

THE DATA DIFFERENTIATOR

HOW IMPROVING DATA QUALITY IMPROVES BUSINESS



IN ASSOCIATION WITH:



CONTENTS

INTRODUCTION.....	2
1. WHAT IS DATA QUALITY, AND WHAT DOES IT LOOK LIKE?	4
• Q&A: DUN & BRADSTREET: SETTING THE STAGE	6
2. DATA AS A DIFFERENTIATOR.....	9
• Q&A: FORRESTER: BEYOND TRANSACTIONAL DATA.....	14
• Q&A: IRONSIDE: CAPITALIZING ON DATA.....	17
3. BRINGING OUTSIDE DATA IN.....	20
• REAL DATA FOR REAL ESTATE.....	21
4. CHOOSING A DATA PARTNER.....	25
CONCLUSION.....	27
ACKNOWLEDGMENTS.....	28

INTRODUCTION

As businesses attempt to prepare for transformational technologies like autonomous vehicles, artificial intelligence and the Internet of Things (IoT), many find themselves grappling with the complexities of data and analytics. The role data plays in enabling these future technologies is critical—but one that will be undermined if businesses do not make data quality a priority.

Executive leadership recognizes the need to focus on data. In fact, according to the Forbes Insights and KPMG “2016 Global CEO Outlook,”¹ data and analytics capability ranks the highest of the top five investment priorities for CEOs today. Yet despite this, faith in the data businesses rely upon is low: The report also found that 84% of CEOs are concerned about the quality of the data they’re basing their decisions on.


Interviews with executives and analysts suggest that confidence in data may be low for various reasons, including silos of information, difficulty in securing

executive buy-in, not to mention the sheer quantity of legacy data a company may possess. However, regardless of the reasons for hesitancy on the data front, the fast pace of technological change and the competitive threat mean that organizations must quickly scale up capacity and integrate new sources of data—including unstructured sources like social media and messaging—as well as learn how to use the data they generate internally to improve product and service offerings, all while improving the quality of data accumulated over the life of the business.





The challenges posed by improving data quality can be daunting and obscure the benefits and possibilities that good-quality data enables, but the costs of doing nothing are high. Gartner measures the average financial impact of poor data on businesses at \$9.7 million per year.² These costs, however, are not solely financial; businesses can see loss of reputation, missed opportunities and higher-risk decision making as a result of low confidence in data.

The key to properly improving data quality and management is to lead with a clear articulation of the business case.  Good data quality is vital to effective Master Data Management and to creating the single customer view that so many organizations are striving for. The insights and conclusions that can be drawn from data will be only as good as the data used to obtain them; the old “garbage in, garbage out” adage. However, the availability of new kinds of big data—extremely large data sets that may be analyzed computationally to reveal patterns, trends and associations—and analytical software have combined with a democratization of data to give ready access to untrained and inexperienced operators. This has made clear that the need for quality data is as important as ever today. And looking toward the future, quality data now will be fundamental groundwork for the adoption of transformative technologies that will enable businesses to innovate and differentiate themselves down the line. Meanwhile, advances like graph databases, cloud computing and the sheer processing power of big data platforms like Hadoop and Spark are improving what can be done with and derived from data.

Data today is often compared with oil, as in its raw form, its uses are limited. It is through refinement that oil becomes useful as kerosene, gasoline and other consumer goods, and similarly it is through the refinement process of cleansing, validation, de-duplication and ongoing auditing that data can become useful in the kinds of advanced analytics that are starting to shape our world.

Based on in-depth research and interviews with industry executives, analysts and real-world users, this paper discusses the benefits of good-quality data and the costs of poor-quality data. It also presents recommendations for evaluating technology partners along the way.



Advances like graph databases, cloud computing and the sheer processing power of big data platforms like Hadoop and Spark are improving what can be done with and derived from data.

¹ “2016 Global CEO Outlook,” KPMG International

² Moore, C.S. (2017, January 24). How to Create a Business Case for Data Quality Improvement. Retrieved April 07, 2017, from <http://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement/>

1
CHAPTER

WHAT IS DATA QUALITY, AND WHAT DOES IT LOOK LIKE?

“Data quality is very much dependent upon the goals of the project as a whole.”

—Paul Hulford
CEO, Attain Insight

What data quality is can be hard to pin down, because so much depends on the needs and objectives of the organization, who is using it, and for what purpose. In fact, it is easier to recognize when data quality is poor, says Paul Hulford, CEO of Attain Insight, a data and analytics solutions and consulting firm in Ottawa, Canada. “People using the data for decision making can pretty quickly tell bad data because it doesn’t match their expectations. Or, more precisely, they see something that doesn’t match something they know to be true,” he explains. “So there’s a very quick reaction to say, ‘Well, can I trust this? I’ve been able to see something that doesn’t match something that I know to be factually true—therefore, what else is wrong?’”

Regardless, there are some general common conditions we can point to that indicate good, or suitable, quality of data (see p. 5).



CONDITIONS THAT INDICATE GOOD, OR SUITABLE, QUALITY OF DATA

ACCURATE

That is, the data is correct—addresses that guarantee mail will be deliverable—or transaction data that properly reflects a customer’s purchase history, for example. And there’s a freshness component to accuracy, too, one that measures the timeliness of data points. If a customer has moved, how long will it be until the new address is reflected in the data set used to communicate with that customer?

Additionally, as Anthony Scriffignano, Dun & Bradstreet’s chief data scientist, notes, accuracy itself can be situational—for example, among 10 telephone numbers that might be associated with a business, the one for investor relations is inappropriate if you are trying to reach the main switchboard for a local office. The point being that measuring accuracy, to a certain extent, depends in part on context. “In many cases, we talk about accuracy as if there’s some grand truth against which it can be measured, and that’s not always appropriate,” says Scriffignano, adding: “The whole concept of accuracy is really nuanced, and it has to be taken in the context of the particular attribute that you’re talking about, and sometimes in the case of the ultimate usage of that data, in order to measure it.”

COMPLETE

The presence or absence of data, or, as Dan Adams, vice president of data product management at Pitney Bowes, puts it: “How well you’ve populated everything you want to capture.” If purchase histories miss half the purchases made, then customer value may be underestimated and the customer experience affected negatively. If a retailer can’t accurately identify its highest-value customers, then it may miss valuable opportunities to build a relationship with them. Therefore, completeness of a data set needs to be evaluated.

STANDARDIZED

This means finding meaningful ways to compare data sets that may differ. For example, inputs and formats of names and addresses can vary widely—an address in Japan is constructed very differently from an address in North America. So the ability to standardize input to the correct format, even in the presence of input errors, is important. Likewise, it is

important to identify and match duplicate records. In many cases, postal or other standards do not exist, or vary widely within a large region, so an important driver becomes the ability to consistently apply transformations to the data so that one can compare like for like.

AUTHORITATIVE

Finally, data sources must be authoritative, credible and fit for purpose. In other words, the source must have credibility to provide the data in an accurate and complete way, whether that source is internal, like a retailer’s own sales records, or external, like a census bureau for certain demographic information. Without credible inputs, output may be of limited use because it won’t be based on the best information.



If purchase histories miss half the purchases made, then customer value may be underestimated and the customer experience affected negatively.



Dun & Bradstreet
has the largest commercial
database of its type in
the world.”

—Anthony Scriffignano
Chief Data Scientist,
Dun & Bradstreet

DUN & BRADSTREET: SETTING THE STAGE

Anthony Scriffignano is the chief data scientist and a senior vice president at Dun & Bradstreet, the company responsible for the global repository of business data, as indexed by the unique business identifier, the D-U-N-S Number. With 35 years of experience in business, Scriffignano is often called upon to speak on issues surrounding big data and to consult on emerging trends. We spoke to him about the importance of getting data quality right now, to prevent bigger issues later and set the stage for speedier adoption of transformational technology like artificial intelligence. Dun & Bradstreet is a customer, supplier and partner to Pitney Bowes. The following conversation has been edited for clarity and length.

For those who aren't familiar with Dun & Bradstreet and the D-U-N-S Number, can you give an overview?

[Dun & Bradstreet] has the largest commercial database of its type in the world. It is a collection of information about businesses and, to some extent, people in the context of business. We collect data from nearly every country in the world into a massive connected context. The data comes in different languages, and different writing systems, and different ontologies, and must be collected with respect to widely varying laws about the use, storage and transmission of data. We curate it largely to provide an opinion of total risk or total opportunity for our customers, so they can do things like issue credit or do sales and marketing activities. The data is also highly dynamic, updated millions of times a day.

“If we don’t get [data] right, we may not even realize it until it’s too late.”

—Anthony Scriffignano
Chief Data Scientist,
Dun & Bradstreet

The D-U-N-S Number is the primary way you reference an entity [within that database]. The idea is that the D-U-N-S Number persists throughout the life of the business at a location, even after the business stops operations. We link D-U-N-S Numbers together to build the corporate family trees of related businesses. We can combine all of that together through our multilingual identity resolution capabilities and compare it to information that comes in from the outside world—like lawsuits, liens, judgments, newspaper articles or anything else that may connect that data—and then we associate that insight to our customers’ inquiries.

What’s driving the need, for the businesses you work with, for improved data quality?

We define quality internally as an agglomeration of things: accuracy, completeness, timeliness and cross-border consistency (see p. 5). So how does the need for increased quality manifest itself in our customer use cases? There are a number of things driving that. One, I would say, is the speed of business itself. There is so much technology now that’s doing things faster, we have the ability to make a mistake and have that mistake propagate itself across a business much, much faster than it [was] possible before.

The need to get information right the first time and not spend a lot of time correcting and fixing errors is partly driven by the fact that we can have a much greater impact if something is wrong. For example, we may not have time to fix it: We don’t have time to go and chase it and correct it. Data permeates all aspects of modern workflow. Specifically, with our customers, a lot of them are making

more automated decisions, more global decisions, and decisions with greater impact to their enterprise.

Another driver is big data. It’s important that we have the ability to connect data and make sense out of it and separate it from the noise of the other data that’s out there. There are many nuances to this ability, such as the fact that all true data is not simultaneously true. All data collected a minute ago is not one minute old. Data is highly nuanced.

How dependent is a company’s successful adoption of future technologies—like artificial intelligence or the Internet of Things—upon getting data management right now, versus at some point down the line?

There are always challenges with emerging technologies, which can have disruptive influences on business and industries. When we start looking at artificial intelligence and related evolutions in areas like the Internet of Things and FinTech, for example—if we don’t get things right from the outset, we may not even realize it until it’s too late.

The Internet of Things is a great example. We had, last year, close to half the internet in North America taken down by a denial of service attack, using devices in unintended ways. This is a great warning to get a little bit better about the security of our “things.” Unfortunately, often there is a rush to build new things and bring them to market as fast as possible. The risk will get greater. Tomorrow, those things are going to be able to more easily discover each other and have conversations and share data with each other that maybe doesn’t go back to a central point. So the importance of being able to understand what’s happening with this data and the sorts of casual inference that might happen from this data [means that] getting the questions right and really formulating the behaviors of these things is critical as we start to use [them] to operate emerging technologies, like drones or autonomous self-driving vehicles.

I want us to be better than we are today—a lot better. Orders of magnitude better. And a big part of that is getting the data right.

Q&A

How can companies successfully differentiate between data sources and their appropriateness for a given purpose?

I use a mental model. Imagine a series of blocks named discovery, curation, synthesis, fabrication and delivery. Those are the five steps in the mental model that I use, and then alongside those, quality assurance and governance. So every one of those steps has a quality assurance step. Every one of those has a governance step.

What I try to think about in terms of how companies can take advantage of data is, think about each of those steps individually. Let's take discovery: How do you know when a new set of data becomes available that might be useful to the enterprise? There should be an answer to that question. The answer shouldn't be, 'Well, I'll wait until somebody tells me about it.' Curation is the next step, like putting things in a museum. You put data with data that makes sense, you put it in a place where you can care for it and treat it like an asset, not like some cost that you have to drive down.

Progression of the data in the model continues, so these steps have an interplay with each other and collectively. Later on in the life cycle, synthesis is the sense that we make out of the data, and fabrication is a form of context-specific storytelling. Fabrication might be producing a product, but in the case of some other organization that is a consumer of data, it might just be making that data make sense to an internal customer. Delivery is, how do you get the answer back? How do you speak to an external party or an external system in a way that party or system will understand and consume the answer?

“Never lead with a data set; lead with a question.”

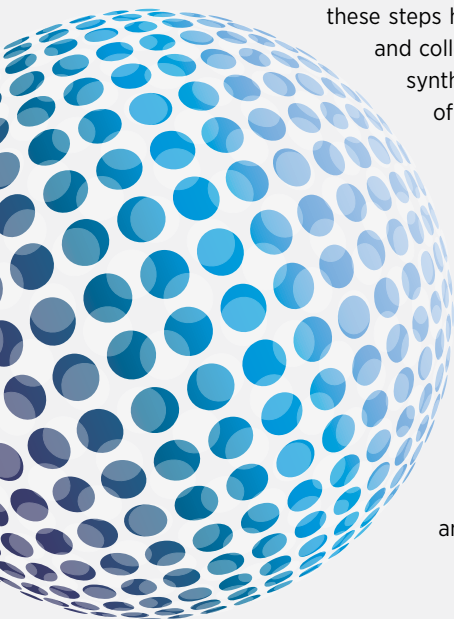
—Anthony Scriffignano
Chief Data Scientist,
Dun & Bradstreet

What's your advice to businesses looking to partner on data?

Number one: Get the question right before you start interrogating the data. Understand what the problem is that you're trying to solve. Never lead with a data set; lead with a question.

The second thing I would say is, understand the drivers of quality that are important to you in that data. Don't let somebody tell you how accurate their data is, or any other measure. It's fine to look at those measures initially, but understand what's important to you from your perspective, not what's important to a provider in selling it to you from their perspective.

The third thing is, always have some sort of closed-loop process. Never engage in dialogue that feels like, "Give me some of your data, I'll test it, and I'll let you know if I'm going to buy it." Consider instead a dialogue that might progress like this: "Let me tell you what I'm trying to do with the data, and let's look at a stratified, representative sample of your data. Then let me tell you what I saw, and let's create some sort of a closed-loop process to get this to a steady state that's the best we can make it. Then I'll decide whether that's good enough." Don't fall victim to the "dipstick test"—simply taking a convenient sample and trying to reach a conclusion, because that almost never works, especially in complex data sets.



DATA AS A DIFFERENTIATOR

As organizations look to adopt the new wave of coming technologies, like automation, artificial intelligence and the Internet of Things, their success in doing so and their ability to differentiate themselves in those spaces will be dependent upon their ability to get data management right. This will become increasingly important as connected devices and sensors proliferate, causing an exponential growth in data—and a commensurate growth in opportunity to exploit the data. It's unsurprising then that KPMG found 41% of top-performing organizations are making leadership in data and analytics a strategic priority.³

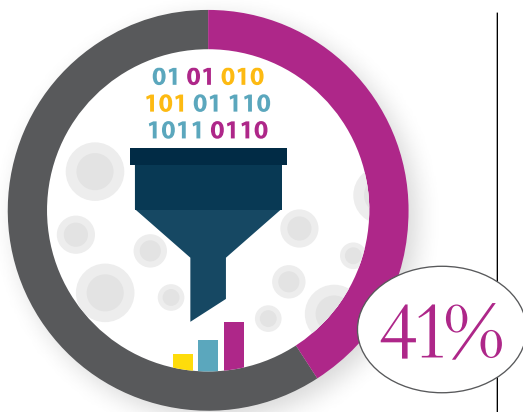
Those that position their organizations to manage data correctly and understand its inherent value will have the advantage. In fact, we may see leaders pull so far in front that it will make the market very difficult for slow adopters and new entrants. Joe Francica, managing director of location intelligence at Pitney Bowes, puts it this way: “Information obtained through the collection of data from transactions or other sources may not be perfect. But if your company has put a premium on correcting addresses, location and other contextual data, your information will be inherently better and a competitive advantage.”


“If your company has put a premium on correcting addresses, location and other contextual data, your information will be inherently better and a competitive advantage.”

—Joe Francica
Managing Director of
Location Intelligence,
Pitney Bowes



³ “2016 Global CEO Outlook,” KPMG International



41% of top-performing organizations are making leadership in data and analytics a strategic priority. 

—“2016 Global CEO Outlook”
KPMG International

Although potential sources and uses of data are still emerging in many cases, there are a few ways we already see data being used to help organizations differentiate themselves:

Customer experience: Data is central to the single customer view that organizations require to better understand their customers and tailor hyper-personalized engagement, especially as organizations seek to differentiate themselves on the experience they provide. Businesses that can attain better understanding of customers by harnessing not only their own data, but also the unstructured data their customers create in the form of social media and other types of messaging, will win at loyalty and engagement.

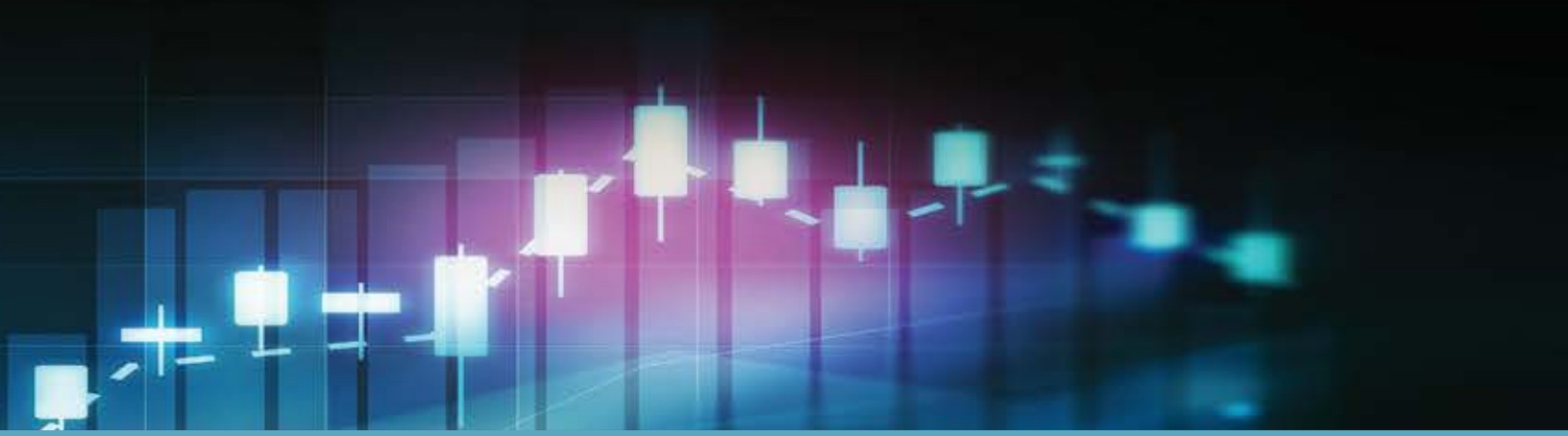
New products and revenue streams: Organizations that become adept at generating and managing their own data may discover new revenue streams by licensing the data to other users. Internally, data about how, when and where products are used may help designers or engineers refine products or launch entirely new ones, while other companies may exploit potential in existing data sets. Pokémon Go is a high-profile example of the latter, taking existing location data and street maps to create a compelling and popular game, and consequently creating a new channel for retailers to reach customers.

Operational efficiency: Improving data quality can improve efficiency in various ways. De-duplicating customer records, as one example, could result in lowered mailing costs.

BENEFITS OF GOOD-QUALITY DATA

In general, good-quality data can impact organizations in several invaluable ways:

Decision making: The better the data quality, the more confidence users will have in the outputs they produce, lowering risk in the outcomes and increasing efficiency. The old “garbage in, garbage out” adage is true, as is its inverse. And when outputs are reliable, guesswork and risk in decision making can be mitigated.



Productivity: Good-quality data allows staff to be more productive. Instead of spending time validating and fixing data errors, they can focus on their core mission. “If you’re using poor-quality data and you’re expecting the engineers to make up for it, you’re asking a lot,” says Adams.

Compliance: In industries where regulations govern relationships or trade with certain customers, especially in finance, maintaining good-quality data can be the difference between compliance and millions of dollars in fines. Compliance must be an ongoing focus as new regulations continue to evolve in regions around the world and wherever a company conducts business. Graph databases are emerging as an important tool for finance firms to understand the complex relationships among their customers and comply with anti-money laundering regulations.

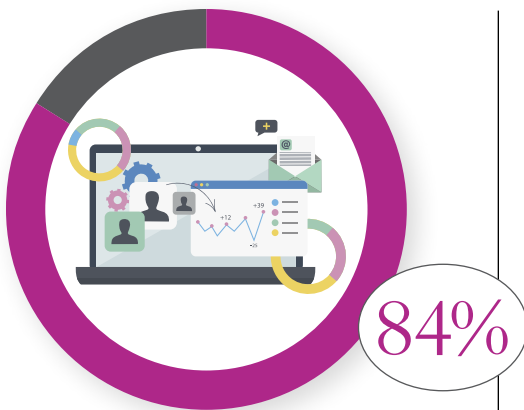
Marketing: Better data enables more accurate targeting and communications, especially in the omnichannel environments many organizations are striving toward. “Good-quality data allows us to do things like advanced analytics, which can drive our strategies around how we should market to our customers, which customers we should be marketing to, targeted

“The cost for companies that put less emphasis on analytics and more emphasis on instinct is that they’ll be putting the organization at greater risk.”

—Paul Hulford
CEO, Attain Insight

Data is central to the single customer view that organizations require to better understand their customers and tailor hyper-personalized engagement, especially as organizations seek to differentiate themselves on the experience they provide.






84% of CEOs are concerned about the quality of the data they're basing their decisions on.

—“2016 Global CEO Outlook”
KPMG International

marketing,” says Aaron Wallace, principal product manager for customer information management at Pitney Bowes. “The positive impact of feeding those processes with high-quality data is that they will produce relevant insights and drive the right kinds of strategy.”

COSTS OF POOR-QUALITY DATA

The costs of poor-quality data are more than simply the reverse of the benefits. Negative impacts of poor-quality data can include:

Undermining confidence: According to KPMG’s “2016 Global CEO Outlook,”⁴ 84% of CEOs are concerned about the quality of data they’re basing decisions on —and when there’s a lack of trust in data quality, confidence in the results it provides is quickly eroded. That can cause obstacles to gaining executive buy-in, dampening enthusiasm for further investment in data and quality improvement initiatives. “Like anything else, when people don’t trust the system, they look to other ways of making decisions,” says Attain Insight’s Hulford. “They may move back to traditional techniques—opinions from trusted individuals, instinct, experience—and while those things are very important, they’re also open to interpretation and vulnerability....So the cost for companies that put less emphasis on analytics and more emphasis on instinct is that they’ll be putting the organization at greater risk.”

Missed opportunities: “If you’re not doing it, your competitors are,” says Adams. “If they’re gaining more insights out of the data they have on hand and their ability to process it, they’re going to have insights that you don’t.” That might mean a company misses a critical opportunity for new product development or customer need that a competitor with a more mature understanding of data may capitalize upon. Adams says companies should treat data as an asset and manage it to maintain quality in order to derive insights that can lead to competitive advantage.

Lost revenue: Poor data can lead to lost revenue in many ways—communications that fail to convert to sales because the underlying customer data is incorrect, for example. Pitney Bowes’ Wallace points out that, in

⁴ “2016 Global CEO Outlook,” KPMG International



insurance, bad property information could cause revenue to be lost on premiums if they are set too low because of the data. One example is where property locations are estimated, instead of precisely specified. In most cases, that might not matter, but where the difference is a property—or a whole neighborhood—located inside or outside of a flood zone, revenue losses could be significant.

Reputational damage: Reputational costs range from the small, everyday damage that organizations may never be aware of to large public relations disasters. As an example, recall Apple’s widely panned Maps rollout in 2012. At the time, it quickly became clear that much of the underlying data was inaccurate or missing, resulting in a product that TechCrunch later called “barely usable.”⁵ Efforts to improve customer experience may also be undermined by bad data resulting in an incorrect spelling of a customer’s name, or obviously sending communications to a deceased customer. On the larger end, poor data in banking, for example, could lead to inadvertent trade with sanctioned governments or suspected terrorist financiers if institutions don’t have accurate enough information about the customers they’re trading with, resulting in PR fallout on top of punitive fines.



⁵ Dillet, R. (2016, August 08). The Apple Maps launch fiasco led to the iOS public beta program. Retrieved April 25, 2017, from <https://techcrunch.com/2016/08/08/the-apple-maps-launch-fiasco-led-to-the-ios-public-beta-program/>

“The cost for companies that put less emphasis on analytics and more emphasis on instinct is that they’ll be putting the organization at greater risk.”

—Paul Hulford
CEO, Attain Insight

FORRESTER: BEYOND TRANSACTIONAL DATA

Michele Goetz is a principal analyst at Forrester, whose expertise spans data management, data governance and data strategy research, with part of that particularly looking at data quality from both a practice and technology standpoint. Forbes Insights spoke with her about differentiating businesses through data applications that move beyond the transactional. The following conversation has been edited for clarity and length.

How and why is data quality critical for business?

Most businesses recognize that data quality is really central to their business. They know if they can't trust their data, if they can't determine how meaningful that data is, if they can't put it into context of how they would use that data—they realize they're not getting the most out of it.

[Data is critical] both from a strategic standpoint, as well as a more operational standpoint: How do you satisfy customers that are calling in to your support center with questions or complaints? How do you work with them quickly? How do you better cross-sell and upsell if you are working in, say, the utilities industry and there's bad weather that brings down your power grid in certain areas? How quickly can you respond to that and fix the infrastructure to bring power back to residents and businesses? All of that is data-driven, and if you can't trust your data, your business just can't operate at the speed of the markets and at the speed of demand anymore.

What are the innovative uses you're seeing for data, and what does that tell us about what's possible for the rest of the field?

Some of the interesting things starting to emerge are [to take] steps beyond transactional systems. Data quality isn't just applied to transactional and application data anymore—structured data. It's also applied to unstructured data, like call notes. It's applied to content, like contracts and information. Take the insurance industry. That can be very transactional: Simple questions can be readily answered—you ask to see your policy, your policy comes up. What if you have questions about how would I bundle my policies so that I have a better rate, for example?

If you can put that into a digital experience and be able to collect information to understand how customers ask for information—which is really the voice of the customer and deciphering call logs and putting that information into a recognizable language and lexicon to work off of. If you can understand search terms and search phrases that your customers are using as they're trying to find information on your website, if you can start interpreting—not only transactional information that's been coming through—but call notes and details of the policy contracts themselves, that becomes a lot more interesting.



Now you could potentially work hand-in-hand with a customer online, without an agent in the middle.

Now you could potentially work hand-in-hand with a customer online, without an agent in the middle. That, in terms of a quality perspective, means that how you not only manage cleanliness of that information, but also are able to carry it, to decipher that information and to put it into something that's meaningful and useful, it's also part of the data quality processes and strategy to move forward.

Do you see data as a differentiator for businesses?

It is absolutely a differentiator. More and more, companies are looking at how do they join the data economy, how do they take off this information they have, package it up, and turn it into a new product or offer? You see a lot of these insight services starting to become available—[but] the first thing you want to know is, what is the quality of that data? How can I trust it?

“If you can't trust your data, your business just can't operate at the speed of the markets and at the speed of demand anymore.”

—Michele Goetz
Principal Analyst,
Forrester

Q&A

Not only do your standards have to be very high, from a quality perspective, but the governance and transparency has to be there as well, otherwise you may not be able to enter the data economy in a way where you are perceived as providing value, or—certainly if you’re selling these services for revenue stream—your revenue stream isn’t going to be well supported if there’s a lot of churn due to the quality of your data. For organizations who haven’t come out of the world of data, or aren’t data brokers but want to enter that space, they really need to think about that. It’s not just the fact that they’ve got great data that somebody wants access to. At the end, data quality has to be there as a way to further differentiate it so that it’s trusted and you get the value added, both in a relationship and potential revenue stream.

What are some of the potential hard and soft costs associated with poor-quality data?

You can’t anticipate sometimes how people work around the system, or work around a process to bring information in—but you should at least have some business rules running behind the scenes in an automated fashion to do the right checks to ensure the right information is within the right field. I’ve seen, a number of times within financial institutions and even healthcare, where social security numbers have been entered into fields that are not PII (personally identifiable information) protected. You can now start accidentally releasing personally identifiable information about your customers and those that you work with. That’s another problem. There are privacy regulation hits that you can take, there are customer experience and relationship hits that you can take.

What advice do you have for businesses looking to invest or partner in data improvement?

The first thing, internally, is they have to recognize—beyond, “how do I integrate data together, across different systems”—but really think about it in the context of how is that data going to provide value? Putting that data into context of a business process, around a customer interaction, around a partner relationship and things of that nature, gives you a much better understanding about what level of quality you need, or what quality really means: the impact of not having good data, but also the impact of having great data.

The technology is honestly the easiest thing to purchase. It’s really, really important that organizations establish the benchmark and the understanding of their policies ahead of time and put the technology against that to ensure that it’s going to support what’s necessary. Also recognize that what you define today for your quality logic and standards will most likely change as you bring in more and more types and variety of data, or have new business objectives—so you want to have a governance program as well as technology that can adapt quickly and flexibly to those changing requirements.

It really begins with establishing that scaffolding of: What is data going to drive within my organization to help me grow my business, or is it helping me mitigate risk and drive efficiencies? And then in those contexts for specific areas, what should the data look like?

IRONSIDE: CAPITALIZING ON DATA

Ironside is an enterprise data and analytics solutions firm and systems integrator, and a Pitney Bowes partner. **Greg Bonnette**, vice president for strategy and innovation, spoke with Forbes Insights about some of the challenges around data quality, and some of the ways businesses are differentiating themselves through innovative uses of data. The following has been edited for clarity and length.

What are some of the challenges to getting data quality initiatives right? Why is this a difficult thing?

It is really, really hard. It's hard to get right, and it definitely takes time. And not everybody's willing to put in the effort or to wait for these things to evolve.

Why these sorts of things fail is because a lot of times people will embark on these initiatives without really understanding why, or how we look at this as an investment, right? So, for an organization to have any chance at success with data quality, they have to understand what their strategic objectives are with data and analytics, and how improving the quality of a certain domain of informa-

tion is going to help them [achieve those objectives]. It needs to be top-down.

What we see really frequently is that if the data is wrong or incorrect in some way, IT or data management and those types of roles will typically bear the brunt of that problem. But in reality, never are they really empowered to make a difference—because the cause of most of those quality issues typically traces back to broken business processes. And there's really not a lot an IT organization can do to change business processes. They can make recommendations, but without executive alignment that says, 'We are all working towards improving our data quality,' nothing's ever going to happen there, and nothing's going to improve.



For an organization to have any chance at success with data quality, they have to understand what their strategic objectives are with data and analytics.”

—**Greg Bonnette**
Vice President
Strategy & Innovation, Ironside



“The greater the quality and completeness of the data, the more feasible and accessible the opportunities that will exist for an organization.”

—**Greg Bonnette**
Vice President
Strategy & Innovation,
Ironside

What are some of the common characteristics of organizations that are ripe for data quality improvement?


There are a bunch of really soft, cultural indicators. Do people trust the data? Do people make decisions based on information in the organization today with confidence? If the answer is no, or they do it half-heartedly, or numbers are just flat-out discounted because there have been quality issues in the past or things have been brought into question—you may have a systemic data quality issue. So, that is number one: Is there a culture of trust in information for decision-making purposes?

Do you see strong data as a potential differentiator for businesses?

Absolutely. Yes, definitely. Having high-quality data allows you to be more precise in terms of how you market to your customers, how you recommend products to your customers. There's a real advantage there that can kind of be capitalized upon. Then there's the situation which we see developing more and more, which is the concept of data monetization; where you have organizations that are transactional processors and they're starting to realize that there's other value in this information. But if you're purporting that, you know, this is valuable information and you want to sell this to somebody else, you really have to be providing some level of guarantee around the quality of information.

What kinds of innovative uses of data are you seeing organizations undertake?

We see the greatest opportunities for data-driven innovation coming from using advanced analytics and data science techniques to improve existing business processes, or in some cases invent new processes and revenue streams. For example, using machine learning and other probabilistic methods, we've helped health insurers to identify cases of fraud, waste and abuse; police departments to optimize deployment of patrol offices to prevent crime from occurring; manufacturers to increase production efficiency and quality; and energy utilities to more accurately forecast natural gas demand during peak usage. These may lack the buzz factor of more recent bleeding-edge artificial intelligence applications, but they are still proven and attainable forms of business innovation that have generated hundreds of millions of dollars in top-line growth and efficiency gains. While advanced analytics is a powerful lever to solve business problems with tangible results, the potential to use these methods—and the energy, effort and cycle times required to get to a result—are hugely impacted by data quality. The greater the quality and completeness of the data, the more feasible and accessible the opportunities that will exist for an organization.



5 STEPS TO SUCCESSFUL DATA QUALITY INITIATIVES

While there can be varying approaches to data quality, interviewees outlined several key steps essential for success:

1. OBJECTIVE-SETTING

Any approach to improving data quality and management must be led by a clear articulation of the business use case, including alignment and agreement among all stakeholders on the objectives to be achieved. Executive sponsorship is also critical to success at this stage. “If the business leadership doesn’t see the value in the exercise, there is little hope for success,” says Ironside’s Bonnette. “Technologists and even technology executives will struggle to drive this agenda on their own.” It is critical to understand why data quality improvements are needed, to which data, who will benefit and what sorts of outcomes are desired. It is also the time to consider who owns leadership of the project, whether it will sit under one function or be shepherded by a cross-functional team. Understanding the objectives comes first, and then the model can be defined to serve those objectives.

2. DATA DISCOVERY AND INVENTORY

Data can be housed in numerous places and forms within a single organization: on paper, in filing cabinets, on spreadsheets, in multiple CRM systems and more. Understanding exactly what kinds of data an organization owns, where it currently exists and in what format is a critical next step toward data quality. This step can help organizations not only identify where the data is but also understand whether there are duplication and standardization issues, the completeness of the data and its suitability for the objectives as defined.

3. EVALUATION

Once the data is inventoried, an evaluation of the quality issues that exist can be completed. This can assess the processes and tools required to make it fit for purpose, most likely including a need to validate the data, standardize it, and match and de-duplicate records where necessary.

4. ESTABLISHING PARAMETERS

The evaluation leads into the next step, which establishes parameters around cleaning the data. It’s an “if this, then that” scenario, Bonnette explains. “If we see data that looks like this, then we need to do this to clean it up on the back end.” Concurrently, data should be looked at to understand root causes of issues—for example, whether users are working around an input problem or consistently entering incorrect information. And subsequently, the knowledge gleaned from that interrogation informs recommendations for data management on an ongoing basis, so that perhaps UI changes are made or further training provided to avoid issues in the first place.

5. BENCHMARKING AND AUDITING

Because data quality is an ongoing requirement, the final step may be one that is repeated often. It involves benchmarking data quality in a way that can be tracked over time and setting conditions for monitoring and alerts when quality deviates too much from those benchmarks. “It’s not one and done,” says Wallace at Pitney Bowes. In fact, he says it is more likely to be continual, especially as organizations move from overnight or weekend batch processing to real-time validation and de-duplication.

3
CHAPTER

BRINGING OUTSIDE DATA IN

“ I don’t know that organizations fully appreciate the growth and the speed at which data is becoming available, and the quality that’s available externally, but that will happen quickly enough.”

—Paul Hulford
CEO, Attain Insight

There is a vast amount of external data available that businesses can import to enrich their own data. This can be from government agencies and open data initiatives, other corporate enterprises that have monetized their own internal data or commercial data vendors. The number and variety of data sets and sources is quickly proliferating, presenting organizations with a rich opportunity to add context to their own data. “I don’t know that organizations fully appreciate the growth and the speed at which data is becoming available, and the quality that’s available externally,” says Hulford at Attain Insight. “But that will happen quickly enough.”



FIRST THINGS FIRST

Regardless of the source of external data sets, there is a certain amount of housekeeping that should be performed internally to make sure external data can be properly integrated and meaningfully used. A few considerations:

- What are the limits of existing internal data, and what gaps need to be filled?
- What modification, if any, needs to be made to IT infrastructure to incorporate external data? As Hulford explains: “If you can’t trust your own internal data sources and you want to bring external information in, you’re kind of at an impasse because you need internal culture and competence in order to use external data sources.”
- What is the quality of existing data sets? Because, as Dun & Bradstreet’s Scriffignano explains, errors in data sets are multiplied, not averaged out, especially in complex, data-dependent processes. “Maybe a simple way to say this is, before you start bailing all the water out of the boat, you should probably find the holes,” he says.



REAL DATA FOR REAL ESTATE

We spoke with three online real estate businesses, all clients of Pitney Bowes, for their experiences with external data sets and data in general. Each business relies on the roughly 750 multiple listing services (MLS) in the U.S. to provide it with basic listing data, which is then enriched with additional data sets imported from vendors, and then used to power different offerings to customers. All quotes have been edited for clarity and length.

• **Henry Behnke** is senior director, channel partnerships, at **Placester**, a Boston-based company creating software for real estate agents and brokers, including websites and lead management software.

• **Jake Lyman** is chief product operations officer at **ZapLabs**, which builds white-label, agent-, broker- and consumer-facing B2B and B2C software platforms and mobile apps.

• **Andy Woolley** is vice president of industry development for **Homes.com**, a nationwide online marketplace for home buyers and sellers.

“The core data set is entered by a human—a real estate agent. And so we spend a lot of time confirming information and using data sets to validate or replace human-entered data.

—**Jake Lyman**
Chief Product
Operations Officer,
ZapLabs

On data quality

Henry Behnke: It's maybe not a dirty little secret, it's probably a wide-open fact about the industry, real estate data can get pretty hairy. But sometimes the address information is incomplete, sometimes bath and bedroom information is incomplete, or price information is stale. And what we do is, we layer information that we get from [Pitney Bowes] on top of the listing to make it more accessible.

Jake Lyman: The quality of the external data that we integrate into our technology is really important. But the core data set is entered by a human—a real estate agent. And so we spend a lot of time confirming information and using data sets to validate or replace human-entered data. A good example would be when an agent enters a home's school zone into our platform, but for a firm understanding of the actual zoning



“We layer information that we get from [Pitney Bowes] on top of the listing to make it more accessible.”

—**Henry Behnke**
Senior Director,
Channel Partnerships,
Placester



“There are three components to measuring the quality of data: accuracy, latency, and the third would be the completeness or comprehensiveness of the set.”

—Andy Woolley
Vice President,
Industry Development,
Homes.com

boundary, we rely on an external data set from a company like Pitney Bowes. We move forward with the data set that we are confident in.

Andy Woolley: There are three components to measuring the quality of data: accuracy, latency, and the third would be the completeness or comprehensiveness of the set. So number one, accuracy, we always strive to compile from the most reliable source. The second challenge is latency; you want to make sure the data is updated frequently, that there's a minimal time delay between when the data changes and when it's reflected on our website. There are a lot of challenges involved in making sure the latency is minimized, how quickly we can process and publish changes, and that plays heavily into data quality. And then the third aspect, the completeness, it's more than just the physical attributes of a property. Schools are an important factor, neighborhood information is an important factor. So the challenge for us is in making sure that we have a high level of data quality in those three challenges. And that

requires us to invest heavily in making sure we have the right business partnerships, the technical infrastructure and a skilled team to operate it all efficiently.

JL: The core data set that we use comes from the local MLS, but lots of brokers and companies also use this. So for us, where we create a rich experience lies in where we customize the data and make enhancements. Some of the things we think about are how we will enhance and present data. What kind of user experience can we provide that makes it really easy for the end consumer? Having the MLS data alone is table stakes. How we enhance the data is really, really paramount.

On challenges

JL: One of the initial challenges is to understand how to merge and bring together data from multiple and varied sets. You want to apply analysis on the data and ensure it represents what it was originally intended to. One of the things we do is bring all the data sets together, cull and customize it. Once you have that data set and you start thinking about layering on additional components, for us, finding the value in those additional components is the most important aspect. It's more of an opportunity to say, okay, if we're going to bring in something like school zones or attendance zones, what kind of value does that bring to the end consumer? What does that tell us about consumer behavior? So it's a challenge to get it right and to interpret it correctly, but it also provides a huge opportunity for us to improve our business and to improve the overall user experience.

On features

HB: One of the key features on our websites is our natural language search. So, keeping an eye on that and making sure that searches are either returning relevant results or returning results at all has been something that we've paid attention to...Without data, it wouldn't function. The thing we've noticed with real estate data is it doesn't always have neighborhood data attached to it, and unfortunately, people tend to search via neighborhood in order to find a place to live. So the relationship we have with [Pitney Bowes] allowed us to birth natural language, but again, adding more data makes that even more useful over time.

On evaluating data vendors

HB: There's accuracy and there's completeness. And sometimes you can find a partner or a vendor that gives you both, sometimes you have to make a call between one or the other. So completeness, or making a determination around completeness, probably relies on you, the company, having some understanding of what's out there and then being able to audit what you may be offered...But I think making the call on whether or not you need all the information, or whether or not you need very accurate information, allows you to make the right decision and pick the right path forward.

JL: Coverage is primarily what we think about when evaluating our vendors—do they have everything we need? How broad is their set? How often and frequently is it reviewed, modified and updated? And what does the process look like? When there are discrepancies, how are those addressed? And how do you bring those discrepancies back to your company and address them with your partners in the future?

“There's accuracy and there's completeness. And sometimes you can find a partner or a vendor that gives you both, sometimes you have to make a call between one or the other.”

—Henry Behnke
Senior Director,
Channel Partnerships,
Placester

So the advice I give is to be fastidious when vetting data. That's just the first step, and you need to make sure that your team does a great job with that. Having a team that really understands how to use data to improve business is key. A big data set for your business is not going to provide the insights needed to really set a business apart and differentiate from the competition.

AW: It gets back again to what we think the three tenets are to data quality: accuracy, latency and completeness. Startups have come around, certainly over our 20 years of existence, and they've tried to take the path of least resistance to go find somebody that's aggregating this information instead of pounding the pavement and doing the hard work that's involved. And if you're going to have accurate data with minimal latency and as comprehensive as possible, it's going to take a lot of hard work to be able to dial in all three of those things, right? To really have good-quality data takes time, investment and a lot of hard work.

CHOOSING A DATA PARTNER

As a result of the challenges and complexities of data discussed in this report, any data partner brought on board needs to have a clear and accountable plan for working with an organization to improve data quality. Our interviewees weigh in on some of the other considerations at play in choosing a data partner:

PUT IN YOUR OWN AUDITING PRACTICE

As a first step, Michele Goetz suggests businesses really need to have their own auditing practice in place for evaluating external sources of data from vendors. It might be fine for the vendors to maintain the databases and platforms for simple email or mailing campaigns, but for a really holistic view of the customer, that data needs to be brought in-house. “Which means that all of the matching logic, all of the hierarchies, enrichment rules, all of the business logic and clean metadata that the service provider is doing on your behalf, all needs to get translated back into your organization,” she says. “You need to have a data quality engine at the front end that is running business rules and business logic and checking that the data is complying with your expectations to either one.”



“You need to have a data quality engine at the front end that is running business rules and business logic and checking that the data is complying with your expectations to either one.”

—Michele Goetz
Principal Analyst,
Forrester

REPUTATION

Reputation and longevity aren't everything when it comes to selecting a data partner, but they are important to get a sense of in the decision-making process. "Looking at the reputation of the company itself, how well known they are for quality versus price, freshness of content, etc. Those are the kind of standard things to do," recommends Pitney Bowes' Dan Adams.

METADATA

What does the vendor provide in terms of metadata about its data set? The latency of its data, how often it is maintained, the number of records it contains—and, ideally, suggests Pitney Bowes' Wallace, there's the opportunity to evaluate it in a proof of concept phase before writing any checks. Adams says the existence and availability of metadata is, in itself, a good sign. "Simply knowing [metadata about a set] exists goes a long way to ensuring that the suppliers themselves are going to be paying attention," he says.

SUPPLIER COMMITMENT OVER TIME

Data is in a state of constant flux, meaning that what was true and accurate in the past—whether six months ago or six minutes ago—may not still be true in the present. Even beyond the currency of the data, the world around us changes, and so does the industry. Consequently, it's important to partner where there's a commitment to maintaining data and bringing new innovations and products to customers over time, and not form a partnership that ends when the main implementation is complete.

SOURCE AUTHORITY

What is the authenticity of the data that is required and the reliability of the provider? Not all data sets are created equal. Pitney Bowes' Joe Francica gives the example of satellite data—with a multitude of satellites in operation in Earth's orbit, many owned and operated by different governments, research facilities and corporations, any given satellite could be considered to be an authoritative source. Some applications of satellite

data, however, may require specific spatial and spectral resolution. Existing satellites in orbit may not be necessarily fit for each purpose or application. Therefore, more flexible platforms, like unmanned aerial vehicles that are more ubiquitous, are being considered but may not be operated by an authoritative source.

LICENSE STRUCTURE AND WIND-DOWN CLAUSES

When bringing in external data sets, businesses should understand the commercial terms of use and termination—and in particular, what rights remain to use work derived from that data after the end of a contract.

TOTAL COST OF OWNERSHIP AND TIME TO VALUE

Total cost of ownership is about more than the price of purchase. Buyers should also factor in the cost of keeping software maintained and updated—and how well it supports agile, iterative methodologies that speed time to deployment. Time to value depends strongly on the particulars of a solution and an organization's needs, but Wallace suggests that agile solutions typically see at least some partial deployment in a working environment within three to six months, versus the typical 18 months to three years for a more traditional waterfall implementation method.

CUSTOMIZABILITY

Some software comes shipped with preset schemas, which may or may not fit an organization's purposes over time. Some vendors have a "black box" approach that allows little, if any, modification, so the ability to customize schemas should be ascertained up front. Wallace points out: "What we've heard from almost everyone is that those [canned schemas] are nice as a starting point, but they almost always wind up having to be modified extensively—and modifying them is not necessarily the easiest thing to do."

CONCLUSION

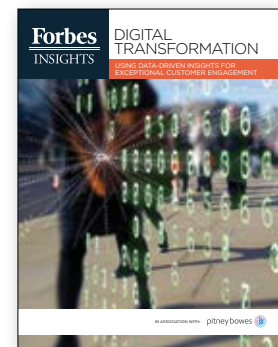
Improving data quality can be a daunting task, for many of the reasons outlined in this paper—sheer volume, rapid pace of change, legacy systems—and others that may be more specific to a particular business or industry.

Yet not only is it critical that businesses begin to tackle the quality of their underlying data to position themselves to take advantage of coming transformational

technologies, but there are also real rewards—in terms of improving customer experience, confidence in decision making and lowering operational costs—that can be realized now. Likewise, the costs of poor-quality data are real and present: reputational harm, missed opportunities and potentially even violations of the law when it comes to financial compliance.

Data is only going to increase in importance for businesses as technological change continues its push toward a more digitized world. But this provides an opportunity for those that capitalize upon it early to differentiate themselves and pull out in front. Laggards may find that it becomes exponentially harder to catch up the longer they delay on the data front.

ADDITIONAL REPORTS FROM FORBES INSIGHTS AND PITNEY BOWES:



For more information about Pitney Bowes, the Craftsmen of Commerce, visit
<http://www.pitneybowes.com/us/data>

ACKNOWLEDGMENTS

Forbes Insights and Pitney Bowes would like to thank the following for their time and expertise:

- **Dan Adams**, Vice President, Data Product Management, Pitney Bowes
- **Henry Behnke**, Senior Director, Channel Partnerships, Placester
- **Greg Bonnette**, Vice President, Strategy & Innovation, Ironside
- **Joe Francica**, Managing Director, Location Intelligence, Pitney Bowes
- **Michele Goetz**, Principal Analyst, Forrester
- **Paul Hulford**, Chief Executive Officer, Attain Insight
- **Jake Lyman**, Chief Product Operations Officer, ZapLabs
- **Anthony Scriffignano**, Senior Vice President, Chief Data Scientist, Dun & Bradstreet
- **Aaron Wallace**, Principal Product Manager, Customer Information Management, Pitney Bowes
- **Andy Woolley**, Vice President, Industry Development, Homes.com

Forbes

INSIGHTS

ABOUT FORBES INSIGHTS

Forbes Insights is the strategic research and thought leadership practice of Forbes Media, a global media, branding and technology company whose combined platforms reach nearly 75 million business decision makers worldwide on a monthly basis. By leveraging proprietary databases of senior-level executives in the *Forbes* community, Forbes Insights conducts research on a wide range of topics to position brands as thought leaders and drive stakeholder engagement. Research findings are delivered through a variety of digital, print and live executions, and amplified across *Forbes'* social and media platforms.

FORBES INSIGHTS

Bruce Rogers, Chief Insights Officer
Erika Maguire, Director of Programs
Andrea Nishi, Project Manager

EDITORIAL

Kasia Wandycz Moreno, Director
Hugo S. Moreno, Director
Lynda Brendish, Report Author
Kari Pagnano, Designer

RESEARCH

Ross Gagnon, Director
Kimberly Kurata, Senior Analyst
Sara Chin, Research Analyst

SALES

North America

Brian McLeod, Executive Director
bmcLeod@forbes.com
Matthew Muszala, Manager
William Thompson, Manager

EMEA

Tibor Fuchsel, Manager

APAC

Serene Lee, Executive Director

