



The 5 Myths of Advanced Analytics

“Over 60% of models developed with the intention of operationalizing, were never actually operationalized.”

—Gartner Inc.

Data science is a multi-disciplinary approach to finding, extracting, and surfacing patterns in data through a fusion of analytical methods, domain expertise, and technology. Machine learning (ML) is currently the most popular area within data science. To monetize and realize the value, organizations need to infuse analytics into business processes and operations. Data science and machine learning is a key strategy for enterprises that want to take advantage of the massive insights buried in their data marts, data warehouses, Apache Hadoop lakes, and spreadsheets. But, despite the millions of dollars invested in analytics technologies, the majority of companies still struggle to establish an efficient and programmatic way to do analytics at scale. According to Gartner Inc., over 60% of models developed with the intention of operationalizing them were never actually operationalized.¹ Why are these investments failing to meet expectations? In this paper, we delve into today’s most common data science and ML myths and offer potential solutions.

¹ Idoine, Carlie, Peter Krensky, Erick Bretenoux, and Alexander Linden. Magic Quadrant for Data Science and Machine Learning Platforms. Gartner, Inc. January 28, 2019.

MYTH #1: DATA SCIENTISTS ARE DATA JANITORS

A typical enterprise has 6–10 databases and thousands of employees with great ideas on how to leverage their data. However, at least 70% of this information is siloed and difficult for data science teams to discover and work with. It's been an accepted requirement that 80% of a data scientist's work is spent on data acquisition, movement, and transformation. These are the most tedious steps that a data science team must go through in order to start extracting value from their organization's data.

The issue with “data janitor” work is that it's an extremely linear and poorly integrated process. Data teams traditionally rely on data engineers to extract data. The extracted data is then moved into a new location, at which point data scientists use their tools and methods to clean and format the data in preparation for analysis. Finally, the data scientists present the results to business analysts who use yet another set of tools to transform, blend, and analyze the outputs. This lack of integration between people, processes, and tools results in a terribly inefficient approach to analytics, requiring constant back and forth communication, delays, and errors.

Modern approaches have transformed data logistics, transformation, and enrichment into a highly collaborative and integrated process. New solutions enable data engineers, data scientists, and citizen data scientists to engage early in the process so that each role can iteratively determine what data needs to be cleaned and analyzed, as well as which role along the analytics continuum is most appropriate for each task.

TRUTH #1

Data scientists don't have to be isolated data janitors. They are constantly forced to do janitor work today because legacy technologies require them to sample data and package it into formats that are accessible by the subsequent groups in the traditional data pipeline. Modern approaches to data science include all roles (engineers, scientists, and analysts) earlier in the data pipeline, each participating in the process of cleaning, formatting, and enriching data. With various roles collaborating on the most time-consuming parts of the data analytics process, like data blending and processing, businesses can ramp up the productivity of their analytics teams.

MYTH #2: AUTO ML WILL REPLACE DATA SCIENTISTS

Data scientists have been at the center of a controversial debate over the past few years. First, Google's chief economist, Hal Varian, called the job the “sexiest job on earth” in a New York Times interview. Second, McKinsey released a report predicting that the US would face a shortage of 140,000 to 190,000 people with deep analytical skills.

Soon enough, software vendors leveraged this industry buzz to turn against data scientists and declared that their tools were so simple to use that “data scientists would not be needed anymore.” The fate of data scientists is not that simple, however. Data scientists are the foundation of the predictive enterprise, and companies that want to win with data shouldn't compromise their data science values.

“Data scientists spend up to 80% of their time mired in the mundane labor of collecting and preparing unruly data, before it can be explored for useful nuggets.”

— New York Times

Although industry studies have shown that true data scientists are rare, no one ever suggested that organizations should use data scientists alone as their analytical backbones. For each data scientist, there are another 10 or 100 citizen data scientists and business analysts involved in the end-to-end ML process. And for each business analyst team, there is another set of executives driven to use data to make better decisions.

“Auto ML will augment the work of data scientists, not replace it.”²

TRUTH #2

Data scientists are not going away. They are the nucleus of the predictive enterprise and should be supported by tools, technologies, and practices that help them scale their knowledge and talent. Over the next 10 years, we predict that data scientists will be supported by technological innovations that augment their expertise as well as the natural evolution of our data-driven culture. Here are some interesting trends:

MORE DATA-NATIVES

Data science is now available in more schools for undergraduates and soon will be made broadly available to younger generations. Similar to physics, data science is built on math, data interpretation, and experimentation.

MORE DECISIONS VIA AUTOMATION

In his book, *Think Twice*, Michael J. Mauboussin makes a great case for the compartmentalization of decision-making. Some decision tasks are better operated by humans. Some are better automated via machines. As humanity surrounds itself with more intelligent devices, we will grow more comfortable with empowering machines to make some decisions on our behalf.

MORE PEOPLE, MORE DECISIONS

Gartner predicts that the number of data and analytics experts in business units will grow at three times the rate of experts in IT departments, which will force companies to rethink their organizational models and skill sets.³

Predicting the “death of the data scientist” is misguided. They are important today, and they will be important tomorrow.

MYTH #3: MOVING DATA IS A REQUIREMENT FOR MACHINE LEARNING

Historically, companies have provisioned data into siloed environments that were separated from their central analytics environments. This practice created multiple copies of the same information across the organization. The proliferation of new data islands brought delays and errors, amplifying the already existing challenge of working with big data.

It also forced teams to work with sample datasets because the entirety of the data they wanted to analyze, although available, was often too large to move into their analytics environment. Leveraging subsets of data ended up limiting companies’ perspectives and the utility of their insights.

The introduction of new technological paradigms allows teams to work with data at scale, without ever having to move it or compromise the depth of insight because of data size.

² <https://www.tibco.com/blog/2019/03/28/love-it-or-hate-it-auto-ml-is-here-to-stay/>

³ Gartner. Maximize the Value of Your Data Science Efforts by Empowering Citizen Data Scientists. June 12, 2018.

TRUTH #3

Moving data into siloed analytics environments is a thing of the past. It introduces delays, errors, and potential security issues in the data pipeline. The companies that succeed with big data have created an open and secure environment that allows users to send instructions to the data where it sits. A key technological innovation in this field is “in-cluster processing.” Instead of pulling the data into separate analytics environments, in-cluster processing allows for instructions to be pushed down into databases or Hadoop clusters. This capability allows data analysts and scientists to run algorithms at scale without having to worry about moving data or optimizing their work based on the data source that their data sits in.

“70 to 80% of the world’s data is trapped in silos with no secure and reliable way to access it.”

MYTH #4:**YOU CAN’T SHARE & PROTECT DATA AT THE SAME TIME**

Data is arguably the most valuable asset that the modern enterprise has, regardless of the industry. Data must be guarded and treated with the highest level of security. Typically, database administrators and big data leads are charged with the difficult tasks of making data accessible within their enterprise, while making sure this access adheres to their company’s governance and security policies.

But in order to be successful, analytics teams need fast and easy access to their organization’s data. Provisioning data to analytics teams has become a daunting and time-consuming task, resulting in constant back-and-forth communication for data access. The consequence is a reduction in the organization’s speed of discovery and execution.

TRUTH #4

Analytic leaders have learned to build processes and deploy tools that allow role-based access and enable fluid communication between data engineers, data scientists, analysts, and line-of-business leaders. Business employees should be able to search and explore datasets immediately. Items such as data, people, models, and projects should be at the fingertips of each employee to promote creativity and optimize business practices. This role-based access respects the security and privacy of the data according to the organization’s rules and regulations.

MYTH #5:**DATA SCIENCE AND ML IS JUST ABOUT THE ALGORITHMS**

When investing in data science and ML solutions, organizations are primarily focused on the quality and complexity of their algorithms. While this is certainly an important and necessary step, neglecting the other aspects of creating a data-driven culture has led to poor insight adoption.

The ways in which data is consumed and implemented across the enterprise often dictates the success or failure of an analytics initiative. Here are three key questions that every advanced analytics project leader should address when planning analytics initiatives:

- What is the business decision to be made?
- How will your business process change as a result of predictive analytics?
- How and where will the analytics be deployed, monitored, refreshed, and maintained?

“Platforms must support not only model building but also model operationalization: The full benefit — including business value — of data science and ML will not be achieved unless models are both: 1. Embedded in business processes and 2. Maintained, monitored and managed over time.”

—Gartner MQ for Data Science and Machine Learning

TRUTH #5

While beginning data science and ML projects, analytics leaders need to shift their organization's focus towards communication, collaboration, and cross-functional teams. Analytical insights need to be infused within business processes and made transparent throughout each area of an organization, from the most senior executive to the most junior associate. Moreover, analytical insights should be embedded into tools through a visual and collaborative interface, so they can be digested by decision-making business executives.

The companies that drive sustainable change understand that while they must take advantage of data by ingesting large amounts of it, they must use advanced analytics to filter the noise and surface only the insights that are relevant at decision time. Bottom line: it's not just about the algorithm. It's about getting the power of the algorithm into the hands of more people, at the point of decision.



Global Headquarters
3307 Hillview Avenue
Palo Alto, CA 94304
+1 650-846-1000 TEL
+1 800-420-8450
+1 650-846-1005 FAX
www.tibco.com

TIBCO fuels digital business by enabling better decisions and faster, smarter actions through the TIBCO Connected Intelligence Cloud. From APIs and systems to devices and people, we interconnect everything, capture data in real time wherever it is, and augment the intelligence of your business through analytical insights. Thousands of customers around the globe rely on us to build compelling experiences, energize operations, and propel innovation. Learn how TIBCO makes digital smarter at www.tibco.com.

©2018-2019, TIBCO Software Inc. All rights reserved. TIBCO and the TIBCO logo are trademarks or registered trademarks of TIBCO Software Inc. or its subsidiaries in the United States and/or other countries. Apache and Hadoop are trademarks of The Apache Software Foundation in the United States and/or other countries. All other product and company names and marks in this document are the property of their respective owners and mentioned for identification purposes only.
05/24/19