

# Open Standards and Cloud Computing

## KDD-2009 Panel Report

Michael Zeller<sup>1</sup>  
Zementis

Robert Grossman  
DMG &  
Open Data Group

Christoph  
Lingenfelder  
IBM

Michael R. Berthold  
KNIME.com &  
University of Konstanz

Erik Marcadé  
KXEN

Rick Pechter  
Microstrategy

Mike Hoskins  
Pervasive Software

Wayne Thompson  
SAS

Rich Holada  
SPSS

### ABSTRACT

At KDD-2009 in Paris, a panel on open standards and cloud computing addressed emerging trends for data mining applications in science and industry. This report summarizes the answers from a distinguished group of thought leaders representing key software vendors in the data mining industry.

Supporting open standards and the Predictive Model Markup Language (PMML) in particular, the panel members discuss topics regarding the adoption of prevailing standards, benefits of interoperability for business users, and the practical application of predictive models. We conclude with an assessment of emerging technology trends and the impact that cloud computing will have on applications as well as licensing models for the predictive analytics industry.

### Categories and Subject Descriptors

H.2.8 [Database Management]: Data Mining; G.3 [Probability and Statistics]: Statistical Computing, Statistical Software; I.5.1 [Models]: Statistical Models, Neural Nets.

### General Terms

Management, Performance, Standardization, Languages.

### Keywords

Open Standards, Cloud Computing, Predictive Model Markup Language, PMML, Predictive Analytics, Data Mining.

## 1. INTRODUCTION

Over the past decade, we have seen tremendous interest in the application of data mining and statistical algorithms, first in research and science, and more recently across various industries. Impacting scientific and business applications alike, interoperability and open standards still lack broader adoption in the data mining community. Additionally, emerging trends in cloud computing and Software-as-a-Service (SaaS) will play a critical role in promoting the effective implementation and widespread application of predictive models.

<sup>1</sup> Panel Moderator: As an active member of the PMML standard, it has been the privilege of Zementis to organize this panel.

This panel discusses various topics related to open standards and cloud computing, with a particular focus on the practical use of statistical algorithms, reliable production deployment of models and the integration of predictive analytics within other systems.

With broad industry support and a new version 4.0, the Predictive Model Markup Language (PMML) standard has reached significant maturity stage. The Data Mining Group (DMG) [1] has succeeded, uniquely, in establishing a widely accepted standard that allows users to exchange predictive models among various software tools. However, user awareness of PMML still lacks the current vendor adoption. This panel session provides an excellent forum to highlight the benefits of such a common standard for researchers as well as business users and IT.

Open standards and cloud computing [2, 3] not only have the power to enable more data mining applications across science and industry, but more importantly they will lower the total cost of ownership and, therefore, allow the community at large to focus on the essence of algorithms, avoiding proprietary issues and incompatibilities among systems.

## 2. PANEL QUESTIONS

### 2.1 Please comment on the importance of open standards for the data mining community and for your organization.

**Robert Grossman.** You can argue that open standards are more widely used and followed in the database community, as compared to the data mining community, and that the database community has benefited because of this. In my opinion, over the long run, the benefits from a community embracing open, consensus driven standards far outweigh the of costs developing them and the inconveniences and compromises that sometimes arise when following them.

**Christoph Lingenfelder.** Open standards are important in general, because they ensure interoperability in a well-designed formalism. For both the data mining community and for IBM, data mining standards make mining models better to compare. Many of IBM's customers use more than just one data mining tool for modeling, but want to deploy these models consistently inside their database.

**Michael Berthold.** I think standards are crucial for the enterprise wide adoption of data mining tools. Users don't want to be tied to a particular flavor of an algorithm of a particular vendor but want to

be able to use the same model in different platforms. I believe that also the entire analysis process (e.g. including ETL) will need to adhere to standards, not just the analytical parts.

**Erik Marcadé.** Standards are key for a wider adoption of data mining. Standards emerge only at a time when a domain achieves some maturity, allowing competitors to come to a single table and discuss their common concepts, so this is a good sign for data mining. There are mainly two standards that are live for now: The Java Data Mining API (JDM) and PMML. KXEN is part of and compliant with both standards.

**Rick Pechter.** Open standards are key to enabling broader deployment and adoption of data mining. Similar to how SQL and ODBC opened up database access thereby improving the development and deployment of data processing applications, data mining standards will allow users to extend historical reporting to include vendor-agnostic predictive analysis.

**Mike Hoskins.** We expect data mining to continue its expansion from historically narrower usage by scientists/experts, into much wider adoption by practitioners of all sorts, from all industries. After all, who couldn't benefit from the extra analytic horsepower? Open standards are clearly a facilitator (and even accelerator) of this adoption - avoiding traditional pitfalls from proprietary vendor "lock-in" strategies, and allowing for faster, healthier and more robust growth of a vibrant, but still diverse ecosystem of vendors and users. At Pervasive Software, we have been supporting standards-based tooling for years, and will do so in our upcoming DataRush-based massively parallel data mining offerings.

**Wayne Thompson.** SAS has been one of the first vendors to recognize the need for operationalizing analytics, i.e., moving beyond just knowledge discovery. Open standards have been very critical to deploy analytical models in a heterogeneous environment.

**Rich Holada.** Extremely important. Open-standards enable all parties involved – practitioners, consultants, software vendors – the ability to focus on applying the most robust and user-friendly modeling techniques and visualizations. As predictive analytics becomes more mainstream and organizations regularly use predictive models to drive business decisions, it is important that the delivery of that intelligence be achieved through open standards into operational systems. Whether that be through a scoring web service; combining the results of business rules or optimization technologies with predictive modeling; PMML; or scoring a large volume data in a database; the more open the technology the easier it will be for organizations to make use of the predictive models. A lack of standards merely exacerbates model preparation time – already estimated at 75-80% of the effort and contributes to the formation of silos. PMML in particular is beginning to play an important role, as we see the emergence of scoring engines that consume PMML from multiple sources.

## 2.2 What are the main benefits of open standards for the practitioners building models (analyst, scientist) and for the practical application in a business environment?

**Robert Grossman.** Standards make it easy to interoperate with other software so that best of breed software and systems can be

used. Standards also make it more likely that over time statistical models are not lost, if the specific software that created them is no longer available. Standards can also make it easier to develop specialized custom components for data mining and predictive modeling systems. Standards give modelers choices if they ever have to move from one data mining or predictive modeling system to another.

**Christoph Lingenfelder.** There are two types of standards, standard API's and standards for data exchange. API standards, such as SQL/MM or JDM allow applications to function independent of the (usually just one) underlying data mining engine. This enables smaller software providers to reach a larger market. Exchange standards, such as PMML allow collaboration of more than one data mining software. Model creation and model deployment can be performed by different data mining software; models from different engines can easily be compared to find the best predictive model or to add other aspects, for instance in clustering models.

**Michael Berthold.** Ultimately people are creating a lot of sophisticated IP when producing advanced data processing and analysis workflows. They hate to see those tied to one proprietary platform -- plus they want to combine the best of several worlds, which makes it important to adhere to common, open standards. Reusability of proven analytical processes will become increasingly important as well.

**Erik Marcadé.** I am not sure that standards can be leveraged by scientists, except when there are freeware or open source projects that are fully compliant with these standards. Of course, this is very different for analysts and practical applications. The fact that it should be easy to use PMML to transfer a model from a development environment to a deployment environment is important, the fact that an application may switch data mining provider because they implement the same API lowers the risk for the integrators and is very important for the spread of data mining technology in operations.

**Rick Pechter.** Data mining efforts are often confined to a small group of individuals within an organization, usually those with statistical knowledge and access to specific tools. Limiting the power of predictive analysis to a select group in the data mining "silo" has the effect of isolating data mining from the rest of the organization. Open standards will help democratize data mining by allowing models to be more easily deployed to any person and any touch point in the enterprise.

**Mike Hoskins.** Open standards at the product metadata/model level (e.g.: PMML) give practitioners maximum power and freedom, allowing them to exchange definitions and rules between data mining products, both at design time and run time. This promotes crucial re-use, and minimizes the risk of vendor "lock-in" inherent in non-standard proprietary offerings. Open standards at the data format level (e.g.: XML, Weka ARFF, etc.) Allow for much easier exchange of data sets among tools and practitioners, and also allow the building up of re-usable front-end data conditioning and pre-processing routines.

**Wayne Thompson.** The core benefit is a standardized schema for sharing data mining models across vendors. A secondary benefit is

to foster distributed modeling and scoring mechanism for small and medium businesses. In a similar vein, it will encourage collaboration between researchers active in scientific knowledge discovery across geography and computing environments. PMML under DMG is a well guided and crafted standard.

**Rich Holada.** As with most standards, one of the primary benefits is future-proofing. By having a common representation for operationally important artifacts, practitioners have some flexibility to move models from one environment to another. PMML has its limits, however, as it only captures the trained model and does not yet fully represent the source data used to build the model and any specifics of the actual training process. As mentioned before, it is about faster deployment of intelligence into operational systems.

### **2.3 Please comment on your current and future support for PMML as the de-facto standard for model exchange.**

**Robert Grossman.** PMML is useful not only for moving statistical and data mining models between different systems, but also for creating repositories of models, and for long-term archiving of models. It has taken almost a decade, but over this time PMML has emerged as the dominant standard for data mining that allows models to be moved between different data mining and predictive modeling systems and applications.

**Christoph Lingenfelder.** IBM has been an active contributor to PMML for many years. Currently I represent IBM in the DMG standards meetings. IBM supports PMML version 3 with its Infosphere Warehouse product, both as a producer and as a consumer. Version 4 has just been published, adding, for example, model explanation functionality, which we currently support as an extension to PMML. The new version enables us to convert this to a standard format.

**Michael Berthold.** KNIME is already supporting PMML for the most prominent models (since v2.0), we plan to increase the coverage in the near future and are also trying to find ways to map the PMML preprocessing into KNIME. The latter is more of a challenge since not all of KNIME's preprocessing can be mapped onto PMML and the nature of a pipelining tool (possibly with parallel branches) also does not fit the current PMML standard all that well.

**Erik Marcadé.** KXEN is part of two standard committees, JDM (1 and 2) and PMML. KXEN is a producer of PMML models, with no extensions, and we have been doing so since PMML 2.0. KXEN is one of the many PMML producers, and is working with PMML consumers. It is interesting to see the shift through time of the industries implementing PMML consumers. It first begun with databases (KXEN never believed it would be a very interesting transfer since it is more efficient to generate directly SQL or UDF - User Defined Function - for in-database scoring. Most of databases vendors implemented their PMML interpreter as SQL generators.), to Business Intelligence vendors (the most innovative and compliant being today MicroStrategy), and will end its course with a distributed environment (such as the Zementis offer on Cloud infrastructures). Our effort in the PMML committee is nevertheless

focused on the compliance process, since PMML, being a semi-open standard - with an extension mechanism and without any formal compliance process - allows every one to claim being compliant with PMML without a formal process and regulation. This is a problem in practice, since someone who wants to work with a producer and a consumer has to check for actual compatibility of the two technology providers before making any move.

**Rick Pechter.** MicroStrategy has been an active contributor to the PMML standard for over five years and is the first major Business Intelligence company to offer vendor-independent deployment of data mining models. As a result, our customers have the flexibility to develop and deploy predictive models in any fashion they see fit. PMML allows MicroStrategy to create models on the fly, score models on the fly and refresh models with no need to re-work the reports, dashboards and scorecards showcasing predictive results. Along with our industry colleagues, we've helped improve features within the PMML spec and we are actively working to promote the standard through demonstrated inter-operability.

**Mike Hoskins.** We will be announcing our DataRush-based parallel data mining solution this summer. Shortly thereafter we expect to announce support for reading and writing PMML, both at design time (model import/export) and run time (model consumption and execution).

**Wayne Thompson.** SAS currently produces PMML 3.1 models. SAS plans to adopt PMML 4.0 and additional versions as ratified by DMG.

**Rich Holada.** PMML is extremely important - akin to SQL for the RDBMS market. Having an adhered to standard - even with proprietary extensions - facilitates an open exchange of tools and techniques. SPSS is fully committed and has been investing since the inception of the standard - that commitment will remain equally strong in the future, as it reflects SPSS philosophy about predictive analytics.

### **2.4 In your opinion, what will be the next disruptive technology which will expand the horizon for data mining algorithms and predictive analytics? What new opportunities will it bring and how will it change the way the community will apply algorithms in science or industry.**

**Robert Grossman.** I think cloud computing is the disruptive technology that will have the greatest impact on data mining and predictive modeling in the near term.

**Christoph Lingenfelder.** The accessibility of mining technologies to a larger group of people will see a new cycle of mining users and subsequent results and decision making processes. As this takes place, the maturity of this new user group and the decisions they make may significantly impact the direction of mining into the future. An important technology will be distributed mining, not to speed up the process, but to be able to tap larger amounts of data

whose owners are not willing or legally not allowed to share the details. Privacy-preserving data mining allows the investigation of distributed data stores without having to merge the data. As an example, consider patient data. Patients are reluctant to allow their data being sent out, but using distributed techniques, it may still be possible to detect commonalities in cancer cases.

**Michael Berthold.** I see two big trends: a) Not really "disruptive" but I strongly believe people will want to have much better integrated tools. Data Mining relies on nice data sitting around somewhere, which is hardly ever the case. People will want to have tools where they work on the same data view irrespective of their tasks (reporting, data mining, event detection...). b) I think data mining will move away from pure models but become more of an explorative analytics area: users need ways to find new insights in their data without being required to know the question before they start. Part of the exploration is to come up with new, interesting questions! Current tools really don't support this all that well - one needs to know well what kind of questions to ask and most tools are not very flexible when quickly trying out something different. KNIME is aiming to support such flexibility by offering visual assembly of the processing and analysis workflow as well as interactive views on the data. The user can quickly change something and instantly explore the effects.

**Erik Marcadé.** KXEN recently released a Social Network Analysis which can be seen as a new source of data in order to improve the results of predictive analytics. KXEN is working on even higher dimensional problems. Since KXEN customers are already making models with thousands of attributes, some of them are looking to higher dimensional space which requires some help from the database vendors. Working on robust algorithms on data streams is also a challenge we are interested in. But the real key for us has always been automation. Especially in the economy we live in, it seems that doing more (modeling) with less is key for the coming years. There is a tremendous growth opportunity ahead of us for automated predictive analytics.

**Rick Pechter.** Over the past decade, we have seen data mining features appear in applications typically not considered statistical tools. From spreadsheets to databases, desktops to laptops and even cell phones and PDAs, analytics usually limited by compute power and licensing are today being adopted and deployed to a wide variety of platforms. Open standards helps to commoditize the building blocks required for new and innovative solutions that include predictive analytics.

**Mike Hoskins.** Without doubt, the single most disruptive technology relevant to data mining is the advent of multi-core, and with it the availability of massively parallel, but shockingly inexpensive hardware. The good news is that we now have at our disposal this massive hardware parallelism - the bad news is that almost the entire data mining software industry (in fact, whole software industry) is unprepared to fully exploit this bounty. The vast majority of code is either single-threaded, or parallel in only very limited ways, e.g., simple data parallelism or loop unrolling. This challenge is also an opportunity. We expect to see new data mining products and technologies emerge that are massively parallel at the (fine-grained, multi-threaded) software level, and consequently able to fully exploit this commodity hardware and

enjoy hundreds or even thousands of times higher throughput and scalability than with traditional products and solutions.

Another very important trend is the introduction of altogether new computational models - ones that we feel will be very liberating for data mining. One of the frightening specters currently threatening data mining, particularly with today's massive data volumes, is the dependence on memory-resident computational models. The explosion in data volumes and dimensions is making this dead-end of memory-based computation painfully obvious. Much more powerful and efficient computational models are now at hand, for example Dataflow. Whether Dataflow is implemented in more coarse-grained, loosely-coupled, divide-and-conquer, one-pattern models like Hadoop's Map/Reduce, or in a much more fine-grained, multi-threaded, multi-pattern model like DataRush, you can implement classic data mining algorithms, and they will easily scale for any data volumes.

**Wayne Thompson.** SAS believes there are many opportunities for predictive analytics such as sensor data analysis, temporal data mining, social network analysis. There are new technology challenges such as distributed and parallel data mining, relational data mining, etc. Persona-based analytics is another area where SAS sees value in extending the adoption of data mining. Our vision as directed by Dr. James Goodnight is no business or research problem should be too large or complex for SAS to solve.

**Rich Holada.** We believe that there will be a resurgence in AI-techniques, particularly self-learning algorithms. Additionally, we are likely to see examples of "hybrid analysis" where multiple types of analyses are combined to provide a holistic approach to solving business problems, e.g., text mining and emotion analysis, speech interpretation and motion detection combined with sentiment information - along with the usual sources (transactional, attitudinal, demographics). Lastly, we believe we will see innovative techniques for data reduction and intelligent merging as means to address data volume growth and myriad sources. Another opportunity is our approach to distributed model building and incremental model building which has the potential to put predictive analytics much closer to points of interaction, for example directly on a set of routers in a cyber-security application, or directly on a set of nodes in a highly scalable database warehouse appliance. Stemming from that point - what may become very important in the near future is the emergence of powerful data federation capabilities that allow multi-source data to be collected, transformed, and pumped into modeling engines with very little work on the part of the analyst.

## 2.5 What value can cloud computing bring to data mining and predictive analytics?

**Robert Grossman.** Using clouds for data mining and predictive analytics offers several advantages. Cloud computing simplifies the scoring of data using statistical and data mining algorithms. If the model is expressed in PMML, then scoring can be done in cloud with any pre-installed PMML compliant scoring system. If the data is small enough, then (public) cloud-based environments can also be used for building statistical and data mining models.

As the data grows in size, moving it into clouds for modeling can be more of a challenge. In this case, private clouds that also host the data can be used for modeling or, alternatively, disks of data can be shipped to the cloud.

**Christoph Lingenfelder.** From a user perspective, cloud computing provides a means of acquiring computing services via the internet. From an organization perspective, cloud computing delivers services in a simplified way. Because of the sensitive nature of the data I expect that data mining will be restricted to private (intranet) clouds or to the analysis of publicly available data, certain research data or census data. Without the need to maintain a data mining software environment, a larger number of departments in a company will gain access to sound analytic methods. The method and discipline surrounding mining in the cloud will have a significant impact on the value the cloud will bring to mining and analytics, especially when critical decision making relies on it.

**Michael Berthold.** Cloud Computing makes it feasible for users to apply scoring/prediction methods to gigantic repositories. However, I think that cloud computing without databases on the cloud will limit its usefulness and restrict it to scoring rather than the real mining. But that's also a problem of data mining research in general, not a lot of attention is being paid to distributed mining and knowledge discovery algorithms so far.

**Erik Marcadé.** Working on the cloud, as Zementis has done, to promote new large scale deployment environments of models, especially targeting real time applications, such as the one found on the Web, is certainly a plus. I am more doubtful for the model training phase. Of course, using the cloud to train models will go faster, but I prefer to work on the algorithmic side to achieve scalability rather than brute force parallelization. We can go further this way and attack much bigger problems.

**Rick Pechter.** Cloud computing offers the potential to spread the use of advanced analytics without large up-front investments in hardware and software. One can already see innovative solutions that leverage the cloud for data collection, advanced analysis and low-cost access to supercomputer-class processor grids.

**Mike Hoskins.** Cloud computing brings multiple advantages to data mining and analytic applications. Users no longer have to buy and manage their own large data processing complexes; and you can start small, and grow incrementally, both from a resources and budget point of view. In addition, managing and scaling for extremely large datasets becomes easier, given the dynamic nature of resource allocation "in the cloud". Finally, sharing of data and best-of-breed applications becomes more realistic, since they are inherently more "open" as opposed to being "trapped" behind organizational firewalls.

**Wayne Thompson.** Lower implementation cost, ability to scale with demand, increased reliability, consolidation of data centers with intranet clouds, etc.

**Rich Holada.** Four primary benefits: 1) to attack problems on a much larger scale - an acknowledgement of the exponential growth of accumulated data; 2) a reduction in set-up time; 3) lower IT expenditures and capital investments; 4) a democratization of predictive analytics to wider audience made of SMBs. A pay-as-

you-go model – combined with open standards – will, we believe, place leading cloud vendors in the same place as telco operators - low consumer switching costs. We believe that those vendors with proprietary formats and techniques will wither and eventually disappear.

## 2.6 Are you exploring cloud computing or are you already offering cloud computing solutions to your clients?

**Robert Grossman.** The National Center for Data Mining has developed an open source cloud for data intensive computing called Sector. Sector includes a wide area storage cloud called the Sector Distributed File System (SDFS) and a programming framework in which you can process a Sector dataset with an arbitrary User Defined Functions (UDF) to produce a new Sector dataset. With Sector, it is very easy to code up statistical and data mining algorithms, even if the datasets are large or are distributed. Sector is available from <http://sector.sourceforge.net>

At Open Data Group, we have developed a Python based open source data mining system called Augustus that supports PMML and can run easily in Amazon-style cloud computing environments. Augustus is available from <http://www.sourceforge.net/projects/augustus>

**Christoph Lingenfelder.** IBM offers a large variety of infrastructure tools, solutions and services for cloud computing. There are, for example IBM Computing on Demand <http://www-03.ibm.com/systems/deepcomputing/cod/> and Remote data protection <http://www-935.ibm.com/services/us/index.wss/offering/bcrs/a1029249> to name just two. More information on the growing set of cloud offerings is available on <http://www.ibm.com/ibm/cloud/Developers> can get started on cloud computing on developerWorks, see <http://www.ibm.com/developerworks/spaces/cloud>

**Michael Berthold.** KNIME does not have immediate plans for cloud support. That is also a reason why we partner with Zementis who do put the interesting parts (predictive analytics) on the cloud already. See "values of open standards" above.

**Erik Marcadé.** Not at this point. Our clients have very classical IT infrastructures centered on a data warehouse in 95% of the time.

**Rick Pechter.** We are exploring the cloud computing model.

**Mike Hoskins.** We have been exploring cloud computing from two angles. First, by getting our DataRush product (SDK and Engine) up and running on both our "Pervasive DataCloud" where we already provide over 130 cloud-based Data Solutions to our customers, as well as the Amazon EC2 cloud. This latter project is quite far along, and we expect to make an announcement around DataRush availability on the Amazon Cloud later this year. Second, by building an interface between our DataRush massively parallel data engine, and the "cloud-ready" Hadoop/HBase infrastructure. So far the "fit" feels really good, since the ability of DataRush to fully exploit extreme fine-grained software parallelism on multicore

"fat nodes" nicely complements Hadoop, or any more coarse-grained distributed file system and workflow/dispatch model.

**Wayne Thompson.** SAS® OnDemand for Academics gives students and teachers easy access to SAS software including SAS Enterprise Miner over hosted SAS servers whenever, wherever they please. Essentially, users access the processing power of SAS over the Internet. Benefits include minimal software installation, configuration and support requirements. SAS also offers hosted solutions such as SAS® Solutions OnDemand: Drug Development. Additional solutions are planned for this year. SAS also is building a 70 million dollar cloud computing facility. Please see <http://www.sas.com/news/preleases/CCF2009.html> for more details.

**Rich Holada.** We are exploring that capability, as we strongly believe in its benefits. We are analyzing business and technical issues related to SPSS technologies in the Cloud. Scoring and delivery of analytics seem to be the natural first areas for us. Model building and data exploration is more challenging due to 1) the volume of data involved and 2) security concerns around data in the cloud.

## **2.7 Cloud computing provides more flexibility to leverage computing capacity, but it also facilitates various software licensing models: Software license, subscription model, Software as a Service (SaaS), open source: Which licensing model will become more important for the community and for software vendors?**

**Robert Grossman.** I don't have a good sense of which licensing models will become most important in the long term, except that I think there will always be room for *both* proprietary and open source software in data mining and predictive modeling, whether cloud-based or not.

**Christoph Lingenfelder.** This is a tough question, and without solid data material, it is not even possible to apply learning methods for predictive analysis. My personal expectation is that open source will become prevalent. This should not be confused with free software, however. The reason is that users of ever more complicated systems will only trust software for critical applications if it can, theoretically, be verified.

**Michael Berthold.** I think open source tools will grow but the final verdict on the best license has not yet been reached. I also think that SaaS is going to gain prominence, likely coupled with Open Source. You know what you are running but you do not want to host it yourself.

**Erik Marcadé.** Being in the predictive world for so many years, I know when it is not safe to predict. I make the distinction between two kinds of predictive analytics. There is the one that is taught for many years on university benches, talking about well know algorithms (such as logistic regression, decision trees, neural networks and the like). These algorithms are well known as well as their limitations and cannot be used as foundations for predictive

analytics automation. This part of predictive analytics is likely to be targeted with almost free licensing (all databases provide these algorithms nowadays for example, another good example is the power of the Excel 2007, seen as almost 'free' for most organizations - since they must have it anyway). Then, there is the other part of predictive analytics where new companies offer new solutions: difficult to say what will emerge from this: agility is the key word.

**Rick Pechter.** Cloud computing creates new licensing approaches beyond the traditional software models of the past. This flexibility should increase the market reach of predictive analysis by allowing incremental investments for new adopters and new business models for all vendors, including start-ups and established veterans alike. While many companies may be slow to change from traditional software acquisition approaches, increased flexibility should result in improved economies for customers and new market niches for vendors.

**Mike Hoskins.** As software moves to delivery "in the cloud", we feel the licensing models will be varied, but likely all have a "variable" component (as opposed to traditional perpetual models). Some likely candidates are time-based subscriptions, payment based on transaction counts, and payment based on data volumes which seems to align well with current cloud computing trends.

**Wayne Thompson.** SAS® is actively working with customers, and analysts to define the right licensing models for our various user groups.

**Rich Holada.** All those models are important for various reasons. Some organizations will always insist on traditional license models and we plan on continuing to offer that model; subscription for those customers looking to deploy mission-critical operations/activities. The SaaS business model for those customers needing processing power on an episodic basis can reflect the usage of the technology – either on a user/consumer basis, transaction basis, scoring basis, time-allocation basis, as the technologies allows various fine-grained consumption metering possibilities.

## **3. CONCLUSIONS**

It is refreshing to see that the panel members express a broad consensus on their support for open standards. This emphasizes that the overarching benefits are significant, not only for software vendors but even more so for the data mining community in general, driving the adoption of predictive algorithms across science and industry, and ultimately multiplying the opportunities for practitioners to apply their skills.

PMML has generally been accepted as the de-facto standard to exchange models between different applications and systems – not only vendor to vendor, but also for moving models from the development environment to production deployment. We expect an even broader vendor support in the near term, but also a significant adoption by users as they begin to reap in the benefits of PMML-based interoperability.

To ultimately succeed and accelerate user adoption, however, vendors must improve PMML cross-vendor compliance and invest in educating the industry about the various benefits of the standard.

Finally, cloud computing [2] will allow us to lower the cost for data mining and provide a roadmap for data mining models to take their place in new applications and industries. The cloud will become a conduit for software and service offerings, making deployment and execution faster and easier, through minimizing the common IT overhead on one side or by providing unprecedented scalability in other cases.

As algorithms for statistical analysis and data mining become accepted best-practices across various industries [4], their real-time execution as part of an intelligent Enterprise Decision Management [5] strategy will usher in a new area of smart business applications.

#### 4. ACKNOWLEDGEMENTS

The panel organizer would like to thank all panel members for their active participation and the KDD committee for allowing us the opportunity to bring these topics to the attention of the data mining community. It is our privilege to speak for many of our colleagues in industry and science who commit their valuable time and energy crafting and promoting open standards for data mining.

#### 5. REFERENCES

- [1] Data Mining Group Website. <http://www.dmg.org>
- [2] N. Carr. The Big Switch: Rewiring the World, from Edison to Google. W. W. Norton and Company, Inc., 2007.
- [3] R. Grossman and Y. Gu. Data mining using high performance data clouds: Experimental studies using sector and sphere. In Proceedings of the 14<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008). ACM, 2008.
- [4] R. Nisbet, J. Elder, and G. Miner. Handbook of Statistical Analysis and Data Mining Applications. Academic Press, 2009.
- [5] J. Taylor and N. Raden. Smart Enough Systems: How to Deliver Competitive Advantage by Automatic Hidden Decisions. Prentice Hall, 2007.

#### 6. ABOUT THE PANELISTS

**Robert Grossman.** Robert Grossman is the Managing Partner of Open Data Group, which he founded in 2002. Open Data Group provides management consulting and outsourced services focused on analytics.

Grossman is also the Director of the Laboratory for Advanced Computing (LAC) and the National Center for Data Mining (NCDM) at the University of Illinois at Chicago (UIC). The Lab and Center perform research, sponsor standards, manage an international data mining testbed, and engage in outreach activities in the areas of data mining and data intensive computing.

Robert Grossman became a faculty member at the University of Illinois at Chicago in 1988 and is a Professor of Mathematics, Statistics, and Computer Science. From 1984-1988 he was a faculty member at the University of California at Berkeley. He received a Ph.D. from Princeton in 1985 and a B.A. from Harvard in 1980. He has had a half time appointment at UIC since 1996.

He is also an Associate Senior Fellow at the Institute for Genomics and Systems Biology at the University of Chicago.

Prior to founding the Open Data Group, he founded Magnify, Inc. in 1996. Magnify provides data mining solutions to the insurance industry. Grossman was Magnify's CEO until 2001 and its Chairman until it was sold to ChoicePoint in 2005.

He has published over 150 papers in refereed journals and proceedings and edited six books on data mining, data intensive computing, Internet technologies, high performance computing, high performance networking, business intelligence, e-business, and related areas.

<http://www.opendatagroup.com/>

**Christoph Lingenfelder.** Dr. Christoph Lingenfelder is part of the development team for IBM's data mining software at IBM Germany Research & Development in Böblingen. He just returned from a research assignment at the IBM Watson Research Center in Yorktown Heights, New York, doing research in the field of mathematical data analysis and data mining. Christoph represents IBM in a number of standardization groups (DIN, ISO, OMG, JCP, and DMG) with database and data mining topics. As editor for ISO SC32, he published the first version of the data mining standard in SQL/MM (ISO/IEC 13249-6). He also participates in the development of PMML in the Data Mining Group. Since the first version in 1996, Christoph has been involved in the development of Intelligent Miner for Data. Originally he was responsible for the development of a clustering algorithm and a numeric prediction algorithm (RBF). Previously, he did research in the area of knowledge representation at the IBM Scientific Center in Heidelberg, Germany. Christoph studied physics and mathematics at the universities of Karlsruhe, Germany, and Maryland, USA. Subsequently he got a PhD in artificial intelligence from the University of Kaiserslautern, Germany.

<http://www.ibm.com>

**Michael Berthold.** Michael R. Berthold is currently the Nycomed-Professor for Bioinformatics and Information Mining at Konstanz University and CEO of KNIME.com GmbH, Zurich. He spent many years in the US at academic and industrial research institutions and has published extensively in the areas of machine learning methods for the explorative analysis of large, heterogeneous information repositories. He is one of the creators of the open source data analysis pipeline platform KNIME.

<http://www.knime.org>

**Erik Marcadé.** With over 17 years of experience in the neural network industry, Erik Marcade, founder and chief technical officer for KXEN, is responsible for software development and information technologies. Prior to founding KXEN, Mr. Marcade developed real-time software expertise at Cadence Design Systems, accountable for advancing real-time software systems as well as managing "system-on-a-chip" projects. Before joining Cadence, Mr. Marcade spearheaded a project to restructure the marketing database of the largest French automobile manufacturer for Atos, a leading European information technology services company. In 1990, Mr. Marcade co-founded Mimetics, a French company that processes and sells development environment, optical character recognition (OCR) products and services using neural network technology. Prior to Mimetics, Mr. Marcade joined Thomson-CSF Weapon System Division as a software engineer and project manager.

working on the application of artificial intelligence for projects in weapons allocation, target detection and tracking, geo-strategic assessment, and software quality control. He contributed to the creation of Thomson Research Laboratories in Palo Alto, CA (Pacific Rim Operation—PRO) as senior software engineer. There he collaborated with Stanford University on the automatic landing and flare system for Boeing, and Kestrel Institute, a non-profit computer science research organization. He returned to France to head Esprit projects on neural networks development.

Mr. Marcade holds an engineering degree from Ecole de l'Aeronautique et de l'Espace, specializing in process control, signal processing, computer science, and artificial intelligence.

<http://www.kxen.com>

**Rick Pechter.** Rick Pechter is Senior Director of MicroStrategy's Pacific Technology Center in Carlsbad, CA. Rick and his team have responsibility for several key portions of MicroStrategy's Business Intelligence platform including MicroStrategy Office, MicroStrategy Web Services and MicroStrategy Data Mining Services. He has over twenty years experience in the computer and data processing industries, with degrees in Electrical Engineering (BSEE, UC Irvine) and Engineering Management (MSEM, National Technological University).

<http://www.microstrategy.com/>

**Mike Hoskins.** Mike Hoskins brings more than 20 years of experience in developing and managing software companies to his roles as Chief Technology Officer of Pervasive Software and General Manager of Pervasive's Integration business unit. Mike joined Pervasive through its December 2003 acquisition of Data Junction Corporation, where Mike served as president for the prior 15 years.

As Chief Technology Officer, Mike champions Pervasive's technology direction, evangelizes the company's industry-leading low-TCO approach to data management and integration, and directs the company's Innovation Labs. Mike received the AITP Austin chapter's 2007 Information Technologist of the Year Award for his leadership in the development of Pervasive DataRush™. In his role as General Manager, Mike is responsible for the growth of the Integration business unit and its strategic leadership position in agile, embeddable integration.

Graduating summa cum laude in 1977, Mike recently received the Distinguished Alumni Award from Bowling Green State University in Ohio. In 1982 he formed SaudiSoft, a leading software company in the Middle East. Mike is a widely respected thinker in the integration industry, has been featured in a variety of publications, and speaks worldwide on innovations in data management and integration. Topics which he has explored from an industry perspective include Software as a Service, ROI in the integration space, the last mile of integration, integration out of the box and the multi-core revolution.

<http://www.pervasive.com/>

**Wayne Thompson.** Wayne Thompson is an Analytics Product Manager at SAS. His primary efforts are centered around bringing relevant product/solution feedback from customers to the SAS

Enterprise Miner and SAS Model Manager development teams to extend SAS® Institute's leadership position in the data mining market. He has been employed at SAS® since 1992. During his tenure at SAS, Wayne also served as a Statistical Services specialists for the Education Division in which he developed and taught applied statistical courses as well as collaborated on several data analysis projects for clients from many industries including pharmaceuticals, financial services, health care, telecommunications, and insurance. Wayne has over 16 years of experience in CRM, eCRM, data mining, database marketing and strategic planning working with a broad range of customers. Wayne received his Ph.D. and M.S from the University of Tennessee in 1992 and 1987, respectively. During his PhD program, he was also a visiting scientist at the Institut Supérieur d'Agriculture de Lille, Lille, France where he taught applied data analysis courses and assisted students during with their thesis research.

<http://www.sas.com>

**Rich Holada.** As Senior Vice President of Technology, Rich Holada's role is to drive and execute the product and technology vision of SPSS in our products and for our customers. Holada joined SPSS in November 2006 as Senior Vice President of Research & Development. Holada brings with him nearly 20 years of deep software research and development background and diverse industry experience in the technology industry. Previously, he was Vice President of Industry Development at Oracle Corporation, where he earlier held the post of Vice President of CRM Development. Holada has also held senior research and development positions at PeopleSoft, Inc., Trimark Technologies, Inc., and Intelligent Trading Systems, Inc., as well as earlier systems engineering positions at Sun Microsystems, Inc.

<http://www.spss.com>

## 7. ABOUT THE ORGANIZER

**Michael Zeller.** Dr. Zeller is currently the CEO of Zementis, a software company focused on predictive analytics and advanced decisioning technology. His mission is to combine science and software to create superior business and industrial solutions that leverage predictive models and rules in real-time.

Previously, Dr. Zeller served as CEO of OTW Software, a company focused on implementing structured software engineering processes and delivering object-oriented analysis and design services. Prior to his engagement at OTW, he held the position of Director of Engineering for an aerospace firm, managing the implementation of IT solutions for major aerospace corporations.

Dr. Zeller received a Ph.D. in Physics from the University of Frankfurt (Germany), with emphasis in the development of neural networks, robotics, and human-computer intelligent interaction. He received a visiting scholarship from the University of Illinois at Urbana-Champaign and was the recipient of a Presidential Postdoctoral Fellowship from the Computer Science Department at the University of Southern California.

<http://www.zementis.com>