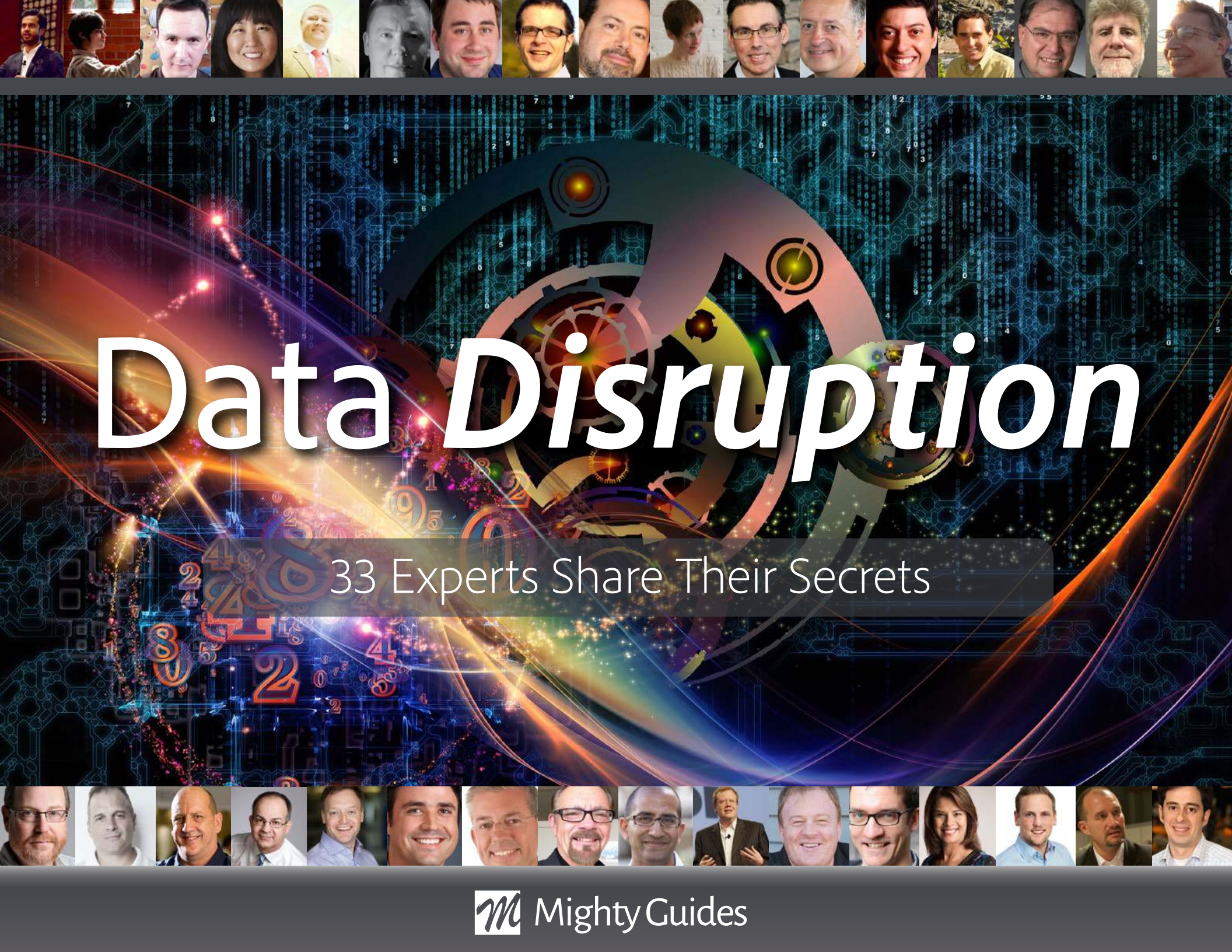

Mighty Guides

TABLE OF CONTENTS

INTRODUCTION

It is impossible to ignore the fact that we live in a data-driven civilization. Not only is the amount of data in the world doubling every two years, but the percentage of these data that are becoming valuable because of advanced analytics is also growing. The entire field of data capture and analysis is evolving so rapidly that organizations have difficulty keeping up. Yet data-driven business processes, competition, and the rewards of faster and more intelligent operations leave us with no other choice.

For a long time, our ability to capture data outpaced our ability to process it. This meant that large quantities of data were stored in data warehouses until some future time when tools would be available to find value in them or until they were discarded all together. Several things have happened in recent years to change this dynamic. One is the exponential growth in data; the other is the emergence of new platforms and technologies that make it possible to process data sets of almost unlimited size economically while lowering the cost and increasing the speed of analysis. These elements, combined with new analytic techniques and a growing use of machine learning to accelerate analytic methods, is changing almost every aspect of our lives.

To gain a fuller understanding of how modern analytical methods are being used in visible and not-so-visible ways, we approached data analytics experts from many fields and industries. I asked them to contribute essays about their experiences applying big data analytics. This e-book is a compilation of those essays. In it you will find discussions about new analytics technologies, how organizations can more effectively use their data assets, and many interesting use cases. The essays have been grouped into five sections:

- **Business Change.** Essays in this section speak to how advanced analytics are changing the way businesses operate. It is much more than a story about increased productivity and efficiency: it is a story about the complete transformation of traditional business models into something new and totally data driven.
- **Technology Platforms.** Essays in this section take a closer look at some of the tools and platforms that are making advanced analytics economical for organizations of all sizes.

INTRODUCTION

- **Industry Examples.** This section continues the discussion of transformative analytics technologies in the context of specific business and public-sector use cases.
- **Research.** This section focuses on how new-age analytics are changing the way scientists are conducting research and how they are speeding knowledge acquisition.
- **Marketing.** This section focuses on advanced, analytics-driven marketing strategies and techniques. These techniques are being used for everything from brand marketing to personalization to public relations to attribution techniques that enable companies to analyze their most effective marketing activities in real time.

It is my hope that assembling knowledgeable insights and experiences from so many different perspectives will provide a valuable glimpse into this rapidly evolving technology. I have found many of these essays both eye-opening and thought provoking. There is no question that advanced analytics will continue to play an increasingly important role in business, government, health care, knowledge acquisition, and a broad spectrum of human endeavor.

All the best,
David Rogelberg
Publisher

Mighty Guides

Mighty Guides make you stronger.

These authoritative and diverse guides provide a full view of a topic. They help you explore, compare, and contrast a variety of viewpoints so that you can determine what will work best for you. Reading a Mighty Guide is kind of like having your own team of experts. Each heartfelt and sincere piece of advice in this guide sits right next to the contributor's name, biography, and links so that you can learn more about their work. This background information gives you the proper context for each expert's independent perspective.

Credible advice from top experts helps you make strong decisions. Strong decisions make you mighty.

Business Change

CYCLES OF INNOVATION ARE BECOMING SHORTER

RANDY BEAN
CEO/Managing Partner,
NewVantage Partners

Randy Bean is CEO and managing partner of NewVantage Partners, a management consulting firm that he co-founded in 2001. He is a recognized industry thought leader and writes a monthly column on big data for *The Wall Street Journal*. Randy is a contributor to the *MIT Sloan Management Review* and *Harvard Business Review*. Randy holds a B.A. degree from Washington University in St. Louis, Missouri.

Twitter | Website

I have been involved in applying big data to business strategy for most of my career—a career that goes back long before "big data" became the popular technology idea it is today. Over the years, I have seen how technical innovations have changed the way businesses operate, but I have also seen how those new technologies often take much longer to work their way into the business process than enthusiasts imagine they will. Even so, it is clear that modern data analytics tools are changing the pace of innovation. It is also true that traditional approaches to the kinds of disruptive change big data analytics enables, such as resisting the change and protecting the franchise, are not so effective in a digital economy. Here are a couple of examples.

KEY LESSONS

1. **A LARGE ENTERPRISE CAN FORGO INNOVATION FOR A LONG TIME, BUT CYCLES OF INNOVATION ARE BECOMING SHORTER, WHICH PUTS THE NONINNOVATING COMPANY AT A DISADVANTAGE.**
2. **TO REMAIN A MAJOR PLAYER IN ITS GIVEN INDUSTRY, A COMPANY NEEDS TO DEVELOP NEW BUSINESS MODELS BASED ON BIG DATA–DRIVEN ENGAGEMENT.**

Not long ago, I moderated a roundtable discussion among technology executives from leading banking and financial services institutions. I asked the executives to talk about how they saw big data affecting them. Each spoke about how he or she envisioned big data affecting operations, and most said predictable things about improving efficiencies. They talked a good bit about how as competitors they could use big data in strategies that would enable them to differentiate themselves from each other.

> "Traditional approaches to the kinds of disruptive change that big data analytics enables, such as resisting the change and protecting the franchise, are not so effective in a digital economy."

CYCLES OF INNOVATION ARE BECOMING SHORTER

When it came time for one executive from one of the largest banks to weigh in on this topic, he looked around the room and said, "I don't see any of my competitors of tomorrow in the room today." He went on to list about 15 companies growing entirely new business models around big data–driven digital engagement, all of them carving out rapidly growing niches at the expense of traditional financial services. Another example is Uber, a totally data-driven enterprise that is turning the traditional taxi business on its head. Already, Uber has its own competitors.

Big data and modern analytics tools rapidly process vast quantities of data from every source possible: transactional data, sensor data, social data, historical data—all kinds of structured and unstructured data. These tools enable faster market response and faster innovation. They provide an analytical sandbox that enables business innovators to fail faster, fail better.

"Big data analytics tools provide an analytical sandbox that enables business innovators to fail faster, fail better."

ALL BUSINESS IS BECOMING CUSTOMER-CENTRIC

SVEN DENECKEN

SAP Global Vice President, Co-Innovation and Strategy SAP S/4HANA, SAP SE

Sven Denecken is the global vice president for Co-Innovation and Strategy at SAP SE, where he assesses customer and market requirements and supports SAP´s strategy. Through Co-Innovation projects, Sven uncovers key trends and best practices in the application of new technologies. Working with his teams, he supports alignment with customers, the ecosystem, and SAP´s field organization. Through Co-Innovation, he works to facilitate and enable sustainable relationships among SAP, customers, industries, and partners.

Twitter | Website

The ability to collect and analyze large amounts of relevant data rapidly enables businesses to operate in ways they have never operated before. We see this improvement in function in four key areas:

- **Customer-centricity.** This term refers to building a business model based on a complete view of customers' interests, behaviors, and preferences. Working on this model becomes possible only when businesses look at large amounts of structured and unstructured customer-related data.
- **Operational excellence.** This excellence includes a high degree of real-time automation, touchless responses to customer demands, and business process integration. It is possible only through rapid, in-memory analysis of data from many sources.
- **Global networking.** The Internet has set the expectation for a business network in that it is responsive and functional and provides global commercial reach. Such networking has also become the means of capturing the data that customers and connected things generate.
- **Rapid innovation.** By exploring large data streams and engaging in analytical sandboxing, businesses can quickly test and adopt new business strategies.

KEY LESSONS

1. BY EXPLORING LARGE DATA STREAMS AND ENGAGING IN ANALYTICAL SANDBOXING, BUSINESSES ARE ABLE TO QUICKLY TEST AND ADOPT NEW BUSINESS STRATEGIES.
2. BUSINESSES THAT THINK OF THEMSELVES AS TRADITIONAL BUSINESS-TO-BUSINESS OPERATIONS ARE INCREASINGLY RESPONDING TO THE DEMANDS OF THEIR CUSTOMERS' CUSTOMERS.

> "The ability to collect and analyze large amounts of relevant data rapidly enables businesses to operate in ways they have never operated before."

ALL BUSINESS IS BECOMING CUSTOMER-CENTRIC

The combined effect of these data-driven capabilities is the creation of business opportunities that disrupt markets and traditional businesses. For instance, some traditional asset-centric businesses are turning into service-oriented businesses. In a traditional taxicab business, the taxi company invests in cars, and then depends on revenue that cab drivers generate to ensure a good return on that investment. In contrast, Uber simply delivers a transportation service. The taxicab assets are no longer part of the business equation.

Another example of a business model in transition is order fulfillment. Traditionally, a customer accesses a digital storefront where he or she clicks a button to make an online purchase. That storefront sits on top of a traditional fulfillment operation, which in turn relies on manual picking and packing of the purchased item. This kind of operation compensates for uncertainties in buying patterns by maintaining a large inventory. The challenge is to use structured and unstructured customer-centric data to anticipate buying patterns, automate fulfillment, and integrate the supply chain more tightly. Real-time analytics platforms are beginning to make that possible, but they require extending digital integration from end to end.

The key to digital business is the increasing use of customer-centric business models. Even businesses that think of themselves as business-to-business operations will find themselves responding to the demands of their customers' customers, and these businesses will model all their activities—marketing and sales, production, supply chain, even finance—around their customers' needs.

> "Some traditional asset-centric businesses are turning into service-oriented businesses."

CONNECTING THE DOTS

MARSHALL SPONDER
Lecturer, Zicklin School of Business, CEO, WebMetricsGuru INC

For more than a decade, Marshall Sponder has influenced the development of the digital analytics industry with his WebMetricsGuru writings, which focus on social media metrics, analytics, and media convergence. Marshall teaches Web Intelligence at Rutgers University and the Zicklin School of Business, where he is a faculty lecturer. He is the author of *Social Media Analytics: Effective Tools for Building, Interpreting, and Using Metrics* (McGraw-Hill, 2011); he is currently working on his second textbook on Digital Analytics for Marketers to be published by Routledge in late 2017.

Twitter | Website

The implications of superfast, in-memory computing became clear to me recently during an event at one of the universities where I teach marketing.

I was speaking that day with a training manager of a leading programmatic advertising vendor. For those that are not familiar with the term "*Programmatic*"; it is one of the newest evolutions of digital advertising that increases efficiency while reducing costs of reaching targeted audiences by collecting and processing big data through superfast computing. The goal of Programmatic Advertising is to harvest eyeballs at the precise moment a consumer is ready to make a purchase decision.

One of the educational institutions I teach at had been losing potential students to other universities that charge much higher tuitions for similar curriculum and programs.

I mentioned that situation to the vendor, saying that some academics at the university in question blamed demographics trends for the declining enrollment. However, based on a conversation with the aforementioned Programmatic training manager suggested a different reason: our competitors' digital marketing agencies had established business relationships with Programmatic Advertising vendors that allowed them to actually do Programmatic marketing of their educational programs at scale—something my institution hasn't yet done (for a variety of reasons). *It was a revelation moment for me!*

> "Ask any taxicab company how disruptive Uber has been."

KEY LESSONS

1. **SUPERFAST, IN-MEMORY COMPUTING WILL TRANSFORM THE COMPETITIVE LANDSCAPE AND WILL BE MASSIVELY DISRUPTIVE.**
2. **PAY ATTENTION. YOUR BUSINESS MODEL MUST BE INFORMED BY DEVELOPMENTS IN THE SUPERFAST COMPUTING SPACE.**

CONNECTING THE DOTS

By overlooking the impact of digital programmatic advertising as part our university's branding we were losing ground to other universities who had the means and insight to employ such methods at their disposal. I believe higher educational institutions using Programmatic technologies attracted more students precisely because they put their targeted messaging in front of the right audience of students just as they were deciding which school to attend.

On price alone, we could not compete. That's disruption. As I look around the digital landscape, I see many similar examples where superfast, in-memory computing transformed the competitive landscape. Here are a few examples:

- **Uber.** Already hugely successful, this company could never have existed before. It is entirely based on big data crowd-sourcing, allowing anyone who has a car to offer a ride to anyone else in any city where the company operates. Ask any taxicab company how disruptive Uber has been.
- **Auto-driving cars.** In the not too distant future vehicles will require so many sensors to move safely without a human driver. I suspect this new technology will create disruptions in cities and states as well as in the federal government as parts of roads and highway lanes begin to be reserved for auto-driving vehicles.
- **Rich media.** A big bugaboo in search engines is their failure to automatically detect, analyze, and tag billions of online images and videos. Big data in-memory processing will change that. Facebook can already accurately tag people's faces while Google has systems that can annotate images with a short description. Eventually similar technology capabilities will be applied to video and audio files.

> "Taking it back to the world of superfast in-memory computing - you cannot afford *not* to know what is going on in this space."

My advice to business leaders is simple: just pay attention to the rapid advances in technology yielding all of these changes. Taking it back to the world of superfast in-memory computing - you cannot afford *not* to know what is going on in this space. Your business thinking—your business model—must be informed by what is happening in this space. When you understand the impacts of these new technology offerings, you won't have to worry so much how it all applies to you. That will be obvious.

DISTRIBUTED HUMAN PROCESSING

SATYEN SANGANI
CEO, Co-founder, Alation

Satyen Sangani is the CEO of Alation. Before Alation, Satyen spent nearly a decade at Oracle, ultimately running the Financial Services Warehousing and Performance Management business, where he helped customers get insights into their systems. Prior to Oracle, Satyen was an associate with the Texas Pacific Group and an analyst with Morgan Stanley & Co. He holds a master's degree from the University of Oxford and a bachelor's degree from Columbia College, both in economics.

Twitter | Website | Blog

Compute and storage are two orders of magnitude cheaper than they were 10 years ago. As a consequence, we now collect and analyze an unprecedented array of data forms; data, too, are unconstrained.

So, where is today's bottleneck? It's within the line-of-business (LOB) workers and their ability to reliably self-serve analytics.

We have given high-end data scientists tools that allow them to use our unconstrained compute, storage, and data resources, but there are probably 500,000 to 1 million such people in the world. We have not extended similar tools to the broader worker population.

KEY LESSONS

1. **TAP INTO THE POWER OF DISTRIBUTED HUMAN PROCESSING.**
2. **INNOVATION IS FUNDAMENTALLY DRIVEN BY INSIGHTS. IF YOU GIVE ALL OF YOUR STAFF THE POWER TO BE DATA ANALYSTS, IMAGINE THE INNOVATION AND DISRUPTION THAT COULD DRIVE.**

The question then becomes, How do I make superfast data available to the 4,000 LOB people in my company? Put another way, how do we make *every* one of these 4,000 people as smart as any one of these 4,000 people? My answer: data literacy.

To understand the concept, think about Yelp. Everyone on Yelp has the ability to review restaurants because they know how to use the interface and they can see everybody else's past comments. They have both the means and the context to contribute.

> "The question becomes, How do I make superfast data available to the 4,000 LOB people in my company?"

DISTRIBUTED HUMAN PROCESSING

You can create an entire staff of analysts in a similar way, simply by giving them the tools and the context for using them.

The wife of friend works as an attorney at a payments company. She is not conversant with database queries and so had to rely on analysts to produce the data she needed. Even when the matter was urgent, getting the data sometimes took weeks.

Frustrated, she started studying past answers to her previous requests. By reverse-engineering those answers, she was able to find what analysts had done to arrive at them. She effectively started by copying their work to get her own answers. In this case, she used our product to help her, so she had both the context and the tools to generate her own insights.

> “Arm people not only with data but also information about how to use those data.”

We must tap into the power of distributed human processing. How?

- **Distribute information with context.** Arm people not only with data but also information about how to use those data. You cannot give people a big file full of information that could potentially contain PII, expecting them to do the right thing. Train them, and offer them the tools they need to gain insights.
- **Measure measurability.** Make sure people back up their reviews, their decisions, and their ideas with data. It is not good enough to make data available, and then allow people to do whatever they want with it. Expect measurable results. That has to be a cultural process.
- **Foster collaboration.** Ultimately, it is that collaborative element that generates data literacy because it allows you to create context socially—by having people talk to each other.

Competition effectively forces every participant in an industry to innovate. Innovation is what gives you the one up, and innovation is fundamentally driven by insights. A staff of 4,000 data analysts? This is the big disruption.

SUPERFAST DATA PROCESSING MAKES PEOPLE BETTER AT WHAT THEY DO

QUENTIN CLARK

Chief Technology Officer, Member of the SAP SE Global Managing Board, SAP

Quentin Clark is responsible for driving SAP's technology vision and leading the company's efforts to build and innovate world-class products that affect people, organizations, and customers. With more than 20 years of enterprise experience, he has been instrumental in developing and driving product strategy as well as leading industry-disruptive product launches. Before SAP, Quentin held various leadership positions at Microsoft, most recently overseeing product development for its entire suite of data products.

Twitter | Website

Humans can no longer keep up with the information flow on our planet. With analytics models that make sense of massive amounts of data, however, it becomes possible to keep up and use growing amounts of information. In a business context, it becomes possible to go from high-level views of a process, drill down to the minutest details in real time, and build analytical models into that view to give it meaning. So, the applications make their users smarter and enable them to make better choices.

For example, if a new paper is published in the medical field at 9:00 AM, and if at 9:05 AM a doctor has completed a patient examination and is interacting with his knowledge base system as part of his evaluation, he wants the information in that paper to be available to him. Nobody wants to hear that that doctor never got around to reading that paper because he doesn't subscribe to that publication and therefore he didn't know about a treatment that would have helped his patient. With an analytics-driven knowledge base, a doctor can keep up with the volumes of new research taking place continuously. The machine does not replace the doctor, but it does help the doctor wade through a vast body of information to focus on the right aspects of a problem.

KEY LESSONS

1. AN ANALYTICS-DRIVEN KNOWLEDGE BASE CAN HELP A DOCTOR WADE THROUGH A VAST BODY OF INFORMATION TO FOCUS ON THE IMPORTANT ASPECTS OF A PROBLEM.
2. APPLICATIONS THAT INCLUDE BUILT-IN ANALYTICS MODELS ENABLE USERS TO MAKE BETTER CHOICES.

> "Humans can no longer keep up with the information flow on our planet."

A much more mundane example would be a retail store in which an employee fails to show up for work. When this happens, the store manager needs to go into his or her office, launch a website, and review a variety of employee records to figure out who would be the right person to fill the shift. This is a costly solution because it might take an hour to bring someone in to fill the shift, and the highest-paid person in the store must be off the floor to solve the problem.

An alternative scenario is that the store manager receives a message before the shift begins saying that this person will not show up. The system has already analyzed the availability of off-duty employees and their skill sets, reviewed specials the store is running that day, verified the level of business, and assessed other factors. Based on all those data, the system makes a recommendation for who should be called to fill the shift. The store manager can agree or work with the system until the optimum worker is selected, and then the system makes the call. The entire process has taken less than a minute of the store manager's time, and the staffing gap is minimal. An analytics model drives everything: it considers all the employee data, sales data, product line goals, work schedules, and other information that go into selecting the best employee to fill that shift, and the analysis happens instantly.

Analytics models applied to large amounts of granular data that are processed instantly make people in decision-making roles better decision makers. Information makes people better at their jobs.

“Analytics models applied to large amounts of granular data that are processed instantly make people in decision-making roles better decision makers.”

DISRUPTION IS PART OF ADVANCEMENT

KIRK BORNE
Principal Data Scientist,
Booz Allen Hamilton

Kirk Borne is a member of the NextGen Analytics and Data Science initiative within the Booz Allen Hamilton Strategic Innovation Group and an advisor for several other firms. Previously, he was professor of astrophysics and computational science at George Mason University, where he did research, taught, and advised students in the graduate and undergraduate Informatics and Data Science programs. Prior to that, he spent nearly 20 years supporting large scientific data systems at NASA.

Twitter | Website | Blog

In the field of analytics, one of the latest and greatest algorithms people talk about is one called *deep learning*. Deep learning is computing that allows you basically to discover implicit data, such as what is in an image, without first providing an example. A few years ago, facial recognition was a hot technology. You could feed this algorithm an image, and it would perform facial recognition to identify that image, whether it was a person or an object, a picture or a video. That was explicit data discovery.

Let's use the 2014 Boston Marathon bombing to understand how implicit data discovery—deep learning—might work in a disaster situation. During the aftermath of the bombing, fast data processing could have made it possible to tap into surveillance cameras and perform a fast pattern-recognition sequence among all the chaotic things that were happening to identify people in medical distress. This could be done just based on facial expressions, body motion, and body movement. We could have quickly identified the people who needed the most attention.

This is what this type of data processing enables. A human can look at those images and see who needs help, but there could be thousands or hundreds of thousands of images to look through. There's no way even a team of people could do that rapidly.

> "You want fast data collection and storage with superfast processing so that you can get immediate benefit from the data you are analyzing."

KEY LESSONS

1. **SUPERFAST DATA PROCESSING, SPECIFICALLY IN THE FORM OF DEEP LEARNING, IS GOING TO ALLOW ORGANIZATIONS TO TAP INTO IMPLICIT DATA, NOT JUST EXPLICIT DATA.**
2. **BEFORE ORGANIZATIONS CAN TAKE ADVANTAGE OF SUPERFAST DATA PROCESSING, THEY MUST LEARN TO RELINQUISH CONTROL OVER DATA STORES TO SHARE ACROSS THE ORGANIZATION AND POSSIBLY ACROSS OTHER ORGANIZATIONS, AS WELL.**

DISRUPTION IS PART OF ADVANCEMENT

It's more effective to employ fast processing with minimal human intervention. That's the general idea here. You want fast data collection and storage with superfast processing so that you can get immediate benefit from the data you are analyzing.

That kind of data collection is already happening. What's missing is the actual storage of those data in a large cluster, like Apache Hadoop. To move forward, we need extra computational power on top of that storage unit. It's not just storage for future playback: the software is playing it back. In this example, each hospital is going to need its own high-performance computing center as part of its operations.

Of course, that is not what's happening. That's where this whole concept of *crowd computing* comes into play. Data centers and communities pool their resources and have shared services; many different organizations can share one big compute center and data center. Then, any organization can basically rent those services.

In my disaster example, hospitals may need these capabilities only for a disaster response. It's an unpredictable spike in their need, and that's when the crowd computing model works so well. They pay only for the computing and data storage when they are needed.

> "We have to share information and data across the boundaries of our organizations and even across different organizations. Silo smashing is disruptive."

I think business disruption will happen when we start doing all the things fast computing makes possible. We have to share information and data across the boundaries of our organizations and even across different organizations. Silo smashing is disruptive. The IT department wants to own its resources. Every other business unit wants to own its own data. No one wants to share.

That's not beneficial to the business, but it is human nature. The willingness to share resources and data will be beneficial to organizations, but people are worried about obsolescence. As we move forward, the work we do will look different. This disruption, which people perceive as a negative thing, is scary, but the benefits are going to outweigh the cost of any disruptions.

DATA: A COMPULSORY OBSESSION

STEEN KJØNG PAULSEN
Business Intelligence Consultant, Inspari A/S

Steen Kjøng Paulsen has, throughout his professional life, been engaged in the business intelligence (BI) and analytics industry. He has taken the journey as software developer, head of research, and consultant. He is currently in the consulting business, where he leads the initiatives on Apache Hadoop–based projects. He is strongly engaged in emerging technologies within BI and analytics and thus has a wide area of expertise using big data.

Twitter I Website

I am data obsessive. Many computers in my home collect data on my water and energy usage, what times the home security alarm is activated, what the weather patterns are in my home country of Denmark. This is no passive habit: data influence my behavior.

For instance, weather data combined with fuel economy data have altered my driving habits. I allow myself to drive a little faster on days when neither heat nor air conditioning are required to keep me comfortable in my car. The extra speed will not reduce my fuel efficiency: I have the data to prove it.

Perhaps you are beginning to see why I think businesses also should become data obsessive.

Some are, of course. Take the power company I recently worked with. It was plagued by periodic shutdowns. The problem was easily fixable, but no one had a clue what caused it. Those shutdowns were expensive, not to mention upsetting to customers.

> "In-memory computing and superfast data processing at scale can disrupt industry by enabling the commoditization of previously unavailable or unusable data."

KEY LESSONS

1. **BECOMING DATA DRIVEN BEGINS WITH BECOMING DATA OBSESSIVE.**
2. **EXECUTIVES NEED NOT BE DATA OBSESSED THEMSELVES, BUT THEY SHOULD SURROUND THEMSELVES WITH PEOPLE WHO ARE.**

The company used big data analytics to investigate the problem. Sensors on its equipment, many of which take readings 50 times a second, provided a flood of data that, once successfully analyzed, showed that a small, inexpensive water valve was the culprit. Without big data analytics, no one would have suspected that a cheap valve could cause full-scale shutdowns.

In-memory computing and superfast data processing at scale can disrupt industry by enabling the commoditization of previously unavailable or unusable data. We no longer must rely on the tightly formatted, structured data of relational databases. With technologies like Apache Hadoop, Apache Spark, and Scala, we can mine and analyze masses of unstructured video, audio, tweets, newspaper articles, and machine sensor data, integrating them all into our daily work lives in real time.

Think back to the power company. By stuffing its sensor data into a Hadoop cluster, it used vast computational power to aggregate and analyze information captured over a period of months. It generated value from previously useless data and added real power to its business process.

“Data that historically were impossible to use because of slow analytics systems can now be game changers.”

These arguments and lessons hold across industries. Data that historically were impossible to use because of slow analytics systems can now be game changers. To my mind, it goes without saying that you should strive to become data driven. Perhaps it is less intuitive to suggest this: becoming data driven begins with becoming data obsessive.

If as an executive you cannot quite generate that level of fascination, then certainly you should hire and strategically deploy some obsessive-minded data nerds in your midst. If your organization does not have a cultural fascination with or flair for using big data, your attempt to become data driven will devolve into just another boring task. It will never quite be completed.

Base your decision on facts, not feelings. Make your industry data driven. I can promise that it will improve your business because you will finally know why things do what they do.

MANY ORGANIZATIONS AREN'T SURE HOW TO USE FAST DATA TECHNOLOGY—YET

PAUL HAWKING

Associate Professor, Information Systems; SAP Mentor, Victoria University

Paul Hawking is an associate professor in Information Systems at Victoria University. He is considered one of the leading commentators on enterprise resource planning systems and business intelligence—specifically, SAP solutions. His knowledge is well respected in both industry and academia, and he is often asked to assist companies with their SAP strategies and solutions. Paul has presented at leading SAP and academic conferences around the world and was the first academic to achieve SAP Mentor status.

Twitter | Website

Currently, the problem with the fast data industry is that a lot of companies know that it's important, but they do not know what to do with it. I see three different approaches to this issue:

- I see one group that has adopted in-memory computing just to do traditional tasks faster. From a business intelligence perspective, they have adopted SAP HANA technology to speed up their reporting and processing. They are actually digging deeper into the data because of the speed at which the data appear.
- Another group is taking advantage of the functionality indirectly because SAP is releasing new products built on HANA functionality. It's not speeding up traditional tasks but new solutions built on fast data technology.
- The third group consists of companies that are using the in-memory technology to create new business models.

An example of that third set of companies is a distribution or logistics company I am working with. This company is using the Internet of Things to monitor its trucks and using the geospatial information and the speed of those trucks to identify and try to warn truck drivers about over-speed situations or dangerous corners. The company is also gathering data about the length of idle time the truck has. Then, it's analyzing that data to see if the idle time results from a holdup at the customer's end for unloading the trucks or traffic congestion.

KEY LESSONS

1. THE ORGANIZATIONS THAT INVEST TIME INTO FAST DATA PROCESSING TECHNOLOGIES WILL BE THE ONES TO LEAD THE WAY FOR OTHER COMPANIES WAITING TO SEE WHAT COMES FROM THOSE TECHNOLOGIES.
2. ORGANIZATIONS NEED A ROADMAP TO GUIDE THEM THROUGH THE STEPS OF ADOPTING AND IMPLEMENTING FAST DATA TECHNOLOGIES, WITH THE UNDERSTANDING THAT WE ARE JUST BEGINNING TO SEE THE FULL POTENTIAL OF THE TECHNOLOGY.

> "Currently, the problem with the fast data industry is that a lot of companies know that it's important, but they do not know what to do with it."

This information helps the business find better contractual ways to reroute the trucks and improve green emissions. If a truck is being held up for 30 minutes at a location, then the company can renegotiate its contract based on how long a truck sits before it's unloaded. For green emissions, if the truck is idling for a prolonged period of time, emissions are far greater. The business can use the data it is collecting to ask why the truck is sitting there and what can be done to improve that time? That improves emissions.

If you look at these three sets of users together, there is a maturity model. Most companies have no idea what to do with any new technology. The challenge is the confusion about the fast data offerings. Numerous opportunities are associated with these technologies. Making existing processes more efficient is the easiest and safest approach. These efficiencies can result from faster processing and integration or through improved analytics. Companies could use these technologies as a game changer.

All this is based on companies being aware of what game they are in and where the opportunities are. Some suggest that companies look at the underlying fundamentals or assumptions of their business and investigate how the business can use this new technology to turn such assumptions upside down. If fast data technologies were available then, how would things be done differently?

“Some suggest that companies look at the underlying fundamentals or assumptions of their business and investigate how the business can use this new technology to turn such assumptions upside down.”

I don't think companies realize the extent of the change that is coming. They know they are going to head down this path, they know that things will improve, but they do not realize the power of what's coming. To start taking advantage of fast data applications, organizations must use in-memory technologies to improve business analytics bottlenecks and implement business process solutions based on in-memory technology. Only then can they investigate how to use their knowledge of the technologies to differentiate themselves from their competitors. With this technology, organizations will be more empowered. It will be a big paradigm shift in the way people do their jobs. Users will not be pulling the data from a transactional environment. Instead, the data will be there as users do their job. This brave new computing world will provide greater insight into what people can already do.

Technology Platforms

HIGHER PROCESSING SPEED AT LOWER COST IS CHANGING EVERYTHING

SCOTT SHAW

Solutions Engineer, Hortonworks

Scott Shaw has more than 15 years of experience in data management and business analytics. He has worked as an Oracle and Microsoft SQL Server database administrator; he was a principal consultant and headed a data management practice. Scott worked closely with clients to design and model complex data warehouses. He now works as a solutions engineer for Hortonworks, providing architecture and deep learning around open source Apache Hadoop. Scott lives in Saint Louis with his wife and two children.

Twitter | Website

Companies have been processing, dissecting, aggregating, and analyzing data for decades. Massive database systems or mainframes take in large data sets, and business analysts use various tools to drill through the data to create operational or forecasting reports. These reports drive the business and help companies determine the current state of affairs and develop insights into new business opportunities. Database systems excel at putting data sets into memory for fast processing. As memory prices drop, larger and larger data sets are pushed into growing memory caches. None of this is especially new, and it is still the primary means most companies use for their analytics. So what has changed?

KEY LESSONS

1. **HADOOP PROVIDES INEXPENSIVE, HORIZONTAL, AND LINEAR SCALABILITY AT THE STORAGE AND COMPUTE LAYERS.**
2. **DIFFERENT DATA SOURCES CAN NOW BE COMBINED AND VIEWED THROUGH A SINGLE PANE OF GLASS.**

We have undergone a paradigm shift in data platforms. This shift started at the lowest levels: storage and compute. Apache Hadoop provides inexpensive, horizontal, and linear scalability at the storage and compute layers, which means that companies now have an analytics platform that easily stores and processes data in parallel across all systems. Complex data-management systems (typed data types, relational constraints) no longer hinder data ingest, and tool set and scalability limitations no longer hinder data egress. Traditional relational systems were built specifically for random read–write, row-level access. If you wanted to provide analytics that required longer, sequential access, you had to build a new system. Hadoop allows for a mixture of data access (batch, interactive, real time) all on the same platform, all acting on the same data.

> “Hadoop allows for a mixture of data access (batch, interactive, real time) all on the same platform, all acting on the same data.”

With this capability, web log data can live next to customer relationship management data. Sensor data can live next to internal product data. Streaming data can be stored next to relational data. These data sources can then be combined and viewed through a single pane of glass. In machine learning, these different types are referred as *features*. The idea is that the more features you introduce, the better your model. Take for example home energy use. Thermostat data are fine, but these data do not complete the picture. What if you also introduced heating, ventilation, and air conditioning sensors; weather data; historical energy use; and square footage data? By adding features, you build more accurate models. The challenge is that adding features increases the demand on processing.

Hadoop solves this issue in two ways. First, it provides the means to store massive amounts of data at scale, which is the foundation of the model. Second, through a compute process such as Spark, companies can quickly build and train models in memory. Now, the system can look at all those data in real time and provide more accurate projections for system maintenance, power consumption, and other control functions.

With faster data processing tools, not only can companies react faster to a changing market, but they can begin to run their analytics pretransaction. Companies using Hadoop and in-memory processing can build more accurate predictive models that anticipate action in the marketplace. Companies can promote to customers most likely to buy their products. Companies can target activities most likely to be fraud. Physicians can engage with patients who are most likely at risk. In a sense, the flexibility of distributed storage and the speed of in-memory processing have created a competitive race for companies in which the realization of future gains is determined by how far and how deeply they look into the past.

> “Companies can build more accurate predictive models that anticipate action in the marketplace.”

A DATA RIVER RUNS THROUGH IT

JUSTIN LANGSETH
CEO,
Zoomdata

Justin Langseth is CEO of Zoomdata, developers of the fastest visual analytics for big data. Zoomdata is Justin's fifth startup: he had previously founded Strategy.com, Claraview, Clarabridge, and Augaroo. Justin is an expert in big data, business intelligence, text analytics, sentiment analytics, and real-time data processing. He graduated from MIT with a degree in management of information technology and holds 14 patents. He is eagerly awaiting the singularity to increase his personal I/O rate, which is currently and frustratingly pegged at 300 baud.

Twitter | Website

Data these days are so big and fast that, for me, the only practical way to handle them is as a continuous stream, end to end. That, of course, is a major disruption to the way we have been doing things.

One casualty of that disruption is the traditional extract, transform, load method of batching and transforming data. If you depend on what used to be called *batch windows*, where you stop everything and cascade in a bunch of integration steps, you will find that batch windows fail if they are overloaded with rapid-fire massive data. Then, you must re-run and debug the batches.

Data warehousing—the consolidation of data in a single, carefully designed storage system—is also fading away. By harnessing the power of Apache Spark, Apache Hadoop and big data, and then in-memory processing to logically join data as needed, there is no need to gather it and store it in advance.

KEY LESSONS

1. **TREATING IT AS A CONTINUOUS, REAL-TIME STEAM IS ONLY PRACTICAL WAY TO HANDLE BIG DATA.**
2. **THE OLD APPROACHES, SUCH AS THE ETL METHOD OF BATCHING AND TRANSFORMING DATA, AND DATA WAREHOUSING ARE FADING AWAY.**

> "Afterward, the customer enjoyed a newfound capacity to handle, manage, and act on IoT data."

A DATA RIVER RUNS THROUGH IT

My company's technology has Spark embedded as a native application. It is built to be completely stream-native, end to end. Because we are a startup and because the open source Spark environment is so robust these days, we were fortunate that we could develop that way. It allows us to help customers deal with real-time sensor, equipment, and device data from the Internet of Things (IoT).

One client is a computer device company that has hundreds of thousands of devices deployed globally. Those devices generate massive continuous streams of data daily. Our customer needed a way to harness and analyze those data, to identify device anomalies and problems and respond in real time. Previously, the client would have spent millions building a system to ingest all those data nightly, and then carefully massage, clean, and organize it to eventually generate actionable information. Such an approach was no option for our client. The data transport and extraction, transformation and enrichment, storage, reporting, and visualization all needed to be continuous and real time.

> "If your system hasn't been revamped in the past two years, it probably needs to be re-architected now."

To facilitate that goal, our client ripped out and replaced almost everything it had in place, implementing newer technologies capable of handling and distributing unprecedented data flows. But afterward, the client enjoyed a newfound capacity to handle, manage and act on its IoT data.

Data volumes and speeds are growing exponentially just as computer-processing capabilities are rapidly expanding. The old approaches just can't compete. With that in mind, my advice is that if your system hasn't been revamped in the past two years, it probably needs to be re-architected now. Things have changed that much that fast.

If you do consider building something new or simply decide to refactor what you already have, think hard about ways to keep your data moving, end to end, as a stream. If you accomplish that goal, get ready for a huge surge of power.

SEAN OWEN
Director, Data Science, Cloudera

Sean Owen is director of Data Science at Cloudera in London. Before Cloudera, he had founded Myrrix Ltd (now, the Oryx open source project) to commercialize large-scale real-time recommender systems on Apache Hadoop. He is an Apache Spark committer and co-wrote *Advanced Analytics on Spark*. Sean was a committer and vice president for Apache Mahout and co-author of *Mahout in Action*. Previously, he was a senior engineer at Google.

Twitter | Website

For me, Peter Norvig's "The Unreasonable Effectiveness of Data" was the turning point.

Norvig is the director of research for Google, and his 2009 essay made a convincing argument that using sophisticated models to analyze small data sets is ineffective compared to applying simple algorithms to increasingly large data sets. Big data, in other words, wins.

That was a rallying cry for a new line of advance. Instead of thinking about how to marginally improve algorithms, why not scale up simple approaches and make them applicable to big data sets distributed over multiple nodes?

KEY LESSONS

1. **BIG DATA TRUMPS SMALL DATA, EVEN WHEN RUN THROUGH SIMPLE ALGORITHMS.**
2. **KNOW BEFORE YOU BUY WHAT *SUPERFAST* REALLY MEANS FOR YOUR BUSINESS.**

That insight got me interested in Apache Hadoop, a Java-based open source software framework for distributed storage and processing, and Apache Mahout, a project for building an environment for creating scalable machine-learning applications quickly. Roughly speaking, those have been the twin themes of my work ever since.

To illustrate what it all means, let's look at recommender engines.

> "Using sophisticated models to analyze small data sets is ineffective compared to applying simple algorithms to increasingly large data sets."

In early iterations, Amazon's recommender engine was based on little more than purchase history: you bought this, so you might be interested in that. But purchase history is only one data point and not a particularly useful one. As a result, the engine's predictions were sometimes wildly off the mark.

Over time, storage and memory became less expensive, so it became easier to aggregate and analyze much more data—every click on an Amazon customer's browser, for example. Many more bits of information were fed into a recommender engine. That engine learned and became much smarter with every click.

> "Caching data in high-speed flash storage or memory is simply a good thing to do."

This is one example of superfast in-memory computing's disruptive influence. It wouldn't be possible without the increasingly abundant resources that allow us to access and exploit massive data sets. A few years ago, even if I knew how they might benefit me, it didn't matter. Big data sets were just too massive to exploit. Therefore, caching data in high-speed flash storage or memory is simply a good thing to do. It avoids the transfer of files across networks and keeps files off slower-spinning disks, making retrieval much faster. That improvement has qualitatively changed some technology and network design choices.

The human factor is important, too. Clerical grunt-work tasks that highly paid analysts had once performed can be done through brute computational force; tuning model parameters, and selecting model features. Now, we can just empirically let the machines decide. That translates into savings in human time and avoidance of human error. It makes the machine an extension of human talent rather than making analysts slaves to their machines.

If all this sounds exciting, it should. But a word of caution. You will hear a lot of market pitches about real-time business and business transformation that follow in the wake of a superfast, in-memory computing investment. The story there is real—more speed, more resources, and bigger scale can concretely change your business. But think strategically. Know before you buy what *superfast* really means for your business before you take the leap.

THERE IS NO FREE LUNCH

MARK HIMELSTEIN
Chief Performance Architect,
Splice Machine, Inc.

Mark Himelstein is the chief performance architect at Splice Machine, the premier scale out RDBMS. Previously, he was CTO and co-founder of Graphite Systems, Inc. Prior to joining Graphite, Mark was the CTO of Quantum Corp.; before that, he was vice-president of Solaris Engineering at Sun Microsystems, where he led several major revision releases of Solaris and spearheaded the development of the DTRACE, ZFS, and Zones features.

Twitter | Website

Two years ago, a malware attack on a major US retailer swept up about 40 million debit and credit cards—a massive security lapse that was entirely preventable. What is sad is that the retailer had all the data in its system that it needed to detect and prevent the breach: it simply did not analyze and act on those data.

The business world is moving toward querying structured and unstructured data simultaneously. This retailer's case is just an extreme example: it relied on security software that could not scale with the volume of unstructured events. The company never saw what was coming.

Sadly, this is not the only example. One of the world's biggest insurance companies told us that it detects a billion security events a day. How many does it hold onto for later analysis? Zero. It tosses all of them out at the end of each day, for similar reasons.

Many companies like these lack the advanced software and hardware infrastructure they need to analyze and report massive, near–real-time structured and unstructured queries simultaneously. So, here is my point: if you plan to ride the big data wave of industry disruption without having it capsize your business, you must do the hard work. Look beyond the current crop of off-the-shelf, big data analytics solutions for both the hardware and software employed.

> "The retailer had all the data in its system that it needed to detect and prevent the breach: it simply did not analyze and act on those data."

KEY LESSONS

1. **THE BUSINESS WORLD IS MOVING TOWARD QUERYING STRUCTURED AND UNSTRUCTURED DATA SIMULTANEOUSLY.**
2. **HARDWARE-AWARE, HOLISTIC SOLUTIONS ARE THE KEY TO TRANSFORMING BUSINESS.**

THERE IS NO FREE LUNCH

One problem with current-generation platforms like Apache Hadoop and Spark—which, incidentally, are excellent for certain jobs—is often they are not quite superfast enough. Hadoop, for instance, is written in Java—a language that does not currently take advantage of the latest high core count extreme NUMA multi-socket computers that represent the current class of enterprise-level processing microarchitecture such as Intel Ivy Bridge and Haswell. Traditional languages like C and C++ can take full advantage of today's advanced processors.

Still, everyone expects that by implementing Hadoop or Apache Spark on enough nodes he or she will handle big data. It rarely works that way. Customers tell me that they are not fulfilling their service level agreements: queries that must be done in less than one minute take an hour. Queries that need to be done in 10 seconds take five minutes. These days, that hurts.

“Everyone expects that by implementing Hadoop or Apache Spark on enough nodes he or she will handle big data. It rarely works that way.”

A hardware-aware, holistic solution is my answer to transforming the way business decisions and therefore money will be made in the future. This is not free or easy. Developing the right solution for your problem takes hard work.

It is no surprise, then, that my company is working on such solutions. But we are not alone. An entire set of companies is working to exploit the flexibility of next-generation software interfaces while respecting and taking advantage of the latest hardware. You would do well to research them all.

People are leaving both security and moneymaking opportunities on the table because their common-wisdom infrastructure choices cannot accommodate big data computing. Many don't even know it. They are trying to do rocket science with only a ruler.

Bottom line: there is no free lunch.

THE DOORS OF PERCEPTION

ADAM GIBSON

Co-founder & CTO, Skymind

Adam Gibson is the co-founder of Skymind and creator of the open source libraries Deeplearning4, a distributed deep-learning framework for the Java Virtual Machine (JVM), and ND4J, a scientific computing library for the JVM. He is the author of *Deep Learning: A Practitioner's Approach* (O'Reilly, forthcoming) and currently an advisor to the data science master's program at GalvanizeU. He studied computer science at Michigan Technological University and lives in the Bay Area, near San Francisco.

Twitter | Website | Blog

Deep learning is a form of machine perception that uses neural networks and huge data sets to teach computers how to solve problems in near–real time through classification and clustering.

OK, but what does that really mean? Say you have a big stack of family photos, a percentage of which have the names attached. A deep-learning framework will rapidly figure out which names go with all those faces. That's *classification*. Now, say you have masses of unlabeled data, like years' worth of raw newspaper articles or a mass collection of unidentified photos. The framework will identify which are the most similar. That's *clustering*.

KEY LESSONS

1. **DEEP LEARNING IS A FORM OF MACHINE PERCEPTION NOT UNLIKE HUMAN PERCEPTION.**
2. **DEEP LEARNING IS A SIMPLE, POWERFUL IDEA THAT CAN INCREASE PROFITS, CUT COSTS, AND HELP CREATE NEW PRODUCTS.**

It is not unlike human perception. Sensory data come in; the brain adds layers of meaning to them, and then decides that the haze of photons in front of me represents my mother. That, too, is an exercise in classifying.

Real-time data processing is the Holy Grail. After all, when we talk about making data go fast, we are really talking about our ability to respond to the world quickly. We are talking about disruption. Take a few random examples that we think will benefit hugely from deep learning:

- **Banking.** If you are a bank, it matters a lot whether you identify fraud now or a month from now. A month from now, you will have lost a lot of money.

> "If you are a bank, it matters a lot whether you identify fraud now or a month from now."

- **Security.** If you are a security firm or government agency monitoring an airport, it matters a whole lot if you identify a person of interest now or five hours from now—after that person is on a plane over the ocean.
- **E-commerce.** E-retailers want to serve the customer the right ad at the right time, preferably in the instant before a purchase decision. If they only figure out what the customer wanted to see after the customer has left their website, they fail. Getting the customer now is what counts, especially if the customer is using a mobile device with a geolocation app and is walking just outside the store.
- **Automotive.** If you're a carmaker, it matters a lot whether you can identify risks and respond immediately. For example, pedestrians and other vehicles have to be recognized in real time – not minutes later.

> "One of the smartest things you can do is engage the open-source community."

Deep learning can work in various ways. I can best describe our platform. In our case, deep learning basically serves as an analytics layer on top of Apache Spark and Hadoop. Hadoop is a kind of data-management platform, while Spark is a form of MapReduce. Those technologies help orchestrate distributed computing, which is basically synonymous with "fast." I see three ways in which that combination affects business:

- By increasing profits or cutting costs
- By serving as the basis of powerful new products
- By associating the company in the public's eye with innovation—something many businesses feel is important. You have to stay ahead of the curve.

Deep learning is a simple, powerful idea. The trick, if you're going in that direction, is to be smart about it. One of the smartest things you can do is engage the open-source community. Hire people who know how the community operates and who can contribute to deep learning's open-source development. That way, your company can gain credibility with the vitally important open-source community while making sure its needs are met as this powerful platform evolves.

THE FUTURE DEMANDS TECHNOLOGICAL MUSCLE

GEORGE GILBERT

Senior Analyst, Big Data and Analytics, Wikibon

George Gilbert is big data analyst for Wikibon/theCUBE. He was a big data analyst for Gigaom Research and has been profiled on the front page of *The Wall Street Journal* and published as a guest author in a major overview of cloud computing in *The Economist*. Previously, George was the lead enterprise software analyst for Credit Suisse First Boston, one of the top investment banks to the technology sector, and a product manager on Notes at Lotus Development. George received his B.A. degree in economics from Harvard University.

Twitter | Website | Blog

Historically, analytics and transaction processing have been separate. For example, in call center applications and e-commerce websites, tracking what was happening was very much separate from trying to analyze what should happen or what would be a better outcome. The performance limitations of databases meant you had to change between capturing transactions and analyzing the data. That was slow. When the systems were in operation, data had to be extracted from the transaction systems, transformed into something the analytic system could work with, and then actually loaded into the analytic system. This "pipeline" could introduce a delay anywhere from hours to days or even weeks.

If the analytic system needed new data to answer different questions, the delay was much worse. This took a full development cycle, because someone had to determine in advance which questions to ask. The developers then had to locate those data in the transaction system, transform them into a format that could be used for analysis, and load the resulting data set into analytics systems. The process took too long and was too brittle.

You could not easily change what you wanted to ask, and you could not get the information back when you needed it: the time of the transaction.

> "The next generation's business applications are going to look a lot more like the leading-edge consumer websites of the past five years—LinkedIn, Netflix, Amazon."

KEY LESSONS

1. **WHAT DISTINGUISHES LEADING-EDGE CONSUMER WEBSITES FROM TRADITIONAL APPLICATIONS IS THE INTEGRATION BETWEEN DATA CAPTURE AND ANALYTICS.**

2. **MAINSTREAM COMPANIES NEED A CUSTOMIZABLE PLATFORM TO TAKE ADVANTAGE OF SUPERFAST DATA COLLECTION AND ANALYSIS BECAUSE THEY DO NOT HAVE THE TECHNOLOGICAL MUSCLE TO BUILD IT IN HOUSE.**

THE FUTURE DEMANDS TECHNOLOGICAL MUSCLE

In contrast, the next generation's business applications are going to look a lot more like the leading-edge consumer websites of the past five years—LinkedIn, Netflix, Amazon. What distinguishes those sites from traditional business applications is that the analytics and the transaction data capture are integrated. In the case of Netflix, while you are browsing through the catalog one night, Netflix's engine is learning what type of movies you like. The next time you log on, Netflix acknowledges your preferences and either shows a different catalog or organizes the existing catalog differently.

At Wikibon, we call these new applications *systems of intelligence*—systems that learn from past behavior, and then anticipate and influence the consumer in a business-to-business-to-consumer–type application. Think of it like a banking application or a retailer: someone is interacting with the app, but that interaction could be coming through any channel or touch point. At the intersection of all these interactions and at any point in real time we want our app to realize, "This is what the consumer is likely to do or wants to do, and here's how we can best influence them."

One of the first enterprise systems to do that was Harrah's Casino, which implemented a sophisticated loyalty system. Customers were issued a loyalty card, and Harrah's collected data to learn more about customer preferences. If a particular customer had a bad night, Harrah's might offer free tickets to a show it knew the customer liked or provided complementary meals at a restaurant the customer frequented so that the customer would have a better experience and stay longer or stay loyal. For Harrah's, the outcome was better: greater loyalty and profitability related to that customer.

> "That type of integration between the transaction and the analytics takes new underlying technology."

Many organizations want to do the same thing with their business applications, but the Herculean effort that goes into building such websites is not affordable for the "great unwashed"—the Fortune 1,000. The platform technology has to get better so that mere mortals can build these sites.

Most mainstream companies' architectures are based on traditional scale-up Structured Query Language (SQL) databases. Leading-edge consumer websites such as Twitter, Facebook, and Google have much more advanced platform technology. So, we have to make our traditional SQL databases better at this sort of thing. That type of integration between the transaction and the analytics takes new underlying technology. Mainstream customers are going to want to see database providers deliver this as a platform on which applications that can be built and customized.

DREAM BIGGER

ABHISHEK MEHTA

CEO,
Tresata

Abhishek Mehta is the founder and CEO of Tresata, a predictive analytics software company redefining business by automating complex human processes. He has built Tresata into a leading analytics innovator with a vision to use data to help enrich life. His history is a combination of radical technology expertise and practical, in-the-trenches business leadership. He was an executive in residence at MIT Media Lab, managing director at Bank of America, and in client-facing leadership positions at Cognizant Technology Solutions and Arthur Andersen.

Twitter | Website

While attending my 6th consecutive Hadoop Summit earlier this year, looking around to get a read on the pace of innovation in the larger Apache Hadoop/big data space, I was surprised and a little bit disappointed.

Just as in past years, the 2015 conference focused on data infrastructure & transformation—finding new ways to develop clean, organized, formed, and relevant data for decision making. If this is the hot topic even after 6 years, I thought, we are in trouble. Data transformation should be a given, almost a birthright in the new world of Big Data. We need to dream bigger.

Here is my advice for moving the collective conversation to the next level:

- Look for areas to automate data management functions that prepare data for monetization. We call this concept the *data factory* and predict that these functions can and inevitably will all be automated.
- Data scientists & engineers must seek to automate complex business–human processes. Examples of these processes - antifraud and anti–money-laundering operations, which are 90 percent manual and 10 percent automated. They should be entirely automatic. The value for large-scale transformation across all enterprises lies in automating these processes.
- Look for ways to combine big data with real-time processing to solve massive business problems. If we do not leverage these capabilities, we will fail.

> “Data transformation should be a given, almost a birthright.”

KEY LESSONS

1. **THE VALUE FOR LARGE-SCALE TRANSFORMATION ACROSS ALL ENTERPRISES LIES IN AUTOMATING COMPLEX BUSINESS–HUMAN PROCESSES SUCH AS ANTIFRAUD AND ANTI–MONEY LAUNDERING.**
2. ***BATCH* AND *REAL TIME* SOON WILL BECOME OUTDATED TERMS: EVERYTHING WILL BE REAL TIME.**

DREAM BIGGER

My prediction is that we soon will stop using the computing terms *batch* and *real time*. They will be meaningless distinctions. Everything will be real time. After all, the holy trinity of enterprise software solutions—speed, quality, and low cost— unachievable for so long, is finally a reality today. That disrupts everything we know.

The global technology market has been estimated at $3 trillion a year. Roughly 80 percent of that—$2.4 billion—comes from enterprise software. In my opinion, all of it is up for grabs.

Technology is going through a Darwinian era: the enterprise technology space has been built on data stacks which are finally getting decimated. We stack everything—storage, databases, analytical tools, and virtualization tools. For the past 50 years, those data stacks have held all the value.

Stacks have a fundamental problem, however—something I call the "*data hop*" problem. Stacked data must be transported around each tier of the stack. But you cannot move big data. The fastest wires in the world are 10 gigabytes per second. Do the math: moving petabytes (1 Million Gigabytes) becomes herculean. If Moore's law keeps reducing hardware costs and doubling capacity/power, why then can we not improve software to process data as well as store, analyze, and organize data without moving them?

> "The ability to deliver products and services at the right time in the right place to the right customer instantly is the future."

The good news is – we finally can. The answer is Hadoop (and the larger ecosystem associated with it including Spark). We do a big disservice to big data by conceptualizing Hadoop as a storage platform. People do not see it for what it really is: a massively parallel computational engine.

I have called this Big Data era powered by Hadoop the start of the Second Industrial Revolution because data is now the core asset for every enterprise. The ability to deliver products and services at the right time in the right place to the right customer instantly is the future. The technology finally exists to make it happen.

We have to discover and deliver value for each and every customer – whether they are an individual, a small business or a large company. We would then have fundamentally improved how we live. So dream big!

Industry Examples

AN AUTOMOTIVE REVOLUTION

FABIAN WILCKENS

EMEA Solutions Architect, MapR

Fabian Wilckens is a solutions architect at MapR, where his focus is on helping companies build next-generation data platforms. Prior to joining MapR, Fabian worked as a strategic architect at VMware, implementing Cloud Computing and Data Analytics solutions. Previously, Fabian started his own e-learning company and was also a senior consultant for Unisys Corporation. Fabian holds a B.S. in Computer Science.

Twitter | Website | Blog

Everywhere you turn in the world today you can see the impact that real-time big data analytics is having on industry. One of the most-affected industries is the automotive industry.

Our relationships with automobiles have not fundamentally changed in 50 years. Cars have become much safer; they have much better fuel economy, and one-time luxury features have become standard equipment. Still, a car is just a car. Big data is beginning to make fundamental changes that affect the entire automotive ecosystem.

Every car today is becoming a connected device that generates enormous amounts of data. These data are available for remote monitoring and analysis. For instance:

- All cars now have one or more computers to control the transmission; antilock brakes; and engine operation based on sensors that measure air and engine temperature, air pressure, oxygen, and other inputs. Performance data from these systems can be monitored remotely.

KEY LESSONS

1. **EVERY CAR TODAY IS BECOMING A CONNECTED DEVICE THAT GENERATES ENORMOUS AMOUNTS OF DATA.**
2. **THIS INFORMATION CAN BE USED TO ALERT THE CAR OWNER OF AN IMPENDING SERVICE REQUIREMENT AND SCHEDULE A SERVICE CALL AT THE OWNER'S PREFERRED SERVICE CENTER.**

> "Big data is beginning to make fundamental changes that affect the entire automotive ecosystem."

- Some cars have hundreds of built-in sensors that provide data about the use and wear of important components. By applying predictive maintenance algorithms to these data, vehicles can anticipate mechanical problems before they occur. The vehicle can then use this information to alert the owner of an impending service requirement. The same data could also be used to notify and schedule a service call at the owner's preferred service center and trigger shipment of the failing part to that service center so that it is there when the car arrives for service. The manufacturer can use these same data to compare the actual service life of car components in real time to projections based on engineering models and perhaps make design adjustments based on that analysis.
- Many cars have connected navigation systems capable of providing navigation and optimal routing information based on real-time analysis of traffic patterns. These systems are already beginning to include commercial information, such as locations of preferred stores and restaurants; they are capable of supporting individualized special offers in real time from preferred establishments on a particular route.
- Some cities are experimenting with on-street parking sensors and systems that provide that information directly to navigation systems so that drivers can request and receive direction to the nearest available parking spot.

> "Real-time analysis of data from automobiles is changing the way automobile manufacturers, service centers, car owners, and cities engage with each other."

These are just a few examples of how real-time analysis of data from automobiles is changing the way automobile manufacturers, service centers, car owners, and cities engage with each other. This technology also opens the door to new car ownership and service package models. For instance, automobile manufacturers or dealers might one day offer a monthly subscription service that covers data packages, maintenance, and insurance. Rates might vary depending on individual driving habits as measured by the car itself.

All of this is possible, and I have not even touched on the next big big data thing for automobiles: self-driving cars.

PUBLIC USES OF PRIVATE DATA

RANDAL SCOTT KING

Managing Partner,
Brilliant Data, LLC

Randal Scott King is the managing partner of Brilliant Data, a global consultancy specializing in big data, analytics, and network architecture. During his 16-year career in IT, he has done work for such industry-leading clients as Sprint, Lowe's Home Improvement, Gulfstream Aerospace, and AT&T. Scott lives with his children on the outskirts of Atlanta, GA.

Twitter | Website | Blog

Big data analytics makes it possible to see everything differently based on real-time analysis of streaming data. It allows us to perform predictive analytics on historical data side-by-side with streaming analysis and machine learning. The question of the impact of big data is really the visible and not-so-visible ways it is changing our lives.

We already take for granted the custom ads that pop up on our Google search pages. Most people are not even aware of the real-time fraud protection that happens at the check-out every time they swipe a credit or debit card, nor do they think about the smart power grid automatically balancing its loads when they turn on an appliance in their house. These are all big data–driven functions that have become a daily part of our lives, to the point we don't even think about how they happen.

KEY LESSONS

1. **BIG DATA ANALYTICS IS PLAYING A GROWING ROLE IN SOCIAL SERVICES, TRANSIT AND LAW ENFORCEMENT.**
2. **THE USE OF BIG DATA COMES WITH THE UNRESOLVED ISSUE OF ETHICS, WHICH IS HANDLED DIFFERENTLY DEPENDING ON CULTURAL ATTITUDES.**

There are many examples where public-sector agencies are beginning to use big data, but whether some of these use cases benefit society is up for debate. For instance:

- Social services agencies are adopting big data analytics to identify outliers in their service allocation data that may suggest fraudulent activity.
- Municipal transit agencies have benefitted from analysis of their data, leading to improvements in efficiency that have resulted in better, faster service.

> "The question of the impact of big data is really the visible and not-so-visible ways it is changing our lives."

PUBLIC USES OF PRIVATE DATA

- Big data is also playing a growing role in law enforcement. For example, an application called *Predpol*, which stands for *predictive policing*, is being used in many communities across the United States. Predpol analyzes the place, time, and type of past crimes in a community, and then predicts where and when crimes are most likely to happen—before they actually occur. Police departments can use that information to adjust the frequency of their patrols in certain areas. Recently this has caused civil rights groups to speak out, claiming that human biases about minorities are finding their way into data analysis methods.
- In Russia, The Center for Research in Legitimacy and Political Protest is rumored to have developed an application call *Laplace's Demon* that monitors individual and organizational posts on Facebook and Vkontakte, a Russian social network. By analyzing sentiment and content, the tool can predict the place and time of "unauthorized" protests. It will soon begin to monitor Twitter, as well.

The last two points lead to a discussion of the biggest gap in big data: ethics. There are significant cultural differences in attitudes about the use of the vast quantities of data being captured and analyzed. What would be completely permissible in China might be frowned upon in the United States and outright forbidden in the European Union. Each region operates according to its own best practices, but the Internet knows no boundaries. The application of big data is growing as time passes, and both companies and government entities are operating in the grey space left by a lack of ethical standards.

“The application of big data is growing as time passes, and both companies and government entities are operating in the grey space left by a lack of ethical standards.”

MIKE KAVIS
VP/Principal Architect, Cloud Technology Partners

Mike Kavis has served in numerous technical roles, from CTO and chief architect to VP. He has more than 25 years of experience in software development and architecture and has been a pioneer in cloud computing, having led a team that built the world's first high-speed transaction network in Amazon's public cloud—a network that won the 2010 AWS Global Startup Challenge. Mike is an analyst and blogger at *Forbes* and *The Virtualization Practice* and is the author of Architecting the Cloud: Design Decisions for Cloud Computing Service Models (IaaS, PaaS, SaaS).

Twitter | Website | Blog

Not long ago, I communicated with a wind farm company that produced energy for the open market. That industry requires solid forecasting to predict levels of power generation over given periods. Accurate forecasting equals lower risk, which in turn equals greater profits. The company wanted to conduct preventative maintenance on its turbines to keep up capacity, so it began attaching sensors. That decision effectively made the turbines Internet of Things (IoT) devices, capable of detecting and recording among other things wind and weather patterns that affect energy output. The sensor data allowed the company to correct the turbine actuators and tilt the blades to precise angles, taking maximum advantage of available wind.

KEY LESSONS

1. **OPTIMIZED BUSINESS PROCESSES BRING MORE INTELLIGENCE UP FRONT, AND THAT MEANS THAT YOU CAN MAKE IMPROVEMENTS EARLIER.**
2. **THE TECHNOLOGY NOW MAKES ONCE-INTRACTABLE BUSINESS PROBLEMS READILY SOLVABLE AND CREATES BUSINESS OPPORTUNITIES THAT HAD NOT EXISTED BEFORE.**

Over time, the devices collected loads of small data from the sensors and pooled them into a data lake for analysis. Tactically, the company used that information to keep making miniscule equipment adjustments. Strategically, on the back end, it did the data mining, batching, and analysis to bring the bigger picture into focus: Why does this vendor's piece break down so often? When I see an electrical storm, what should I anticipate? Over time, the machines learned. So did the company, which has gotten out of the wind turbine business. Realizing that its true value proposition lay in the data it was collecting, the business is now a software provider.

> "The sensor data allowed the company to correct the turbine actuators and tilt the blades at precise angles, taking maximum advantage of available wind."

Connectivity has improved. The costs of sensors, software, and storage have plummeted. The cloud era means that an organization does not have to buy 50 petabytes of physical storage on day one; it can just consume storage as data trickle in and pay only for what it uses. What's more, the organization can realize lightning-fast cluster computing through innovations like Apache Hadoop and Spark. Application program interfaces liberate coders from the scourge of auto-scaling, so now the company does not have to worry if someone sends it 10 billion rows instead of 1 billion. It does not have to worry about more clusters popping up. All that is done automatically.

> "Everything is changing, and you have to move at the new speed of business or be crushed."

The final, possibly crucial element that has emerged and that is bringing the industrial IoT to life is machine learning. The IoT really works when technology is set loose to detect business patterns and instruct the business to exploit or adjust them. That is a lot better than having some poor soul run 5,000 queries in analytics and come up with little more than a hypothesis.

To me, therein lies the future. The more the machine learns, the faster you can do things like change the precise angles of your wind turbine blades, or offer personalized insurance discounts to safe drivers, or streamline your delivery truck service. We have reached an age where technology makes once-intractable business problems readily solvable. Industries are discovering that they can act on data in near–real time, providing new service levels that drive price down and improve customer service.
In other words, we can:

- **Optimize business processes.** We now get much more intelligence up front and can make process improvements earlier in the life cycle, resulting in better products or services at a lower cost.
- **Understand our data earlier in the life cycle.** That creates business opportunities that had not existed before. Just ask those wind farm guys.

The world is running at a speed we have never seen before. Everything is changing, and you have to move at the new speed of business or be crushed. Understanding data is the new gold.

THE REVOLUTION WILL BE VISUALIZED

GIORGIA LUPI
Design Director,
Accurat

Giorgia Lupi is an information designer. Her work in information visualization frequently crosses the divide between digital and print, exploring visual models and metaphors to represent dense and rich data-driven stories. She is co-founder and design director at Accurat, a data-driven research, design, and innovation firm based in Milan and New York.

Twitter | Website | Blog

My company works in information design, so we conceive and build custom visual analytics tools that can explore and analyze disparate data types at various scales. Recently, we worked with a multinational group to process geo-referenced financial data from multiple cities. To do that, we built a next-generation visual data–workflow integration engine using Apache Spark, Scala, and Apache Kafka combined with real-time geographic information system analysis.

One of our analysts, while actually working on something else, accidentally noticed something strange in a crucial district within one of our cities. Visually comparing current data from that locality to data from the same month the previous year and the same day the previous month in the same place, he spotted major unexplained revenue fluctuations. Something was wrong.

We immediately analyzed the anomaly. It turned out that a vendor contract, which had been placed on hold, was affecting our client's bottom line far more than anyone expected. Corrective measures were taken within hours.

KEY LESSONS

1. **DATA VISUALIZATION CAN HELP BUSINESSES CRUNCH MOUNDS OF DATA RAPIDLY AND TAKE FAST CORRECTIVE ACTION.**
2. **NEW TOOLS, POWERED BY VISUAL ANALYTICS, WILL SOON EMERGE, WITH CAPACITIES WE CAN SCARCELY IMAGINE.**

> "Visual analytics blends art and science to communicate meaning."

THE REVOLUTION WILL BE VISUALIZED

In a way, visual analytics blends art and science to communicate meaning within organizations or externally with their stakeholders. Our data visualization interfaces provide insights, guidance, and decision-making support to different audiences.

Visual analytics allows business users and analysts to explore and find value in big data. Users can instantly execute analytic correlations on billions of data rows, and then present results cogently in intuitively designed dashboards. They can identify new patterns, trends, and relationships in their data and quickly act on them.

Doing what I do, I believe that superfast data processing will soon lead to an explosion of visual analytics solutions that take advantage of unprecedented, real-time connections. Our tools monitor large-scale real-time processes. They can track incidents on a worldwide distributed computer network. They can monitor real-time credit card transactions across an entire nation. They can aggregate vast amounts of information with minimal delay. They are scalable and adaptable. Many new, innovative tools based on these technologies are certain to emerge.

> “A superfast, in-memory computer processing revolution is under way.”

It is a complex task to visualize millions of data points, network mappings, and cluster analyses, and we have been accustomed to long processing times. But I believe that will soon change. A superfast, in-memory computer processing revolution is under way. Because of that, I think we will eventually see the emergence of visual analytics tools, operating seamlessly in time-sensitive environments that at present we scarcely imagine.

Watch this space.

THE BIG DATA HAMMER

IAN HOWELLS
Chief Marketing Officer, Argyle Data

Dr. Ian Howells is a recognized thought leader on the transition to big data machine learning applications. He wrote the book *Fighting Future Fraud: A Strategy for Using Big Data, Machine Learning and Data Lakes to Fight Communications Fraud* and has written widely on the subject on fraudtechwire. Howells has spent his career in senior marketing roles, building companies based on disruptive technology from their early stages to IPO or acquisition.

Twitter | Website | Blog

Data volumes are orders of magnitude bigger than anything businesses ever witnessed in the old relational database days. Fortunately, the Apache Hadoop distributed storage and processing model is fundamentally better at managing these big data stacks—so good, in fact, that Hadoop is changing the game. I firmly believe that every enterprise application will eventually be rewritten in a data-driven way and operate in real time—at Hadoop scale.

My company performs real-time fraud analytics for major mobile communications providers like AT&T and Vodafone. Astonishingly, the mobile industry loses $46 billion per year to fraud—2 percent of its revenue. Its problems do not end there. If a fraud incident hits the press, the provider loses subscribers. It also finds itself dealing with a major brand problem.

KEY LESSONS

1. **MACHINE LEARNING MUST NOT BE OVERLOOKED. IT CAN DETECT ANOMALOUS PATTERNS NEVER BEFORE SEEN, INCLUDING NEW, UNKNOWN FRAUD ASSAULTS.**
2. **EVERY INDUSTRY SHOULD INVESTIGATE NATIVE HADOOP AND SPARK APPLICATIONS.**

Traditionally, mobile providers have coped by summarizing data into piles and batch-importing those piles into rules-based systems. Then, human analysts would sift though endless data to detect fraud. Mostly what they found were false positives and mental exhaustion.

Rules fail. By definition, rules define only what is already known. If a provider got hit with a new, unknown fraud assault, rules-based systems could never detect that. Criminals have plenty of time to re-stage attacks to elude any rules adjustments.

> “Rules fail. By definition, rules define what is already known.”

A data-driven approach is the answer. If I see someone placing high volumes of calls to the United States from abroad, for example, I should be able to detect whether that is a long-time enterprise customer who has acted that way for years or someone new trying to scam customers into placing calls or texts with premium services.

Today, we can ingest massive amounts of data into data lakes and machine-scan them for anomalies. Then, we can compare the findings against customer relationship management and billing data. Through machine learning, we can even detect anomalous patterns never before seen.

Compared to rules-based antifraud systems, the big data approach detects 280 percent more fraud incidents, with 25 times fewer false positives. We can even pick up clues that a crime ring is involved, something that was never before possible.

“Compared to rules-based antifraud systems, the big data approach detects 280 percent more fraud incidents, with 25 times fewer false positives.”

These lessons apply not only to the mobile industry, however. Every data-intensive business should investigate made-to-Hadoop and made-to–Apache Spark applications. Over time, prepackaged, data-intensive applications built on those platforms will emerge, in turn driving greater demand for the platforms themselves.

Machine learning must not be overlooked, either. Think of it this way: if you have 1,000 times more data to deal with, you would never meet that challenge by hiring 1,000 more human analysts, particularly if you are scanning for fraud. You would only find more false positives, and the analysts would drive themselves batty.

New methods are needed. Machine learning, we feel, is the only way.

RETAIL IS BECOMING AN ANALYTICS-DRIVEN BUSINESS

TAMAS SZIRTES

Director, Innovation and Technology, SOA People Nederland

Dr. Tamas Szirtes is director of Innovation and Technology at SOA People Nederland. He is an SAP Mentor, HANA Distinguished Engineer, speaker, and author. He has 17 years of consulting experience from international projects in major Fortune 500 companies. He is active in mobile and in-memory (big data) technology consulting. Tamas has a Ph.D. in knowledge management, a master's of science degree in business administration, and a master's of science degree in computational engineering.

Twitter | Website | Blog

By using an in-memory computing platform like SAP HANA with Apache Hadoop and other analytics tools for processing big data from many sources, including business data, it becomes possible to more intelligently run every aspect of the retail business. For example, analytics makes it possible to personalize interactions with customers in a store. You can provide them with alerts or shopping recommendations based on an analysis of their location in the store, their purchasing history, and any marketing campaigns the store may be running at that time.

Data-driven engagement also helps retailers deliver the best shopping experience possible for customers. Part of this ability involves creating a uniform online and in-store experience for customers, who can search for products online, order online, and pick them up in the store or order online and have a product shipped. Some mobile store apps not only locate products in the store but also direct a user to the product. It is even becoming possible for users to build shopping lists, and the app will provide them with an optimum route through the store to minimize the time it takes to find all the items on the list. Along the way, the app can provide recommendations and promotions.

KEY LESSONS

1. **USING NEW ANALYTICS TOOLS THAT PROCESS BIG DATA FROM MANY SOURCES MAKES IT POSSIBLE TO MORE INTELLIGENTLY RUN EVERY ASPECT OF THE RETAIL BUSINESS.**
2. **ARMED WITH A COMPLETE PICTURE OF CUSTOMERS AT THE GRANULAR LEVEL, RETAILERS ARE IN A STRONG POSITION TO NEGOTIATE THE BEST PRICES FROM THEIR WHOLESALE SUPPLIERS.**

> "Data-driven customer engagement helps retailers deliver the best shopping experience possible for customers."

RETAIL IS BECOMING AN ANALYTICS-DRIVEN BUSINESS

Customers can scan items with their phone to get pricing or other product information. This entire process is data driven in that it combines sales and marketing data with inventory information and personalized information that the customer provides in the form of product lists, shopping history, and real-time scanning inputs. In this way, analytics improves the shopping experience for customers while relying less on in-store staff to help customers find what they are looking for.

Analytics also enables retailers to optimize their internal operations. By analyzing customer engagement data alongside other kinds of information, such as social networking data, events, weather, and even pricing, it becomes possible to apply accurate forecasting models to optimize other retail processes, such as in-store staffing, returns handling, product pricing, and inventory management.

Many stores offer guaranteed lowest prices. Customers often compare competitor pricing on their smartphone while in the store to ensure that they are getting the lowest price. Retailers themselves analyze competitor pricing and use those data with real-time adjustments of promotional strategy and inventory management to optimize the margin on merchandise they carry.

A complete picture of customers at the granular level gives retailers great power in their negotiations with suppliers. They can now know with a high degree of certainty what customers will pay for a product and how much of it they will buy next week or next month. Armed with this knowledge, retailers are in a strong position to negotiate the best prices from their wholesale suppliers. Retail is becoming an end-to-end analytics-driven business.

“Retailers can apply accurate forecasting models to optimize retail processes such as in-store staffing, returns handling, product pricing, and inventory management.”

THE ROLE OF DATA IN GLOBAL TRADING

PETER LANGNER
Consultant,
Adventas Consulting

Since 2006, Peter Langner has been a consultant advising mainly trading and finance companies on SAP implementation and upgrades. He is an experienced project manager, business consultant, and developer. Previously, Peter consulted in banking, business process modelling, and design and implementation of processes with SAP. He later joined a well-known German retailer as a project manager tasked with changing the systems landscape in such a way that the business could be rolled out to other countries.

Twitter I Website

In a global trading business, traders buy product from a seller, and then sell the product to their customer—typically, a wholesaler. For example, one global trader purchases all the meat that McDonald's Germany uses. This trader purchases meat from suppliers in other parts of the world and sells it to its customers. The trader also manages shipping and other logistics related to getting that product to the customer. McDonald's Germany is just one ordinary customer for that trader.

Global traders are often family-run businesses that have been operating for generations. Many are large, but you do not hear much about them because they are silent; they focus on their business, which is trading large quantities of goods. Global traders tend to be slow to change their basic business processes. Some traders focus on one product or a few kinds of goods, and that is all they trade. That may be all they have ever traded for the last hundred years. Others are quicker to respond to new opportunities, sometimes deciding to take on a new product group on short notice. When those kinds of traders make a change, it can be challenging for the IT department. For example, a global trader dealing in textiles may suddenly decide that it is going to start trading cutting machines. This is an entirely different kind of product group with its own forecasting and planning needs, its own sales cycles, sourcing requirements, and distribution logistics.

KEY LESSONS

1. NEW ANALYTICS TOOLS WILL CERTAINLY PROVIDE A COMPETITIVE ADVANTAGE TO TRADERS THAT ADOPT THEM, BUT GLOBAL TRADERS ARE CAUTIOUS ABOUT ADOPTING NEW TECHNOLOGIES LIKE THESE.
2. RETAILERS ARE USING NEW ANALYTICS TOOLS TO PERSONALIZE THEIR CUSTOMER ENGAGEMENTS, WHICH HAS ALREADY HAD AN IMPACT ON THE WAY THEY BUY PRODUCT FROM THEIR WHOLESALE S UPPLIERS.

> "Trading is a goods-driven business; such businesses must acquire and deliver goods on time to meet customers' needs."

THE ROLE OF DATA IN GLOBAL TRADING

Trading is a goods-driven business; such businesses must acquire and deliver goods on time to meet customers' needs. To do that, traders must be highly knowledgeable about their markets. They depend on demand forecasts, but they also need to know what is in storage, what is in transit, what goods are on which ship, and where those ships are located. Sometimes, deals take place and a ship that is en route will be asked to unload cargo that was sold after it was loaded onto the ship.

Traders depend on tools like SAP's Global Trading Management solution to manage many of the buying and selling transactions, tracking, and costs associated with trading, such as freight, loading costs, storage, commissions, insurance, and fees. Traders must also pay attention to currency exchange rates and commodity futures. The data structures that support global trading are complex.

In-memory computing makes it possible to run trading management applications faster and generate reports more quickly, which is important to traders that must track so many variables to ensure profitability. Some of the new analytics tools will certainly provide a competitive advantage to traders that adopt them, but global traders tend to be cautious about adopting new technologies like these. Near–real-time reporting will enable global traders to more effectively support their customers, especially traders whose customers supply retailers. Retailers are using new analytics tools to personalize their customer engagements, and this has already had an impact on the way they buy product from their wholesale suppliers.

“The data structures that support global trading are complex.”

Research

THE AGE OF MORE OPEN DATA

JONATHAN SCHWABISH
Senior Research Associate,
Urban Institute

Jon Schwabish is an economist, writer, teacher, and creator of policy-relevant data visualizations. He has written on various aspects of how best to visualize data, including technical aspects of creation, design best practices, and how to communicate social science research in more accessible ways. He is considered a leading voice for clarity and accessibility in how researchers communicate their findings. He is currently writing a book with Columbia University Press on presentation design and techniques.

Twitter | Website | Blog

In my role conducting public policy research, I see how new tools capable of quickly querying big data sets are changing the way we think about research. This is especially true when we work with administrative data. For example, I once worked on an analysis involving Social Security data that contained more than a billion observations—a good-sized data set—and I was able to query all those data in different ways and learn some interesting things.

It is now possible, however, to query data quickly from multiple sources. For instance, it would be possible to create a massive data set made of Social Security data, with longitudinal earnings for people over time. You could also add Internal Revenue Service data that list dependents and begin to see cross-generational connections between children and earnings. You might then link to health data to see how changes in health affect earnings over time and whether intergenerational correlations exist. Tools that make it easier to analyze large data sets allow you to break down walls between organizational silos and discover new insights that you can find only by analyzing those data together.

KEY LESSONS

1. **NEW ANALYTICAL TECHNOLOGIES MAKE IT EASIER TO CORRELATE DIFFERENT BUSINESS DATA SETS FROM DIFFERENT SILOS IN THE ENTERPRISE, REVEALING MORE VALUABLE INSIGHTS.**
2. **NEW ANALYTICS TOOLS ARE MOVING US TO THE NEXT GENERATION OF OPEN DATA, WHICH INVOLVES RENDERING VAST QUANTITIES OF COLLECTED DATA INTO A FORM THAT ANYONE CAN ACCESS AND UNDERSTAND.**

> “Tools that make it easier to analyze large data sets allow you to break down walls between organizational silos and discover new insights.”

These analytical tools are also more readily available to more organizations and people, which means that more people can look at data with their own questions and perspectives. The result is a growing movement toward more open data. Open data can be publicly available data, such as some types of government data, but they can also be "organizationally" open. For example, data that are proprietary to a business might once have been available only to the operational unit within the business that collected them and the business intelligence experts who analyzed them, but they can now be made available to every unit within the organization. These new analytical technologies make it easier to share different business data sets from different silos within the organization, revealing more valuable insights.

The availability of open data is creating demand for the tools to analyze those data, and the tools are creating demand for more open data. Another evolutionary trend in the world of open data is the nature of the data themselves. In the past, a great deal of open data was in a form unavailable for analysis, such as documents and PDF files. Today, most data are machine readable but often not intelligible to people. New analytic tools and platforms are moving us to the next generation of open data, which involves quickly rendering vast quantities of collected data into a form that anyone can access and understand.

"The availability of open data is creating demand for the tools to analyze those data, and the tools are creating demand for more open data."

TRADITIONAL BUSINESSES ARE TURNING INTO DATA BROKERS

SCOTT GNAU

Chief Technology Officer, Hortonworks

Scott Gnau is responsible for the Hortonworks' global technology strategy, leading innovative product directions and providing expertise and leadership across the organization's research and development programs. He has spent his entire career in the data industry, most recently as president of Teradata Labs, where he ran research, development, mergers and acquisitions, and sales support activities related to Teradata's integrated data warehousing, big data analytics, and associated solutions. Scott holds a BSEE degree from Drexel University.

Website

I see data as the primary disruptor. For instance, the data warehouse and business intelligence industries as we know them today are really second-order results of the enterprise resource planning deployments from the late 1980s and 1990s that enabled businesses to digitize transactional and business process information. That became the foundation for analytics that we take for granted today, such as using data to reduce customer churn, optimize inventory, and streamline the supply chain. If a business is not doing these things today, it is not operating at scale.

The past few years have seen a dramatic increase in the amount of collected data, which in turn have spawned a whole new generation of tools needed to make sense of them. For instance, the "schema on read" approach to handling data in Apache Hadoop represents something of a paradigm shift from the more traditional "schema on write" approach. It is true that schema on read enables a more agile analytics process, but that in itself is not why it was invented. The fact is, schema on read is the only practical way to make timely sense out of high-volume data coming from a variety of sources, some of which may be business systems and others data sources outside the company.

KEY LESSONS

1. **THESE TOOLS MAKE IT POSSIBLE TO DERIVE MEANING FROM LARGE DATA SETS, AND THEY REALLY ENABLE US TO LOOK AT PROBLEMS DIFFERENTLY.**
2. **ONE COULD EXAMINE ALL MEDICAL RECORDS FOR BOTH HISTORICAL AND CURRENT MEDICAL INFORMATION ACROSS THE ENTIRE POPULATION AND USE A DATA-DISCOVERY APPROACH TO ACCELERATE TESTING HYPOTHESES.**

> "I see data as the primary disruptor."

TRADITIONAL BUSINESSES ARE TURNING INTO DATA BROKERS

These tools make it possible to derive meaning from large data sets, and they really enable us to look at problems differently. For instance, if every person in the United States had an electronic medical record that detailed information about their health and any treatments they might be receiving, it would become possible to apply the entire data set to exploring the efficacy of the treatments. Rather than setting up a clinical trial and running it for some period of time to derive a statistically meaningful result for the trial class, one could examine all medical records for both historical and current medical information across the entire population and use a data-discovery approach to accelerate testing hypotheses.

The value of big data in business is much greater than improvements in work process efficiencies, however. A company's business information and intellectual property are really becoming its most valuable asset, and this is turning businesses of all kinds into data brokers. For example, a company might create a product by bending and adding features to a piece of sheet metal. That is the product, but the real value is in information about who needs the product, when they will purchase it, how it functions, when it functions, when it will fail, and other data that increase its value and the customers' quality of experience. That information is as valuable as the product—maybe more so. Those data are the company's true competitive differentiation.

> "A company's business information and intellectual property are really becoming its most valuable asset, and this is turning businesses of all kinds into data brokers."

THE POWER OF DATA-DRIVEN SCIENCE

MICHAEL FRANKLIN

Chair, Computer Science Division, AMPLab, University of California, Berkeley

Michael Franklin has more than 30 years of experience in the database, data analytics, and data management fields as an academic and industrial researcher, teacher, lab director, faculty member, entrepreneur, and software developer. He is also the director of Berkeley's Algorithms, Machines, and People Laboratory (AMPLab), which is known for creating the popular open source big data systems Apache Spark, Mesos, GraphX, and MLlib—all parts of the Berkeley Data Analytics Stack.

Twitter | Website | Blog

Data have always played a central role in scientific research. Historically, good data were difficult to obtain, so scientific methods had been used to identify and generate meaningful data. For instance, an *experimental approach* to science involves devising a controlled experiment, usually with a hypothesis in mind, and then observing the results. The experimental data either prove or disprove the hypothesis. A *theoretical approach* to science involves developing mathematical explanations of phenomena, and then over time collecting data that support or contradict the formula. A *computational approach* to science involves developing computer models of complex phenomena, such as weather or models of social behavior, and then comparing predictions to actual outcomes. All of these methods depend on limited data sets.

KEY LESSONS

1. THE EXPLOSION OF DATA IS THE RESULT OF SO MUCH INFORMATION NOW BEING DIGITIZED.
2. EASY ACCESS TO DATA AND LOW-COST ANALYTICAL TOOLS HAVE BECOME THE BASIS FOR A NEW SCIENTIFIC PARADIGM: DATA-DRIVEN SCIENCE.

In recent years, however, there has been an explosion of data. So much information is now digitized, from the vast amounts of data that scientific equipment generates to the contents of libraries; the world's Internet activity; all the business and transactional data generated continuously; government data; and the data generated by sensors built into phones, cars, machines, and buildings. The amount of data is doubling every two years, and that rate of data accumulation is accelerating. With this huge growth in data comes the recent development of tools like Apache Spark that make it possible to analyze large data sets quickly at low cost. This change has become the basis for a new scientific paradigm: *data-driven science*.

> "The amount of data is doubling every two years, and that rate of data accumulation is accelerating."

Here are a few examples of how data-driven science is affecting scientific research:

- **Astronomy.** Space- and land-based observatories generate enormous amounts of data. By analyzing data in near–real time, it is possible to identify unusual events, and then coordinate the observations of different equipment optimized for different spectra to gain a more complete picture of the observed event. Correlating observational data from different telescopes is an important technique in verifying the discovery of new exoplanets.
- **Particle physics.** The Large Hadron Collider in Switzerland generates more than 40 terabytes of data every day. Those data must be analyzed to extract the results of particle collisions, and then the collision data are further analyzed to test various theories of particle physics.
- **Medical research.** High-speed big data analytics is making it possible to research the genetic basis for diseases such as cancer across large population samples. Analysis that would have been impossible only a few years ago because of the size of the data sets can now be conducted in hours.
- **Humanities.** Through the digitization of libraries and language analysis, it becomes possible to analyze changes in the use of words or ideas over long periods of time and correlate those changes to culture movements.

Data-driven science makes it possible to view the world as a living laboratory that can be observed in real time. In this way, the cycle of hypothesis and testing happens much faster, which means learning happens faster, too.

“Data-driven science makes it possible to view the world as a living laboratory that can be observed in real time.”

ALLEN DAY

Chief Scientist,
MapR Technologies

Allen Day is chief scientist at MapR. His primary career objective is to improve the quality of human life by innovating at the intersection of genetics, computer science, mathematics, and IT. Allen is inspired by the natural world, where the most advanced designs can usually be found as algorithms encoded in DNA that run as a massively parallel network of chemical reactions.

Twitter | Website

Computers are all about simulations and testing what-if scenarios. We can create realistic three-dimensional simulations in physics engines because we understand the underlying theory—the *laws*—of physics.

We do not have that advantage in biology, which has no underlying theoretical basis. Today's biology is a descriptive science and this limits our ability to test hypotheses and get to the roots of many medical problems that continue to baffle us, particularly at the molecular level. We can, however, collect massive volumes of high-density biological data, especially from sensors. With the arrival of superfast, in-memory computation, we begin to process and analyze those data.

KEY LESSONS

1. **HIGH-DENSITY SENSOR DATA ARE MOVING US CLOSER TO CRACKING THE "BLACK BOX" OF BIOLOGY.**
2. **BIG DATA SCIENCE IS GOING TO BE A MAJOR PART OF THIS DISRUPTION, BUT ULTIMATELY WE WILL NEED NEW COMPUTER ARCHITECTURES.**

I sometimes think of biology as a black box that we are trying to reverse-engineer. Without good, high-resolution descriptions of what goes into and comes out of the black box, we cannot hope to understand its inner workings. High-density sensor data are giving us a more complete picture of the black box, and that is helping us reverse-engineer it, but even if we succeed, the job will not be done. If we ever manage to crack the black box of biology at the level of neuroscience, DNA sequencing, and synthetic biology, we will want to move on to in-depth analysis, to what-if computer simulations.

We can do some of that today. It works, but it is slow even with advanced computing systems like the Genomics Analysis Toolkit (GATK) and Apache Hadoop.

> "I sometimes think of biology as a black box that we are trying to reverse-engineer."

I think something more interesting is afoot. At some point, I believe, we will be able to run what-if simulations not in computers, but *in vivo*—within living cells. Cells, after all, are essentially computers built out of organic molecules.

The fundamental limiting component to all this is DNA synthesis. Rather than just reading out DNA, we want to encode and print synthetic DNA strands. When you can do that, you can inject the code into a cell. Automatically, it will begin to execute because DNA is basically a computer program.

That is where we can complete the loop and begin to develop the missing theory. We can place sensors and collect more data about our new intercellular computer program. Then, we can analyze what it is doing, testing our hypotheses about what is happening in the cell. Over time, we may begin to understand biology's laws.

Obviously, big data science is going to be a large part of this disruption to biological science. Ultimately, I think it will not be enough, though. Very quickly, I believe, we will find that the current breed of superfast computation will not keep up with all the data coming from all our biological experimentation. We will need new computer architectures, something like quantum computers, to reach a tractable solution.

You may wonder what any of this has to do with your industry. I see a connection.

At some point, computers' ability to test and verify what-if simulations will outpace the rate and amount of data collection required to build a workable model. We're already seeing the result in applications of narrowly focused artificial intelligence. Industries will increasingly compete on the basis of computational simulations, and the winners will be those with the models of the real world that are most accurate.

“Industries will increasingly compete on the basis of computational simulations, and the winners will be those with the models of the real world that are most accurate.”

Marketing

DATA CAN TELL MANY STORIES—IF YOU KNOW HOW TO LISTEN

SHONALI BURKE
President and CEO,
Shonali Burke Consulting

Shonali Burke was named to *PRWeek*'s inaugural top "40 Under 40" list of United States–based PR professionals, is one of 25 women who rock social media, and is the 2015 recipient of AWC-DC's Matrix Award. As president and CEO of Shonali Burke Consulting, she uses measurable social PR to take business communications from corporate codswallop to community cool™. Shonali teaches at The Johns Hopkins and Rutgers Universities, is founder and publisher of the popular PR community blog Waxing UnLyrical, and creator and curator of the #measurePR hashtag and Twitter chat.

Twitter | Website | Blog

Application of big data analytics is beginning to disrupt the public relations (PR) industry. New analytics tools are an important part of this disruption, but it is also largely because of easy access to all those data. In PR, we focus on several categories of metrics. Output metrics measure messaging goals such as hits or impressions, but there a limits to what those metrics tell you. Outtake metrics measure what people are actually taking away from your messaging, and outcome metrics measure the ultimate outcome, such as a change in behavior. These data can tell so many stories through trends and patterns, but if you are not looking for them, you will not see them.

A simple example illustrates how data can change a PR strategy. I worked with a company that was rolling out a new online offering; its goal was to get people to sign up. The strategy consisted largely of a traditional outreach campaign with online ads, tracking URLs, and a limited reliance on social information.

KEY LESSONS

1. **FEW PUBLIC RELATIONS AGENCIES THINK OF USING DATA ANALYTICS TO REFINE THEIR STRATEGY.**
2. **DATA SHOW WHAT IS WORKING AND WHAT IS NOT, AND IT PROVIDES CLEAR INSIGHTS INTO WHAT STRATEGIES YOU SHOULD USE.**

> "New analytics tools are an important part of this disruption, but it is also largely because of easy access to all those data."

DATA CAN TELL MANY STORIES—IF YOU KNOW HOW TO LISTEN

Data showed us early on that the traditional outreach was not accomplishing its goal of moving traffic to the website. Realizing that there was no online discussion related to our customer, we decided to try Facebook and Twitter ads along with social engagement. We worked with our customer to start an online discussion of the topic. Data showed that practically all the traffic coming to the new website was coming from Twitter and the online initiatives. We decided to cut back on traditional outreach and focus almost entirely on online events that got people talking. The analytics showed that that is what would work—and it did, much to our client's delight.

> "The analytics showed what would work—and it did, much to the customer's delight."

Many PR agencies do not think of using data analytics to refine their strategy. Some big creative agencies are building data capabilities, but many large PR firms are not, partly due because big agencies find it difficult to invest the time to build analytics skills and use analytics in their process. When you work in an agency, you are a slave to your utilization and availability, and the larger the agency gets, the harder it must work to maximize utilization. These large PR organizations rely more heavily on where their bread-and-butter projects come from, which is traditional media, and that makes it more difficult to develop an analytics capability.

The industry is changing, however, as companies spend more on digital strategies and recognize how important analytical measurements are to seeing their effectiveness. Data show what is working and what is not, and they provide clear insights into what strategies you should use. Businesses must map that knowledge back to business objectives and the story they need their audience to react to amidst everything else happening around them.

FRANÇOIS GARILLOT
Big Data Engineer,
Swisscom

François Garillot worked on the Scala type system in 2006 and earned his Ph.D. from École Polytechnique in 2011. He has worked in online advertising and on interactive interfaces to the Scala compiler while nourishing a passion for data analytics in his spare time. In 2014, Apache Spark let him fulfill this passion as his main job. In November 2014, François because the first developer in the world to receive Databrick's Spark Certification.

Twitter | Website | Blog

There are two primary ways to advertise on the Internet. One is to analyze browsing behavior, often with other personal information, and based on that analysis pop up a product-specific advertisement in real time that is appropriate for that person. This technique, performance advertising, has become the most common form on the Internet.

Another approach to Internet advertising is *brand advertising*. In that process, the goal is to reach every possible person with a brand ad that tells a story—something difficult to do on the Internet because there are so many outlets where to reach someone. Attempting to saturate the Internet with an ad would be prohibitively expensive. However, new analytics models are making Internet brand advertising across large populations both possible and effective. Here is an example.

KEY LESSONS

1. **MODELING USING MACHINE LEARNING CAN ALLOW YOU TO IDENTIFY TARGET GROUPS MORE QUICKLY AND EASILY THAN EVER BEFORE.**
2. **DEFINING THESE SEGMENTS CAN RESULT IN MORE EFFECTIVE BRAND ADVERTISING STRATEGIES ACROSS LARGE POPULATIONS.**

In 2014, I worked with a team tasked with creating a brand advertising strategy that would work, say, for a coffee machine manufacturer. The Internet advertising company wanted to offer this branding strategy to customers in the United Kingdom. But because there are far more tea drinkers than coffee drinkers in Britain, it would be a waste of money to present coffee brand advertising to everyone. So, the challenge was how to identify coffee drinkers in a predominately tea-drinking population, and then figure out how to reach them with coffee machine brand advertising.

> "The challenge was how to identify coffee drinkers in a predominately tea-drinking population."

We devised an unsupervised clustering model that would analyze the whole UK population to define groups with shared browsing patterns. The model is unsupervised in that we use an algorithm to train it, with which it analyzes browsing data in a way that is completely guided by the data, with no input from our particular desires or biases. In this way, the model quickly found many population groupings, including some we did not expect to find. For instance, one group that popped up quickly consisted of singles. We had had no idea in advance that we would find this large cohort appear in mobile browsing data.

Another practical effect of this kind of modeling is the identification of more groups than we know what to do with. It is up to the company to decide whether a group we discovered is a suitable target for its brand advertising, but we can select certain groups based on criteria such as group size, location, or combinations of characteristics members of the group shared. It even becomes possible to segment groups. For example, in defining coffee drinkers, it's possible, based on data, to witness a group that is interested in coffee; but within that group, we could see sub-segments pop up within people who primarily like to drink coffee, such as those who are extreme coffee connoisseurs. The vanilla coffee drinkers will be more interested in elegant, easy-to-use machines that might use coffee packets. The coffee connoisseurs are more interested in the esthetics of coffee making and are likely to be interested in high-end espresso machines. The interesting part is that finding the segments as revealed in the data results in a more effective brand advertising strategy than hoping for those segments to appear.

“The result was a solution that changed brand advertising from advertising to everyone to advertising only to relevant — though previously unidentified — groups.”

It took us just five weeks to develop a solution scalable to a population of millions of people. We did it by using Scala as our development language combined with open source libraries such as Apache Spark and an interactive shell. This foundation enabled us to develop and test routines quickly for analyzing and comparing subsets of data, writing only a few lines of code at a time. The entire development process was fast, and the result was a solution that changed brand advertising from advertising to everyone to advertising only to relevant — though previously unidentified — groups.

ALLISON LLOYD
Editor in Chief and Conference Director, DOCUMENT Strategy Media & Forum

Allison Lloyd serves as the editor of *DOCUMENT Strategy Media*, a management publication for executives, directors, and managers involved in the core areas of communications, enterprise content management, and information management strategies. Building on her highly respected editorial, she also helped launch the DOCUMENT Strategy Forum, a prestigious management conference for high-level executives involved with corporate communications and information management. Regularly addressing C-suite–level decision makers and enterprise executives, she delivers thought leadership on strategic solutions for managing effective communications with consumers.

Twitter | Website | Blog

A leading Fortune 100 company I work with has built a highly engaged online community comprising both current and prospective customers, who provide feedback on the company's marketing collateral. The company does a lot of data collection and segmentation and a ton of testing—both qualitative and quantitative—through its online channel, which gives it focused marketing intelligence about both its offline and online marketing touchpoints. Data from the company's offline touchpoints, however, such as newspaper and TV ads, call centers, and direct marketing, are more difficult to quantify and use because they are not as trackable as online data.

This issue is largely a problem of big data. Executives tell me all the time that they are data rich and information poor. They collect a lot of data, but that data does not always translate to actionable information. A customer's buying pattern is complex, and the journey from awareness to conversion can be the result of a marketing effort's many channels, messages, and touchpoints. Therefore, choosing the right method of attribution when using advanced analytics and data modeling can power intelligent decision making on your marketing spend. It can help you choose how to prioritize engagement channels, allocating credit percentages to those channels that lead to eventual sales.

KEY LESSONS

1. **COLLECTING A LOT OF DATA DOES NOT TRANSLATE TO ACTIONABLE INFORMATION: YOU NEED TO BUILD PROCESSES TO RETRIEVE DATA THAT PROPELS DECISION-MAKING.**

2. **ATTRIBUTION IS REALLY ABOUT UNDERSTANDING THE MARKETING CHANNELS IN WHICH TO INVEST.**

> "Executives tell me all the time that they are data rich and information poor."

Marketers need to move away from last-click attribution to multi-dimensional attribution models that can provide better context. In other words, how can marketing touches along the buying journey eventually work together to trigger the desired action? Taking that holistic view requires a shift in mindset, a move from product sales to customer relationships.

To that end, it is essential that marketing organizations build an effective business case that clearly outlines return on investment, risk, and cost. It must build up its IT infrastructure and processes to classify and rapidly retrieve information that propels decision-making. It must make investments in data-collection strategies and analytics to create an optimal mix of channels and touchpoints and to gain real insight in the long term. To me, attribution is really about understanding the channels in which to invest. Which are the most interactive? How are those interactions propelling action?

Many companies struggle to develop effective data-collection methodologies that provide the granularity and segmentation necessary to show clearly whether messaging in a given channel is actually triggering the desired customer actions. Therefore, a good place to start is determining the data-collection methodology that best suits the organization's needs and stated goals. All data collected must be connected to actual decision-making.

“It is essential that marketing organizations build an effective business case that clearly outlines return on investment, risk, and cost.”

THE END OF DEMOGRAPHICS

KIRK BORNE
Principal Data Scientist,
Booz Allen Hamilton

Kirk Borne is a member of the NextGen Analytics and Data Science initiative within the Booz Allen Hamilton Strategic Innovation Group and an advisor for several other firms. Previously, he was professor of astrophysics and computational science at George Mason University, where he did research, taught, and advised students in the graduate and undergraduate Informatics and Data Science programs. Prior to that, he spent nearly 20 years supporting large scientific data systems at NASA.

Twitter | Website | Blog

I'm a white guy, over 50 years of age, living in a particular ZIP code with a lot of other people who answer to similar descriptions. A unified marketing campaign should reach all of us with equal effectiveness, and we should all be equally responsive. Right? Well, maybe not.

That plan breaks down as soon as you look past the first demographic layer. For instance, I'm a big college football fan. The guy next door loves indie rock concerts. A guy down the street loves football, too, but he's an NFL guy. His wife couldn't care less. She's a fan of "The Real Housewives of Atlanta." Suddenly, these people aren't so identical. That uniform marketing campaign is running into trouble.

Enter *divisive clustering*. You start with whole populations, and then separate groups into smaller and smaller clusters that share common characteristics. If you continue drilling all the way down to individuals, you will have erased the concept of demographics—the lifeblood of traditional marketing.

KEY LESSONS

1. **CLUSTERING POPULATIONS INTO SMALL GROUPS OF PERSONAS CAN LEAD TO HIGHLY PERSONALIZED MARKETING.**
2. **TO TRANSFORM A MARKETING ORGANIZATION INTO A DATA-DRIVEN ORGANIZATION, USE ALL THE DATA YOU COLLECT, TAKE ADVANTAGE OF PUBLICLY AVAILABLE DATA FOR CONTEXT, AND MOVE WELL BEYOND DEMOGRAPHICS.**

The problem? Some businesses have millions of customers. None can afford to conduct millions of personalized marketing campaigns, but they can cluster those individuals into groups of prototype "personas." For instance, "this person likes to buy sporting goods on the weekends."

> "Your company is already capturing all the data it needs to define personas."

THE END OF DEMOGRAPHICS

Say your company wants to target 100 million American adults. How many personas does the company need to define? My estimate for that scenario is roughly 17. Splitting each dimension of an individual's data into two distinct segments (Positive/Negative, True/False, or High/Low), we find that 2 to the 17th power equals about 100 million unique combinations in 17 dimensions. So, 17 campaigns (based on one unique choice in each of the 17 dimensions) times 2 (to distinguish both Positive and Negative segments) is all you need. That becomes doable but still highly personalized.

Your company is already capturing all the data it needs to define personas, by the way. Call it *mouse clicks, impressions, click-through rate*, all of the above and then some—whatever you like. You're tracking it. Most likely, however, you're not using it.

You can mine clicks and web histories, drilling into what individuals look at, what they're typing into your site's search box, which pages they view and in what sequence. All that is evidence you can leverage to develop a model of that person's interest. Then, you can make your pitch directly, based on the persona model.

> "Marketing is a forensic science: it should be based on evidence, not just the opinion of the highest-paid person in the room."

Marketing is a forensic science: it should be based on evidence, not just the opinion of the highest-paid person in the room, which often becomes the decisive metric. It's just like your favorite cop show. You can't ask the dead body to name the killer; you must collect and analyze the evidence.

Here is my advice for transforming a marketing organization into a data-driven organization:

- Use all the data you collect.
- Take advantage of open, publicly available data for context. Lots of cities, counties, and states are collecting and presenting publicly a rich mine of deep-dive demographics and census data that reach all the way down to the household level.
- Race recklessly to the end of demographics.

It's science—but it is not rocket science. We just need to use what's already in front of us.

ANALYTICS IS THE KEY TO GETTING STARTED

CHRIS CONREY
Director of Sales and Marketing, Expanded.io

Chris Conrey is a writer, speaker, and coach on the new world of sales. Beginning with the writing of the *Post Modern Sales Manifesto*, Chris began moving sales away from being a dirty word. He's brought that focus to other companies as a team member while also spending time helping small businesses build their sales process. Chris spends his non-sales time playing with his three daughters, reading, and watching his beloved Red Sox.

Twitter | Website | Blog

First, start with analytics and start small. That gives you an action item, a place to begin without being overwhelmed. Take one piece of information, one metric, and track everything in relation to it. You can determine whether it's wrong later, but you have to start somewhere.

When you find the one thing you know you can measure right now, track it relentlessly. When you have some data, you can then go through and figure out what the "whys" are. Look for patterns. Even if you can't explain why the pattern is what it is just yet, you'll get there. For example, say you always do a lot of business right around the second week of the month: What does that tell you about your customers?

You want to get to the point where you can start understanding which touch points are driving conversions. Does your ad placement here make more sense than another option? The best way to start proper attribution is to tag your assets so that your analytics platform will do the heavy lifting. UTM tagging is a simple process, and there are plenty of tools to help you build such tags. Building correct links allows you to analyze properly and accurately identify which sources in your strategy are working and which are ineffective.

> "You want to get to the point where you can start figuring out attribution—which touch points are driving conversions."

KEY LESSONS

1. **USING ANALYTICS TO DETERMINE ATTRIBUTIONS AND CONVERSIONS DOESN'T HAVE TO BE OVERWHELMING. START SMALL, WITH A SINGLE METRIC, AND BUILD FROM THERE.**

2. **ANALYTICS YOU SET AND FORGET ARE USELESS. GET INTO THE HABIT OF REVIEWING ANALYTICS REGULARLY TO TRULY SEE WHAT WORKS AND WHAT DOESN'T.**

ANALYTICS IS THE KEY TO GETTING STARTED

We had a client that was selling chocolates and had planned ramp-ups in known chocolate holidays—Valentine's Day, Easter, Christmas, Mother's Day. We were seeing ads driving sales at the beginning of the week. So, we split the ads up and tried to run a few a little heavier toward the front of the week. It worked: those ads created more conversions. That's the type of pattern for which you are looking.

Adopting a data-driven process requires a bit of change from the way you're used to doing things. You have to get comfortable with your analytics platform of choice, get to know more than just the stuff you can figure out in the first hour on your own. Really dig into it. Your data will be running on these analytics; your analytics program allows you to see conversions, and then grow from there.

Then, it's a matter of creating a routine and building a process that requires you to look at the data regularly. All the analytics in the world aren't going to do you any good if you set them and forget them.

A once-a-week report won't tell you what you need to know. If that's what you're using, then you're missing a lot of opportunities to learn more about your needs and behaviors. So, look at a better solution for your clients, and look *like* a better solution for your clients.

The only way you're going become a data-driven organization is to take the time to collect the data and look at it yourself. Noticing a pattern that leads to better conversions goes a long way toward moving your organization forward.

> "All the analytics in the world aren't going to do you any good if you set them and forget them."

THE VALUE OF CREATING A DATA-DRIVEN MARKETING ORGANIZATION

ANDREW C. SANDERSON
Managing Partner, Pawn Global Venture Capital Consultants

Andrew Sanderson is a millennial entrepreneur and international strategy executive focused on solutions to problems and innovative steps to reach quantitative team goals. A trained sales executive and leader whose business development skills are inline with todays socially connected business minds, his unique ability to bridge industries and connect like-minded executives along a common goal is what makes him indispensable.

Twitter | Website

To turn your marketing organization into a data-driven marketing organization, you must first understand the value of attribution as it relates to your business. *Attribution* is assigning or ascribing a value or category to something like a number or data point. Ask yourself this simple question: "What are the numbers that I'm currently tracking telling me right now, and why do I attribute it to that specific information?" If you can't answer that question, then you're not truly living in the data-driven marketplace. Most small and midsized businesses in the United States are using few metrics other than their own sales numbers to try to gain a better understanding of the market and industries they're involved in. For those who are looking to gain greater insight into their customer base and engage with them to establish those metrics worth tracking, I suggest starting out by using the social media platforms out there that allow the paid user to view data points.

KEY LESSONS

1. **EMBRACE SOCIAL MEDIA OUTLETS—EITHER FREE OR PAID—THAT PROVIDE FREE DATA POINTS THAT HELP YOU GATHER MORE DATA ABOUT YOUR CUSTOMERS.**
2. **LOOK AT THE INFLUENCERS IN YOUR AUDIENCE: ONE OF THEM MAY BE THE CEO OF YOUR COMPETITION.**

First, the free model: How are you using social media channels that don't charge for publication, fan connection, or customer interaction? What are you doing with those free channels that provide metrics and analytics that you can use? Then the premium model: If you are a LinkedIn member, are you a premium member? Do you pay to have access to those statistics and the data that allow to you analyze your current marketing?

> "For those who are looking to gain greater insight into their customer base . . . I suggest using the social media platforms out there that allow the paid user to view data points."

The prepaid model is pay-per-click advertising—the paid leads, if you will. If I search for your company name, what comes up at the top of the search results? Who else is there besides your company? When you know whom you want to reach and the message you want to send them, it becomes imperative that you then use those same platforms to purchase specific pay-per-click ads into the market you want.

Attribution of the data you gather and the assignment of values and categories you choose to put them in are an important issue in a data-driven environment. By assigning each data point or metric to a specific category, you are able to categorize like values or metrics into a more discernable chart. It requires analyzing the differences in performance among the various pipelines you have in place so that you can make adjustments in your strategy, shifting the budget to higher-performing channels if need be.

“By assigning each data point or metric to a specific category, you are able to categorize like values or metrics into a more discernable chart.”

Sometimes, you encounter surprises that make you reevaluate your approach to attribution from an entirely different angle. I once had an experience like that where an industry influencer was concerned. My company was, as a best practice, regularly looking at metrics on our Twitter performance, such as impressions and re-tweets, to make sure our strategy was as finely tuned as possible. Then, one day at a meeting, an executive told us that he had analyzed the interactions we were having on Twitter with his competitors and other executives in the field. He perceived us as a potential ally when he looked at the data we were gathering and the people we were collaborating with, and so he proposed a strategic alliance with us.

It’s difficult to calculate the value of a single LinkedIn update or Twitter tweet until you gather the market responses that reaffirm that your marketing is attracting new revenue through found business. The tricky part of the data, viewed in aggregate, is the anonymity that comes with it. Unless you’re carefully looking at the influencers within your audience, you might not know if one of the 1,000 people who saw your tweet was the CEO of your competition. That was the one impression it took to convert that person to a potential ally or partner, and it was powerful for us.

RUSS MERZ

Professor, Research Scientist, Analytics Consultant, Eastern Michigan University, Blab Predicts

Russ Merz is an experienced research scientist, analytics consultant, and professor. He has subject matter expertise in market research methods, as well as in the development and application of customer experience analytics to marketing management problems in the areas of advertising, public relations, branding, retailing, social media, and e-commerce.

Twitter | Website

Recently, I worked with a packaged goods company that had an active social media presence. Its goal was to figure out what effect its social media programs had in driving traffic to its website. To quantify this, we analyzed page views, time onsite, clicks, and other metrics with the aid of an analytics engine. We demonstrated positive, empirical linkages between the client's social media activity and changes in website patterns. We even categorized social media conversation types and showed which ones most strongly influenced web activity.

Linking social media to website behavior was a necessary building block for developing algorithms to predict consumer behavior. It's an example of applying attribution measurement methods that go well beyond classic last-click attribution.

KEY LESSONS

1. **ALGORITHMIC ATTRIBUTION MODELS ARE ONLY AS GOOD AS THE DATA INJECTED INTO THEM: ALGORITHMS NEED A VERIFIED FOUNDATION OF ACCURATE DATA.**
2. **APPLY THE FOUR MS—MAKE A STRATEGY, MOVE ON IT, MEASURE IT, MONETIZE IT.**

Algorithmic attribution models are only as good as the data injected into them: algorithms need a verified foundation of accurate data. Without the right model, it's difficult to establish advanced algorithmic attribution.

> "Without good models, it's virtually impossible to allocate resources in a way that will have any true impact on ROI."

To that end, my consulting team and I have developed what we call *exposure models*, which empirically assess the effect of various experiential influences on targeted outcomes, like customers' intent to purchase. Often, we rely on survey data linked to web metrics, but other data might work equally well.

Clearly, sound empirical data are crucial for understanding and shaping the marketing decisions that drive attribution. Without good models, it's virtually impossible to allocate resources in a way that will have any true impact on ROI. You have to understand what you're trying to accomplish.

I've got a heuristic that I use with clients, a framework I call *The Four Ms*. It goes like this: make a strategy, move on it, measure it, and monetize it. It's simple but effective.

Working through a framework like that helps you understand how your strategy and tactics might change your key performance indicators. With a framework in place, you can measure how those changes affect sales, profit, market share—whatever metrics are most important to you. The key to this approach is building an integrated measurement system that links all these elements.

In simplest terms, here's my advice:

- Create a framework.
- Quantify the various pieces of the framework.
- Look for empirical linkages between the pieces.

“With a framework in place, you can measure how those changes affect sales, profit, market share—whatever metrics are most important to you.”

FOUR STEPS TO BETTER DATA-DRIVEN MARKETING

GREG BONSIB

Director, Channel Marketing,
Zenith Products Corp.

Greg Bonsib is a B2C channel marketing expert. He has extensive experience working in senior marketing roles at Owens Corning and Newell Rubbermaid and is currently leading channel marketing at Zenith Products Corp. He specializes in selling consumer products through mass retailers like Wal-Mart, Target, Home Depot, Lowe's, and Amazon. Greg also publishes an industry-leading blog on channel marketing.

Twitter | Website | Blog

Data has always played an important role in marketing. What has changed in recent years is the sheer volume of data now available from so many different sources. If we were once challenged to see the forest through the trees, we're now challenged to see the forest through the leaves.

Here at Zenith, we primary sell our products through our network of retail partners. We do have a product website, but it is not a major source of our sales. In my role at Zenith Products Corp. I look at data strategically, with the essential premise that the customer is the ultimate focus of interest. With that in mind, I have a four-step approach to using data in customer marketing:

1. **Start small.** For example, we often use Amazon Marketing Services (AMS) ads to test the effectiveness of promotional strategies. We will focus on a product or a phrase or an idea and watch how it performs. By doing so, we can create a test with a focused objective, and we can keep the cost of our tests very low. Amazon Webstore provides an excellent data environment in which to try something small at low cost and capture meaningful performance data around it.

> "I look at data strategically, with the essential premise that the customer is the ultimate focus of interest."

KEY LESSONS

1. **WHEN TESTING IDEAS, IT IS IMPORTANT TO CREATE A TEST THAT HAS A FOCUSED OBJECTIVE AND TO KEEP THE COST OF TESTS LOW.**
2. **THE POWER OF DATA-DRIVEN MARKETING IS THAT WHEN YOU HAVE DATA THAT PROVE YOUR STRATEGY, YOU CAN GAIN CREDIBILITY AND RESOURCES TO DO IT AGAIN.**

2. **Test and learn.** By keeping the test small and the cost low, we can try many different variations. For instance, I used AMS to test category ads, product ads, brand ads, ads based on search words, ads based on competitors, even ad placement that AMS recommended. The good thing about this method is that Amazon provides data that allow us to track ad performance directly to Amazon Webstore sales. In this way, we can identify ad strategies that perform well and those that are total duds. Failures are as important as successes. The best-performing ads returned $5 in sales for every $1 spent on marketing—a 500 percent return on investment (ROI) on marketing spend.

3. **Build on success.** Using the proof point of a 500 percent marketing ROI, I can then build that ad concept into a larger marketing strategy that involves product packaging, in-store promotions, and many other things. Then, I start looking at other kinds of performance data. For instance, if a store is running a promotion on a product, I pay close attention to the halo effect of that product on the sales of other products.

4. **Document your knowledge.** It is incredibly important to document what you do. Documentation enables you to build a body of knowledge about what works and also to share that knowledge through your organization.

The power of this kind of data-driven marketing is that when you have data that prove your strategy, you can gain credibility and resources to do it again.

“In this way we can identify ad strategies that perform well and those that are total duds. Failures are as important as successes.”

Mighty Guides

Mighty Guides make you stronger.

These authoritative and diverse guides provide a full view of a topic. They help you explore, compare, and contrast a variety of viewpoints so that you can determine what will work best for you. Reading a Mighty Guide is kind of like having your own team of experts. Each heartfelt and sincere piece of advice in this guide sits right next to the contributor's name, biography, and links so that you can learn more about their work. This background information gives you the proper context for each expert's independent perspective.

Credible advice from top experts helps you make strong decisions. Strong decisions make you mighty.