10 MYTHS ABOUT DATA SCIENCE

And Why They Shouldn't Hold You Back

Daniel Jebaraj
Vice President, Syncfusion, Inc.

Syncfusion®

Introduction

Data science is now being used as a competitive weapon. As with other technologies and processes that can transform the way companies operate, there's a lot of contradictory information about it that's causing considerable confusion.

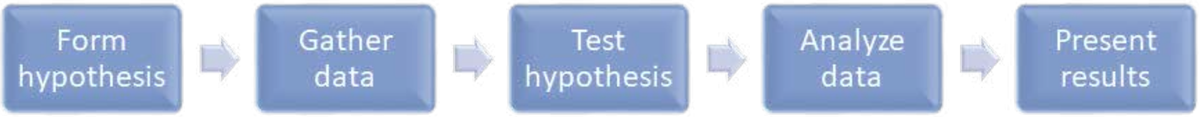
Most of today's business leaders have heard that data science can improve operational efficiency and customer relationships, but it isn't always clear how data science should be implemented or what the specific business benefits might be.

This white paper addresses some of the misunderstandings individuals and organizations have about data science. It also includes tips developers can use to enable data science capabilities in their organizations.

What Is Data Science?

Data science is an umbrella term comprising some of the today's hottest topics such as machine learning, analytics, modeling, and data visualization. In practice, data science is a process. It starts with a hypothesis, and then data is gathered with the hope of producing valuable insights. Once the data is collected, it is used to test the hypothesis and build models. Finally, the results are analyzed and presented to decision makers as reports or dashboards.

The models, which tend to approximate events or behaviors in the real world, are used to make important decisions. For instance, churn detection models are often used to predict which customers are at the highest risk of defecting to a competitor so the business can take preventative action. Depending on the circumstances, the preventative action may take the form of a phone call from a manager, a discounted subscription renewal rate, or coupon.



Unfortunately, there is no single definition of data science, but many data scientists and vendors describe it as a process, similar to the definition and workflow presented above. Some consider data science synonymous with statistical modeling or analytics (identifying patterns in data and presenting the results via a dashboard), which only adds to the confusion. Modeling and analytics are subsets of the data science process.

The good news is that businesses can choose how they implement data science in their organizations because there's no "right" way to do it. How data science is implemented depends on many things, including the expertise, tools, and data available to the organization. The most effective implementations of data science tend to start with, and align with, business goals.

Seasoned data scientists understand such nuances. Such understanding promotes clarity. Unfortunately, there are many myths around data science that serve as roadblocks in the path toward clarity. By confronting these myths, it is our hope that more organizations, especially organizations with development teams, will implement data science.

Myth #1: It's Hard to Find Data Scientists

The shortage of data scientists is well documented in the media. In fact, Fast Company and others cited a report from McKinsey that predicts a shortfall of 250,000 data scientists in the U.S. alone by 2024. Many of today's companies are competing for "real" data scientists or "unicorns." Unicorns are rare creatures who have a graduate degree in math or statistics (Ph.D preferred), strong programming skills, and solid domain expertise. Few candidates have deep expertise in all three areas, which is why there is a shortage of data scientists. To overcome that obstacle, some organizations are trying to develop a data-science practice that combines the expertise of several people.

A common mistake is to hire specialized expertise, such as a Ph.D.-level statistician or data scientist, before it's necessary. Company decision makers believe the company needs such a person to gain a competitive advantage, but it is unclear what that person should do and for whom. Lacking a mission and purpose, the statistician or data scientist who longs to make a positive impact on the business but can't will likely resign with a better offer in hand from another employer. That's why it's often easier to hire specialized talent than to keep it.

Most organizations can start reaping the benefits of data science without highly specialized expertise or expensive software, but quite often they don't know where to start. We recommend looking inwards and starting with software development teams. It is our experience that software development teams can be trained to take on data science tasks.

Myth #2: Data Science Is Suited Only for Large Organizations

Large organizations typically have the financial resources necessary to build a formal data science practice. However, that does not mean their data science practice will succeed.

When those large organizations are successful, the media likes to use them as examples of what companies can achieve, such as competing more effectively, improving operational efficiency, and even disrupting an entire industry. Because large, brand-name companies are often positioned as the leaders of their industries, small and medium businesses (SMBs) may believe that data science requires hefty investments in expensive software and the expertise needed to use that software.

In fact, data science requires neither of those things. In this domain, vast resources do not guarantee success. Smart resources do. Organizations of all sizes can succeed in their data science activities if they are implemented correctly by a competent team.

Myth #3: Data Science Is Just a Buzzword

Business leaders, journalists, and industry analysts are quick to use the latest jargon. The resulting noise can make it difficult to discern between industry hype and technologies or processes that can stand the test of time. Given the extreme hype about data science these days, it's not surprising that some consider it just another buzzword or fad.

Data science isn't a buzzword or fad, however. It's a confluence of time-tested disciplines, including statistics and forecasting, that have existed in some form for centuries. For example, actuaries and meteorologists have long used models to predict risks and weather, respectively. Now, businesses in virtually every industry are trying to use data to improve their performance.

A few things that distinguish data science from its predecessors, including actuarial science and statistics, are access to massive amounts of data that can be stored cheaply, robust computing power, and quick access to predefined models. Compared to yesteryear, organizations can learn more about themselves, their markets, and their customers than ever before because the data they need is plentiful, easily duplicated, easy to share, and relatively easy to process. Those capabilities, coupled with today's powerful programming environments, give developers considerable control over how data is manipulated, cleansed, preprocessed, analyzed, and visualized.

Myth #4: Complex Models Are Better Than Simple Models

Decision trees, statistical regression, and linear regression are not new, so the media pays less attention to them than deep learning and neural networks. Deep learning and neural networks use complex models that are considerably more sophisticated than the models used to solve simpler problems because they are attempting to emulate arbitrarily complex functions.

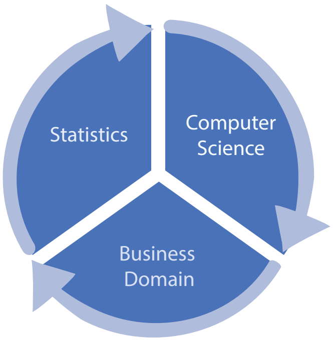
Complex models are not necessarily better than simpler models for a few reasons. First, a complex model can be less efficient than a simpler model if the problem is relatively simple. Second, complex models can be costly in terms of processing power. Finally, complex models can lead to black-box approaches that are difficult or impossible to explain. While the results of a black-box solution may be "good," black-box solutions don't allow users to explore how a result was derived. If users can't explore how a result was derived, they can't understand what went into it. If they can't understand what led to the result, they can't explain the details, which is not good, particularly in an audit scenario.

Simpler models are easier to understand and explain. For example, a relatively simple logistic regression model can be used to predict which of your prospects will likely buy your product.

A common mistake is to think that complex models necessarily yield better results than simple models in all situations. However, unnecessary complexity can result in diminishing returns. When that's the case, it's better to spend less time tweaking the model and more time understanding and cleansing the data.

Myth #5: Data Science Requires a Deep Understanding of Statistics and Statistical Methods

While it's true that data science requires an understanding of statistics, businesses can take advantage of data science without having a statistician on staff. Most developers have a basic understanding of statistics because they took at least one course in college.



If you're a developer who has been tasked with building data science capabilities in your organization, or you want to start building the capability yourself, it's wise to refresh or augment your knowledge of statistics so that you can understand the fundamentals commonly used to develop models.

You do not have to take a formal course. You do not have to pursue a graduate degree. The e-books and other resources referenced at the end of this white paper will help you understand the basics. Armed with that knowledge, you'll be able to build models that are meaningful to your organization.

If you want to modify the model later, you may need to learn a little bit more so you can understand how particular assumptions affect what you're doing.

Myth 6: Regulated Companies Can't Take Advantage of Data Science

Regulated companies have to be careful about the information they use and how they use it. However, those limitations do not mean regulated companies cannot take advantage of data science or build models.

For example, hospitals are using data science to improve patient care, emergency triage, and cost control. Similarly, companies in other regulated industries such as financial services, oil and gas, and pharmaceuticals are also benefitting from data science without using information that is prohibited by law.

Be mindful of inference, however. Your company may be prohibited from using certain types of information, such as personally identifiable information (PII), for specific purposes. It is nevertheless possible to infer sensitive information by combining other data points that are not restricted. Such uses could expose your company to regulatory fines and damages.

You can minimize the likelihood of such risks by avoiding unnecessary attributes that allow personal information to be inferred, which could be prohibited by law. For example, if it were illegal to use income as the basis for discrimination, one could nevertheless infer a person's approximate income level from her zip code, car make and model, etc.

Even if certain types of personal information are not prohibited by law, their use can be brand-damaging. For example, Forbes reported that Target inferred a teenage girl's pregnancy based on her purchasing habits. Based on that insight, Target sent relevant coupons to the girl's home address where they were discovered by her unsuspecting father.

Because inference can open the door to legal and other risks, organizations should understand what can be inferred by their data and what the associated risks are.

Myth #7: Data Science Tools Are Too Expensive

Some of the most sophisticated data science products are extremely costly to buy and difficult to use. However, it is not necessary to invest millions of dollars in software in order to benefit from data science.

For one thing, there are many open-source tools, such as R and Apache Spark, that are not difficult to set up and use. There are also a lot of commercial support options available for such tools, given their popularity.

There are also commercial products available that are far less expensive than traditional solutions. For example, Syncfusion has a deployment environment that can extract a model built with R or Spark so it can be used in a C# application.

Syncfusion also offers an Apache Hadoop distribution that allows developers to take advantage of proprietary and open-source models including machine learning models that are supported in the Apache Spark environment. R is a well-respected open-source product that offers great support for data processing, modelling, and display. You do not have to budget for expensive tools to take advantage of data science.

Myth #8: Data Science Requires Massive Computing Power

The big data and AI hype have created the impression that data science requires massively parallel GPU-accelerated machines or huge clusters. While large deep learning and neural networks do sometimes require that kind of computing power, many use cases do not.

Problems that can be solved with simple models may only require a PC with 64 GB or 128 GB of RAM. If that's not enough, two or three hours spent on a cloud may be all that's necessary to build and test a model. A cloud environment, such as AWS or Microsoft Azure, may also be necessary if the data processing or data cleansing requirements exceed the capacity of a single node.

Essentially, it's more cost effective to scale computing resources as necessary than to over-engineer a computing environment that is more complex and costly than the problem requires.

Myth #9: Data Cannot Be Monetized Because It's in a Hard-to-Use Format

Data-first companies such as Google and Facebook are masters at monetizing data. They have collected treasure troves of information that are sold to various parties at a handsome profit.

Some small to medium businesses think that data monetization is something only industry giants can do because they are data-first companies. However, most businesses have valuable customer data that could be used to improve company operations and perhaps drive new revenue streams. For example, most companies have transactional information, whether it's customer orders or credit card sales. They probably also have customer service records from their website or call center, and support tickets. Yet, many businesses aren't able to leverage that data effectively, let alone monetize it.

In fact, it's unclear what might be discerned from the data by modeling or analyzing it. Worse, the data may not be readily accessible because it's stored in various databases, on paper, or in business systems that have not been interconnected yet.

Part of the problem can be solved using a data integration platform. Using an integration platform, organizations are able to connect the dots, which means their insights transcend the data stored in any one system. Using that approach, organizations are in a better position to optimize business processes and customer journeys. Common connections include sales, marketing, and customer support, although that information can also be connected with supply-chain information and information from other systems, as appropriate.

Trend information, such as weather, traffic, and customer buying patterns are commonly bought and sold to improve sales, marketing, or operational effectiveness. The companies monetizing such data typically transform it so it can be consumed easily by other applications (which is part of what data integration platforms do). The data is then made available to third parties via APIs.

In short, data integration platforms lower barriers to information sharing and monetization.

Myth #10: Data Science Is Hard to Adopt Because It's Complicated

Data science can be a very complex undertaking, but it doesn't have to be. In fact, it's better to start simply, drive success with that, and then expand your capabilities.

Many organizations start by aggregating data they think is valuable, gleaning some insights from it, and pushing those insights out to decision-makers via reports and dashboards. Later, they start building models on top of the data to drive new and finer-grained insights.

Although there is no single “right” path to data science adoption, the wrong path is inevitably overcomplicating the problem when a simpler solution is more elegant, effective, and cost-efficient.

Conclusion

Data science doesn't have to be a complex and expensive undertaking that requires a formidable staff of Ph.Ds. The software development capabilities you have today can produce valuable insights that you once considered impossible without heavy investments in additional resources.

One way you can overcome your organization’s obstacles is to supplement your computer science and business domain expertise with a basic understanding of statistics so that you can start building models that benefit your organization. As the needs of your business grow, you can expand your knowledge to help move your company down a successful data science path.

Syncfusion can help.

Recommended resources

RESOURCE	LINK	DESCRIPTION
Andrew Ng's Machine Learning course on Coursera	www.coursera.org/learn/machine-learning	An excellent machine learning course with exercises suitable for any knowledge level, it uses the **GNU Octave** environment, which is similar to MATLAB but available for free. Syncfusion's **MATLAB Succinctly** is a great way to become familiar with this environment.
An Introduction to Statistical Learning	www-bcf.usc.edu/~gareth/ISL/	A great introductory textbook that builds learning from a statistical basis. This book uses the R environment. Syncfusion's book on R may help you get started.
Brandon Foltz's videos on Statistics and Probability		If you need more on the fundamentals of statistics, this is a good place to start. We also recommend **Statistics Fundamentals Succinctly**, published by Syncfusion.
Mathematical Monk's Machine Learning Playlist	www.youtube.com/watch?v=yDLKJtOVx5c&list=PLD0F06AA0D2E8FFBA	This definitely requires more mathematical understanding, but not too much. If you stick with the course, you will be rewarded with an excellent understanding of how things work.
Python Machine Learning by Sebastian Raschka	www.packtpub.com/big-data-and-business-intelligence/python-machine-learning	This is a great end-to-end book, especially if Python is your preferred language.
Turn the Wheel: Street-Smart Stats	www.turnthewheel.org/learn/street-smart-stats/	An online tutorial from Succinctly series author Katie Kormanik (Statistics Fundamentals Succinctly).

Select books from Syncfusion's *Succinctly* series that are useful when starting with machine learning.

RESOURCE	LINK	DESCRIPTION
Statistics Fundamentals Succinctly	www.syncfusion.com/resources/techportal/details/ebooks/Statistics-Fundamentals-Succinctly	By Katie Kormanik, it provides a foundation for the theories and methodologies behind statistical procedures.
Neural Networks Using C# Succinctly	www.syncfusion.com/resources/techportal/details/ebooks/neuralnetworks	By James McCaffrey, it teaches encoding and normalizing data, activation functions and how to choose the right one, and ultimately how to train a neural network to find weights and bias values that provide accurate predictions.
Machine Learning Using C# Succinctly	www.syncfusion.com/resources/techportal/details/ebooks/machine	Also by James McCaffrey, it covers several different approaches to applying machine learning to data analysis and prediction problems by demonstrating different clustering and classification techniques, and explaining the many decisions that must be made during development that determine how effective these techniques can be.
MATLAB Succinctly	www.syncfusion.com/resources/techportal/details/ebooks/matlab	By Dmitri Nesteruk, it explains the essential skills needed to use the flexible MATLAB system.

Total Sales

$787,328