Data Science: Executive Briefing

by Matthew Renze

Why Is Data Science Important?

Introduction

Hi. I'm Matthew Renze, data science consultant, author, and public speaker. Over the past 2 decades, I've taught over 150 thousand software developers and IT professionals how to make better decisions with data science. Today, I'd like to provide you with a quick introduce to data science geared towards executives and management. I'll provide you with all of the information you need to quickly get up to speed on this exciting new topic. The purpose of this executive briefing is to answer the following three questions. What is data science? Why is it important for you and your business? And how do I get started becoming a data- driven organization with data science? However, in order to keep things interesting, rather than starting with what data science is, we're going to start with why data science is important for you. My hope is that this will show you the value of data science before we go into the details of what it is and how it works. So let's get started.

Why Is Data Science Important?

So why is data science important? What is it about this approach to solving problems that has everyone so excited? To answer this question, I'd like to show you three examples of how data science solves common problems that exist in your business. Then, I'll show you a series of examples of how three distinct industries have been completed transformed by the emergence of data science.

Making Better Decisions

First, let's learn how we can use data science to make better decisions for our business. Imagine that you're sitting in a product development meeting. You're trying to decide if your company should create a new product or not. Some of your team members are arguing that the product will be a huge success, while others are arguing that it's completely unnecessary and will likely be a total failure. How do you, as a decision maker, decide whether to invest thousands or potentially millions of dollars developing this new product or not? In traditional decision making, to make this kind of decision, we'd typically rely on guesswork, gut instinct, anecdotal evidence, human biases, and, in many cases, the flip of a coin. We all know that these are less-than-ideal ways to make critical business decisions. However, many of us are unaware that there's a better way to make these key decisions. Data science solves these problems in an entirely different way. Rather than using guesswork or coin flips, we use the scientific method and data to provide evidence for or against our hypothesis. The scientific method is a process of answering questions that involves a few steps. We ask a question, form a hypothesis, design an experiment, collect data, analyze the data, draw a conclusion, and take action. This method is evidence-based, objectively measureable, reproducible, and transparent to everyone involved. It's a much better state of affairs than guesswork and coin flips. In fact, initial findings from MIT Research show that companies that are implementing data-driven decision making practices had 4% higher productivity and 6% higher profits than traditional businesses. Some examples of how we can use data science to make better decisions include using customer segmentation to create highly personalized marketing campaigns, predicting customer churn, and intervening to keep customers loyal to our brand, running A/B tests on our website to maximize click-through rates and customer conversions, performing sentiment analysis to determine how our customers respond to our products on an emotional level, and providing portfolio managers with recommendations on investments in asset diversification. The key difference here is that we're using data and the scientific method in order to help us make critical business decisions instead of traditional, flawed decision-making practices. This is just one of the many ways that data science can help our business.

Creating Smarter Products

Next, let's learn how we can use data science to create smarter products. Imagine you just bought a brand new coffee maker. You open the box, remove it from its packaging, and try to plug it in to the power outlet. However, you quickly discover that it has no power cable. In fact, it doesn't even run on electricity. Instead, you have to heat the water on the stove to brew the coffee. In addition, without temperature sensors to regulate the heat of the warming plate, the coffee always turns out tasting burnt. Without an electric timer, you can't automatically start brewing the coffee in the morning or automatically turn off the heat for safety. Oh and the coffee grinder, yep, you guessed it. Without electricity, you're stuck grinding that coffee by hand. You think to yourself, why does this have to be so difficult? Why couldn't they have just used electricity to make this product more convenient for me? You can imagine how frustrating this would be. Well, in the very near future, your customers are going to be equally frustrated if you aren't building data- driven artificial intelligence into your products to make them more convenient to use. Let's face it. Traditional products are dumb. By that, I mean they can't solve complex problems on their own, they don't anticipate their users' needs, and they can't provide us with insight into our customers, their behaviors, or how our products are actually being used. With data science, our products can become intelligent. We can use machine learning to create products that solve complex problems for our users automatically. We can use anticipatory design to anticipate the needs of our users, minimize the number of choices they need to make, and provide them with a user experience custom-tailored to their needs. In addition, we can use product telemetry data to better understand our customers, their preferences and behaviors, and how they actually use our products so that we can improve their design and operation. We can embed this intelligence directly into our products to provide real value to our customers. Some examples of products made smarter with data science include Google Home turning on the lights in your house based on a voice command, Apple's iPhone recognizing your face and automatically unlocking your smartphone, Nest thermostats learning your schedule and preferences and adjusting the temperature of your house accordingly, Netflix learning your movie-watching habits and creating playlists of movies that you're likely to enjoy, and Microsoft application insights providing you with statistics about your customers' interactions with your products in real time. With data science, we can make our products smarter. Smarter products can provide more value to our customers. In addition, they can delight our customers by going beyond what's expected of them. Finally, smarter products allow us to learn more about our customers and their behaviors by providing us with data on their interactions with our products. This allows us to improve the

design and operation of our existing products and create new products that will provide even more value to our customers.

Automating Manual Processes

Now let's learn how we can automate manual processes with data science. Imagine you're an employee reviewing credit applications. Each day you clock in at 8:00 a. m., review the same credit application forms all day long, and clock out at 5:00 p. m. It's definitely not the work you were hoping to do with your education and certainly not what you want to be doing for the rest of your life. Unfortunately, this work needs to be done, and there's someone willing to pay you to do it simply because there's no alternative. Also, as an employer, you would much rather have this employee working on tasks that can provide the company with more value than reviewing paper forms all day long. In addition, it costs quite a bit of money to pay this person to review these credit applications every day, money that could definitely be better spent elsewhere. However, the job needs to be done, and currently it requires a human to perform it. The first problem here is that for employees, repetitive mental tasks are boring, uninspiring, and highly prone to errors. And repetitive physical tasks have all of the same issues, but can be physically exhausting and lead to repetitive stress injuries. The second problem here is the high cost of labor. For most business, labor is our biggest expense. So we want our employees to be working on tasks that will provide the most value for our company. The simple fact is that much of the work that we do every day isn't enjoyable or cost-effective. However, data science has led to the emergence of a new era of automation, automation based on data-driven AI. In the same way that we automated manual work on assembly lines with industrial robots, data science is automating knowledge work with a new set of tools. With machine learning, deep learning, reinforcement learning, and artificial intelligence, we are now able to automate a variety of tasks that were previously only able to be completed by humans. This frees us humans to work on tasks that they actually want to be working on, tasks that are creative, complex, interpersonal, and meaningful. In addition, it allows our business to significantly lower our cost of human labor. Machines can run 24 hours a day, 7 days a week without needing breaks, vacations, or employee benefits. Finally, it can reduce the number of defects in our work caused by human error, lack of precision, and miscommunication. Some examples of automation made possible by data science include automatically approving credit applications based on a customer's demographic data and financial history, automatically detecting fraudulent credit card transactions before the transaction has completed, automating tech support with chatbots that can answer frequently asked questions, automating quality control with machines that can inspect products for visible

defects, and automating simple programming tasks by applying deep learning to program synthesis. With data science, we now have the ability to automate tasks that were previously only the domain of humans. This can be good for our business, as well as for our employees. However, there will likely be significant social and economic side effects caused by this level of automation that we, as a society, will need to address over the next few decades.

Industry Examples

Finally, to wrap things up for this module, let's take a look at a few industries that have been completely disrupted by the adoption of data science. My hope is that this will help you see what a data science transformation may look like within your own industry. First, the financial industry. Just a few decades ago, the trading floors of Wall Street and other exchange markets were filled with people yelling and screaming to buy or sell stocks, bonds, and other securities. At the intersection of every trade were people using a combination of information, guesswork, and gut instinct to make trading decisions. However, the trading room floors are almost completely empty these days. Stock trading was taken over by quantitative analysts sitting in back offices using data science to predict the performance of their investments. In recent years however, these quants, as they're called, have replaced themselves with automated trading algorithms. Today, lightning-fast, high-frequency trading algorithms now account for the majority of all trades in the modern stock market. As a result, the financial industry has been changed forever by data science. Next, we have baseball. Prior to 2002, picking a lineup of players in a Major League baseball draft relied largely upon traditional measures of player performance, like stolen bases, runs batted in, and batting averages. In addition, the scouting profession was full of human biases that distorted the true value that each player brought to the baseball team. However in 2002, the Oakland Athletics baseball team used data science to select players that were undervalued by the market. Through rigorous statistical analysis, the Oakland As discovered that metrics like on-base percentage and slugging percentage were better indicators of offensive success. Despite their significantly smaller budget for player salaries, the Oakland As created a team that was competitive against the must better funded teams. This data-driven strategy contributed to the Oakland As setting a record 20- consecutive wins and brought them to the playoffs in 2002 and 2003. As a result, the game of baseball and how players are valued has been changed forever by data science. Finally, let's look at retail sales. We all know the impact that the internet has had on retail sales. We've seen the rise of online stores and the decline of brick-and-mortar stores for almost 2 decades now. However, what most people don't know is that the internet was just the beginning for this trend. Unlike transactions that take place in a physical store, transactions that

occur on the internet provide a massive trail of data we call data exhaust that can be mined for actionable insight. Online retailers have leveraged these data to revolutionize the world of retail sales. They are able to recommend products based on your previous shopping history, your geographic location, and upcoming holidays. They can use these product purchases to predict major life events, like when you're getting married or when you're about to have a baby. And they can run fine-grained experiments in order to increase the likelihood of impulse purchases. In addition, major online retailers are now in the process of automating their warehouses. By using autonomous warehouse robots, they can eliminate the need for many human jobs and significantly reduce their operating costs while dramatically increasing their efficiency. In the future, it's anticipated that these highly automated warehouses will involve little to no human labor. As we can see, the world of retail sales is currently being changed forever by data science.

What Is Data Science?

What Is Data Science?

So now that we know why data science is important, we need to answer the next question. What is data science? Data science is an interdisciplinary field composed of computer science, math and statistics, and domain knowledge that seeks to derive insight from data. Data science is the intersection of these three respective disciplines. Another way to think about it is that data science is the intersection of data engineering and the scientific method. With data science, we're using large-scale data systems to drive the scientific method. The goal of data science is to transform data into knowledge, knowledge that can be used to make rational decisions so that we can take actions that help us achieve our goals. We refer to this process as transforming data into actionable insight.

The Data Science Team

So what is a data scientist? A data scientist is someone who performs data science as a profession. However, a data scientist is more than just a scientist, more than just a data analyst, and more than just a software developer. They possess all three sets of these skills. Individuals with all three sets of skills and the proper credentials are currently very rare in our industry. In small organizations, a data scientist may work independently within the organization, moving from project to projects as needed and working to develop a culture of data science within the

organization. However, within larger companies, a data scientist is often part of a data science team. This team will work together to solve business problems with data science. Other roles you may find on a data science team include a data engineer, someone who has a deep understanding of the data, the database, and the infrastructure who can find, transform, and clean data. And a subject matter expert. Someone who has a deep understanding of the business who can provide guidance in context from a business perspective for the data, the analysis, and the results. You may also find other roles, including a project manager to oversee the project and remove impediments for the team, software developers to implement production code for data science products, or designers to create user interfaces and user experiences for data science products. However, it's important to know that data science doesn't only belong to the data science team. Everyone in your organization can use data science to make better decisions, to create smarter products, and to automate manual processes. While there are some tasks that should only be performed by trained data scientists, basic data science skills to improve the work that they do. So if you want to become a data-driven organization, it's important to provide basic data science training to all of your employees.

Skills of Data Science

The next question we need to address is what set of skills are necessary for data science? In general, the skills commonly associated with data science are programming computers using programming languages, like SQL, Python, and R, working with data, that is collecting, cleaning, and transforming data, creating and interpreting descriptive statistics, that is numerically analyzing data, creating and interpreting data visualizations, that is visually analyzing data, creating statistical models and using them for statistical inference, hypothesis testing, and prediction, handling big data, data sets that are of volume, velocity, or variety beyond the limitations of conventional computing architecture, automating decision-making processes using machine learning algorithms, and deploying data science solutions into production or communicating results to a wider audience.

Tools of Data Science

So what tools are used to perform data science? We'll take a look at the most popular programming languages, relational databases, big data platforms, spreadsheets, reporting, and business intelligence tools. For programming languages, the most popular language is SQL, or structured query language. SQL is a very important language in data science because of how

much time data scientists spend exploring, transforming, and cleaning data. Next, we have Python followed by R. R and Python are the two workhorses of the data science world. Most data scientists will generally use either one or the other in addition to SQL to perform the majority of all data science tasks. However, they each have their own strengths and weakness, so I recommend that data scientists learn both. For relational databases, we have MySQL followed by Microsoft SQL Server and then PostgreSQL. MySQL has become very popular for data science because it's a free, open-source database that's easy to use, easy to maintain, and runs on a variety of operating systems. For big data platform, we have Spark, Hive, MongoDB, and Amazon Redshift. Spark has become the big data platform of choice for data scientists in recent years because of its speed, scalability, ease of use, and open source license. Finally, we have spreadsheets, business intelligence tools, and reporting tools. The most popular tool in this category by far is Excel. Many data scientists have a general dislike for Excel and prefer more powerful and statistically rigorous tools for working with data. However, the reality is that Excel is very easy to use and quite powerful. As a result, much of the business world today is still run on Excel spreadsheets.

The Data Science Process

Finally, we need to address the question, how is data science performed? The general data science process works like this. First, we find a question that we want to answer. This can be a hypothesis we want to test, a decision we want to make, or something we want to attempt to predict. Second, we collect the data for our analysis. Sometimes this means designing an experiment to create new data. Other times, the data already exists. We just need to find them. Third, we prepare the data for analysis, a process often referred to as data munging or data wrangling. We need to clean and transform these data to get them into a form suitable for analysis. Unfortunately, this step often accounts for roughly 80% of the work in data science. Fourth, we create a model from our data. In the most generic sense, this could either mean a numerical model, a visual model, a statistical model, or a machine learning model. We use this model to provide evidence for or against our hypothesis, to help us make a decision, or to predict an outcome. Fifth, we evaluate the model. We need to determine if the model answers our question, helps us make a decision, or creates an accurate prediction. In addition, we need to make sure that our model is appropriate given our data and the context surrounding our data. Finally, if everything looks good, we deploy our model into production. This could mean communicating the results of our analysis to others, making a decision and acting upon our decision, or deploying an application into a production environment. Then, we repeat this process

for each business problem in our data science backlog. We perform this process in a highly iterative and agile manner. We use feedback from our previous results to help guide our process moving forward.

How Do I Get Started?

Ingredients of a Data-driven Organization

So now you might be wondering, how do I get started with data science? How do I begin leveraging data science on my team and within my organization? First, what are the ingredients of a data- driven organization? What is in the DNA of companies like Google, Facebook, and Amazon that make them so successful with data and insights. In general, there are five key ingredients to data-driven enterprises. Let's take a look at each of these in more detail. First, we need a strategy for becoming a data-driven organization. Our data science strategy defines key aspects of how we do business, like what people we need on our data science team, what data to collect and how it's collected, what technology will be required to achieve our goals, and how we will grow a culture of data science to become a data-driven organization. Second, we need the right people to help us execute this strategy. We want people that value evidence-based reasoning, that think like scientists and can use data to drive business outcomes. This means acquiring talent with a background in data science and training our existing employees to think more like data scientists. The third ingredient is data. We need the right data in order to derive actionable insight. In addition, we need our data to be accurate, reliable, timely, and available. So we need to invest time and energy collecting, cleaning, storing, and organizing our data. Data is an investment in the future of our business, so we need to treat it like a strategic asset. Fourth, we need the right technology. Technology is necessary to build our data science pipeline, to analyze our data, and to automate decision-making processes. Choosing the right technology can make a big difference in the long-term costs, benefits, and overall return on investment of our data science strategy. Finally, we need to create a data-driven culture. We need to change the mindset of our organization to one that values experimentation, evidence-based reasoning, and data-driven decision making. And this is probably the most important of the five ingredients. However, it's also the most difficult because it can't be bought, acquired, or hired. It has to evolve over time. We do this by encouraging certain data- driven behaviors and discouraging other non data-driven behaviors.

Data Science Hierarchy of Needs

So what is the process for becoming a data- driven organization? Unfortunately, we can't just throw all of these ingredients together and expect a data-driven enterprise to emerge fully baked. Data-driven organizations are grown and evolve over time. They go through various stages of growth as they reach maturity. These stages are based upon a hierarchy of data-driven needs. Essentially, we can't get to the next stage of development until we've sufficiently satisfied our lower, more primal needs. The most basic need of a data-driven organization is the need to collect data. This starts with basic data collection activities, for example recording transactions, logging errors, and digitizing analog data. Then as the company evolves, this can lead to more advanced forms of data collection, for example gathering telemetry data from applications running in production, running experiments to create new data, and acquiring data from external sources. Next, we have a need to organize our data. We need to get our data into a form suitable for analysis. This starts with basic data organization tasks, for example transforming, cleaning, and storing data. Then as the company matures, this may often lead to building more robust solutions, for example creating a data ETL pipeline, a data warehouse, or a data lake. Third, we have a need to analyze our data. We need to use our data to explain what's happening in our organization and why it's happening. This generally starts with basic data analysis tools, for example reports, dashboards, and KPIs. Then, as the company matures, this can lead to more powerful forms of data analysis, for example data mining, descriptive analytics, and diagnostic analytics. Fourth, we have a need to make predictions. We want to know what will likely happen in the future and how we should respond to these potential scenarios should they occur. This generally involves more advanced types of data analysis, for example predictive analytics, prescriptive analytics, and machine learning. Finally, we have a need to automate. This is where we close the data science loop and remove the human from the process. This involves advanced technologies, like artificial intelligence, deep learning, and reinforcement learning. Data science is essentially a stepping stone on the road to data-driven artificial intelligence. However, in order to become an AI-driven organization, we first to become a data- driven organization. We need to cultivate organizational knowledge and adoption of these core data science practices before we can achieve the transformative effects of modern, artificial intelligence.

Next Steps

So what's the next step? How do you get your team and your organization started on the journey to becoming a data-driven organization? Fortunately, we have some great resources available to help you get started. First, I recommend that you and your employees begin by watching online

courses on data science. Start by having everyone watch Data Science: The Big Picture. This short course will cover what data science is, why it's important, and how to get started for a general audience. Next, for more hands-on training, I recommend watching Data Science with R. This course will walk you through each step in the data science process using the programming language R. Beyond this, Pluralsight has a variety of other courses on data science topics, like data analysis, data visualization, and more advanced topics, like deep learning. You can find all of my courses on data science at the link below. Next, while online courses work great for teaching a variety of topics, they may not be appropriate for all topics, audiences, and situations. Some scenarios require more formal classroom-style training. For these cases, I recommend either onsite or offsite training. For example, I frequently teach full-day and multi-day workshops on data science, machine learning, client locations and off site at conferences and events around the world. Kicking off a data science initiative first with classroom-style training and following up with online training can be a powerful combination for organizational transformation. Third, taking the first step on your data science journey is often the most difficult part. So you may want to consider hiring a data science consultant or advisor to help you get started. For example, I often help companies develop their data science strategy, hire the right people, design data pipelines, choose the right technology, and begin cultivating a data science culture. You might find similar services beneficial to help you get started as well. Finally, before we wrap things up for this course, feedback is very important to me. I use your feedback to improve each and every one of my courses. So please be sure to rate this course, ask questions in the discussion board, leave comments to let me know what you found valuable and what you think could be improved, and feel free to send me a tweet on Twitter if you'd like to provide me with feedback in public. My twitter handle is @matthewrenze.

Conclusion

Today we answered three important questions about data science. We learned what data science is, why it's important for you and your business, and how you can get started becoming a data-driven organization. I believe that we're just beginning to see the true potential of data science. Data science is going to have a fundamental and profound impact on software and the IT industry for many years to come. So to end, I'd like to propose three new questions for you. Are you prepared for this, is your team prepared, and is your organization prepared? If you, your team, or your organization would like help preparing for the coming data science revolution, please reach out to me. I would love to help. Thank you. We hope you enjoyed this course. If you're interested in more content for technical leaders, content we keep short and focused with the up-to-date

information you need to be informed and make decisions, but without getting buried in the details, find other courses like this at plrsig. ht/exec.

Course author

Matthew Renze

Matthew is a data science consultant, author, and international public speaker. He has over 17 years of professional experience working with tech startups to Fortune 500 companies. He is a...

Course info

Level	Beginner
Rating	★★★★½ (90)
My rating	★★★★★
Duration	0h 27m
Released	16 Nov 2018

Share course