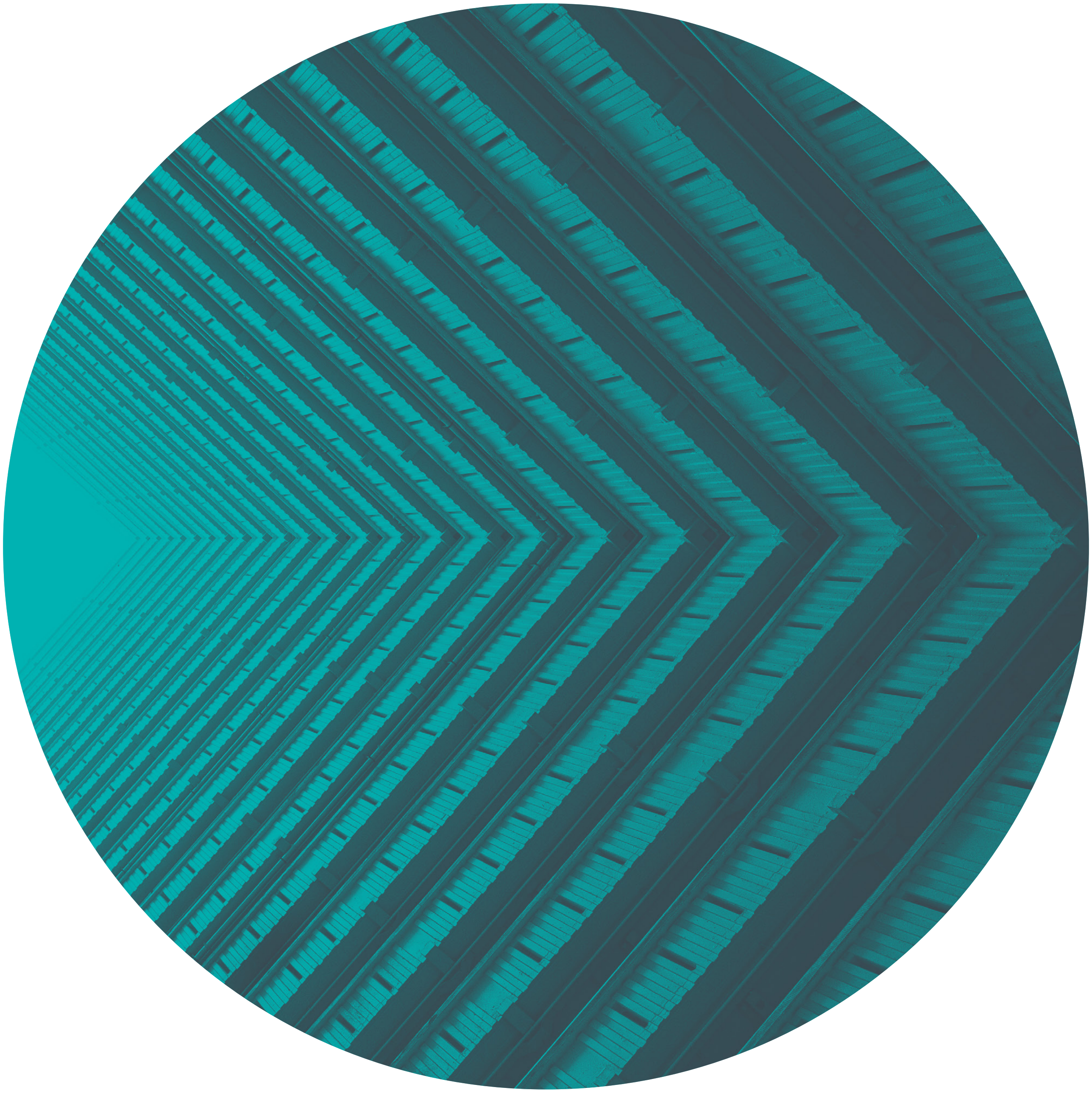
WHITE PAPER

Bring Your Own Model

A new approach to successfully building and deploying predictive analytics at scale



By Dr. Chris Hillman, Data Science Director, Teradata
and Alexander Smirnov, Principal Data Scientist, Teradata

11.20 / DATA ANALYTICS / WHITE PAPER

teradata.

Table of Contents

Executive summary

The role of model training in creating machine learning, AI and predictive analytics strategies for organisations can sometimes be over emphasised. The Teradata Analytics 123 strategy ensures that sufficient focus is placed on the crucial feature engineering, data management and deployment aspects of a successful data analytics process. These are parts 1 and 3 of the Analytics 123 approach. However, the second step remains central to the whole process. It is undeniable that the building of performant models is the essence of data driven businesses and the core purpose of the data science team. This paper focuses on the challenges faced at this point, and outlines Teradata's innovative Bring Your Own Model approach.

At the heart of the challenge is the need to balance the desire of data scientists to use tools, language and approaches that they are skilled in and feel are the best fit for a specific task, and the need for businesses to manage the dependencies created by using many different tools. Friction between organisations keen to standardise on one modelling approach and environment and individual data scientists more familiar, and more expert in another, is common. Different tools are better suited to different projects and data scientists are understandably keen to use those they feel are the best fit. Deep experience with a particular language can also increase productivity making the decision to enforce the use of another harder to justify. Yet, supporting every preferred tool and approach coupled with frequent version updates of opensource software, can build significant obstacles into the journey from building and testing a model to scoring and deploying it for use in the business. Creating and supporting numerous different environments, each used to score a handful of models is neither scalable nor commercially viable. The common practice of moving entire data sets out of live production databases for scoring in specific tool-centric modelling environments creates significant cost and time barriers. Often these prove insurmountable and effectively kill off a model before it gets deployed. This certainly contributes to the 65% of predictive analytics projects that Gartner suggests never make it out of the lab.

By moving models to score in database, rather than moving databases to be scored in the modelling environment, BYOM makes deployment easier, faster and more cost effective

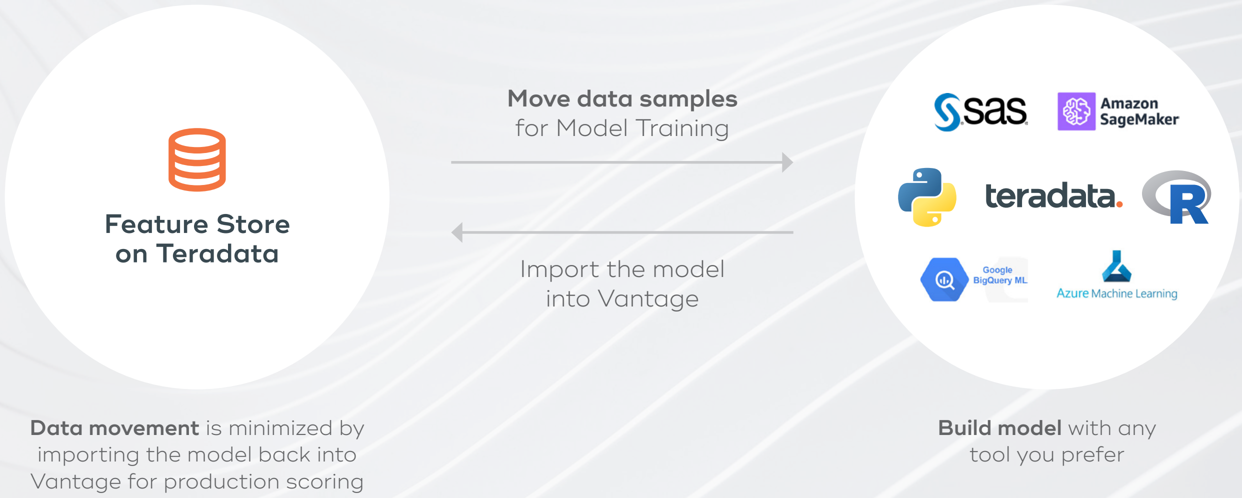
Teradata's Bring Your Own Model (BYOM) approach offers a solution to all these challenges. By changing the approach and moving models to score in database, rather than moving databases to be scored in the modelling environment, BYOM makes deployment easier, faster and more cost effective. It changes the relationship between data science teams and the business, prevents wasteful, pointless and risky movement of vast quantities of data, and gives all parties the flexibility to work productively and at their best.

Advantages of the Teradata BYOM approach

- Data science teams can select and use the tools that they prefer, and which they feel are best suited to the task at hand.
- They can use any environment including cloud-based resources to build and train models.
- Using standard libraries, they can convert models to score in database – breaking dependencies between modelling and deployment code support.
- Bring Your Own Model to the data massively reduces the need for data movements saving time and cost.
- Scoring in database makes it easier to embed results into business processes to drive value.
- Models move more easily from experiment to business value generators.
- As a result, data science teams are better connected (and recognised) for business outcomes.

Train a model with any tool

Allowing the Data Science freedom to use their preferred tools



Feature Store on Teradata

Move data samples for Model Training

Import the model into Vantage

Data movement is minimized by importing the model back into Vantage for production scoring

Build model with any tool you prefer

Breaking dependencies between models and production code

Organisations seeking to put data analytics at the heart of their business often face challenges in agreeing standardized approaches to model creation. Data scientists want the flexibility to use a wide range of tools for model creation, but IT and business audiences do not want to support a myriad of languages in order to have those models work in production environments. Breaking the dependency between modelling languages and production is essential to improve the percentage of 'experiments' that lead to real value across the business.

Use your choice of language

Typically, the approach has been to enforce a defined set of supported languages, but this can be a barrier to innovation and limit the performance of data scientists. Most have a favourite among the commonly used coding languages like Python, R, SAS, Knime and SparkML and others. They have deep experience and skills with these languages developed over several years and naturally feel most comfortable with their chosen tool. But data scientists also want the flexibility to use whichever approach is most suited to the task in hand. Polyglot programming is embraced by the data science community as the most effective approach to deliver the diverse predictive analytics outcomes demanded by today's businesses. Efforts to restrict this can not only cause conflict but lead to 'shadow' IT environments as teams create environments and workarounds in order to use their favourite tools and deliver the models the business needs.

Ultimately the best tool is the one you know how to use, and it is likely that data scientists will be more efficient and deliver better models using the tools which they are familiar. Plus, different languages are better in different circumstances; selecting those that are well matched to the task in hand is part of the data scientist's expertise. However, this can lead to 'pipeline jungles'[1] of code that are difficult to document and manage and undermine the agility of the organisation and its flexibility to change to meet fast-evolving requirements.

Teradata's BYOM approach in conjunction with an Enterprise Feature Store solves this conundrum. Rather than trying to impose uniformity on data scientists' choice of how to work, the Teradata approach advocates diversity and equality of tools. The Teradata Bring Your Own Model concept does not impose any requirements on which model training tool should be used but instead, provides a simple way for any model to be moved into production irrespective of approach and language used in its creation.

The Teradata approach advocates diversity and equality of tools rather than trying to impose uniformity on data scientists' choice of how they work

Using pre-packed, standard, proven data features available in an Enterprise Feature Store[2] data scientists can quickly and easily pull sample data into whichever tool they prefer, including the latest versions and most recently updated libraries. Then, once the model is tested and proven, it can be seamlessly imported into the production system to score live production data without creating any dependencies for its native code to be present.

Pluggable Models

Teradata has developed in-database functionality to realise the concept of 'pluggable models'. Essentially it gives data scientists the freedom to develop models in whichever language they prefer, and then simply 'plug' them into Teradata for scoring. There are three different ways in which this can be done, each supporting simple and seamless integration of models and production data.

Predictive Model Mark-up Language (PMML) has been developed by the Data Mining Group, a consortium of over 30 leading players, including Teradata, to simplify

[1] https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

[2] https://assets.teradata.com/resourceCenter/downloads/WhitePapers/Efficiency-Productivity-and-Speed-to-Deployment-MD007074.pdf

the deployment of analytics models. Using PMML data scientists can quickly and easily interpret a wide range of models for scoring in production environments. PMML is an established and proven standard serialisation format that interprets leading machine learning platforms including Python, R, SparkML, H2O, SAS, SPSS, Knime, XGBoost, CatBoost, LightGBM among hundreds of others. It also supports many of the most common and useful model types including: Tree, Naïve Bayes, Support Vector Machine, Regression, Time Series, Clustering, Bayesian Network and Neural Network models among others.

To implement PMML to integrate a tested model for scoring with live data in a production system, data scientists simply use one of the freely available PMML exporter libraries to get their model into PMML format and then import the model into the database as a BLOB field.

Models can also be converted to SQL to score data in production. Translating the modelling language (R, Python etc) into SQL means the model can run in database and is no-longer dependent on the tools and libraries used in its creation. Teradata Vantage provides automatic SQL statement generation from multiple popular tools so that data scientists can simply optimise their models for scoring in Teradata data warehouses.

The small minority of models that cannot be converted using either PMML or SQL can still be used in-database using bespoke code. Although a more complex process, and one that still requires source languages and libraries to be installed in the production environment, it does mean that even in these unusual cases models developed in virtually any language can be moved and deployed in database to deliver real results to the business.

Don't fight data gravity, take your model to the data

Not only does this approach reduce friction throughout the data analytics value chain, and allow data scientists to deploy their best and most favoured approaches to model building and testing but it drastically reduces the need to move data back and forth between test and production systems.

Current practices see data scientists build their own silos of data, often on a model by model basis, by taking data from source systems and producing a pipeline that does the data wrangling and feature engineering in a series of steps. For model training, the first optimisation that Teradata recommend is the use of an Enterprise Feature Store to make the data pipeline process far more efficient by elimating duplication and redundancy. This means that for predictive model testing and training, samples of data can be transferred out of the EFS into an external environment with a minimum of effort while retaining control over data governance.

Avoid pointless data movements

When it comes to scoring the actual data, many chose to move the entire database into a production version of the same environment used for model training. This creates a whole host of issues. Offloading, preparing and scoring data in the modelling system before reloading the results into the business systems for use involves significant time and resources. One customer has told Teradata that it used to spend 3 hours offloading data, 20 minutes scoring it, and then almost as much time again reloading the outputs back into the main database.

Network bandwidth constraints can make this a lengthy process and deprive other systems of access to resources. Systems that were fast, cost effective and performant with small subsets of data quickly become slow and expensive when scoring entire data sets. Shifting large amounts of data into modelling environments for scoring can quickly inflate initially low costs for cloud storage and compute. To maintain performance, it may be necessary to spin up different instances and duplicate data across them to score. The cost per query increases rapidly, response times slow and outputs are harder to reintegrate to provide insights to the business. This is costly, time-consuming and hinder usability of outputs even for single models. However, leading business already deploy hundreds of models, many of which are scoring entire databases on a daily or even more frequent basis. The business of the future will deploy thousands of models daily – clearly the current approach is not a scalable solution.

Moving all the data to every model every time makes no sense, it simply adds cost and time with no added value. Faced with these hurdles, plus the 'refactoring' process often implemented by engineering teams as they take the data scientists' models and re-write them before they can be used on production systems, it is hardly surprising that only 20% of data analytics projects make it through to deliver a business outcome.[3]

Moving models to the data is more efficient and will lead not only to faster and lower-cost queries, but also puts outputs closer to the business systems that can use them. BYOM incorporates the concept of Data Gravity; keeping data where it is and moving applications and models to it. Importing models to score records in database removes all this wasted effort. It also prevents duplication and the risk of pollution of crucial data. Models are still built and trained on a subset of data extracted from the database, but then it is the model that moves to the database for scoring and production.

Real-time scoring in Teradata

The BYOM approach manages the whole process, freeing data scientists to focus on creating the best possible models. They can leverage the best tools and cost-effective compute and storage to test models knowing that they can take them to the main data for scoring. The advantages go beyond speed and cost. Scoring in a Teradata database means incredible scalability. Data sets with billions of rows can be quickly and cost effectively scored in a batch and because that data is still within the database it is guaranteed to be up-to-date and accurate. Teradata's always-parallel-all-the-time architecture also makes it an ideal platform for scoring predictive models at hyper-scale and in near real-time. Customer feedback suggests that moving the model to score in database, rather than shifting entire databases into modelling environments for scoring reduces costs by a factor of between five and ten times!

Critically the results of the model are produced in the same environment as the business processes that can use them. Data analytics are only of value if they can be leveraged to create beneficial business outcomes. The outputs of the model matter, but so does the ease in which they can be used. So, for example, it is useful to create a model that can predict when a specific part in a complex piece of machinery like a locomotive engine, will fail. It can give engineers and customer service representatives huge insights. But it is far more valuable if that insight sits in the same systems that manage inventory, scheduling and product tracking so that a whole business process can be created to actually service the component cost effectively and just in time. Solving complex business issues rather than narrow data science problems is the ultimate benefit of performing data analysis in the Teradata database. Seamlessly embedding predictive models into business processes makes them much more likely to succeed and be used by the business to create value.

Solve complex business issues rather than narrow data science problems by using BYOM to score in database

[3] https://blogs.gartner.com/andrew_white/2019/01/03/our-top-data-and-analytics-predicts-for-2019/

From science-fair to business value

This is the ultimate value of the BYOM approach. It dissolves the barriers between data science teams and the business consumers of their work. It simultaneously allows data scientists to concentrate on what they do best – building performant and proven models that use data analytics to predict outcomes – and gives them a simple way to prepare for effective deployment of those models into the business.

Data scientists face two main issues in building models that solve business problems. The first is getting access to data simply and quickly – reducing as much as possible the ‘data wrangling’ necessary to find, assemble and engineer the data they need to build models. The Enterprise Feature Store concept handles this – providing proven, reusable features directly from databases. The other challenge sits at the other end of the data scientist’s role – taking proven models into production, scaling a model from test data of a few hundred rows to score live data that could run to millions in not billions of rows.

This is where most data science experiments fall-down. It is either too difficult or too expensive to deploy models at scale. Many of the reasons have already been articulated – the dependencies created between code used to model the code in production systems; the disagreements over which languages and libraries to support to do this. The problems of scale and speed when models are scored against complete data sets and the friction and cost of moving large data sets to modelling environments for scoring. Add to this the additional demands of governance, security and process demanded by IT and risk officer and it is no wonder that many data scientists see this last, vital step as a significant hurdle.

The BYOM approach minimises these issues for the data scientist, allowing them to concentrate on building models to solve business problems. Customers typically host data from live applications in Teradata already. Moving models to Teradata Vantage for scoring not only ensures scalability, speed and performance, but creates direct connections between the model, its outputs and the core business processes that keep organisations running. BYOM puts predictive analytics at the heart of the business process – without creating extra work for the data scientist.

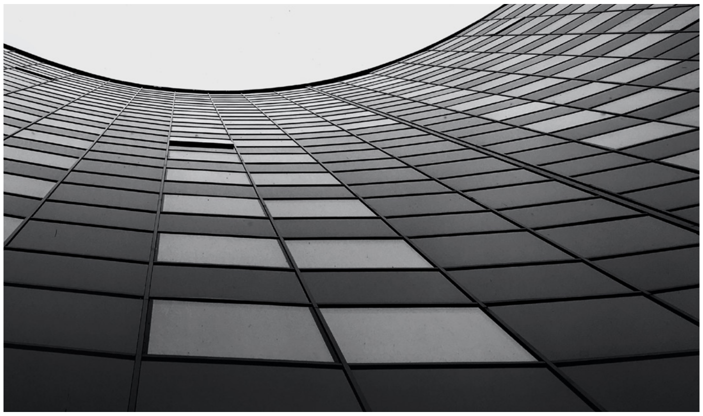
BYOM puts predictive analytics at the heart of the business process – without creating extra work for the data scientist

One model for batch and real-time in database scoring

The process of bringing the model to the data and translating it from the powerful modelling languages preferred by data scientists to the SQL code running fundamental business processes integrates models fully into the database. Once in SQL the same models can be applied to both batch and real-time scoring. So, for example, instead of building, testing and supporting different models to batch score customers eligible for a credit offer, and to real-time score applicants for those product, the same model can be used to accomplish both.

Teradata’s massively parallel architecture and O(1) access to localized data, enables extremely high throughput and low-latency for tactical queries, for both batch and near real-time model scoring. This means that a model can score billions of rows of data quickly and cost effectively to support a batch process. The same architecture allows a query analysing hundreds of data points related to a single customer to be calculated in near-real time which opens up the capability to score models on demand. The same model can be used in both instances. The architecture of Teradata also ensures that mission-critical operational workloads (like near real-time model scoring) can co-exist with complex, resource-intensive processing (including data labs, data preparation and model training).

BYOM can create a far more flexible and platform-independent approach to model building. Data scientists can use the tools they want, on whatever platforms and environments they want to get the best models designed, built and tested as fast as possible. But they are not 'locked in' to any particular tool, vendor or environment, Different individuals and different teams can even use different environments to create their own models. Once it comes to scoring, all those models can be moved to score data in the database irrespective of where and how they were created. Teradata can translate models built in the cloud, in proprietary environments or on individual servers to provide the near-real time performance and massive scalability needed to deliver value at enterprise scale. BYOM creates the freedom to use the best combinations of tools and environments available, whilst avoiding vendor lock-in and ensuring easy deployment to production systems no matter the size of live data sets.



BYOM, at the centre of the Analytics 123 strategy

Teradata's BYOM approach is a key aspect of its Analytics 123 vision. It closely connects to the feature preparation and management delivered by the Enterprise Feature Store massively reducing the overhead required to find and engineer the data to build models. BYOM facilitates a more flexible, fast and effective approach to data model building. Not only does it give individuals and data science teams the flexibility to use the tools they prefer, but it streamlines the path from testing to scoring and deployment. In database scoring eliminates the need to move vast quantities of data in and out of test and modelling environment – improving performance, lowering costs and putting the results exactly where they can be most effectively used by the business to generate real value.

As such BYOM represents more than a technical approach. It can foster a change in mindset across organizations and promote a deeper connection between data science and business processes. Many businesses have invested in data science teams and the tools and environments they need to create predictive, machine learning and AI models. But few have seen significant ROI from these efforts. The technical challenges, cost and time required to move from a great model to using outputs from scoring live data with it prevent the vast majority from ever being used.

BYOM – Freedom to develop using any tool and productionize with Teradata

Radical architectural simplification, performance at-scale, reduced TCO

Convert models to SQL queries

EMEA Bank: round-trip credit scoring in < 2s running 15 jobs in parallel.

Use language-specific model format with in-DBMS interpreters

R python

Global Bank: complex income estimation models scored for 7M customers in < 23m (previous approach would not scale past 20k rest records).

Use common serialization format and IVSM accelerator

PMML Predictive Model Markup Language

Asian Lottery: estimated 10x reduction in model scoring costs on Teradata compared with current Databricks-based approach.

New goals

Bring Your Own Model subtly but importantly shifts the endpoint of data science teams involvement in creating valuable predictive analytics for the business. Rather than using the models they created to score data and then pass the outputs back to the business BYOM allows data scientists to move their models and score in database as part of a business process. The scored results are now automatically already in the systems where they can be used to influence, integrate and deliver value alongside the main production systems of the business. Data scientists can see directly and clearly how the fruits of their work have an impact on the business – creating a stronger connection between the models and the business value they create.

Bridging this gap, and finding ways to more deeply embed data science teams into the business has been one of the most significant challenges as businesses look to put predicative analytics, machine learning and AI at the heart of their businesses. Few in the business understand the complexities and challenges of data science and so those that are experts often feel remote and cut off from the rest of the business. With BYOM they can more easily point to the outputs of their models as they drive business process improvement. Whilst few in the business will understand how the model was built, they will be able to see and understand how it is working in practice. Understanding, and crucially trust, in the output of predictive model among business owners is a big contributor to the success of those models in a production environment

BYOM provides data scientists with a highly productive and effective, streamlined process for model building, testing and scoring. They can use the best-fit tools for the job and avoid the bottlenecks in scoring entire databases. But the job is not done when the model scores its first data set. Businesses have hundreds if not thousands of models, scoring multiple databases at a whole range of frequencies. Keeping track of all of these models, tuning and ensuring they remain performant, is just as important as adding new models. It is is also vital that models are tested against challenger models so that the most accurate and effective continue to be used whilst ensuring that lessons, code and insights are not lost or locked away in redundant models. Organisations must guard against a build-up of technical debt as different teams build models in isolation unaware of potential overlaps and dependencies. Finally, models must be retired when no longer effective.

Model lifecycle management is emerging as the new unseen challenge for the business of tomorrow, and one that could occupy significant time and resource in data science teams. The final element of the Teradata Analytics 123 approach is to automate and manage as much of this lifecycle as possible. The next white paper outlines its approach.

About Teradata

With all the investments made in analytics, it's time to stop buying into partial solutions that overpromise and underdeliver. It's time to invest in answers. Only Teradata leverages all of the data, all of the time, so you can analyze anything, deploy anywhere, and deliver analytics that matter most to your business. And we do it on-premises, in the cloud, or anywhere in between. We call this pervasive data intelligence. It's the answer to the complexity, cost and inadequacy of today's analytics. And how we transform how businesses work and people live through the power of data. Get the answer at **Teradata.com**.

About the Authors

Dr. Chris Hillman is the Director of Data Science for the EMEA Advanced Analytics team.

Alexander Smirnov is a Moscow-based Principal Data Scientist for the EMEA Advanced Analytics team.

17095 Via Del Campo, San Diego, CA 92127   Teradata.com