VOLUME 25 NUMBER 2

Business Intelligence Journal

THE LEADING PUBLICATION FOR DATA MANAGEMENT AND ANALYTICS

Overcoming BI Self-Service Obstacles During and After a Pandemic

Bhupesh Malhotra and Dan Columbus

The Case for Data Transparency: Four Key Areas Where Deeper Visibility Is Critical

Doug Kimball

The Need for Explainable Processes and Algorithms

Hugh Watson

tdwi Transforming Data With Intelligence™

CERTIFIED BUSINESS INTELLIGENCE PROFESSIONAL

TDWI CERTIFICATION

Get Recognized as an Industry Leader

Advance your career with CBIP

"Professionals holding a TDWI CBIP certification command an average salary of $130,518—$19,576 higher than the average for non-certified professionals."

2018 TDWI Salary, Roles, and Responsibilities Report

cbip
Certified Business Intelligence Professional

Distinguishing yourself in your career can be difficult yet rewarding. Let your résumé show you have the powerful combination of experience and education that comes from the BI, DW, and analytics industry's most meaningful and credible certification program.

Become a Certified Business Intelligence Professional today! Find out how to advance your career with a BI certification credential from TDWI. Take the first step: visit tdwi.org/cbip.

tdwi.org/cbip

tdwi

VOLUME 25 • NUMBER 2

Business Intelligence Journal

THE LEADING PUBLICATION FOR DATA MANAGEMENT AND ANALYTICS

tdwi ONLINE LEARNING

ONLINE COURSES FROM TDWI

Did you know that TDWI's industry-leading curriculum is accessible through interactive, online courses? TDWI offers individual courses and bundled course packages, all with flexible access that allows you to learn when and where you are most comfortable. Earn a certificate of completion plus a LinkedIn badge at the end of each course!

Deepen your understanding of analytics and data management through interactive courses taught by industry experts.

Want Maximum Value?

Check out our new multi-week courses that combine the benefits of online learning with the added value of deadline-driven assignments and one-on-one instructor interaction.

We also offer an annual premium subscription for unlimited access for those looking to stay on top of our new courses as soon as they are released.

Visit go.tdwi.org/online-learning to sign up for course updates and special offers.

TDWI offers a wide variety of online courses on:

- Business Intelligence
- CBIP Exam Prep
- Data Science
- Data Analytics
- Data Management
- Data Governance and Quality

Extend your learning with self-directed online courses and

SAVE 30%*

WITH COUPON CODE

JOURNAL30

*Valid for courses only. Certain exclusions apply. Expires 12/31/2020

tdwi.org/online-learning

From the Editor

What's the secret to getting the most from your enterprise data? From data models to AI algorithms, data lineage to data governance, users have to feel confident that the data is collected, created, and cleaned properly.

Senior editor Hugh Watson focuses on how explainable AI (XAI) and predictive models build trust in data. In addition to exploring technology options, Watson looks at four steps enterprises must take to ensure that analytics applications meet the demands of management, laws, and regulators as well as calls from the public for greater explainability.

Another technique is explored by our experts John O'Brien, James Taylor, and Coy Yonce: ModelOps—automating model building and implementing and updating processes. They offer suggestions for scaling analytics in terms of people, data, models, technologies, and processes.

Doug Kimball explores data transparency, offering four key areas where enterprises must provide deeper visibility, driven in part by a host of new regulations. He offers tips for leveraging data management technology for the transparency businesses and their customers demand.

Once you trust your data, how can you use it to make truly data-driven decisions? Gaurav Anand looks at how the lack of a framework for viewing data and generating insights impedes decision making. He proposes a business management system to help enterprises overcome the challenges.

Putting analytics in the hands of business users is at the heart of the self-service BI movement, and in our cover story, Bhupesh Malhotra and Dan Columbus describe how your enterprise can overcome BI self-service obstacles during and after a pandemic.

AI and ML can also help users develop insights faster. Arup Duttaroy details the machine learning tasks enterprises can automate to realize error-proof productivity gains and promote democratization of data science.

Of course, data never stands still, and neither do data sources. Eric J. Peters explains how a DW/BI development team can architect new source systems into a data warehouse so users can produce the reports and make the decisions they need quickly. He offers a framework to keep end users nimble while maintaining the integrity and durability of the enterprise data warehouse.

We welcome your thoughts about this issue; write me at jpowell@tdwi.org.

James E. Powell

Business Intelligence Journal

THE LEADING PUBLICATION FOR DATA MANAGEMENT AND ANALYTICS

VOLUME 25 • NUMBER 2

Advertising Sales: Kim Ryan, kryan@tdwi.org, 415.913.9038.

List Rentals: 1105 Media, Inc., offers numerous email, postal, and telemarketing lists targeting business intelligence and data warehousing professionals, as well as other high-tech markets. For more information, please contact our list manager, Merit Direct, at 914.368.1000 or www.meritdirect.com.

Reprints: For single article reprints (in minimum quantities of 250–500), e-prints, plaques, and posters, contact: PARS International, phone: 212.221.9595, email: 1105reprints@parsintl.com, www.magreprints.com/QuickQuote.asp.

tdwi Transforming Data With Intelligence™
tdwi.org

President	Meighan Berberich
Vice President, Education and Research	Fern Halper
Sr. Director of Marketing	Sabrina Bozek
Graphic Designer	Rod Gosser

1105 MEDIA

Chief Executive Officer	Rajeev Kapur
Chief Technology Officer	Erik A. Lindgren
Executive Vice President	Michael J. Valenti
Executive Chairman	Jeffrey S. Klein

Reaching the Staff
Staff may be reached via email, telephone, fax, or mail.

Email: To email any member of the staff, please use the following form: FirstinitialLastname@1105media.com

Renton office (weekdays, 8:30 a.m.–5:00 p.m. PT)
Telephone 425.277.9126; Fax 425.687.2842
555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

Corporate office (weekdays, 8:30 a.m.–5:30 p.m. PT)
Telephone 818.814.5200; Fax 818.734.1522
6300 Canoga Avenue, Suite 1150,
Woodland Hills, CA 91367

Business Intelligence Journal **(article submission inquiries)**
Peter Considine
Email: journal@tdwi.org
tdwi.org/journalsubmissions

TDWI Premium Membership (inquiries and changes of address)
Email: membership@tdwi.org
tdwi.org/Premium-Membership
425.226.3053
Fax: 425.687.2842

The Need for Explainable Processes and Algorithms

Hugh Watson

Hugh J. Watson is a professor of MIS and holds the C. Herman and Mary Virginia Terry Chair of Business Administration in the Terry College of Business at the University of Georgia. He is a TDWI Fellow and senior editor of the *Business Intelligence Journal*.
hwatson@uga.edu

INTRODUCTION

With the expanded use of analytics comes a growing need to understand and explain how models are built, how they work, and what output they generate. This requirement isn't new. Dating back to the hey-days of management science and operations research, managers were hesitant to implement models they didn't understand. There was simply too much risk for their companies and personal reputations. In regulated industries such as financial services and insurance, models used to determine, for example, credit ratings or insurance rates have a long history of review by regulators.

The demand for explainable development processes and models has expanded recently. Because models are increasingly used in ways that significantly affect people's lives, additional stakeholders are demanding fairness and transparency. The media regularly reports stories of companies (e.g., Target, Facebook) that seem to be using personal data and algorithms inappropriately. Popular books such as Cathy O'Neil's *Weapons of Math Destruction* describe how algorithms can increase inequality and even threaten democracy. New laws and regulations, such as the EU General Data Protection Regulation (GDPR) and the California Consumer Protection Act (CCPA), place requirements on how companies can use personal data and algorithms; these laws and regulations are only going to increase over time.

Business and analytics managers, business analysts, and data scientists must be aware of the calls for greater explainability and use technologies and methodologies that satisfy these demands.

THE FOCUS ON PREDICTIVE MODELS

Although there are descriptive, predictive, and prescriptive analytics, the concerns are primarily with predictive models, and especially those with strong impacts on individuals. People have relatively little interest or concern about non-threatening applications such as online product recommendations or AI-powered chatbots, but they do care about:

- A prediction that a person is likely to commit another crime and is given a prison sentence rather than probation
- The decision whether a person is given insurance and at what cost based on an algorithm
- A prediction that a person would be a good employee and is granted an interview based on an algorithm's analysis of the person's resume
- A prediction that an individual isn't a good credit risk and is denied a loan
- A treatment plan created based on an algorithm's analysis of a patient's test results and symptoms

People want to know that models are fair, trustworthy, secure, and explainable and that remediation processes are in place when errors are made.

APPROACHES TO PREDICTIVE ANALYTICS

Analytical approaches to building predictive models include statistics, artificial intelligence (AI), and machine learning (ML). Statistical modeling is generally well understood and includes, for example, various kinds of regression, factor, and discriminant analysis. Artificial intelligence has many definitions, but the essence is a computer performing tasks associated with human intelligence—natural language processing, vision, solving problems, making decisions, and the like. Machine learning uses algorithms to parse through data, find patterns, make predictions, and learn without being explicitly programmed.

The distinctions between statistics and especially AI and ML are not always clear. There is overlap, and many people use the terms interchangeably. For example, ML may employ statistics and AI-like heuristics in its algorithms. Neural networks and deep learning are sometimes mentioned as examples of both AI and ML.

The definitional distinctions are not normally important, but some of the approaches (e.g., neural networks, deep learning) create "black box" models whose inner workings are difficult to understand and explain to stakeholders. These are the targets for the calls for greater explainablity.

HOW WE GOT HERE

Predictive models have existed for many years, but in the 1970s companies began to use statistical models and expert/rules-based systems. Both are relatively easy to understand and explain because the importance and relationships among the variables can be clearly identified. For example, with regression analysis the variables

and their contributions (e.g., weights) to the prediction are shown in the regression equation.

Expert/rules-based systems typically create a model in the form of a tree diagram (i.e., IF/THEN statements) which reveals the variables used and the relationships among them. As an example, several years ago I developed an expert system to screen applications to the University of Georgia's Veterinarian Science program. The model simulated the assessments made by the director of admissions and was formulated as a tree diagram. The models' variables included the usual suspects, including grade point average and test scores, but also had a branch for applicants who came from rural Georgia. Atlanta has all the veterinarians it needs for dogs and cats, but there is a shortage of vets to care for large animals (e.g., cows and horses) in South Georgia. Applicants from rural areas were given special consideration because they are considered more likely to return home to work after graduation.

Though their origins also trace back many years, the 1990s saw an increased use of artificial intelligence and machine learning. They were enabled by advances in computational power, the availability of massive amounts of data, new algorithms, and business opportunity.

The inner workings of many AI and ML models are unclear. For example, consider hierarchical neural networks (also called deep learning). Once data for the independent variables are input, each neural network layer feeds its analysis results to the next one in the hierarchy, picking up predictive power as it goes along. The consequence of this analysis approach is the significance of the various input variables and their relationships are difficult to identify and explain in human terms. Explaining the inner workings of these models is the equivalent of having to "show your work" when solving a math problem and can be difficult to do. Black-box models are perceived by some people to be mysterious and potentially ominous.

EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

Due to the growing importance of having explainable development processes and algorithms, business managers, analytics managers, and analytics professionals (e.g., data scientists) must have a broadened awareness of the factors that are important to analytics work. They need to recognize that simply building a model with a high level of predictive accuracy may not be enough. Some of the necessary understandings and actions are organizational in nature (e.g., governance) and are discussed later (see *4 Steps Companies Must Take*), while others are potentially addressed though technology. Let's consider some of the technology options.

Primarily coming out of universities, government research agencies (e.g., DARPA), and high-tech companies such as Google, research on a portfolio of techniques sometimes referred to as transparent, interpretable, or explainable AI (XAI) has been accelerating since the 1990s. The research addresses the needs of two different audiences: (1) data scientists who need to understand the workings of their models and (2) the general public impacted by the use of the models.

Some XAI techniques focus on providing "local" explanations; others strive to provide "global" explanations. A local explanation reveals why an algorithm arrived at a decision in a particular situation (e.g., a loan application was denied); a global explanation provides an overall under-

standing of the functioning of an algorithm. Local explanation techniques normally generate more accurate information about a model than global ones do because they can focus on a specific part of a model. The best overall insights about a model can be obtained by using both local and global explanations.

There are several major research streams on XAI: (1) the development of methods/algorithms that explain other algorithms (also referred to as deep explanation), (2) new algorithms whose inner workings are more transparent and interpretable, and (3) methods to systematically experiment with black box models in order to infer their inner workings (called model induction) (Dickson, 2019; Hall and Gill, 2019). Let us briefly consider each of these approaches to XIA.

Deep Explanation

Semantic associations are used with neural networks and provide an excellent example of deep explanations. With this approach, an algorithm identifies semantic attributes within the hidden layers of the neural network. For example, in the classic identifying cats in YouTube videos, a hidden layer may reveal the object has whiskers (the semantic attribute). In a lending model, the hidden layer might associate "Homeowner" as an attribute (i.e., variable) in arriving at the loan decision.

This ability to recognize attributes requires associating hidden, labelled nodes with known attributes using ontologies (e.g., knowledge, understandings, and vocabulary about a phenomenon). Also, when nodes are prominent but unlabeled, they can be identified by the algorithm so they can be labeled and added to the ontology.

Companies that need explainability may create their own algorithms to satisfy specific needs or requirements. For example, the lead-scoring function in Salesforce's Sales Cloud Einstein provides insights into how particular leads are scored, which is important to sales teams.

To meet the requirements of the Fair Credit Reporting Act, Equifax must be able to tell consumers the top four reasons why they did not get a perfect credit score and give reasonable recommendations (i.e., remediation) for how to improve their score. In response, Equifax created and patented NeuroDecision to meet these regulatory requirements for explainable credit scoring.

For a more wholistic understanding of the entire model, clusters or combinations of prominent nodes can be identified and interpreted using the ontology. This approach to explainability is narrow in the sense that explanations are tied to a specific ontology.

New Algorithms

Because of the growing importance of explainability, researchers are focusing on developing algorithms that are both accurate and interpretable. Explainable neural networks (ENNs), explainable gradient boosting machines (GBMs), and scalable Bayesian rule lists are examples of these new algorithms. They make model debugging, explanations, and fairness auditing easier.

Explainable GBMs are especially useful with applications where reasons are required for actions or decisions, such as those in regulated industries. The algorithms function by creating splits in a decision tree such that splits of a variable in one direction always increases the average value of the dependent variable while splits in the other direction decreases the dependent variable's average value. This results in a model, for example, where being a homeowner is consistently assessed with being a better credit risk.

Model Induction

Local interpretable, model-agnostic explanation (LIME) is an example of a popular technique that uses model induction. LIME can be employed with any predictive model, including ML models built using text, image, or tabular data. To illustrate its use, one of the more popular data sets for learning about and experimenting with ML uses mushroom data. The predictive challenge is to identify which mushrooms are poisonous and which are not, based on the mushroom's attributes.

From an explanation perspective, the issue is why a particular mushroom is classified by the model as poisonous or not. LIME provides this explanation by building an interpretable (simpler) model around the instance that is to be explained, and the simpler model is used to provide explanations. The explanation provided is local in that it is applicable to only the specific instance examined (a specific mushroom).

THE CAPABILITIES OF CURRENT AND FUTURE ANALYTICS PLATFORMS

Through the years, analytics software vendors made significant advances in the capabilities of their analytics platforms and workbenches. Figure 1 shows the workflow for current and future analytics platforms. The left-to-right flow across the columns corresponds with the analytics application development process. It also reflects the built-in workflow process provided by vendors and is consistent with development methodologies such as the cross-industry process for data mining (CRISP-DM) and SAS' sampling, exploring, modifying, modeling, and assessing (SEMMA). Some of the capabilities are not widely available in current generation platforms (no X in the cells) but are logical future extensions that are starting to appear in analytics platforms (e.g., H_2O).

Connect to Data Sources	Data Profiling	Data Blending	Data Transformation	Feature Generation	Model Building	Model Testing	Model Explanation	Explanation Interface
X	X	X	X	X	X	X		

Figure 1. Current and future analytics platform capabilities.

First, let us consider the more well-known, established capabilities and then delve into what is likely in the future through the addition of XAI capabilities.

This movement to XAI is starting to emerge in the products of current analytics platform vendors (e.g., H_2O), will be more common in the analytics platforms offered by other established vendors, and will be part of new platforms created by startups (e.g., Fidler Labs, Kyndi). Although XAI capabilities are the most interesting from a new technology perspective, current platform capabilities are critical to carrying out and documenting model development and maintenance processes.

Connect to Data Sources

All analytical platforms support the sourcing of data for analysis purposes, whether it is from a data mart/warehouse, SAP, Salesforce, desktop, or elsewhere. The biggest challenge is identifying which data source(s) and variables to use. By way of contrast, connecting to a specific source is relatively easy—the platform does the work. It is important to document and be able to show that appropriate data sources were used in building a model.

Data Profiling

Data profiling is used to better understand the data sources and to assess the data's quality. For example, the data profiling capabilities may include calculating a data quality index for the data sources, indicating the number of missing values in each field, showing the distribution of data in each field, and the like.

Problems with the data, such as missing values, are often addressed now (e.g., deleting records with missing data) or handled later in the data transformation step (e.g., estimating and replacing missing values). Without data of appropriate quality, there is not much trust in the models that are built. Once again, it is important to know what data was analyzed and any changes made to it.

Data Blending/Integration

When data comes from multiple sources, it must be integrated into a single table using a join based on a primary-foreign key combination (for relational data bases). This activity used to be a major stumbling block for users unfamiliar with relational concepts and SQL and it is why SQL skills are critical for analysts and data scientists.

Today, modern platforms better support data blending/integration by detecting possible table joins and executing the joining of data from multiple data sources. The joined tables are no better, however, than the quality of the data sources from which the tables are created. All data blending activities need to be recorded.

Data Transformation

It is commonly necessary to transform data prior to analysis. For example, after the data profiling and integration steps are completed, there may be a need to recategorize auto-detected field types (e.g., converting ZIP codes from numerical to categorical), normalize skewed variables (e.g., perform logarithmic transformations), and reformat data (e.g., how date is represented). Modern platforms provide a rich set of capabilities for transforming data. The lineage of all data used in models should be documented, including any transformations.

Feature Generation

Feature generation is related to data transformation and involves creating new variables (i.e.,

features) for analysis. Users with strong business domain knowledge are often important in this step because they often have insights for variables that can be productively used in an analysis. For example, in a credit risk assessment model, the data set may contain (1) monthly debt payments and (2) monthly income, and the business professional (or analyst) might suggest that the ratio of debt-to-income might be a good predictor of risk.

Feature generation is especially important because of the potential to improve the accuracy of predictive models. To illustrate, winners of the annual Kaggle competition are normally differentiated by the features they create rather than by the algorithms used.

Because feature generation is so important, analytics platforms are increasingly providing auto-feature generation capabilities. Some platforms examine existing fields, identify possible new features, test their predictive powers in various algorithms, and suggest the best new features to use. A few also allow the analyst to control the level of explainability of any auto-generated features. Modelers should be careful, of course, that new features are not based on variables (e.g., race) that are protected by laws and regulations.

Model Building

Analytics platforms:

- Offer a wide variety of algorithms for both supervised and unsupervised learning
- Create a model(s) using the prepared source data, either automatically or using the algorithm specified by the user (e.g., logistic regression)
- Provide statistics (e.g., F-value, ROC curve) related to the model(s) built
- Generate a large number of alternative models with good accuracy

The choice of which algorithms to implement should be based on potential accuracy *and* on whether they can be explained. Also, contemporary platforms (e.g., H_2O) are increasingly including capabilities that allow modelers to create only models that can be explained, which is especially important in regulated industries.

Model Testing/Validation

Virtually all models can make accurate predictions using the same data upon which they were built. The real test is whether they can make accurate predictions using data they have not seen before. Data splitting is a common method for obtaining data for model building and testing. With this approach, most of the data is used for building models while the rest is held out for testing purposes. Modelers need to be able to explain and defend the building and testing of their models.

Analytics platforms help automate the testing/validation process. For example, the platforms often use a default split that is typically 80/20, with 80 percent of the data randomly used for building the model and the remaining 20 percent used to test the model. Users can change this to different splitting percentages and how the two data sets are created (e.g., the first 80 percent of the data for training, the remaining 20 percent for testing). The platforms then provide information about the model's predictive abilities (e.g., a confusion matrix for categorical predictions).

It is also important to see how sensitive a model's behavior and outputs are to changes in input data. A model's predictive accuracy should remain the same across a variety of data inputs. For example, problems can occur with regression models when independent variables are correlated. A contemporary platform should provide a variety of model sensitivity analysis and model debugging tools.

Disparate impact testing is needed with black-box models. This testing investigates whether a model is having a negative impact on specific groups of people (e.g., women, minorities) beyond some reasonable threshold. It is needed to ensure that protected classes are not unintentionally discriminated against or simply not treated fairly.

Model Explanation

A model should be understandable to humans, and explanation capabilities are just starting to emerge in vendors' platforms (e.g., H_2O), but they will become more common. This capability should serve two audiences: (1) data scientists and (2) end users (e.g., customers and regulators). Data scientists need to be able to explore, understand, and debug the models they build. Because of their technical background and needs, the explanation techniques need to be varied and robust, with ease of use an important but secondary consideration. End users, on the other hand, need explanations they can easily understand (delivered through the explanation interface discussed next).

As mentioned, vendors' platforms contain a wide variety of algorithms; some are easy to understand and explanations are readily generated. To illustrate, categorical decision trees can be used with a lending application to generate customer-specific adverse action notices or reason codes for why a loan was denied.

Machine learning algorithms vary in their interpretability. Some can be analyzed to determine the importance of particular variables, and reason codes can be automatically generated. Others are much more inscrutable and require a combination of explanation techniques. Some of the newer algorithms that are both powerful and explainable will be included on future platforms.

Initially, vendors' platforms will focus on providing a variety of explanation techniques that data scientists can use, including algorithms explaining other algorithms. Over time, the use of these techniques is likely to become more tightly coupled and automated, with specific techniques associated with specific algorithms. It is analogous to how the data visualization and analysis platforms (e.g., Tableau) evolved from simply providing a variety of charts to recommending the best one(s) to use. The further evolution will see more specialized linking of algorithms to explanation techniques to support specific applications, such as for lending decisions.

Explanation Interface

The explanation interface connects the system's explanation capabilities to users and can take a variety of forms. Data scientists can gain a better understanding of the inner workings of the models they develop. It can be used to show business managers what the models are doing. Explanation and remediation processes can be automated.

With a natural language interface, users may be able to have a dialog and ask questions and receive answers directly, much like with Siri

or Alexa. To illustrate, after going through an authentication process, a caller might have the following conversation with a software agent:

> **Caller:** Why was my loan application denied?
>
> **Agent:** There were several reasons, but the major one is that your debt-to-income ratio is too high. Your credit card debt is also high.
>
> **Caller:** What can I do to improve my chances of getting approved next time?
>
> **Agent:** Reduce your credit card debt and improve your credit score.
>
> **Caller:** Thanks, I appreciate your help.
>
> **Agent:** You are welcome, and I look forward to doing business with you in the future.

The design of the explanation interface will also reflect the current research into how to ensure human trust and confidence.

4 STEPS COMPANIES MUST TAKE

In addition to the use of technology, organizational actions in governance and model development processes can help ensure that analytical applications meet the demands of management, laws and regulators, and calls by the public for greater explainablity.

1. **Understand how important explainablity is in your industry and specific applications**
   Although XAI is now receiving considerable attention, the importance of explainability varies with the industry and the application. If a company is high profile (e.g., Facebook) or in a regulated industry (e.g., financial services, healthcare), explainability is critical. In other industries, such as ecommerce and fintech, explainability may be slightly less important but there may be applications where explainability is a concern. The need for explainability is important for any organization using algorithms that significantly impact people's lives.

   Explainability is not always a concern. To illustrate, FirstData, a credit card processing company (recently acquired by Fiserv), uses an algorithm to detect potentially fraudulent in-store and online transactions. The public is generally unaware of its use, other than with the occasional false positive for fraud. The fraudsters are the only people negatively affected and there is no need to be concerned about or to explain anything to them.

   Even for companies where explainability is not currently a major issue (other than for internal purposes), as more algorithm-driven decisions become automated, demand to explain and justify any algorithms used will increase.

2. **Build explainablity into the entire model development process**
   Analytics managers and professionals should recognize the importance of explainability before, during, and after model development. Explainability should be an up-front consideration and not a post hoc, possibly biased justification for the work that was done. The entire model development process should stand up to an external audit of how the model was built and the model's performance and impacts on individuals and groups.

Explainability should be in every step of model development, and much of the responsibility for explainability rests with the data scientists who build the models. They need to internalize that building models with high predictive accuracy is a necessary but not sufficient condition. Document any time data is touched (e.g., accessed) and used. Record any data cleansing and transformations. Document any data used for model training and testing. Record various model versions and updates as well as the performance (i.e., accuracy) of the model. In general, every step in the process must be defendable. Fortunately, analytics platforms are increasingly providing capabilities that support explainability.

3. Satisfy the legal requirements for explainablity

For many years, interest or scrutiny of the use of algorithms has been confined to regulated industries. This has obviously changed with the passage of laws and regulations such as the GDPR and CCPA. Many of the changes have focused on the collection, use, and sharing of personal information, but there are also requirements for explainability.

Article 22 of the GDPR provides a striking example of the emerging legal restrictions on automated algorithmic decision making. It states that *"The data subject (i.e., individual) shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."* The cumulative effect of new laws and regulations is that companies must be prepared to explain how automated decisions are made, explain why any specific decision was made, provide opportunities for remediation when errors are found, and allow individuals to opt out of automated decision-making processes.

4. Expand governance to include explainability

Governance can be thought of in terms of the people, committees, and processes needed to ensure rules, norms, and activities are followed and performed correctly. The legal and societal demands for greater explainablity require an expansion of governance that affects the people involved, how algorithmic models are developed, and the review processes that are put in place.

Governance committee members should have a mix of business, legal, ethical, statistical, modeling, and systems engineering backgrounds and should ideally be led by a C-level executive such as a CIO, chief data officer, or director of data science. Assign responsibility to track emerging regulations, concerns, and issues. Companies should have a business manager or professional who is always asking and answering, "Is this fair to the consumer?"

Good governance also involves explicit policies that are regularly reviewed and updated to ensure they are consistent with new analytics initiatives. A close collaboration of all stakeholders is required to know what explanations are available for analytical applications. ●

REFERENCES

Dickson, Ben, "Inside DARPA's Effort to Create Explainable Artificial Intelligence," *TechTalks*, January 10, 2019. https://bdtechtalks.com/2019/01/10/darpa-xai-explainable-artificial-intelligence/

Hall, Patrick and Navdeep Gill, *An Introduction to Machine Learning Interpretability*, Second Edition, O'Reilly, 2019. https://get.oreilly.com/ind_introduction-to-machine-learning-interpretability-2e.html

The Case for Data Transparency: Four Key Areas Where Deeper Visibility Is Critical

Doug Kimball is vice president for global industry and solution strategy at Stibo Systems.
doki@stibosystems.com

Doug Kimball

ABSTRACT

A growing number of factors have made deeper data transparency critical for businesses across industries: an increasingly complex manufacturing and distribution process, new state and federal privacy regulations, and a growing desire for more socially conscious brands are just a few examples. A global economy means consumers have a greater number of brands with which to build loyalty; businesses risk falling behind if they don't communicate openly or fail to provide easy access to the product information their customers seek.

Businesses must consider data transparency in four key areas: product, supply chain, AI/data protection and privacy, and corporate social responsibility. This article offers tips for leveraging data management technology to achieve the transparency businesses and their customers require in today's marketplace.

INTRODUCTION

Today's consumers have the world at their fingertips. Perhaps no other industry has been as impacted by the sea changes of the Information Age as retail. Amazon led the way in making choice and flexibility key factors in how customers shop. To stay competitive and customer-focused, many businesses have been forced to adapt (e.g., grocery stores are rolling out new ordering, pick-up, and delivery offerings) or retire (e.g.,

video stores made way for streaming and video-on-demand services).

We often discuss choice at a surface level: in retail, that might include a customer asking in what stores the product is available or can it be shipped to a store closer to me? Is my size available? What colors are in stock? Maintaining clean, consistent product data is vital to meeting consumer demand. Ensuring visibility into supplier data has also rapidly gained prominence in recent months as supply chains respond to pandemic-related disruptions.

However, businesses must recognize it's not just surface-level choices that impact a buyer's decision. As stores look deeper into aspects of their products and business operations that build long-term loyalty, they find that data plays an even more important role. Customers want the stores where they shop to practice transparency—to know that the products they're purchasing were manufactured and delivered ethically and sustainably.

When a business is open to sharing information about the materials in its shirts or the ingredients in its food and makes that data easily accessible, it builds customer relationships based not just on product availability but on shared values—a far firmer foundation. The numbers support a transparency policy: 86 percent of Americans believe business transparency is more important than ever, and 73 percent of consumers are willing to pay more for brands that guarantee product transparency (Sprout Social, 2019).

If numbers aren't proof enough, consider the general direction of global policy surrounding data transparency. More countries are passing their own version of the General Data Protection Regulation (GDPR). The California Consumer Privacy Act (CCPA) brought stronger consumer protections, and the Brazilian General Data Protection Law (LGPD) also became enforceable in early 2020. Businesses that don't comply face steep fines. Data transparency initiatives make it far easier for businesses to respond to customer requests for their data as well as notify consumers in the event of a data breach. Of course, it's best to focus on data transparency for the carrot rather than the stick.

As businesses keep up with both customer demand and changing privacy expectations, understanding the four key areas where data transparency can make a positive impact will become critical:

- **PRODUCT TRANSPARENCY**. Do customers have access to product availability? Can they see product specs such as materials used?
- **SUPPLY CHAIN TRANSPARENCY.** How effective is the business' track-and-trace capability? Is its data visibility powerful enough to quickly resolve a crisis such as a recall or shift production gears after a natural disaster?
- **DATA PROTECTION AND PRIVACY.** What steps is the business taking to keep customer data private, even as more customers seek personalized offers from brands? Are those steps in line with global data protection policies?
- **CORPORATE TRANSPARENCY.** How is the business leveraging data to improve corporate social responsibility (CSR) initiatives—for example, actions taken to increase sustainability or support human rights? How is the business making CSR initiatives visible to the public?

Before organizations improve their transparency initiatives, they must first consider the goal they most want to achieve by strengthening their data. A more extensive examination of each area is required to understand the impact of data transparency on a retailer's operations and how a strategy rooted in achieving a single view of data can make transparency possible.

Product Transparency: Every Attribute Is Important

Prior to 2020, complete product transparency was an important, if not yet critical, goal. COVID-19 created the turning point. The global pandemic has driven more consumers to select home delivery or buy online, pick up in store (BOPIS) options from small *and* larger retailers. E-commerce sales growth soared mid-pandemic, with 78 percent year-over-year growth in May, maintained at 76 percent in June (Crets, 2020). Even as growth slowed a bit—55 percent year-over-year growth in July—Adobe predicted 2020 sales would match 2019's by early October, right before the holiday season (Crets, 2020). Although some of these trends may abate now that a COVID-19 vaccine is available, expect many converts to stick with digital shopping based on the convenience and wealth of options.

Complementary to BOPIS, buy online and return in store (BORIS) numbers have increased as well. As consumers purchase more goods online, the probability of an item not being what was expected, not fitting properly, being the wrong color, or just not actually needed goes up. With regulations for physical in-store returns varying by geography, BORIS can be challenging by location. As a result, it becomes even more important to have rich data about the product to properly set customer expectations. Accurate data about availability by location and making the return process as seamless as possible are key to making for a smooth e-commerce process.

The increase in online shopping doesn't necessarily translate to more sales for a business, however. In fact, e-commerce enables greater window shopping. Instead of driving to a mall, buyers can simply browse multiple websites from their home. They can compare prices between a manufacturer and a big box store and gather information from hundreds of product reviews.

Committing to product transparency can help cauterize e-commerce customer bleed and help build loyalty.

Even convincing a customer to complete the sale is a complicated process. An aggregate of studies puts the number of shopping carts abandoned at nearly 70 percent (Baymard Institute, 2019). Surprises such as added costs or qualifying conditions for free shipping can cause buyers to leave the site. Concerns about product quality, conflicting product specifications gathered from different sites, or the lack of product images can also raise second thoughts.

Committing to product transparency can help cauterize e-commerce customer bleed and help build loyalty. Look first at product availability: SKU data connected to the e-commerce app or website must maintain real-time consistency with what's available in store or at the distribution center. Customers don't want a surprise

when they arrive to pick up an item only to discover it's actually out of stock. They don't want to order an item for next-day shipping and receive an email explaining that the product is delayed a week. Conversely, delivering an item on time that looks just as it did online—not, for example, an image of a light purple shirt that's actually deep purple—will help establish trust.

Once a business achieves consistent product availability, it should focus on providing a broader range of product attribute information across all platforms. Consumers have a variety of needs and wants, and their brand loyalty can be severely impacted if they are asked to search the website—or even review websites to gather basic product info. For example, food, beverage, and consumer packaged goods brands need to make ingredients readily visible on product pages so customers can determine whether an item matches dietary restrictions or preferences. Similarly, health and beauty brands should identify whether a product was tested on animals.

Because it directly impacts customer retention on sales platforms, data transparency should be addressed immediately.

To take it a step further, as businesses reopen across the globe, with varying regional regulations, social acceptance of exposure and a drive to engage—transparency becomes important to businesses seeking to pull consumers back. Dining establishments need to share location information (e.g., what are their open hours, do they have distanced seating, are they offering pay and pick-up, do they have handicapped parking, etc.) All of these should be made available to consumers, whether online, in an app, or ideally both. Brick-and-mortar stores want to share whether they are now taking returns in person, at curbside, what shopping hours are available, and much more. Creating this transparency allows businesses to safely and effectively respond to a growing number of in-person visits and enables people to feel comfortable with their choices.

Because it directly impacts customer retention on sales platforms, data transparency should be addressed immediately. Just as important is achieving transparency across product manufacturing and transportation.

Supply Chain Transparency: Every Node Is Important

From a customer perspective, supply chain transparency is perhaps the most significant ask to come out of the Amazon revolution. When the e-commerce giant introduced its Prime program, free two-day shipping became an industry standard; pressure continues to grow as options such as same-day delivery are introduced.

The number of consumers seeking same-day delivery options jumped dramatically between 2018 and 2019; more than 90 percent want free one-day delivery via the fastest way possible (including via drone, driverless car, or messenger) compared to 43 percent who wanted such options in 2018 (Oracle, 2019). Brands that make such promises and don't deliver face a steep cost—the same study found 13 percent of customers whose orders arrived late would never buy from the brand again (Oracle, 2019).

As track-and-trace solutions become more commonplace, hardware that enables them—such as GPS devices, barcode printers, IoT beacons, and RFID technologies—and data management software have also advanced, allowing customers deeper visibility into their order status as soon as a business receives it. Feeding this information across platforms is important: customers will expect apps, websites, and customer service representatives to have consistent updates on their order if it gets lost or delayed.

Enabling customers to track the progress of their orders, with near real-time updates, is almost expected and certainly is important in driving loyalty. The same is true for returns: customers want to know when their item is received, whether a refund is approved, and when the money will be back in their bank account.

Achieving supply chain transparency is also key for supplier relations. Lacking details about product movement across the supply chain can pose a multitude of challenges for replenishment and inventory management. As manufacturing processes become more complex and may even cross borders, organizations must understand what products and how many of each are located at each link along the supply chain. Not knowing where an item or shipment is during its journey can impede a company's ability to shift suppliers in case of disruptions or shortages—as a major event such as a data breach or a warehouse fire could take a critical partner offline for days or weeks.

Furthermore, it can slow recall resolutions. Supply chain data is critical for pinpointing which materials are tainted or broken and which suppliers are impacted. The bottom line is that without the clarity from having agreed-upon data, monitored by a supply chain "control tower," and providing advanced analytics, decision making is seriously diminished.

Supporting such deep level transparency requires shared, trusted data between an organization and its suppliers—at every node in the supply chain, ensuring full visibility into activities both upstream and downstream. That trusted data provides a variety of opportunities to develop deeper insights into product development and flow, as well as visibility. With improved visibility gained from centrally managed supplier data, businesses can more effectively monitor performance and assess the value of their supplier relationships.

Implementing technology such as a master data management solution can allow businesses to securely provide data to each of their suppliers (and vice versa), giving partners a 360-degree view of production from raw materials to store shelves, and even to the customer's front door. It allows the business to more effectively maintain cash flow expectations, adjust production and labor planning, reduce downtime caused by a recall, and manage warehouse operations.

However, with so much sensitive information moving between a business, its customers, and its suppliers, ensuring privacy is another critical data transparency issue.

Data Protection and Privacy: Every Profile Is Important

The previous section briefly mentioned the threat of a data breach. Although it seems like a nightmare scenario, it's unfortunately becoming a more common occurrence as malware grows more advanced and hackers grow bolder. Over

half (57 percent) of companies experienced a data breach between 2017 and 2019, while 24 percent suffered a breach in the first half of 2019 alone (Bitdefender, 2019).

Not every breach makes headlines such as the City of Atlanta's ransomware attack or Target's credit card data hack. In fact, attacks don't have to come from an outsider. IBM research determined malicious insiders are responsible for almost 45 percent of cyberattacks (Armstrong, 2016). Still, enough consumers have had their credit card numbers—or, even worse, their identities—stolen during high profile hacks that transparency in data privacy is now essential. In these cases especially, data transparency can engender trust, or at the very least prevent situations from becoming worse if an enterprise doesn't have or share information.

Compounding the challenge is personalized advertising driven by artificial intelligence, which has made consumers more aware of the amount of data they generate each day. Between the laptops we use for work, the phones we use for play, and the virtual assistants we use for convenience, businesses have access to astoundingly deep customer profiles. They use that information to send targeted emails featuring offers specific to an individual customer, display ads that pop up as customers surf the web, and run ads before online videos based on customer search histories.

Yes, consumers want more relevant ads: two-thirds of buyers believe personalized experiences equate to more positive relationships with retailers (Stibo Systems, 2019). Almost all (99 percent) retail marketers believe personalization plays a role in advancing customer relationships, and 92 percent say customers expect a personalized experience (Evergage, 2020).

Despite such great opportunities for personalization that consumers are eager for, these same consumers don't want those targeted offers to come at the expense of data misuse. As many of us have likely experienced, receiving an ad on our phone for an item we didn't search for but rather spoke about in a person-to-person conversation can be both jarring and disturbing.

These challenges and others led state and national authorities to introduce data regulations—and companies doing business in both the United States and the European Union are responsible for ensuring their respective operations are in compliance. The GDPR in particular offers guidelines that would cause headaches for companies that lack deep data visibility. For example, businesses have 72 hours after they discover a data breach to inform all impacted parties and are required to find and delete all of a customer's personal information upon that person's request.

Adopting technology that enables data transparency simplifies privacy reporting and compliance while allowing businesses to safely leverage data into more personalized advertising efforts. By gaining visibility into all customer data through a single source, businesses can find, present, and delete data soon after receiving a request. They can also quickly determine which records might have been breached during a cyberattack, notify affected customers, and work to make things right. There's no question a data breach will impact brand trust, but pursuing a policy of data transparency allows

businesses to respond as effectively as possible—and win back customers quicker.

Adopting technology that enables data transparency simplifies privacy reporting and compliance while allowing businesses to safely leverage data into more personalized advertising efforts.

Call it the era of the empowered consumer. Just as customers want businesses to be good stewards of their data, they also want brands to be good stewards of their communities. That leads us to our final area: corporate transparency.

Corporate Transparency: Every Action Is Important

As information sharing becomes global, consumers have increased their awareness of cultural and environmental concerns. They're following stories closely, researching potential solutions using the vast amounts of data available, and meeting like-minded people on social media.

This rapid spread of information allows new movements to form, with goals ranging from combatting racism to improving working conditions and reducing the impacts of climate change. They're also making more informed decisions about where to shop, eat, and live based on which option best aligns with their values.

As businesses build brand loyalty among more socially active consumers, it's clear that a few words and the occasional donation are no longer an adequate strategy. Buyers want their brands to stand for something; nearly two-thirds (64 percent) were "buying on belief" in 2018—meaning they're patronizing businesses that openly support causes important to the buyer—and those buyers are just as likely to demonstrate purchasing intent after learning about a brand's values as they were after learning about a specific product (Edelman, 2018). The rise of brands such as Toms, which donates a pair of shoes for every pair sold, demonstrates consumer desire to support companies making a difference.

Corporate data transparency is important for two reasons. First, it enables brands to achieve the manufacturing and logistical goals contained within many corporate social responsibility (CSR) campaigns. Consider a brand seeking to reduce its carbon footprint. Among the steps it could take:

- Determine how much of its raw material is recyclable and whether onboarding new suppliers will improve that rate
- Understand energy consumption at manufacturing plants and warehouses and consider how to modify operations to reduce energy waste
- Examine its transportation strategy to determine the emissions produced by its trucks and calculate the number of empty miles it could reduce while still delivering goods efficiently

Such an initiative can bring in considerable new business from sustainability-minded consumers—not only because the products are more sustainable but because the company follows sustainable business practices. As a

companywide effort, each section of the business must have visibility into operational data to best determine how to make an impact.

Second, pursuing a policy of data transparency ensures customers have easy access to information about how the brands they patronize act as global citizens. Presenting this information prominently on all sales channels—and keeping it up to date—builds trust and makes it clear the business is living its CSR initiatives rather than simply paying them lip service.

Pursuing a policy of data transparency ensures customers have easy access to information about how the brands they patronize act as global citizens.

However, keeping this data clean can be a challenge without a dedicated data management strategy. All benchmarks and supporting data—whether that's product, customer, or supply chain—must be visible through one system, simplifying the process of gathering information and keeping CSR program details consistent across platforms.

Building a Better Data-Transparency Strategy

It's ironic: enabling simplified access to accurate data is one of the most complicated tasks a business can undertake. As businesses collect mountains of external and internal data each day, determining how to sort through it and leverage it ethically—while keeping customers apprised of how their data is being used—requires a strategy to break down data silos and enable easy sharing across the company.

There are many ways to approach this challenge. One is master data management (MDM), which can enable this visibility by providing a single, dedicated system for accessing data across the business, giving anyone who needs it (whether in marketing, R&D, or supply chain) visibility anytime, anywhere. MDM allows businesses to syndicate data to all platforms, simplifying keeping product sheets and website and app pages up to date and enabling visibility into deeper product information before the customer places an item in their cart.

MDM helps businesses share data with suppliers and see any operational challenge along the supply chain before it significantly impacts business. Most important, it allows businesses to be more transparent with customers about their activities, whether that's keeping their personal data safe or making the world a better place.

MDM is neither an initiative to be taken lightly nor a short-term option. Instead of a single project, it often occurs in waves as businesses master thousands or even millions of customer, product, location, supplier (and other) data records. It's a radical overhaul of how businesses approach data sharing and accessibility, and it requires buy-in across the business. Mastering product, customer, and location data delivers a higher ROI when that data is continually leveraged in manufacturing and marketing tasks.

Once an MDM initiative is underway, businesses must consider complicated questions around privacy (which employees should have access to what data?) and oversharing (how can

a business become more transparent with its customers while avoiding disclosure of proprietary data?).

However, the benefits of adopting a data transparency strategy based on MDM far outweigh the complex implementation process. Such a strategy allows customers to see the brand as a trustworthy source, to know exactly what they're getting when they purchase an item, to know their personal information is kept safe, and that the business shares their values. Ultimately, those are the expectations of businesses operating in the era of the empowered consumer. Only data transparency—both among internal audiences and with external customers—helps businesses meet those expectations. ●

REFERENCES

Armstrong, Martin (2016). "Most Cyber Attacks Are An Inside Job." Statista, retrieved from https://www.statista.com/chart/4994/most-cyber-attacks-are-an-inside-job/

Baymard Institute (2019). "41 Cart Abandonment Rate Statistics." Retrieved from https://baymard.com/lists/cart-abandonment-rate

Bitdefender (2019). "Hacked off!" Retrieved from https://www.bitdefender.com/files/News/CaseStudies/study/285/Bitedefender-Hacked-Off-Report.pdf

Crets, Stephanie (2020). "Online Sales Taper Off in July as Retail Stores Reopen." DigitalCommerce360, retrieved from https://www.digitalcommerce360.com/article/coronavirus-impact-online-retail/

Edelman (2018). "Earned Brand 2018." Retrieved from https://www.edelman.com/earned-brand

Evergage (2020). "2020 Trends in Personalization." Retrieved from https://www.evergage.com/resources/ebooks/trends-in-personalization-survey-report/

Oracle (2019). "Retailers and Customers Don't See Eye-to-Eye on Returns or In-Store Experiences." Retrieved from https://www.oracle.com/corporate/pressrelease/returns-in-store-experiences-101519.html

Sprout Social (2019). "#BrandsGetReal: Social Media and the Evolution of Transparency." Retrieved from https://sproutsocial.com/insights/data/social-media-transparency/

Stibo Systems (2019). "Stibo Systems Recent Survey Reveals Important Gap." Retrieved from https://www.stibosystems.com/press-releases/news/recent-survey-reveals-gap-between-demand-for-deeper-consumer-experiences

Overcoming BI Self-Service Obstacles During and After a Pandemic

Bhupesh Malhotra and Dan Columbus

Bhupesh Malhotra is a product manager for Wyn Enterprise.

BhupeshM@grapecity.com

Dan Columbus is director of enterprise sales at Wyn Enterprise.

dan.columbus@grapecity.com

A decade ago, *ComputerWorld* discussed the merits of self-service business intelligence [Horwitt, 2011]. In particular, the article addressed a car dealership that implemented a business intelligence platform to uncover early warning signs of the Great Recession a few years before. These signals of impending problems gave management time to prepare—all without the need for technical assistance from the IT department. Consequently, they weathered the downturn when so many other car dealerships faltered because they held too much inventory when demand slowed.

Ten years later, businesses face another economic downturn. During the global COVID-19 pandemic, most businesses requested employees who could work from home to do so. Many companies also needed to shift the way they conducted business to cope with social distancing guidelines. For example, retailers and restaurants needed to shift their focus from in-person to online sales quickly. For these businesses to even attempt to survive, IT teams needed to race to offer increased e-commerce options (that may have barely existed before). Additionally, thousands of companies worldwide moved out of their brick-and-mortar offices to begin working in remote environments. IT teams needed to transition complete infrastructures from physical offices to the cloud.

The ability to overcome these challenges will be the driving distinction between businesses that persevere and succeed during this unprecedented time and

companies that will suffer an economic collapse. Many countries lag technologically and face intense challenges with abrupt technological shifts. Numerous IT directors are working 12-18-hour days—creating and monitoring remote environments, managing new digital collaboration environments, and an excess of newly assigned remote workforce tasks. IT departments are overworked.

Businesses have been asked to change, and self-service BI is driving this change. The need for real-time data and data visualization tools are needed now more than ever before.

SELF-SERVICE BI IN A CHANGING LANDSCAPE

Underscoring the assertion that businesses understand the need for fast, flexible business intelligence platforms, Dressner Advisory Services published interim results from their recent and still ongoing "COVID-19 Impact Summary." [Dressner, 2020]

The survey found:

- **BUSINESS CHALLENGES:** All respondents said the pandemic had affected their business. Most perceived either a stable or declined operating climate. Few reported seeing any signs that the environment might improve soon. Some common challenges included sourcing, changing customer behavior, and even rules and mandates governing how they can operate.
- **INCREASED BI INVESTMENTS:** Despite setbacks, just about half of all respondents said their companies launched new BI initiatives. Also, the importance of both collaborative and self-service BI platforms has risen sharply in perceptions.

Even during a global pandemic, business intelligence (more specifically, self-service BI) supports companies by saving time. With self-service business intelligence, users can be more independent and efficient.

Benefits of self-service BI include:

- **SAVES TIME.** Before the pandemic, end users would request data visualization (dashboards and reports) from IT departments. Self-service BI changes this request process. Self-service platforms can speed up the method of data gathering (without having to wait for IT to process requests) and give analysts and managers better control of both their information and the way they view it. This can prove especially critical in an age when many people have begun to consider remote work the "new normal."
- **REAL-TIME INSIGHTS.** Self-service BI can also provide the critical early-warning signals that companies require in such an unpredictable environment where they have to deal with issues that range from unreliable supply chains to changing consumer behavior. The abruptness of the changes during COVID-19 emphasizes the need for fast, flexible ways to gain and use information. Self-service BI enables the collaborative, remote, and self-sufficient reporting and analytics that all businesses need in a continuously changing economic landscape.

OBSTACLES TO SELF-SERVICE BUSINESS INTELLIGENCE

As mentioned by Jim Gallo on TDWI about two years ago, the idea of self-service BI didn't suddenly arise from social distancing measures during the COVID-19 pandemic [Gallo, 2020].

Both IT departments and users have long discussed benefits, available platforms, and the challenges they would need to overcome to offer their employees these services.

For example, these are some of the top-of-mind benefits to consider:

- Information technology professionals could save time if they didn't need to create many ad hoc queries and reports. They could focus more on creating remote environments, managing digital collaboration environments, training end users working within new remote environments, troubleshooting e-commerce shopping carts, creating cloud infrastructures, and managing dispatch help desks for end users.
- From analysts' and managers' perspective, they could work more self-sufficiently because they wouldn't have to wait for a busy IT department to get to their report or revision. Users could get answers they needed faster.

Of course, on the other side of the argument, some employees have already taken it upon themselves to create their own self-service queries and reports. Commonly, they turn to standalone apps or isolated Excel sheets. Unfortunately, these spreadsheets and apps often fall under the realm of shadow IT—the use of data, applications, or related services without explicit approval from the security and governance groups within IT.

In the past, companies have viewed shadow IT with uncertainty. Its use demonstrates that employees outside of IT want to use tech to solve their own productivity issues proactively. However, surveys of security professionals found that almost half of them view the rise of shadow IT from remote workers during the coronavirus as a looming threat [Muncaster, 2020]. IT teams cannot protect information, apps, or devices that are not supported by the organization's central IT department. In a recent study, 80 percent of employees say they use applications on the job that aren't approved by IT. Shadow IT can present serious security risks to your organization through compliance violations, infrastructure vulnerabilities, and exposure of privacy-sensitive information through data leaks.

As just one industry example, *Bank Info Security* reported the massive and sudden shift to remote working had turned the existing problem of employees downloading and using their own unapproved applications into an even more critical security issue [Goswami and Nandikotkur, 2020]. One security expert called the situation "chaotic" as employees increasingly turn to insecure apps without understanding the danger.

These concerns lead to exploring the significant benefits that self-service business intelligence platforms can offer an organization.

MAJOR ADOPTION OBSTACLES TO SELF-SERVICE BUSINESS INTELLIGENCE

Considering that many businesses understand the benefits of self-service BI, why is it not implemented throughout the organizational structure from the start?

The following are a few of the major obstacles:

Data Security

When it comes to self-service, the main risk is that data will be misused. Businesses can't afford to compromise on data security and governance in the face of rising threat levels. Sacrificing security could mean losing valuable information, getting slapped with government penalties, and even generating poor press over data breaches. Some businesses fear that adding self-service applications will increase security vulnerabilities. They may approve the platform but lose control of how employees use the data.

OVERCOMING DATA SECURITY ISSUES: You can overcome data security obstacles by establishing governance practices. Governance can be mistaken as control or preventing access. This is not the case; data governance is about organizing, accessing, and maintaining data securely in a democratized manner. Data democratization will safely extend the access to an organization's data (from the department level to the corporate level to the local level). Instead of restricting access to a select few, employees can have secured access to data with implemented regulations throughout a company.

Although data governance is a business process that requires planning and implementation, it is essential to bring data securely to business users. It starts with organizing and defining the data that needs to be secured. This helps establish a standard definition of the data (while defining rules and regulations).

Speed-To-Value

Some potential users may fear that it will take them too much time to create a usable and complete dashboard or report. Some users may even worry that it will take them too long to learn how to use the new platform to produce the sorts of results they require. This is further amplified if users are working remotely. Business users may not be able to attend live training sessions. Users also may feel that individual training sessions will distract from their essential job functions.

OVERCOMING SPEED-TO-VALUE ISSUES: New software can be overwhelming to non-technical users whose focus is to keep the business running smoothly. To help overcome the fear of inadequately using a new business application, one of the best options is to incorporate the functionality into an existing system that the users are comfortable using. This ensures that self-service BI is blended with regular business processes.

For example, provide an embedded BI experience so users do not need to log in to a separate system. Keep the self-service environment simple and easy to use; create and filter prebuilt dashboards by slicing and dicing prefixed options. If you build a simple, easy-to-use system (that integrates into the same app that the business users are using daily), users will feel less hesitant with no additional learning curve or outside applications to navigate.

Little Commitment to Change

Implementing changes to an established business process is complicated and time-consuming. Not only is there training involved, but there is often a perception that the new approach is more complicated and will take more effort. Continuous leadership action is required to keep everyone motivated (and on track) to make new systems an intrinsic part of what

they are used to daily. Particularly in a difficult or unpredictable business climate, some leaders may find it difficult to justify changes from established processes and investment in new ones. They often fear that the learning curve or the time it takes to create dashboards will consume too many resources or produce ineffective results.

OVERCOMING RELUCTANCE TO CHANGE: To overcome the reluctance to change, leadership needs to show users all the benefits of using self-service BI tools. Ask senior management to become involved in the self-service BI implementation initiative. Have senior management (outside of IT) evangelize the benefits and lead the initiative. Have a top-down approach emphasizing how executive management uses self-service and how they benefit from the change. Managers in all departments can follow, personalizing each message, discussing use cases, and explaining interdepartmental benefits.

Some employees will still want to hold onto the spreadsheets and apps they have always used to gather and analyze information. These employees need to understand that old-fashioned methods create even more data stores, with all of the associated problems of shadow IT. For example, attaching a spreadsheet to an email generates a common versioning problem when suddenly, multiple versions of the same spreadsheet keep getting passed around with no way to control updates.

Overcoming general resistance to change will require end-user education about how each employee will personally benefit from the use of self-service BI:

- Having access to their own BI dashboard gives decision makers the chance to produce exactly what they need when they need it.
- The use of these platforms helps employees better understand their own information to utilize it more productively.
- The new platforms have intuitive interfaces that do not require special technical skills to learn to use. Instead of requiring coding, they work off of data sources and parameters. Once employees observe use cases and receive orientation or training with the new technology, they tend to enjoy using it (and even feel thankful for their business's modern tools that help empower them). Ethan Conner's video demonstrating how to create a BI dashboard in five minutes offers an excellent illustration [Conner, 2019].

Issues With the Speed of Data Collection, Validation, and Cleaning

Sometimes, IT professionals still work to tune established applications to meet growing dataset demands. In response, they may need to generate smaller sets of more structured data. In turn, they're worried that the sources for any subsets they create may get outdated before they complete the project because of the time it takes to prepare usable datasets.

Indeed, an analytics program can only perform as well as the quality of its data sources. Plenty of businesses already struggle with massive stores of dark and outdated information that they haven't cleaned. Beginning with self-service BI hardly seems like the first step in what they believe will be a major project.

OVERCOMING CONCERNS OVER EXISTING DATA QUALITY: In a data-driven organization, everyone needs to understand and work with data in some way or another. Data literacy thus becomes paramount to ensure everyone understands and views the data in the same context. Educating business users about where the data originates and interpreting it helps build confidence and establish a common data language. Showing the data lineage for the underlying data points in a dashboard further inculcates trust in the validity of the information displayed.

As users start trusting the data and the processes to collect, clean, and validate the data points available to them, they will be more confident using the data in their own self-service BI dashboards and reports.

If your data needs to be cleaned before it is visualized, data unification tools will clean and validate outdated data stores. When users interact with data (that has already been cleaned and validated), it will build and enhance user trust.

Reluctance to Learn New Applications

Even though employees often demonstrate their willingness to use technology and adapt to new situations, most of them have already faced many challenges as they adjusted to working remotely or at least working with extra precautions. Even though they understand the benefits, they already feel overwhelmed and less eager to confront something they perceive as just one more burden that management has added on their plate. Many times, users are not consulted to clarify their needs to ensure the BI system's benefits match their requirements.

OVERCOMING RELUCTANCE TO LEARN NEW APPLICATIONS: Typically, with self-service BI implementations, the more you can prove the benefits and ease of use, the less reluctance you will face from end users. Therefore, start with a smaller focus group of users who have clear and immediate use cases to confirm immediate results.

The users in this small group can be from different departments with different user needs. When selecting the individuals, it is essential to choose people who are not overwhelmed by learning new techniques. Especially with a predominantly remote workforce, those who have adjusted to the new environment are the best candidates.

Many times, users are not consulted to clarify their needs to ensure the BI system's benefits match their requirements.

After the initial group has submitted their feedback, adjust your BI deployment and approach if needed. Then repeat with the next group with a varying set of challenges. As you move from one group to the next, ensure that you keep the early adopters involved and updated with new learnings.

These focus groups have always been a great tool to help gather feedback about business processes and implementing process changes. Focus groups will help build momentum for the new process and prove the benefits of

adoption. These groups also help grow the BI knowledge of the organization.

Starting with small groups also gives IT teams and BI analysts an advantage—they can focus their efforts on a small group of people rather than the whole organization. In the current work environment, where most of the workforce is remote, this dramatically reduces the effort that the IT teams and analysts need to start a self-service initiative and keep it going. For the organization, the ROI of such an approach is also evident as the investment is relatively small and focused.

Misunderstanding How to Use Self-Service BI Effectively

Some employees may view the data without a good understanding of how to use it.

OVERCOMING MISUNDERSTANDING HOW TO USE SELF-SERVICE BI EFFECTIVELY: Communicate the need to know your audience. Additionally, understand where the most significant benefits will arise.

With self-service BI you can choose the audience: who is consuming the data. As an example, a CEO's dashboard might provide revenue and sales data, but that CEO's office assistant might not know how to best use that information. A BI platform can generate revolutionary change. At the same time, businesses mostly use these platforms to give managers or analysts a way to empower quick decisions or present their leadership cases. A manager or executive may prefer to see data visualized in a chart, graph, or even infographic. In contrast, analysts may want to see more of the hard numbers used to generate the report.

The benefits of self-service business intelligence include efficiency, speed, and the ability to make better decisions. Organizations can improve their resiliency, agility, and even their culture. To overcome all of these obstacles, show the benefits of self-service BI, personalized by the department. Communicate how self-service BI gives every user precisely what they find most useful.

BENEFITS OF SELF-SERVICE BUSINESS INTELLIGENCE

Even before the pandemic, surveys of companies emphasized the importance of self-service applications for business intelligence. Businesses understand the need to save time and become more self-sufficient. They understand the need for better collaborative tools.

Improves Efficiency

From the perspective of busy IT departments, giving their users self-sufficiency can help free up their technical talent from backlogged reporting tasks. The growing demands of managing new software and increased security issues have already stretched technical talent thin during the pandemic. These issues only lengthen turnaround times.

Self-service BI gives end users independence. Users can produce their own dashboards and reports without waiting for their turn in the queue (or resorting to shadow IT). Self-service BI tools allow users to extract and refine their results instantly. BI tools enable remote workers to access their information anytime and anywhere. Well-secured, BI tools allow remote workers to collaborate (even if they're not in the same building or even in the same country). With the right credentials and a secure internet connection, users can access

their platforms from a browser on just about any device, including laptops, smartphones, or tablets—all completely secured.

Provides a Tool for Fast and Flexible Insights

As *Fast Company* mentioned, businesses should take at least one lesson away from the pandemic [Baiya 2020]. Agility, flexibility, and resilience aren't just buzzwords but also real advantages that organizations need to survive. According to a University of Minnesota report, more working Americans had lost jobs than during any downturn since the Great Depression [Soucheray, 2020]. This suggests that many companies have not achieved these often-stated goals as well as previously thought.

Some companies may have reacted too slowly or taken the wrong path when they did respond. These organizations needed to find cheaper, faster, or more effective ways to operate in a dramatically and abruptly transformed business environment.

The McKinsey report noted that accelerating digitization had widened the gap between leaders and laggards even more than 10 years ago.

A year ago, McKinsey's quarterly report warned businesses that they should not waste time trying to predict when the next downturn will occur [Hirt, Laczkowski, and Mysore, 2019]. Instead, they should spend more time preparing to cope when it comes. McKinsey also spent time looking into the reasons why only about 10 percent of businesses flourished when most companies struggled during the Great Recession.

McKinsey distilled the behaviors of the winners into three main points:

- They created a safety buffer by taking such steps as divesting themselves earlier of underperforming businesses, finding alternative financing, or accumulating more liquid safety nets
- They began to cut costs as much as a year before the downturn in response to business and economic signals that provided warnings
- They found and invested in growth opportunities, before and during the downturn

The current economic downturn doesn't precisely mirror the Great Recession (or just about any economic decline in recent history). It is fair to argue that no one had much advance warning of this particular crisis. The consequences of COVID-19 emphasize the need to have tools that offer an increased ability to understand an unexpected situation in order to react adequately.

With self-service BI, enterprises can tailor the way they design and produce data visualization and analytics for their immediate needs. Additionally, they can use fast and flexible platforms to add time-to-value insights by using their own information—and not having to wait for IT (or anybody else) to provide them with what they require.

Empowers Data-Driven Business Decisions Through Independence and Time Savings

The McKinsey report noted that accelerating digitization had widened the gap between leaders and laggards even more than 10 years ago. As an example of hard numbers, they found that traditional cost-cutting measures that may have served businesses well enough during the Great Recession do not tend to perform as well as those measures derived from in-depth data analytics.

From McKinsey's own analysis:

- Such conventional tactics to cut costs (such as reducing labor costs) only reduce expenses by an average of two percent
- In contrast, personalized, data-driven measures typically uncover cost-saving measures that account for reduced expenses of closer to seven percent

Self-service BI products with a low barrier to entry for non-technical employees can ultimately turn almost every decision maker into a data-driven analyst. Businesses can use real-time data to make strategic decisions faster, providing them with a competitive edge in discerning both overall market trends and those within their own organization.

ENJOYING THE BENEFITS OF SELF-SERVICE BI PLATFORMS

Everybody understands the pandemic forced businesses into challenging situations because of numerous abrupt changes. It is fair to mention that these aren't the best times to force all companies into making more dramatic, pervasive changes. Simultaneously, implementing self-service BI platforms will not necessarily have to fall into the category of a major risk or 180-degree pivot.

As with many innovations, companies might begin with those users or departments who have proven most receptive to technological change or, in particular, self-service BI in the past. Project planners could also take the path of least resistance by finding the kinds of projects or areas that are most prepared for the change by already having fairly clean, structured data sources to use. Once this subset of employees works with IT to resolve common issues, they can serve as ambassadors for this kind of transformation in more areas.

Companies can also adhere to best practices that will make it easier and more efficient to use the platforms. Examples include creating catalogs of preset reports and using parameters to make it easy to change certain variables to suit the purpose. Some users can benefit from their BI platform by doing little more than selecting their report and choosing parameters from drop-down lists.

INVOLVE IT FOR SUPPORT AND GOVERNANCE

Even with a self-serve BI platform, businesses still need IT support to create data sets and provide help. IT can often set up automated refreshes of information sources so it is more efficient (without the need for constant requests for new or revised reports). The best of the new platforms come with built-in security, governance, and collaboration tools that will help alleviate concerns that IT departments may have had about other user departments' methods.

BI systems can create virtual data sets with many different views, and these are easy to

generate for specific uses. They have built-in intelligence that can make order out of either structured or unstructured data from a variety of sources. Also, they can refresh their data and their views to keep data updated.

HOW TO BEGIN EXPLORING SELF-SERVICE BUSINESS INTELLIGENCE

Although most companies encounter challenges during this uncertain time, several hurdles can be minimized using self-service BI. The use of self-service data visualization makes information more valuable by uncovering opportunities to reduce expenses, improve business processes, and increase revenue. As the ability to make data-driven decisions increases and moves upward within an organization, a fundamental change happens. A company's culture can transform as individuals are empowered to become more data-driven. ●

REFERENCES

Baiya, Evans [2020]. "Why agility is key to companies surviving the pandemic," *Fast Company*, June 18. https://www.fastcompany.com/90516862/why-agility-is-key-to-companies-surviving-the-pandemic

Conner, Ethan [2019]. "Create a BI Dashboard in 5 Minutes (Video)," July 22. https://wyn.grapecity.com/blogs/create-a-bi-dashboard-in-5-minutes

Dressner Advisory Services [2020]. "Research Insight: New Findings on How COVID-19 Impacts Businesses, Budgets, and Projects," May 8. https://www.patreon.com/posts/new-findings-on-36892231

Gallo, Jim [2018]. "Self-Service BI: Barriers, Benefits, and Best Practices," TDWI, April 24. https://tdwi.org/articles/2018/04/24/bi-all-self-service-bi-barriers-benefits-and-best-practices.aspx

Goswami, Suparna, Nandikotkur, Geetha [2020]. "Addressing Shadow IT Issues During COVID-19 Crisis," *Bank Info Security*, April 17. https://www.bankinfosecurity.com/addressing-shadow-issues-during-covid-19-crisis-a-14137

Hirt, Martin, Laczkowski, Kevin, and Mysore, Mihir [2019]. "Bubbles pop, downturns stop," *McKinsey Quarterly*, May 21. https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/bubbles-pop-downturns-stop

Horwitt, Elisabeth [2011]. "Self-Service BI Catches On," *Computerworld*, Jan 24. https://www.computerworld.com/article/2549837/self-service-bi-catches-on.html

Muncaster, Phil [2020]. "Shadow IT Represents Major #COVID19 Home Working Threat," *Info Security*, April 7. https://www.infosecurity-magazine.com/news/shadow-it-covid19-home-working/

Soucheray, Stephanie [2020]. "US job losses due to COVID-19 highest since Great Depression," University of Minnesota Center for Infectious Disease Research and Policy, May 8. https://www.cidrap.umn.edu/news-perspective/2020/05/us-job-losses-due-covid-19-highest-great-depression

Enterprise Business Management System for Strategic Decision Making

Gaurav Anand is the head of business finance at Google Cloud.
gaurav.anand@gmail.com

Gaurav Anand

ABSTRACT

Most enterprises today want to be data-driven, and yet most large organizations also agree that they are still far from realizing this goal. What organizations lack is a framework for viewing data and generating insights as well as a common language to speak while making data-driven decisions. In this article, we propose a framework called the business management system (BMS) to aid in decision making. We show how BI teams can implement, manage, and drive adoption of this framework.

This article explains how a business unit could design, build, and put the BMS into practice to allow for a more effective and data-driven way to make decisions. We also show how to bring together business analysts, business intelligence, operations, and leadership to instill and manage a data-driven culture for the larger workforce. As an example, we've mocked up a BMS for the sales function, but the concept is applicable in most areas.

INTRODUCTION

Most enterprises today want to be data-driven. The general consensus is that competitive advantage and growth can be realized by using data more effectively. Yet most large organizations also agree that they are still far from realizing this goal. In fact, in a 2019 survey by New Vantage Partners, 72 percent of executives admitted that they have yet to forge

a data-driven culture. Over half (52 percent) responded that they are not competing on data and analytics.

Some of the key reasons companies are not able to drive towards a more data-oriented culture include:

- Lack of skills and talent
- Lack of executive sponsorship
- Cultural challenges

There are many areas where data-driven decision making can lead to competitive advantages, cost reductions, and faster speed of execution. Troy Hiltbrand speaks about continuous intelligence and the need to automate or partially automate the scaled decision-making process and, in some cases, augment it with data-based insights to create better human decisions (Hiltbrand, 2019). In this article we focus on strategic decision making in enterprises and how we can instill data-based rigor and discipline in the process.

CURRENT PROBLEMS WITH STRATEGIC DECISION MAKING

More often than not, in business reviews and strategy meetings, executives have positions about business performance and its future course based on past data. Differences in opinion arise because the data has been presented at different times in different contexts to different executives and leaders are inclined to question the data's veracity rather than understand the information.

The second issue organizations face is introducing changes based on a decision to the affected workforce. We see a pattern of management communicating decisions to employees in a conference, meeting, or email that leads to a scramble to update tools, guidelines, data, and processes. This can result in several rounds of noncompliance until the teams actually understand how to implement the changes, by which time the decision appears less effective.

Finally, there is usually no feedback loop to inform executives in real time about performance impacts due to the changes introduced. Key performance indicators (KPIs) and metrics are usually designed to create a more top-down view and are ill-equipped to provide bottom-up insights.

We propose a framework called the business management system (BMS) to aid in decision making to solve these problems. We also show how BI teams can implement, manage, and drive adoption of this framework. A business management system is defined as "a set of policies, practices, procedures, and processes used in developing and deploying strategies, their execution, and all associated management activity" (CIO Wiki, 2018). An enterprise BMS is designed to:

- Provide a common language for executives to understand, measure, and discuss business performance
- Create tools and processes that embed desired changes in the incentive system for the workforce almost instantaneously
- Better measure and attribute results in a statistical manner

Our framework includes three key steps: design, build, and implement.

BUSINESS MANAGEMENT SYSTEM (BMS) DESIGN

The design phase of BMS is usually led by business analysts or strategy analysts and is based on building a KPI framework that borrows heavily from Avinash Kaushik's KPI framework (n.d.). The framework is further solidified as the "Top-down framework" in an article by Douglas Holmes (2018). It consists of:

1. Identifying top-level KPIs to measure the business objective
2. Segmenting the business based on KPI behavior
3. Identifying drivers that affect performance of the segments
4. Calibrating actions that move the drivers

1. Identifying Top-Level KPIs

Any business performance measurement design usually revolves around key performance indicators. For BMS, we start with the most important strategic objective of the business unit: the reason for the existence of the organization/team.

A smart way to identify business objectives is to inspect the incentive systems. The industry-standard incentives are designed to maximize performance towards achieving the key business objective. For example, the key business objective for the sales team is to increase overall sales for the company. For marketing, it could be customer acquisition. For product organizations, the goal is usually about engagement; customer service organizations usually strive to increase customer satisfaction. Control functions (such as finance) are tasked to ensure the incentives are devised so the success of one unit is not detrimental to another.

There are many well-established methods for identifying strategic objectives for a business unit. Some common frameworks include SWOT analysis, balanced scorecard, OKR methodologies, and Porter's five forces analysis.

In this article, we will not discuss how to identify your primary strategic objective. Rather we will assume that this step has already been completed and move to a discussion about managing the business keeping the objective in mind.

Once the business objective is identified, the next step is to create a KPI that aligns with the incentive attached to the business objective. If no incentive is attached, the KPI should directly measure the business objective. This is the one metric that can be described as the true North Star for the business unit. This KPI must be given a target during the planning cycles and most of the initiatives should be geared to achieve that target. Examples here could include revenue forecast (versus targets) for sales, opportunity size created for marketing, or monthly active users (MAU) for product teams. (See Figure 1.)

2. Segmentation

Segmentation is one of the trickier steps within building this framework and requires an innate understanding of the data underlying the business performance. To define the task at hand here, we need to create segments within our customer/user/product base that *meaningfully differ in behavior with respect to the top-level KPI*. Customer segmentation is a well-researched concept. According to the steps laid out by Judy Bayer and Marie Taillard (2013) in

	BUSINESS OBJECTIVE EXAMPLES	TOP-LEVEL KPI EXAMPLES
Sales	Maximize revenue and future revenue	Forecast versus quota
Marketing	Acquire potential customers with the highest chance of conversion	Conversion rate opportunity generated
Product	Ensure best possible experience for the customers using the product	Monthly active users (MAU) Daily active users (DAU)
Customer support	Support and retain customers	Customer satisfaction

Figure 1. Business objective and KPI examples for select business units.

their article "A New Framework for Customer Segmentation," we must:

- Identify the context in which customers are using the company's products
- Combine information about transactions and customer behavior in the contexts to describe each of the jobs to be done
- Map individual customers to jobs using the data

To contextualize the steps for our purposes, the "job to be done" here is to optimize business performance based on the top-level KPI. We must thus identify segments that behave differently as measured by this KPI. Bear in mind that segmentation may differ between functions within the same company. What works for product teams may be different from the segments used by marketing.

Consider sales as an example. Most companies segment their customers based on size, primarily because different sales strategies are applied to those segments. Usually larger customers get high-touch treatment while smaller ones are managed using a more scaled approach. Although this works well in most cases and is aligned with marketing, which is "feeding the funnel," these segments may themselves not be enough. The larger customer base may then have to be further segmented into customer life-stage (e.g., new, growing, mature, declining) or based on portfolio usage (how many products the customer uses regularly). The scaled segment might be further divided based on a customer potential model.

The objective here is to flesh out segments that differ in how they contribute to sales achieving their bookings targets, which is our KPI for sales. (See Figure 2.)

The final exercise is to officially define these segments from a data perspective so the systems and the business intelligence teams can tag the customers accordingly in the data and then assign targets to each of the segments. For example, we could define "Greenfield Large" customers as those identified as having a large revenue potential but which currently spend

less than $10K annually. Customers identified as "Growing" could be defined as those who made recent purchases and exhibit high growth rates because they are still ramping up their use of the product. "Mature Large" customers could include product users who seem to have tapered off in growth (regardless of whether further opportunity exists). "Declining" customers are those whose revenue/product usage has declined over two consecutive cycles.

Each segment can be assigned a portion of the overall target so executives can understand how their business performance risk is distributed. Target assignment is generally based on:

- Historical trends
- Overall company objectives
- Competitive benchmarking

This target must be broken down by individual teams/actors that can deliver against their own target. This is usually done using a model that takes into account the historical performance of the actor, the resources allocated, and the business requirement. For example, a sales rep is usually assigned a quota that is a portion of the overall target for the sales organization.

We'll deep dive into the sales example shortly, although BMS could be operationalized for any type of business.

3. Identifying Drivers

With the segmentation exercise complete and segments and targets defined, we now need to understand what drives these segments so we can take appropriate actions to measure and improve performance. We need to ask the following questions to understand the drivers:

- What are the indicators that the segment is performing on track to meet its targets?
- Is performance well-distributed across the segment or are one or two outliers skewing it?
- If the segment is not on track, what course correction is needed?

These business drivers are not actions but metrics that have been generated using either knowledge of the industry (qualitative) or

TOP-LEVEL KPI	INITIAL SEGMENT	FINAL SEGMENT
Sales (Forecast)	Large customers	Greenfield
Marketing (Conversion)	Midsize customers	Growing
Product (MAU)	Small customers	Mature
		Declining

Figure 2. Fleshing out KPI segments.

certain historical analysis (quantitative). They are usually function-specific and follow the workflows innate to the industry and function. We can also think of these drivers as nodes on a decision tree, which dictate or at least heavily influence the segment performance. In some cases, the drivers are evident, while in others, drivers need to be created using data insight and business consensus.

For the enterprise sales case, Figure 3 illustrates how we could create drivers at the segment level. For each driver, we must also assign targets we expect to see over the course of the business cycle corresponding to the targets set at the segment level for each team/actor.

4. Calibrating Actions

The final step in designing the BMS is determining what actions need to be taken to move the drivers in the desired direction. These are representative of the activities we want the organization to undertake and focus on doing well. These actions are what ultimately trickle down as operational guidance from the top-level strategic decisions.

LARGE CUSTOMER SEGMENTS	DEFINITION	ANALYSIS	DRIVER
Greenfield	A customer with no spend in past 12 months (an arbitrary time period; the actual number would depend on the product adoption rates and sales cycles)	To increase greenfield segment bookings, we must acquire more new customers and sign them to larger deals	Number of new customers signed with a deal size above $100,000
Growing	A customer who purchased in the past 12 months and is presumably growing fast	To ensure we hit our target for this segment, these customers need to maintain a specific growth rate as a cohort	Number of customers lagging expected growth
Mature	A customer who has purchased over 12 months ago and exhibits no growth	Because targets for this segment will be higher than current growth rate, we need to up-sell/cross-sell to induce growth	Number of customers with an up-sell/cross-sell deal in pipeline for current cycle
Declining	A customer who exhibits decline in revenue over the past two quarters	Targets for such segments are invariably set with lower expectations, but the goal is to explore if we can stabilize revenue decline and prevent churn	Number of customers with an up-sell/cross-sell deal in pipeline

Figure 3. Creating drivers at the segment level.

Actions should be determined and/or verified using historical data. There must be significant confidence that a chosen action will significantly advance the driver metric. Statistical methods such as regressions or decision trees can be used to examine the correlations and causality. If we want to maximize the number of large deals won, and we want to understand what actions to provide to our sales reps to maximize this number, we can create a decision tree using all the features that may affect the win rate.

In Figure 4, it is clear that conducting a proof of concept with a customer is the biggest driver for winning new deals and would be the action metric for the win-rate driver. We would ideally come up with a set of three to five primary actions that we believe affect our most important driver metrics to achieve their targets. We would then also create targets for the action metrics based on the targets derived for the drivers at individual levels. For our sales example, we could end up with the structure shown in Figure 5.

BUSINESS MANAGEMENT SYSTEM: BUILD

Once we have the design framework in place, we need to build out the data sets and tools to support the implementation of the BMS. The business intelligence teams need to work closely with their strategy and operations counterparts to deliver a product that is based on consistent data and reflects the metrics built out in the design phase. The tool here could refer to a set of dashboards or a more interactive UI that distills the essence of the design into actionable insights.

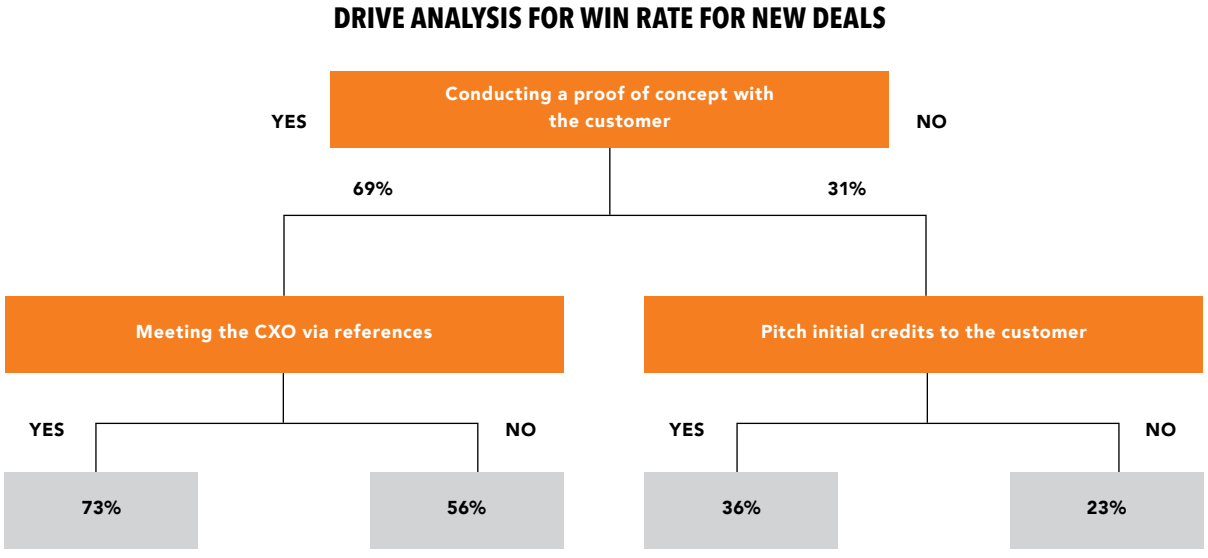


Figure 4. Decision tree for a driver.

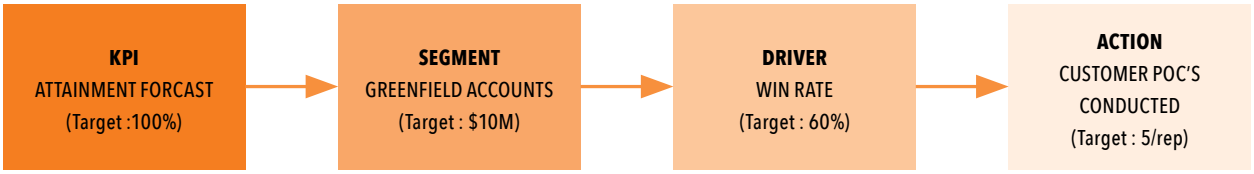


Figure 5. Action sequence for a KPI.

There are three features to keep in mind for the build:

- Data reliability
- Scorecard
- Drill-down ability

1. Data Reliability

An O'Reilly survey shows how data quality is still plaguing organizations with top factors being "Too many data sources and inconsistent data" and "Disorganized data stores and lack of metadata" (Magoulas and Swoyer, 2020). For an organization to leverage a framework such as the BMS, data quality must reach a level that encourages both senior leadership and rank-and-file members to trust it. Those making strategic decisions and those taking actions based on those decisions need to speak the same data language and not be constantly distracted by data quality issues.

To instill this confidence in the BMS, the BI teams must:

- Create a central data mart that serves as the source of truth for all reporting, including the BMS
- Create comprehensive documentation explaining the definitions and calculations of the metrics
- Provide a way to quickly correct inconsistencies in data as part of the data governance process

2. Scorecard

The UI for the BMS finally distills down to a scorecard that connects the targets for each individual team/actor at all levels of design. The scorecard should show the actor the top-level targets, driver targets, and the targets for the actions. By providing a condensed view of targets and focusing more on actions (we need not share the targets at segment level), we're placing emphasis on individuals performing their actions with the appropriate rigor—with the conviction that our framework is sound enough to capture the effects of those activities on business results. Figure 6 shows a sample sales scorecard.

3. Drill-Down Ability

An aggregate metric is great for showing directionality, but for individuals to understand where to target their next action, they need to drill down to a much more granular level based on how the metric was created. An action metric designed to spur an action at the

customer level should be drillable right down to individual customers so individuals can see what customers to target. For an action that focuses on reducing SKU backlogs, details of individual SKUs must be provided. Ideally this drill-down functionality is directly linked into the scorecard so the scorecard can act as the control panel; the drill-down guides the actor to what actions to take. (See Figure 7.)

BUSINESS MANAGEMENT SYSTEM: IMPLEMENTATION AND MONITORING

The BMS implementation and monitoring phase can begin in parallel with the build phase. This phase consists of several tasks.

Align Metrics at the Leadership Level

This is the most important implementation step and may require multiple sessions with the executive team. In this step, business analysts create a view of the entire BMS (top line KPI, segments, drivers, and actions) sliced by key dimensions (usually mirroring the organizational structure such as region and market). This view (either in the form of a dashboard or presentation) should be presented to leaders, carefully explaining why each metric was chosen and how the drivers are correlated with actions. The primary objective is to ensure that the entire leadership team understands and buys into the framework and is incentivized to focus on chosen actions for their teams.

Launch the BMS Tool with Documentation and Training

Once the tool is ready, marketing should promote the tool and its benefits. Because we want the scorecard to be the primary view our workforce uses, we must deprecate or remove other dashboards that may repeat the same information or be distracting. Additionally, sufficient training and documentation must be put in place so the workforce can understand

Q3'19 Scorecard

	Target	Current	%	WoW	MoM
Forecast	5M	3M	60%	+0.5M	+1M
New Customer Pipeline	3M	1M	33%	0M	+0.7M
# of POCs	5	2	40%	+1	+1
# of CXO meetings	10	5	50%	+2	+3
Upsell Pipeline	1M	0M	0%	0M	0M
# of Sales engineers / customer	2	0.8	40%	+0.2	+0.4
# of CXO meetings	20	8	40%	+1	+3

Figure 6. Sample sales scorecard.

the actions expected of them, the targets for those actions, and the effect those actions will have on the enterprise's KPIs.

Establish Regular Business Reviews

Business reviews for the leadership team are run based on the BMS framework. Depending on the frequency of business reviews, the analyst team needs to prepare the BMS views with numbers well in advance. During the preparation phase, the business teams should weigh in on the numbers to explain anomalies or lags in the business and prepare a response about how they plan to fix them. The business review conversation should then focus on leaders putting forth plans to manage their businesses and sharing best practices.

At this point, all leaders in the review should be comfortable with the BMS and the metrics and targets in the framework. They should be questioning their managers about why certain indicators are lagging or why particular activity levels are anemic so as to drive the data-based rigor in the day-to-day functioning of the organization. Driving a management by exception method is key to adopting this framework at all levels.

Use the BMS in Planning

Because the BMS is the primary view of the business, business planning should use the BMS framework so the appropriate business targets can be chosen. Logic suggests that we place more resources where we get the most return on investment. Correlation coefficients between drivers and actions could suggest where to put more resources, assuming there is sufficient business opportunity. As an example, if revenue is higher in one region than the other, one idea would be to put more sales resources in that region. Another idea would be to assign a few enablement resources in the lagging region so they can train the salesforce to improve this metric.

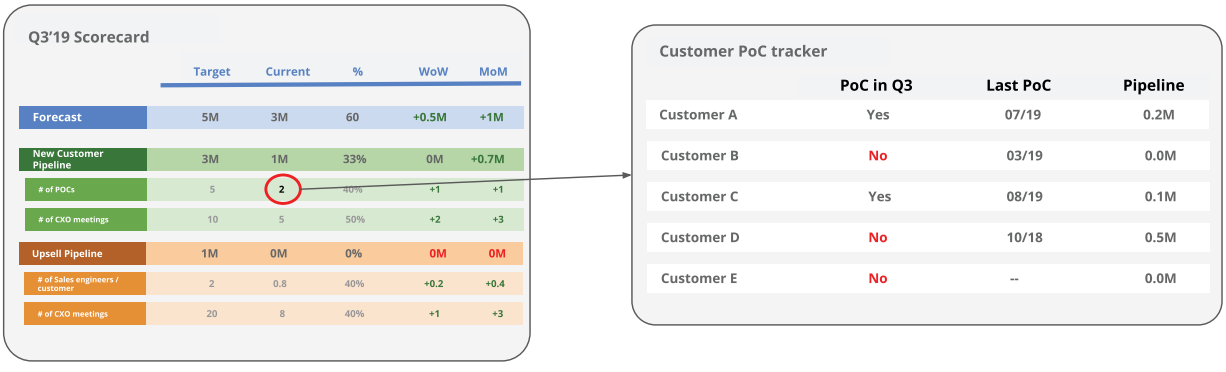


Figure 7. Scorecard drill-down.

Monitor the KPI Framework and Change Management

Similar to a machine learning model that degrades in performance over time if not frequently trained, the BMS framework could get less effective if the dependencies between metrics change and the framework is not adjusted. Actions such as providing credits to declining customers may have worked in the past, but over time this strategy may stop working as competitors catch up to the practice and/or drop pricing. New drivers and new actions must be constantly considered as part of this monitoring to ensure the framework remains valid and the organization is not being guided towards ineffective actions.

When leadership makes decisions that change organizational priorities, it also changes the underlying KPIs, metrics, and actions. Once the BMS is ingrained into the daily functioning of the workforce, allowing the change to percolate is much easier by changing the BMS metrics and targets than without such a mechanism. Change is better communicated when the supporting tools and processes reflect the change quickly, which is what the BMS provides.

COMPARATIVE FRAMEWORK ANALYSIS

Many frameworks serve as a strategic management tool. The most widely used closed-loop frameworks, that include feedback from actual performance to make course corrections, include the balanced scorecard; objective and key results (OKR) methodology; results-based management, European Foundation for Quality Management (EFQM); and specific, measurable, achievable, relevant, time-bound (SMART) goals. Figure 8 compares the popular balanced scorecard, OKR methodology, and BMS.

Based on the comparison, it is clear that different frameworks can be used to cater to different audiences. In most cases, a combination of OKR and balanced scorecard or OKR and BMS is beneficial to running the business more effectively and creating a data-driven culture, depending on the type of business. For example, the company board and C-level leaders could use the balanced scorecard to understand and create strategic objectives for a longer period (such as one to two years). OKRs can be used to break down the longer-term objectives into shorter-term goals. The same company could also directly choose their primary objectives using OKR methodology and then use the BMS to cascade objectives using the data-calibrated framework. The BMS works best when:

- The business can measure most or all activities and outcomes
- There is less ambiguity in the interpretation of the objectives so business control—not business understanding—is the key requirement (for example, when well-established functions are looking to run their daily operations more effectively)
- A larger-scaled workforce needs fewer, simpler key performance metrics

The business management system is a consolidated effort between various teams: leadership, business analysts, business intelligence, IT, and strategy analysts. If designed, built, and implemented correctly, BMS provides a common, data-driven language for everyone to understand the business, make informed decisions, and manage change effectively. ●

	BALANCED SCORECARD	OKR METHODOLOGY	BMS
Frequency	Usual target time frames are one year or greater because it aims to align on a longer-term strategy	Usually quarterly, even if part of a larger strategic objective	Biweekly or monthly because it targets a more operational, hands-on approach
Philosophy	Holistic strategic view	Management by objectives	Management by exception
Design Principle	"Four perspective" approach (financial, customer, internal processes, growth) paired with a strategy map	Cascaded strategic objectives interpreted, internalized, and implemented at every level	Hierarchical metric system calibrated on data to provide control and flag exceptions to managers
KPI versus Objective	KPIs follow the objectives	KPIs follow the objectives	Aside from the top-level objective, most objectives built on data-based observations
Cascading	Relevant goals and KPIs from higher-level scorecards are filtered down to lower levels	Objectives are filtered down, with every management level partaking in efforts to achieve the target	Different KPIs provided to different management levels based on what they can act upon
Strength	Provides long-term vision for a company by considering various perspectives	Provides managers flexibility to interpret objectives, choose direction, and create buy-in	Provides visibility and operational course-correction for a well-understood business unit

Figure 8. Three methodologies compared.

REFERENCES

Bayer, Judy and Marie Taillard [2013]. "A New Framework for Customer Segmentation," *Harvard Business Review*, June 12. https://hbr.org/2013/06/a-new-framework-for-customer-s

CIO Wiki [2018]. Definition of business management system, last edited on December 6, 2018. https://cio-wiki.org/wiki/Business_Management_System_(BMS)

Hiltbrand, Troy [2019]. "Reducing Friction for Continuous Intelligence," *Business Intelligence Journal*, Vol 24, No.1

Holmes, Douglas [2018]. "Key Performance Indicator (KPI) Frameworks," blog post, DouglasHolmes.com, January. http://douglasholmes.com/key-performance-indicator-kpi-frameworks/

Kaushik, Avinash [n.d.]. "Digital Marketing and Measurement Model," blog post, Occam's Razor, accessed August 3, 2020. https://www.kaushik.net/avinash/digital-marketing-and-measurement-model/

Magoulas, Roger and Steve Swoyer [2020]. "The State of Data Quality in 2020." https://www.oreilly.com/radar/the-state-of-data-quality-in-2020/

New Vantage Partners [2019]. "Big Data and AI Executive Survey 2019." http://newvantage.com/wp-content/uploads/2018/12/Big-Data-Executive-Survey-2019-Findings-Updated-010219-1.pdf

The 5 "A"s of AutoML

Arup Duttaroy

Arup Duttaroy is an experienced information management and business analytics practitioner currently employed with Larsen & Toubro Infotech Ltd. in Pune, India.

adroy@yahoo.com

ABSTRACT:
Recent advancements in big data, cloud computing, and other technologies have created artificial intelligence (AI) and machine learning (ML) technologies that almost any company can use to find practical answers to difficult business problems. However, AI and ML still have high barriers to entry; they require expertise and resources that few companies can afford on their own.

In recent years, the demand for ML experts and data scientists has outpaced the supply despite the surge of people entering the field. To ease this shortage, automated machine learning (autoML) is gaining interest in the advanced analytics community. Simply stated, autoML is the process of automating the rigorous, end-to-end procedure of applying machine learning to real-world problems.

This article details which machine learning tasks can be really automated to usher in error-proof productivity gains and promote democratization of data science. It also introduces some of the popular ML tools that offer autoML features.

Hype or not, autoML is not only here to stay, but it's clear that autoML is going to be a big part of the future of machine learning.

INTRODUCTION

The term autoML has two parts. The first—auto—stands for automation, which is simply the means of executing manual, repetitive tasks without human intervention using technology. The second part of the term stands for machine learning—an application of artificial intelligence, which is the science

of getting computers to learn and act like humans and improve their learning over time in autonomous fashion by ingesting data and information in the form of observations and real-world interactions.

Machine learning and artificial intelligence have provided many significant breakthroughs in diverse fields in recent years. Financial services, healthcare, retail, and transportation industries (among others) have been using machine learning systems for some time, and the results have been promising. Recent advancements in big data, cloud computing, and other technologies have put AI and ML within reach of many companies, enabling them to find practical answers to difficult business problems.

As companies scramble to get a competitive edge using AI and ML, they're learning an important lesson: these aren't "one strategy fits all" technologies. When it comes to handling large amounts of data, often the machine learning projects don't work in the real world as well as enterprises would like. That's in part because many solutions that incorporate AI/ML capabilities use predetermined algorithms and data handling techniques. In reality, each organization's system data has specific characteristics that might not fit the predetermined machine learning algorithm's configuration.

Machine learning isn't magic. Getting accurate results requires a data scientist who can study the input data, understand the desired output to solve a business problem, choose from dozens of mathematical algorithms, tune those algorithm's parameters (called hyperparameters), and evaluate the resulting models. If the results aren't sufficiently accurate, the data scientist can adjust the algorithm's tuning parameters repeatedly until the machine-learning model produces the desired results. If the results don't improve, the data scientist might even start the process over from scratch, using an entirely different ML algorithm to see if it can better model the training data. Building a model is truly a rigorous process.

AI and machine learning are still fields with high barriers to entry; they require expertise and resources that few companies can afford on their own. In recent years, the demand for machine learning experts or data scientists has outpaced the supply. To address this gap, enterprises now can employ user-friendly machine learning software that can be used by non-experts.

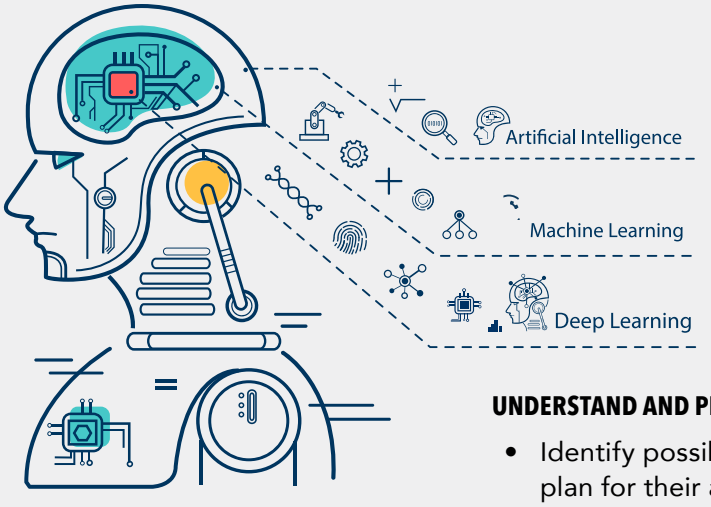
The first step to simplifying machine learning involved developing simple, unified interfaces to a variety of machine learning algorithms. Although these tools have made it easy for non-experts to experiment with machine learning, users still need a fair bit of knowledge about and background in data science to produce high-performing machine learning models. Deep-learning algorithms (such as those for neural networks) are notoriously difficult for a non-expert to tune properly. AutoML may be the answer to such situations.

Before we dive deeper into the details of AutoML, let's examine what data scientists or machine learning practitioners actually do.

THE ROLE OF A MACHINE LEARNING PRACTITIONER

Building a machine-learning-powered solution is a multifaceted and complex task. Table 1 lists

Tasks of a Machine Learning Practitioner



UNDERSTAND THE CONTEXT AND BUSINESS REQUIREMENTS

- Develop an understanding of the business problem, the associated risks, the overall business objective, and the desired outcomes; share this knowledge with other analysts and business users
- Break down business problems into smaller hypotheses that, when solved, give the predictive or prescriptive outcomes sought
- Identify what kind of data the organization has to solve these hypotheses
- Frame and scope the analytics task
- Understand operational constraints (e.g., what data is actually available at inference time)
- Identify potential biases and potential negative feedback loops

UNDERSTAND AND PREPARE THE DATA

- Identify possible sources of data and plan for their acquisition as needed (and if possible)
- Integrate data from multiple sources; often this data has been collected in different formats or with inconsistent naming conventions
- Perform exploratory data analysis, look into data profiles and associated metadata information to understand the granular data elements which will be helpful for the experiments
- Handle missing or corrupted data; de-dupe data if required
- Use data visualizations to identify outliers, data trends and patterns, and distribution characteristics that can potentially infuse unnecessary bias or imbalance
- Identify features that can contribute to the desired outcome of the models
- Perform feature engineering to transform data types as needed for specific models/algorithms

Table 1. ML tasks.

- Create training, validation, and test data sets

BUILD AND TRAIN MODELS

- Choose the appropriate model/algorithm to use
- Identify constraints (e.g., will the completed model need to run on an edge device, in a low-memory or high-latency environment, etc.)
- Choose hyperparameters (e.g., in the case of deep learning, this includes choosing an architecture, loss function, and optimizer)
- Create an appropriate scoring function
- Train and debug the model; this can involve:
  - Adjusting hyperparameters (e.g., the learning rate)
  - Persisting intermediate results to see how the loss, training error, and validation errors are changing over time
  - Identifying underlying errors or issues with the data
  - Realizing the need to change the data cleansing and pre-processing
  - Recognizing the need for augmenting more or different data
  - Trying different models
  - Determining if the models are becoming under- or over-fitted
- Evaluate the accuracy and other model performance measures in order to select the one with the highest accuracy

OPERATIONALIZE AND DEPLOY MODELS INTO PRODUCTION

- Create an API or web app with the model as an endpoint to integrate with downstream applications
- Perform model A/B testing and select the best one
- Track model versions, including which version is deployed; track performance and accuracy of each version
- Export the models into needed format(s)
- Persist the model's output into the underlying data storage for downstream consumption as applicable
- Plan how often the models will need to be retrained with updated data (prediction accuracy deteriorates over time)

MONITOR AND TRACK OF MODELS POST DEPLOYMENT

- Monitor model use and performance
- Monitor input data to identify if it changes with time in a way that would invalidate the models
- Communicate the results (i.e., the model outcomes) to the rest of the organization
- Explain model output
- Develop and implement a plan for monitoring and responding to mistakes or unexpected consequences

many of the tasks machine learning practitioners do during the process.

Table 1's multitude of steps shows how rigorous and effort-intensive the role of the machine learning practitioner is. According to the 2018 *Kaggle ML and Data Science Survey*[1], 15–26 percent of a typical data science project time is devoted to model building or model selection. That means the rest (74-85 percent) of the time is spent on the other activities. Certainly, not every machine learning practitioner needs to do all of the steps shown in Table 1, but components of the process will be part of most machine-learning solution building. Even if one is working on just a subset of these steps, familiarity with the rest of the process will help ensure that you are not overlooking considerations that would hamper project success.

Once we understand the typical day-to-day tasks of a data scientist, we are in a better position to understand and appreciate autoML. Automated machine learning is just what it sounds like—automation of the end-to-end process of applying machine learning to real-world problems. Although there is no substitute for skillful problem definition and data preparation in machine learning, autoML takes on many of the repetitive tasks, reducing the need to understand algorithm parameters and shortening the compute time needed to produce better models.

Although autoML makes machine learning available even to people with no major expertise in this field, it can also be a helpful tool for the most proficient data scientist by providing a simple wrapper function that performs many modeling-related tasks that would typically require many lines of code, and by freeing up their time to focus on other aspects of the data science pipeline tasks such as data-preprocessing, feature engineering, and model deployment.

Let's now look at all those aspects of a machine learning workflow that can really be automated. In my opinion there are five such areas and hence the 5 A's of autoML.

THE 5 "A"S OF AUTOML

Many people feel the autoML field is too overhyped. Let's look closer at five aspects of the machine learning workflow that can be automated in pragmatic ways.

#1: Automated Data Preparation and Ingestion

Once all the data sets needed to solve the hypothesis have been identified and acquired in a central location (such as a data lake or a data warehouse), exploratory data analysis and data preparation can begin.

Dealing with data formatting issues, inconsistencies, and errors is often a messy and tedious process. Frequently the root cause of issues during model training is related to the input data itself—data that data scientists must revise in pre-processing and then input again, which can be time-consuming. Things that can be automated at this stage using autoML capabilities include:

- Automated determination of strategy for pre-processing the data: how to deal with imbalanced or skewed data; how to impute missing values; remove, replace, or keep outliers.

1 See https://www.kaggle.com/headsortails/what-we-do-in-the-kernels-a-kaggle-survey-story/execution

- Automated column type detection; e.g., Boolean, discrete numerical, continuous numerical, categorical, text, date, or time-series data.
- Automated data profiling showing missing counts, descriptive statistical measures, variability measures, histogram for the distribution of continuous data elements, or boxplot showing quartiles and outliers (if any).
- Automated tagging to enterprise data glossary or taxonomy, auto relationship identification between data sets, auto identification of sensitive data elements such as personally identifiable information (PII) fields.
- Automated sampling of training/validating/scoring data. There might be too much data, but it's difficult to know which subset of data to use for training a machine-learning model. In some cases, training using some data variables or predictors can increase training time while actually reducing the accuracy of the ML model. It's not easy to achieve significant size reduction without affecting accuracy, but with care it can be done. AutoML may be a good option here to automate creating small, accurate subsets of data to use for those iterative refinements, yielding excellent results in a fraction of the time.

All this automation will be made possible by using scripts running in the background. Such scripts will probably use rule-based logic at the beginning, but ML/AI-based automation will also be needed, which is offered by some of the toolsets to be discussed later.

#2: Automated Feature Engineering

Feature engineering is the process of selecting and extracting features from raw data using data mining techniques. A feature is an attribute or property shared by all of the independent units on which analysis or prediction is to be done.

Any attribute can be a feature as long as it is useful to the model. The purpose of a feature, other than being an attribute, would be much easier to understand in the context of a problem. A feature is a characteristic that might help solve the problem. These features can be used to improve the performance of machine learning algorithms.

Any attribute can be a feature as long as it is useful to the model.

For example, let's assume a raw data set used for training a time-series model has a column named "Sales_Date" that contains date values in the format MM/DD/YYYY. You can apply feature engineering to generate seven more columns, including Year (YYYY), Month (MM), Day (DD), Quarter (QQ), Days Since_Beginning (NNNNN), Is_Weekend (Boolean, 0 or 1) , Day_of_the_Week (N), which, in turn, can be used as features or independent variables.

Hot encoding for categorical variables is another example. Let's say a column named Gender_of_Earning_Members has four possible values: Male, Female, Both, and Not Applicable. Through feature engineering, you can generate

four additional code fields with a combination of values 1, 2, 3, and 0 as the case may be.

Feature engineering can be considered as applied machine learning itself and as such a vital component of autoML:

Automated feature selection. A business problem's data set might have dozens, hundreds, or even thousands of variables (predictors) that a model can consider, so it's not easy to tell which of the data points are significant for making a decision. The process of selecting the most relevant information to include in a data model is called "feature selection." AutoML can come in handy for generating new features and selecting the meaningful ones.

Automated column intent detection. This can include target/label, stratification field, numerical feature, categorical text feature, or free-text feature. These features can be used in a "step-up/step-down" method of "n" building supervised or unsupervised models, which involves iterations and performance checking before arriving at the best-fit model. All of these iterations can be automated, an important characteristic of autoML.

Automated feature extraction. Feature extraction involves reducing the number of resources required to describe a large data set. When performing analysis of complex data, one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computational power; it may also cause a classification algorithm to overfit training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Many machine learning practitioners believe that properly optimized feature extraction is the key to effective model construction. However, it is a tedious process and involves techniques such as dimensionality reduction, principal component analysis (PCA), datatype conversion, encoding, and vectorizing, to name a few. These techniques can be implemented through rigorous coding in languages such as Python and R and therefore are tedious, error-prone, and time-consuming. However, these techniques are amenable for autoML much to the relief of data scientists, especially non-expert citizen data scientists, such as business analysts.

#3: Automated Model Selection

Building and training the model is at the heart of the data science/machine learning workflow. There are many well-known algorithms for machine learning, and it's not always obvious which algorithm will work best for building real-value prediction, anomaly detection, or classification models for a particular data set.

In complex, real-world situations, a data scientist may need weeks or months to choose the right algorithm and refine the model created using that algorithm. The difficulty of training models deters many beginners, who often wind up feeling discouraged. Even experts frequently complain of how frustrating and fickle the

training process can be. That is where autoML can be a real savior.

Let's examine the areas where autoML improves productivity in the model selection stage:

Automated task detection. Model selection may also refer to the problem of selecting a few representative models from a large set of computational models for decision making. Variables include the business problem to be solved and optimization uncertainty, given the available data. It involves determining the kind of learning algorithm needed to solve the business problem (i.e., whether it needs a supervised or unsupervised learning model, or is it a reinforcement learning use case, and the like). Within each learning category, you must determine which specific algorithm is needed (e.g., binary classification, linear or polynomial regression, clustering, association, ranking, etc.).

Automated model selection. Earlier we explained that about 15 to 26 percent of a typical data science project's time is devoted to model building or model selection. It's a demanding task both in terms of "person-hours" and elapsed hours for computational throughput. If the main objective or underlying data changes (for example, if new features are added), the process needs to be repeated from the beginning. AutoML can help data scientists save time and spend it on more important tasks.

Automated meta learning and transfer learning. AutoML uses machine learning to choose and optimize the machine-learning pipeline, a technique called meta learning. Simply stated, meta learning is a subfield of machine learning where automatic learning algorithms are applied to the metadata about machine learning experiments. The main goal is to use such metadata to understand how automatic learning can become flexible in solving learning problems and improve the performance of existing learning algorithms or to induce the learning of the learning algorithm itself.

Transfer learning, on the other hand, is a powerful technique that lets people with smaller data sets or less computational power achieve state-of-the-art results. Transfer learning takes advantage of models that have been trained on similar, larger data sets. Because the model learned via transfer learning doesn't have to learn from scratch, it can generally reach higher accuracy with much less data and computation time than models that don't use transfer learning. This is another feature in the toolkit of autoML.

Automated creation of ensemble models. Another interesting trick of autoML is the default enablement of stacked ensembles, which is a "model-of-models" method that is based on all previously trained models; another ensemble model can be built using the best model of each family. Such ensemble models are automatically trained on collections of individual models to produce highly predictive ensemble models which, in most cases, will be the top performing models in the autoML leaderboard. The ensemble iterations appear as the final iterations of all autoML batch runs, if enabled. Automated machine learning uses both voting and stacking ensemble methods for combining models:

- The voting method prediction is based on the weighted average of predicted class probabilities (for classification tasks) or

predicted regression targets (for regression tasks).

- The stacking method combines heterogeneous models and trains a metamodel based on the output from the individual models. The current default metamodels are LogisticRegression for classification tasks and ElasticNet for regression/forecasting tasks.

#4: Automated Hyperparameter Tuning and Optimization

In the machine learning workflow, the next step is to tune and optimize your hyperparameters. In machine learning, a hyperparameter is a parameter whose value is set before the learning process begins. To give a clearer picture, some of the hyperparameters are the "learning rate" for training a neural network, "C" and "sigma" values for support vector machine (SVM), or the "k value" in k-nearest neighbors (KNN).

Hyperparameters can be classified as **model hyperparameters** (such as topology or size of neural networks) that cannot be inferred while fitting the machine to the training set because they refer to the model selection task or **algorithm hyperparameters** (such as learning rate or mini-batch size) that in principle have no influence on the performance of the model but affect the speed and quality of the learning process.

Different model training algorithms require different hyperparameters. Using these hyperparameters, the training algorithm learns the parameters from the data. The time required to train and test a model can depend upon the choice of its hyperparameters. A hyperparameter is usually of continuous or integer type, leading to mixed-type optimization problems.

The existence of some hyperparameters is conditional upon the value of others as well; for example, the size of each hidden layer in a neural network can be conditional upon the number of layers. From the beginning, the main focus of autoML has been automated model selection (as discussed above) and this hyperparameter tuning and optimization. Let's look at this aspect in detail.

From the beginning, the main focus of autoML has been automated model selection and hyperparameter tuning and optimization.

Environment readiness for hyperparameter tuning. Hyperparameter tuning is by far the most resource-intensive of the entire autoML batch run. Just because of the sheer number of hyperparameters and permutations and combinations of distinct values of those parameters, the process is time-consuming and computationally heavy. Especially for some of the more complex algorithms, such computation may need to be allocated additional processors and memory resources. Hence, for selection of automated pipeline, you have to be cognizant of the given time, memory, and complexity constraints and plan accordingly.

Automated hyperparameter identification. When tuning how the chosen algorithm works, the process called "hyperparameter tuning" involves considerable trial and error. Complex ML algorithms can have more than a dozen

configurable parameters, each of which can have a large impact on model performance. The following examples of hyperparameters for two popular algorithms will help explain the idea:

- XGBoost (Extensible Gradient Boost) hyperparameters include ntrees, max_depth, min_rows, min_sum_hessian_in_leaf, sample_rate, col_sample_rate, col_sample_rate_per_tree, booster, reg_lambda, reg_alpha
- Deep learning hyperparameters include epochs, adaptive_rate, activation, rho, epsilon, input_dropout_ratio, hidden, hidden_dropout_ratios

Automated hyperparameters optimization. This requires a deep understanding of algorithms and core ML concepts. Every decision in the data-driven pipeline is a hyperparameter. The very idea of autoML is to find such hyperparameters, which could give a good score in a reasonable amount of time. AutoML tunes hyperparameters of the model chosen (for example, number of trees and subsampling for tree-based models or architecture, learning rate and number of epochs for neural networks, choosing an architecture, loss function, and optimizer in the case of deep learning, etc.) by running multiple iterations during which you set an initial value for one or more of the parameters and incrementally change the value for each run until the run with highest accuracy is found.

There can be a huge number of permutations and combinations possible. Beginners often feel they are just guessing as they test different hyperparameters for a model, and automating the process could make this piece of the machine learning pipeline easier, as well as accelerating the process for experienced machine learning practitioners. That's why autoML is the preferred approach. However, although increasing the performance or the accuracy of the model is the main objective, it also has to be completed in a reasonable time. To govern that, autoML itself has some parameters of its own which need to be set judiciously before it is unleashed to run the experiments in auto-pilot mode. It may be regarded as a small downside that one has to deal with these additional parameters of autoML tools, which may need some expertise to be set themselves.

#5: Automated Selection of Evaluation Metrics and Validation Procedures

The ultimate objective of every data scientist is to create a model with the most optimized performance and the highest accuracy. That is not easily achieved and that is why you need multiple model runs with different combinations of hyperparameters using different features and checking the accuracy or a common performance metric at the end of each run. The following are important parts autoML.

Automated selection of evaluation metrics and validation procedures. AutoML output includes a "leaderboard" of models that were trained in the process, including the five-fold cross-validated model performance (by default). If the user wants to score the models on a specific data set, they can specify the leaderboard_frame argument, and then the leaderboard will show scores on that data set instead. The models are ranked by a default metric based on the problem type.

Some additional metrics are also provided for convenience. In fact, at the start of an autoML run, you can decide which specific metric to use for the comparative evaluation or let autoML decide it automatically. At the end of the autoML run, the final leaderboard will display all the models with their respective measurement metrics in descending order of value—the most accurate or best-performing model will appear at the top. Once that model is identified, the next step is to productionize it in the form of APIs or other means.

Automated explanation of the model output. This is yet another hot topic especially in the light of compliance and regulatory requirements and ethical use of AI/ML in general. Whether it is a regression, classification, an association/clustering, or a recommendation model, you must be able to explain what the model has predicted to justify that prediction. This is another important aspect of autoML, which can generate various charts and diagrams to explain how it arrived at the predicted outcome. On top of that, it can integrate natural language processing (NLP) and natural language generation (NLG) techniques to create commentaries about the predictive and prescriptive result.

BENEFITS OF AUTOML

Now that we have seen all the 5 "A"s of autoML in detail, you can easily appreciate the productivity boost autoML can bring to the entire machine learning workflow. AutoML has the potential to play an enabler role for the so-called citizen data scientist to engage in complex ML/AI-powered experiments to bring about disruptive business outcomes. However, autoML can also help seasoned data scientists create innovative solutions to more complex business problems.

One of the benefits of autoML is that it can quickly make a well-educated selection of a suitable ML algorithm and effective initial hyperparameters. AutoML can then test the accuracy of training the chosen algorithms with those parameters, make tiny adjustments, and test the results again. AutoML can also automate the creation of small, accurate subsets of data to use for those iterative refinements, yielding excellent results in a fraction of the time. Instead of having to test a set of parameters against multiple billions of rows of training data, autoML can test against .01 percent of that data without compromising model accuracy, which will result in 100 to 1,000 times faster training time, even on the same data set.

AutoML tools increase data scientist productivity and extend machine learning to non-expert ML users. Furthermore, by saving computing time, using autoML to automate algorithm selection results in cost savings, especially in cloud environments, where the cost of compute time can be directly quantified.

To sum up, the three key advantages of autoML are:

- **Increased productivity** by automating repetitive tasks, enabling a data scientist to focus more on the problem rather than the models
- Automating the ML pipeline helps to **avoid errors** that might creep in manually
- Ultimately, autoML is a step towards **democratizing machine learning** by making the power of ML accessible to everyone ●

BI Experts' Perspective

THE EMERGENCE OF MODELOPS

John O'Brien is the principal advisor and CEO at Radiant Advisors.
john.obrien@radiantadvisors.com

James Taylor is CEO at Decision Management Solutions.
james@decisionmanagementsolutions.com

Coy Yonce is chief technology officer at Vedapointe.
Coy@vedapointe.com

By John O'Brien, James Taylor, and Coy Yonce

Q» A growing number of organizations are using predictive analytics to make product recommendations, identify market segments, calculate customer lifetime value, and more. In the beginning, predictive modeling was a "cottage industry." Only a few people developed, implemented, and maintained a limited number of models. However, the need for models has grown (often to the thousands) in many organizations.

The cottage industry approach is no longer viable, as more people are involved. Data pipelines must be established; model building, testing, and updating must be automated and documented; and the scope of data and analytics governance must be expanded. The ModelOps term is increasingly used to describe the automation of the model building, implementing, and updating processes.

Assume you are a consultant to a chief analytics officer who was recently hired to "take analytics to the next level" in the company. Doing this requires scaling analytics in terms of the people, data, models, technologies, and processes. Please provide your perspective and advice on the following questions and issues:

1. **In terms of people, to what extent can business analysts (often with upgraded skills) and citizen data scientists contribute to and be integrated into a ModelOps environment? A company's staffing isn't likely to be all data scientists.**

2. **The need for data engineers to build data pipelines for operational systems and analytics has grown, resulting in shortages of personnel with these skills. How should a company ensure that it has the data engineering skills it needs?**

3. **As companies use additional data sources, often obtained from data brokers, how can they maintain data quality and ensure that the use of the data does not violate regulations and laws such as the GDPR and CCPA?**

4. **To what extent do current analytics workbenches support ModelOps? What advances do you anticipate in these workbenches? How do programming languages such as R and Python fit into a ModelOps environment?**

5. **What processes need to be put in place for ModelOps?**

John O'Brien

1. Staffing is a common challenge that many fledgling data science programs face. The question asked is: "How can we scale to deliver more analytics models to the business even faster?" The first place to start is to ensure that the highly specialized data scientists are spending their time where they add the most value. This might mean selecting and building analytics models initially, but eventually they should be spending more time mentoring and reviewing the analytics models of citizen data scientists.

The key is to break down stages of the modern analytics life cycle process to enable project managers, business analysts, and data engineers to each do as much of the work as possible. The second place to focus on is efficiency in the model testing, deployment, monitoring, and updating process. Delivering an analytics model into production is more time-consuming than typically anticipated. Continuous deployment and containers can help with the process.

2. To ensure that your organization has the data engineering skills it needs, begin by understanding how data engineers are spending their time on projects. Specifically, how much time are they spending finding data, validating data, and testing the feasibility of the project?

Quite often, a noticeable portion of their time is spent early in the discovery phase of the project. These efforts can be easily performed by their business analyst counterparts—often more quickly because of their business and process knowledge to make iterative, quick decisions.

We take this one step further by understanding how much of the data engineer's time is spent performing architecture- or operations-related work. We institute an architecture pattern book process whereby data engineers spend their time applying the architecture pattern to the business project rather than reinventing the wheel. The goal is to find ways to increase speed and agility for data engineers.

3. When it comes to proper use of data, this is where the data governance office (DGO) gets involved. Self-service data analytics or agile delivery teams should be able to incorporate external or third-party data into the enterprise data lake as needed. In the case where the external data is

going to be used by multiple business groups, we recommend that the DGO include this in their reference data management responsibilities.

At a minimum, we strive to incorporate data governance and data quality at the point of data ingestion into the data lake. Data engineers, business analysts, and data scientists know to do this anytime their source data is not already in the data lake. For the discovery phase, the DGO is notified. For the usage and implementation phase, the DGO facilitates data owners and data stewards as necessary.

4. We are pleased to see that most data science platforms have incorporated ModelOps as part of their life cycle. Only a few years ago, enterprises recognized the need for operationalizing analytics, and it was handled manually.

Most data science platforms continue to recognize that data scientists need to be allowed to work in their programming language of choice, such as Python and R, and data science notebooks are becoming the de facto standard for workbenches.

The importance has shifted from building analytics models to the ongoing work of operationalizing and maintaining those models in production. The work now is to monitor if a deployed analytics model maintains its effectiveness over time, how it competes with other, similar models in production, and whether updating data features or reinforcement training is needed.

5. ModelOps essentially provides the ability to maintain analytics models at scale. ModelOps monitoring capability enables the data science team to quickly see if any analytics models need attention or are trending in the wrong direction. When the data science team needs to adjust and update the model, the team must have a process in place for A/B testing to quickly react and analyze the impact of changes.

In addition, efficient model deployment processes through ModelOps can leverage continuous integration and continuous delivery (CI/CD) so that a change can be made and deployed quickly. Both of these capabilities enable data science teams to react quickly to notifications delivered from an analytics platform that utilizes ModelOps.

James Taylor

The value of predictive analytics lies not in its ability to predict things such as market segments or lifetime value but in its ability to improve decision making. Predicting something is interesting and using that prediction to improve decision making is valuable. Predicting which products a customer might need, their customer market segment, and their lifetime value is only useful if we use those predictions to change the way we make decisions.

Ignoring this obvious truism is at the heart of most failed analytics efforts caused by teams believing the purpose of a predictive analytics project is insight rather than improved decision making.

A chief analytics officer (CAO) taking analytics to the next level has two key problems: making sure the predictive analytics can improve a decision and then making sure that they do. Failing to consider

ModelOps carries the risk of leaving models on the shelf, adding no value. Relying on ModelOps to operationalize models will not succeed if they are the wrong models.

Successful predictive analytics projects work backwards. They don't start with the data, build a predictive analytics model, and then try to find a use case for it. Instead, they start by identifying the decision they want to improve and making sure they can tell if they have improved it.

With the decision clearly in mind, they develop predictive analytics models that will improve decision making. Once they have predictive analytics models that could help, they use ModelOps to integrate those models into production.

The best decisions for predictive analytics are those the company makes over and over. These transactional, operational decisions generate the data needed to build predictive analytics models and reward even small improvements in decision accuracy with high ROI because any improvement multiplies by the number of decisions made. These decisions must be made at scale, often in real time, and so must be integrated into systems that handle day-to-day operations rather than dashboards. ModelOps should be focused on these scenarios.

Taking analytics to the next level will require that the CAO stops thinking about data and starts thinking about decisions–and not just any decisions.

The good news is that more analytics workbenches are supporting ModelOps directly or by integrating with third-party products. In most larger companies, this work must be integrated with existing IT systems and processes.

Data pipelines and the plumbing of ModelOps are standard IT functions. They're a challenge for many analytics teams because those teams don't want to talk to IT! Collaboration with IT will reduce the risk. The CAO should recognize that although predictive analytics models have some unique needs, much of ModelOps is just DevOps for analytics, and reach out to the CIO to coordinate this.

The most important process for the CAO is the overall predictive analytics process because fixing just ModelOps may result in implementing the wrong models quickly. The CAO needs an overall process that works backward from the problem:

1. Begin by understanding the business decision to be improved
2. Identify the predictive analytics models that would be helpful
3. Prepare, integrate, and analyze the available data to build these models
4. Evaluate these models against the decision-making to ensure they are both mathematically accurate and that they will improve the decision making
5. Deploy these models using a modern ModelOps

approach, working closely with IT

6. Continuously monitor and improve deployed models and the decisions being made with them.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a great starting point for such a process. Although a little long in the tooth, it's a well-defined, standard approach to analytics projects that outlines the right steps for success.

A couple of final thoughts:
It is not about making business analysts into citizen data scientists. It is about ensuring that all the company's business analysts see the decisions in their systems and processes and understand what kinds of questions to ask to frame predictive analytics requirements. Only business analysis can find real opportunities for predictive analytics and this is the training they need most urgently.

Data quality is not absolute. The required accuracy of a predictive analytics model also varies based on what's needed to improve the decision. The CAO needs to make sure that the question is always "is the data quality good enough to build the models we actually need to improve this decision?" That assures everyone stays focused on improving decision making not fixing a long list of data problems.

Privacy and regulation are important. Ensuring that decision making is modeled and understood, making sure that policies and regulations are handled using transparent business rules (not ETL code), and using modern techniques to explain how data is used inside the predictive analytics models will make it clear how data is being used to make decisions.

Coy Yonce

As organizations make decisions about how best to apply their data to the problems ahead of them, they need to determine how to scale their predictive and machine learning models.

In some cases, this could mean leveraging resources already available to the organization. In others, it means acquiring new resources to help create a more effective environment for deploying predictive models.

In terms of people, business analysts are impacted by the scaling of predictive models. Traditionally, business analysts focus on turning data analysis into organizational improvement from the perspective of processes, products, and services.

Scaling requires a focus on having the right people, data, technology, and processes to support continually integrating, deploying, and improving their models.

This primary focus should be unchanged. However, with predictive modeling and analytics, business analysts can focus less on massaging data and making it more usable. They can now focus on applying predictive models and the output data to their problems. In theory, this should allow them to spread their skills

across more problems within their organization.

Reliance on the predictive models developed by data scientists means business analysts are key stakeholders in the effectiveness of how models can be applied within the organization. This makes them candidates for helping to improve the models through iterative testing, feedback, and commentary about how models can be enhanced or applied in different ways.

The same is true for citizen data scientists across the organization. More eyes on a model, the data it produces, and the problems it seeks to solve equates to better feedback going into the continual improvement process for the model, the data, and the deployment architecture.

Seeking to recruit more employees to work as citizen data scientists as part of their daily tasks could be a successful strategy for producing effectively designed, well-trained models.

Aside from data scientists and business analysts, data engineers play a significant role in the deployment of predictive models. Similar to SysOps and DevOps engineers, data engineers create the infrastructure, pipelines, and processes required to scale predictive models.

Currently, data engineers are in high demand within the industry because there are so many organizations looking to make the most of their data and analytics platforms. This creates situations where organizations have a hard time finding the talent and skills they need.

To help address this problem, organizations should focus on defining their strategic road map for building data science skills throughout the organization, performing a skill assessment of their existing resources, and then mapping this skill assessment to the skills they require based on their strategic road map.

For any skill gaps, the focus should then be on hiring people who actually satisfy the need the organization has rather than hiring candidates with the best degree. Yes, highly skilled candidates are great, but they may not really be what you need if you aren't performing cutting-edge modeling.

It's best to focus on your strategic road map to determine what you need. This will help you to define the optimal hiring and resource plan for your organization. If you find that you're not sure what your road map should be in relation to data science and data engineering, then perhaps the best plan is to hire a data science lead who can help build that strategy.

With the understanding of a strategic road map, an effective hiring process, and the right team of data engineers in place, your organization can now focus on deploying processes to support ModelOps.

The goal of ModelOps is to support your use of predictive models at scale. This is enabled by borrowing from continuous integration, continuous deployment, and continuous improvement processes leveraged by SysOps and DevOps teams.

There are various subprocesses required for implementing an effective ModelOps CI/CD pipeline:

- Version control for the models, training data, infrastructure, and schemas
- Automated testing of the models based on training data, historic data, and new data
- Automated scaling of compute resources powered by infrastructure-as-code whether deployed on premises or in the cloud
- Monitoring the accuracy, effectiveness, use, and compliance of the model and output data produced; specifically focusing on the need to retrain the model with new data to understand concept drift
- Tuning the model based on the results of automated monitoring and notifications for how the model performs against new data
- Ensuring that security controls are implemented, monitored, and improved through scans focused on adherence to organization standards, as well as regulations and governance. ●

Instructions for Authors

The *Business Intelligence Journal* is a semiannual journal that focuses on all aspects of data warehousing and business intelligence. It serves the needs of researchers and practitioners in this important field by publishing surveys of current practices, opinion pieces, conceptual frameworks, case studies that describe innovative practices or provide important insights, tutorials, technology discussions, and annotated bibliographies. The *Journal* publishes educational articles that do not market, advertise, or promote one particular product or company.

SUBMISSIONS

For more information and complete submissions guidelines, please visit tdwi.org/journalsubmissions.

Materials should be submitted to:
Peter Considine, Managing Editor
Email: journal@tdwi.org

UPCOMING SUBMISSIONS DEADLINES

Volume 26, Number 1
Submission Deadline: February 19, 2021
Distribution: June 2021

Volume 26, Number 2
Submission Deadline: August 20, 2021
Distribution: December 2021

Managing Rapid Data Delivery While Maintaining a Durable Data Warehouse

Eric J. Peters is a data architect and data warehouse manager at Vitalant.
epeters1526@gmail.com

Eric J. Peters

ABSTRACT:
Modern BI toolsets boast the capability to place data blending into the hands of eager end users. Although these mash-up features allow quick time to market for the savvy desktop developer, the influx of various new streams of data into an organization poses a problem for the teams that struggle to manage such requests. As self-service analytics continues to thrive as a means of making data-driven decisions, maintaining a stable single source of the truth becomes an unruly task.

Waiting for a DW/BI development team to complete the tedious work of architecting new source systems into the data warehouse causes valuable time to slip away, impacting an organization's ability to produce the reports that help make those informed decisions.

Is there a way to satisfy both sides of the data delivery equation? This article provides the framework to keep end users nimble while maintaining the integrity and durability of the enterprise data warehouse.

INTRODUCTION

The rise of more powerful analytics tools has allowed business users to combine, join, and correlate data on their own. Data wrangling is the term that's often applied to the act of pulling data together from a variety of sources to churn out analytics at the same

freaky speed that a sandwich shop can turn around a lunch order.

This great power has been placed within the desktops of personnel who seldom grasp the responsibility that comes with it. Where does the ownership lie to ensure that "wrangled" data remains accurate, current, and reliable? Cartesian join pitfalls and attention to row granularity may be conveyed to end users through visual aids and specific business scenarios, but few will retain the importance of these technical rules. This is a skill set that lies largely outside a business user's experience. Wrangled collections of data too often remain as haphazard as the Wild West moniker implies.

Is the solution to lock down the feature-rich environment of these analytics capabilities and take this ability away from users? Is IT ready to reclaim dashboard design and development and place it back into the hands of an already over-taxed data development team? Certainly not. The advantages offered by these tools and the investment made to procure the software and all its features are many.

Does the data warehouse development team throw up their collective hands and say farewell to the traditional data warehouse? Allow thousands of hours of customized extract, transform, and load (ETL) development as well as a robust and well-governed repository of enterprise business wisdom to give way to the ubiquity of user-extracted data, spreadsheets, and one-offs? Certainly not. IT and business teams must acknowledge and embrace the advantages of each approach to delivering on data and realize that the two must co-exist so the enterprise can maintain a strategic business advantage through nimble analytics while maintaining trust and reliability in an enterprise data warehouse.

IDENTIFYING AN OPPORTUNITY

We will start with a business scenario that will drive our discussion. Imagine that the finance director of your organization is tasked by her supervisor to deliver a key performance indictor dashboard before the close of the next fiscal period. The organization already has its BI tool in place and a traditional data warehouse has been implemented with many of the business's key subject areas available. However, requirements call for additional data with sources that do not yet exist in the warehouse.

The prudent finance director studies a mock-up; she identifies gaps and promptly reaches out to you, the DW/BI development manager, to convey the additional data that will be needed in the warehouse. By the end of your planning session, it is clear that your delivery timelines do not align. The director's boss needs the data served quickly (like that sandwich shop) and your timeline spans days, perhaps weeks.

This is a common scenario. It is likely that similar requests will increase dramatically in most organizations, if they have not already. Embrace this environment.

THE FRAMEWORK IN ACTION

At this point, the ETL developers, data architects, and you as a manger will say, "Stop! This data belongs in the warehouse." You would all be correct. The data must be properly sourced, modeled, extracted, transformed, and loaded and, of course, validated.

The business units will say, "We already possess all the data we need, we have the reporting tool that IT helped provision, and we need this report now." They would also be correct. There is nothing but your word to impede the business team from beginning to harness their data in its various formats for their desired outcome.

The finance director's team has received adequate training on the chosen enterprise reporting tool's advanced data blending features. Data blending was arguably the most important software feature when your enterprise was evaluating solutions. The team is eager, they have bought into the concept of independent, end-to-end development, so work begins immediately.

Observe and Approve

Let them build. Leverage your organization's BI tool with all of the self-service capabilities and let the business perform at the speed and agility they can handle. Granting business users the autonomy to build their own analytics allows them to move fast and take the risks they're comfortable with. Allow them to roam the frontier and lasso all the data they need to tell their analytics story.

This philosophy is not new or untested. DARPA, created in 1958, developed a use-inspired research model that allows a dedicated business team to experiment with their own ideas to meet objectives and timelines that have been dictated to them [Dugan, Gabriel, 2013].

Initially, the DW/BI development team will take on the role of observer. Dedicate a resource from DW/BI with the expertise to carry on the role of one who will review the work as business development milestones are achieved. Their observations are converted into requirements for ETL developers and may require changes to the data warehouse. The team's data modeler is involved early in the project and provides guidance where needed.

While the business is in its own design and release phase, the development team can move forward with its own structured design life cycle. This begins with a development plan for long-term integration into the data warehouse. Changes to the data warehouse to support user requirements will begin. Much of the work may be done in parallel, but the business will almost always outpace ETL development. Adherence to shop standards, best practices, documentation, and peer reviews all ensure a data warehouse remains durable.

The process is complete when the development team representative has signed off on the alignment and aggregation of data before reports are released. A representative from data governance should also be involved in the project phase's review.

It's Never as Easy as 1, 2, 3

Let's make our scenario interesting and a bit more realistic. A call comes in from the VP of marketing. A new opportunity has been identified to reward CSRs who convert prospects to customers based on a new promotional campaign that has already launched. Customer or prospect core data exists in the warehouse but the VP has obtained a third-party manual feed showing clickthrough data for respondents to the campaign. His idea to analyze performance will be launched from his division and, if successful, will be expanded across all divisions and repeated often.

Now the scenario is complete. Multiple requests have come to your development team from different functions of the organization and both need their analytics at the same time—that is, now.

Elements in Play

Looking deeper into the two new requests for data, let us assume you will need to incorporate the following new sources:

- A new, delimited flat file with safety recommendations that the enterprise will compare with its actuals. This file changes infrequently.
- A new spreadsheet that acts as a conversion table containing product names from a division that has recently been acquired as part of business expansion. The core ERP system of this new business is not scheduled to be integrated into the enterprise transactional system until sometime next year.
- A large spreadsheet of columnar data that contains metrics on operational throughput. Upon review it is discovered that this may be sourced directly from the relational database of the company's on-premises shipping and receiving application (with the necessary approvals).
- For the marketing request, an extracted file of clickthrough data pulled from a hosted CRM service that, until now, has not shown sufficient strategic value to include in the warehouse.

There may be any number of additional sources such as social media data, unstructured data, and machine data, any of which can follow the same approach.

Build a Repeatable Process

As quickly as they become available, have samples of new data sources sent to the development team to identify where new data will reside in the warehouse. Profile data and explore source systems, removing any barriers to access. Vendor contracts, firewalls, and other security considerations for data housed on hosted environments are common causes of unexpected delays.

Review the report the business group has put together. Examine the relationships defined in the underlying schema and determine where complex joins are taking place. These are designed into aggregate tables. Multiple fact-table joins are converted into report tables and both are included into the refresh process cycle. Other custom queries are optimized and created as views.

New metrics are identified and descriptive attributes are added into respective fact and dimension tables. New tables are created as needed.

The predominately static conversion file, containing product names and number translations, are added into the warehouse decode table.

Some manual attributes, such as goal-setting values, will change over time. For the safety recommendations file needed for the KPI dashboard, the business user is provided with a direct access interface using IT-approved methods for data stewards to update these records on their own.

The application business owner of the shipping and receiving software is contacted. You receive that group's buy in and confirm that the application database can be accessed without violating the terms of the contract. A time has been negotiated when this data can be extracted from the software's database without interrupting transactional operations.

The marketing report will run as a prototype and awaits results; it will eventually be spliced together with HR turnover data and the results of a recent employee satisfaction survey. It is determined this can live in a pre-release state until the decision is made that it will continue, then full-time ETL development is begun. This file remains outside the data warehouse and no unnecessary development is needed at this time.

Now a road map is created to get all new data sets into the warehouse. Any one of these steps can (and often will) take longer than others to complete, but the framework has been established. Meanwhile, the business is off and running and has produced its initial version of the report.

TEST DRIVE THE DATA

This framework advocates prototyping and releasing results to small groups to allow report consumers to "try before you buy" as new data sources are introduced. Sending reports to mid-level leaders, those who know their data best, will ensure data resources have been managed properly before final release.

Take advantage of a BI application's verification or annotation features to mark reports in their various states of validation. Establish a release taxonomy of at least three stages of versioning—for example, Pre-Release, Verified, and Final.

- **PRE-RELEASE:** This is the unverified release, still undergoing approval from a data modeler. Caveat emptor. The design and layout of a dashboard might be complete with artistic and meaningful visuals to tell the story, but the numbers' accuracy is not yet confirmed until architecture review is complete.
- **VERIFIED:** This is the business release of the report with both sides agreeing that the data output is valid and producing accurate results. BI/DW development has signed off on the blending of data and the mash-up of sources. Table joins still exist within the design of the report, but EDW integration is not yet complete.
- **FINAL:** At this stage, the report is running with data sourced from the EDW. Between the Verified and Final stages there should be no discrepancy in the resulting metrics and the impact will be seamless to users.

More detailed notes and tooltips within a report can elaborate areas still undergoing validation as part of the iterative process toward the final EDW state.

Some report consumers may only be interested in a specific slice of the report for their business function or geographic location and can begin to make use of the analytics sooner than others. In other cases, a business unit may have particularly difficult data to validate due to enterprise integrations, such as a recent merger or regional realignment. There is no need to hold back on a release while waiting for the final 10 percent of the data to be verified. Tactically, this approach opens up just-in-time

analytics to the right decision makers at the right time.

At any point in time you may have multiple reports in any one or several of the three stages. Don't let reports get stuck in perpetual Pre-Release. Enact a portion of ongoing governance meetings to review these reports and move them along to their final phase or eliminate them.

NEW NORMAL VERSUS THE STATUS QUO

Issues concerning multiple sources of data that pre-dated the enterprise data warehouse are once again making themselves known through the self-service features of modern BI tools. "[T]he proliferation of user-developed spreadsheets and databases inevitably leads to multiple versions of key indicators within an organization" [Davenport, 2006]. Enhanced collaboration applications for sharing ideas increases the likelihood that these conditions will persist.

Without a framework such as this in place, an organization may be left with a myriad of business-developed data sets that never make their way into a structured environment. These homegrown data sources, in their various formats, become the accepted source of "truth" for the decision makers, with no two versions displaying the same results. Data refresh times are ambiguous and reliant on one or two end users with tacit knowledge that may leave the organization at any time. Over time the data warehouse becomes an underutilized asset.

On the flip side, to restrict self-service causes delays while waiting on delivery of end-to-end integration and can cost the business a strategic advantage for making data-driven decisions at a critical moment. Business users will lose patience and confidence in IT and ultimately stop sending requests to IT, inevitably reverting to a shadow self-service model. The power of the enterprise BI tool is attenuated and only basic functionality is utilized, wasting a costly enterprise investment in licensing, training, and compute power in favor of numerous downloadable trials in as many departments across the organization.

By following this approach, the overarching purpose of the enterprise data warehouse remains intact, to present the data in "standard formats, integrate it, store it… and make it easily available" [Davenport, 2006]. Users will feel empowered to take ownership of their own data. They will understand the data and its relationships, and over time will need less hand holding from developers, freeing up valuable developer time.

As users grasp the full capabilities of the BI tool, user adoption increases, and the likelihood of users branching out and using tools outside of those supported by the organization is minimized.

Developers will understand what the business units want out of the data and can plan accordingly. They begin to think ahead and start to model data that is already presented in ways that will benefit users.

A collaborative team effort is the final critical piece that ties it all together, allowing the framework to persist and succeed. The business acumen and technology skills—along with the artistic ability to generate meaningful dashboards—rarely come in the same package. "No one individual can possess all the skills

needed to transform data into knowledge." [Davenport, Harris, De Long, Jacobson, 2001]

Finally, all data resides inside the data warehouse—a database that is backed-up, optimally tuned, monitored by technical support, and trusted.

CONCLUSION

The 2001 article *Data to Knowledge to Results: Building an Analytic Capability* foreshadowed the need "to get people to do their own analysis and queries." [Davenport, Harris, De Long, Jacobson, 2001]. That time is upon us, as it has been for some time, but it now comes with expectations to deliver at the velocity that ideas spring to life. To ignore this is to diminish the data warehouse and all the advantages it still offers.

We discourage business teams from working apart from DW/BI developers. Left unchecked, self-service analytics creates an abundance of shadow reports reliant on runaway files, skewing numbers and charts, and creating distrust of the same business intelligence tool that the organization has invested heavily in.

By allowing an observed self-service approach, the development team benefits by integrating at a sustainable pace and business users can create swift analytics and adhere to their own deadlines.

On-demand integration road maps identify needed source systems, avoiding the countless hours spent on importing and validating data elements that are seldom (if ever) used in reporting. The hybrid approach operates on known requirements, is focused on the specific details needed for analysis, and delivers what is most strategically useful at that time. Business users are less likely to delay a quality check because they know the usefulness of the data at that moment rather than spending their valuable time trying to validate data they may never request in a report.

Keep your focus on cross-team collaboration and remain informed. The examples provided in this article only succeed because both the finance director and the marketing VP reached out to the DW/BI manager. This communication is vital to the framework's continued success. ●

REFERENCES

Davenport, Thomas H. [2006]. "Competing on Analytics," *Harvard Business Review*, January

Davenport, Thomas H., Harris, Jeanna G., De Long, David W., and Jacobson, Alvin L. [2001]. "Data to Knowledge to Results: Building an Analytic Capability," *California Management Review*, Vol. 43, No. 2

Dugan, Regina E., and Gabriel, Kaigham J. [2013]. "Special Forces Innovation: How DARPA Attacks Problems," *Harvard Business Review*, October

BI StatShots

Use Multiple Development Methodologies for the Most Value

TDWI's annual *Teams, Skills, and Budgets Report* enables data and analytics teams to compare themselves to their peers on a series of organizational and performance metrics. This year's report was recently released, and it included some interesting insights about current use of agile and other development methodologies.

Agile remains the dominant development methodology used by data teams. This year almost half of all responding teams (48 percent) used agile for creating their new solutions. As in the previous year's survey, waterfall holds second place in 29 percent of data teams. This year's survey also found scrum and sprint are popular.

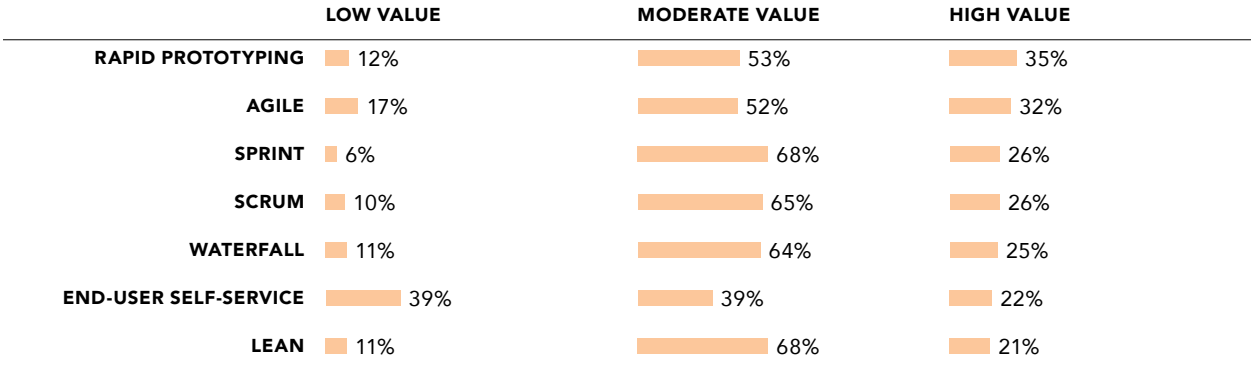
We then cross-referenced this data with respondents' assessments of the value of their data management and analytics programs. Agile showed high value in 32 percent of companies; when adding "moderate value" to the results, agile showed good value in 84 percent of respondent organizations and provided low value in just 17 percent. In comparison, waterfall, used in 29 percent of organizations, achieved high value for only a quarter (25 percent) but high or moderate value in 89 percent of respondents—slightly more than agile.

Fewer respondents reported success with self-service analytics this year; 22 percent of users reported high value and 39 percent reported low value. Sprint, scrum, lean, and rapid prototyping all delivered moderate or high value for a significant majority of users.

Which of the following development methodologies do you use when creating new data management/ analytics solutions?

AGILE: Highly iterative model in which requirements and feedback are gathered throughout development	**48%**
WATERFALL: Sequential model in which most requirements are gathered before coding	**29%**
SCRUM: The most popular form of agile development, which uses a lightweight development process with a focus on productivity	**25%**
SPRINT: An individual development cycle in Scrum	**25%**
LEAN: Alternative to agile development, in which the emphasis is on optimizing use of existing resources while eliminating waste	**15%**
END-USER SELF-SERVICE: A method of allowing business users to set up their own queries and reporting	**15%**
RAPID PROTOTYPING: Rapid development of a BI solution that users can try out and to which they can suggest modifications	**14%**

Data management/analytics value by development methodology



However, organizations will typically use a combination of development methods as the situation demands, and the highest values were found for enterprises that reported using multiple methodologies. Fifteen percent of respondents reported using agile, sprint, *and* scrum, and 95 percent of them reported high or moderate value. Nineteen percent of respondents used *both* waterfall and agile, and they *all* reported either high or moderate success from their data management and analytics projects. These findings underscore the importance of matching your development methodology to your business objectives. ●

TDWI Membership

FOR DATA PROFESSIONALS

Advance Your Projects and Your Career

TDWI Membership is the one-stop learning shop for data leaders who want access to the tools necessary to move their careers, teams, and projects forward.

Gain exclusive access to the latest industry research and resources for analytics and data management, a community of professionals looking to connect and collaborate, and exclusive discounts on valuable skills-development opportunities.

Member-only benefits include:

- All TDWI research (Best Practices Reports, Checklist Reports, Insight Accelerators, etc.)
- All TDWI webinars on-demand
- 10% discounts for event registration and online learning courses
- $150 annual training voucher per person
- Exclusive Member-only content such as Ten Mistakes to Avoid; the annual *Salary, Roles, and Responsibilities Report*; and the *Business Intelligence Journal*

260+
ON-DEMAND WEBINARS, CONFERENCE KEYNOTES & POPULAR SESSIONS

380+
TDWI RESEARCH AND MEMBERS-ONLY PUBLICATIONS

100%
SALES FREE ZONE

Immediately save 10% on TDWI Education

UPCOMING VIRTUAL SEMINARS:

Data Literacy in the Workplace	June 22–24
Data Management Essentials	July 20–22
Data Modeling Essentials	Aug. 3–5
Building Your Enterprise Data Strategy	Aug. 10–11
Agile BI & Analytics	Sept. 14–16
Predictive Analytics for Practitioners	Sept. 21–23
Dimensional Modeling Deep Dive	Sept. 28–30
Visualization, Dashboards & Analytics Adoption	Oct. 5–7
Machine Learning Bootcamp with Python	Oct. 19–21
Modern Data Architectures	Nov. 16–18
Data Management Essentials	Dec. 7–9

SELF-PACED, ON-DEMAND ONLINE LEARNING

Analytics Fundamentals
Business Intelligence
Data Governance
Data Modeling
Data Quality
Data Science, R, and Python
Predictive Analytics
Supervised Machine Learning

Become a member, or learn more at **tdwi.org/membership**

tdwi

VIRTUAL TEAM TRAINING

YOUR TEAM OUR INSTRUCTORS ANYWHERE, ANYTIME

Bring Your Team Together in a Virtual Classroom

The demand for data-driven decision making has not slowed in the wake of the current situation that is impacting us globally. Now, your team's professional development doesn't have to slow down either.

TDWI's virtual training program allows your team to learn, collaborate, and interact with an expert instructor in real time via the safety of a virtual classroom. Flexible scheduling allows you to spread one or more full-day courses across half-day sessions, so you don't have to put time-sensitive work on hold as you build essential skills with an agenda tailored for your enterprise's needs.

Choose from among these key learning categories:

- Custom Workshops
- Business Intelligence
- Data Modeling
- Business Analytics
- Predictive Analytics & Data Science
- Infrastructures & Technologies
- Leadership & Management
- Data Management
- Hands-On Courses
- CBIP Certification

For more information, contact:

Yvonne M. Baho
Director, Enterprise Learning Solutions
Phone: 978.582.7105 | Email: ybaho@tdwi.org

TDWI Partners

These solution providers have joined TDWI as special Partners and share TDWI's strong commitment to quality and content in education and knowledge transfer for business intelligence, data warehousing, and analytics.

tdwi PARTNER

Partners

alteryx

ATTUNITY

cloudera

DATAWATCH

denodo

Google Cloud

HORTONWORKS

IBM

opentext™

ORACLE

panoply

Qlik

SAP

sas

snowflake

+ a b | e a u

talend

THOUGHTSPOT

WhereScape®

wynEnterprise