic modeling is usually a component of a larger process (e.g. marketing campaign). The DM tool will thus be integrated into an architecture including databases and various other tools. To make the integration easy, the tool needs to be compliant with IT architecture standards (COM/DCOM, C++, CORBA, Java API, SOA, Web Services, J2EE) and data mining standards (JDM, PMML);
 - Model results can be produced in two ways, either within the tool (which executes the model on a given data set) or the model is exported to further execute on a different tool. In the first case, a list manager will be needed to manipulate the results produced and possibly import them into the data warehouse. In the second case, the DM tool should produce code in any format (SQL, UDF, Java, C, XML, HTML, SAS, ...) so that the exported code can easily execute within the target IT platform (for example, execute directly in the database through SQL or UDF);
 - A components architecture will allow users to build their analytics architecture progressively, including components only when they are needed;
 - Data Access API should allow to connect to any data format (text files, databases – SQL, DB2, Oracle, Teradata ... – SAS, SPSS, Excel ...).
- Operational
 - Process control & workflow. The DM tool will work within a general IT environment and will thus be inserted into a general process managed through usual IT tools (job scheduler, version control, users rights management, workflow tool ...) The DM tool must then be open and provide generic, well-documented Application Programming Interfaces (APIs).

5. Value

Models must deliver useful, accurate results, so that users can exploit them in their business processes.

- Exploratory modeling. The model is used to understand the data, for example, the key drivers for a customer to churn, or the characteristics for a given segment. The need here is for the DM tool to provide easy-to-understand business-oriented reports and graphics.
- Predictive modeling. After the model is built on existing data, it will be put in production against new data. The performance of the model on the new data is expected to be of the same order as the performance calculated a priori with existing data: the model needs to be robust and trusted. The DM tool must provide both an automated and easy-to-understand quality assessment, and guaranteed robustness.

KXEN Analytical Framework

At KXEN, we build technology for 'benefit creators'. KXEN Analytic Framework (AF) has been built from the ground up, utilizing a new kind of math developed in 1990s, to answer the requirements of Extreme Data Mining:

- Data preparation, recoding, and data quality handling. KXEN AF provides fully automatic optimal coding, and can handle very large data volumes effectively because it does not duplicate data.
- Model production is automated. With KXEN AF, one chooses functions (such as prediction segmentation, forecasting, basket analysis, in line with the JDM standards) - there is no need to choose or tune algorithms.
- Model production is made available to the business analysts through a simple, click-through, business-language interface.
- Full life-cycle of model production is extremely fast (including variable encoding), for example, on a laptop PC, a campaign model with 1,000,000 customers and 220 variables is built in 13 minutes.
- Model quality is guaranteed through built-in robustness, based upon the new math of 1990s, Vapnik's Structural Risk minimization theory [2]. KXEN AF provides an easy to understand robustness percentage KR.
- Model export is automated through the KMX module, capable of exporting the model code to – almost – any format or language.
- Model control can be automated through the use of the 'deviations detection' functionality.
- Model integration is greatly automated. KXEN AF offers access to all data formats and export to all formats. KXEN AF is compliant with DM standards: JDM [8] and [9].

Real World Examples

KXEN customers want to make more decisions intentionally based on knowing the odds, and are deploying *Extreme Data Mining* to fully exploit their investment in data. They have decided that data mining should not be confined to small groups of expert users, but taken to the next level, into the workflow of their businesses.

We now would like to share some examples, omitting protected commercial details.

Cox Communications

Cox Communications [10, 11] started using KXEN in September 2002 in its marketing department, to analyze its customer data base. It now produces hundreds of models for marketing campaigns in 26 regional markets from a data base of 10 million customers and 800 variables.

- Cox believes that *“the ease of using intuitive click-through menus lets analysts focus on creating models without extensive data preparation”*.
- Cox has found that only 4 senior analysts were needed to manage the work, with one *“regular analyst ... self-sufficient in creating and supporting analytic models”* per region.
- Time to produce models, from start to finish has reduced by approximately 80 percent bringing model building time from three weeks to one.
- Since Cox began using KXEN, results have improved : direct mail responses returns have risen from 1.5 to 5.5 %, churn rate has reduced by a percentage point.
- *“By using this tool, the company has realized the return on investment in the (first) two months it has been in service”* [10].

Sears

Sears [12, 13] has been an early adopter of predictive analytics technology, but their initial mainframe-based system became too expensive and inflexible. They wanted to improve cost, performance and quality in their catalog business with low expense and a small staff performing all duties (Modeling / Analytics, Data Operations and Marketing).

- First they integrated data from Sears’ multiple channels, brands, credit data, market demographics and external sources, resulting in a data mart with more than 900 attributes, integrated into the corporate Teradata warehouse.
- Then they used KXEN to automate the data preparation process, including attribute importance and nominal attribute encoding, among others.
- They now build more models with better model quality, while reducing model development time and costs. For example, it now takes a few hours to create robust models where it used to take weeks.
 - Finally, Sears uses KXEN to automatically generate model deployment code for the data warehouse, eliminating the need for hand-coding the models and allowing changes in minutes, instead of hours or days. Now, Sears can score 75 million customer records in 30 minutes.

- Results and benefits identified by Sears are :
 - Creating and implementing models now takes 1-2 days.
 - Expert statisticians are not needed to conduct a campaign or analyze customers.
 - Sears has cut 50%+ off operational costs of analytics.
 - Sears has cut 90%+ from modeling and scoring time.
 - Sears can react to operational changes very quickly.

A major European wireless communications company

The company wanted a complete analytic environment to support the easy creation and deployment of churn and cross sell / up sell models [5].

- They built a Teradata warehouse where their CAR (Customer Analytic Record) contains over 2500 meaningful variables, with monthly aggregates updated on every billing cycle. By building the CAR process once in their Teradata Enterprise Data Warehouse, they reduced the data preparation step from 70% or more of the modeling work effort to almost nothing.
- They identified the need to build more than 700 models per year (Fig. 2) to be used for various business activities (marketing, customer price sensitivity analysis, channel preferences ...).
- KXEN gave the company the ability to move from developing a small number of predictive models per year to hundreds of models per year.
- They found that consistently high quality models can be produced by less experienced analysts.
- KXEN is giving the company the ability to automatically deploy models into production and rescore customers whenever necessary.

Conclusion

A major shift is happening in data mining; from Artisanship to Industrial Model Factories driven by massive data sets and the desire to extract the full benefit out of data assets.

The economics of traditional data mining break down when a business decides to embed predictive analytics into the workflow of its business processes and to make more decisions intentionally based on knowing the odds.

Extreme Data Mining is bringing data mining into every business, process, and activity where data exists. KXEN technology, whether integrated into specialized operational applications, or used interactively by business analysts, is giving businesses the tools they need.

We believe that *Extreme Data Mining* is for benefit creators, people who understand a business issue and are best equipped to create value. *“Data mining is not really an ‘end’ per se, but a means to an end. These ‘means’ will become progressively submerged in the infrastructure of the products they serve until they are as natural to use as standard arithmetic and graphical techniques”* [15].

KXEN is proven new technology for *Extreme Data Mining* deployed at 500 global sites. It has transformed companies' decision-making, or multiplied the great results they have already achieved with analytics. We would be delighted to discuss how we may assist you in achieving great results through analytics – please contact us at www.kxen.com if you would like to schedule a call with your local KXEN Sales Consultant.

References

1. Davenport, Thomas (2006) "Competing on analytics". Harvard Business Review, January.
2. Vapnik, Vladimir (1995) "The Nature of Statistical Learning Theory". Springer.
3. Piatetsky, Gregory (2006) "Poll Results: Top Industries for Data Mining Applications". www.kdnuggets.com/news/2006/n13/1i.html
4. <http://www.motorola.com/motorolauniversity.jsp>, <http://www.quality.nist.gov/>
5. West, Andreas & Bayer, Judy (2005) "Creating a Modeling Factory at Vodafone D2: Using Teradata and KXEN for Rapid Modeling". Teradata Conference, Orlando. <http://www.teradata.com/teradata-partners/conf2005/>
6. Herschel, Gareth (2005) "Right Timing Customer Analysis". Teradata Conference, Orlando. <http://www.teradata.com/teradata-partners/conf2005/>
7. Harris, Matt (2005) "The Journey from Product to Customer Centricity". Teradata Conference, Orlando. <http://www.teradata.com/teradata-partners/conf2005/>
8. Java Community Process (2005) "JSR 73 : Data Mining API". www.jcp.org/en/jsr/detail?id=73
9. Hornick, Mark F., Lei Liu, Marcade, Erik, Venkayala, Sunil, Yoon, Hankil (to appear) "Java Data Mining. Strategy, Standard, and Practice. A practical guide for architecture, design, and implementation". Morgan Kaufmann.
10. Douglas, Seymour (Feb 2003) "Product Review – KXEN Analytic framework". DMReview.
11. Ericson, Jim (Dec 2005-Jan 2006) "Perfect pitch". Business Intelligence Review.
12. Bibler, Paul and Bryan, Doug (Sep 2005) "Sears: A Lesson in Doing More With Less". TM Tipline. http://ga1.org/tmgroupp/notice-description.tcl?newsletter_id=1960075&r=#6
13. Bibler, Paul (2005) "Lifting Predictive Analytics Productivity at Sears". Teradata Conference, Orlando. <http://www.teradata.com/teradata-partners/conf2005/>
14. MMDS 2006 (2006) "Workshop on Algorithms for Modern Massive Data Sets". <http://mmds.stanford.edu>
15. Nisbet, Robert A. (March 2006) "Data Mining Tools: Which One is Best for CRM?" DM Direct Special Report http://www.dmreview.com/editorial/newsletter_article.cfm?articleId=1050627

Contact KXEN

USA and Canada Offices
Headquarters
KXEN, Inc.
201 Mission Street, Suite 1950
San Francisco, CA 94105, USA

Tel: +1 (415) 904-4160
Fax: +1 (415) 904-9041
Email: sales-us@kxen.com

European Offices
Headquarters
KXEN SARL.
25 Quai Gallieni
92158 Suresnes Cedex
France

Tel: +33 (0)1 41 44 88 44
Fax: +33 (0)1 41 44 88 40
Email: sales-eur@kxen.com

United Kingdom
KXEN Limited
400 Thames Valley Park Drive
Thames Valley Park
Reading
Berkshire, RG6 1PT, UK

Tel: +44 (0) 118 9 65 34 24
Fax: +44 (0) 118 9 65 35 25
Email: sales-eur@kxen.com



About KXEN

KXEN provides next generation business analytics software to drive better corporate decisions. KXEN's unmatched speed, ease of use and scalability enable leading companies around the world to expand the use of predictive analytics and enhance corporate performance. Based on breakthrough mathematical theory, KXEN's products offer reliable predictions and deep insight for achieving critical business goals. The company partners with leading systems integrators and software vendors to integrate advanced analytics into enterprise applications and business processes. Founded in 1998, KXEN is headquartered in San Francisco, California, with offices in the USA, UK, and France, and distributors throughout the world. For more information on KXEN, visit the KXEN Web site at www.kxen.com.

www.kxen.com

(c) Copyright 1999-2006 KXEN, Inc. All rights reserved.
KXEN Analytic Framework, K2C, K2R, K2S, KTS, KAR, KEL, KSC, KMX, and KXEN Logo are trademarks of KXEN, Inc.