Data *Disruption*

Research

Mighty Guides

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.

© 2015 Mighty Guides, Inc. I 62 Nassau Drive I Great Neck, NY 11021 I 516 360 2622 I www.mightyguides.com

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 I Website I Blog

Download the full e-book: Data Disruption

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.”

THE AGE OF MORE OPEN DATA

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

Download the full e-book: Data Disruption

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.”

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 I Website I Blog

Download the full e-book: Data Disruption

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.

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.

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.

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

THE POWER OF DATA-DRIVEN SCIENCE

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.*

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.”

WHAT IF

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

Download the full e-book: Data Disruption

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.

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.

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.”

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 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.

> “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.”

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.

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.

© 2015 Mighty Guides, Inc. I 62 Nassau Drive I Great Neck, NY 11021 I 516.360.2622
www.mightyguides.com